# BIOSTATISTICS I

# CQS SUMMER INSTITUTE

Tatsuki Koyama, PhD

Center for Quantitative Sciences

Vanderbilt University School of Medicine

tatsuki.koyama@vanderbilt.edu

8/7/17 $\sim$ 8/11/17

Updated: 4/16/2020 10:36
R version: 3.6.3

# Contents

# Preliminaries

**Director:** Tatsuki Koyama, PhD

**Schedule:** Monday, August 7th - Friday, August 11th from 1:00 PM - 4:00 PM

**Description:** This course will introduce fundamental concepts and techniques for basic statistical analysis, including types of variables, data summary, hypothesis testing, simple linear regression, and power analysis.

- This course covers basic concepts in statistical data analysis.
- This course is not a tutorial on using a statistical software.

**Learning objectives:** After participating in this activity, participants should be able to:

1. Critically read statistical analysis plans and analysis reports.
2. Identify study design appropriate for research question.
3. Select and perform simple statistical analysis for each study design.

**Software:** We will use **R**, a programming language and software environment for statistical computing and graphics. R can be downloaded from `www.r-project.org`.

# Chapter 1

# Introduction

## 1.1 What is (bio)statistics?

**Biostatistics** The application of statistics to a wide range of topics in biology. (Wikipedia)

**Statistics** The study of the collection, analysis, interpretation, presentation, and organization of data. (The Oxford Dictionary of Statistical Terms)

When we have census data, interpretation of the data is straightforward. (A census is the procedure of systematically acquiring and recording information about the member of a given population.)

In medical research, it is almost always impossible to get the data on everyone in the *population* of interest. So we take a representative *sample* from the population and make inference about the population.

Samples vary from one another. Statistics allows us to make inference about the unknown population *parameters* using the data at hand (samples) and computing *statistics*.

The following definitions are from Freedman et. al[1]

**Population** A whole class of individuals on which we want to make a general statement.

**Parameters** Some numerical facts about the population.

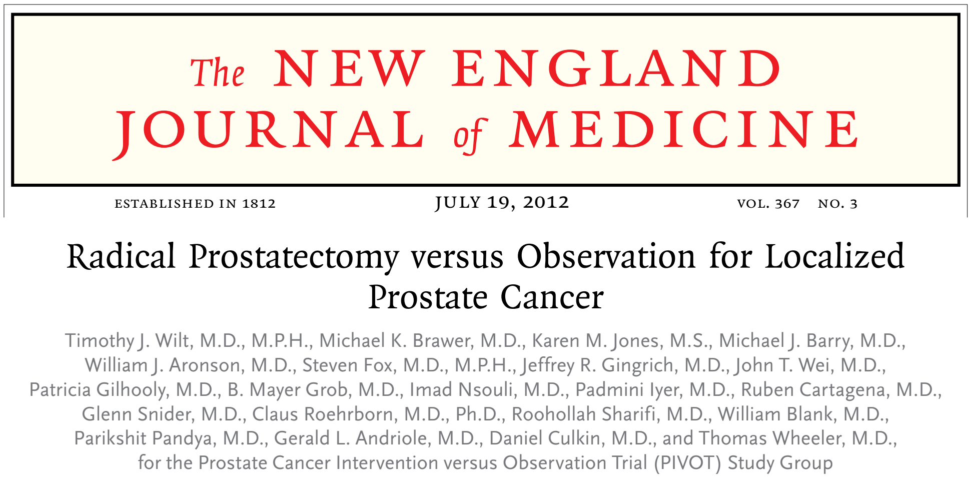
[1] Freedman D, Pisani R, Purves R.(2007): STATISTICS. Fourth Edition. New York. Norton & Company.

**Sample** A part of population that can be examined.

**Statistics** Numbers which can be computed from a sample.

## 1.2 Example

### 1.2.1 Prostate Cancer Intervention Versus Observation Trial (PIVOT)



The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 JULY 19, 2012 VOL. 367 NO. 3

Radical Prostatectomy versus Observation for Localized Prostate Cancer

Timothy J. Wilt, M.D., M.P.H., Michael K. Brawer, M.D., Karen M. Jones, M.S., Michael J. Barry, M.D., William J. Aronson, M.D., Steven Fox, M.D., M.P.H., Jeffrey R. Gingrich, M.D., John T. Wei, M.D., Patricia Gilhooly, M.D., B. Mayer Grob, M.D., Imad Nsouli, M.D., Padmini Iyer, M.D., Ruben Cartagena, M.D., Glenn Snider, M.D., Claus Roehrborn, M.D., Ph.D., Roohollah Sharifi, M.D., William Blank, M.D., Parikshit Pandya, M.D., Gerald L. Andriole, M.D., Daniel Culkin, M.D., and Thomas Wheeler, M.D., for the Prostate Cancer Intervention versus Observation Trial (PIVOT) Study Group

> “From November 1994 through January 2002, we randomly assigned 731 men with localized prostate cancer to radical prostatectomy or observation and followed them through January 2010. The primary outcome was all-cause mortality; the secondary outcome was prostate cancer mortality.”

Inclusion criteria

1. 75 years or younger
2. Localized disease
3. PSA $< 50$mg/mL
4. Diagnosed within 12 months

5. Radical prostatectomy candidate

The following are the population of interest and the sample if we are to make an inference about all prostatectomy patients.

**Population** All prostate cancer patients meeting the inclusion criteria who receive radical prostatectomy.

**Sample** $364$ patients in this study who were assigned to the radical prostatectomy group.

**Parameter** All-cause mortality in the population ($\%$ of the population who die).

**Statistic** All-cause mortality in the sample ($\%$ of $N = 364$ of the sample who die).

Conclusions: $171$ of $364$ $(47.0\%)$ men assigned to radical prostatectomy died.

The number, $47.0\%$, is specific to our sample, and if we had a different sample, the number probably would have been different. Using various techniques in statistics, we sort of understand how the samples (the corresponding statistics) vary. And that understanding allows us to say something about the unknown population parameter, the true all-cause mortality of the all prostate cancer patients meeting the inclusion criteria and received and will receive the radical prostatectomy.

An *estimate* of the all-cause mortality for the prostatectomy patients is $47.0\%$ and its $95\%$ *confidence interval* is $(41.8\%, 52.2\%)$.

“We do not know the true all-cause mortality, but it is probably close to $47.0\%$. Perhaps between $41.8\%$ and $52.2\%$.”

### 1.2.2 Example: Comparative Effectiveness Analysis of Surgery And Radiation (CEASAR)

CEASAR is an *observational study* which recruited men who were diagnosed with prostate cancer from 2011 to 2012.

- CEASAR enrolled $3,691$ men.

- Two of the primary variables are sexual and urinary function scores (UCLA Prostate Cancer Index) based on quality of life survey.
- Quality of life score (QoL) ranges from $0$ to $100$.
- Majority of patients underwent surgery (radical prostatectomy), and other treatment options include radiation and active surveillance.

