BIOS 6321: Clinical Trials and Experimental Design

Tatsuki Koyama, PhD
Department of Biostatistics, Vanderbilt University School of Medicine
tatsuki.koyama@vumc.org

Spring 2020

Last updated: July 15, 2020

Updated: 7/15/2020 21:20
R version: 4.0.1

Contents

Chapter 1

Introduction: Observational studies and experiments

1.1 Observational studies

Observational study A study design in which the investigator does not control the assignment of treatment of individual study subjects (Piantadosi[1])

Experiment A study in which the investigator makes a series of observations under controlled/arranged conditions. In particular, the investigator controls the treatment applied to the subjects by design. (Piantadosi)

Clinical trial A prospective study comparing the effect and value of an intervention against a control in human subjects (Friedman[2])

Advantages of observational studies include:

- Lower cost.
- Greater timeliness.
- A broad range of patients.
- Greater application where experiments would be impossible or unethical.

1.1.1 Example: PCOS

Prostate cancer is the second leading cause of cancer death among American men behind lung cancer. The common treatment choices for localized disease are surgery, radiation, and observation.

[1] Clinical Trials: A Methodologic Perspective
[2] Fundamentals of Clinical Trials

Suppose we are interested in comparing effectiveness of surgery and radiation therapies.

The Prostate Cancer Outcomes Study (PCOS): Subjects were identified through six sites participating in the NCI's SEER (`https://seer.cancer.gov/`) program. (diagnosed with Prostate cancer from 1994/10/1 to 1995/10/31).

```
dim(SR)
```

```
[1] 2091   71
```

```
table(SR$Trt, SR$vital)
```

```
           Alive Dead
  RP only   1034  407
  XRT only   272  373
```

```
prop.table(table(SR$Trt, SR$vital), margin = 1)
```

```
            Alive   Dead
  RP only  0.7176 0.2824
  XRT only 0.4217 0.5783
```

Can we conclude that surgery is better?

```
chisq.test(table(SR$Trt, SR$vital), correct = FALSE)
```

```
Pearson's Chi-squared test

data:  table(SR$Trt, SR$vital)
X-squared = 167, df = 1, p-value <2e-16
```

```
f <- fisher.test(table(SR$Trt, SR$vital))
f$estimate
```

```
odds ratio
     3.481
```

```
f$conf.int


[1] 2.856 4.249
attr(,"conf.level")
[1] 0.95
```

In general, establishing a cause-and-effect association from an observational study is difficult because of **confounders**.

Confounder A prognostic factor that is associated with both response (e.g. survival) and explanatory variable (e.g. treatment choice).

	N	RP only $N = 1445$	XRT only $N = 646$	Test Statistic
stage	2089			χ^2_1=37.51, P<0.001[1]
Clinically Localized		98% (1421)	93% (602)	
Metastatic		2% (23)	7% (43)	
grade	1849			χ^2_2=34.3, P<0.001[1]
1		74% (936)	61% (358)	
2		20% (255)	28% (163)	
3		6% (73)	11% (64)	
race	2091			χ^2_2=12.87, P=0.002[1]
NH White		67% (962)	74% (480)	
NH Black		18% (253)	14% (92)	
Hispanic		16% (230)	11% (74)	
psa.value	1986	5.0 6.8 10.3	6.0 8.6 14.5	$F_{1,1984}$=60.5, P<0.001[2]
age	2091	57 62 67	64 70 74	$F_{1,2089}$=420.4, P<0.001[2]
vital	2086			χ^2_1=166.6, P<0.001[1]
Alive		72% (1034)	42% (272)	
Dead		28% (407)	58% (373)	

$_a\, b\, _c$ represent the lower quartile a, the median b, and the upper quartile c for continuous variables. N is the number of non–missing values. Numbers after percents are frequencies. Tests used: [1]Pearson test; [2]Wilcoxon test

Table 1.1: PCOS study

Let's analyze the data with a method that accounts for the baseline difference in the two treatment groups.

Survival $\sim$ Treatment $*$ (Age + PSA + Stage + Tumor grade)

How about race? comorbidities? sex? smoking?

Many statistical methods exist to establish causal relationship from an observational study such as propensity scores and instrumental variables.

Can observational studies establish a cause-effect association?

- `https://www.jti.com/about-us/our-business/our-attitude-to-smoking`
- `https://www.pmi.com/our-business/about-us/our-views/health-effects-of-smoking-tobacco`

Older versions:

JTI The Ministry of Health's claim that smoking is a risk factor for many diseases is primarily based on epidemiological studies of comparisons between smokers and non-smokers on disease rate. Epidemiological studies are useful in establishing exploratory associations between a disease and risk factors, but they can not establish a cause-and-effect association without controlling for other factors such as genetic factors, diet, exercise and stress. Moreover, epidemiological studies are intended to compare populations and do not reveal the risk of disease for individual smokers.

$\cdots$ no matter how smoking is described, people can stop smoking if they are determined to do so. No one should believe that they are so attached or 'addicted' to smoking that they cannot quit.

PM Smoking causes many serious diseases including cardiovascular disease (heart disease), lung cancer, and chronic obstructive pulmonary disease (emphysema, chronic bronchitis). Smokers are far more likely to become sick with one of these diseases than non-smokers. Smoking is also addictive and can be extremely difficult to stop. These are the views of every leading medical and scientific organization around the world. And they are the views of Philip Morris International.

1.2 Clinical trials

For a study to be considered as a clinical trial, it *must*:

- be designed.
- involve human subjects.
- involve intervention.
- have prospective follow-up for a specified outcome.

It *usually* has comparable treatment groups.

Examples of study designs that are *not* clinical trials are:

- General observational studies.
  - Cross-sectional studies
  - Case-control studies.
  - Case report.
- Animal studies.

Human studies typically have the following characteristics:

- Large variation among subjects.
- Lengthy disease process.
- Non-compliance and dropouts.
- Ethical issues.
- Rare disease.

Finally, *the* advantage of clinical trials is that it can establish a cause-effect association (free of confounding).

1.2.1 Example: PIVOT

In Prostate Cancer Intervention Versus Observation Trial[3] (PIVOT), prostate cancer patients who were good candidates for radical prostatectomy were enrolled from 1994 to 2002. The last observation was made in 2010. The results were presented at American Urological Association Annual Meeting in May, 2011. The *inclusion criteria* for the study were:

- 75 years or younger.
- Localized disease.
- PSA $\leq 50mg/mg$.
- Diagnosed with 12 months.
- Radical prostatectomy candidate.

With the all-cause mortality as the primary endpoint, the primary objective was to answer the following question:

> Among men with clinically localized prostate cancer detected during the early PSA era, does the intent to treat with radical prostatectomy reduce all-cause & prostate cancer mortality compared to observation?

- $13,022$ men entered into screening registry.
- $5,023$ were eligible.

[3] Wilt et al. (PIVOT Study Group). "Radical prostatectomy versus observation for localized prostate cancer". *N Engl J Med*. 2012. **367**(3):203–213.

- $4,292$ declined to participate.
- 731 were randomized. (364 prostatectomy, 367 observation)
- Radical prostatectomy was performed on 281 (77%) of the prostatectomy group and 36 (10%) of the observation group. The following table summarized the assigned and received treatments.

	Actual Treatment			
Assigned Treatment	Surgery	Observation	Other	
Surgery	281 (77%)	53 (15%)	30 (8%)	364
Observation	36 (10%)	292 (80%)	39 (11%)	367
	317 (43%)	345 (47%)	69 (9%)	731

Intention-to-treat analysis compares 364 surgery patients and 367 observation patients based on their assigned treatments.

As-treated analysis compares 317 surgery patients and 345 observation patients based on their received treatments.

Per-protocol analysis compares 281 surgery patients and 292 observation patients who adhered to the protocol.

Conclusions: "Among men with localized prostate cancer detected during the early era of PSA testing, radical prostatectomy did not significantly reduce all-cause or prostate-cancer mortality, as compared with observation, through at least 12 years of follow-up. Absolute differences were less than 3 percentage points."

1.3 Phases of clinical trials

Clinical trials are usually classified into phases (I to IV). This classification is increasingly inadequate as objectives and characteristics of these phases have become less distinguished.

1.3.1 Phase I

The main objective of phase I clinical trial is to establish safety by estimating the maximum tolerable dose (MTD). The MTD is the highest dose that results in dose-limiting toxicities (DLT) with a preset low probability (e.g., 33%). These studies provide information on the *pharmacokinetics* and *pharmacodynamics* of the drug in humans.

Pharmacokinetics What the body does to drug. The process by which the drug is absorbed, distributed, metabolized, and eliminated by the body. Some commonly used parameters to study pharmacokinetics are: Concentration of drug; Biological half-life, C_{max}, the peak plasma concentration of a drug; t_{max}, time to achieve C_{max}.

Pharmacodynamics What drug does to the body. Effects of drugs on living organisms and systems.

Subjects for phase I study are usually normal healthy volunteers. In cancer studies, patients often participate. Single arm dose escalation (dose determination) trials are common in phase I cancer trials. Historically, $3+3$ designs have been frequently used; however, Bayesian designs such as the continual reassessment method (CRM) and the modified toxicity probability interval (mTPI) design are gaining popularity.

1.3.2 Phase II

Phase II trials primarily look for evidence of efficacy (drug activity); however, safety should also be closely monitored. Sometimes they are divided into phase IIa (safety as the primary goal) and phase IIb (efficacy as the primary goal) trials.

Phase II trials in cancer are often single-arm trials, where the response rate is compared to a historical control. Two-stage and multi-stage designs are sometimes applied to expedite a decision of futility.

1.3.3 Phase III

Phase III clinical trials are usually considered pivotal, and the new treatment is compared to the standard treatment of placebo to establish its effectiveness. These trials usually involve many sites (multi-center) sometimes spanning more than one country (multi-national). When there is already an effective conventional treatment, establishing *noninferiority* -as opposed to *superiority*-, is the primary objective.

1.3.4 Phase IV

Phase IV clinical trial is a postmarketing surveillance, often to look for uncommon but serious side effects.

Phase I/II and phase II/III trials have become popular.

Chapter 2

Selected topics in statistics

2.1 Confidence interval for a binomial proportion

- Asymptotic method.
- Wilson score method. (See wikipedia for the formula.)
- Clopper-Pearson (exact) method.

Suppose $X = 8$ is observed from a Binomial$(32, p)$ distribution.

Exercise: Compute 90% confidence intervals for p using these three methods.

```
## Asymptotic method
phat <- x/n
z <- qnorm(0.95)
phat + c(-1, 1) * z * sqrt(phat * (1 - phat)/n)
```

```
[1] 0.1241 0.3759
```

```
## Wilson score method
(1/(1 + z^2/n)) * (phat + z^2/(2 * n) + c(-1, 1) * z * sqrt(phat * (1 -
    phat)/n + z^2/(4 * n^2)))
```

```
[1] 0.147 0.392
```

Given x and N, the Clopper-Pearson confidence interval is (p_L, p_U), where p_L and p_U satisfy the

following:

$$P[X \geq x | p = p_L] = \alpha/2$$
$$P[X \leq x | p = p_U] = \alpha/2$$

This confidence interval contains p^* values such that $H_0^* : p = p^*$ would not be rejected by the observed data. p_L satisfies $P[X \geq 8 | p = p_L] = 0.05$, and p_U satisfies $P[X \geq 8 | p = p_U] = 0.05$.

```
f1 <- function(p.star, x = 8, n = 32, alpha = 0.05) 1 - pbinom(x - 1, n,
    p.star) - alpha
## Note 1-pbinom(x-1, n, p.star) = sum(dbinom(x:n, n, p.star))
uniroot(f1, lower = 0, upper = 1)$root
```

```
[1] 0.1309
```

```
f2 <- function(p.star, x = 8, n = 32, alpha = 0.05) pbinom(x, n, p.star) -
    alpha
## Note pbinom(x, n, p.star) = sum(dbinom(0:x, n, p.star)
uniroot(f2, lower = 0, upper = 1)$root
```

```
[1] 0.4061
```

To solve for p^* in a more straightforward manner, we can use the following relationship between a Binomial random variable and a Beta random variable. If $X \sim Binomial(n, p)$ and $Y \sim Beta(k, n - k + 1)$, then

$$P[X \geq k] = P[Y \leq p].$$

For the lower bound, instead of solving for p_L in $P[X \geq 8 | p = p_L] = 0.05$ iteratively, we can solve $P[Y \leq p_L] = 0.05$, where $Y \sim Beta(8, 32 - 8 + 1)$.

```
qbeta(0.05, 8, 32 - 8 + 1)
```

```
[1] 0.1309
```

For the upper bound, we needed to solve $P[X \leq 8 | p = p_U] = 0.05$, which is $P[X \geq 9 | p = p_U] = 0.95$. Instead, we can solve $P[Y \leq p_U] = 0.95$, where $Y \sim Beta(9, 32 - 9 + 1)$.

```r
qbeta(0.95, 9, 32 - 9 + 1)
```

```
[1] 0.4061
```

```r
binconf(x = 8, n = 32, alpha = 0.1, method = "all")
```

```
           PointEst  Lower  Upper
Exact          0.25 0.1309 0.4061
Wilson         0.25 0.1470 0.3920
Asymptotic     0.25 0.1241 0.3759
```

2.2 Stop when significant

Suppose we would like to test $H_0 : p = 0.10$; $H_1 : p > 0.10$ by taking a random sample of size 40 from a $Bernoulli(p)$. Under H_0, the number of successes out of 40 has a $Binomial(40, 0.10)$ distribution. Moreover, we compute $P_0[X \geq 8] = 0.04$, so the test that rejects H_0 if $X \geq 8$ has a type I error rate of 0.04.

The data were observed sequentially, and the 8th success was observed after $N = 32$, and we decided to reject H_0 and terminate the study. The conclusion was that p was significantly greater than 0.10, and $\tilde{p} = 8/32 = 0.25$.

Questions

- Was type I error rate controlled?
- Was $\tilde{p}$ unbiased?

$\tilde{p}$ is an estimator such that

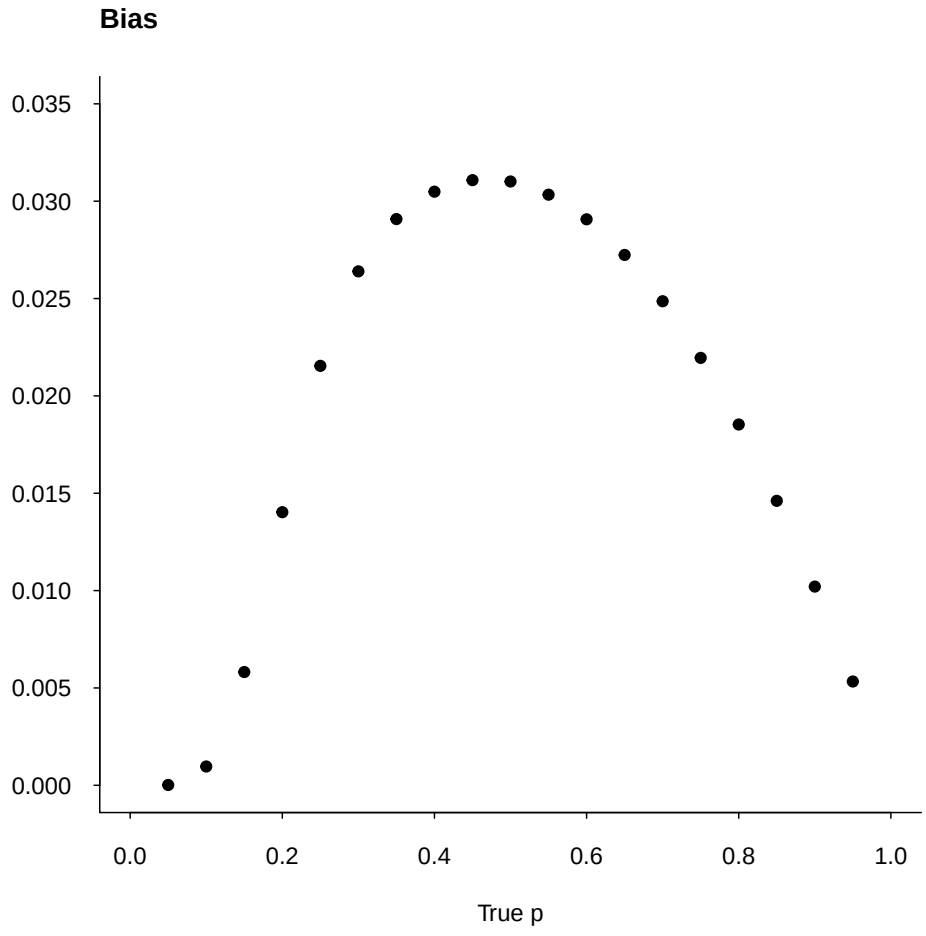
$$\tilde{p} = \begin{cases} 8/Y & \text{if } Y \leq 40, \\ X/40 & \text{if } X < 8, \end{cases}$$

where Y is the number of trials when 8th success happened. (Y has a distribution)

$$E_p[\tilde{p}] = \sum_{y=8}^{40} \frac{8}{y} P_p[Y = y] + \sum_{x=0}^{7} \frac{x}{40} P_p[X = x]$$

$$E_p[\tilde{p}] - p > 0$$

```r
theoretical <- function(p) {
    obs <- 8
    N <- 40
    x <- 0:(obs - 1)
    one <- sum(x/N * dbinom(x, N, p))
    y <- obs:N
    two <- sum(obs/y * dnbinom(y - obs, obs, p))
    ## dnbinom(x,k,p) ; x is number of failures until k-th success.
    ## dnbinom(8,4, 0.2) = choose(8+3,3) * 0.2^3 * 0.8^8 * 0.2
    one + two
}

trueP <- seq(0.05, 0.95, by = 0.05)
p.tilde <- sapply(as.list(trueP), theoretical)
```



This idea is related to *stochastic curtailment*.

2.3 Repeat until significance

Suppose X is a random sample of size 10 from $Normal(\mu, \sigma^2)$. If $H_0 : \mu = 0$ is not rejected, we will take another sample of size 15 and test again with $n = 25$. If each test has one-sided $\alpha = 0.05$, what is the actual type I error rate of this procedure?

- If we have k (independent) significance test of size α, the probability of at least one false positive result is $1 - (1 - \alpha)^k$. For $\alpha = 0.05$,

k	1	2	3	4	5	10
P[false positive]	0.05	0.0975	0.143	0.186	0.226	0.401

Because the first and second tests are not independent, we need to consider a conditional distribution of Z_t given $Z_1 = z_1$.

Under H_0,

$$\begin{aligned} \bar{X}_1 &\sim N\left(0, \sigma^2/n_1\right), & \bar{X}_2 &\sim N\left(0, \sigma^2/n_2\right). \\ Z_1 &= \sqrt{n_1}\bar{X}_1/\sigma & Z_2 &= \sqrt{n_2}\bar{X}_2/\sigma \\ &\sim N(0,1), & &\sim N(0,1). \end{aligned}$$

Let $n_t = n_1 + n_2$ and write

$$\begin{aligned} Z_t = \frac{\sqrt{n_t}}{\sigma}\bar{X}_t &= \frac{\sqrt{n_t}}{\sigma}\frac{n_1\bar{X}_1 + n_2\bar{X}_2}{n_t} \\ &= \frac{\sqrt{n_t}}{\sigma}\frac{\sigma\sqrt{n_1}Z_1 + \sigma\sqrt{n_2}Z_2}{n_t} \\ &= \frac{\sqrt{n_1}Z_1 + \sqrt{n_2}Z_2}{\sqrt{n_t}} \end{aligned}$$

Therefore, given $Z_1 = z_1$,

$$Z_t = \frac{\sqrt{n_1}}{\sqrt{n_t}}z_1 + \frac{\sqrt{n_2}}{\sqrt{n_t}}Z_2,$$

and $Z_t > c$ is equivalent to

$$Z_2 > \frac{\sqrt{n_t}}{\sqrt{n_2}}c - \frac{\sqrt{n_1}}{\sqrt{n_2}}z_1.$$

The *conditional type I error rate given* $Z_1 = z_1$ *is*

$$P_0\left[Z_2 > \frac{\sqrt{n_t}}{\sqrt{n_2}}c - \frac{\sqrt{n_1}}{\sqrt{n_2}}z_1 \middle| Z_1 = z_1\right].$$

And by integrating this conditional type I error rate with respect to the distribution of Z_1, we get unconditional type I error rate for the second stage (α_2):

$$\alpha_2 = \int_{-\infty}^{c}\left[1 - \Phi\left(\frac{\sqrt{n_t}}{\sqrt{n_2}}c - \frac{\sqrt{n_1}}{\sqrt{n_2}}z_1\right)\right]\phi(z_1)dz_1.$$

If we let $n_1/n_2 \to 0$, the above expression tends to

$$\alpha_2 = \int_{-\infty}^{c}\left(1 - \Phi(c)\right)\phi(z_1)dz_1 = \left(1 - \Phi(c)\right)\Phi(c).$$

And if $\alpha = 0.05$, i.e., $c = 1.645$, $p_2 = 0.05 \times 0.95 = 0.0475$.

Now, $n_1 = 10$ and $n_2 = 15$, and

$$\alpha_2 = \int_{-\infty}^{1.645}\left[1 - \Phi\left(\frac{\sqrt{25}}{\sqrt{15}}1.645 - \frac{\sqrt{10}}{\sqrt{15}}z_1\right)\right]\phi(z_1)dz_1$$

```
inside <- function(z1, n1, n2, cv) {
    nt <- n1 + n2
    (1 - pnorm(sqrt(nt/n2) * cv - sqrt(n1/n2) * z1)) * dnorm(z1)
}

additional.error <- function(n1, n2, cv) {
    integrate(inside, lower = -Inf, upper = cv, n1 = n1, n2 = n2, cv = cv)$value
}

additional.error(n1 = 10, n2 = 15, cv = qnorm(0.95))
```

```
[1] 0.03325
```

2.4 Divide by the control mean

Suppose we are interested in comparing two treatment groups. Take random samples of size 9 from each group ($x_{11}, \cdots, x_{19}$; $x_{21}, \cdots, x_{29}$). We want to test $H_0 : \mu_1 = \mu_2$. But we may be worried

that the background noise is different for these two groups, so we take random samples of size 3 representing the background. c_{11}, c_{12}, c_{13} and c_{21}, c_{22}, c_{23}. Then we may "normalize" x's by dividing them by the average of the corresponding control group. $y_{ij} = x_{ij}/\bar{c}_{i\cdot}$. $i = 1, 2$; $j = 1, \cdots, 9$.

Suppose $X_i \sim Normal(\mu_i, \sigma_x^2)$, and $C_i \sim Normal(\nu_i, \sigma_c^2)$.

What is the distribution of Y?

Suppose that $\mu_1 = \mu_2 = 120$, $\nu_1 = \nu_2 = 30$; $\sigma_x = 4, \sigma_c = 3$. Then if we use a t-test on y_{ij}, type I error rate is about ... .

N_x	N_c	σ_x	σ_c	α
9	3	4	3	0.69
9	3	4	0.4	0.08
9	3	4	0	0.05
90	30	4	3	0.72
9	300	4	3	0.05

A simple remedy is to use a regression approach (analysis of variance).

$$Y = \beta_0 + \beta_1 X_g + \beta_2 X_t + \beta_{12} X_g X_t + \epsilon,$$

where $X_g = 0$ if group 1, 1 otherwise; $X_t = 0$ if control, 1 otherwise.

Then the expected group means are

	Control	Treatment
Group 1	β_0	$\beta_0 + \beta_2$
Group 2	$\beta_0 + \beta_1$	$\beta_0 + \beta_1 + \beta_2 + \beta_{12}$

Interpretation of the regression parameters:

β_0**:** Group 1 control mean

β_1**:** Difference of control means

β_2**:** Treatment - Control in group 1

β_{12}**:** Group 2T - Group 1T - Difference of control means

Thus, testing $\beta_{12} = 0$ is testing for the treatment difference taking into account the control difference.

Type I error rate by simulation ($B = 10,000$)

N_x	N_c	σ_x	σ_c	unweighted α	weighted $(1/\sigma^2)$ α	weighted $(1/s^2)$ α
9	3	4	3	0.026	0.051	0.063
9	3	4	0.4	0.002	0.052	0.053
90	30	4	3	0.024	0.050	0.050
200	30	4	3	0.050	0.051	0.053
90	200	4	3	0.083	0.050	0.055
90	200	4	0.4	0.184	0.054	0.054

If 'fold change' is desired like in this example, perhaps, we can take logarithm of all values before fitting the regression model.

$$\log(Y) = \beta'_0 + \beta'_1 X_g + \beta'_2 X_t + \beta'_{12} X_g X_t + \epsilon'$$

Then what does β'_{12} represent?

Chapter 3

Brief review of Bayesian analysis

- What are the philosophical differences between Frequentist and Bayesian statistics?
  To a Frequentist, parameters are fixed and unknown numbers. In the Bayesian paradigm, the parameters have distributions like the data.

- How does a Bayesian inference work?
  - Start with a prior guess of the distribution of the parameter, θ.
  - Update the distribution by combining it with the data X.
  - Obtain a posterior distribution of θ.
  - Make Statistical inferences using the posterior distribution.
- Advantages of Bayesian approaches.
  - The relevant prior information can be incorporated in a straightforward way.
  - Adaptation using the data from the ongoing experiment does not affect the inference in the way Frequentist approaches do.
  - Interpretation of the result is easier.

3.1 Bayes theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

Using this theorem, we can switch the event of interest and condition. For instance, we are usually interested in P[disease | test positive], but the observed data are usually P[test positive | disease]. As long as we can compute P[test positive] and know P[disease], we can make the switch.

In terms of updating distributions, the theorem is written as

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)},$$

where $p(\theta)$ is the prior distribution for the parameter, θ, $p(y|\theta)$ is the likelihood of the observed data, y given θ, and $p(\theta|y)$ is the posterior distribution for the parameter. It is often written as

$$p(\theta|y) \propto p(y|\theta)p(\theta).$$

3.2 Example

3.2.1 Normal distributions

Suppose we were interested in the long-term systolic blood pressure (SBP) in mmHg of a particular 60 year old female. We took two independent readings 6 weeks apart, and their mean was 130. We know that SBP is measured with a standard deviation $\sigma = 5$. (Example taken from "Bayesian Approaches to Clinical Trials and Health-Care Evaluation" by Speigelhalter, et al.)

With a Frequentist approach, a 95% confidence interval is

$$130 \pm 1.96(5/\sqrt{2}) = (123.10, 136.90).$$

Using a Bayesian approach, we can incorporate a prior belief that females aged 60 have a mean long-term SBP of 120 with standard deviation 10. Let's also assume that the prior distribution is normal. (prior = normal, likelihood = normal), that is,

$$\theta \sim N[\theta_0, \sigma^2/n_0], \qquad\qquad y|\theta \sim N[\theta, \sigma^2/m].$$

Then

$$\begin{aligned} p(\theta|y) &\propto p(y|\theta)p(\theta) \\ &\propto exp\left(-\frac{m(y-\theta)^2}{2\sigma^2}\right) \times exp\left(-\frac{n_0(\theta-\theta_0)^2}{2\sigma^2}\right) \\ &= \\ &= \\ &\propto exp\left(-\frac{n_0+m}{2\sigma^2}\left(\theta - \frac{n_0\theta_0 + my}{n_0+m}\right)^2\right). \end{aligned}$$

Therefore, the posterior distribution of θ is

$$\theta|y \sim \text{Normal}\left(\frac{n_0\theta_0 + my}{n_0+m}, \frac{\sigma^2}{n_0+m}\right).$$

In this example, we have as the prior distribution and the likelihood,

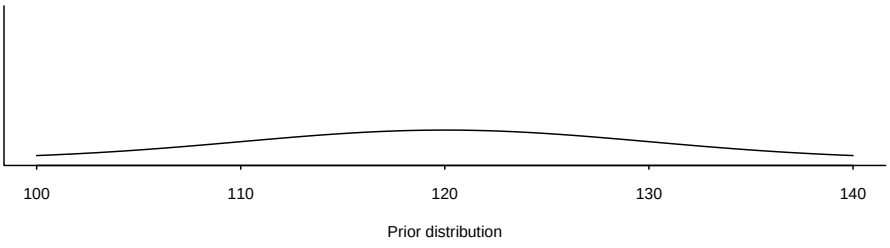
$$\theta \sim \text{Normal}(120, 10^2)$$
$$y|\theta \sim \text{Normal}(130, 5^2/2)$$

How influential was the prior information? Equivalent to $n = 0.25$

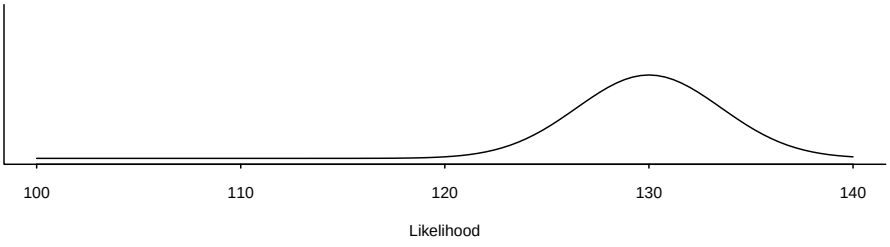
Continuing, we have the posterior mean and variance as follows:

$$\text{Posterior mean} = \frac{n_0\theta_0 + my}{n_0 + m} = \frac{(0.25)(120) + (2)(130)}{0.25 + 2} = 128.89$$
$$\text{Posterior variance} = \frac{\sigma^2}{n_0 + m} = \frac{5^2}{0.25 + 2} = 3.33^2$$

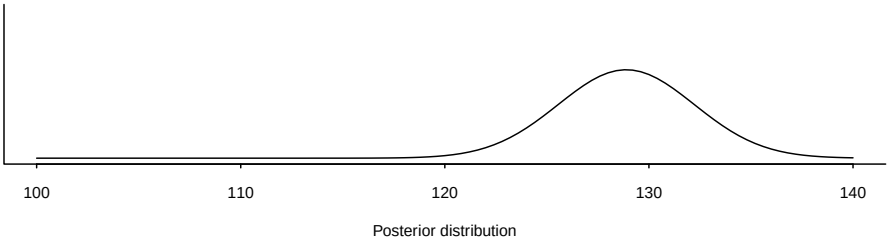
A 95% credible interval is $128.90 \pm 1.96 \times 3.33 = (122.40, 135.40)$, and we say that the probability that θ is between 122.40 and 135.40 is 95%.



Prior distribution



Likelihood



Posterior distribution

3.2.2 Beta-Binomial

In the last example, we started with a normal prior, used a normal likelihood, and arrived at a normal posterior. Things are not always that nice. In general, the posterior distribution does not have a closed form. The last example is a case of a *conjugate* analysis. Conjugate models occur when the posterior distribution is of the same family as the prior distribution. Three common examples are:

prior	likelihood	posterior
Normal	Normal	Normal
Beta	Binomial	Beta
Gamma	Poisson	Gamma

In phase I and II clinical trials, the outcome of interest is often binary, and we would like to use a Binomial distribution to model it. There, a Beta-binomial model can be used.

Example

In a safety study, we aim to estimate the probability of severe adverse events associated with a treatment of interest. We have an access to the data from another similar study that showed 7 out of 117 had a severe adverse event.

Let X be the number of adverse events out of m patients in our study, and we have

$$X|p \sim Binomial(m, p).$$

And let the prior distribution of p be $Beta(a, b)$.

Beta distribution:

$$f(p) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1},$$

where $0 \leq p \leq 1$, $a > 0$, $b > 0$, and $\Gamma(s)$ is defined by

$$\Gamma(s) = \int_0^\infty t^{s-1} e^{-t} dt.$$

It can be shown that the expectation and variance of a Beta random variable are

$$E[p] = \frac{a}{a+b}, \qquad V[p] = \frac{ab}{(a+b)^2(a+b+1)}.$$

The posterior distribution of p is

$$\begin{aligned} f(p|x) &\propto f(x|p)f(p) \\ &= \left[\binom{m}{x} p^x (1-p)^{m-x}\right] \times \left[\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1}\right] \\ &\propto \frac{\Gamma(a+b+m)}{\Gamma(x+a)\Gamma(m-x+b)} p^{x+a-1}(1-p)^{m-x+b-1}. \end{aligned}$$

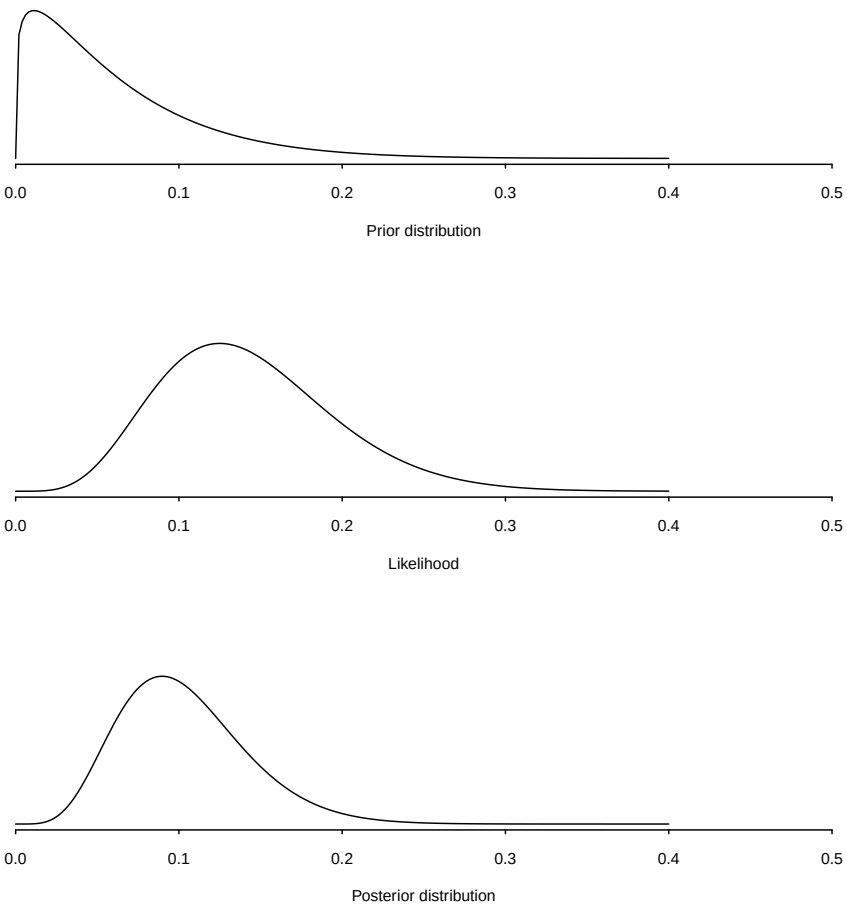
(Using the fact that for an integer k, $\Gamma(a+k) = \Gamma(a)a(a+1)(a+2)\cdots(a+k-1)$.) Thus, the posterior distribution of p is $Beta(a+x, b+m-x)$.

What do a and b mean in relation to x and $m-x$?

If we chose to use $Beta(7, 110)$ as the prior, this prior gives equal weight for the two studies, i.e., a patient in the previous study counts as much as a patient in the new study. If we want to almost completely disregard the prior information, we would use e.g., $Beta(1, 1)$, which is equivalent to having a sample size of 2 (1 adverse event out of 2).