In this course, we use a small subset of the data as a practice data set. (This is from an ongoing project, and data have been altered.)

```r
getwd()

[1] "/Users/tatsukikoyama/Dropbox/CQS/CqsSummerInstitute/2017Biostats1"

d <- read.csv("R/practiceCeasarData.csv", header = TRUE, as.is = FALSE)
d$Race <- factor(d$Race, levels = c("White", "Black", "Other"))
d$Education <- factor(d$Education, levels = c("High school", "Some college", "College graduate",
    "Graduate school"))
d$Income <- factor(d$Income, levels = c("- 30K", "30K - 50K", "50K - 100K", "100K -"))
d$Gleason <- factor(d$Gleason, levels = c("6 or less", "3 + 4", "4 + 3", "8,9,10"))

names(d)  # Variable names

 [1] "Risk"          "Gleason"       "PSA"           "Age"           "Race"
 [6] "MaritalStatus" "Education"     "Income"        "QoL0"          "QoL6"
[11] "Treatment"     "HeartDisease"  "Hypertension"  "Athma"         "Diabetes"

dim(d)  # Number of rows and columns

[1] 200  15
```

Table 1.1: Descriptive Statistics by Treatment

| | N | Radiation | Surgery | Test Statistic |
|---|---|---|---|---|
| | | $N=70$ | $N=130$ | |
| Age | 200 | 63.0 67.0 73.0 | 56.0 63.5 68.0 | $F_{1,198}$=11.8, P<0.001$^{1}$ |
| Race : White | 200 | 69% (48) | 75% (97) | $\chi^2_2$=1.18, P=0.555$^{2}$ |
| Black | | 10% ( 7) | 10% (13) | |
| Other | | 21% (15) | 15% (20) | |
| MaritalStatus : Not married | 200 | 27% ( 19) | 10% ( 13) | $\chi^2_1$=9.95, P=0.002$^{2}$ |
| Education : High school | 200 | 47% (33) | 25% (33) | $\chi^2_3$=11.6, P=0.009$^{2}$ |
| Some college | | 21% (15) | 23% (30) | |
| College graduate | | 20% (14) | 26% (34) | |
| Graduate school | | 11% ( 8) | 25% (33) | |
| Income : - 30K | 195 | 34% (23) | 15% (19) | $\chi^2_3$=13.2, P=0.004$^{2}$ |
| 30K - 50K | | 24% (16) | 20% (26) | |
| 50K - 100K | | 25% (17) | 31% (40) | |
| 100K - | | 16% (11) | 34% (43) | |
| PSA | 200 | 5.00 6.30 9.05 | 4.70 5.80 7.45 | $F_{1,198}$=2.69, P=0.103$^{1}$ |
| Gleason : 6 or less | 200 | 46% (32) | 47% (61) | $\chi^2_3$=3.84, P=0.279$^{2}$ |
| 3 + 4 | | 44% (31) | 34% (44) | |
| 4 + 3 | | 4% ( 3) | 7% ( 9) | |
| 8,9,10 | | 6% ( 4) | 12% (16) | |
| HeartDisease : Yes | 200 | 9% ( 6) | 15% ( 19) | $\chi^2_1$=1.52, P=0.218$^{2}$ |
| Hypertension : Yes | 200 | 67% (47) | 52% (68) | $\chi^2_1$=4.1, P=0.043$^{2}$ |
| Diabetes : Yes | 200 | 17% ( 12) | 18% ( 24) | $\chi^2_1$=0.05, P=0.817$^{2}$ |
| QoL0 | 200 | 32.6 51.9 68.2 | 41.5 64.2 81.8 | $F_{1,198}$=8.98, P=0.003$^{1}$ |
| QoL6 | 200 | 19.6 48.0 73.9 | 10.8 24.7 59.2 | $F_{1,198}$=5.82, P=0.017$^{1}$ |

$a$ $b$ $c$ 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: $^{1}$Wilcoxon test; $^{2}$Pearson test

# Chapter 2

# Basic Concepts

## 2.1 Types of data

**Response variable ($Y$)** (Outcome variable, Dependent variable) Clinical endpoint and lab measurements that represent an effect.

**Explanatory variable ($X$)** (Exposure variable, Independent variable) Something that may be associated with the response variable and of major interest.

**Confounding variable (Confounder)** (Adjustment variable, Effect modifier) A variable not of major interest but may be associated with response and/or independent variables.

A few example:

- Survival $\sim$ Breast Cancer Treatment
  Possible confounders: age, BMI, smoking history.
- Prostate Cancer Treatment $\sim$ Participatory decision making score
  Possible confounders: tumor stage, age, comorbidities.

### 2.1.1 Types of measurements

- Binary (Dichotomous): A variable with $2$ possible categories.

- Categorical (Nominal): More than $2$ categories that are not naturally ordered. e.g., Race, Location.
- Ordinal: A categorical variable with natural order. e.g., Income bracket
- Count: An ordinal variable with no upper limit.
- Continuous: A numeric variable having many possible values.

What type of variables are these?

1. Age: Younger than $65$, $65$ to $75$, older than $75$.
2. Education level.
3. Number of ER visits.
4. Blood pressure.
5. Tumor response: Complete response, Partial response, Stable disease, Progressive disease.
6. Number of stained cells.
7. Age: Younger than $65$, $65$ and older.
8. Age.

A continuous variable can always be categorized or dichotomized, but doing so is *never* a good idea. There is loss of information, and a larger *sample size* is required to yield the same statistical information (precision or *power*).

**Random variable** : A variable whose possible values are numerical outcomes of a random phenomenon. It is usually denoted by a capital alphabet $X$ and $Y$. And its values are usually denoted by $x$ and $y$.

Example:

- $X$ is systolic blood pressure of patients in a study. $x = 122, 124, 141, \cdots$

- $X$ is race of patients. $x =$ white, black, Asian, $\cdots$.

## 2.2 Distribution

The *distribution* of a random variable $X$ is a profile of its variability and other tendencies. A distribution is characterized by:

- Binary variable: the probability (proportion) of “yes” (or one of the categories).

- Categorical variable (more than 2 levels): the probability (proportion) of each category.

- Continuous variable: cumulative probability distribution. $P[X \leq x]$ for all values of $x$. (There are many other ways...) Sometimes, the *mean* and *standard deviation* are sufficient.

## 2.3 Descriptive statistics

- Categorical variables (including binary and ordinal variables) can be described by the proportion of each category.
  Because proportions add up to $1$ $(100\%)$, we only need to report $K-1$ proportions for a variable with $K$ categories.

```
table(d$Race)   # Distribution of ''Race''


White Black Other
  145    20    35

prop.table(table(d$Race))   # Proportions of each category


White Black Other
0.725 0.100 0.175

table(d$Education)
```

```
     High school     Some college College graduate  Graduate school 
              66               45               48               41 

prop.table(table(d$Education))


     High school     Some college College graduate  Graduate school 
           0.330            0.225            0.240            0.205 
```

- To represent (summarize) a continuous variable, a measure of central tendency and a measure of variation are often used.
  - Measure of central tendency

    Mean, median, geometric mean
  - Measure of variation

    Standard deviation, inter-quartile range (IQR)

### 2.3.1 Mean, median, geometric mean

Mean (aka “average”, arithmetic mean)

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- Add all the numbers and divide by the number of items.
- Works well in general but not always.