Suppose we want to use the prior information, but discount it so that it is only equivalent to 20 patients with $P[\text{adverse event}] = 7/117$. So the prior distribution of p is $Beta(1.20, 18.80)$

Let's say in our study with sample of size 40, there were 5 adverse events. Then the posterior distribution is $f(p|x) = Beta(1.20 + 5, 18.80 + 35) = Beta(6.20, 53.80)$.



Using the posterior distribution, $Beta(6.20, 53.80)$, we can compute $E[p] = 6.20/(6.20+53.80) = 0.10$. Moreover we can compute the probabilities:

```
# P[p > 0.1]
1 - pbeta(0.1, post.a, post.b)


[1] 0.489


# P[p > 0.2]
1 - pbeta(0.2, post.a, post.b)
```

```
[1] 0.01704
```

From the data alone, a Frequentist confidence interval on p:

```
data.a <- x
data.b <- n - x

binconf(x, n, method = "all")
```

```
           PointEst   Lower  Upper
Exact         0.125 0.04186 0.2680
Wilson        0.125 0.05460 0.2611
Asymptotic    0.125 0.02251 0.2275
```

And a Bayesian credible interval is:

```
# Lower bound
qbeta(0.025, post.a, post.b)
```

```
[1] 0.04036
```

```
# Upper Bound
qbeta(1 - 0.025, post.a, post.b)
```

```
[1] 0.1911
```

This is narrower than the Frequentist counterpart, and it has a nice interpretation.

Chapter 4

Randomization in clinical trials

4.1 Example: Polio vaccine trial (1954)

In 1954, 1.8 million children participated in the largest clinical trial to date to assess the effectiveness of the vaccine developed by Jonas Salk in preventing paralysis or death from poliomyelitis.

- 1.8 million children in selected school districts throughout the US were involved in this placebo-controlled trial.
  - Why was placebo necessary?
  - $60,000$ cases in 1952; about half of that in 1953.

- Randomized trial.
  - $750,000$ children participated.
  - They required parents' consent.
  - Half of the children with consent were randomized into the vaccine group.
- NFIP design.
  - The National Foundation for Infantile Paralysis (NFIP) conducted a study in which all 2nd graders with consent received the vaccine with 1st and 3rd graders acting as control.
  - $1,125,000$ children participated.
  - The control children did not require consent.
    Systematic difference between groups.
  - No blinding.
  - Polio is a contagious disease!

The results of the SVF trial are tabulated below[1].

[1] Freedman et al. Statistics, second edition

The randomized double-blind design

	Size	Rate*
Treatment	$200,000$	26
Control	$200,000$	71
No consent	$350,000$	46

(*Rate of polio cases per $100,000$)

The NFIP design

	Size	Rate*
Treatment	$225,000$	25
Control	$725,000$	54
No consent	$125,000$	44

(*Rate of polio cases per $100,000$)

4.2 Introduction

Randomization Assignment of patients or experimental subjects to two or more treatments by chance alone.

Main advantages of randomization

- It removes the potential of bias in the allocation of participants to the intervention group or to the control group (allocation bias).
- It tends to produce similar (compatible) groups in terms of measured as well as unmeasured confounders
  confounding by indication in observational studies.

Randomization is considered so important that the intention-to-treat principle considered sacrosanct: "Analyze by assigned treatments irrespective of actual treatment received."

Perceived disadvantages of randomization are often about emotional and ethical issues.
$\rightarrow$ randomization before consent

Predecessors to randomization:

- Alternating assignments (TCTCTCTC...).
- Treatment assignment based on birthday / day of the week.

The primary problems with these non-random assignment are the lack of assurance of comparability

(baseline balance). An additional issue with the "alternating assignments" is that if one is unblinded, all the rest are unblinded, too.

4.3 Simple randomization

For each subject, flip a coin to determine treatment assignment. P[treatment 1] $= \cdots =$ P[treatment k] $= 1/k$.

Problems with simple randomization and how to deal with them.

- Imbalance in treatment allocation.
  - Replacement randomization.
  - Block randomization.
  - Adaptive randomization. (Biased coin / Urn model etc.)
- Imbalance in baseline patient characteristics.
  - Stratified randomization. (Stratified permuted block randomization)
  - Covariate adaptive randomization. (Minimization randomization)

4.4 Imbalance in treatment allocation

If the number of patients, N is 20, $P[10 \text{ and } 10] = 0.18$. The probability of 7 to 13 split or worse is 26%. The treatment effect variance for $7 - 13$ split relative to $10 - 10$ split is

$$\left(\frac{1}{7} + \frac{1}{13}\right) / \left(\frac{1}{10} + \frac{1}{10}\right) = 1.098.$$

7-13 split is only $1/1.098 = 0.92$ as efficient as 10-10 split.

Even if treatment allocation is balanced at the end of trial, there may be a (severe) imbalance at some point. Because we may monitor trials over time, we prefer to have balance over time.

4.4.1 Block randomization

To ensure a better balance (in terms of number of patients) across groups over time, consider a block randomization (random permuted blocks).

Block randomization ensures approximate balance between treatments by forcing balance after a small number of patients (say 4 or 6). For example, the first 4 patients are allocated to treatment A or B sequentially based on $AABB$.

There are 6 sequences of A, A, B, B, and let each sequence have 1/6 chance of being selected.

$$AABB \qquad ABAB \qquad ABBA \qquad BAAB \qquad BABA \qquad BBAA$$

```
for (i in 1:5) {
    cat(i, sample(rep(LETTERS[1:2], each = 2), 4, replace = FALSE), "\n")
}
```

```
1 A B B A
2 B B A A
3 B A A B
4 A B A B
5 A B A B
```

What's wrong with block size of 2? Block size of 200?

Easily applicable to more than 2 groups (A, B, C)

```
for (i in 1:5) {
    cat(i, sample(rep(LETTERS[1:3], each = 2), 6, replace = FALSE), "\n")
}
```

```
1 A B C C A B
2 B B C A C A
3 C A B B C A
4 C A A B C B
5 A B C A C B
```

Easily applicable to unequal group sizes ($N_a = 40$ and $N_b = 20$).

```
for (i in 1:5) {
    cat(i, sample(rep(LETTERS[1:2], c(4, 2)), 6, replace = FALSE), "\n")
}
```

```
1 A B A A B A
2 A B A A A B
3 B B A A A A
4 A A B B A A
5 A B B A A A
```

Why do we want unequal group sizes?

- We may want to have a better estimate of the effect for the new treatment.
- Treatment costs may be very different.
  Given the total sample size and the relative cost of treatment 2 to treatment 1, we can find the optimal allocation ratio to minimize the total cost. (More in sample size computation)
- Variances may be different.
  Suppose the means, μ_1 and μ_2, of treatment groups are being compared using

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}.$$

  For a given $N = n_1 + n_2$, the test statistic is maximized when the denominator is minimized. Solving

$$\frac{\partial}{\partial n_1}\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{N - n_1}\right) = 0$$

  we get

$$\frac{n_1}{N} = \frac{\sigma_1}{\sigma_1 + \sigma_2}.$$

  Therefore, the optimal allocation ratio is $r = n_1/n_2 = \sigma_1/\sigma_2$.

Analysis should account for the randomization scheme but often does not. Matts and McHugh (1978 *J Chronic Dis*) point out that

- because blocking guarantees balance between groups and increases the power of a study, blocked randomization with the appropriate analysis is more powerful than not blocking at all or blocking and ignoring it in the analysis.
- not accounting for blocking in analysis is conservative.

4.4.2 Biased coin and urn model

These techniques are sometimes classified as “adaptive randomization”.

Allocation of i-th patient depends on how many have been randomized to group A (n_a) and group B (n_b).

Any given time, the probability of allocation to group A may be

$$P[A] = \frac{n_b}{n_a + n_b}.$$

Or the rule may be to use $P[A] = 2/3$ when $n_b - n_a > 5$, and $P[B] = 2/3$ when $n_a - n_b > 5$. Characteristics of such a randomization scheme are usually studied by simulations.

An urn model is one type of biased coin randomization.

- Prepare an urn with one Amber ball and one Blue ball.
- Pick one ball and make the corresponding treatment assignment (A/B).
- Put a ball of the opposite color in the urn.

```
urn1 <- function(n) {
    # randomize n patients into A or B.  At any time P[A] = (#B so far + 1)
    # / (#A so far + 1 + #B so far + 1).
    out <- data.frame(matrix(0, ncol = 4, nrow = n + 1))
    out[1, 1] <- 1
    out[1, 2] <- 1
    for (i in 1:n) {
        out[i, 3] <- out[i, 2]/(out[i, 1] + out[i, 2])
        out[i, 4] <- sample(c("A", "B"), 1, prob = out[i, 2:1])
        out[i + 1, 1] <- out[i, 1] + (out[i, 4] == "A")
        out[i + 1, 2] <- out[i, 2] + (out[i, 4] == "B")
    }
    out[, 1] <- out[, 1] - 1
    out[, 2] <- out[, 2] - 1
    names(out) <- c("A so far", "B so far", "P[A next]", "Next")
    out[1:n, ]
}

urn1(n = 10)
```

```
   A so far B so far P[A next] Next
1         0        0    0.5000    A
2         1        0    0.3333    A
3         2        0    0.2500    B
4         2        1    0.4000    B
5         2        2    0.5000    A
6         3        2    0.4286    A
7         4        2    0.3750    B
8         4        3    0.4444    B
9         4        4    0.5000    B
10        4        5    0.5455    A
```

4.5 Imbalance in baseline patient characteristics

Block randomization and biased coin model ensure that the group sizes are reasonably balanced. In order to facilitate the comparison of treatment effects, balance on important baseline variables is

sometimes desired.

- Randomization does not guarantee all the measured variables will be balanced. And imbalance does not mean randomization did not work.

Senn (1994) It is argued that this practice [testing baseline homogeneity] is philosophically unsound, of no practical value and potentially misleading. Instead it is recommended that prognostic variables be identified in the trial plan and fitted in an analysis of covariance regardless of their baseline distribution (statistical significance).

Piantadosi These methods, while theoretically unnecessary, encourage covariate balance in the treatment groups, which tends to enhance the credibility of trial results.

An annonymous reviewer Since this is a randomized controlled trial, comparison of baseline characteristics (Table 1) is not necessary. The problem with this approach is that when comparing baseline characteristics we already know that the null hypothesis is true if the randomization was done correctly. Thus, we would expect 1 test in 20 to give a 'significant' result with $p < 0.05$ by chance alone. The best approach is to specify key prognostic factors to include in multivariable models irrespective of their significance between treatment groups.

4.5.1 Stratified randomization

Stratified randomization is applied to ensure that the groups are balanced on baseline variables that are thought to be significant.

- Create strata based on the variables for which balance is sought.
  e.g., (Male, 65 or younger), (Male, older), (Female, younger), (Female older)
- Randomize to treatments within each stratum. **Use block randomization!**
  What's wrong with
  - using simple randomization within a stratum?
  - using too many strata?
- Stratification should be accounted for in analysis.
  - Pre-randomization stratification and post-randomization stratification (at time of analysis) has no clear winner.
- If trial is large, stratification may not be necessary
- Stratification by center is a good idea from practical viewpoints.
  - Allows randomization to be hosted at each site
  - Allows sites to be removed and still maintains balance
- Block randomization is a special type of stratified randomization where strata are defined by ... .
- If each stratum has a target size, plans need to be in place to close down recruitment based on the baseline characteristics. e.g., "We do not need any more (Male, older)".

4.5.2 Adaptive and minimization randomization

Adaptive randomization can be used to reduce baseline imbalance:

- Define an imbalance function based on factors thought to be important
- Then use a rule to define P[treatment A] so that the next assignment is more likely to reduce imbalance.

For example, the factors to balance are sex (male/female) and hypertension (yes/no), and let the imbalance function be

$$I = 2 \times (\text{sex imbalance}) + 3 \times (\text{hypertension imbalance}).$$

The patients randomized so far are

	Sex		Hypertension	
	Male	Female	Yes	No
Group 1	10	3	8	5
Group 2	8	3	6	5

The next patient is male-non hypertensive. The imbalance will be

$$I = 2 \times (11 - 8) + 3 \times (6 - 5) = 9 \text{ if Group 1,}$$
$$I = 2 \times (10 - 9) + 3 \times (6 - 5) = 5 \text{ if Group 2.}$$

Thus let P[Group 2] $= 2/3$.

Minimization randomization uses the same idea but use P[Group 2] = 1, to eliminate randomness when there is some imbalance. Randomize only when to assign the next patient to either group gives the same value of I.

4.6 Response adaptive randomization

As the name suggests, response adaptive randomization methods use the information about the response so far to allocate the next patient.

Play the winner: The idea is to allocate more patients in the treatment that seems to be working better. To apply these methods, it is necessary to have a response quickly. Urn model can be used to make treatment assignment imbalance based on the results (success/failure) of each treatment so far. (e.g., put one blue ball if the treatment B yields success.)

Instead of updating the probabilities of treatment assignment after each patient, we can update them after a group of patients' results are available to reduce administrative burden. In a phase II clinical

trial, play the winner design may be used to reduce the number of treatments in consideration. (e.g., Only retain the treatment arms that have P[positive response] > 0.4.)

4.6.1 Example: ECMO

Bartlett et al.[2] conducted a randomized study of the use of extracorporeal membrane oxygenation (ECMO) to treat newborns with respiratory failure. A play-the-winner design[3] was used because

- the outcome is known soon after randomization.
- most ECMO patients were expected to survive and most control patients were expected to die.

  Ethically, the investigators felt obligated not to withhold the lifesaving treatment. Scientifically, they felt obligated to perform a randomized study.

The randomization plan:

- The first patient will be randomized to ECMO or the conventional treatment (CT) with equal probability.
- For each patient who survives on ECMO or dies on CT, one ECMO ball is added to the urn.
- For each patient who survives on CT or dies on ECMO, one CT ball is added to the urn.
- The trial will be terminated when 10 balls of one kind have been added, and that treatment will be chosen as the winner.

What actually happened:

P(ECMO)=1/2 Patient 1 was randomized to ECMO and survived.

P(ECMO)=2/3 Patient 2 was randomized to CT and died.

P(ECMO)=3/4 Patient 3 was randomized to ECMO and survived

P(ECMO)=4/5 Patient 4 was randomized to ECMO and survived

P(ECMO)=5/6 Patient 5 was randomized to ECMO and survived

P(ECMO)=6/7 Patient 6 was randomized to ECMO and survived

P(ECMO)=7/8 Patient 7 was randomized to ECMO and survived

P(ECMO)=8/9 Patient 8 was randomized to ECMO and survived

P(ECMO)=9/10 Patient 9 was randomized to ECMO and survived

[2] "Extracorporeal circulation in neonatal respiratory failure: a prospective randomized study". (1985) *Pediatrics*
[3] Zelen (1969) *JASA*; Wei and Durham (1978) *JASA*

P(ECMO)=10/11 Patient 10 was randomized to ECMO and survived

Randomization was stopped when there were 11 ECMO patients who survived and 1 CT patient who died.
Controversies followed because ...

```
fisher.test(cbind(c(11, 0), c(0, 1)))
```

```
Fisher's Exact Test for Count Data

data:  cbind(c(11, 0), c(0, 1))
p-value = 0.08
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.2821    Inf
sample estimates:
odds ratio
       Inf
```

In retrospect it would have been better to begin with two or three pairs of balls, which probably would have resulted in more than one control patient.

Chapter 5

Sample size and power

5.1 Introduction and terminology

- Type I error rate
- Type II error rate
- Power
- Effect size
- Minimum clinical significance

Type I error can be controlled regardless of sample size, so sample size calculation is mostly about statistical power. The regulatory body (e.g., FDA) is concerned about the type I error, and the sponsor is concerned, in addition, about the type II error. FDA's main goal is to prevent an ineffective intervention to be claimed effective. If the sample size is smaller than necessary, then a statistically significant effect may not be observed when the intervention is in fact effective.

Sample size calculations are only approximation because ...

- They are based on (hopefully educated) guesses.
- They are based on a mathematical model that are only approximately true.

It is usually best to be conservative.

There is much to think about when planning a study:

- The main question (hypothesis).
- The endpoint.
- Statistical analysis plan.
- Treatment effect to detect. (Clinical significance)

Statistical power is a function of type I error rate, sample size, treatment effect under the alternative,

variance. Sometimes, it is better to compute sample size based on the desired length of confidence intervals instead of power.

5.2 Sample size calculation for continuous variables

Suppose data are at least approximately normally distributed. Then a test statistic for testing

$$H_0 : \mu_t - \mu_c = \delta_0$$
$$H_1 : \mu_t - \mu_c > \delta_0$$

may be

$$Z = \frac{\bar{X}_t - \bar{X}_c - \delta_0}{\sqrt{\sigma_t^2/N_t + \sigma_c^2/N_c}}.$$

To have a type I error rate of α and power of $1 - \beta$ at δ_1, we need:

$$\frac{\sigma_t^2}{N_t} + \frac{\sigma_c^2}{N_c} = \frac{(\delta_1 - \delta_0)^2}{(z_{1-\alpha} - z_\beta)^2}.$$

- If $\delta_0 = 0$?
- If two-sided alternative?
- If $N \equiv N_t = N_c$?
- If $\sigma^2 \equiv \sigma_t^2 = \sigma_c^2$?

Let $N_t = rN_c$ and solve for N_c to get

$$N_c = \frac{(z_\alpha + z_\beta)^2(\sigma_t^2 + r\sigma_c^2)}{r(\delta_1 - \delta_0)^2}$$

Example: $\delta_0 = 0$, $\delta_1 = 1$, $\sigma_t = 4$, $\sigma_c = 2$, $\alpha = 0.03$, $\beta = 0.10$, and $r = 2$.

$$N_c = \frac{(1.96 - 1.28)^2(4^2 + 2(2^2))}{2(1-0)^2}$$
$$= 127$$
$$N_t = 254$$

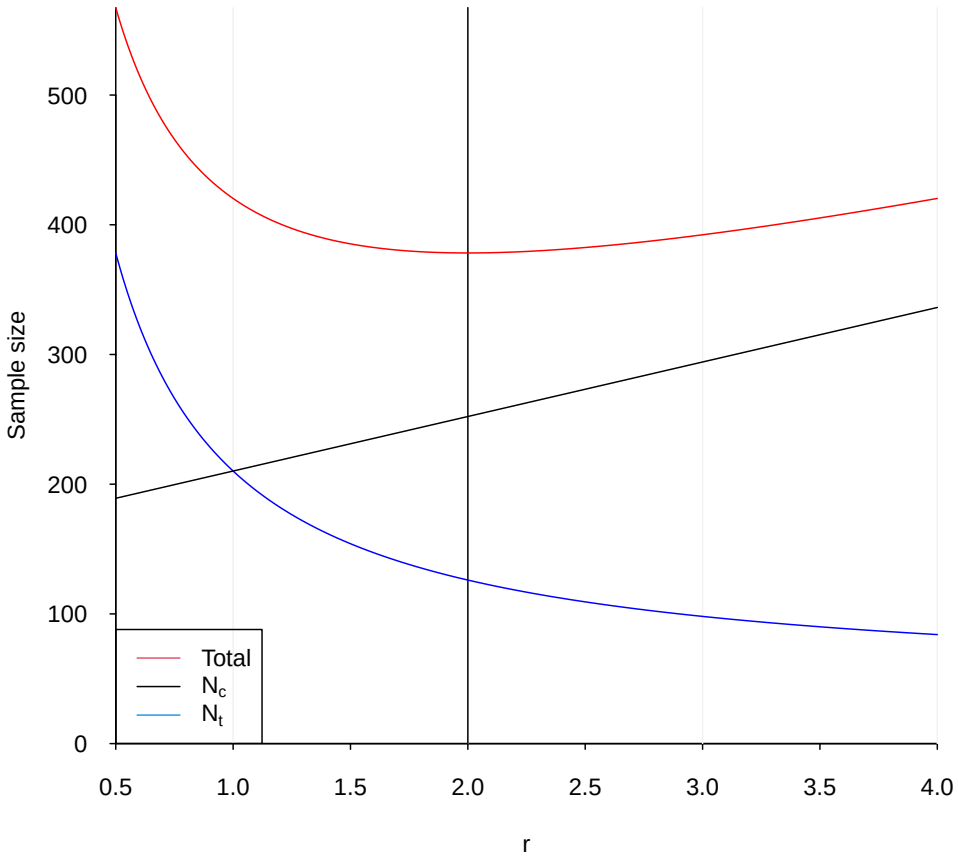
5.2.1 Optimal (minimum) sample size

To find the value of r that minimizes $N \equiv N_t + N_c$, differentiate the expression of $N \equiv N_c + N_t$ with respect to r, and solve $\frac{\partial N}{\partial r} = 0$.

$$
\begin{aligned}
N &= N_c + N_t \\
&= N_c(1+r) \\
&= \frac{(z_\alpha + z_\beta)^2}{(\delta_1 - \delta_0)^2}\left(\frac{\sigma_t^2}{r} + \sigma_c^2\right)(1+r) \\
&= W^2(\sigma_t^2/r + \sigma_c^2 + \sigma_t^2 + r\sigma_c^2). \\
\frac{\partial N}{\partial r} &= -W^2\sigma_t^2/r^2 + W^2\sigma_c^2
\end{aligned}
$$

Setting this to 0 and solve for r yields $r_{opt} = \sigma_t/\sigma_c$.

```
ssNorm <- function(d0, d1, sigt, sigc, al, be, r) {
    (-qnorm(al) - qnorm(be))^2 * (sigt^2 + r * sigc^2)/(r * (d1 - d0)^2)
}

r <- seq(0.5, 4, by = 0.01)
Nc <- ssNorm(d0 = 0, d1 = 1, sigt = 4, sigc = 2, al = 0.025, be = 0.1,
    r = r)
Nt <- Nc * r
```



Now further suppose that the relative cost of the subjects in treatment group compared to the control group is C ($C > 1$ indicates that the treatment group is more expensive).

The total cost (M) is

$$\begin{aligned} M &= N_c + CN_t \\ &= N_c(1 + Cr), \end{aligned}$$

and the optimal allocation in this case is

$$r_{opt} = \frac{\sigma_t}{\sigma_c} \frac{1}{\sqrt{C}}.$$

5.2.2 Sample size calculation for precision

For a given sample size, n_c and n_t, a $(1-\alpha)\times 100\%$ confidence interval has a form,

$$(\overline{X}_t - \overline{X}_c) \pm z_{1-\alpha/2}\sqrt{\sigma_t/n_t + \sigma_c/n_c}.$$

We can solve $M = z_{1-\alpha/2}\sqrt{\sigma_t/n_t + \sigma_c/n_c}$ to compute the necessary sample size to specify the width of this confidence interval. M is the margin of error, and $2M$ will be the length of this confidence interval. Letting $\sigma = \sigma_t = \sigma_c$ and $N = n_t = n_c$ and solving for n gives

$$n = \frac{2z^2_{\alpha/2}\sigma^2}{M^2}.$$

- This is (should be) used to compute the necessary number of simulations to run.
- What is the required sample size to make the width of a confidence interval reduced by half?
- What is the required sample size if σ is doubled?

5.2.3 Sample size calculation for paired observations

With a paired data set, suppose we would like to test $H_0 : \mu_a - \mu_b = 0$. The same formula as two-sample case can be used, but a few points to note are:

- $n_a = n_b$.
- Usually $\sigma_a = \sigma_b$.
- Usually there is a positive correlation between the paired measurements.

Let ρ be the correlation between the paired measurements, and the simple version of the sample size formula (one-sample case) is

$$n_b = n_a = \frac{(z_\alpha + z_\beta)^2\sigma_d^2}{(\delta_1 - \delta_0)^2},$$

where σ_d^2 is the variance of the differences, which is $\sigma_d^2 = \sigma_b^2 + \sigma_a^2 - 2cov(A, B)$.

- When is it better to treat the data as independent even when the data are actually paired?

5.3 Binomial outcome variables

5.3.1 One sample problem

In many early phase clinical trials, the hypothesis of interest is about a response rate and a one-arm trial may use a historic control: $H_0 : p = p_0$.

A test statistic,

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}},$$

has approximately the standard normal distribution under H_0. Under $H_1 : p = p_1$,

$$E[Z] = \frac{p_1 - p_0}{\sqrt{p_0(1-p_0)/n}}.$$
$$V[Z] = \frac{p_1(1-p_1)/n}{p_0(1-p_0)/n}.$$

The same construction as for the normal case leads to

$$n = \frac{\left(z_\alpha\sqrt{p_0(1-p_0)} + z_\beta\sqrt{p_1(1-p_1)}\right)^2}{(p_1 - p_0)^2}.$$

5.3.2 Two sample problem

Now suppose we want to test $H_0 : p_t = p_c$, where p_t and p_c are response rates for the treatment and control groups, respectively. We have

$$X_t \sim Binomial(n_t, p_t),$$
$$X_c \sim Binomial(n_c, p_c).$$

And we have

$$E[\hat{p}_t - \hat{p}_c] = p_t - p_c,$$
$$V[\hat{p}_t - \hat{p}_c] = p_t(1-p_t)/n_t + p_c(1-p_c)/n_c.$$

Thus the test statistic,

$$Z = \frac{\hat{p}_t - \hat{p}_c}{\sqrt{p_t(1-p_t)/n_t + p_c(1-p_c)/n_c}},$$

has the standard normal distribution (when sample sizes are large), and it reduces to

$$Z = \frac{\hat{p}_t - \hat{p}_c}{\sqrt{p(1-p)}\sqrt{1/n_t + 1/n_c}},$$

where $p = p_t = p_c$. p is estimated by $(X_t + X_c)/(n_t + n_c)$, i.e., $\bar{p} = (n_t p_t + n_c p_c)/(n_t + n_c)$, which is $(p_t + p_c)/2$ if $n_t = n_c$.

Under $H_1 : p_t - p_c = p'$, no such simplification is available.

$$n = \frac{\left(z_\alpha\sqrt{2\bar{p}(1-\bar{p})} + z_\beta\sqrt{p_t(1-p_t) + p_c(1-p_c)}\right)^2}{(p_t - p_c)^2}.$$

5.3.3 Adjustment for noncompliance

Here we assume that a new treatment is being compared to a standard treatment.

dropouts those who are randomized to the new treatment but received a standard treatment

drop-ins those who are randomized to the standard treatment but received the new treatment

When the treatment effect is tested with the intention-to-treat principle, these usually dilute the treatment effect.

Let R_{tc} denote the proportion of dropouts (moving from t to c), and R_{ct} denote the proportion of drop-ins.

Recall the sample size formula,

$$n = \frac{\left(z_\alpha \sqrt{2\bar{p}(1-\bar{p})} + z_\beta \sqrt{p_t(1-p_t) + p_c(1-p_c)}\right)^2}{(p_t - p_c)^2}.$$

Dropouts and drop-ins have a small effect on the numerator but pose a major impact on the denominator, which is the expected difference of the proportions under the alternative, and it gets diluted with dropouts and drop-ins.

$$\begin{aligned} p'_t &= E[\hat{p}_t] = p_t(1 - R_{tc}) + p_c R_{tc} \\ p'_c &= E[\hat{p}_c] = p_c(1 - R_{ct}) + p_t R_{ct} \\ (p'_t - p'_c)^2 &= [(p_t - R_{tc}(p_t - p_c)) - (p_c + R_{ct}(p_t - p_c))]^2 \\ &= (p_t - p_c)^2 (1 - R_{tc} - R_{ct})^2 \end{aligned}$$

Thus, the necessary sample size will increase by the factor $1/(1 - R_{tc} - R_{ct})^2$. For example, if we estimate $R_{tc} = 0.05$ and $R_{ct} = 0.10$, then the final sample size will be increased by the factor of $1/(1 - 0.10 - 0.05)^2 = 1.38$.

5.4 Additional topics for binomial responses

5.4.1 Other parameterizations

Hypotheses of interest may be specified in terms of difference or ratio of the two proportions or the odds ratios.

We have seen that even if the hypotheses are written in terms of difference of proportions, we need to specify the location (p_c) to compute the sample size. Once p_c and p_t are set, we can compute the ratio (e.g, risk ratio), $\phi = p_t/p_c$ and odds ratio, $\psi = (p_t/(1-p_t))/(p_c/(1-p_c))$.

If the hypotheses are written in terms of risk ratio or odds ratio, we translate those into statements about difference of proportions and compute the sample size.

$H_0 : \psi = 1$
$H_1 : \psi = 2$
(Odds of the good outcome is twice as likely in the treatment group as in the control group.)

p_c	0.01	0.05	0.10	0.20	0.30	0.40	0.50	0.60
p_t	0.020	0.095	0.18	0.33	0.46	0.57	0.67	0.75
odds ratio	2	2	2	2	2	2	2	2
risk ratio	1.98	1.90	1.82	1.67	1.54	1.43	1.33	1.25
difference	0.010	0.045	0.082	0.13	0.16	0.17	0.17	0.15

5.4.2 Discreteness

Consider the following example. $H_0 : p = 0.2$, $H_1 : p > 0.2$. To have a type I error rate of 5% and 80% power at $p = 0.35$, the necessary sample size is:

$$\begin{aligned} n &= \frac{\left(z_\alpha \sqrt{p_0(1-p_0)} + z_\beta \sqrt{p'(1-p')}\right)^2}{(p'-p_0)^2} \\ &= \frac{(1.645\sqrt{.2 \times .8} + 0.842\sqrt{.35 \times .65})^2}{(.35-.2)^2} \\ &\approx 50 \end{aligned}$$

The critical value in terms of the observed p is:

$$\begin{aligned} C_p &= 0.2 + 1.645\sqrt{.2 \times .8/50} = 0.293 \\ &\approx 0.35 - 0.842\sqrt{.35 \times .65/50} \end{aligned}$$

This translates to the rejection rule that if we observe 15 or more successes, we reject H_0. ($50 \times 0.293 = 14.7$) And the actual type I error rate and power are:

$$P_0[X \geq 15] = \sum_{x=15}^{50} \binom{50}{x} 0.2^x 0.8^{50-x} = 0.061.$$

$$P_1[X \geq 15] = \sum_{x=15}^{50} \binom{50}{x} 0.35^x 0.65^{50-x} = 0.81$$

Brute force search nearby $(50, 15)$ found $(56, 17)$.

$$P_0[X \geq 17] = \sum_{x=17}^{56} \binom{56}{x} 0.2^x 0.8^{56-x} = 0.043.$$

$$P_1[X \geq 17] = \sum_{x=17}^{56} \binom{56}{x} 0.35^x 0.65^{56-x} = 0.81$$

```
bProb <- function(r, n, p0, p1) {
    # Reject if X > r.
    rej0 <- sum(dbinom((r + 1):n, n, p0))
    # also 1-pbinom(r,n,p0)
    rej1 <- sum(dbinom((r + 1):n, n, p1))
    # also 1-pbinom(r,n,p1)
    c(rej0, rej1)
}

findR <- function(n, p0, alpow) {
    # Reject if X > r.  Find r so that P[X>r] < alpow.
    r <- qbinom(1 - alpow, n, p0)
    actual.alpow <- 1 - pbinom(r, n, p0)
    data.frame(r, actual.alpow)
}
```

5.4.3 arc sin method

one-sample case

For a random variable, $X \sim Binomial(n, p)$,

$$Z = 2\sqrt{n}\left(\sin^{-1}\sqrt{\hat{p}}\right).$$

is approximately normally distributed with mean $2\sqrt{n}(\sin^{-1}\sqrt{p})$ and variance 1.

Using the sample size formula for normal distributions, we get

$$n = \frac{(z_\alpha + z_\beta)^2}{4(\sin^{-1}\sqrt{p_1} - \sin^{-1}\sqrt{p_0})^2}$$

For the current example, we have

```
(qnorm(.05)+qnorm(.20))^2 / 4 / (asin(sqrt(.35)) - asin(sqrt(.20)))^2
```

$n = 54$.

two-sample case
Similarly,

$$Z = \sqrt{2m}\left(\sin^{-1}\sqrt{\hat{p}_t} - \sin^{-1}\sqrt{\hat{p}_c}\right)$$

is approximately normally distributed with mean $\sqrt{2m}(\sin^{-1}\sqrt{p_t} - \sin^{-1}\sqrt{p_c})$, where $m = 2n_t n_c/(n_t + n_c)$, and variance 1.

example
$H_0 : p_t - p_c = 0$, $H_1 : p_t - p_c = .2$. $\alpha = .05$, $\beta = .10$. Suppose we want $n_t = 1.5 \times n_c$.

Further suppose $p_c = 0.2$ so that under H_1, $p_t = 0.4$.

The sample size formula is

$$m = \frac{(z_\alpha + z_\beta)^2}{2(\delta_1 - \delta_0)^2},$$

where $\delta_1 = \sin^{-1}\sqrt{p_t} - \sin^{-1}\sqrt{p_c}$ under H_1, and δ_0 is the same quantity under H_0. (0 in this case)

$m = (1.645 + 1.282)^2/(2(0.2211^2)) = 86.7$, so if $n_t = n_c$ is desired, the sample size is 87 each. However, we need to solve for n_t and n_c satisfying $n_t = 1.5 \times n_c$.
$n_c = 73$ and $n_t = 110$.

5.5 Sample size using simulation

A simulation study is a highly useful tool when computing the sample size and learning operation characteristics of a future design.

example
Suppose we want to conduct a study to compare two treatments to test $H_0 : \mu_t - \mu_c = 0$ against $H_1 : \mu_t - \mu_c > 0$ with $\alpha = 0.025$ and power = 0.90 at $\mu_t - \mu_c = 10$. Using information from similar studies, we decide to use $\sigma = 32$.

Then the per-group sample size should be

$$\begin{aligned} N &= \frac{(z_\alpha + z_\beta)^2(\sigma_t^2 + \sigma_c^2)}{(\delta_1 - \delta_0)^2} \\ &= \frac{(1.960 + 1.282)^2(2 \times 32^2)}{(10 - 0)^2} \\ &= 215.2. \end{aligned}$$

The sample size depends on the estimate of σ used. The value of σ and the sample size are summarized below.

```
ss <- function(alp, bet, sigm, delt) (qnorm(alp) + qnorm(bet))^2 * (2 *
    sigm^2)/(delt^2)
po <- function(alp, nnn, sigm, delt) pnorm(sqrt(nnn * delt^2/(2 * sigm^2)) +
    qnorm(alp))

ceiling(ss(0.025, 0.1, c(32, 36, 40, 44), 10))


[1] 216 273 337 407


round(100 * po(0.025, 216, c(32, 36, 40, 44), 10))


[1] 90 82 74 66
```

σ	32	36	40	44
N	216	273	337	407
power if $N = 216$	90%	82%	74%	66%

Because we are not confident about our pre-study estimate, $\sigma = 32$, we plan to (have someone) look at the data and compute the pooled variance, s_1^2, when the total sample size is 50. If $s_1 > 36$ then we will recompute the sample size so that the total sample size is

$$N^* = \frac{(1.960 + 1.282)^2(2 \times s_1^2)}{10^2}$$

Simulation results ($B = 1,000$)
The following tables show the five-number summary (Min-Q1-Med-Q3-Max)

	Under H_0		Under H_1	
σ	type I error rate	N	power	N
32	0.030	216, 216, 216, 216, 415	0.91	216, 216, 216, 216, 407
36	0.024	216, 216, 216, 311, 524	0.86	216, 216, 216, 306, 495
40	0.024	216, 287, 333, 380, 648	0.88	216, 287, 332, 380, 667
44	0.025	216, 347, 396, 460, 727	0.87	216, 349, 402, 455, 700

Perhaps, the timing of the internal look is too early. Change it to when we have 100 observations total.