  Grossly influenced by *outliers*.

  Grossly influenced by *skewness*.

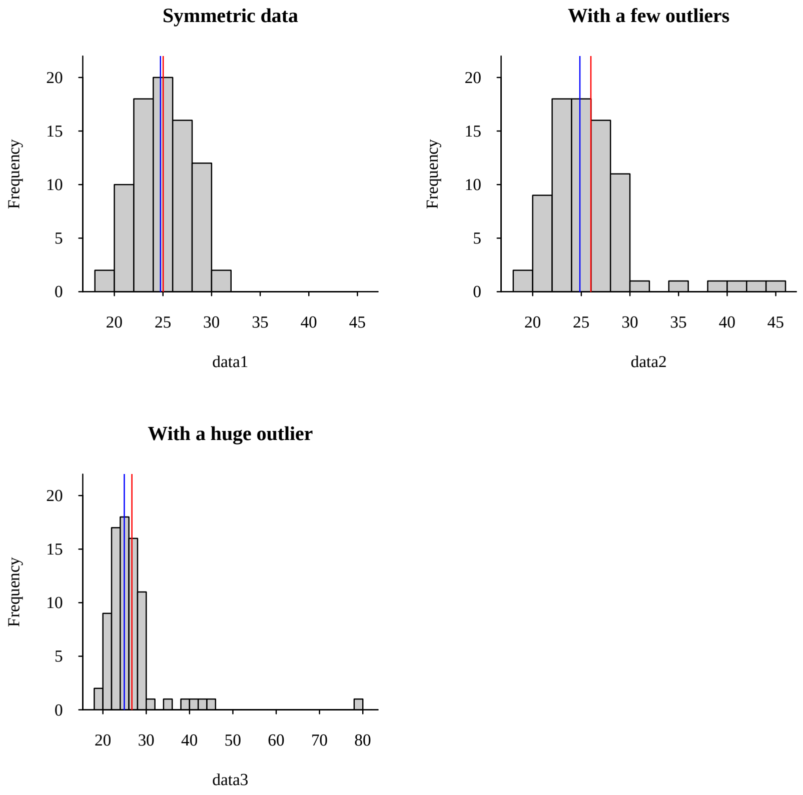
Median: The middle value.

- Works well with outliers and skewness.

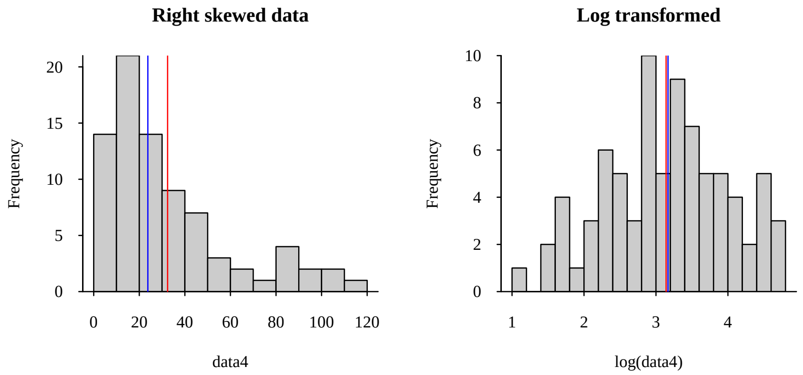
Geometric mean: Multiply all the numbers and take the $n^{th}$ root.

- Works only for positive numbers.

- Works well with skewed data with outliers.
- Used often without being mentioned because it is antilog of the mean of log-transformed data.



| | Mean | Median |
|---|---|---|
| Symmetric data | 25.0 | 24.7 |
| With a few outliers | 26.0 | 24.9 |
| With a huge outlier | 26.7 | 24.9 |



| | Mean | Median | Geometric mean |
|---|---|---|---|
| Right skewed data | 32.4 | 23.8 | 23.1 |
| Log transformed | 3.1 | 3.2 | |

```
x <- rnorm(n = 40, mean = 100, sd = 15)
## Generate 40 random numbers from Normal(100, 15) distribution.

mean(x)  ## mean

[1] 101.3

median(x)  ## median

[1] 101.17

prod(x)^(1/length(x))  ## geometric mean

[1] 100.38
```

```
mean(log(x))  ## mean of log-transformed data

[1] 4.609

exp(mean(log(x)))  ## anti-log of mean of log-transformed data

[1] 100.38
```

### 2.3.2 Variance, standard deviation, standard error, IQR

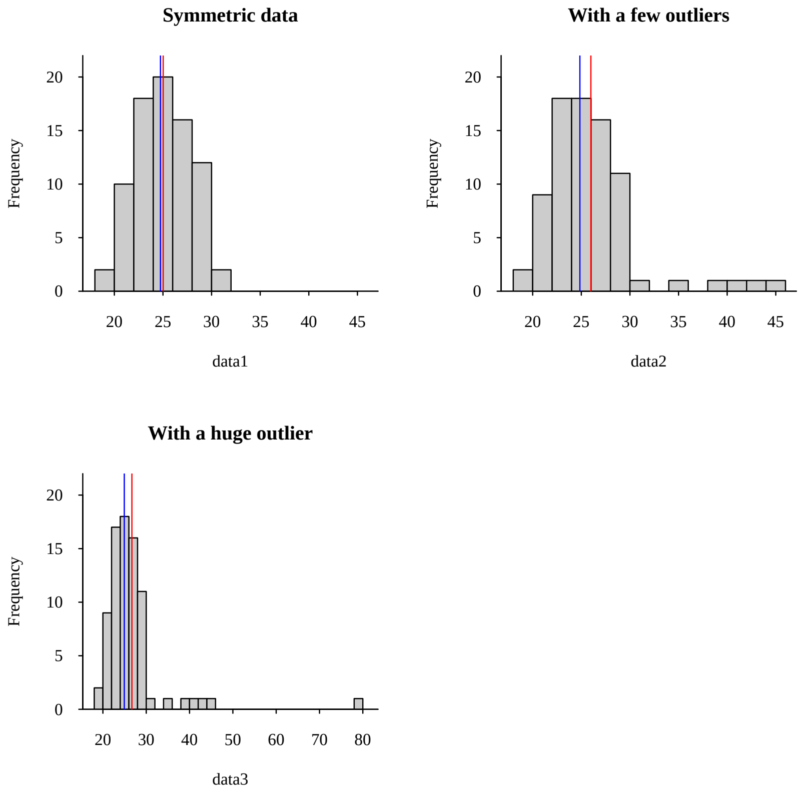
Variance

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

- Variance is average of squared-distance from the mean.
- Denominator is $n-1$ instead of $n$ to make the value a little bigger to compensate for the fact we are estimating the mean. (i.e., more uncertainty)
- The more spread the data the bigger the variance.

Standard deviation $(s)$ is the square root of variance.

- Standard deviation and the mean have the same units as the original data.
- Variance and standard deviation are always positive.
- Variance and standard deviation are $0$ when all the data are the same (constant).
- Variance and standard deviation are grossly influenced by outliers.
- Variance and standard deviation are only useful for symmetric data.



| | Mean | Median | Variance | Standard deviation |
|---|---|---|---|---|
| Symmetric data | 25.0 | 24.7 | 7.2 | 2.7 |
| With a few outliers | 26.0 | 24.9 | 22.9 | 4.8 |
| With a huge outlier | 26.7 | 24.9 | 59.2 | 7.7 |

Standard error ... Later!

IQR

- Quantiles (percentiles): $q$-th sample quantile is the value such that $q\%$ of the data fall below. For example, $20\%$ of the data fall below the $20^{th}$ percentile.
  - $50^{th}$ percentile is the median ($Q_2$).
  - $25^{th}$ quantile is the lower quartile ($Q_1$).
  - $75^{th}$ quantile is the upper quartile ($Q_3$).

- Interquartile range (IQR) is the *difference* of $Q_3$ and $Q_1$. (or $Q_1$ to $Q_3$).
    - IQR is the range such that the middle $50\%$ of the data fall.
    - It is meaningful for any continuous data (skewed, outliers), perhaps except when there are a large number of ties.

| | Mean | Median | Variance | sd | $Q_1$ | $Q_2$ | $Q_3$ |
|---|---|---|---|---|---|---|---|
| Symmetric data | 25.0 | 24.7 | 7.2 | 2.7 | 23.2 | 24.7 | 27.0 |
| With a few outliers | 26.0 | 24.9 | 22.9 | 4.8 | 23.2 | 24.9 | 27.2 |
| With a huge outlier | 26.7 | 24.9 | 59.2 | 7.7 | 23.3 | 24.9 | 27.2 |

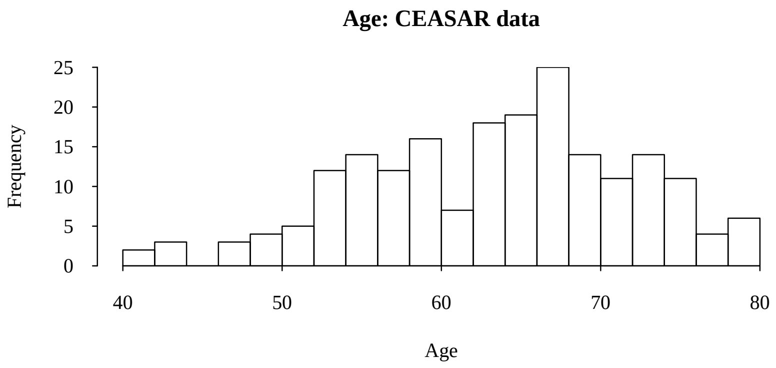
### 2.3.3 CEASAR: summarizing continuous variables

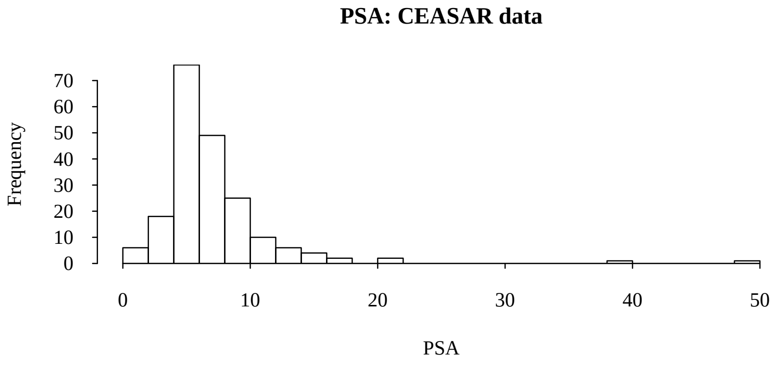
Let's look at age of the patients in the CEASAR data.

```
range(d$Age)

[1] 42 79

mean(d$Age)

[1] 63.92

median(d$Age)

[1] 65

var(d$Age)

[1] 70.92

sd(d$Age)

[1] 8.422

quantile(d$Age, c(0.25, 0.5, 0.75))

25% 50% 75%
 58  65  70
```

And PSA.

```
range(d$PSA)

[1]  0.9 48.2

mean(d$PSA)

[1] 7.099

median(d$PSA)

[1] 6.05

var(d$PSA)

[1] 24.46

sd(d$PSA)

[1] 4.946

quantile(d$PSA, c(0.25, 0.5, 0.75))

 25%  50%  75%
4.80 6.05 8.10
```





Summary

- For symmetric data, mean and standard deviation are good summaries of the distribution.
- For any data, median and quartiles are good summaries of the distribution.

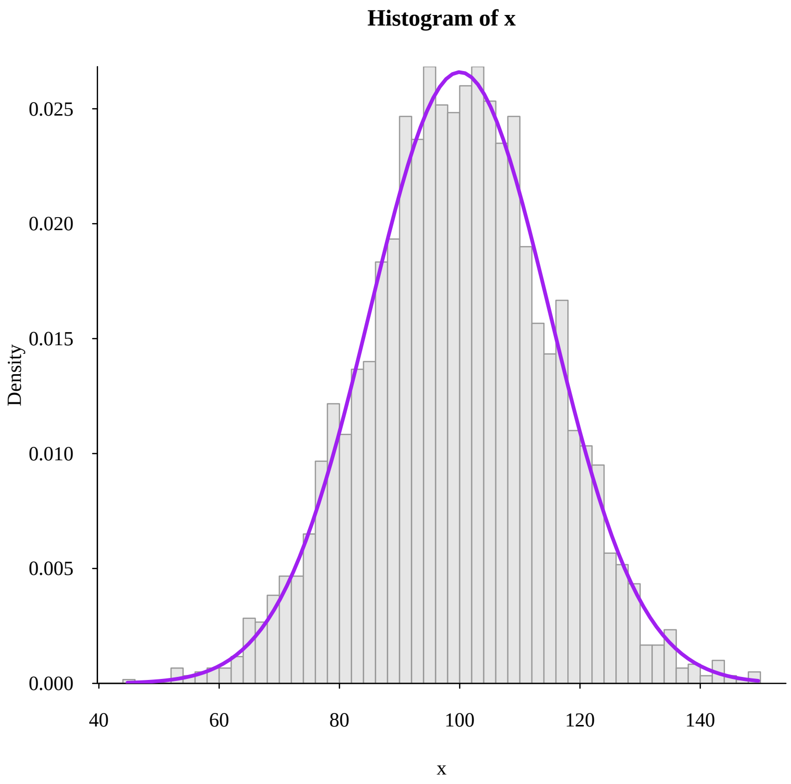
  So why do we even compute the mean and standard deviation?