Simulation results ($B = 1,000$)

	Under H_0		Under H_1	
σ	type I error rate	N	power	N
32	0.031	$216, 216, 216, 216, 327$	0.90	$216, 216, 216, 216, 322$
36	0.017	$216, 216, 216, 296, 449$	0.88	$216, 216, 274, 301, 395$
40	0.031	$216, 299, 333, 368, 497$	0.88	$216, 302, 333, 368, 524$
44	0.027	$216, 366, 404, 443, 641$	0.88	$216, 368, 404, 445, 591$

- Are $1,000$ simulations enough?
- Can we estimate σ^2 well with $n_1 = 50$ or 100?

Chapter 6

Phase I Clinical Trials

6.1 Introduction

Phase I clinical trial is the first study in which a new drug is administered in humans. The primary objectives of phase I studies are 1) to collect pharmacokinetic and pharmacokynamic data, and 2) to establish safety with a specific goal to estimate the maximum tolerated dose (MTD).

- Traditionally, the MTD is the dose level at which the probability of dose-limiting toxicity (DLT) is 33% (or maybe 20%).
- We make an assumption that higher doses are more effective.
  We assume that the highest safe dose is the dose most likely to be effective; in other words, we are using dose-related toxicity as a surrogate for efficacy.

Phase I clinical trials are also known as:

Treatment mechanism Early developmental trial that investigates mechanism of treatment effect. (e.g., pharmacokinetics)

Dose escalation Design that specifies methods for increase in dose for subsequent subjects.

Dose-ranging Design that tests some or all of a prespecified set of doses (fixed doses)

Dose-finding Design that titrates dose to a prespecified optimum based on biological or clinical considerations

In cancer and AIDS trials, only patients participate in phase I trials; however, in other disease areas, safety data may be gathered on healthy volunteers. The primary reason to recruit patients into a phase I clinical trial is known toxicity. (cytotoxic drug)

Who are the healthy volunteers?

- usually 18-35 years
- usually male
- non-smoker / non substance abuse
- no symptoms of disease
- no laboratory abnormalities

FDA uses the terminology, “normal volunteers” (Guidance for industry: General considerations for the clinical evaluation of drugs), but who are normal?

“With respect to the use of ‘normal’ subjects it should be recognized that few people are literally normal in all respects. This term should be interpreted with caution and should mean volunteers who are free from abnormalities which would complicate the interpretation of the experiment or which might increase the sensitivity of the subject to the toxic potential of the drug.”

6.2 Non-cancer, non-AIDS phase I clinical trials

Most phase I studies are placebo controlled randomized study to reduce observer bias and facilitate comparison between active drug and placebo.

In a typical design, subjects are assigned to a cohort of size 8 to 10; 6 to 8 subjects are assigned to the active treatment (same dose) and 2 to placebo. If deemed safe, the next cohort of the same size are given one higher dose. The trial is stopped when an unacceptable number of adverse events is observed; the highest safe dose is the target dose that will be recommended for future trials.

Starting dose
One popular starting dose is based on the dose that causes 10% mortality in rodents on a mg/m^2 (per body surface) basis (LD_{10}). We use $LD_{10}/10$ as the starting dose. An FDA document (Guidance for industry. Estimating the maximum safe starting dose in initial clinical trials for therapeutics in adult healthy volunteers) provides a conversion table for this purpose.

Species	mg/kg to mg/m^2 Multiply by	animal mg/kg to HED mg/kg Multiply by
Human	37	–
Child $(20kg)^*$	25	–
Mouse	3	0.08
Hamster	5	0.13
Rat	6	0.16
Ferret	7	0.19
Dog	20	0.54
Monkey	12	0.32
Baboon	20	0.54

6.3 Frequentist approaches in oncology phase I trials

6.3.1 Up-and-down designs / 3 + 3 designs

Many variations exist to these so-called "up-and-down designs" / "3+3" designs, but the basic idea is:

1. 3 patients are allocated to a dose level.
   (a) If there is 0 toxicity reaction then 3 patients will be assigned to the next dose level.
   (b) If there is 1 toxicity reaction then 3 patients will be assigned to the current dose level.
   (c) If there are 2 or 3 toxicity reactions, then the current dose will be closed (too toxic) and 3 patients will be assigned to the previous dose level.
2. Continue until
   - the next dose already has had 6 patients.
   - there is no more higher/lower dose
3. The MTD is the highest dose with at most 1 toxicity reaction out of 6. (Generally, the MTD has to have data from 6 patients.)

In general there are four components to the typical dose ranging design:

1. selection of a starting dose
2. specification of the dose increments and cohort sizes
3. definition of dose limiting toxicities
   Toxicities that, due to their severity or duration, are considered unacceptable and limit further dose escalation within the subject. (These need to be pre-defined.)
4. decision rules for escalation and de-escalation

Notes:

- The starting dose is usually the lowest dose, but it does not have to be.

- The dose level for the next patient may not be known when he/she is available to be allocated.
- The best outcome is selection of the dose level that is closest but does not exceed the MTD.
- This design is not motivated with statistics in mind. The probability of selecting the right dose can be very low. The right dose may not even have the highest probability of being selected.
- Estimation of the true dose or the true probability of a toxicity reaction for a given dose is difficult.This process usually underestimates the MTD.
- The frequency of stopping escalation at a certain dose level depends on toxicity rate at that dose as well as the rates at all levels below.
- On average the dose chosen by the 3+3 design has the probability of toxicity of about 20% to 25%. The operating characteristics studied by Reiner et al. (1999); Lin and Shih (2001); Kang and Ahn (2001, 2002).

Use `Three3.R` function to simulate...
See my presentation on 3+3 designs.

Modified Fibonacci dosing
In addition to arithmetic and geometric sequence, the Fibonacci sequence is often used.

$$F_n = 0,\ 1,\ 1,\ 2,\ 3,\ 5,\ 8,\ 13,\ 21, \cdots$$

Based on this sequence, one possible dose escalation is

Dose	D	$2 \times D$	$3 \times D$	$5 \times D$	$8 \times D$	$13 \times D$	$21 \times D$
Relative increment		100%	50%	67%	60%	63%	62%
$\downarrow$							
Modified Fibonacci							
Relative increment		100%	67%	50%	40%	33%	33%
Dose	D	$2 \times D$	$3.3 \times D$	$5 \times D$	$7 \times D$	$9.3 \times D$	$12.4 \times D$

6.3.2 Example: Dose Escalation Plan

The Phase I Dose Escalation Trial of DRUG.A Administered in Combination with DRUG.B in Patients with .... Cancer

The dose of DRUG.B will remain constant while testing six escalating doses of DRUG.A. Dose escalation will proceed according to a standard 3+3 design and each cohort of 3 patients will be fully evaluated for dose limiting toxicity (DLT) before the next cohort can be enrolled. The first dose level of DRUG.A will be $20mg$. Dose escalations will proceed as follows.

Three patients will be treated at the first dose level. If none of these patients experience a DLT, a new cohort of 3 patients will be treated at the next higher dose level. If 1 patient experiences a DLT at the first dose level, a second cohort of patients will be treated at the first dose level. Dose

escalation will precede as long 0 in 3 patients or 1 or fewer patients in 6 patients treated experience a DLT. If 2 or more DLTs are observed in 3 or 6 patients treated, the MTD will be exceeded. The MTD is defined as the highest dose level tested in which fewer than 2 patients in 6 treated at that dose level experience DLT. The MTD determined in the trial will be the recommended phase II dose for subsequent testing. Since the declaration of the MTD requires 6 patients, the minimum sample size would be 6 patients and the maximum sample size required for this phase I study is 36 patients.

Phase I Dose Escalation Table of DRUG.A
DRUG.A is an oral tablet and will be administered daily for 3 consecutive days, followed by 4 days off on a 21 day cycle. The patient will receive DRUG.A on days 1, 2, 3, 8, 9, 10, 15, 16, and 17 on a 21 day cycle. The Level 1 dose will start at $20mg$ a day. We will enroll three patients per dose level and observe for DLT throughout the first cycle of therapy. Dose escalation increments will be determined by the toxicity observed at the previous dose level according to Table 1A until the MTD and DLT are defined. We will closely monitor patients for all toxicities that may appear during later cycles of treatment.

Dose Level	DRUG A dose	Number of patients
-2	$5mg$	
-1	$10mg$	
1	$20mg$	$3-6$
2	$30mg$	$3-6$
3	$45mg$	$3-6$
4	$90mg$	$3-6$
5	$140mg$	$3-6$
6	$210mg$	$3-6$

Table 6.1: The phase I dose escalation schedule of DRUG.A in combination with DRUG.B

6.4 CRM

Because of their sequential nature, “3+3” design and its variant tend to yield a biased underestimate of the target dose when estimating the MTD. One dose-finding (ranging) design that is not subject to this bias as much is the continual reassessment method (CRM).

One obvious advantage of the CRM is its use of an explicit mathematical model describing the relationship between dose and toxicity. Parameters underlying a dose-toxicity curve are given as priors. These prior values are updated sequentially and used to find the current “best” estimate of dose that would produce the acceptable risk of a toxic event.

The CRM is an algorithm for updating the best guess regarding the optimal dose. It does not require

a set of fixed dose levels. The CRM algorithms make no assumptions about

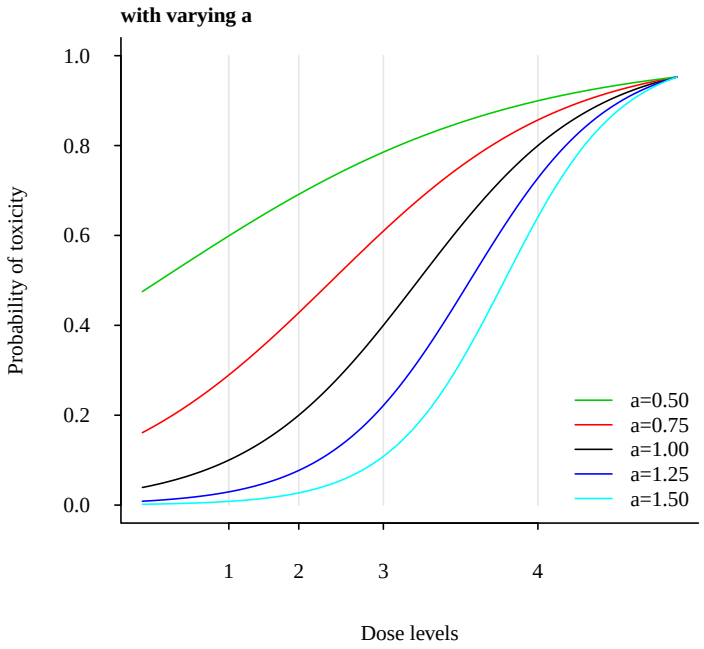
1. the actual dose used
2. the cohort size
3. ordering of doses
4. integer responses

6.4.1 Example

(Dougherty *et al.* (2000))
Suppose we want to find the dose that is associated with 20% toxicity, and the available doses are 0.25, 0.50, 0.75, and 1.00. The dose-toxicity relationship is written with a one-parameter logistic response model

$$\log\left(\frac{p_i}{1-p_i}\right) = 3 + \alpha d_i,$$

where d_i is the dose level ($i = 1, 2, 3, 4$) and p_i is the toxicity probability for dose i, and α is an unknown parameter.

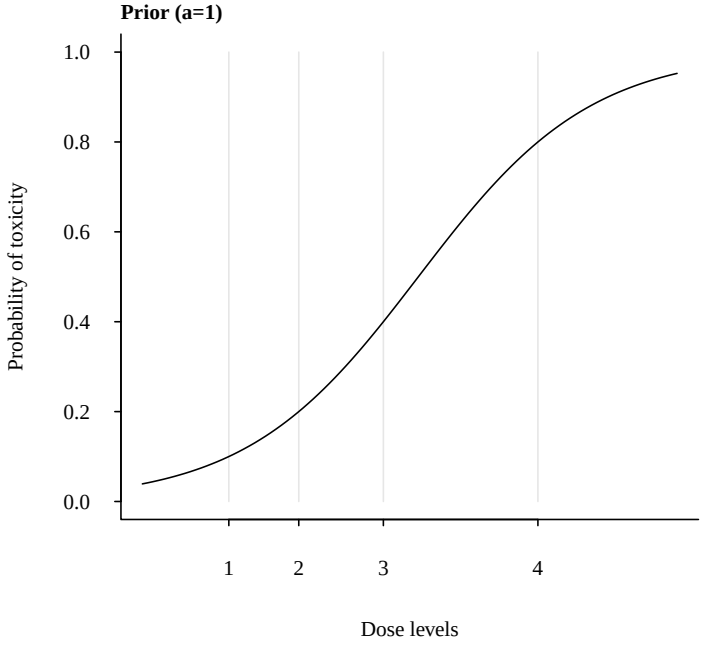


For the prior distribution of α we choose to use an exponential distribution with mean 1. The actual prior information is usually written in terms of the prior probabilities of toxicity at each dose. In this

example, let's say we have 10%, 20%, 40%, and 80% for each of the four dose levels (p_i^0). At i^{th} dose, we have

$$d_i = logit(p_i^0) - 3$$

Dose		Prior	
Level	Actual dose	p_i^0 (prior guess)	d_i
1	0.25	0.10	-5.20
2	0.50	0.20	-4.39
3	0.75	0.40	-3.41
4	1.00	0.80	-1.61



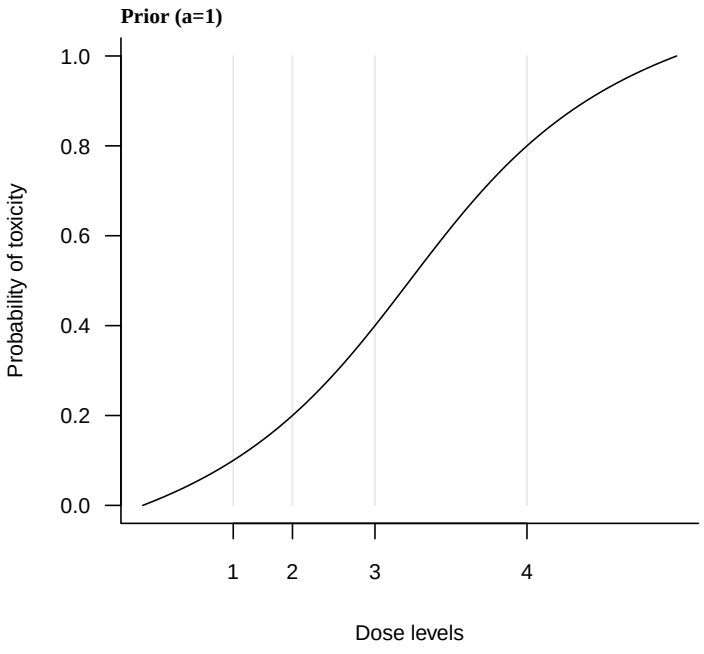
Other possible model is

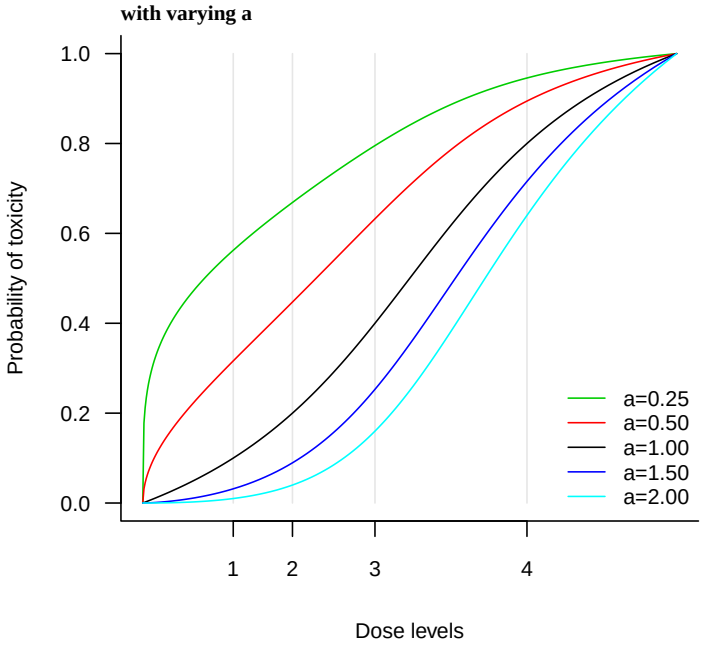
$$p_i = \{(\tan^{-1} x_i + 1)/2\}^a.$$

If we let the prior value of $a = 1$, we can solve for the corresponding x_i of each p_i,

$$x_i = \tan(2p_i^0 - 1)$$

Dose		Prior	
Level	Actual dose	p_i^0 (prior guess)	x_i
1	0.25	0.10	−1.03
2	0.50	0.20	−0.68
3	0.75	0.40	−0.20
4	1.00	0.80	0.68





Steps for updating information

1. Treat a cohort of 1 patient at the lowest dose.
2. Obtain a posterior distribution of α.
3. Find the optimal dose that gives 20% toxicity using the posterior distribution.
4. Treat another patient at the dose closest to the optimal dose.
5. Repeat steps 2, 3, 4.

The trial continues until a predetermined fixed sample size is reached, or some other trial terminating condition is satisfied.

With regard to step 2 (finding the posterior distribution), we only need the expected value of α so that it can be plugged into the dose-toxicity equation, and expectation of α is

$$E[\alpha] = \frac{\int_0^\infty \alpha L(\alpha; \tilde{d}_j, \tilde{t}_j) g(\alpha) d\alpha}{\int_0^\infty L(\alpha; \tilde{d}_j, \tilde{t}_j) g(\alpha) d\alpha},$$

where L is likelihood function for the data, defined as

$$L(\alpha; \tilde{d}_j, \tilde{t}_j) = \prod_{j=1}^{J} [\phi(d_j; \alpha)]^{t_j} [1 - \phi(d_j; \alpha)]^{1-t_j},$$

where d_j is the dose for the j^{th} patient, and $t_j = 1$ if toxicity 0 otherwise for the j^{th} patient; and

$$\phi(d_j; \alpha) = \frac{exp(3 + \alpha d_j)}{1 + exp(3 + \alpha d_j)}.$$

This is usually beyond analytical solution, so we use simulations. One result from such simulation is given below:

Dose		Prior		Observed data		Posterior	
Level	Actual dose	p_i^0 (prior guess)	d_i	# patients	# toxicity	mean	sd
1	0.25	0.10	−5.20	4	0	0.10	0.05
2	0.50	0.20	−4.39	18	3	0.19	0.08
3	0.75	0.40	−3.41	3	2	0.38	0.09
4	1.00	0.80	−1.61	0	0	0.79	0.03

- The posterior means of p_i show strong agreement with the prior.
- The actual doses do not enter into the model.
- A tolerability for dose 4 is estimated with considerable accuracy, even though no one was ever given the dose.
  - $\rightarrow$ This method should be used with great caution.

The CRM method is sometimes criticized for begin falsely precise. However, it has been demonstrated that the CRM is more efficient and less biased than classic designs.

Using R

A number of packages and functions are available in R. Over a hundred packages are listed on `cran.r-project.org/web/views/Bayesian.html`. For analysis (as opposed to design) of the data, `bcrm` (Bayesian continuous reassessment method) seems comprehensive and easy to use.

Let's continue with the same example with four dose levels. Recall the prior guess of toxicity probabilities are 10%, 20%, 40%, and 80% for each of the four dose levels. Doses are 0.25, 0.50, 0.75, 1.00.

Chapter 7

Phase II Clinical Trials

7.1 Introduction

Phase II clinical trial A clinical trial designed to test the feasibility of, and level of activity of, a new agent or procedure. (safety and activity)

Some typical characteristics of a typical phase II clinical trial include:

- It includes a placebo and two to four doses of the test drug.
- When the response is observed quickly, adaptive designs may be beneficial and used because they may
  - improve quality of estimation of the MED (minimum effective dose (lowest dose of a drug that produces the desired clinical effect).
  - increase number of patients allocated to MED.
  - allow for early stopping for futility.

The primary objectives of phase II trials are:

- To determine whether the drug is worthy of further study in phase III trial. Significant treatment effect? / dose-response relationship?
- To gather information to help design phase III trial.
  - Determine dose(s) to carry forward
  - Determine the primary and secondary endpoints
  - Estimate treatment effects for power/sample size analysis
  - Estimate recruitment rate
  - Examine feasibility of treatment (logistics of administration and cost)
  - Learn about side effects and toxicity

In phase II clinical trials, parallel group designs, crossover designs, and factorial designs are often

used.

7.2 Phase II trials in oncology

A phase II clinical trial in oncology generally uses a fixed dose chosen in a phase I trial. The primary objective is to assess therapeutic response to treatment. In the simplest case, a single treatment arm is compared to a historical control. In other cases, a control group and/or multiple doses are included.

The treatment efficacy is often evaluated on surrogate markers for a timely (quick) evaluation of efficacy.

Surrogate outcome An outcome measurement in a clinical trial that substitutes for a definitive clinical outcome or disease status.

- CD4 counts in AIDS study.
- PSA (prostatic specific antigen) in prostate cancer study.
- Blood pressure in cardiovascular disease.
- 3 months survival (binary) for survival.
- Tumor shrinkage for survival.

Tumor response to treatment is evaluated according to Response Evaluation Criteria in Solid Tumors (RECIST)

Complete response (CR) Disappearance of all target lesions.

Partial response (PR) At least a 30% decrease in the sum of the longest diameter (LD) of target lesions, taking as reference the baseline sum LD.

Stable disease (SD) Neither sufficient shrinkage to qualify for PR nor sufficient increase to qualify for PD, taking as reference the smallest sum LD since the treatment started.

Progressive disease (PD) At least a 20% increase in the sum of the LD of target lesions, taking as reference the smallest sum LD recorded since the treatment started or the appearance of one or more new lesions.

Generally, objective tumor response is defined as CR or PR in RECIST so that the response variable has a binary endpoint. In the rest of chapter, we will consider a single arm trial with a binary response. The hypothesis of interest is one-sided $H_1 : p > p_0$, and the type I error rate is usually 5 to 10%. The power is usually 80 to 90%.

7.3 Classical (old) two-stage designs

It is crucial that these phase II studies have an opportunity to stop early for toxicity, and that is accomplished by Data Monitoring Committee (DMC), aka, Data and Safety Monitoring Board (DSMB). It is also desired to discard ineffective treatment early, and two-stage designs with a futility stop has been popular.

We will discuss the designs proposed by Gehan (1961), Fleming (1982), and Simon (1989), using the following unified notation:

- stage I sample size $\cdots n_1$.
- stage I data $\cdots X_1 \sim Binomial(n_1, p)$.
- stage I critical value $\cdots r_1$ so that if $X_1 \leq r_1$ then terminate the study for futility.
- stage II sample size $\cdots n_2$.
- stage II data $\cdots X_2 \sim Binomial(n_2, p)$.
- total sample size $\cdots n_t = n_1 + n_2$.
- total data $\cdots X_t \equiv X_1 + X_2$.
- stage II critical value $\cdots r_t$ so that if $X_t \leq r_t$ then terminate the study for futility, otherwise conclude efficacy.

7.3.1 Gehan's design

It is old (1961) and outdated but may be ok to use in limited situations. The design calls for the first stage with $n_1 = 14$ and $r_1 = 0$, i.e., if no positive response is observed in 14, then stop for futility. The rational is that if true response rate is at least 20%, then $X_1 = 0$ is unlikely. In fact, it is 0.044. The second stage sample size depends on the desired precision for estimating p, and it ranges between 1 and 86. A typical n_2 is 14 so that $n_t = 28$.

7.3.2 Fleming's design

Fleming (1982) proposed a multistage design for phase II clinical trials. One of its key characteristics is stopping early for efficacy.

Example
$H_0 : p = 0.15$, $H_1 : p = 0.30$. (powered at 0.30)
$\alpha = .05$, $\beta = .2$
(Reject H_0 in stage 1 if $X_1 \geq s_1$.

n_1	r_1	s_1	n_t	r_t	α	$1-\beta$	$E_0[N]$	$E_1[N]$
29	4	9	47	10	0.0490	0.8013	36.6	36.9

7.4 Simon's design

In his 1989 paper, Simon introduced two criteria to choose a 2 stage design for single arm and one sided test. The optimal design has the smallest expected sample size under H_0 ($n_1 + E_{p_0}[n_2]$), and the minimax design has the smallest total sample size ($n_1 + n_2$). For $p_0 = 0.15$ and $p_1 = 0.30$,

	n_1	r_1	n_t	r_t	α	$1-\beta$	$E_0[N]$	pet_0	$E_1[N]$	pet_1
optimal	19	3	55	12	0.048	0.801	30.4	0.68	50.2	0.13
minimax	23	3	48	11	0.046	0.804	34.5	0.54	46.7	0.05
single stage	$--$	$--$	48	11	0.048	0.819	48.0	0.00	48.0	0.00

7.4.1 Conditional power

To find a good design (sample sizes and critical values), we need to understand the *conditional* power of a design. The conditional power is the probability of rejecting H_0 (in stage 2) given the stage 1 result, i.e., conditioned on $X_1 = x_1$. Clearly, when $X_1 > r_t$, conditional power is 1, and when $X_1 \leq r_1$ (futility stop), conditional power is 0.

$$\begin{aligned} CP(x_1) &= P[\text{Reject in stage 2}|x_1] = P[x_1 + X_2 > r_t|x_1] \\ &= P[X_2 > r_t - x_1|x_1] \\ &= \sum_{x_2=r_t-x_1+1}^{n_2} \binom{n_2}{x_2} p^{x_2}(1-p)^{n_2-x_2} \end{aligned}$$

Conditional power is a function of p, x_1 and n_2 as well as r_t

To obtain the unconditional power, we need to integrate (sum) the conditional power over all possible x_1 values.

$$\rho(p) = \sum_{x_1=0}^{n_1} CP(x_1) P_p[X_1 = x_1]$$

$$\rho(p) = \sum_{x_1=r_1+1}^{n_1} CP(x_1) \binom{n_1}{x_1} p^{x_1}(1-p)^{n_1-x_1}.$$

Given α and β a design is chosen so that $\rho(p_0) \leq \alpha$ and $\rho(p_1) \geq 1-\beta$.

Unlike in a single-stage situation, there may be more than one *good* design. Simon used the *optimal* and *minimax* to choose two reasonable designs among many satisfying the type I error rate and power constraints.

Expected sample size under the null can be written as

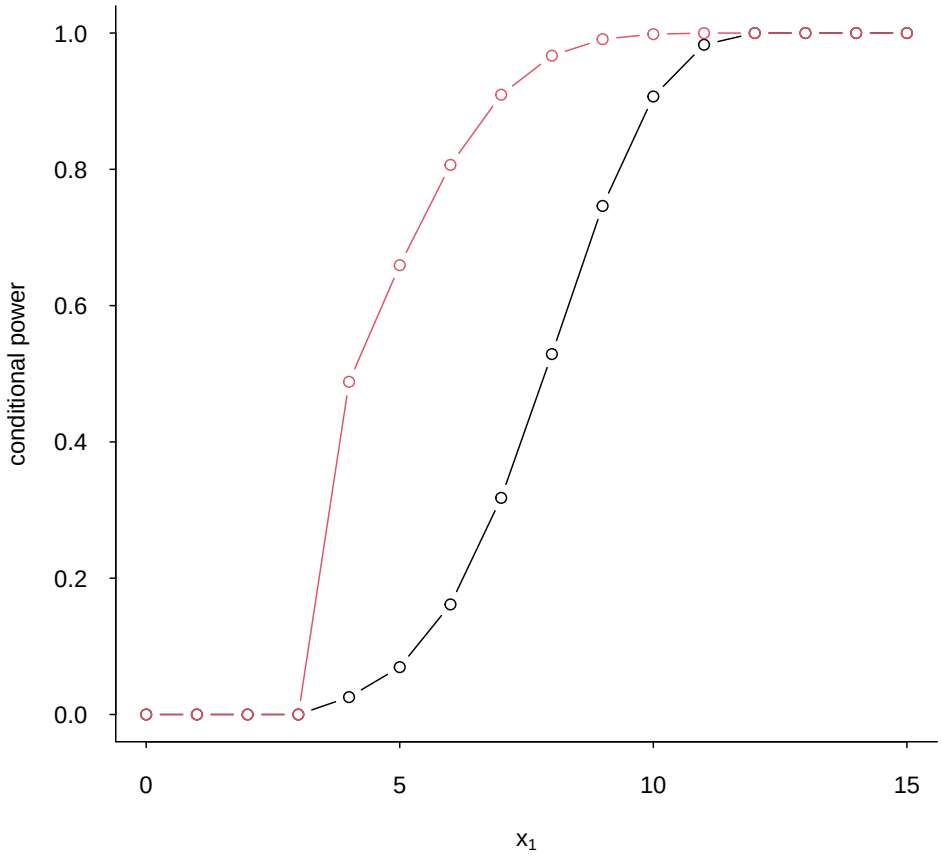
$$\begin{aligned} E_{p_0}[n_t] &= n_1 + n_2 P[\text{continue to stage 2}|p_0] \\ &= n_1 + n_2 \times P[X_1 > r_1|p_0] \\ &= n_1 + n_2 \times \sum_{x_1=r_1+1}^{n_1} \binom{n_1}{x_1} p_0^{x_1}(1-p_0)^{n_1-x_1}. \end{aligned}$$

7.4.2 Computing design characteristics

```
simon.d <- function(n1, r1, nt, rt, p0, p1, pl = TRUE, simple = FALSE) {
    # x1 <= r1 stop for futility xt <= rt conclude futility
    R4 <- function(x) {
        round(x, 4)
    }
    R1 <- function(x) {
        round(x, 1)
    }

    x1 <- 0:n1
    pst1.0 <- dbinom(x1, n1, p0)
    pst1.1 <- dbinom(x1, n1, p1)

    cp0 <- 1 - pbinom(rt - x1, nt - n1, p0)
    cp1 <- 1 - pbinom(rt - x1, nt - n1, p1)
    cp0[x1 <= r1] <- 0
    cp1[x1 <= r1] <- 0
    cp0[x1 > rt] <- 1
    cp1[x1 > rt] <- 1

    pow0 <- sum(pst1.0 * cp0)
    pow1 <- sum(pst1.1 * cp1)

    keep <- pmax(pst1.0, pst1.1) > 9e-05
    out1 <- data.frame(x1 = R4(x1), pst1.0 = R4(pst1.0), pst1.1 = R4(pst1.1),
        cp0 = R4(cp0), cp1 = R4(cp1))[keep, ]

    pet0 <- pbinom(r1, n1, p0)
```

```
    pet1 <- pbinom(r1, n1, p1)
    en0 <- n1 + (1 - pet0) * (nt - n1)
    en1 <- n1 + (1 - pet1) * (nt - n1)

    out2 <- data.frame(n1, r1, nt, rt, p0 = formatC(p0, digit = 2, format = "f"),
        p1 = formatC(p1, digit = 2, format = "f"))
    out3 <- data.frame(pow0 = R4(pow0), pow1 = R4(pow1), en0 = R1(en0),
        en1 = R1(en1), pet0 = R4(pet0), pet1 = R4(pet1))
    if (pl) {
        plot(out1$x1, out1$x1, type = "n", las = 1, ylim = c(0, 1), bty = "L",
            xlab = expression(x[1]), ylab = "conditional power")
        lines(out1$x1, out1$cp0, col = 1, type = "b")
        lines(out1$x1, out1$cp1, col = 2, type = "b")
    }
    out <- list(out1, out2, out3)
    if (simple)
        out <- list(out2, out3)

    out
}

simon.d(n1 = 23, r1 = 3, nt = 48, rt = 11, p0 = 0.15, p1 = 0.3)


[[1]]
   x1 pst1.0 pst1.1    cp0    cp1
1   0 0.0238 0.0003 0.0000 0.0000
2   1 0.0966 0.0027 0.0000 0.0000
3   2 0.1875 0.0127 0.0000 0.0000
4   3 0.2317 0.0382 0.0000 0.0000
5   4 0.2044 0.0818 0.0255 0.4882
6   5 0.1371 0.1332 0.0695 0.6593
7   6 0.0726 0.1712 0.1615 0.8065
8   7 0.0311 0.1782 0.3179 0.9095
9   8 0.0110 0.1527 0.5289 0.9668
10  9 0.0032 0.1091 0.7463 0.9910
11 10 0.0008 0.0655 0.9069 0.9984
12 11 0.0002 0.0332 0.9828 0.9999
13 12 0.0000 0.0142 1.0000 1.0000
14 13 0.0000 0.0052 1.0000 1.0000
15 14 0.0000 0.0016 1.0000 1.0000
16 15 0.0000 0.0004 1.0000 1.0000
```

```
[[2]]
  n1 r1 nt rt   p0   p1
1 23  3 48 11 0.15 0.30

[[3]]
    pow0   pow1  en0  en1   pet0   pet1
1 0.0455 0.8035 34.5 46.7 0.5396 0.0538
```



Given a design, computing operational characteristics such as type I error rate, power, expected sample size is not difficult; however, solving for the optimal, minimax, and other preferable designs is not trivial. Simon's original papers show how to do this.

A very good webpage by Anastasia Ivanova at UNC is at `http://cancer.unc.edu/biostatistics/program/ivanova/SimonsTwoStageDesign.aspx`.

7.4.3 Something in between

The two criteria, optimal and minimax, give two designs that are extreme, and neither may fit the investigators' needs. For example, for testing $H_0 : p = 0.3$ with $\alpha = 0.05$ and $\beta = 0.10$ at $p_1 = 0.45$, the optimal design and minimax designs are:

	n_1	r_1	n_t	r_t	α	$1-\beta$	$E_0[N]$	pet_0
optimal	40	13	110	40	0.048	0.901	60.8	0.70
balanced	53	18	106	39	0.043	0.903	64.4	0.78
minimax	77	27	88	33	0.050	0.901	78.5	0.86

The optimal design tends to have a small n_1 and the minimax design tends to have a large n_1. Therefore, a simple approach to find a good alternative design is to force $n_1 = n_2$. (balanced design of Ye and Shyr, 2007)

```
simon.d(n1 = 40, r1 = 13, nt = 110, rt = 40, p0 = 0.3, p1 = 0.45, pl = FALSE,
    simple = TRUE)


[[1]]
  n1 r1  nt rt   p0   p1
1 40 13 110 40 0.30 0.45

[[2]]
    pow0   pow1  en0   en1   pet0   pet1
1 0.0482 0.9012 60.8 104.7 0.7032 0.0751


simon.d(n1 = 53, r1 = 18, nt = 106, rt = 39, p0 = 0.3, p1 = 0.45, pl = FALSE,
    simple = TRUE)


[[1]]
  n1 r1  nt rt   p0   p1
1 53 18 106 39 0.30 0.45

[[2]]
    pow0   pow1  en0   en1   pet0   pet1
1 0.0431 0.9028 64.4 102.4 0.7844 0.0687
```

A more systematic approach is to express the criteria for optimization as

$$q(w) = w \times (n_t) + (1 - w) \times E_0[N],$$

where $0 \leq w \leq 1$. $q(0)$ and $q(1)$ correspond to the optimal and minimax designs, respectively. Computation shows that the minimax design is the best design with respect to $q(w)$ for $w \in (0.827, 1]$.

In between the optimal and minimax designs, the following "admissible" designs exist that optimize $q(w)$ for certain ranges of w. (Jung, Lee, Kim, George, 2004)

	n_1	r_1	n_t	r_t	α	$1-\beta$	$E_0[N]$	pet_0	w
optimal	40	13	110	40	0.048	0.901	60.8	0.70	$(0, 0.006)$
admissible 1	43	14	104	38	0.050	0.903	60.8	0.70	$(0.006, 0.136)$
admissible 2	48	16	101	37	0.050	0.901	61.3	0.75	$(0.136, 0.182)$
admissible 3	40	12	94	35	0.048	0.902	62.8	0.58	$(0.182, 0.303)$
admissible 4	46	14	91	34	0.049	0.902	64.1	0.60	$(0.304, 0.827)$
minimax	77	27	88	33	0.050	0.901	78.5	0.86	$(0.827, 1)$
single stage	$--$	$--$	90	34	0.045	0.900	90.0	0.00	

7.5 Data analysis following a two-stage design in phase II clinical trials

The primary objective of a (cancer) phase II clinical trial is to make a correct "go/no-go" decision; however, making a good inference for p is advantageous for planning the following phase III trial.

We have seen before that when we terminate a study based on an interim summary of the data, a usual statistic that we often compute may be biased. In this section, we will look at the issue of bias in two-stage design in phase II clinical trial in detail. Simon's design will be our focus, but many general discussions can be applied to other designs as well.

7.5.1 p-value

If we ignore the fact that the data were gathered in a two-stage design and compute a p-value as if $X \sim Binomial(n_t, p)$, it is bigger than the true p-value with the following definition/interpretation.

p-value the probability under the null hypothesis that we would observe the data *as or more extreme* than what we have observed

The term "as or more extreme" can be interpreted as "as big or bigger evidence against H_0". In a simple single-stage design, the meaning of this is usually straightforward. We can all agree that $Z = 2.0$ is more extreme (more evidence against H_0) than $Z = 1.9$. However, in two-stage designs, understanding the definition of p-value sometimes gets tricky.

Example:
$H_0 : p = 0.3$, $H_1 : p > 0.3$; $\alpha = 0.05$ and the power is 0.80 at $p = 0.5$. Then the optimal design is: $n_1 = 15$, $r_1 = 5$, $n_t = 46$, $r_t = 18$.

```
simon.d(n1 = 15, r1 = 5, nt = 46, rt = 18, p0 = 0.3, p1 = 0.5, pl = FALSE,
    simple = TRUE)


[[1]]
  n1 r1 nt rt   p0   p1
1 15  5 46 18 0.30 0.50

[[2]]
    pow0   pow1  en0  en1   pet0   pet1
1 0.0499 0.8032 23.6 41.3 0.7216 0.1509
```

Now suppose we observe $X_1 = 7$ in stage 1 so that we move on to the second stage. And in stage 2, we observe additional 12 positive responses in $n_2 = 31$ patients (19 in 46 total) so that H_0 is rejected because $X_t = 19 > r_t$.

If we compute a p-value without taking into account the study design, we might use $X \sim Binomial(46, 0.3)$ and compute

$$p_{\mathsf{c}} = P_0[X \geq 19] = \sum_{i=19}^{46} \binom{46}{i} 0.3^i (1 - 0.3)^{46-i}$$

where p_{c} is a *conventional* p-value. H_0 is rejected but this p-value is greater than α as shown below:

```
1 - pbinom(18, 46, 0.3)


[1] 0.06805
```

To see this inconsistency clearly, we will rewrite above as

$$\begin{aligned} p_{\mathsf{c}} &= P_0[X \geq 19] \\ &= \sum_{x_1=0}^{15} P_0[X_2 \geq 19 - x_1 | X_1 = x_1] P_0[X_1 = x_1]. \end{aligned}$$

From this expression we see that in computing p_{c}, we include sample paths that can not be realized with this Simon's design, namely, $X_1 = 0$, $X_2 \geq 19$; $X_1 = 1$, $X_2 \geq 18$; $\cdots$; $X_2 = 5$, $X_2 \geq 14$. A *proper* p-value that takes into account the actual sampling scheme used may be

$$p_{\mathsf{p}} = \sum_{x_1=6}^{15} P_0[X_2 \geq 19 - x_1 | X_1 = x_1] P_0[X_1 = x_1].$$

In general, for Simon-like two-stage designs, p-value should be calculated

$$p_{\mathsf{p}} = \sum_{x_1=r_1+1}^{n_1} P_0[X_2 \geq x_t - x_1 | X_1 = x_1] P_0[X_1 = x_1],$$

if $x_1 > r_1$ (i.e., if there is a second stage).

The following simple R script computes this p-value:

```
pp <- function(n1, r1, nt, rt, x1, xt, p0) {
    x1v <- (r1 + 1):n1
    p.val <- sum((1 - pbinom(xt - x1v - 1, (nt - n1), p0)) * dbinom(x1v,
        n1, p0))
    pc <- 1 - pbinom(xt - 1, nt, p0)
    if (x1 <= r1) {
        p.val <- pc <- 1 - pbinom(x1 - 1, n1, p0)
    }
    c(p.val = p.val, pc = pc)
}

pp(n1 = 15, r1 = 5, nt = 46, rt = 18, x1 = 7, xt = 19, p0 = 0.3)
```

```
  p.val      pc
0.04987 0.06805
```

When $x_1 \leq r_1$ so that the trial is terminated in stage 1, we can define

$$p_{\mathsf{p}} = P_0[X_1 \geq x_1].$$

Thus we think that "moving on to the second stage" has more evidence against H_0 than "terminating in the first stage for futility", which makes sense.

The *proper* p-value (p_{p}) has the following characteristics:

- It is always smaller than or equal to p_{c}.
- It is consistent with the hypothesis testing, i.e., $p_{\mathsf{p}} \leq \alpha$ if and only if H_0 is rejected.
- If $X_t = r_t + 1$, then p_{p} is equal to the level of the test (so-called the *actual* type I error rate).
- It does not distinguish different sample paths that lead to the same X_t. That is, evidence against H_0 is identical if x_t is the same regardless of x_1.
  For example, $X_1 = 8$, $X_2 = 12$ and $X_1 = 10$, $X_2 = 10$ yield the same p-values.

When does this (p_{p}) break down?
It breaks down when we allow n_2 to be different for various values of X_1. In some modifications of

Simon's design (e.g., Banerjee A, Tsiatis AA. Stat Med 2006), the stage 2 sample size varies with x_1. Then, p_{p} can not be computed because we cannot order the sample paths simply based on X_t.

A bigger concern is that this p_{p} cannot be used when n_2 is changed from that planned. An even bigger concern is if the actual n_2 is different from that planned, how can we re-compute the critical value, r_t, to control type I error rate? The answer is not simple!

7.5.2 Point estimate

Because the results from a phase II clinical trial are often used in planning a phase III clinical trial, a good estimate of p is often of interest.

MLE

In a single stage design, the MLE of p is $\hat{p} = x/n$. For a Simon's design, we can write the likelihood, letting Y_i denote the individual datum from a $Bernoulli(p)$ population, as follows:

$$L(p|Y) = \begin{cases} \Pi_{i=1}^{n_1} p^{y_i}(1-p)^{1-y_i} & \text{if } \sum_i^{n_1} y_i \le r_1 \\ \Pi_{i=1}^{n_t} p^{y_i}(1-p)^{1-y_i} & \text{if } \sum_i^{n_1} y_1 > r_1 \end{cases}$$

$$l(p|X) = \begin{cases} x_1 log(p) + (n_1 - x_1) log(1-p) & \text{if } x_1 \le r_1 \\ x_t log(p) + (n_t - x_t) log(1-p) & \text{if } x_1 > r_1 \end{cases}$$

Therefore, the MLE for π is

$$\hat{p}(x) = \begin{cases} x_1/n_1 & \text{if } x_1 \le r_1 \\ x_t/n_t & \text{if } x_1 > r_1 \end{cases}$$

We have seen before that this $\hat{p}(x)$ has a downward bias, i.e., $E_p[\hat{p}(x)] \le p$. A simple explanation is that when $\hat{p}$ is small at the end of stage 1, we tend to terminate the study, and this downward bias tends to remain; however when $\hat{p}$ is large at the end of stage 1, more data are gathered and the upward bias of stage 1 tends to be corrected.

Example: $p_0 = 0.3$, $p_1 = 0.5$, $\alpha = 0.05$, $\beta = 0.2$. Then the minimax design is ($n_1 = 19$, $r_1 = 6$, $n_t = 39$, $r_t = 16$). Further suppose $X_1 = 8$ and $X_2 = 12$ so that $X_t = 20$.

$$\hat{p} = \frac{20}{39} = 0.513.$$

Whitehead

We can write the bias of the MLE estimator as:

$$B(p) = E_p[\hat{p}(x)] - p.$$

So a good estimator would be

$$\breve{p} = \hat{p} - B(p).$$

However, $B(p)$ is unknown, so we need to estimate it. Let's use the current estimate of p in $B(p)$. That is

$$\hat{p}_w = \hat{p} - B(\hat{p}_w).$$

This is Whitehead's estimator (1986 Biometrika). We can write

$$\hat{p}_w = \hat{p} - E_{\hat{p}_w}[\hat{p}(x)] + \hat{p}_w,$$

which leads to

$$E_{\hat{p}_w}[\hat{p}(x)] = \hat{p}.$$

To find $\hat{p}_w$, we need to numerically solve for $\hat{p}_w$ that satisfies

$$\begin{aligned} E_{\hat{p}_w}[\hat{p}(x)] &= \sum_{x_1=0}^{r_1} \frac{x_1}{n_1} P[X_1 = x_1 | p = \hat{p}_w] + \sum_{x_1=r_1+1}^{n_1} \sum_{x_2=0}^{n_2} \frac{x_1 + x_2}{n_t} P[X_1 = x_1 | p = \hat{p}_w] P[X_2 = x_2 | p = \hat{p}_w] \\ &= \hat{p} \end{aligned}$$

In the current example, $\hat{p}_w = 0.520$.

Koyama

We can write the bias of the MLE estimator as:

$$B(p) = E_p[\hat{p}(x)] - p.$$

So a good estimator would be

$$\breve{p} = \hat{p} - B(p).$$

However, $B(p)$ is unknown, so let's use $B(\hat{p})$, that is

$$\hat{p}_k = \hat{p} - B(\hat{p}).$$

This is simpler and more straightforward than Whitehead's estimator. We can write

$$\begin{aligned}\hat{p}_k &= \hat{p} - E_{\hat{p}}[\hat{p}(x)] + \hat{p} \\ &= 2\hat{p} - E_{\hat{p}}[\hat{p}(x)].\end{aligned}$$

Solving for $\hat{p}_k$ is considerably easier. First compute

$$E_{\hat{p}}[\hat{p}(x)] = \sum_{x_1=0}^{r_1} \frac{x_1}{n_1} P[X_1 = x_1 | p = \hat{p}] + \sum_{x_1=r_1+1}^{n_1} \sum_{x_2=0}^{n_2} \frac{x_1 + x_2}{n_t} P[X_1 = x_1 | p = \hat{p}] P[X_2 = x_2 | p = \hat{p}],$$

then subtract it from $2\hat{p}$. In the current example, $\hat{p}_k = 0.521$.

Unbiased estimator

For a general multistage design with early stopping for futility and efficacy, Jung and Kim (2004 Stat Med) found the unbiased estimator of p. They showed that the pair (M,S), where M is the number of stage (when terminated) and S the number of successes, is complete and sufficient for p. And clearly x_1/n_1 is unbiased for p, the uniformly minimum variance unbiased estimator (UMVUE) is found through Rao-Blackwell theorem.

The expression for $\hat{p}_{ub}$ is complex, but for Simon's two-stage design (two-stage with only futility stop), it can be written as

$$\hat{p}_{ub} = \frac{\sum_{x_1=(r_1+1)\vee(x_t-n_2)}^{n_1 \wedge x_t} \binom{n_1-1}{x_1-1}\binom{n_2}{x_t-x_1}}{\sum_{x_1=(r_1+1)\vee(x_t-n_2)}^{n_1 \wedge x_t} \binom{n_1}{x_1}\binom{n_2}{x_t-x_1}},$$

where $a \wedge b = min(a, b)$ and $a \vee b = max(a, b)$.

For the current example, $max(r_1 + 1,\, x_t - n_2) = max(6 + 1,\, 20 - 20) = 7$, and $min(n_1,\, x_t) = min(19, 20) = 19$, and

$$\begin{aligned}\hat{p}_{ub} &= \frac{\sum_{x_1=7}^{19} \binom{18}{x_1-1}\binom{20}{20-x_1}}{\sum_{x_1=7}^{19} \binom{19}{x_1}\binom{20}{20-x_1}} \\ &= 0.517.\end{aligned}$$

Median estimator

Another simple estimator is the value, p_0^* such that the p-value for testing $H_0 : p = p_0^*$ is 0.5 by the realized sample path. Many adaptive designs for phase II clinical trials were originally motivated as a hypothesis testing procedure, and computing this estimator should be fairly simple in many designs.

If the test statistic is continuous, this estimator is known as the median unbiased estimator (Cox and Hinkley 1974). It is unbiased for the true median. The proof uses the fact that the p-value is distributed $Unif(0,1)$ under H_0.

We need to find p_0^* such that

$$\begin{aligned} p_{\mathsf{p}} &= \sum_{x_1=r_1+1}^{n_1} P_{p_0^*}[X_2 \geq x_t - x_1 | X_1 = x_1] P_{p_0^*}[X_1 = x_1] \\ &= \sum_{x_1=7}^{19} P_{p_0^*}[X_2 \geq 20 - x_1 | X_1 = x_1] P_{p_0^*}[X_1 = x_1] \\ &= 0.5. \end{aligned}$$

$p_0^* = 0.500$.

Comparisons

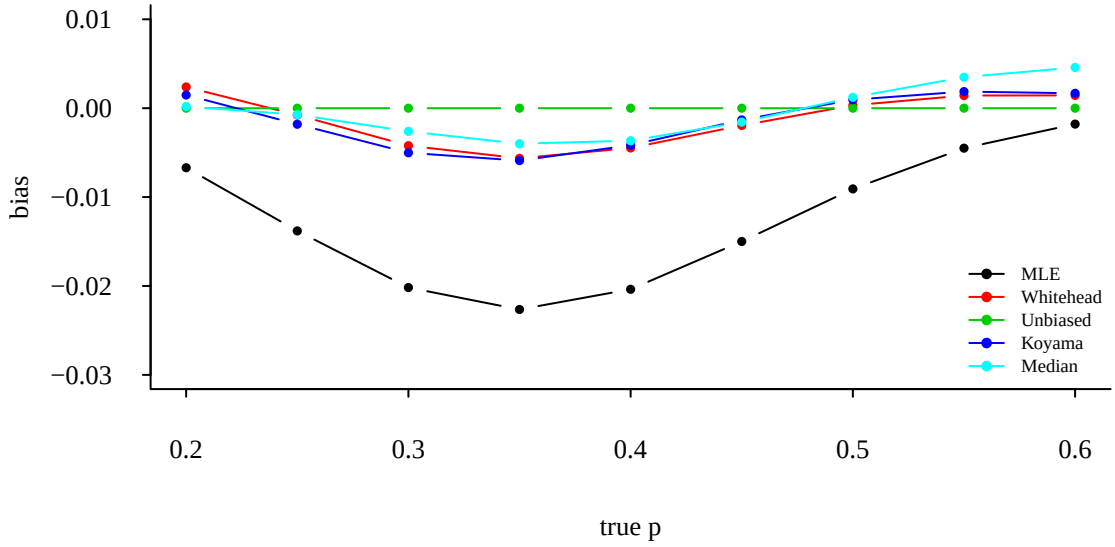
To compare these methods, we compute the bias of each estimator for various true values of p. Use bias and mean squared error = var + bias^2 to compare them. For each estimator, compute $\hat{p}(X)$ for every sample path (defined by X in $[0, n_t]$) and compute

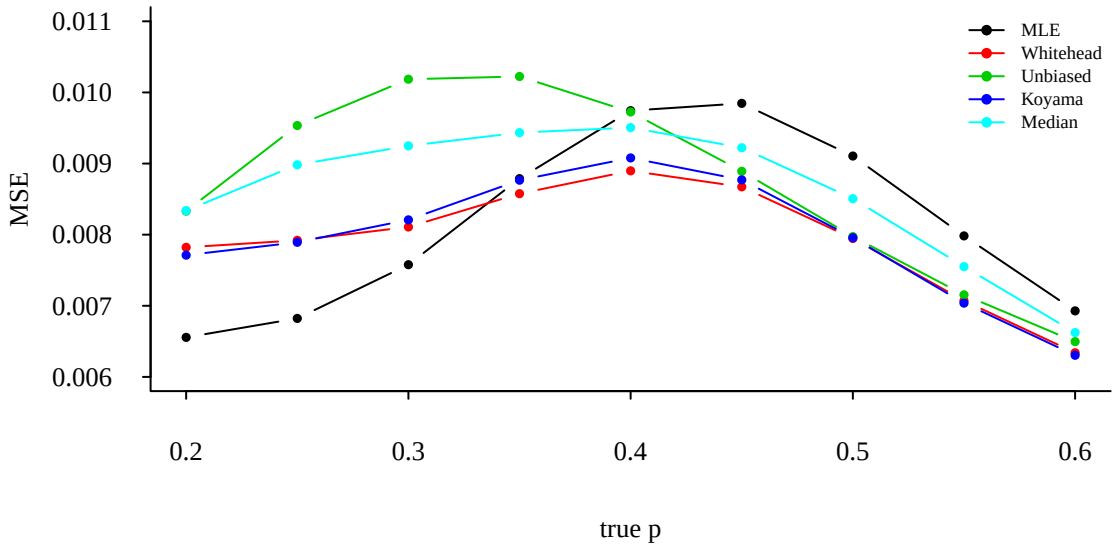
$$E_p[\hat{p}(X)] = \sum_{x=0}^{n_t} \hat{p}(x) P_p[X = x].$$

Mean squared errors can be computed by

$$\begin{aligned} MSE_p[\hat{p}(X)] &= E_p[(\hat{p}(X) - p)^2] \\ &= \sum_{x=0}^{n_t} (\hat{p}(x) - p)^2 P_p[X = x]. \end{aligned}$$

The following two plots show bias and MSE for the current example.


0.01
0.00
−0.01
−0.02
−0.03
bias
0.2
0.3
0.4
0.5
0.6
true p
MLE
Whitehead
Unbiased
Koyama
Median


0.011
0.010
0.009
0.008
0.007
0.006
MSE
0.2
0.3
0.4
0.5
0.6
true p
MLE
Whitehead
Unbiased
Koyama
Median

Chapter 8

Non-inferiority

8.1 Introduction

In a phase III clinical trial, the objective may be to show the experimental treatment is *non-inferior* to the *active control*.

Active control is a conventional treatment that has been shown to be effective.

Non-inferiority trial aims to show that the experimental treatment is clinically and statistically not inferior in effectiveness compared to an active control.

When there is no proven treatment, placebo-controlled trials are generally not controversial; however, when a proven effective treatment exists, the ethics of placebo-controlled clinical trials are questionable.

> Declaration of Helsinki
> (`https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medi`
> is a set of ethical principles regarding human experimentation developed for the medical community by the World Medical Association. It is regarded as the cornerstone document on human research ethics. (wikipedia)

Article II.3 of Declaration of Helsinki stated (until 2000), “In any medical study, every patient - including those of a control group, if any- should be assured of the best proven diagnostic and therapeutic method. This does not exclude the use of inert placebo studies where no proven diagnostic or therapeutic methods exists”.

Currently, paragraph 33 states, “The benefits, risks, burdens and effectiveness of a new intervention must be tested against those of the best proven intervention(s), except in the following circumstances:

- Where no proven intervention exists, the use of placebo, or no intervention, is acceptable; or
- Where for compelling and scientifically sound methodological reasons the use of any intervention less effective than the best proven one, the use of placebo, or no intervention is necessary to determine the efficacy or safety of an intervention
- ...

The terms, ”equivalence” and “non-inferiority” are sometimes used interchangeably; however, their objectives are different.

8.2 Hypotheses and multiplicity

Superiority The objective is to show that the test treatment is better than the control.

$$H_0 : \mu_t - \mu_c = 0$$
$$H_1 : \mu_t - \mu_c > 0$$

Power is set at some $\delta_s > 0$.

Non-inferiority The objective is to show that the test treatment is not inferior to the control. (“Not much worse than”)

$$H_0 : \mu_t - \mu_c = -\delta_I$$
$$H_1 : \mu_t - \mu_c > -\delta_I$$

for some $\delta_I > 0$. Power is usually set at 0. The value, δ_I is called, “non-inferiority margin” and needs to be specified a priori.

Equivalence The objective is to show that the test treatment is equivalent to the control. This is rarely used for efficacy, but in studies of pharmacokinetics, equivalence hypothesis testings may be useful (Bioequivalence).

$$H_0 : \mu_t - \mu_c < -\delta_e \text{ or } \delta_e < \mu_t - \mu_c$$
$$H_1 : -\delta_e \leq \mu_t - \mu_c \leq \delta_e$$

for some $\delta_e > 0$. Power is usually set at 0. Sample size for an equivalence trial ($\delta_e = \delta$) with α and β (power at 0) is equal to that for a superiority test with α and $\beta/2$ (power at δ).

Equivalence trials are inherently different from the other two; superiority and non-inferiority can be asked sequentially.

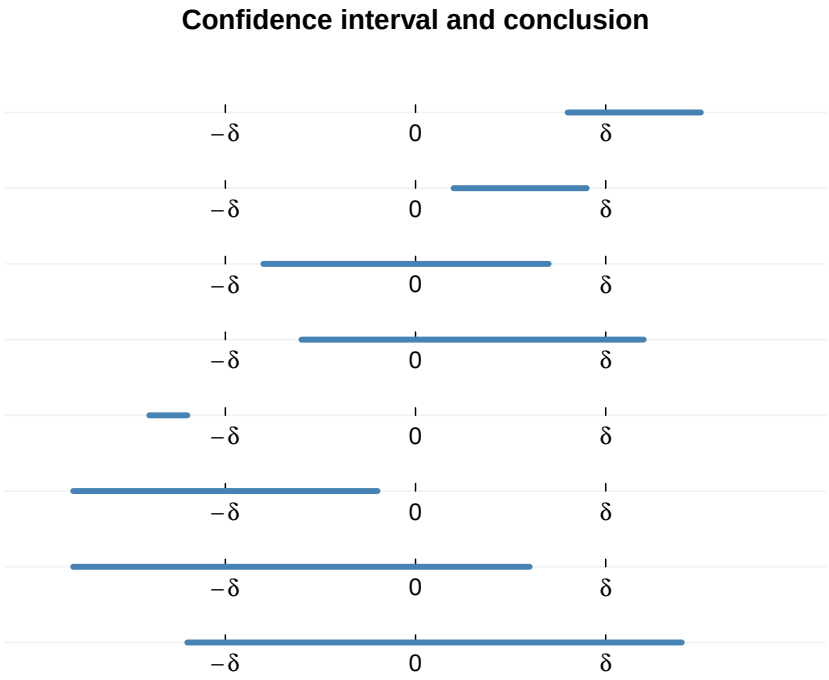
- Now that we have shown `T` is non-inferior to `C`, can we test superiority using the same data?
- We did not have enough evidence to say `T` is better than `C`, can we try and at least say `T` is non-inferior to `C`?

There is no penalty in terms of type I error rate inflation, and superiority should always be tested after non-inferiority is established. Non-inferiority can be tested when superiority is not shown **if** the non-inferiority margin is defined a priori and stated in the protocol.

There is no multiplicity when testing both superiority and non-inferiority because

- only one (superiority or non-inferiority) can be true.
- we can test both hypotheses with 1 confidence interval.

Conclusion / Truth	Superiority	Non-inferiority but not superiority	Inferiority
Superiority	Correct	type II error	type II error
Non-inferiority but not superiority	type I error	Correct	type II error
Inferiority	type I error	type I error	Correct



8.3 Unique problems with non-inferiority trials

Choice of non-inferiority margin

- The non-inferiority margin must not be larger than the control treatment effect. Otherwise, the new treatment may be as effective as placebo and yet non-inferior to the active control.
- The non-inferiority margin should be based on both statistical reasoning and clinical judgement.

The first point will require us to be conservative in selecting the non-inferiority margin (pick a small δ_I). This leads to a large sample size.

`http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/`

WC500003636.pdf

Assay sensitivity is a property of a clinical trial defined as the ability of a trial to distinguish an effective treatment from a less effective or ineffective intervention.

In a superiority test, assay sensitivity is not an issue because concluding superiority automatically establishes assay sensitivity. However, in a non-inferiority trial, showing non-inferiority does not show assay sensitivity. In other words, T may be non-inferior to C, but it may not be better than a placebo. Including a placebo arm will address the issue if it is ethical to do so.

Assay sensitivity is not shown but assumed.

- Historical evidence of sensitivity to drug effects.
  - Appropriately designed and conducted trials in a particular disease with a specific active drug reliably show an effect. This is an abstract concept about trials of a drug in a particular disease whereas assay sensitivity is a characteristic of a particular trial.
- There are some conditions where treatment responses are large and clearly greater than placebo effect, and there are other cases evidence from many studies are consistent.
  - However, this may not be applicable to the new clinical trial.
- High quality study. Factors that may reduce assay sensitivity includes
  - Poor compliance
  - Crossovers (switching treatments)
  - Poor diagnostic criteria (patient reported outcomes)

Example

Suppose we are to show non-inferiority of Intervention T to an active control C. The data are assumed to come from $X \sim Normal(\mu_t, \sigma^2)$ and $X \sim Normal(\mu_c, \sigma^2)$, where $\sigma = 32$. The non-inferiority margin is set at $\delta_I = 10$, that is,

$$H_0 : \mu_t - \mu_c = -10$$
$$H_1 : \mu_t - \mu_c > -10$$

The clinical trial is designed with the one-sided type I error rate of $\alpha = 0.03$ and power of 0.90 at $\mu_t - \mu_c = 0$. The test statistic is

$$Z = \frac{\overline{X}_t - \overline{X}_c + \delta_I}{\sqrt{2}\sigma/\sqrt{n}}$$

And the required sample size is

$$\begin{aligned} n &= \frac{2\sigma^2(z_\alpha + z_\beta)^2}{\delta_I^2} \\ &= \frac{2(32^2)(1.96 + 1.28)^2}{10^2} = 216. \end{aligned}$$

We now check the power using a simple simulation.

```
simulation.NI <- function(muc, mut, sig, n, del0 = 10, B) {
    ## del0 > 0.  muc and mut are true values.
    p.value <- numeric(B)
    for (i in 1:B) {
        xt <- rnorm(n, mut, sig)
        xc <- rnorm(n, muc, sig)
        z <- (mean(xt) - mean(xc) + del0)/(sqrt(2) * sig/sqrt(n))
        p.value[i] <- 1 - pnorm(z)
    }
    p.value
}
set.seed(4791)
outNULL <- simulation.NI(muc = 20, mut = 10, sig = 32, n = 216, del0 = 10,
    B = 1000)
table(outNULL < 0.025)
```

```
FALSE  TRUE
  976    24
```

```
set.seed(7914)
outALT <- simulation.NI(muc = 20, mut = 20, sig = 32, n = 216, del0 = 10,
    B = 1000)
table(outALT < 0.025)
```

```
FALSE  TRUE
  107   893
```

Now suppose that 10% of group `T` switched to `C`, and 12% of group `C` switched to `T`.

```
intention2treat <- function(muc, mut, sig, n, del0 = 10, Rtc, Rct, B) {
    ## Rtc is % of T group who switched to C.  Rct is % of C group who
    ## switched to T.
    p.value <- numeric(B)
    ntc <- ceiling(n * Rtc)
    nct <- ceiling(n * Rct)
    for (i in 1:B) {
        xt <- c(rnorm(n - ntc, mut, sig), rnorm(ntc, muc, sig))
```

```
        xc <- c(rnorm(n - nct, muc, sig), rnorm(nct, mut, sig))
        z <- (mean(xt) - mean(xc) + del0)/(sqrt(2) * sig/sqrt(n))
        p.value[i] <- 1 - pnorm(z)
    }
    p.value
}
set.seed(22117)
outNULL2 <- intention2treat(muc = 20, mut = 10, sig = 32, n = 216, del0 = 10,
    Rtc = 0.1, Rct = 0.12, B = 1000)
table(outNULL2 < 0.025)
```

```
FALSE  TRUE
  899   101
```

```
set.seed(71122)
outALT2 <- intention2treat(muc = 20, mut = 20, sig = 32, n = 216, del0 = 10,
    Rtc = 0.1, Rct = 0.12, B = 1000)
table(outALT2 < 0.025)
```

```
FALSE  TRUE
   99   901
```

In a non-inferiority trial, the intention-to-treat analysis is anti-conservative.

Chapter 12

Treatment effects monitoring

12.1 Introduction

A phase III clinical trial (comparative treatment efficacy phase) is a type of trial design that assesses the efficacy of a new treatment relative to an alternative, placebo, standard therapy, or no treatment.

DSMB Data and safety monitoring board

DMC Data monitoring committee

TEMC Treatment effects monitoring committee

- DMC should have no formal involvement with subjects or investigators.
- DMC should interact actively in data analysis, request additional analyses if necessary.
- DMC usually meets two to three times a year (or after set number of patients contribute the data).

Motivations for monitoring treatment effects

- Check protocol compliance (baseline variables).
  Baseline imbalances alone are not likely to be a cause for much concern, but it can undermine the credibility of a trial, some intervention might be proposed to correct them.
- Review accrual
  Accrual tends to be slow at the beginning of the trial. The dropout rate may be higher than expected and/or the event rate may be lower than planned. Remedial actions include to prolong the accrual length and to add study centers.
- Review resource availability
  Money! Human resources (loss of irreplaceable expertise), difficulty obtaining rare drugs.

- Review data quality
  DMC checks patient eligibility (minor deviations are common), minor deviations in baseline data acquisition, randomization, misdiagnosis. Deviations occurring more than 10% of all patients may be a sign of internal quality control problem.
  Treatment compliance/adherence.
- Report adverse events
  Frequent side effects of low intensity may trigger dose reduction.
  A rarely occurring fatal toxicity could be intolerable in studies where patients are basically healthy / have a long life expectancy.
- Monitor treatment efficacy
  After all the previously mentioned checks are cleared, TEMC assesses efficacy differences.
- **Check treatment efficacy**

Specific questions for TEMC

- Should the trial continue?
  There may be secondary outputs from the trial such as secondary questions, database.
- Should the protocol modified?
  e.g., terminating one of many arms; adjusting timing or frequency of diagnostic tests changing consent process, improving quality of data collection, ...
- Does the TEMC need other views of the data?
- Should the TEMC meet more/less frequently? If the timing is based on “information time” the meeting may not occur at the recommended intervals (calendar time).

Reasons for stopping a trial (Table 14.1 in Piantadosi)

- Treatments are found to be convincingly different.
- Treatments are found to be convincingly not different.
- Side effects are too severe.
- The data are of poor quality.
- Accrual is too slow.
- Definitive information about the treatment becomes available making the study unethical or unnecessary.
- The scientific questions are no longer important.
- Adherence to the treatment is unacceptably poor.
- Resources to perform the study are no longer available.
- The study integrity has been undermined.

Factors to consider before terminating a study

- Delays in reporting.
- Baseline differences.
- Bias in response assessment.
- Missing data.
- Credibility of results if stopped early.

Reasons for not stopping early:
Increasing precision and reducing errors
Subgroup analyses, interaction effects, secondary endpoints

12.1.1 Composition and organization of TEMC = DMC

TEMC is intellectually and financially independent of the study investigators so that it can provide objective assessments.

Who should TEMC make **recommendations** to? Trial sponsor or trial investigators or both?

"The TEMC has an obligation to inform the investigators of their opinions and recommendations about actions that carry ethics implications."

FDA's 1989 guideline has a very brief description of data monitoring and DMCs.

NIH policy (1998)

- All sponsored trials must have a monitoring system for safety, efficacy and validity.

ICH guidelines (1998)
"When a sponsor assumes the role of monitoring efficacy or safety comparisons and therefore has access to unblinded comparative information, particular care should be taken to protect the integrity of the trial..."

"Any interim analysis that is not planned appropriately (with or without the consequences of stopping the trial early) may flaw the results of a trial and possibly weaken confidence in the conclusions drawn. $\cdots$ If unplanned interim analysis is conducted, the clinical study report should explain why it was necessary, the degree to which blindness had to be broken, provide an assessment of the potential magnitude of bias introduced, and the impact on the interpretation of the results."

"The IDMC should have written operating procedures and main records of all its meetings, ..."

"The IDMC is a separate entity from an Institutional Review Board (IRB) or an Independent Ethics Committee (IEC), and its composition should include clinical trial scientists knowledgeable in the appropriate disciplines including statistics."

DMC membership
Data monitoring is a complex decision process and requires a variety of expertise in medicine, basic science, biostatistics, epidemiology, and medical ethics. (Additionally representative from a regulatory body)

DMC confidentiality
In general, interim data must remain confidential. DMC rarely releases interim data, and its

members must not share interim data with anyone outside of DMC. Data leaks may affect

- Patient recruitment.
- Protocol compliance.
- Outcome assessment.
- Market value.

Then why not have DMC only use blinded data?
Complete objectivity $\neq$ ethical

Revisit the question, "Should DMC include the study investigators?"

FDA draft guidance
"Knowledge of unblinded interim comparisons from a clinical trial is not necessary for those conducting or those sponsoring the trial; further, such knowledge can bias the outcome of the study by inappropriately influencing its continuing conduct or the plan of analyses. Therefore, interim data and the results of interim analyses should generally not be accessible by anyone other than DMC members."

Guessing the between-group difference only using blinded data.
Suppose in order to check data quality, the pooled variance was computed and reported in the DMC. For example, "We originally anticipated $\sigma^2 = 20$; however, the pooled variance after $n_1 = 50$ from each group was $s^2_{1p} = 25$. If the clinical trial scientist or the sponsor has an access to this information how bad is it?

By itself it is not too bad; if data are from normal populations, variance estimate and mean estimate are independent.

However, without breaking the blind, the overall mean $\bar{X}_{1\cdot}$ and variance s^2_{1o} can be computed.

$$\hat{\delta}^2 = \left(4 - \frac{2}{n_1}\right) s^2_{1o} - \left(4 - \frac{4}{n_1}\right) s^2_{1p}.$$

DMC meeting format

- Open session
  - Monitor progress with blinded data
  - Sponsor, Executive committee, DMC, SAC.
- Closed session
  - Unblinded data.
  - DMC and SAC.
  - Sponsor?
- Executive session
  - DMC only.
- Debriefing session
  - DMC chair, Sponsor representative, Executive committee representative.

Chapter 13

Group Sequential Method

13.1 Introduction

Fully sequential method A test of significance is repeated after each observation.