  They are *very* useful when they are useful! (More about this later: Normal, Central Limit Theorem)

# Chapter 3

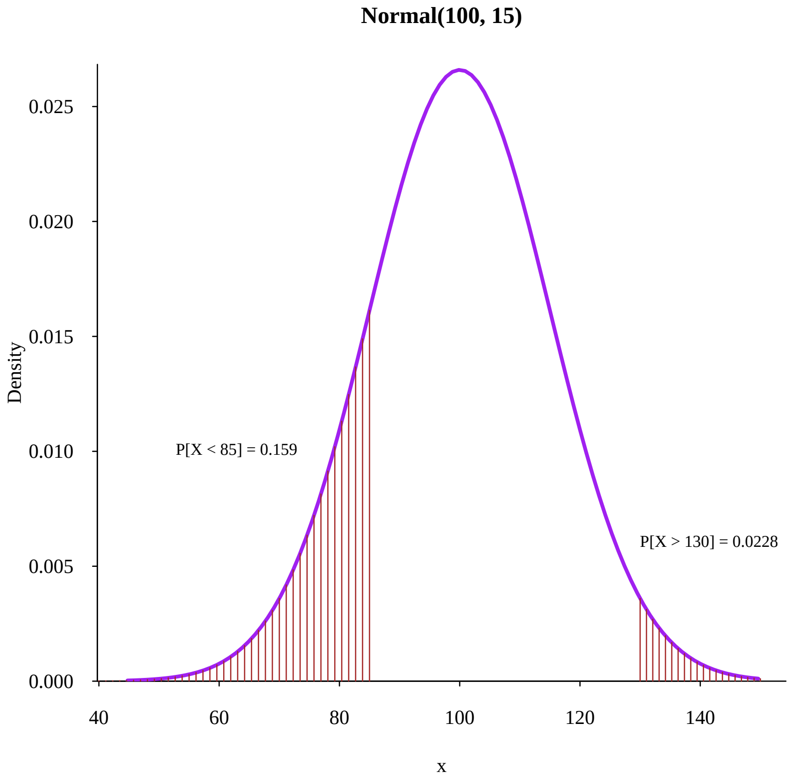
# Normal Distribution

Many random variables (are assumed to) have a Normal distribution. A normal distribution is symmetric and is bell-shaped.



This is a Normal distribution with mean$= 100$ and sd$= 15$. $X \sim Normal(100, 15)$ or $X \sim Normal(100, 15^2)$.

- A normal distribution can be uniquely specified by the mean and standard deviation.
- $50\%$ of the data are above (below) the mean.
- About $2.5\%$ of the data are above $2$ standard deviations above (below) the mean.
- About $15\%$ of the data are above $1$ standard deviation above (below) the mean.



```
pnorm(85, mean = 100, sd = 15, lower.tail = TRUE)

[1] 0.1587

pnorm(130, mean = 100, sd = 15, lower.tail = FALSE)

[1] 0.02275
```

```
1 - pnorm(130, 100, 15)

[1] 0.02275
```

Normal distribution is very useful because ...

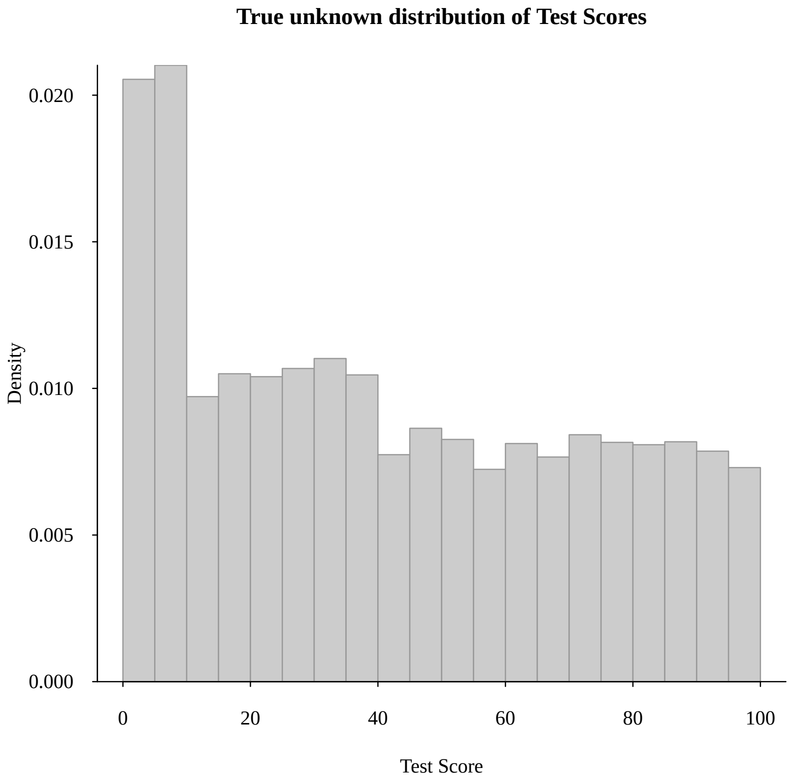
- when the data come from a normal distribution the distribution of *sample means* is normal.
- the distribution of *sample means* is normal even when the data do not come from a normal distribution.

  There are exceptions.

  True only when *sample size* is large.

A (made-up) example:

Suppose that a psychological test score has the true (unknown) distribution shown below.



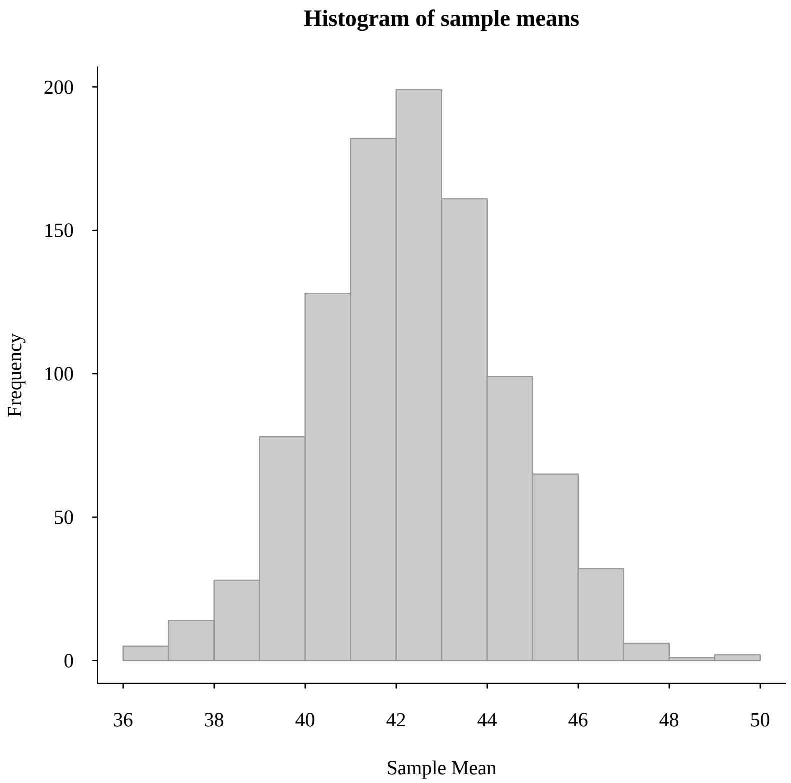
| | mean | sd | Q1 | median | Q3 |
|---|---|---|---|---|---|
| Test Score | 42.35 | 30.17 | 14.26 | 37.88 | 68.66 |