Group sequential method A test of significance is repeated after a group of observations.

Some basic characteristics of a group sequential method

- The response variable needs to be observed immediately.
- Number of stages (or looks) can be 2 to 20.
- Looks are equally spaced. (This is not a critical requirement.)
- At each interim (and final) analysis, compute summary statistic based on the cumulative data.
- A group sequential method is a strategy to stop early as opposed to an “adaptive design”, which is often viewed as a strategy to extend the study if necessary.
- A set of critical values are computed so that the overall α is as specified.
  - Haybittle-Peto (1971)
    This is an *ad hoc* method in which a very conservative critical value (e.g., $Z > 3$) is used at every interim test. At the final analysis, no adjustment is used (i.e., $Z > -1.96$)
    It is highly unlikely to stop early.
  - Pocock (1977)
    A ”repeated test of significance” at a *constant* significance level to analyze accumulating data.
  - O’Brien-Fleming (1979)
    The significance levels increase as the study progress.

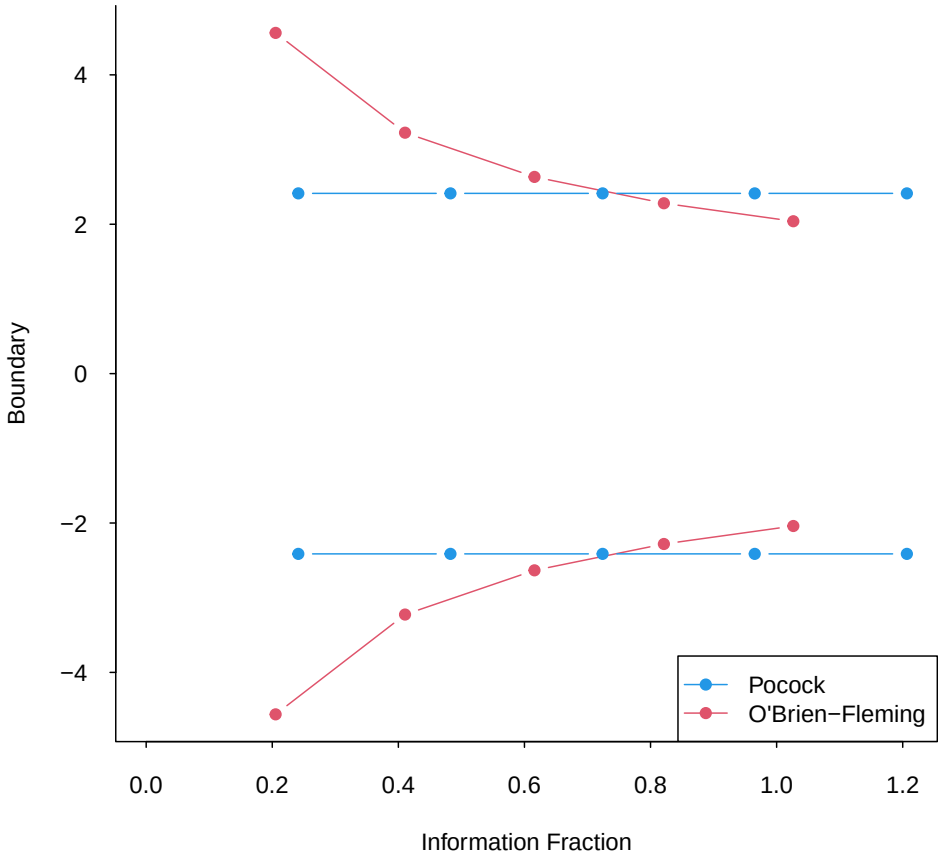
13.2 Example

For testing

$$\mu_t - \mu_c = 0$$
$$\mu_t - \mu_c > 0$$

With $\alpha = 0.025$ and power$= 0.90$ at $\delta_1 = 0.25$ ($\sigma^2 = 1$), 5-stage group-sequential designs are:

```
library(gsDesign)

x.of <- gsDesign(k = 5, test.type = 2, alpha = 0.025, beta = 0.1, delta0 = 0,
    delta1 = 0.25, n.fix = 1, sfu = "OF")
x.po <- gsDesign(k = 5, test.type = 2, alpha = 0.025, beta = 0.1, delta0 = 0,
    delta1 = 0.25, n.fix = 1, sfu = "Pocock")
```



Sample size is expressed in terms of ratios to the sample size of the conventional single-stage design.

```
## Pocock
x.po


Symmetric two-sided group sequential design with
90 % power and 2.5 % Type I Error.
Spending computations assume trial stops
if a bound is crossed.

           Sample
            Size
  Analysis Ratio*  Z   Nominal p  Spend
```

```
         1  0.241 2.41    0.0079 0.0079
         2  0.483 2.41    0.0079 0.0059
         3  0.724 2.41    0.0079 0.0045
         4  0.965 2.41    0.0079 0.0037
         5  1.207 2.41    0.0079 0.0031
     Total                       0.0250

++ alpha spending:
 Pocock boundary.
* Sample size ratio compared to fixed design with no interim

Boundary crossing probabilities and expected sample size
assume any cross stops the trial

Upper boundary (power or Type I Error)
          Analysis
  Theta      1      2      3      4      5 Total   E{N}
  0.000 0.0079 0.0059 0.0045 0.0037 0.0031 0.025 1.1767
  3.241 0.2059 0.2603 0.2086 0.1402 0.0851 0.900 0.6849

Lower boundary (futility or Type II Error)
          Analysis
  Theta      1      2      3      4      5 Total
  0.000 0.0079 0.0059 0.0045 0.0037 0.0031 0.025
  3.241 0.0000 0.0000 0.0000 0.0000 0.0000 0.000


## O'Brien-Fleming
x.of


Symmetric two-sided group sequential design with
90 % power and 2.5 % Type I Error.
Spending computations assume trial stops
if a bound is crossed.

           Sample
            Size
  Analysis Ratio*  Z   Nominal p  Spend
         1  0.205 4.56    0.0000 0.0000
         2  0.411 3.23    0.0006 0.0006
         3  0.616 2.63    0.0042 0.0038
         4  0.821 2.28    0.0113 0.0083
         5  1.026 2.04    0.0207 0.0122
```

```
     Total                        0.0250

++ alpha spending:
 O'Brien-Fleming boundary.
* Sample size ratio compared to fixed design with no interim

Boundary crossing probabilities and expected sample size
assume any cross stops the trial

Upper boundary (power or Type I Error)
          Analysis
  Theta     1      2      3      4      5 Total   E{N}
  0.000 0.000 0.0006 0.0038 0.0083 0.0122 0.025 1.0191
  3.241 0.001 0.1244 0.3421 0.2840 0.1484 0.900 0.7503

Lower boundary (futility or Type II Error)
          Analysis
  Theta 1     2      3      4      5 Total
  0.000 0 6e-04 0.0038 0.0083 0.0122 0.025
  3.241 0 0e+00 0.0000 0.0000 0.0000 0.000
```

- In the tables of the critical values, `Nominal p` is simply $P[Z > z]$, where $Z \sim Normal(0, 1)$.
- `Spend` is the type I error probability that has been spent by the end of each stage, and it is based on *conditional* probability.
  For example, for the second stage of the Pocock design, it is 0.01. It can be computed as follows:

$$P[Z_2 > 2.41 | -2.41 \leq Z_1 \leq 2.41].$$

13.3 General applications

Let $k = 1, \cdots, K$ be denote the stages so that we have

$$\begin{aligned} \bar{X}_t^{(k)} - \bar{X}_c^{(k)} &= \frac{1}{n_{tk}} \sum_{i=1}^{n_{tk}} X_{ti} - \frac{1}{n_{ck}} \sum_{i=1}^{n_{ck}} X_{ci} \\ &\sim Normal\left(\mu_t - \mu_c, \frac{\sigma^2}{n_{tk}} + \frac{\sigma^2}{n_{ck}}\right), \end{aligned}$$

where n_{tk} and n_{ck} are the *cumulative* sample sizes for the treatment and control groups. Note that this is not a conditional distribution but a marginal distribution.

Define "information" as $I_k = (\sigma^2/n_{tk} + \sigma^2/n_{ck})^{-1}$. Roughly speaking, information is square of what appears in the denominator of the test statistic, Z. When $n_k = n_{tk} = n_{ck}$, $I_k = (2\sigma^2/n_k)^{-1}$.

The test statistic for stage k is

$$Z_k = \frac{\bar{X}_t^{(k)} - \bar{X}_c^{(k)}}{\sqrt{2\sigma^2/n_k}} = (\bar{X}_t^{(k)} - \bar{X}_c^{(k)})\sqrt{I_k}.$$

The vector, $(Z_1, \cdots, Z_k)$, has a multivariate normal distribution because each Z_k is a linear combination of the independent normal variates X_{ti} and X_{ci}. The marginal distribution of Z_k is

$$Z_k \sim Normal\left((\mu_t - \mu_c)\sqrt{I_k}, 1\right).$$

How about the covariance of Z_{k_1} and Z_{k_2} for $k_1 < k_2$?

$$\begin{aligned}
Cov(Z_{k_1}, Z_{k_2}) &= Cov\left(\{\bar{X}_t^{(k_1)} - \bar{X}_c^{(k_1)}\}\sqrt{I_{k_1}},\ \{\bar{X}_t^{(k_2)} - \bar{X}_c^{(k_2)}\}\sqrt{I_{k_2}}\right) \\
&= Cov\left(\{\bar{X}_t^{(k_1)} - \bar{X}_c^{(k_1)}\},\ \{\bar{X}_t^{(k_2)} - \bar{X}_c^{(k_2)}\}\right)\sqrt{I_{k_1}}\sqrt{I_{k_2}} \\
&= \left[Cov\left(\bar{X}_t^{(k_1)}, \bar{X}_t^{(k_2)}\right) + Cov\left(\bar{X}_c^{(k_1)}, \bar{X}_c^{(k_2)}\right)\right]\sqrt{I_{k_1}}\sqrt{I_{k_2}}
\end{aligned}$$

$$\begin{aligned}
Cov\left(\bar{X}_t^{(k_1)}, \bar{X}_t^{(k_2)}\right) &= Cov\left(\frac{1}{n_{k_1}}\sum_{i=1}^{n_{k_1}} X_i,\ \frac{1}{n_{k_2}}\sum_{i=1}^{n_{k_1}} X_i + \frac{1}{n_{k_2}}\sum X_i\right) \\
&= \frac{1}{n_{k_1}}\frac{1}{n_{k_2}} Var\left(\sum_{i=1}^{n_{k_1}} X_i\right) = \frac{1}{n_{k_2}}\sigma^2 \\
Cov(Z_{k_1}, Z_{k_2}) &= \sigma^2\left(\frac{1}{n_{k_2}} + \frac{1}{n_{k_2}}\right)\sqrt{I_{k_1}}\sqrt{I_{k_2}} \\
&= \sqrt{I_{k_1}/I_{k_2}}.
\end{aligned}$$

Therefore,

- $(Z_1, \cdots, Z_K)$ is multivariate normal.
- $E[Z_k] = (\mu_t - \mu_c)\sqrt{I_k}, \qquad k = 1, \cdots, K$, and
- $Cov(Z_{k_1}, Z_{k_2}) = \sqrt{I_{k_1}/I_{k_2}}, \qquad 1 \le k_1 \le k_2 \le K.$

General decision rule for a group sequential design is

After group $k = 1, \cdots, K-1$
if $|Z_k| \geq c_k$ stop and reject H_0.
otherwise continue to group $k+1$.

After group K
if $|Z_k| \geq c_K$ stop and reject H_0.
otherwise stop for futility.

The test's type I error rate can be expressed as

$$P\left\{|Z_k| \geq c_k \text{ for some } k = 1, \cdots, K\right\}.$$

The critical values, c_k, are chosen so that the above probability is equal to α. And the power of the study at δ_1 is

$$P\left\{\bigcup_{k=1}^{K}\left(|Z_j| < c_j, \text{for } j = 1, \cdots, k-1 \text{ and } |Z_k| \geq c_k\right)\right\}.$$

Evaluation of this probability requires knowing the distribution of $(Z_1, \cdots, Z_K)$. Refer to tables of c_K values or a computer software.

- For a Pocock method, the critical values are constant, so $c_k = C_P(K, \alpha)$. That is, specifying α and K uniquely determines the critical values.
- For the previous example, $C_P(5,\ 0.03) = 2.41$.
- For an O-Fleming method, the critical values have the form, $c_k = C_B(K, \alpha)\sqrt{K/k}$
- For the same example, $C_B(5,\ 0.03) = 2.04$. And the other critical values are: $2.04\sqrt{5/4}$, $2.04\sqrt{5/3}$, and so on.

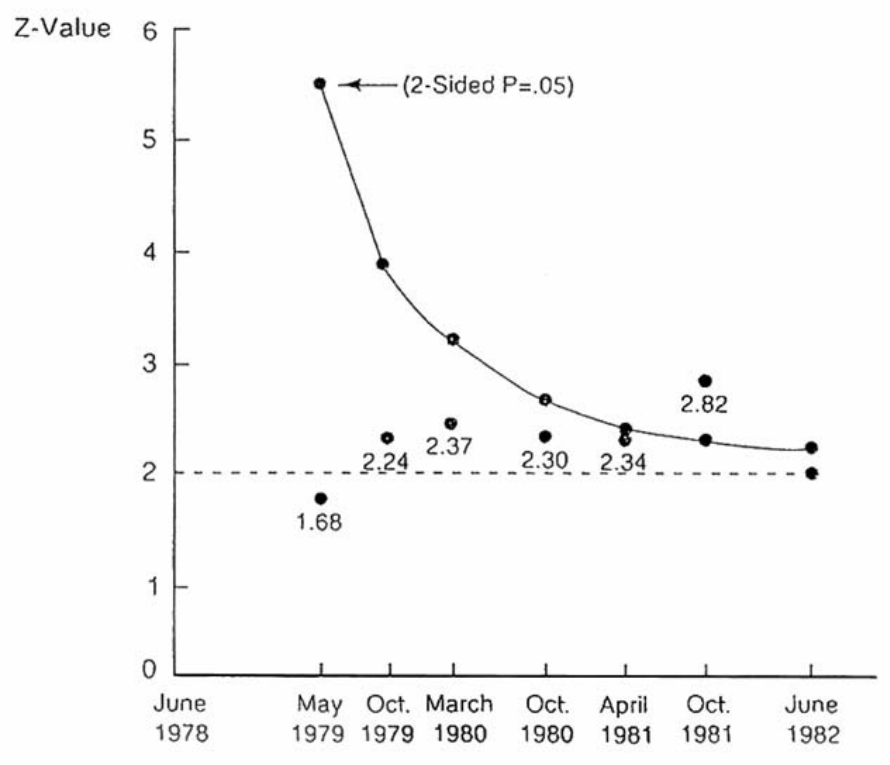
```
(K.of)

[1] 2.04

K.of * sqrt(5/(5:1))

[1] 2.040 2.281 2.634 3.226 4.562
```

- More generally, if stage sample sizes are different, use I_k, that is, $c_k = C_B(K, \alpha)\sqrt{I_K/I_k}$.

13.3.1 Beta blocker heart attack trial

Seven analyses (including the final one) were planned (corresponding to the timing of the Data Monitoring Committee meetings) using O'Brien-Fleming bounds with two-sided type I error rate of

5%. The primary outcome was survival, and log-rank test was used.



If Pocock boundary had been used, $N = 7$ and $\alpha = 0.05$ give $Z = 2.485$. Therefore, the trial would have been stopped at the same point.

13.3.2 non-Hodgkin's lymphoma

Pocock 1983 *Clinical Trials: A Practical Approach*. A trial was conducted in patients with non-Hodgkin's lymphoma for two drug combinations (cytoxanprednisone -CP- and cytoxan-vincristine-prednisone -CVP-). The primary endpoint was tumor shrinkage (Yes/No).

Statistical analyses were planned after approximately 25 patients. With 5 looks and one-sided $\alpha = 0.05$. The Pocock procedure requires a significance level of 0.02 at each analysis. χ^2 tests without the continuity correction were performed at each of the 5 scheduled analyses.

```
gsDesign(k = 5, test.type = 1, alpha = 0.05, n.fix = 1, sfu = "Pocock")


One-sided group sequential design with
90 % power and 5 % Type I Error.
          Sample
```

```
            Size
  Analysis Ratio*  Z   Nominal p  Spend
         1  0.246 2.12    0.0169 0.0169
         2  0.491 2.12    0.0169 0.0117
         3  0.737 2.12    0.0169 0.0087
         4  0.982 2.12    0.0169 0.0069
         5  1.228 2.12    0.0169 0.0057
     Total                       0.0500

++ alpha spending:
 Pocock boundary.
* Sample size ratio compared to fixed design with no interim

Boundary crossing probabilities and expected sample size
assume any cross stops the trial

Upper boundary (power or Type I Error)
           Analysis
  Theta      1      2      3      4      5 Total   E{N}
  0.000 0.0169 0.0117 0.0087 0.0069 0.0057  0.05 1.1968
  2.926 0.2510 0.2574 0.1900 0.1249 0.0767  0.90 0.6678
```

	Tumor shrinkage		
	CP	CVP	p-value
Analysis 1	$3/14$	$5/11$	0.40
Analysis 2	$11/27$	$13/24$	0.50
Analysis 3	$18/40$	$17/36$	1
Analysis 4	$21/54$	$24/48$	0.58
Analysis 5	$26/67$	$31/59$	0.30

The CVP appeared better than the CP, but difference was not statistically significant. Further analyses of secondary endpoints convinced the researchers that the CVP was better than the CP.

13.4 Alpha-spending

"Classical" group sequential designs have equal information (sample size) at every stage, but we may want to be a little more flexible. And when I_k is not a constant we might want to change α spent accordingly.

Decompose the rejection region.

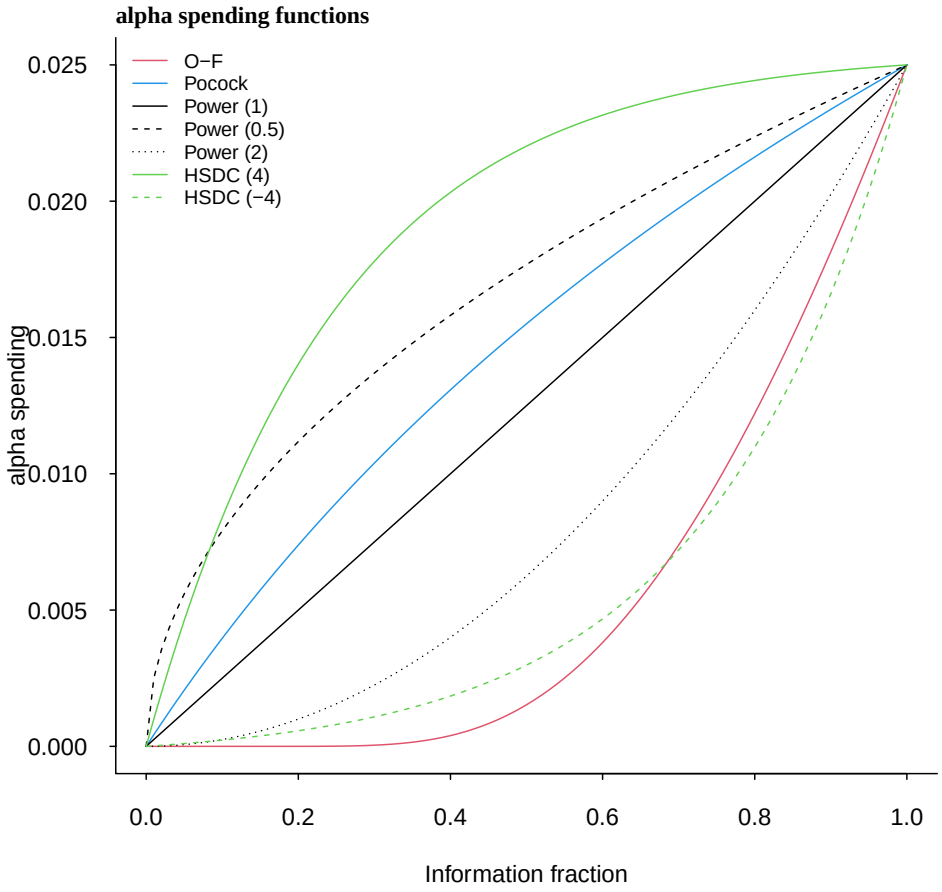
$$
\begin{aligned}
R &= P\{|Z_k| \geq c_k \text{ for some } k = 1, \cdots, K\} \\
&= P\{(|Z_1| \geq c_1) \text{ or } (|Z_1| < c_1 \text{ and } |Z_2| \geq c_2) \text{ or } \cdots\} \\
&= P\{|Z_1| \geq c_1\} + P\{|Z_1| < c_1 \text{ and } |Z_2| \geq c_2\} + P\{|Z_1| < c_1 \text{ and } |Z_2| < c_2 \text{ and } |Z_3| \geq c_3\} + \cdots \\
&= \alpha(I_1) + (\alpha(I_2) - \alpha(I_1)) + (\alpha(I_3) - \alpha(I_2) - \alpha(I_1)) + \cdots
\end{aligned}
$$

The biggest advantage of alpha-spending approach is its flexibility; neither the number nor timing of the interim analyses need to be specified in advance. The monitoring plan can be changed during the trial and still type I error rate is preserved. The power depends relatively little on the number and timing of the interim looks[1].

Alpha-spending functions

O'Brien-Fleming	$\alpha(t) = 2\left[1 - \Phi\left(z_{\alpha/2}/\sqrt{t}\right)\right]$
Pocock	$\alpha(t) = \alpha \log(1 + (e-1)t)$
Kim-DeMets (Power)	$\alpha(t, \theta) = \alpha t^{\theta}$ (for $\theta > 0$)
Hwang-Shih-DeCani	$\alpha(t, \phi) = \alpha \frac{1 - e^{-\phi t}}{1 - e^{-\phi}}$ (for $\phi \neq 0$)

[1] "Fundamentals of clinical trials (4th ed)" by Friedman LM, Furberg CD, DeMets DL



13.5 One-sided test

If "stop for futility" is not an option, the same boundary can be used. If a futility stop is an option, then

After group $k = 1, \cdots, K-1$
- if $Z_k \geq b_k$ stop and reject H_0.
- if $Z_k \leq a_k$ stop for futility (accept H_0).

After group K
- if $Z_k \geq b_K$ stop and reject H_0.
- if $Z_k < a_K$ stop for futility.

Note that $a_K = b_K$ ensures that the test terminates at analysis K.

13.6 Repeated confidence intervals

If we compute unadjusted confidence intervals $\bar{X}_{\text{so far}} \pm 1.96\sigma/\sqrt{n_{\text{so far}}}$ at the end of each stage, we get low coverage probabilities. Armitage, McPherson, Rowe ("Repeated significance tests on accumulating data". *JRSS-A* 1969) computed the actual coverage probabilities (Table 2).

Number of looks	Overall probability that all intervals contain θ
1	0.95
2	0.92
3	0.89
4	0.87
5	0.86
10	0.81
20	0.75
50	0.68
∞	0

The idea of repeated confidence intervals (RCIs) is to use an adjusted value, $c_k(\alpha, K)$, instead of 1.96 so that the overall coverage probability is $1 - \alpha/2$. The value of $c_k(\alpha, K)$ is the critical value (border) for each stage and depends on α and K if Pocock boundary is used, and additionally k if O'Brien-Fleming boundary is used.

Example: Suppose we use a 6-stage group sequential design of O'Brien-Fleming type with a two-sided $\alpha = 5\%$. The critical values are:

```
gsDesign(k = 6, test.type = 2, alpha = 0.025, sfu = "OF")


Symmetric two-sided group sequential design with
90 % power and 2.5 % Type I Error.
Spending computations assume trial stops
if a bound is crossed.

           Sample
            Size
  Analysis Ratio*  Z   Nominal p  Spend
         1  0.172 5.03    0.0000 0.0000
         2  0.343 3.56    0.0002 0.0002
         3  0.515 2.90    0.0018 0.0017
         4  0.686 2.51    0.0060 0.0047
         5  0.858 2.25    0.0123 0.0079
         6  1.030 2.05    0.0200 0.0105
     Total                       0.0250
```

```
++ alpha spending:
 O'Brien-Fleming boundary.
* Sample size ratio compared to fixed design with no interim

Boundary crossing probabilities and expected sample size
assume any cross stops the trial

Upper boundary (power or Type I Error)
          Analysis
  Theta     1      2      3      4      5      6 Total   E{N}
  0.000 0e+00 0.0002 0.0017 0.0047 0.0079 0.0105 0.025 1.0218
  3.241 1e-04 0.0487 0.2350 0.2915 0.2088 0.1159 0.900 0.7393

Lower boundary (futility or Type II Error)
          Analysis
  Theta 1     2      3      4      5      6 Total
  0.000 0 2e-04 0.0017 0.0047 0.0079 0.0105 0.025
  3.241 0 0e+00 0.0000 0.0000 0.0000 0.0000 0.000
```

The critical values are

```
[1] 5.028 3.556 2.903 2.514 2.249 2.053
```

First, let's confirm that the critical values have the form $c_k = C_{OB}(K, \alpha)\sqrt{I_K/I_k}$. The final critical value $C_{OB}(6, \alpha) = 2.05$, and assuming the looks are equi-distant (same group sample size), we have:

$$c_1 = 2.05\sqrt{6/1} = 5.03 \qquad c_2 = 2.05\sqrt{6/2} = 3.56$$
$$c_3 = 2.05\sqrt{6/3} = 2.90 \qquad c_4 = 2.05\sqrt{6/4} = 2.51$$
$$c_5 = 2.05\sqrt{6/5} = 2.25 \qquad c_6 = 2.05\sqrt{6/6} = 2.05$$

Then after stage 1, we would use 5.03 in place of the regular 1.96 when computing a 95% confidence interval. In general after stage k ($k = 1, \cdots, 6$),

$$(\bar{x}_{kt} - \bar{x}_{kc}) \pm c_k \frac{\sqrt{2\sigma^2}}{\sqrt{mk}},$$

where m is per-group sample size for each stage.

This method (RCI) is consistent with the corresponding hypothesis testing. Only when is H_0 rejected in stage k, the confidence interval for that stage will exclude the null value. Thus, we can use the idea of "inverting hypothesis test" to get the same confidence interval. (more later)

13.7 P-values

Recall how we construct a proper p-value for a Simon's two-stage design in phase II methodology. We needed to define "more or as extreme as the observed data". To be able to do this, we need to have an ordering of all the sample paths. In a simple single-stage design, the ordering is usually based on z-values (or absolute value of z-values if two-sided test), i.e., the bigger the observed z, the stronger the evidence against H_0. Then a one-sided p-value is computed by

$$p = P_0[Z \geq z].$$

With a group sequential design, or more generally, with a multi-stage design with pre-specified group-wise sample sizes, the following orderings have been proposed. Notation: $(k', z') \succ (k, z)$ to denote (k', z') is above (k, z).

- Stage-wise ordering.
  $(k', z') \succ (k, z)$ if any of the following is true:
  1. $k' = k$ and $z' \geq z$.
  2. $k' < k$ and $z' \geq b_{k'}$ (upper critical value).
  3. $k' > k$ and $z' \leq a_k$ (lower critical value).
- MLE ordering.
  $(k', z') \succ (k, z)$ if $z'/\sqrt{I_{k'}} > z/\sqrt{I_k}$. Originally proposed in connection with a test for a binomial proportion The bigger value of the MLE gets a higher order. Sometimes called "sample mean ordering" because this is equivalent to ordering based on the sample mean (one-sample) or the difference of sample means (two-samples).
- Likelihood ratio ordering.
  $(k', z') \succ (k, z)$ if $z' > z$. (Stages do not matter.)
- Score test ordering.
  $(k', z') \succ (k, z)$ if $z\sqrt{I_{k'}} > z\sqrt{I_k}$.

Whichever ordering is used, we can compute a one-sided p-value is

$$P_0[(T, Z_T) \succ (k^*, z^*)]$$

For example, if we use a stage-wise ordering and test terminates in the $K-1$ stage with $Z_{K-1} > b_{K-1}$ (reject H_0).

$$p = \int_{b_1}^{\infty} g_1(z; 0)dz + \cdots + \int_{z^*}^{\infty} g_{K-1}(z; 0)dz.$$

In the above expression, $g_k(z; \theta)$ is a density function of z in stage k. Conceptually, the density function of z in k stage depends on all the data in the previous stages, $1 \cdots k-1$, requiring multivariate integration.

Armitage, McPherson, Rowe (1969) derived a recursive formula so that the computation is much simplified, requiring only a succession of univariate integrations. For $k = 2, \cdots, K$,

$$g_k(z;\theta) = \int_{C_{k_1}} g_{k-1}(\mu;\theta) \frac{\sqrt{I_k}}{\sqrt{\Delta_K}} \phi\left(\frac{z\sqrt{I_k} - \mu\sqrt{I_{k-1}} - \Delta_k\theta}{\sqrt{\Delta_k}} \right) d\mu,$$

where C_{k_1} is the continuation region of the stage k_1, and Δ_k is the increment information, $I_k - I_{k-1}$.

If stage-wise ordering is used, it automatically ensures that item the p-value is less than the significance level α if and only if H_0 is rejected.

Once we define the ordering to use with the group sequential test then we can compute a p-value for testing $H_0 : \theta = 0$ by "inverting hypothesis test". A $(1 - \alpha/2)$ confidence interval is a collection of θ_0' such that $H_0' : \theta = \theta_0'$ would be accepted with the observed sample path. (More details with general adaptive designs.)

Chapter 14

Two-stage adaptive designs

14.1 Introduction

Much of discussion in the literature for flexible designs in phase III clinical trial methodologies revolves around 2 stage designs. Practically speaking, implementing flexible clinical trials beyond two stages is difficult, and perhaps these multi-stage flexible designs add only little to the designs with just two stages. Moreover, phase III clinical trials are for confirmatory purposes, and adaptively changing the design more than once in the middle of a confirmatory trial is not seen favorably by the regulatory figure.

So we will only consider two-stage adaptive designs. Two-stage group sequential designs are examples of such designs.

14.2 Background

We will look at unmasked (unblinded) two-stage designs in which all the information from stage 1 is available. Design of the second stage (sample size and critical value) may be specified as functions of stage 1 data. If both sample size and critical value are constants in stage 1 data, then it reduces to a two-stage group sequential design.

Adaptive designs can be categorized into the following two types:

- **Prespecified designs**
  Design of the second stage (e.g., sample size and critical value) is specified before the first stage. There is nothing to decide at the end of stage 1. The design of the second stage is defined flexibly as functions of stage 1 data. Group sequential designs fall into this category.

- **Unspecified designs**
  Design of the second stage is not specified in advance and determined after stage 1 data are observed.

Characteristics of these types of designs:

Prespecified designs

- Type I error can be controlled.
- Type II error can be controlled.
- You can compute design characteristics of the design (e.g., Expected and maximum sample sizes) prior to initiation of the study.

Unspecified designs

- Type I error can be controlled.
- These designs give much flexibility to handle unexpected situations (e.g., Variance is much bigger than anticipated).

Something in between: Not specifying the stage 2 sample size is unrealistic because it makes it impossible to budget such a clinical trial. Instead of leaving the stage 2 unspecified, maybe we should specify the maximum sample size. And perhaps, we may want to specify the minimum *conditional* power for the second stage, P[Reject H_0 in stage 2 | Stage 1 data].

With these specifications, unspecified designs start to look like prespecified ones.

What do they say about adaptive designs

PhRMA (2006) "... a clinical study design that uses accumulating data to decide how to modify aspects of the study as it continues, without undermining the validity and integrity of the trial."
"... Changes are made by design, and not on an ad hoc basis; therefore, adaptation is a design feature aimed to enhance the trial, not a remedy for inadequate planning."

EMA (2006) "A study design is called 'adaptive' if statistical methodology allows the modification of a design element (e.g. sample-size, randomisation ratio, number of treatment arms) at an interim analysis with full control of type I error rate."
"... adaptive designs should not be seen as a means to alleviate the burden of rigorous planning of clinical trials."

FDA (2010) "... Adaptive design clinical study is defined as a study that includes a prospectively planned opportunity for modification of one or more aspects of the study design and hypotheses based on analysis of data (usually interim data) from subjects in the study."

14.3 Set up

What are good two-stage adaptive designs?

- What do we use to compare different designs?
  - Power between μ_0 and μ_1
  - Expected sample size at different μ's.
- Design of stage 1 tends to be more influential in terms of these characteristics.
- "Optimality" is not the only driving force to choose a design. A design with a very small n_1 may have different objectives than those with a large n_1. (Ambitious designs vs. Insurance-type designs)

To test $H_0 : \delta = 0$, where $\delta = \mu_t - \mu_c$, we take random samples from

$$X_t \sim Normal(\mu_t, \sigma_t^2) \qquad\qquad X_c \sim Normal(\mu_c, \sigma_c^2).$$

Assume the true variances are equal and known: $\sigma_t^2 = \sigma_c^2 = \sigma^2$. Also assume the sample sizes are equal in the control and treatment groups: $n_{1t} = n_{1c} = n$. Then

$$\bar{X}_{1t} \sim Normal(\mu_t, \sigma^2/n_1) \qquad\qquad \bar{X}_{1c} \sim Normal(\mu_c, \sigma^2/n_1)$$

and

$$Z_1 = \frac{\sqrt{n_1}(\bar{X}_{1t} - \bar{X}_{1c})}{\sqrt{2}\sigma}.$$

The distribution of Z_1 is $Z_1 \sim Normal(\sqrt{n_1}\xi, 1)$, where $\xi = \delta/\sqrt{2}\sigma$.

Also define $\zeta = \sqrt{n_1}\xi$ for the stage 1.

In stage 1, we observe Z_1 and use the following decision rule:

- If $Z_1 < k_1$, stop for futility.
- If $Z_1 > k_2$, stop and reject H_0.
- If $k_1 < Z_1 < k_2$ then continue to stage 2.

In stage 2, we take a sample of size $n_2(z_1)$ from each arm, and define

$$Z_2 = \frac{\sqrt{n_2(z_1)}(\bar{X}_{2t} - \bar{X}_{2c})}{\sqrt{2}\sigma}.$$

Conditioned on $Z_1 = z_1 \in (k_1, k_2)$,

$$Z_2 \sim Normal(\sqrt{n_2(z_1)}\xi, 1).$$

The decision rule at the end of stage 2 is:

- If $Z_2 \leq c(z_1)$, stop and conclude futility.
- If $Z_2 > c(z_1)$, stop and conclude efficacy.

We can use Z_1 and Z_2 to construct a two-stage design, and we can also construct a test statistic that combines the test statistics from both stages.