Because the data are not normally distributed, it would be *incorrect* to say things like “$2.275\%$ of the data are $2$ standard deviations above the mean.” (This is only true for normal distributions.) It is still correct to say $25\%$ of the data are above the third quartile $(68.66)$. (This is always true.)

Now suppose that we took a sample of size $n = 225$ from this distribution and compute the sample mean. And it was $41.5$. If we take a different sample of size $n = 225$, the new sample mean will be different from $41.5$. In reality, we only take one sample, but here, let’s say we keep taking samples of size $n = 225$ and keep computing the sample means. And the first 10 are:

```
sampleMeans[1:10]

 [1] 41.45 43.78 42.70 38.99 41.35 42.06 42.92 39.93 40.69 44.04
```

The distribution of $1,000$ sample means looks like:



| | mean | sd | Q1 | median | Q3 |
|---|---|---|---|---|---|
| Test Score | 42.35 | 30.17 | 14.26 | 37.88 | 68.66 |
| Sample Average | 42.35 | 2.04 | 40.99 | 42.30 | 43.70 |

If the data come from a distribution with mean $\mu$ and sd $\sigma$, the sample means have $Normal(\mu, \sigma/\sqrt{n})$ distribution, where $n$ is the sample size. This is the *Central Limit Theorem*.

$30.17/\sqrt{225} = 2.01$.

- When making an inference about the population mean, we can assume that its distribution follow $Normal(\mu, \sigma/\sqrt{n})$ regardless of the underlying (true) distribution of the data.
- The quantity $\sigma/\sqrt{n}$ is *standard error* $=$ standard deviation of the sample mean.
- The mean of a skewed data may not be that interesting because it is not a good measure of central tendency.

### Standard deviation or standard error?

- A distribution can be summarized with “mean & standard deviation” or “mean & standard error”.
- With standard deviation, you can compute “mean $+$ sd” and say, “Ok about $15\%$ of the data are above this number.
- With standard error, you can compute “mean $+$ se” and say, “...” You can’t say much. “mean $+$ se” is an important number when making an inference (later), but as a summary of a distribution, there isn’t much use.
- “median & quartile” is a better combination to report, anyway.

So far, we have always assumed that the true distribution to be known, but obviously, it is not the case in reality. What we want to do is to make an inference about the unknown population using the information from samples. The 2 key components in statistical inference are estimation and hypothesis testing. Perhaps the former is a little bit more important, but we’ll talk about hypothesis testing because it’s easier to explain!

# Chapter 4

# Hypothesis Testing

## 4.1 Fundamentals

**Hypothesis** Usually a statement about the population parameters (such as the population mean, difference of the population means, and the population proportions). Note that the population parameters are the unknown truth.

- $\mu = 100$ (The population mean is $100$.)
- $\pi = 0.2$ (The population proportion is $20\%$.)
- $\mu_1 - \mu_0 = 0$ (The mean of population $1$ is the same as the mean of population $0$.)

It can be about population distributions, but that is rare. (e.g., The true distribution is Normal.)

**Null hypothesis** is the statement you hope to reject/dismiss. ($H_0$)

- "Probability of success is $20\%$." when you want to say that the probability of success is *greater than* $20\%$.
- "There is no difference in the group means." when you want to say that the true means are different.

**Alternative hypothesis** is the statement you want to use as a conclusion. ($H_1$ or $H_a$)

- $H_0 : \pi = 0.2$

  $H_1 : \pi > 0.2$

  This is an example of one-sided alternative.

- $H_0 : \mu_1 = \mu_2$

  $H_a : \mu_1 \neq \mu_2$

  This is an example of two-sided alternative.

How we think when we conduct a hypothesis testing.

1. Compute the probability of acquiring the data we actually acquired assuming that $H_0$ is true. Strictly speaking, "acquiring the data we actually acquired *or something more extreme*.
2. If that probability (p-value) is small, we say that something is wrong...
3. The data we have cannot be wrong, so what's wrong must be our assumption (i.e., $H_0$).

**P-value** is the probability of observing the data actually observed or something more extreme under $H_0$.

- Note that a p-value can be computed without ever referring to the alternative hypothesis.
- We can just compute the p-value, but a lot of times, we are required to make a go/no-go decision. So using the data, we decide to either "reject $H_0$" or "not reject $H_0$".
- The null hypothesis is a statement about (true but unknown) population parameter, and it can be either true or false. $H_1$ is a complement of $H_0$, and it can be either true or false.
- Sometimes rejecting $H_0$ is correct, and sometimes it is not. The following table summarizes what happens when we reject or fail to reject $H_0$.

|  | Truth |  |
|---|---|---|
| Conclusion | $H_0$ is true. | $H_0$ is not true. |
| Reject $H_0$ | Type I error | Correct |
| Fail to reject $H_0$ | Correct | Type II error |

**Type I error** is an error of rejecting a true $H_0$.

We use $\alpha$ to denote the probability of such an error.

**Type II error** is an error of failing to reject a false $H_0$.

We use $\beta$ to denote the probability of such an error.

**Power** is $1-\beta$: probability of correctly rejecting a false $H_0$.

- Customarily, we set $\alpha$ to $5\%$.
- Note that we can reject $H_0$ when the p-value is less than $5\%$. That is, if the probability of observing what we observed (or something more extreme) is less than $5\%$ that is an evidence against $H_0$.
- **Warning**: $H_0$ can never be shown to be true (believable), i.e., even a p-value of $95\%$ does not allow us to say "$H_0$ is shown to be true". Not even "$H_0$ seems to be true/believable/credible."

## 4.2 SPADI example

After the rotator cuff repair surgery using a new technique, the Shoulder Pain and Disability Index (SPADI) is measured on each patient. We would like to test the average SPADI is higher than $72$, which is the known average for the conventional surgical technique. We also know (or assume) that the true standard deviation is $8$.

Let $\mu$ be the true (but unknown) mean SPADI for the new technique. We'd like to test

$H_0 : \mu = 72$

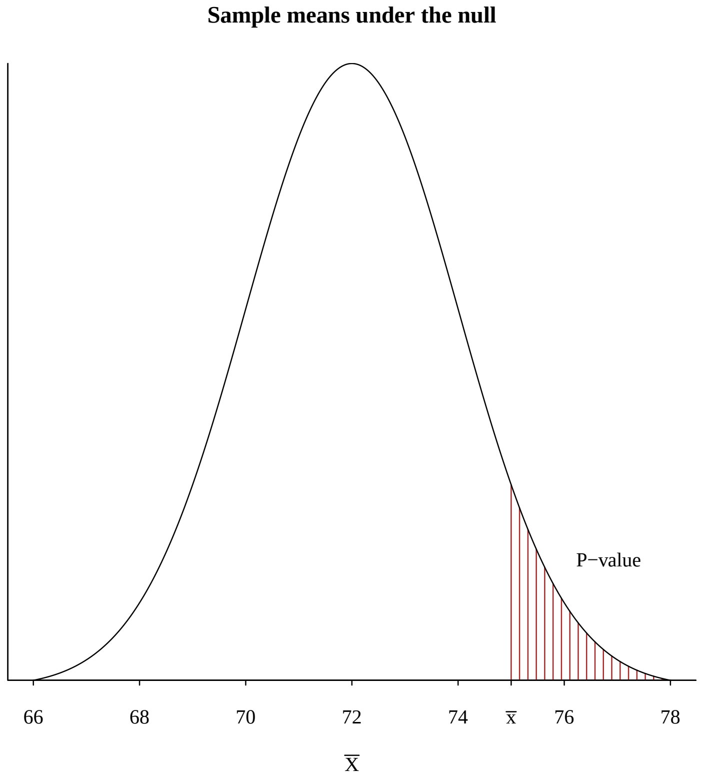
$H_1 : \mu > 72$

Note: Perhaps it is more appropriate to write $H_0 : \mu \leq 72$. Either is acceptable for a one-sided alternative hypothesis.

To test these hypotheses, a random sample of size $16$ was taken from the population of patients. We know that the sample average $\overline{X}$ has the null distribution,

$$\overline{X} \sim Normal\left(72, \frac{8}{\sqrt{16}}\right).$$

This is the distribution of $\overline{X}$ assuming that the null hypothesis is true. Suppose that the observed sample mean was $75$.



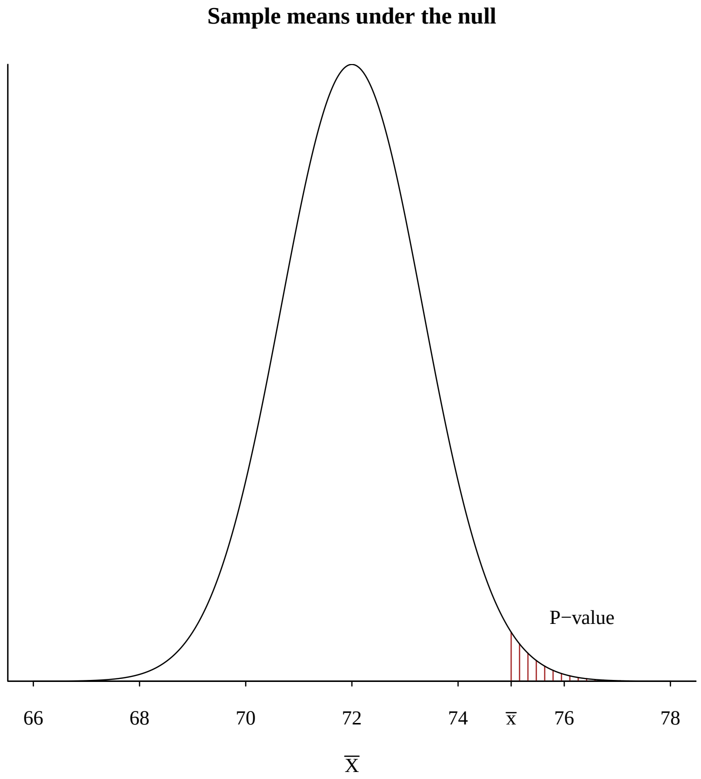
```
## p-value ##
pnorm(75, mean = 72, sd = 8/sqrt(16), lower.tail = FALSE)

[1] 0.06681
```

P-value is $0.067$, and it is not smaller than $5\%$, so we do not have strong enough evidence to conclude that the true average SPADI for the new technique is higher than $72$.

- Can we say the true average is $72$? -No.
- Can we say the true average is lower than $72$? -No.
- So what can we say? -Nothing.

When we do not reject $H_0$, we cannot conclude anything other than “The sample size was too small.” or “We didn’t do the experiment right.” More on this later.

Now suppose that the sample size is $36$. The sample average was still $75$.



```
## p-value ##
pnorm(75, mean = 72, sd = 8/sqrt(36), lower.tail = FALSE)

[1] 0.01222
```

With these data, we reject $H_0$. We have enough evidence to claim that the true mean is greater than $72$ (with type I error rate of $5\%$.).

Statistical hypothesis testing

Q: Is the true mean less than $72$?

A: *If* the true mean is less than $72$, then the probability of observing what we observed is very small $(0.012)$.

Q: So are you saying that the true mean is less than $72$?

A: No.

So what is the true mean? That is a more interesting research question than "Is the true mean greater than $72$?".

$\Rightarrow$ Estimation

## 4.3 Multiplicity

- When type I error rate is controlled at $5\%$, we conduct a hypothesis test with a $5\%$ probability of making an erroneous conclusion (reject $H_0$ that is true).
- If we conduct more than one hypothesis test, the probability of making *at least one* erroneous conclusion becomes more than $5\%$.

When we test $K$ (independent) hypotheses, probability of rejecting at least one true null hypothesis is

$$\text{P[At least one type I error]} = 1-(1-0.05)^K$$

| $K$ | 1 | 2 | 3 | 4 | 10 | 20 | 50 |
|---:|---|---|---|---|---|---|---|
| P[At least one type I error] | $5\%$ | $9.75\%$ | $14.26\%$ | $18.55\%$ | $40.13\%$ | $64.15\%$ | $92.31\%$ |
| To make P[family-wise error] $= 5\%$ | $5\%$ | $2.53\%$ | $1.70\%$ | $1.27\%$ | $0.51\%$ | $0.26\%$ | $0.01\%$ |

So if we want to make P[at least one error] controlled at $5\%$, we need to make each test more stringent. Solve

$$0.05 = 1 - (1 - \alpha)^K$$

for $\alpha$. This is controlling *family-wise* type I error rate, and it is a general but conservative multiplicity control.

**Analysis of Variance** A generalization of two group comparison to comparisons of more than $2$ groups. Sometimes, all pair-wise comparisons are of interest; however, oftentimes, you can pre-specify comparisons of interest.

**Kruskal Wallis test** The non-parametric counterpart to ANOVA.

**Multiple comparison procedures**

**Bonferroni** Very general; applicable to many situations but conservative (not very powerful). Holm's procedure is uniformly better than Bonferroni.