Let

$$Z_w = \frac{Z_1 + Z_2}{\sqrt{2}}.$$

If Z_1 and Z_2 are independent

$$Z_w \sim Normal\left(\xi\frac{\sqrt{n_1} + \sqrt{n_2(z_1)}}{\sqrt{2}}, 1\right)$$

Under H_0, Z_w has the standard normal distribution. Z_w is rarely used because it weights stage 1 and stage 2 data differently. To give equal weight to every datum, we should construct a test statistic that uses

$$\begin{aligned}
Y &= \frac{\left(\sum_{i=1}^{n_1} X_{1ti} + \sum_{i=1}^{n_2(z_1)} X_{2ti}\right) - \left(\sum_{i=1}^{n_1} X_{1ci} + \sum_{i=1}^{n_2(z_1)} X_{2ci}\right)}{n_1 + n_2(z_1)} \\
&= \frac{\left(n_1\bar{X}_{1t} + n_2(z_1)\bar{X}_{2t}\right) - \left(n_1\bar{X}_{1c} + n_2(z_1)\bar{X}_{2c}\right)}{n_1 + n_2(z_1)} \\
&= \frac{n_1(\bar{X}_{1t} - \bar{X}_{1c}) + n_2(z_1)(\bar{X}_{2t} - \bar{X}_{2c})}{n_1 + n_2(z_1)} \\
&= \frac{\sqrt{2}\sigma(\sqrt{n_1}Z_1 + \sqrt{n_2(z_1)}Z_2)}{n_1 + n_2(z_1)} \\
&\sim Normal\left(\sqrt{2}\sigma\xi,\ \frac{2\sigma^2}{n_1 + n_2(z_1)}\right).
\end{aligned}$$

Thus if we let

$$Z_u = \frac{\sqrt{n_1 + n_2(z_1)}\,Y}{\sqrt{2}\sigma},$$

then we have

$$Z_u \sim Normal\left(\sqrt{n_1 + n_2(z_1)}\xi,\ 1\right),$$

if Z_1 and Z_2 are independent.

It is useful to write Z_u as a weighted average of Z_1 and Z_2 as follows:

$$Z_u = \frac{\sqrt{n_1}}{\sqrt{n_1 + n_2(z_1)}}Z_1 + \frac{\sqrt{n_2(z_1)}}{\sqrt{n_1 + n_2(z_1)}}Z_2$$

Because even the existence of the second stage depends on Z_1, we need to think about stage 2 conditioned on stage 1 data.

Conditioned on $Z_1 = z_1$,

$$\begin{aligned} Z_u|(Z_1 = z_1) &= \frac{\sqrt{n_1}}{\sqrt{n_1 + n_2(z_1)}} z_1 + \frac{\sqrt{n_2(z_1)}}{\sqrt{n_1 + n_2(z_1)}} Z_2 \\ &\sim Normal\left(\frac{\sqrt{n_1}}{\sqrt{n_1 + n_2(z_1)}} z_1 + \frac{n_2(z_1)}{\sqrt{n_1 + n_2(z_1)}} \xi,\ \frac{n_2(z_1)}{n_1 + n_2(z_1)} \right). \end{aligned}$$

The original decision rule at the end of stage 2 was written in terms of Z_2, i.e., $Z_2 > c(z_1)$ then reject H_0. This can be written in terms of Z_u. Suppose the critical value that goes with Z_u is $c_u(z_1)$. Conditioned on $Z_1 = z_1$ we have

$$c_u(z_1) = \frac{\sqrt{n_1}}{\sqrt{n_1 + n_2(z_1)}} z_1 + \frac{\sqrt{n_2(z_1)}}{\sqrt{n_1 + n_2(z_1)}} c(z_1).$$

The decision rule at the end of stage 2 can be written in terms with Z_2 and Z_u.

There exist many different normalization schemes. For example, Z_u can be rescaled to have a variance of 1.

$$\frac{\sqrt{n_1 + n_2(z_1)}}{\sqrt{n_2(z_1)}} Z_u \sim Normal\left(\frac{\sqrt{n_1}}{\sqrt{n_2(z_1)}} z_1 + \sqrt{n_2(z_1)}\xi,\ 1 \right)$$

14.4 Conditional power functions

Conditional power is the probability of rejecting H_0 in stage 2 conditioned on the first stage data. Let $A(z_1, \xi)$ to denote the conditional power at ξ given $Z_1 = z_1$. Then we have

$$A(z_1, \xi) = P_\xi[Z_2 > c(z_1)|Z_1 = z_1].$$

The conditional distribution of Z_2 given $Z_1 = z_1$ is

$$Z_2|(Z_1 = z_1) \sim Normal\left(\sqrt{n_2(z_1)}\xi,\ 1 \right),$$

and

$$A(z_1, \xi) = 1 - \Phi\left[c(z_1) - \sqrt{n_2(z_1)}\xi \right].$$

The conditional type I error rate is $A(z_1, \xi_0)$. Specifically, when $\xi_0 = 0$, we have

$$A(z_1, 0) = 1 - \Phi\left[c(z_1) \right].$$

The conditional power at the alternative, ξ_1, is

$$A(z_1, \xi_1) = 1 - \Phi\left[c(z_1) - \sqrt{n_2(z_1)}\,\xi_1\right].$$

To specify a design that has type I error rate of α, we need to pick the conditional power functions (and other design parameters such as critical values and sample sizes) so that

$$\begin{aligned}\alpha &= \int_{k_2}^{\infty} g_1(z_1, \xi_0)\, dz_1 + \int_{k_1}^{k_2} A(z_1, \xi_0) g_1(z_1, \xi_0)\, dz_1 \\ &= \alpha_1 + \int_{k_1}^{k_2} A(z_1, 0) g_1(z_1, 0)\, dz_1,\end{aligned}$$

where $g_1(z_1, \xi)$ is the probability density function of $Z_1 \sim Normal(\sqrt{n_1}\xi, 1)$. Similarly, to ensure power of $\rho \equiv 1 - \beta$, we need

$$\begin{aligned}1 - \beta = \rho &= \int_{k_2}^{\infty} g_1(z_1, \xi_1)\, dz_1 + \int_{k_1}^{k_2} A(z_1, \xi_1) g_1(z_1, \xi_1)\, dz_1 \\ &= \rho_1 + \int_{k_1}^{k_2} A(z_1, \xi_1) g_1(z_1, \xi_1)\, dz_1,\end{aligned}$$

Given stage 1 is already designed (n_1, k_1 and k_2), we can choose to use any $A(z_1, 0)$ and $A(z_1, \xi_1)$ as long as they satisfy these α and ρ conditions. Then we can find the critical value and sample size for stage 2 using

$$\begin{aligned}A(z_1, \xi_0) &= 1 - \Phi\left[c(z_1) - \sqrt{n_2(z_1)}\,\xi_0\right] \\ A(z_1, \xi_1) &= 1 - \Phi\left[c(z_1) - \sqrt{n_2(z_1)}\,\xi_1\right]\end{aligned}$$

This relationship can be used to derive the following:

$$\begin{aligned}n_2(z_1) &= \frac{\left(z_{A(z_1,\xi_0)} - z_{A(z_1,\xi_1)}\right)^2}{(\xi_1 - \xi_0)^2} \\ c(z_1) &= z_{A(z_1,\xi_0)} + \frac{\xi_0}{\xi_1 - \xi_0}\left(z_{A(z_1,\xi_0)} - z_{A(z_1,\xi_1)}\right)\end{aligned}$$

In above, we specified two $A(z_1, \xi)$ functions and solve for $n_2(z_1)$ and $c(z_1)$; however, we can specify any two of the four "design elements" and solve for the remaining two. Perhaps, we want to specify $A(z_1, \xi_0)$ so that the type I error rate is controlled and $n_2(z_1)$ so that the sample size is controlled. In this case, we have

$$\begin{aligned}c(z_1) &= z_{A(z_1,\xi_0)} + \xi_0\sqrt{n_2(z_1)} \\ A(z_1, \xi_1) &= 1 - \Phi\left[z_{A(z_1,\xi_0)} - \sqrt{n_2(z_1)}(\xi_1 - \xi_0)\right]\end{aligned}$$

```
sig <- 4
alp <- 0.025
bet <- 0.1
pow <- 1 - bet

del0 <- 0
del1 <- 1

delM <- 0.5
```

Example: To test $H_0 : \mu_t - \mu_c = 0$ and $H_1 : \mu_t - \mu_c > 0$. Assume σ is known to be 4. We want one sided α to be 0.03 and power to be 0.90 at $\mu_t - \mu_c = 1$.

```
xi0 <- del0/(sqrt(2) * sig)
xi1 <- del1/(sqrt(2) * sig)

xiM <- delM/(sqrt(2) * sig)
```

$\xi_0 = 0$
$\xi_1 = 1/\sqrt{2}\sigma = 1/4\sqrt{2} = 0.18.$

```
N <- ceiling((qnorm(alp) + qnorm(bet))^2/xi1^2)
```

For a single stage design, the sample size is

$$N = \frac{(z_{0.03} + z_{0.10})^2}{0.18} = 337$$

```
n1 <- round(N * 0.4)
```

Let's decide to look at the data when $n_1 = 135$ observations are available from each group. (approximately 40% of N)

```
alp1 <- 0.01
bet1 <- 0.025
```

We also need to decide how much of α and β we want to "spend" in stage 1. Let's choose $\alpha_1 = 0.01$ and $\beta_1 = 0.03$.

$$\begin{aligned} \alpha_1 &= P_0[\text{Reject } H_0 \text{ in stage 1}] \\ &= P_0[Z_1 > k_2] \qquad \text{where } Z_1 \sim Normal(0, 1) \end{aligned}$$

```
(k2 <- qnorm(1 - alp1))
```

```
[1] 2.33
```

Then $k_2 = 2.326$.

```
(zet1 <- xi1 * sqrt(n1))
```

```
[1] 2.05
```

$$\begin{aligned}\beta_1 &= P_1[\text{Accept } H_0 \text{ in stage 1}] \\ &= P_0[Z_1 < k_1] \qquad \text{where } Z_1 \sim Normal(\sqrt{n_1}\xi_1, 1) = Normal(2.05, 1)\end{aligned}$$

```
(k1 <- qnorm(bet1, zet1, 1))
```

```
[1] 0.094
```

Then $k_1 = 0.094$.

```
nMax <- 500
```

Also let's set the maximum sample size to be 500 (approximately 50% increase from N).

First, let's look at some stage 1 design characteristics:

```
(zeta0 <- sqrt(n1) * xi0)
```

```
[1] 0
```

```
(zetaM <- sqrt(n1) * xiM)
```

```
[1] 1.03
```

```
(zeta1 <- sqrt(n1) * xi1)
```

```
[1] 2.05
```

```
## Futility ##
fut0 <- pnorm(k1, zeta0)
futM <- pnorm(k1, zetaM)
fut1 <- pnorm(k1, zeta1)

## Rejection ##
rej0 <- 1 - pnorm(k2, zeta0)
rejM <- 1 - pnorm(k2, zetaM)
rej1 <- 1 - pnorm(k2, zeta1)

## Continue ##
cont0 <- 1 - fut0 - rej0
contM <- 1 - futM - rejM
cont1 <- 1 - fut1 - rej1
```

Some stage 1 characteristics

			Stage 1		
μ	ξ	ζ	Accept	Continue	Reject
0	0	0	0.54	0.45	0.01
0.50	0.09	1.03	0.18	0.73	0.10
1	0.18	2.05	0.03	0.58	0.39

Next, we will select the conditional power functions.

The unconditional type I error rate for the second stage is

```
(alp2 <- alp - alp1)
```

```
[1] 0.015
```

Similarly, the unconditional power (at the original alternative) for the second stage is

```
(pow2 <- pow - rej1)
```

```
[1] 0.507
```

If we are to use a flat $A(z_1\,\xi_0)$, we need the conditional type I error rate to be

```
(cp0 <- alp2/cont0)
```

```
[1] 0.0331
```

```
(cp1 <- pow2/cont1)
```

```
[1] 0.871
```

$$A_{flat}(z_1, \xi_0) = 0.02/0.45 = 0.03$$

Similarly, the flat conditional power is

$$A_{flat}(z_1, \xi_1) = (0.90 - 0.39)/0.58 = 0.87$$

These A-functions give rise to the following $n_2(z_1)$ and $c(z_1)$ regardless of Z_1 value.

$$n_2(z_1) = \frac{(z_{0.0331} - z_{0.8710})^2}{(0.1768 - 0)^2} = 282$$
$$c(z_1) = z_{0.0331} = 1.837$$

Design with flat A_0 and flat A_1

			Stage 1			Stage 2	This design			Single stage	
μ	ξ	n_1	Accept	Continue	Reject	Reject	Power	$E[N]$	Max N	Power	N
0	0	135	0.538	0.452	0.010	0.015	0.025	262.5	417	0.025	337
0.50	0.088	135	0.175	0.728	0.097	0.263	0.360	340.1	417	0.367	337
1	0.177	135	0.025	0.582	0.393	0.507	0.900	299.1	417	0.900	337

Now consider A-functions of the form $A(z_1, \xi_0) = a_0 + a_1(z_1 - k_1)^2$ and $A(z_1, \xi_1) = b_0 + b_1(z_1 - k_1)$. We can use any A functions as long as they satisfy α and power conditions.

```
a0 <- 0.002
```

First we pick a_0 (the value of $A(z_1, \xi_0)$ at $z_1 = k_1$ to be 0.00 and solve for a_1 so that

$$\alpha_2 = 0.015 = \int_{k_1}^{k_2} A(z_1, \xi_0) g_1(z_1, \xi_0)\, dz_1.$$

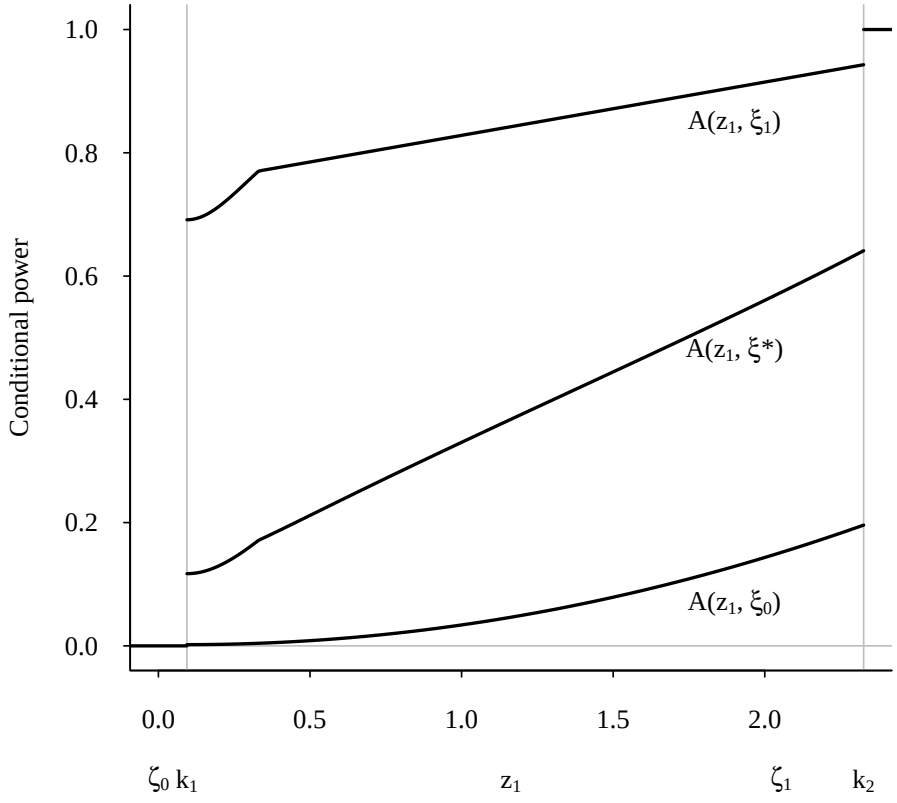
```
Afun0 <- function(z, k1, a0, a1) a0 + a1 * (z - k1)^2
integrand <- function(z, k1, a0, a1, zeta) Afun0(z, k1, a0, a1) * dnorm(z,
    zeta, 1)

unconditionalAlpha <- function(k1, k2, a0, a1, zeta) {
    integrate(integrand, lower = k1, upper = k2, k = k1, a0 = a0, a1 = a1,
        zeta = zeta)$value
}

f <- function(a1, k1, k2, a0, zeta, alp2) alp2 - unconditionalAlpha(k1,
    k2, a0, a1, zeta)

(a1 <- uniroot(f, lower = 0, upper = 0.1, k1 = k1, k2 = k2, a0 = 0.002,
    zeta = 0, alp2 = alp2)$root)


[1] 0.0389
```

The above numerical integration found $a_1 = 0.04$.

$$A(z_1, \xi_0) = 0.00 + 0.04(z_1 - 0.09)^2.$$

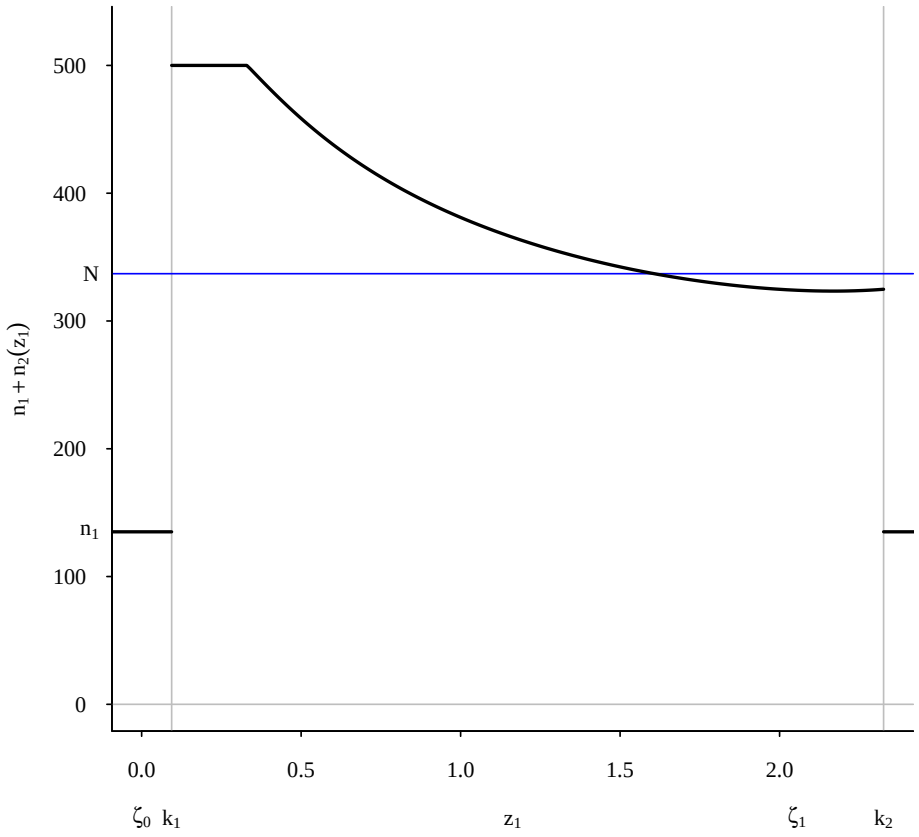
Similarly for $A(z_1, \xi_1)$, by specifying $b_0 = 0.75$, we find $b_1 = 1.004$ so that

$$A(z_1, \xi_1) = 0.75 + 1.004(z_1 - k_1).$$



Using these two A functions, we can compute $n_2(z_1)$, and it turns out $\max\{n_1 + n_2(z_1)\} > 500$, and we need to modify the design a little. It is relatively simple to make small modifications to the design because we understand how the design elements $A(z_1, \xi_0)$, $A(z_1, \xi_1)$, $n_2(z_1)$, and $c(z_1)$, are interrelated.

First while fixing $A(z_1, \xi_0)$, we "tap" $n_2(z_1)$ so that $\max\{n_1 + n_2(z_1)\} = 500$. This action changes $A(z_1, \xi_1)$ slightly resulting a smaller power than 0.90. To make the power 0.90 again, we add a constant to the new $A(z_1, \xi_1)$ but capping the resulting $n_1 + n_2(z_1)$ at 500. The final design is shown below graphically.



Design with flat A_0 and flat A_1

			Stage 1			Stage 2	This design			Single stage	
μ	ξ	n_1	Accept	Continue	Reject	Reject	Power	$E[N]$	Max N	Power	N
0	0	135	0.538	0.452	0.010	0.015	0.025	262.5	417	0.025	337
0.50	0.088	135	0.175	0.728	0.097	0.263	0.360	340.1	417	0.367	337
1	0.177	135	0.025	0.582	0.393	0.507	0.900	299.1	417	0.900	337

Design with the quadratic A_0 and linear A_1

			Stage 1			Stage 2	This design			Single stage	
μ	ξ	n_1	Accept	Continue	Reject	Reject	Power	$E[N]$	Max N	Power	N
0	0	135	0.538	0.452	0.010	0.015	0.025	264.2	500	0.025	337
0.50	0.088	135	0.175	0.728	0.097	0.264	0.360	318.4	500	0.367	337
1	0.177	135	0.025	0.582	0.393	0.507	0.900	264.7	500	0.900	337

In the literature, many specific A functions have been proposed. A few examples include:

- Proschan & Hunsberger (1995)

$$A_{\text{PH}}(z_1, \xi) = 1 - \Phi\left[\sqrt{n_1}\sqrt{(k_2 - \xi\sqrt{n_1})^2 - (z_1 - \xi\sqrt{n_1})^2}\right]$$

- Chen, DeMets & Lan (2004)

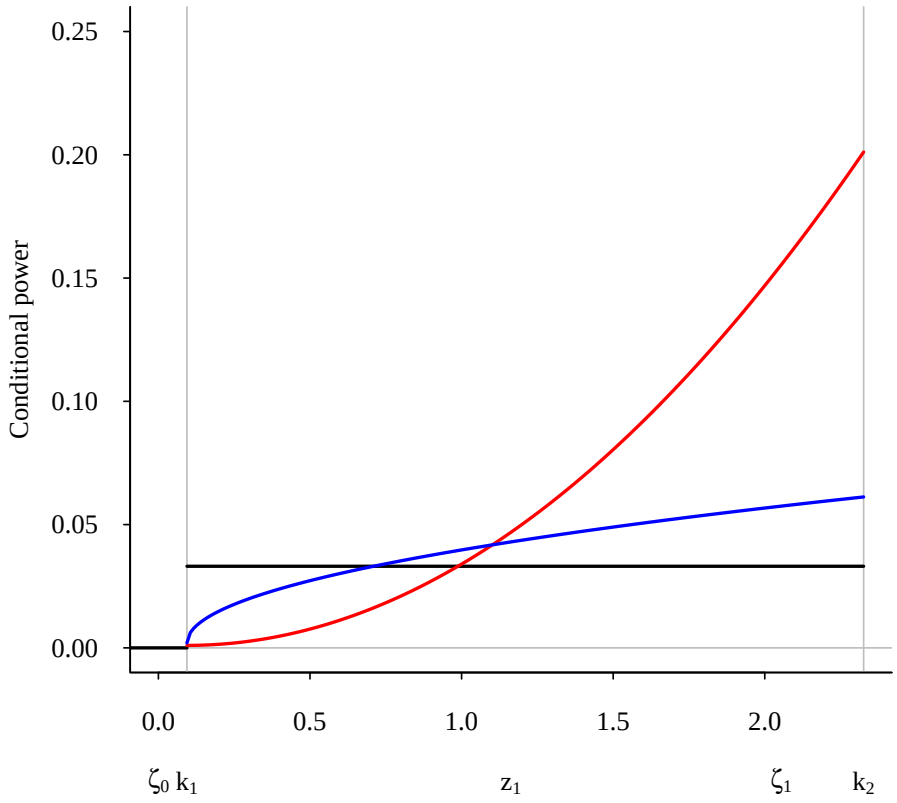
$$A_{\text{CDL}}(z_1, \xi) = 1 - \Phi\left[\sqrt{2}z_\alpha - z_1 - \xi\sqrt{n_1}\right]$$

$A(z_1, z_1)$ "Conditional power under the current trend"

14.5 Unspecified designs

The minimum requirement to control type I error rate is to pre-specify $A(z_1, \xi_0)$ function that satisfies α condition. Then after the first stage, when the actual z_1 from the data are available, pick $n_2(z_1)$ so that conditional powers at any a value of ξ (other than ξ_0) can be set.

If we allow even $A(z_1, \xi_0)$ to be specified after the fist stage, type I error rate cannot be controlled. There exist many (in fact infinite number of) $A(z_1, \xi_0)$ functions that give the desired value of α_1 (0.01 in our example). Depending on z_1 the required sample size to guarantee a certain conditional power differs. We cannot choose an $A(z_1, \xi_0)$ function that gives the minimum sample size for the observed z_1. Roughly speaking, when the conditional type I error rate at the observed z_1 is large, the required sample size is small for the same conditional power. All three conditional type I error rates in the following plot give $\alpha = 0.025$.

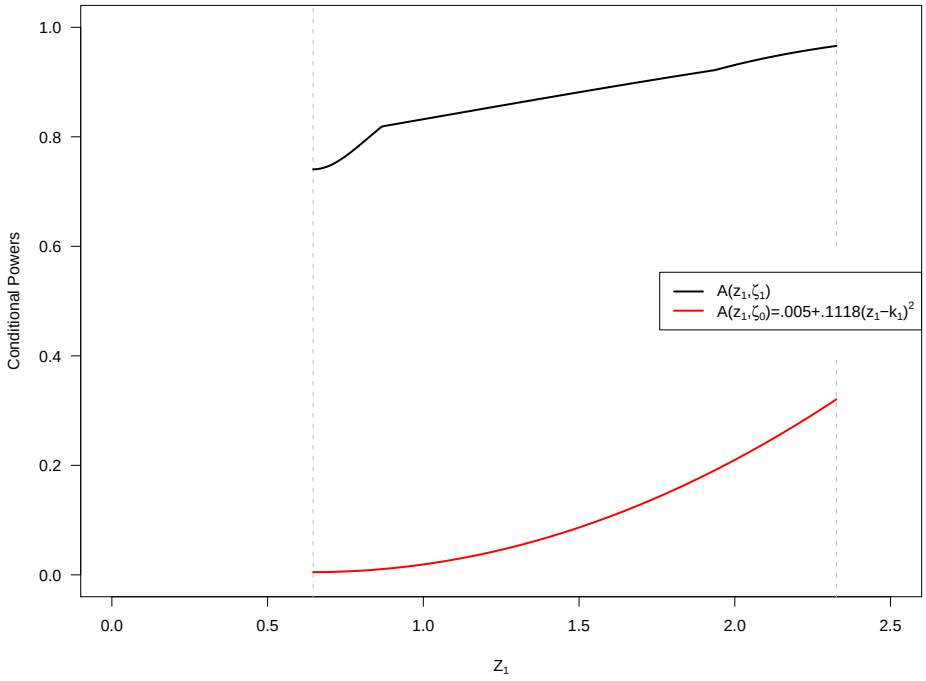


14.6 Ordering of sample space

To compute p-values and confidence interval (through inverting hypothesis tests), we need to define an ordering of sample space. However, this task is difficult because of sample size difference for potential values of z_1.

One useful fact (not too difficult to show) is that the decision rule, "reject if $Z_2 > c(z_1)$" is equivalent to the rule "reject if stage 2 *conditional* p-value is less than $A(z_1, \xi_0)$ evaluated at the observed z_1." So we can compute a *conditional* p-value just using the stage 2 data, $P_0[Z_2 > z_2]$, and compare it to the *conditional* type I error rate computed at the observed z_1.

Different example! Suppose the following conditional type I error rate is used:

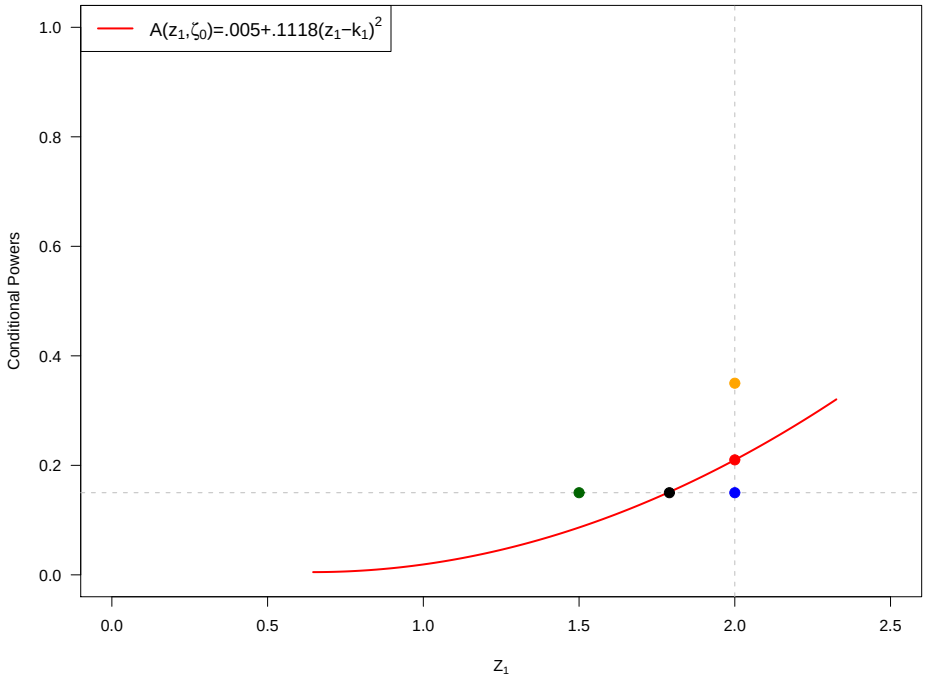


- if $z_1 = 2.0$ and stage 2 conditional p-value is 0.10 then H_0 will be rejected because this p-value is less than $A(z_1, \xi_0)$ i.e., below the red line.
- if $z_1 = 1.0$ and stage 2 conditional p-value is 0.10 then H_0 will not be rejected.

Therefore, we need an ordering of the sample space that takes into account not only the sample size of stage 2, $n_2(z_1)$, but also the conditional type I error rate for the stage 2, $A(z_1, \xi_0)$.

Suppose we choose to use $A(z_1, \xi_0)$ and $A(z_1, \xi_1)$ shown in the plot below.

And let's consider the following 5 sample paths indicated by the conditional p values. Can we order the strength of evidence against H_0 for these data?



- When z_1 is the same, the second stage sample size is the same, so it should be simple to order the sampling paths. The smaller the p value, the stronger the evidence against H_0.
  Blue $\succ$ Red $\succ$ Yellow.
- When the conditional p values are the same, then we can order them by the strength of evidence in the first stage.
  Blue $\succ$ Black $\succ$ Green.
- The black and red dots should indicate equal strength of evidence because they both result in "just" rejecting H_0.

So in the above picture, the only unclear ordering is between Green and Yellow.

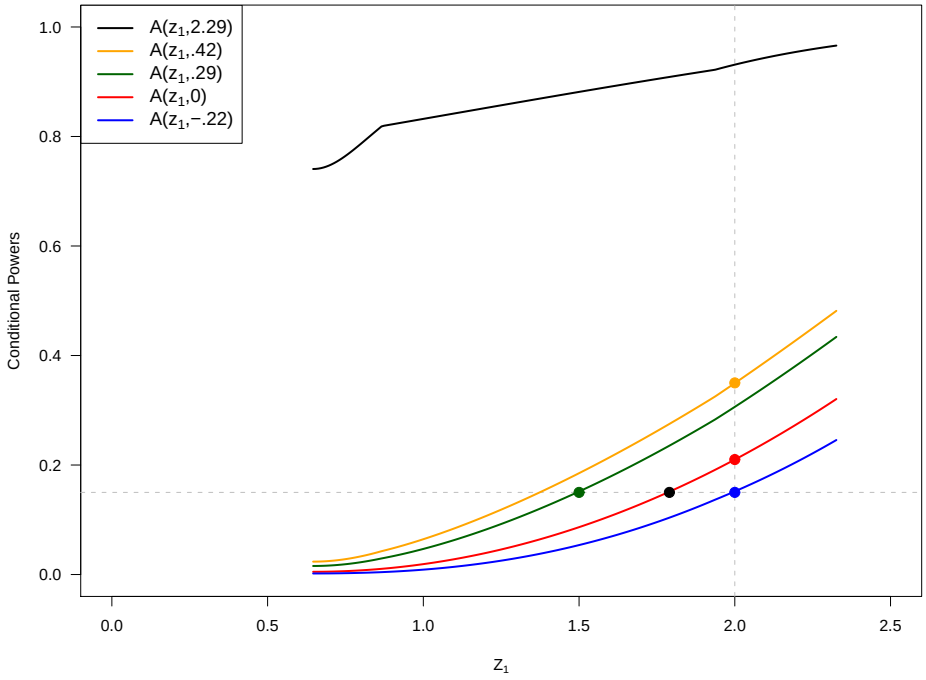
The third rule gives a hint as to how to proceed; the data leading to the black and red dots indicate that those data have just enough evidence to reject H_0. The blue dot is for a sampling path that gives stronger evidence against H_0; we could reject H_0' that is more extreme.

We can find a value of ξ_0^* (equivalently, μ_0^*) so that $A(z_1, \xi_0^*)$ goes through the blue dot, and say we could have rejected $H_0^* : \mu = \mu_0^*$.

Technically speaking, we can find ξ_0^* by solving the following:

$$p = A(z_1, \xi_0^*) = 1 - \Phi\left[c(z_1) - \sqrt{n_2(z_1)}\,\xi_0^*\right]$$

Note that once a value of z_1 is observed, we can evaluate $c(z_1)$ and $n_2(z_1)$, so the only unknown quantity in the above expression is ξ_0^*.



From the picture above , we know the ordering is: Blue $\succ$ Red $=$ Black $\succ$ Green $\succ$ Yellow. And "some as or more extreme" than the observed is anything on and below the line, and we can compute the p-value by computing

$$\int_{k_2}^{\infty} g_1(z_1, \xi_0)\, dz_1 + \int_{k_1}^{k_2} A(z_1, \xi_0^*) g_1(z_1, \xi_0)\, dz_1$$

This method (ordering) guarantees that the p-value and the corresponding hypothesis testing are consistent (p-value $< \alpha$ iff H_0 is rejected).

And it can be shown that when $n_2(z_1)$ and $c_u(z_1)$ {critical value for the combined statistic} are constants, this ordering reduces to the stage-wise ordering.

14.7 Predictive power

With an unspecified design, some people are reluctant to use the conditional power to determine the design of the second stage. One issue is that where to compute the conditional power is not always clear.

The original alternative is usually a reasonable choice ($A(z_1, \xi_1)$). However, when the observed z_1 is much different (smaller) from ξ_1 we may not be interested in the conditional power at ξ_1 but at some smaller value that is still clinically meaningful. (Minimum clinically relevant alternative $= \xi_1^\dagger$)

Another popular choice is $\hat{\xi} \equiv z_1/\sqrt{n_1}$ ("alternative under the current trend").

Or maybe we should compute the conditional power at somewhere in between ξ_1 and $\xi_1^\dagger$. Average? Now we are talking like a Bayesian because we are talking about an average of ξs which are, for a frequentist, parameters. Maybe we have a prior distribution of ξ (or equivalently μ). And a posterior distribution of ξ after the first stage, $\pi(\xi|z_1)$, and we can compute a weighted average of the conditional powers with respect to the posterior distribution. Something like

$$\int_{-\infty}^{\infty} A(z_1, \xi)\pi(\xi|z_1)\, d\xi,$$

and this is often called a predictive power given the stage one data.

The conditional power is a frequentist concept, and it is computed at one value of ξ. The predictive power is a Bayesian concept, and it is a weighted average of the conditional power with respect to a posterior distribution of ξ.

Chapter 15

Factorial design

15.1 Introduction

Factorial clinical trials (Piantadosi) Experiments that test the effect of more than one treatment using a design that permits an assessment of interactions among the treatments

The simplest example of a factorial design is 2 treatment, 2 treatment groups (2 by 2) designs. With this design, one group receives both treatment, a second group receives neither, and the other two groups receive one of A or B.

	Treatment B		
Treatment A	No	Yes	Total
No	n	n	$2n$
Yes	n	n	$2n$
Total	$2n$	$2n$	$4n$

Four treatment groups and sample sizes in a 2×2 balanced factorial design.

Alternatives to a 2×2 factorial design

- Two separate trials (for A and for B)
- Three arm trial (A, B, neither)

Two major advantages of factorial design (but not simultaneously):

- Allows investigation of interactions (drug synergy).

  Drug synergy occurs when drugs interact in ways that enhance effects or side-effects of those drugs.

- Reduces the cost (sample size) if the drugs do not interact.

Some requirements for conducting a clinical trial with factorial design:

- The side effects of two drugs are not cumulative to make the combination unsafe to administer.
- The treatments need to be administered in combination without changing dosage of the individual drugs.
- It is ethical not to administer the individual drugs. A and B may be given *in addition* to a standard drug so all groups receive some treatment.
- We need to be genuinely interested in studying drug *combination*, otherwise some treatment combinations are unnecessary.

Some terminology

- Factors (how many different treatments are in consideration)
- Levels (2 if yes/no)
- 2^k factorial studies have k factors, each with two levels (presence/absence)
- Full factorial design has no empty cells.

- Unreplicated study has one sample per cell (obviously not very common in clinical studies)
- Fractional factorial designs (some cells are left empty by design)
- Complete block designs / Incomplete block designs
- Latin squares

15.2 Notation and assumptions

	Treatment B	
Treatment A	No	Yes
No	μ	$\mu + \beta$
Yes	$\mu + \alpha$	$\mu + \alpha + \beta + \gamma$

With this formulation, α is the effect of treatment A, β is the effect of treatment B, and γ is the interaction effect. (If the effects of A and B are additive with no interaction, then $\gamma = 0$.)

For a continuous outcome and large sample sizes (may be different for each group), we have the following for the observed sample cell means.

$$\begin{aligned}
\overline{Y}_0 &\sim Normal(\mu, \sigma^2/n_0) \\
\overline{Y}_A &\sim Normal(\mu + \alpha, \sigma^2/n_A) \\
\overline{Y}_B &\sim Normal(\mu + \beta, \sigma^2/n_B) \\
\overline{Y}_{AB} &\sim Normal(\mu + \alpha + \beta + \gamma, \sigma^2/n_{AB})
\end{aligned}$$

We assume $n = n_0 = n_A = n_B = n_{AB}$.

15.3 Test for the interaction effect

In a factorial design, we usually test the presence of interaction effect first.

$$H_0 : \gamma = 0$$
$$H_1 : \gamma \neq 0$$

The observed mean responses are:

	Treatment B	
Treatment A	No	Yes
No	$\overline{Y}_0$	$\overline{Y}_B$
Yes	$\overline{Y}_A$	$\overline{Y}_{AB}$

The interaction effect may be estimated by

$$\hat{\gamma} = (\bar{Y}_{AB} - \bar{Y}_B) - (\bar{Y}_A - \bar{Y}_0),$$

and

$$Var(\hat{\gamma}) = \frac{4\sigma^2}{n}. \qquad \text{(Why is this problematic?)}$$

Thus, if we assume σ^2 is known, then

$$Z = \frac{\hat{\gamma}}{2\sigma/\sqrt{n}}$$

has $Normal(0, 1)$ distribution under H_0.

If we have to estimate σ^2 and assume within-group variances are equal, we use a pooled sample variance, $s_p^2 = (s_0^2 + s_A^2 + s_B^2 + s_{AB}^2)/4$.

The test statistic

$$t = \frac{\hat{\gamma}}{2s_p/\sqrt{n}}$$

has a t distribution with $df = 4(n - 1)$ under H_0.

15.4 Treatemnt effect

15.4.1 $\gamma \neq 0$

The treatment A effect can be estimated as

$$\hat{\alpha} = \bar{Y}_A - \bar{Y}_0,$$

and its variance is

$$Var(\hat{\alpha}) = \frac{2\sigma^2}{n}.$$

And we have

$$Z = \frac{\hat{\alpha}}{\sqrt{2}\sigma/\sqrt{n}} \sim Normal(0,1)$$

under H_0. We can estimate σ^2 by $s_p^2 = (s_A^2 + s_0^2)/2$. (Note: $2(n-1)s_p^2/\sigma^2 \sim \chi^2_{2(n-1)}$.) Then

$$t = \frac{\hat{\alpha}}{\sqrt{2}s_p/\sqrt{n}}$$

has a t distribution with $df = 2(n-1)$ under H_0. Constructing the test for β is exactly the same.

15.4.2 $\gamma = 0$

If no interaction is present then $\gamma = 0$, and $\tilde{\alpha} = \bar{Y}_{AB} - \bar{Y}_B$ can also be used to estimate α. If we use the average of $\hat{\alpha}$ and $\tilde{\alpha}$ to estimate α, this estimator has a smaller variance.

$$\breve{\alpha} = \frac{\hat{\alpha} + \tilde{\alpha}}{2} = \frac{(\bar{Y}_A - \bar{Y}_0) + (\bar{Y}_{AB} - \bar{Y}_B)}{2}$$
$$Var(\breve{\alpha}) = \frac{1}{4}Var\left(\bar{Y}_A - \bar{Y}_0 + \bar{Y}_{AB} - \bar{Y}_B\right) = \frac{\sigma^2}{n}$$

Similarly to before,

$$Z = \frac{\breve{\alpha}}{\sigma/\sqrt{n}} \sim Normal(0,1), \text{and}$$
$$t = \frac{\breve{\alpha}}{s_p/\sqrt{n}} \sim t_{df}$$

under H_0. (What's the df?) Here we use

$$s_p^2 = (s_0^2 + s_A^2 + s_B^2 + s_{AB}^2)/4,$$

and

$$\frac{4(n-1)}{\sigma^2} s_p^2 \sim \chi^2_{4(n-1)}$$

Constructing the test for β is exactly the same.

In order to have the same efficiency in a two-arm trial (A vs placebo), we would need $2n$ patients in each treatment arm.

$$var\left(\hat{\alpha}_1\right) = \frac{2\sigma^2}{2n} = \frac{\sigma^2}{n}.$$

So if we were to test A and B in two separate experiments we would need $2n$ per arm $\times$ 4 arms (A and placebo, B and placebo), totaling $8n$ subjects. Noticing we are repeating the placebo in these hypothetical experiments, we decide to use a 3-arm experiment with A, B, and placebo arms. Then we would require a total of $6n$ subjects for the same precision.

15.5 Examples

The group means are:

	Treatment B	
Treatment A	No	Yes
No	10	40
Yes	30	60

If there is a synergistic effect, then $\eta_{11} > 60$.

	Treatment B	
Treatment A	No	Yes
No	10	40
Yes	30	80

	Treatment B	
Treatment A	No	Yes
No	10	40
Yes	30	120

In the last situation, the treatment effects may be multiplicative.

	Treatment B	
Treatment A	No	Yes
No	$\log(10) = 1$	$\log(40) = 1.60$
Yes	$\log(30) = 1.48$	$\log(120) = 2.08$

Suppose the samples of size 20 yield the following estimates of the cell means.

	Treatment B	
Treatment A	No	Yes
No	9.83	40.05
Yes	28.94	59.76

Assuming no interaction, to estimate the drug A effect we compute either

$$\hat{\alpha}_1 = \bar{Y}_A - \bar{Y}_0 = 28.94 - 9.83 = 19.11$$

or

$$\tilde{\alpha}_1 = \bar{Y}_{AB} - \bar{Y}_B = 59.76 - 40.05 = 19.71$$

or their average $(19.11 + 19.71)/2 = 19.41$.

How bad is it to estimate α_1 this way when there is actually a significant interaction?

$$\begin{aligned} E[(\hat{\alpha}_1 + \tilde{\alpha}_1)/2] &= \frac{1}{2} E[(\bar{Y}_A - \bar{Y}_0) + (\bar{Y}_{AB} - \bar{Y}_B)] \\ &= \frac{1}{2} \left((\mu + \alpha_1) - \mu + (\mu + \alpha_1 + \beta_1 + \gamma_{11}) - (\mu + \beta_1) \right) \\ &= \alpha_1 + \frac{\gamma_{11}}{2} \end{aligned}$$

15.5.1 Example: the Physician's Health Study I (1989)

Read all about it on `http://phs.bwh.harvard.edu/`.
The Physician's Health Study was a randomized clinical trial designed to test the following two theories:

- Daily low-dose aspirin use reduces the risk of cardiovascular disease.
- Beta carotene reduces the risk of cancer.

Population hierarchy:

- 261,248 US male MDs aged 40 to 84.
- 112,528 responded to questionnaires.

- 59,285 willing to participate.
- 33,332 willing and eligible MDs enrolled in run-in (18 weeks of active aspirin and beta-carotene placebo).

  Run-in period Eligible patients are monitored for treatment compliance.

- $22,071$ randomized

	Beta-carotene		
Aspirin	Active	Placebo	Total
Active	$5,517$	$5,520$	$11,037$
Placebo	$5,519$	$5,515$	$11,034$
Total	$11,036$	$11,035$	$22,071$

Major findings:

- The trial's DSMB stopped the aspirin arm several years ahead of schedule on 1988/1/25 because it was clear that aspirin had a significant effect on the risk of a first myocardial infarction. (It reduced the risk by 44%.)
  - Did it change the sample sizes for the Beta-carotene components? (Next homework?)
- There were too few strokes or deaths to base sound clinical judgement regarding aspirin and stroke or mortality.
- The beta-carotene arm terminated as scheduled on 1995/12/12 with the conclusion that 13 years of supplementation with beta-carotene produced neither benefit nor harm. Beta-carotene alone was not responsible for the health benefit seen among people who ate plenty of fruits and vegetables.
- Over 300 other findings have emerged from the trial so far.

15.6 Treatment interactions

Factorial designs are the only way to study treatment interactions. Recall the interaction term is estimated by $\hat{\gamma} = (\bar{Y}_{AB} - \bar{Y}_B) - (\bar{Y}_A - \bar{Y}_0)$, and its variance is $Var(\hat{\gamma}) = 4\sigma^2/n$. This variance is 2 times as large as that of A and B main effects, and to have the same precision for an estimate of an interaction effect, the sample size has to be 4 times as large. This means, the two main advantages of the factorial designs (efficiency and interaction objectives) cannot be satisfied simultaneously.

When there is an AB interaction, we cannot use the estimators, $\breve{\alpha}$ and $\breve{\beta}$, which are only valid with no interaction effect. In fact, we cannot talk about an overall main effect in the presence of an interaction. Instead, we can talk about the effect of A in the absence of B,

$$\alpha = \bar{Y}_A - \bar{Y}_0,$$

or the effect of A in the presence of B

$$\alpha' = \alpha + \gamma = \bar{Y}_{AB} - \bar{Y}_B.$$

Some additional notes

- In the $2 \times 2 \times 2$ design (2^3 design), there are 3 main effects and 4 interactions possible. The number of high order interactions will grow quickly with k, but oftentimes, they are (assumed to be) 0.
- A "quantitative" interaction does not affect the direction of the treatment effect. For example when treatment B is effective either with or without treatment A, but the magnitude of its effectiveness changes.
- With a "qualitative" interaction, the effects of A are reversed with the presence of B. In this case, an overall treatment A effect does not make sense.
- The factorial design can be analyzed with linear models (analysis of variance models).

Limitations of factorial designs

- A higher level design can get complex quickly.
- Test for interaction requires a large sample size (or have a very low power if the study is powered for the main effects).
- Combination therapy may be considered as a treatment in its own right.

Of further interest...

- Partial (fractional) factorial designs have missing cells by design (especially when higher order interactions are assumed to be zero)

Chapter 16

Crossover design

Crossover trials are those in which each patient is given more than one treatment, each at different times in the study, with the intent of estimating differences between them.

In a simple 2×2 design (or AB/BA design), patients are randomized to either “A then B” group or “B then A” group.

	Period	
Group	I	II
AB	Treatment A	Treatment B
BA	Treatment B	Treatment A

2 Treatments / 2 Periods / 2 Sequences

	P_1	P_2	
S_1	A	B	n_1
S_2	B	A	n_2

---

2 Treatments / 2 Periods / 4 Sequences

	P_1	P_2	
S_1	A	B	n_1
S_2	B	A	n_2
S_3	A	A	n_3
S_4	B	B	n_4

2 Treatments / 4 Periods / 2 Sequences

	P_1	P_2	P_3	P_4	
S_1	A	B	A	B	n_1
S_2	B	A	B	A	n_2

16.1 Some characteristics of crossover design

- All subjects receive more than one treatment (not simultaneously).
- Each subject acts as own control. Therefore, the treatment groups are comparable without relying on randomization.
  - Treatment periods (order of A and B) are often randomly assigned.
  - Baseline characteristics are identical with regard to many patient characteristics, but not with regard to their recent history of exposure to other potentially effective treatments. **carryover effects**
  - The comparability of the treatment groups is not guaranteed by the structure of the trial alone. The investigators need to estimate the carryover effects.
- Crossover designs are not used ...
  - with any condition that treatment could effect considerable change.
  - for acute illness.
- Crossover designs are most suitable for treatments intended for rapid relief of symptoms in chronic diseases, where the long-term condition of the patient remains fairly stable.

Precision
The primary strength of crossover trials is increased efficiency. Suppose the treatment effects are

$$Y_t \sim Normal(\mu_t, \sigma^2),$$
$$Y_c \sim Normal(\mu_c, \sigma^2),$$

and we are interested in $\mu_t - \mu_c$. In a parallel design (with per group sample size of n), we have

$$\hat{\Delta} = \overline{Y}_t - \overline{Y}_c \sim Normal\left(\mu_t - \mu_c, \frac{2\sigma^2}{n}\right).$$

With a TC/CT crossover design with sample size of n,

$$\begin{aligned} var(\hat{\Delta}) &= \frac{2\sigma^2}{n} - 2cov(\overline{Y}_t, \overline{Y}_c) \\ &= \frac{2\sigma^2}{n}(1 - \rho_{tc}), \end{aligned}$$

where ρ_{tc} is the within-subject correlation of responses on treatments T and C. Therefore, a crossover design is more efficient than a parallel design given $\rho_{tc} > 0$.

Recruitment
Some patients may hesitate to participate in a clinical trial if there is a 50% probability of not receiving any effective treatment. With a crossover design, everyone is guaranteed to receive the test drug.

On the other hand, the patients may hesitate to participate in a crossover trial because they will go through more than one treatment, especially when outcomes are assessed with diagnostic procedures such as X-ray, blood drawing, lengthy questionnaires.

Carryover effects
The biggest concern is the possibility that the treatment effect from one period might continue to be present during the following period. A sufficiently long "washout" period between the treatments may prevent significant carryover effects (but how long is sufficiently long?). If there are baseline measurements that represent patient's disease status, this can be checked against their baseline levels.

If the treatment effects a permanent change or cure in the underlying condition, the treatment given after could look artificially superior.

Dropouts
In a crossover design, the trial duration tends to be longer than a comparable study using independent groups, which may cause more dropouts. Also because every patient take more than one treatment, dropouts due to severe side effects may also increase. The consequences of dropouts are more severe in crossover trial; a simple analysis cannot use only the data from the first period.

16.2 Analysis of 2×2 crossover design

	P_1	P_2
$S_1 = AB$	$\mu_{A1} = \beta_0$	$\mu_{B2} = \beta_0 + \beta_1 + \beta_2$
$S_2 = BA$	$\mu_{B1} = \beta_0 + \beta_1$	$\mu_{A2} = \beta_0 + \beta_2 + \beta_3$

$\beta_0 \cdots$ Treatment A effect
$\beta_1 \cdots$ Increment of treatment effect due to B.
$\beta_2 \cdots$ Carryover effect of treatment A
$\beta_3 \cdots$ Increment carryover effect of treatment B

Treatment B effect is $\beta_0 + \beta_1$, and the carryover effect due to treatment B is $\beta_2 + \beta_3$.

The primary hypotheses to test are:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

And how to conduct this test (how to estimate β_1) depends on whether $\beta_3 = 0$. (We'll see why later.)

Step 0: Assumptions

1. Sample size is n for S_1 and S_2.
2. $\overline{Y} \sim Normal\left(\mu, \sigma^2/n\right)$.
3. $Cor(Y_{A1}, Y_{B2}) = Cor(Y_{B1}, Y_{A2}) = \rho$.

Data:

	P_1	P_2
$S_1 = AB$	$\overline{Y}_{A1} = \hat{\beta}_0$	$\overline{Y}_{B2} = \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2$
$S_2 = BA$	$\overline{Y}_{B1} = \hat{\beta}_0 + \hat{\beta}_1$	$\overline{Y}_{A2} = \hat{\beta}_0 + \hat{\beta}_2 + \hat{\beta}_3$

Step 1: Test $\beta_3 = 0$
Note that under $H_0 : \beta_3 = 0$, we have $\mu_{A1} + \mu_{B2} = \mu_{B1} + \mu_{A2}$. Thus, we can use

$$Z_1 = \frac{(\overline{Y}_{B1} + \overline{Y}_{A2}) - (\overline{Y}_{A1} + \overline{Y}_{B2})}{\sqrt{Var(\overline{Y}_{B1} + \overline{Y}_{A2}) + Var(\overline{Y}_{A1} + \overline{Y}_{B2})}}$$

to test these hypotheses (assuming σ^2 is known).

Why is this good (convenient)? Because we don't have to worry about the correlations when computing Z_1. Instead, we can compute the within-subject ~~difference~~ sum. Let's say $\omega_{i1} = y_{A1i} + y_{B2i}$ and $\omega_{j2} = y_{B1j} + y_{A2j}$. Then we have

$$Z_1 = \frac{\overline{\omega}_2 - \overline{\omega}_1}{\sqrt{Var(\overline{\omega}_2) + Var(\overline{\omega}_1)}}$$

If we don't assume σ^2 is known, we can use a simple two-sample t-test.

$$t_1 = \frac{\overline{\omega}_2 - \overline{\omega}_1}{\sqrt{2} s_\omega / \sqrt{n}},$$

where $s^2_\omega = (s^2_{\omega_1} + s^2_{\omega_2})/2$. What's the degree of freedom?

Why is this bad?

Step 2a (If $\beta_3 = 0$)

We can estimate β_1 and test if $H_0 : \beta_1 = 0$. Let's say $\delta_{i1} = y_{B2i} - y_{A1i}$ and $\delta_{j2} = y_{A2j} - y_{B1j}$. Let's confirm that the following test statistic can be used to test this hypothesis.

$$t_2 = \frac{\overline{\delta}_1 - \overline{\delta}_2}{\sqrt{2} s_\delta / \sqrt{n}},$$

where $s_\delta^2 = (s_{\delta_1}^2 + s_{\delta_2}^2)/2$. What's the degree of freedom?

We are interested in estimating β_1. Note that $\hat{\beta}_1 = (\overline{\delta}_1 - \overline{\delta}_2)/2$. So a 95% confidence interval for β_1 is

$$\hat{\beta}_1 \pm t_{0.975,2(n-1)} se(\hat{\beta}_1)$$
$$\frac{\overline{\delta}_1 - \overline{\delta}_2}{2} \pm t_{0.975,2(n-1)} \sqrt{s_\delta^2/(2n)}$$

Step 2b (If $\beta_3 = 0$)

We can estimate β_2 and test if $H_0 : \beta_2 = 0$. Note that $\hat{\beta}_2 = (\overline{\delta}_1 + \overline{\delta}_2)/2$ and $se(\hat{\beta}_2) = se(\hat{\beta}_1)$. Thus a 95% confidence interval for β_2 is

$$\frac{\overline{\delta}_1 + \overline{\delta}_2}{2} \pm t_{0.975,2(n-1)} \sqrt{s_\delta^2/(2n)}$$

But we are not that interested in estimating β_2.

Step 2c (If $\beta_3 = 0$)

Maybe we want to estimate β_0. We have two estimates of β_0, and we can take the average of them to get

$$\begin{aligned}
\hat{\beta}_0 &= \frac{1}{2}\left(\overline{Y}_{A1} + (\overline{Y}_{B1} + \overline{Y}_{A2} - \overline{Y}_{B2})\right) \\
&= \frac{1}{2}\left(\overline{Y}_{B1} + \overline{Y}_{A2} - (\overline{Y}_{B2} - \overline{Y}_{A1})\right) \\
&= \frac{1}{2}(\overline{\omega}_2 - \overline{\delta}_1)
\end{aligned}$$

This means we can test to see if $\beta_0 = 0$ by testing

$$H_0 : \mu_{B1} + \mu_{A2} = \mu_{B2} - \mu_{A1}$$
$$H_1 : \mu_{B1} + \mu_{A2} \neq \mu_{B2} - \mu_{A1}$$

Constructing the relevant t test statistic is not as straightforward as the previous steps because we cannot assume the true variances of ω_2 and δ_1 are equal. We can use

$$t = \frac{\overline{\omega}_2 - \overline{\delta}_1}{\sqrt{\frac{s_{\omega_2}^2}{n} + \frac{s_{\delta_1}^2}{n}}},$$

which follows a t distribution approximately with estimated degrees of freedom given by the Satterthwaite formula.

To estimate β_0 with a confidence interval, compute a confidence interval

$$(\overline{\omega}_2 - \overline{\delta}_1) \pm t_{1-\alpha/2,df}\sqrt{\frac{s^2_{\omega_2}}{n} + \frac{s^2_{\delta_1}}{n}}$$

first (applying the unequal-varince t-test), and divide it by 2.

Step 2d (If $\beta_3 = 0$)

Maybe we want to estimate $\beta_0 + \beta_1$. Again we can use the average of two estimates:

$$\begin{aligned}\hat{\beta}_0 + \hat{\beta}_1 &= \frac{1}{2}\left(\overline{Y}_{B1} + (\overline{Y}_{B2} - \overline{Y}_{A2} + \overline{Y}_{A1})\right) \\ &= \frac{1}{2}\left(\overline{Y}_{A1} + \overline{Y}_{B2} - (\overline{Y}_{A2} - \overline{Y}_{B1})\right) \\ &= \frac{1}{2}\left(\overline{\omega}_1 - \overline{\delta}_2\right)\end{aligned}$$

Follow the same process as in step 2c.

Step 3 (If $\beta_3 \neq 0$)

Because the carryover affects S_1 and S_2 differently, we cannot eliminate β_2 as we did before.

$$\begin{aligned}\overline{Y}_{B2} - \overline{Y}_{A1} &= \hat{\beta}_1 + \hat{\beta}_2 \\ \overline{Y}_{B1} - \overline{Y}_{A2} &= \hat{\beta}_1 - \hat{\beta}_2 - \hat{\beta}_3\end{aligned}$$

Taking the within-individual difference is not going to help, so we cannot take an advantage of the correlated endpoint. In this case, we ignore the data from the second period.

$$\begin{aligned}\hat{\beta}_1 &= \overline{Y}_{B1} - \overline{Y}_{A1} \\ Var(\hat{\beta}_1) &= \frac{2\sigma^2}{n}\end{aligned}$$

Treat the study as a two sample test with sample size $= n$ per group.

Moreover, we can estimate the treatment effects with

$$\begin{aligned}\hat{\beta}_0 &= \overline{Y}_{A1} \\ \hat{\beta}_0 + \hat{\beta}_1 &= \overline{Y}_{B1}\end{aligned}$$

16.2.1 Variance of β

Step 1

$$\hat{\beta}_3 = (\overline{Y}_{A2} - \overline{Y}_{A1}) - (\overline{Y}_{B2} - \overline{Y}_{B1})$$
$$Var(\hat{\beta}_3) = \frac{4\sigma^2}{n}(1+\rho)$$

Step 2 (If $\beta_3 = 0$)

$$\hat{\beta}_1 = \frac{1}{2}\left((\overline{Y}_{B2} - \overline{Y}_{A1}) + (\overline{Y}_{B1} - \overline{Y}_{A2})\right)$$
$$Var(\hat{\beta}_1) = \frac{\sigma^2}{n}(1-\rho)$$
$$\hat{\beta}_2 = \frac{1}{2}\left((\overline{Y}_{B2} - \overline{Y}_{A1}) - (\overline{Y}_{B1} - \overline{Y}_{A2})\right)$$
$$Var(\hat{\beta}_2) = \frac{\sigma^2}{n}(1-\rho)$$
$$\hat{\beta}_0 = \frac{1}{2}\left((\overline{Y}_{A1} - \overline{Y}_{B2}) + (\overline{Y}_{B1} + \overline{Y}_{A2})\right)$$
$$Var(\hat{\beta}_0) = \frac{\sigma^2}{n}$$
$$\hat{\beta}_0 + \hat{\beta}_1 = \frac{1}{2}\left((\overline{Y}_{B1} - \overline{Y}_{A2}) + (\overline{Y}_{A1} + \overline{Y}_{B2})\right)$$
$$Var(\hat{\beta}_0 + \hat{\beta}_1) = \frac{\sigma^2}{n}$$

Step 3 (If $\beta_3 \neq 0$)

$$\hat{\beta}_1 = \overline{Y}_{B1} - \overline{Y}_{A1}$$
$$Var(\hat{\beta}_1) = \frac{2\sigma^2}{n}$$
$$\hat{\beta}_0 = \overline{Y}_{A1}$$
$$Var(\hat{\beta}_0) = \frac{\sigma^2}{n}$$
$$\hat{\beta}_0 + \hat{\beta}_1 = \overline{Y}_{B1}$$
$$Var(\hat{\beta}_0 + \hat{\beta}_1) = \frac{\sigma^2}{n}$$

In step 2 and 3, variances of $\hat{\beta}_0$ and $\hat{\beta}_0 + \hat{\beta}_1$ are the same.

$Var\{\hat{\beta}_3\}$ is at least twice as large as $Var\{\hat{\beta}_2\}$ for $\rho \geq 0$. Therefore, any crossover trial designed to detect the differential treatment effects will have lower power for difference of the carryover effects, which is critical to detect the subsequent analysis and interpretation of the trial will be different. With the presence of a clinically important carryover effect difference, a crossover design is no more efficient than an independent-groups trial.

A two-stage procedure may be used: the difference of carryover effects is tested first with a type I error rate of $10 \sim 20\%$ before moving on to the primary hypothesis testing of the treatment effects. Estimates will be different depending on the conclusion from the first stage.

16.3 Examples

Capecitabine/Erlotinib Followed of Gemcitabine Versus Gemcitabine/Erlotinib Followed of Capecitabine
`http://clinicaltrials.gov/ct2/show/NCT00440167`

This crossover trial is performed in advanced and metastatic pancreatic cancer not previously exposed to chemotherapy. The study compares a standard arm with gemcitabine plus erlotinib to an experimental arm with capecitabine plus erlotinib. It is the first trial of its kind to incorporate second-line treatment into the study design. Patient who fail on first-line therapy are switched to the comparator chemotherapy without erlotinib. The trial therefore not only compares two different regimens of first-line treatment, it also compares two sequential treatment strategies.

Colchicine Randomized Double-Blind Controlled Crossover Study in Behcet's Disease
`http://clinicaltrials.gov/ct2/show/study/NCT00700297`

Method: patients were randomized at the study entry to take either colchicine or placebo. At 4 months, they were crossed over. Those who were taking colchicine went on placebo and those on placebo went on colchicine. Each patient tried therefore, both colchicine and placebo. The primary outcome was the effect of colchicine on the disease activity index, the IBDDAM (16-17). To calculate the overall IBDDAM of the baseline, the IBDDAM of the last 12 months (prior to the study) of each manifestation was calculated and added together. The overall disease activity index was then divided to the number of months (12 months) to have the mean activity index per month. IBDDAM was then measured every 2 months (in the middle and at the end, in each arm of the study). The total IBDDAM of the 4 months was then divided by 4 to have the mean activity index per month. The secondary outcome was to see how the individual symptoms responded to colchicine (IBDDAM of each manifestation).

Statistical analysis: The analysis was done by the intention to treat method. As the difference between IBDDAM before and after treatment had normal distribution Student T test for paired samples were used to evaluate the outcome in the colchicine and the placebo group. As the

Levene's test showed the homogeneity of variance, ANOVA (one way) was used to test the effect of treatment (colchicine and placebo) and gender on patients' outcome. The dependent variable was the difference between IBDDAM (before and after the treatment). The independent variables were the treatment, and the gender. SPSS 15 was used for all statistical calculations.

A Placebo-Controlled, Cross-Over Trial of Aripiprazole
`http://clinicaltrials.gov/ct2/show/record/NCT00351936`

Primary endpoint: Evaluate the effects of aripiprazole on weight, Body Mass Index (BMI), and waist/hip circumference.

This study is a ten-week, placebo-controlled, double-blind, cross-over, randomized trial of the novel antipsychotic agent, aripiprazole, added to 20 obese stable olanzapine-treated patients with schizophrenia or schizoaffective disorder. The advantage of the crossover design is that each subject will act as their own control and fewer subjects will be required.

The double-blind, placebo-controlled, crossover study will consist of two random order 4-week treatment arms (aripiprazole 15 mg or placebo) separated by a 2-week adjuvant treatment washout. Following baseline, subjects will be randomized, double-blind, to either aripiprazole or placebo for 4 weeks. After the initial 4 weeks of medication patients will be reassessed, have a 2-week washout period and then crossover to the other treatment for another 4 weeks.

Data management and statistical analysis will be provided by Dr. David Schoenfeld from the Massachusetts General Hospital, Biostatistics Center.

16.4 Examples

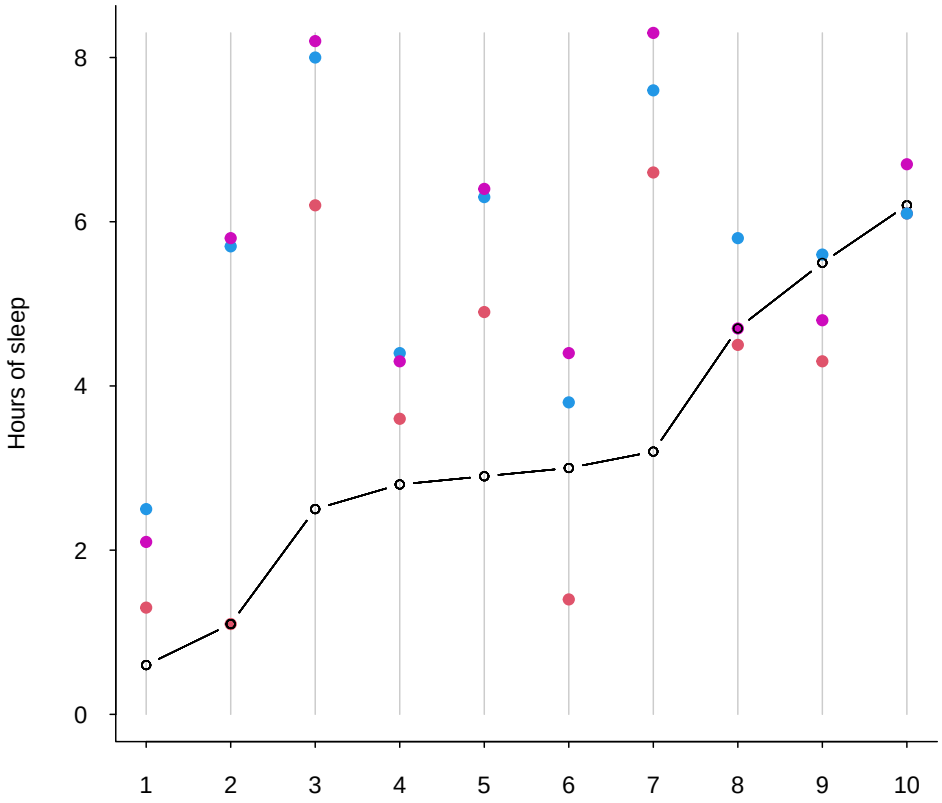
16.4.1 Cushny and Peebles

Cushny AR, Peelbes AR (1904) "The action of optial isomers II: Hyoscines" *J Physiology*. **32**: 501-510.

- Clinical trial of the effect of 3 hypnotic drugs on duration of sleep
  - Study population: inmates of the Michigan Asylum for Insane
  - Patients were given an active treatment on each alternate evening. A typical treatment plan was:
    X C X C X C Y C Y C Y C Z C Z C Z C X Y Z X Y Z X Y Z,
    where 'C' is the control evening where no treatment was given.
- These data were used in "The probable error of a mean" (1908) *Biometrika*. **6**(1): 1-25.

```r
sleep <- cbind(c(0.6, 3, 4.7, 5.5, 6.2, 3.2, 2.5, 2.8, 1.1, 2.9), c(1.3,
    1.4, 4.5, 4.3, 6.1, 6.6, 6.2, 3.6, 1.1, 4.9), c(2.5, 3.8, 5.8, 5.6,
    6.1, 7.6, 8, 4.4, 5.7, 6.3), c(2.1, 4.4, 4.7, 4.8, 6.7, 8.3, 8.2, 4.3,
    5.8, 6.4))
sleep <- data.frame(sleep)
names(sleep) <- c("cont", "X", "Y", "Z")

sleep
```

```
   cont   X   Y   Z
1   0.6 1.3 2.5 2.1
2   3.0 1.4 3.8 4.4
3   4.7 4.5 5.8 4.7
4   5.5 4.3 5.6 4.8
5   6.2 6.1 6.1 6.7
6   3.2 6.6 7.6 8.3
7   2.5 6.2 8.0 8.2
8   2.8 3.6 4.4 4.3
9   1.1 1.1 5.7 5.8
10  2.9 4.9 6.3 6.4
```



Student's paper

```
x <- sleep$X - sleep$cont
mean(x)
```

```
[1] 0.75
```

```
sd(x)
```

```
[1] 1.79
```

```
mean(x)/sd(x)
```

```
[1] 0.419
```

```
t.test(x)
```

```

One Sample t-test

data:  x
t = 1, df = 9, p-value = 0.2
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.53  2.03
sample estimates:
mean of x
     0.75
```

```
y <- sleep$Y - sleep$cont
mean(y)
```

```
[1] 2.33
```

```
sd(y)
```

```
[1] 2
```

```
mean(y)/sd(y)
```

```
[1] 1.16
```

```
t.test(y)
```

```

One Sample t-test

data:  y
t = 4, df = 9, p-value = 0.005
alternative hypothesis: true mean is not equal to 0
```

```
95 percent confidence interval:
 0.898 3.762
sample estimates:
mean of x
     2.33
```

Remarks

- Ethical issues? Potential risks of giving drugs to the mentally ill subjects.
- The primary interest was to study differences between treatments; treatment sequences were of incidental interest.
- Drug X seems to have no effect, and Drugs Y and Z seem to have about the same positive influence in inducing sleep.
- Sequences were not wisely chosen.
- Patients did not receive an equal number of treatments (missing data?)

16.4.2 Hills and Armitage

Hills M, Armitage P (1979) "The two-period cross-over clinical trial". *Br J Clin Pharmacol*. **8**: 7-20.

- Children with enuresis were treated with a new drug or placebo for 14 days
- The primary data are number of dry nights out of 14.

An estimate of within-subject differences (treatment effects) is $\delta = Y_A - Y_B$. The carryover effects may be estimated by

$$Z_1 = \frac{\overline{\delta}_1 - \overline{\delta}_2}{\sqrt{var(\overline{\delta}_1) + var(\overline{\delta}_2)}},$$

and Z is approximately normally distributed under H_0. Similarly the overall treatment effect can be estimated by

$$Z_2 = \frac{\overline{\delta}_1 + \overline{\delta}_2}{\sqrt{var(\overline{\delta}_1) + var(\overline{\delta}_2)}},$$

and this is approximately normal under H_0.

```r
d0 <- c(8, 5, 12, 11, 14, 10, 8, 0, 6, 8, 9, 7, 11, 6, 13, 9, 3, 5, 8,
    8, 6, 0, 8, 9, 0, 0, 4, 8, 8, 14, 13, 12, 2, 4, 10, 2, 7, 5, 8, 13,
    13, 13, 8, 10, 9, 7, 7, 7, 9, 0, 7, 10, 10, 6, 2, 2, 7, 6)
pat <- rep(1:29, each = 2)
period <- rep(1:2, 29)
placebo.first <- c(2, 5, 8, 10, 12, 14, 15, 17, 20, 23, 26, 29)
group <- rep(1, 29)
group[placebo.first] <- 2
trt <- matrix(1:0, nrow = 29, ncol = 2, byrow = TRUE)
trt[placebo.first, 1] <- 0
trt[placebo.first, 2] <- 1

(d <- data.frame(id = pat, group = rep(group, each = 2), period = period,
    trt = c(t(trt)), dry = d0))
```

```
   id group period trt dry
1   1     1      1   1   8
2   1     1      2   0   5
3   2     2      1   0  12
4   2     2      2   1  11
5   3     1      1   1  14
6   3     1      2   0  10
7   4     1      1   1   8
8   4     1      2   0   0
9   5     2      1   0   6
10  5     2      2   1   8
11  6     1      1   1   9
12  6     1      2   0   7
13  7     1      1   1  11
14  7     1      2   0   6
15  8     2      1   0  13
16  8     2      2   1   9
17  9     1      1   1   3
18  9     1      2   0   5
19 10     2      1   0   8
20 10     2      2   1   8
21 11     1      1   1   6
22 11     1      2   0   0
23 12     2      1   0   8
24 12     2      2   1   9
25 13     1      1   1   0
26 13     1      2   0   0
```

```
27 14     2      1   0   4
28 14     2      2   1   8
29 15     2      1   0   8
30 15     2      2   1  14
31 16     1      1   1  13
32 16     1      2   0  12
33 17     2      1   0   2
34 17     2      2   1   4
35 18     1      1   1  10
36 18     1      2   0   2
37 19     1      1   1   7
38 19     1      2   0   5
39 20     2      1   0   8
40 20     2      2   1  13
41 21     1      1   1  13
42 21     1      2   0  13
43 22     1      1   1   8
44 22     1      2   0  10
45 23     2      1   0   9
46 23     2      2   1   7
47 24     1      1   1   7
48 24     1      2   0   7
49 25     1      1   1   9
50 25     1      2   0   0
51 26     2      1   0   7
52 26     2      2   1  10
53 27     1      1   1  10
54 27     1      2   0   6
55 28     1      1   1   2
56 28     1      2   0   2
57 29     2      1   0   7
58 29     2      2   1   6
```

```r
# Group 1: trt -> placebo
g1 <- subset(d, group == 1)
g2 <- subset(d, group == 2)

ms <- function(v) c(mean(v), sd(v)/sqrt(length(v)))

g1.diff <- matrix(g1$dry, ncol = 2, byrow = TRUE)
ms(g1.diff[, 1] - g1.diff[, 2])
```

```
[1] 2.824 0.841
```

```
g2.diff <- matrix(g2$dry, ncol = 2, byrow = TRUE)
ms(g2.diff[, 2] - g2.diff[, 1])
```

```
[1] 1.250 0.863
```

$$z_1 = \frac{2.82 - 1.25}{\sqrt{0.8412^2 + 0.8627^2}} = 1.30$$
$$z_2 = \frac{2.82 + 1.25}{\sqrt{0.8412^2 + 0.8627^2}} = 3.38$$

16.5 A two-period crossover design for the comparison of two active treatments and placebo

By GG Koch, IA Amara, BW Brown, T Colton, and DB Gillings (1989).

Consider sequences of treatments TT, TC, and CT.

1. The first period is parallel group design to address direct use in all patients
2. The second period for TT versus TC is a parallel group comparison design to address T versus C for patients who received T during the first period.
3. The second period for TT versus CT enables "delayed start" assessment of T relative to C if dropout during the first period is minimal and non-informative.
4. The second period for CT versus TC is for assessment of T relative C if carryover effects are small.
5. If T $-$ C from 1, 2, 4 are similar (carryover effects of T to T, T to C, C to T are small), then an overall analysis of treatment effect differences have a very high power.
6. More patients are allocated to receive T within each period.

	P_1	P_2
$S_1 = CT$	β_0	$\beta_0 + \beta_1 + \beta_2$
$S_2 = TC$	$\beta_0 + \beta_1$	$\beta_0 + \beta_2 + \beta_3$
$S_3 = TT$	$\beta_0 + \beta_1$	$\beta_0 + \beta_1 + \beta_2 + \beta_3 + \tau$

$\beta_0 \cdots$ Treatment C effect
$\beta_1 \cdots$ Increment of treatment effect due to T.
$\beta_2 \cdots$ Carryover effect for C
$\beta_3 \cdots$ Increment of carryover effect for T

τ could represent additional treatment effects for longer duration.

Period 1 comparison between T and C is for primary treatment effects, and period 2 comparisons address effects of delayed start (CT vs. TT) and of long-duration effects.

Now consider TT, TC, CT, and CC.

1. This design can estimate all the parameters in the TT, TC, CT case.
2. CC vs. CT enables estimation of treatment effects with run-in period.
3. Relatively unethical to have many patients assigned to receive C.

	P_1	P_2
$S_0 = CC$	β_0	$\beta_0 + \beta_2$
$S_1 = CT$	β_0	$\beta_0 + \beta_1 + \beta_2$
$S_2 = TC$	$\beta_0 + \beta_1$	$\beta_0 + \beta_2 + \beta_3$
$S_3 = TT$	$\beta_0 + \beta_1$	$\beta_0 + \beta_1 + \beta_2 + \beta_3 + \tau$

$\beta_0 \cdots$ Treatment C effect
$\beta_1 \cdots$ Increment of treatment effect due to T.
$\beta_2 \cdots$ Carryover effect for C
$\beta_3 \cdots$ Carryover effect for T

τ could represent additional treatment effects for longer duration.

Example: Pincus T *et al.* (2004) “Patient preference for placebo, acetaminophen (paracetamol) or celecoxib efficacy studies (PACES): two randomised, double blind, placebo controlled, crossover clinical trials in patients with knee or hip osteoarthritis”. *Ann Rheum Dis*. **63**: 931-939.

16.6 Latin squares

When there are k treatments and each patient is to receive all k treatments. Then there are $k!$ possible sequences. Three treatments yield 6 sequences, four treatments yield 24, and five yield 120.

$k = 3$: ABC, ACB, BAC, BCA, CAB, CBA

The idea is to use a reduced number of sequences (reduced sample size) but maintain a good "representation", i.e., every treatment is represented in every period with the same frequency.

	P_1	P_2	P_3
S_1	A	B	C
S_2	B	C	A
S_3	C	A	B

	P_1	P_2	P_3
S_1	A	C	B
S_2	B	A	C
S_3	C	B	A

There are $6!/(3!)(3!) = 20$ ways to choose 3 sequences from 6, but only 2 of those are Latin squares.

16.7 Optimal designs

There is an extensive literature on optimal choice of sequences for measuring treatment effects in the presence of carryover.

- More advanced theory $\cdots$
- Optimality depends on assumptions about carryover effects

Concerns about carryover can be reduced by using designs with more than two periods. (Laska E, Meisner M, Kushner HB. (1983) "Optimal crossover designs in the presence of carryover effects". *Biometrics*. **39**(4): 1087-1091.

Consider treatments A and B in two sequences: $AABB$ and $BBAA$. This design is not uniquely optimal, but it can be used to estimate treatment effects with more efficiency than using data from period 1.

	P_1	P_2	P_3	P_4
AABB	$\mu_{11} = \mu + \pi_1 + \tau_a$	$\mu_{12} = \mu + \pi_2 + \tau_a + \lambda_a$	$\mu_{13} = \mu + \pi_3 + \tau_b + \lambda_a$	$\mu_{14} = \mu + \pi_4 + \tau_b + \lambda_b$
BBAA	$\mu_{21} = \mu + \pi_1 + \tau_b$	$\mu_{22} = \mu + \pi_2 + \tau_b + \lambda_b$	$\mu_{23} = \mu + \pi_3 + \tau_a + \lambda_b$	$\mu_{24} = \mu + \pi_4 + \tau_a + \lambda_a$

Note μ is the overall mean, π is the period effect, τ is the treatment effect, and λ is the carryover effect.

To obtain an unadjusted (for carryover effect) treatment effect ($B - A$), use the following weights.

	P_1	P_2	P_3	P_4
AABB	$-1/4$	$-1/4$	$1/4$	$1/4$
BBAA	$1/4$	$1/4$	$-1/4$	$-1/4$

- Weights sum to 1 for B and -1 for A to form a contrast $B - A$.
- Weights sum to 0 over sequence and period.

$$
\begin{aligned}
&-\frac{1}{4}\mu_{11}-\frac{1}{4}\mu_{12}+\frac{1}{4}\mu_{13}+\frac{1}{4}\mu_{14}+\frac{1}{4}\mu_{21}+\frac{1}{4}\mu_{22}-\frac{1}{4}\mu_{23}-\frac{1}{4}\mu_{24}\\
&=(\tau_b-\tau_a)+(\lambda_b-\lambda_a)/4
\end{aligned}
$$

When carryover effects are present, we can construct weights so that carryover effects will be eliminated.

	P_1	P_2	P_3	P_4
AABB	$-w_1$	$-w_2$	w_3	w_4
BBAA	w_1	w_2	$-w_3$	$-w_4$

Constraints on w's.

- $w_1+w_2+w_3+w_4=1$
- $w_2-w_3+w_4=0$

$$
\begin{aligned}
&-w_1\mu_{11}-w_2\mu_{12}+w_3\mu_{13}+w_4\mu_{14}+w_1\mu_{21}+w_2\mu_{22}-w_3\mu_{23}-w_4\mu_{24}\\
&=-w_1\tau_a-w_2(\tau_a+\lambda_a)+w_3(\tau_b+\lambda_a)+w_4(\tau_b+\lambda_b)+w_1\tau_b+w_2(\tau_b+\lambda_b)-w_3(\tau_a+\lambda_b)-w_4(\tau_a+\lambda_a)\\
&=(w_1+w_2+w_3+w_4)\tau_b-(w_1+w_2+w_3+w_4)\tau_a-(w_2-w_3+w_4)\lambda_a+(w_2-w_3+w_4)\lambda_b\\
&=\tau_b-\tau_a
\end{aligned}
$$

Let σ^2 be the within-patient variance and n be the number of patients per sequence. The variance of the unadjusted estimator is

$$
2\left\{\left(\frac{1}{4}\right)^2+\left(\frac{1}{4}\right)^2+\left(\frac{1}{4}\right)^2+\left(\frac{1}{4}\right)^2\right\}\frac{\sigma^2}{n}=0.5\frac{\sigma^2}{n}.
$$

And for adjusted estimator:

$$
2\left\{w_1^2+w_2^2+w_3^2+w_4^2\right\}\frac{\sigma^2}{n}.
$$

If we pick $w_1=4/10$, $w_2=2/10$, $w_3=3/10$, $w_4=1/10$, we have

$$
2\left\{\left(\frac{4}{10}\right)^2+\left(\frac{2}{10}\right)^2+\left(\frac{3}{10}\right)^2+\left(\frac{1}{10}\right)^2\right\}\frac{\sigma^2}{n}=0.6\frac{\sigma^2}{n}.
$$

If we only use data from the first period

$$
2\left\{1^2+0^2+0^2+0^2\right\}\frac{\sigma^2}{n}=2\frac{\sigma^2}{n}.
$$

The adjusted estimator has slightly higher variance, but it is unbiased with presence of carryover.

William's square
When an even number of treatments are considered in the same number of periods, William's square gives an optimal design. (Williams EJ (1949). "Experimental designs balanced for the estimation of residual effects of treatments". *Australian Journal of Scientific Research. Series A2*. 149-168.) It is a Latin square design in which every treatment precedes every other treatments exactly once.

	P_1	P_2	P_3	P_4
sequence 1	A	B	C	D
sequence 2	B	D	A	C
sequence 3	C	A	D	B
sequence 4	D	C	B	A

Latin square designs are a special type of incomplete block design.

Example:
An experiment was conducted to study the effects of different types of background music on the productivity (Y) of bank tellers. The treatments were defined as five combinations of temp and style of music:
A: slow, instrumental and vocal
B: medium, instrumental and vocal
C: fast, instrumental and vocal
D: medium, instrumental only
E: fast, instrumental only

There are 120 possible sequences of these treatments.

Chapter 17

Meta analysis

17.1 Introduction

Meta analysis is a comprehensive re-analysis of published and unpublished studies, based on obtaining individual patient data or summary statistics, to investigate and quantify consistency or lack of consistency among study results.

It is also referred to as “overview”, “systematic review”, and “research synthesis”.

Some characteristics of meta analysis

- There is rich history in social sciences, relatively new to medical research.
- The number of randomized clinical trials probably approaches 10,000 per year, so synthesizing results can be informative but difficult and confusing.
- When individual patient data are available, a meta analysis will be simpler.
- Cochrane collaboration (`www.cochrane.org`).
- The primary purpose of a meta analysis is analytic rather than descriptive (as opposed to experts’ review of the field).
  - Use formal statistically sound methods to combine treatment effects. (Do not plan an underpowered study just for a meta analysis.)
  - Statistical power of a meta analysis is usually very high, and reliable assessment of secondary endpoints may be possible.

Basic steps

1. Formulation of a purpose (hypothesis) and specification of an outcome.
2. Identification of relevant studies.
3. Establishing inclusion/exclusion criteria of studies.
4. Data abstraction and acquisition.

5. Data analysis.
6. Dissemination of results and conclusions

17.2 Literature search and publication bias

MEDLINE, EMBASE, the Cochrane Controlled Trials Registry, and ClininicalTrials.gov are some of very useful databases. In all cases, well-defined terms such as Medical Subject Heading (MeSH) should be used.

The published literature is not a complete repository of studies actually performed.

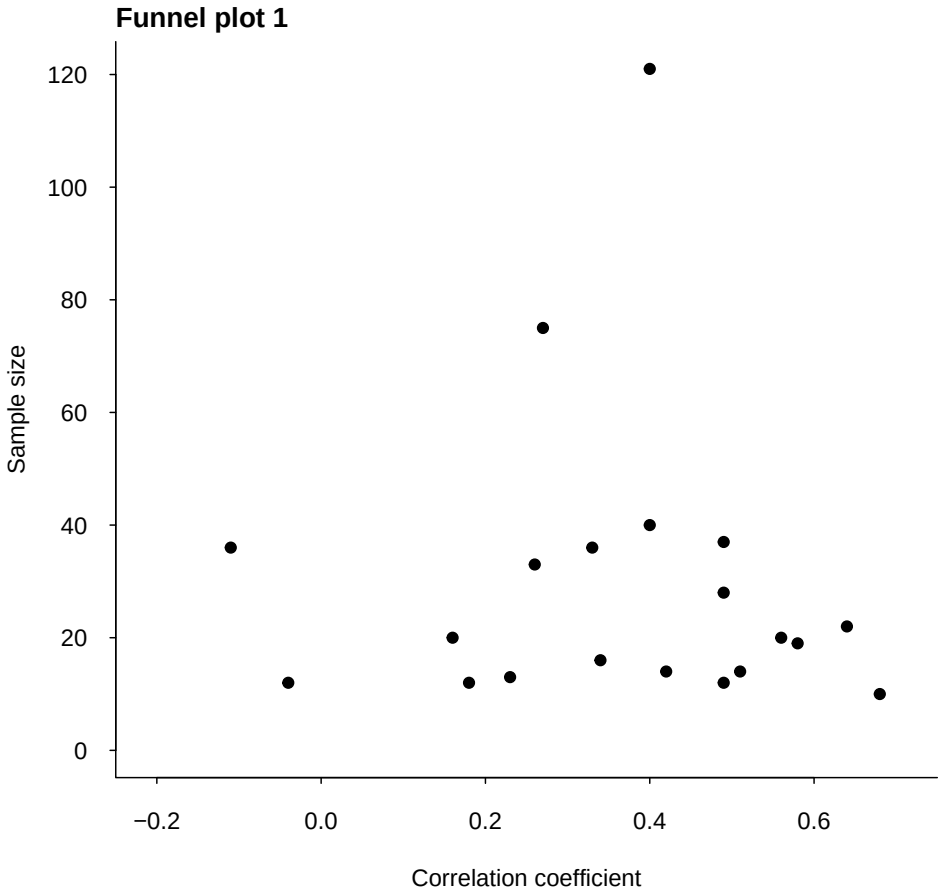
Publication bias A selection bias in the published literature such that publication of research depends on the nature and direction of the study results.

Studies that failed to show positive results are less likely to be published, and combining the information from published studies may be biased in the positive direction.

17.2.1 Funnel plot

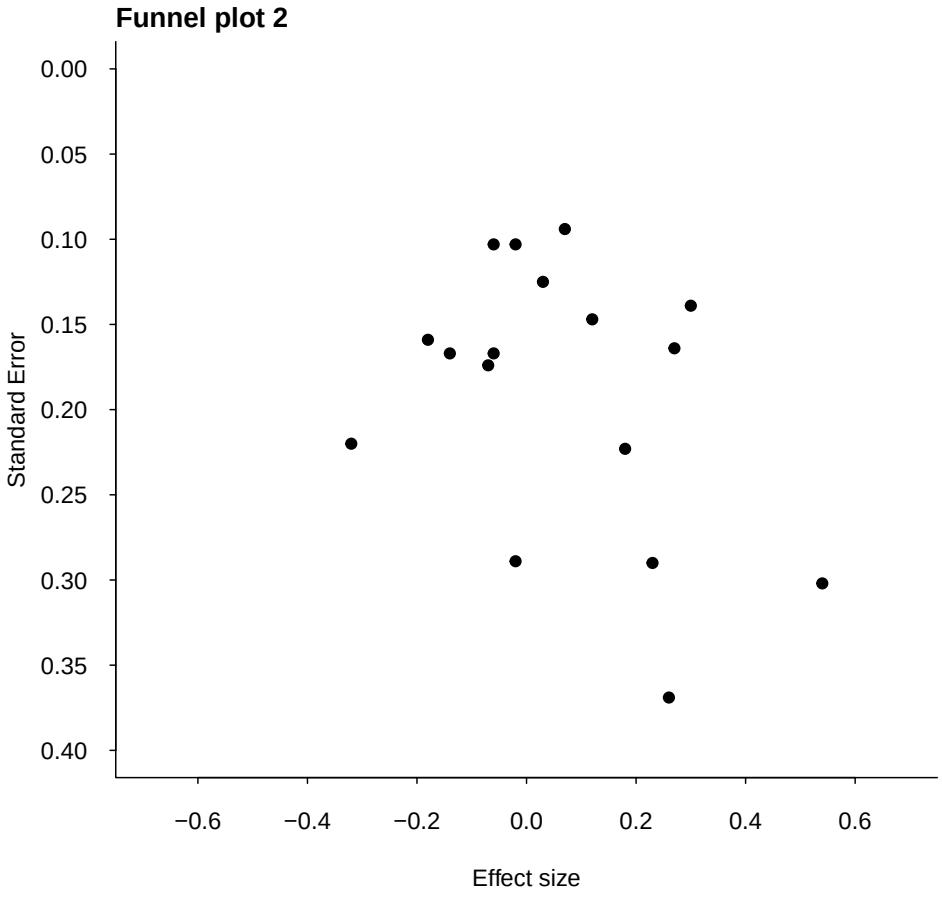
- Sample size tends to be associated with publication bias.
  - Small studies produce highly variable effect size estimates, and if every study is published, there should be an association between reported effect size and sample size.
  - A very large study without statistically significant result may still be published, but a small non-significant study may not be published.
- A funnel plot is one way to visually examine publication bias.

The first funnel plot shows the association between the effect sizes (observed correlation coefficients) and sample sizes.



The data are from a meta analysis of studies of student ratings of their instructor and average achievement (grades) of the students.

- If no bias is present, this plot should be shaped like a funnel, with the spout pointing up. A broad spread of points for the highly variable small studies at the bottom and decreasing spread as the sample size increases.
- The mean effect size should be the same regardless of sample size. One should be able to draw a vertical line through the mean effect size, and the points should be distributed on either side for all sample sizes.
- For this particular example, the range of sample sizes is so narrow that the demonstration of a trend would be unlikely.



The second example is from a meta analysis of the effects of teacher expectancy on pupil IQ.

- The standard error of the effect size is plotted against the effect size.
- Large studies (i.e., small standard error) seem to be clustered around the null.
- Small studies (i.e., large standard error) at the bottom are skewed toward a positive effect.
- A classic pattern of publication bias where small and/or negative studies are not published.

17.2.2 The file-drawer method

Rosenthal R (1979), “The ‘file-drawer problem’ and tolerance for null results”. *Psychological Bulletin*. **86**. 638-641.

Suppose a meta analysis results in a statistically significant effect, using a test based on combining

the z scores of the individual (published) studies. That is, the overall z score,

$$Z = \sum_{i=1}^{k} \frac{Z_i}{\sqrt{k}},$$

is greater than $z_{1-\alpha/2}$.

The goal of the 'file-drawer' method is to determine the number of unpublished studies with an average observed effect of 0 to make the z score greater than $z_{1-\alpha/2}$ so that the test is no longer significant.

We need to find k^* such that

$$Z^* = \sum_{i=1}^{k} \frac{Z_i}{\sqrt{k+k^*}} < z_{1-\alpha/2}.$$

If k^* is sufficiently large for the relevant research domain that it is unlikely that so many unpublished studies exist, then we can conclude that the significance of the observed effects is unchallengeable.

For the data shown in the second funnel plot,

$$\sum_{i=1}^{k} \frac{Z_i}{\sqrt{k}} = \frac{11.02}{\sqrt{19}} = 2.53$$

It is simple to solve for k^* to get

$$k^* > 13.$$

So if there are at least 13 unpublished studies with average effect size at the null value, then the apparent statistical significance of the analysis would be reversed.

While this method is simple to apply and to interpret, a disadvantage is that it relies on the assumption that the results of the missing studies are centered on the null hypothesis.

17.3 Study selection

Once all relevant studies are identified, the researcher needs to establish eligibility criteria for including studies in their meta analysis.

- Relevant studies may have slightly different treatment regimen, various dose levels, various study populations. Oftentimes, studies with broader definition of "treatment" may be included in a meta analysis.

- the Early Breast Cancer Trialists' Collaborative Group (EBCTCG) performed a large and important meta analysis looking at the benefit of chemotherapy in breast cancer, and for some of their analysis, the chemotherapy combination did not have to be identical or given at the same dose. (EBCTCG (1990). "Treatment of early breast cancer, vol I: Worldwide evidence 1985-1990". Oxford University Press)
- Sometimes it is reasonable to exclude studies employing a very small sample size or those with limited follow-up.

17.4 Statistical analysis

Some of the primary issues in statistical analysis are

- Choosing an effect estimate.
  - Standardized mean differences
  - Risk ratios / hazard ratios
  - Correlations
  - p-values
- Deciding on the unit of analysis (trial or individual patient).
- Quality scoring of studies (the newer the better? Often subjective)
- Selecting statistical methods.

17.4.1 Summarizing the data using observed and expected

The "observed minus expected" method can be based on statistics such as number of events, survival rates, or other clinical endpoints.

Suppose that number of events is the primary outcome measure for studies comparing treatment versus control. For the i-th study,
O_i is the random variable (number of successes in treatment)
N_i is the total sample size
n_i is the sample size of the treatment group
K_i is the total number of event of interest
p_i is the overall event rate $p_i = K_i/N_i$.

For the i-th study

Under the null assumption of no difference, the number of observed events in the treatment group, O_i follows a hypergeometric distribution so that

$$P[O_i = x] = \frac{\binom{K_i}{x}\binom{N_i - K_i}{n_i - x}}{\binom{N_i}{n_i}}. \qquad P[O_i = a_i] = \frac{\binom{a_i + c_i}{a_i}\binom{b_i + d_i}{b_i}}{\binom{m_i}{a_i + b_i}}.$$

	Result		
Group	Success	Failure	Total
Treatment	$O_i = x$		n_i
Control			$N_i - n_i$
Total	K_i	$N_i - K_i$	N_i

	Result		
Group	Success	Failure	Total
Treatment	a_i	b_i	$a_i + b_i$
Control	c_i	d_i	$c_i + d_i$
Total			m_i

The expected value is

$$E[O_i] = n_i p_i, \qquad E[O_i] = \frac{(a_i + c_i)(a_i + b_i)}{m_i}$$

and the variance is

$$V[O_i] = n_i p_i (1 - p_i) \frac{N_i - n_i}{N_i - 1}. \qquad V[O_i] = \frac{(a_i + c_i)(b_i + d_i)(a_i + b_i)(c_i + d_i)}{m_i^2 (m_i - 1)}$$

Finally, let $O = \sum_{i=1}^{M} O_i$, $E = \sum_{i=1}^{M} E_i$, and $V = \sum_{i=1}^{M} V_i$. Then the overall test statistic from M studies is

$$Z_{\text{MH}} = (O - E)/\sqrt{V},$$

which has a standard normal distribution under H_0 (independence of columns and rows).

Example 41 (p535 Piantadosi "Clinical Trials")

```
   T.x  T.n C.x  C.n O.minus.E   Var
1   58  615  76  624      -8.5  29.9
2  129  847 185  878     -25.2  64.2
3  244 1620  77  406     -12.7  43.3
4  154 1563 218 1565     -31.9  82.0
5  395 2267 427 2257     -16.9 168.2
6   88  758 110  771     -10.2  43.1
7   39  317  49  309      -5.6  18.9
8  102  813 130  816     -13.8  49.8
9   38  365  57  362      -9.7  20.7
10  65  672 106  668     -20.8  37.3
11   9   40  19   40      -5.0   4.6
```

As an example, the first row is computed as follows:

```
Tx <- 58
Tn <- 615
Cx <- 76
Cn <- 624
```

```
ai <- Tx
bi <- Tn - Tx
ci <- Cx
di <- Cn - Cx
mi <- ai + bi + ci + di

EO <- (ai + ci) * (ai + bi)/mi
ai - EO


[1] -8.51


(VO <- (ai + ci) * (bi + di) * (ai + bi) * (ci + di)/mi^2/(mi - 1))


[1] 29.9
```

$$\begin{aligned} Z_{\text{MH}} &= (O - E)/\sqrt{V} \\ &= -160.30/23.70 \\ &= -6.80 \end{aligned}$$

The Peto-Yusuf method estimates the pooled odds ratio, OR_{p}, as

$$OR_{\text{p}} = \exp[(O - E)/V],$$

and the $(1 - \alpha/2) \times 100\%$ confidence interval is

$$\exp\left(\frac{O - E}{V} \pm \frac{z_{1-\alpha/2}}{\sqrt{V}}\right).$$

```
## Odds ratio
(peto.yusuf.or <- exp(OminusE/V))


[1] 0.752


(peto.yusuf.sd <- 1/sqrt(V))


[1] 0.0422
```

```
## 95% confidence interval ##
exp(OminusE/V + c(-1, 1) * qnorm(0.975) * peto.yusuf.sd)


[1] 0.692 0.817
```

In general, when we have

Group	Success	Failure
Treatment	n_{11}	n_{12}
Control	n_{21}	n_{22}

An estimate of log odds ratio with a continuity correction is

$$\hat{\theta} = \log\left((n_{11}+0.5)(n_{22}+0.5)/(n_{12}+0.5)(n_{21}+0.5)\right),$$

and its variance is

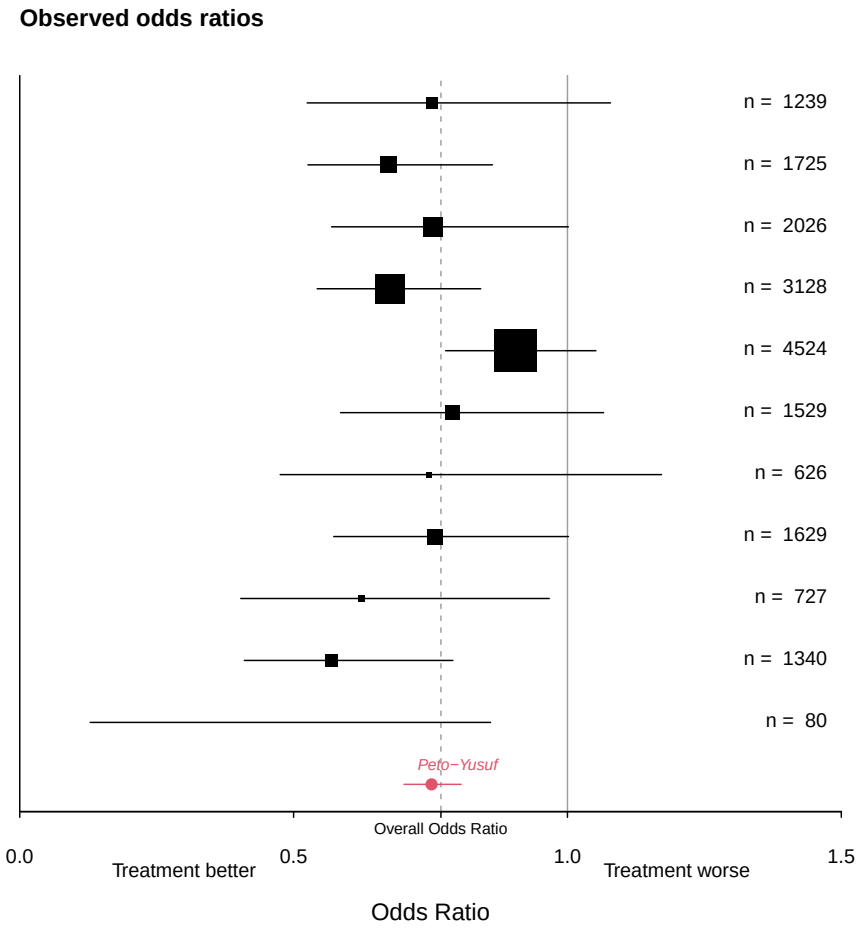
$$\widehat{V}(\hat{\theta}) = \frac{1}{n_{11}+0.5} + \frac{1}{n_{12}+0.5} + \frac{1}{n_{21}+0.5} + \frac{1}{n_{22}+0.5}.$$

Thus, a $(1-\alpha/2)\times 100\%$ confidence interval can be obtained by

$$\exp\left(\hat{\theta} \pm z_{1-\alpha/2}\sqrt{\widehat{V}(\hat{\theta})}\right).$$

```
orf <- function(xt, nt, xc, nc) xt * (nc - xc)/((nt - xt) * xc)  ## odds.ratio
orv <- function(xt, nt, xc, nc) 1/xt + 1/(nt - xt) + 1/(nc - xc) + 1/xc  ## var(odds.ratio)

xt <- d$T.x
xc <- d$C.x
nt <- d$T.n
nc <- d$C.n
est <- orf(xt + 0.5, nt + 1, xc + 0.5, nc + 1)  # with continuity correction
sdt <- sqrt(orv(xt + 0.5, nt + 1, xc + 0.5, nc + 1))  ## sd for log odds ratio
lb <- exp(log(est) - qnorm(0.975) * sdt)
ub <- exp(log(est) + qnorm(0.975) * sdt)
(out <- data.frame(N = d$T.n + d$C.n, est, sdt, lb, ub))


     N   est    sdt    lb    ub
1 1239 0.752 0.1838 0.525 1.079
2 1725 0.674 0.1263 0.526 0.863
3 2026 0.755 0.1441 0.569 1.002
4 3128 0.676 0.1118 0.543 0.842
```

```
5  4524 0.904 0.0771 0.778 1.052
6  1529 0.790 0.1528 0.586 1.066
7   626 0.746 0.2301 0.475 1.172
8  1629 0.758 0.1424 0.573 1.002
9   727 0.625 0.2229 0.404 0.967
10 1340 0.569 0.1676 0.410 0.791
11   80 0.333 0.4846 0.129 0.860


OverallXt <- sum(d$T.x)
OverallXc <- sum(d$C.x)
OverallNt <- sum(d$T.n)
OverallNc <- sum(d$C.n)

## Overall odds ratio
(theta.hat <- orf(OverallXt, OverallNt, OverallXc, OverallNc))


[1] 0.769


peto.yusuf.or


[1] 0.752


(sd.hat <- sqrt(orv(OverallXt, OverallNt, OverallXc, OverallNc)))


[1] 0.0412


peto.yusuf.sd


[1] 0.0422


## 95% confidence interval ##
exp(log(theta.hat) + c(-1, 1) * qnorm(0.975) * sd.hat)


[1] 0.709 0.834
```


Observed odds ratios
n = 1239
n = 1725
n = 2026
n = 3128
n = 4524
n = 1529
n = 626
n = 1629
n = 727
n = 1340
n = 80
Peto−Yusuf
Overall Odds Ratio
0.0
0.5
1.0
1.5
Treatment better
Treatment worse
Odds Ratio

17.4.2 Methods for summarizing significance values

k is the number of studies, and α^* is the significance level of the meta analysis.

Method	Reject H_0 if
Minimum p	$min(p_1, \cdots, p_k) = p_{[1]} < \alpha = 1 - (1 - \alpha^*)^{1/k}$
Sign test	$(r - r')^2/(r + r') > \chi^2_1(\alpha^*)$ where r is the count of p-values < 0.5, and $r' = k - r$.
Mean p	$\sqrt{12k}(0.5 - \sum p_i/k) = \sqrt{12k}(0.5 - \bar{p}) > z(\alpha^*)$
Sum of zs (Stouffer)	$\sum z(p_i)/\sqrt{k} > z(\alpha^*)$
Weighted sum of zs	$\sum w_i z(p_i)/\sqrt{\sum w_i^2} > z(\alpha^*)$
Mean z	$\sum(z(p_i)/k)/(s_z/\sqrt{k}) = \bar{z}/s_{\bar{z}} > t_{k-1}(\alpha^*)$ where s_z is the standard deviation of $z(p_i)$.
Sum of logs (Fisher)	$-\sum 2\log(p_i) > \chi^2_{2k}(\alpha^*)$
Weighted sum of logs	$-\sum w_i \log(p_i) > C(\alpha^*)$ C is a function of w_i.
Logit	$-(k\pi^2(5k + 2)/3(5k + 4))^{-1/2} \sum \log(p_1/(1 - p_i)) > t_{5k+4}(\alpha^*)$
Weighted logit	$-\sum w_i log(p_i/(1 - p_i)) > C_v t_v(\alpha^*)$ v and C_v depend on w_i.