**Dunnett** Applicable when multiple groups are compared to a common control group.

**Tukey-Kramer** Applicable when all pair-wise comparisons are sought.

**Scheffé** Applicable with general contrasts (e.g., $\mu_1 - (\mu_2 + \mu_3)/2 = 0$, i.e., "The mean of groups $2$ and $3$ is equal to the mean of group $1$.")

- If only pair-wise comparisons are of interest, Tukey-Kramer is preferred.

**Holm** Applicable in general.

- Suppose that there are $K$ comparisons of interest. Compute p-value for each comparison.
- Compare the smallest p-value to $\alpha/K$. If the p-value is smaller, reject the corresponding $H_0$ and continue. Otherwise end.
- Compare the second smallest p-value to $\alpha/(K-1)$. Continue in the same manner as long as $H_0$'s are getting rejected.

    - The largest p-value is compared $\alpha$ if all other p-values are smaller than the respective threshold.

**FDR** A different concept. Popular method when testing many, many hypotheses.

    - Benjamini-Hochberg
    - Benjamini-Hochberg-Yekutieli

- Bonferroni, Tukey-Kramer, and Scheffé are used as post-hoc tests for ANOVA, which tests an *overall* hypothesis. $H_0 : \mu_1 = \mu_2 = \cdots = \mu_K$ (All the means are the same.); $H_1$ : At least one mean is different.
- If the overall test is not significant, no group-wise comparisons are granted.

```
p.values <- c(0.002, 0.58, 0.0015, 0.077, 0.002, 0.14, 0.0014, 0.33)
pv.ordered <- sort(p.values)
Bonf <- p.adjust(pv.ordered, method = "bonferroni")
Holm <- p.adjust(pv.ordered, method = "holm")
Bh <- p.adjust(pv.ordered, method = "BH")
Bhy <- p.adjust(pv.ordered, method = "BY")

data.frame(pv = pv.ordered, Bonf, Holm, Bh, Bhy)

      pv   Bonf   Holm     Bh     Bhy
1 0.0014 0.0112 0.0112 0.0040 0.01087
2 0.0015 0.0120 0.0112 0.0040 0.01087
3 0.0020 0.0160 0.0120 0.0040 0.01087
4 0.0020 0.0160 0.0120 0.0040 0.01087
5 0.0770 0.6160 0.3080 0.1232 0.33484
6 0.1400 1.0000 0.4200 0.1867 0.50733
7 0.3300 1.0000 0.6600 0.3771 1.00000
8 0.5800 1.0000 0.6600 0.5800 1.00000
```

# Chapter 5

# Estimation

## 5.1 Point estimates

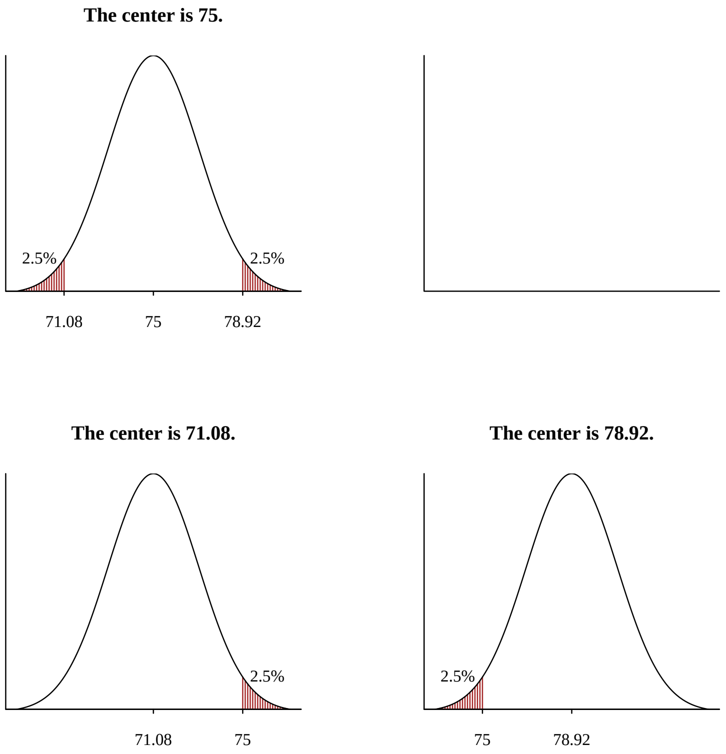
When we want to make inference about the *population parameters*, we take a (representative) *sample* from the population and compute a *statistic* using the data from the sample. Using the sample statistic, we estimate the population parameter.

- Sample mean to estimate the population mean.
- Sample proportion to estimate the population proportion.
- Sample correlation to estimate the population correlation.
- Sample difference of means/proportions to estimate the population difference.

## 5.2 Confidence intervals

### SPADI example

Recall that the sample mean was $75$ from a sample of size $16$. The population standard deviation was known to be $8$. We estimate the population mean to be $75$ (sample mean).


The center is 75.

The center is 71.08.

The center is 78.92.

If the true (unknown) mean is $71.08$, observing a sample mean of $75$ is barely plausible. And if the true mean is $78.92$, observing a sample mean of $75$ is barely plausible. Any values between these $2$ numbers would make observing $\bar{x} = 75$ not very unusual. Let's call this interval $(71.08, 78.92)$ a $95\%$ confidence interval.

Interpreting a $95\%$ confidence interval.

- "Probability that the true mean is between $71.08$ and $78.92$ is $95\%$" is wrong.

  The true mean is unknown but a constant. It is a regular number, so it is either in an interval or it is not. There is no probability (randomness) about it.

- Remember that the numbers, $71.08$ and $78.92$, are specific to the particular samples we observed. With different samples we will get a different confidence interval.

- If we imagine repeating this experiment many times, each sample will give us a different

> confidence interval. Most of $(95\%$ of) these confidence intervals contain the true, unknown mean, but some of them $(5\%)$ do not.

“We don’t know if the true mean is in $(71.08, 78.92)$. But we are using a process that produces intervals, $95\%$ of which include the true mean. (We don’t know if the one that we have is one of them...)

```r
## Recall that the population standard deviation is 8.
sig <- 8
## Sample mean is 75.
x.bar <- 75
## Sample size is 16.
n <- 16

## 2.5% of the data will fall below this number.

qnorm(p = 0.025, mean = x.bar, sd = sig/sqrt(n))

[1] 71.08

qnorm(p = 0.975, mean = x.bar, sd = sig/sqrt(n))

[1] 78.92
```

# Chapter 6

# Comparing Means

## 6.1 One sample test for mean

In CEASAR data, first we will look at the Surgery group, and we want to test if the true baseline QoL is $55$. $\sigma$ is assumed to be $35$.

```
table(d$Treatment)


Radiation   Surgery 
       70       130 

surgery <- subset(d, Treatment == "Surgery")
# surgery <- d[ d$Treatment == 'Surgery', ]
dim(surgery)

[1] 130  15
```

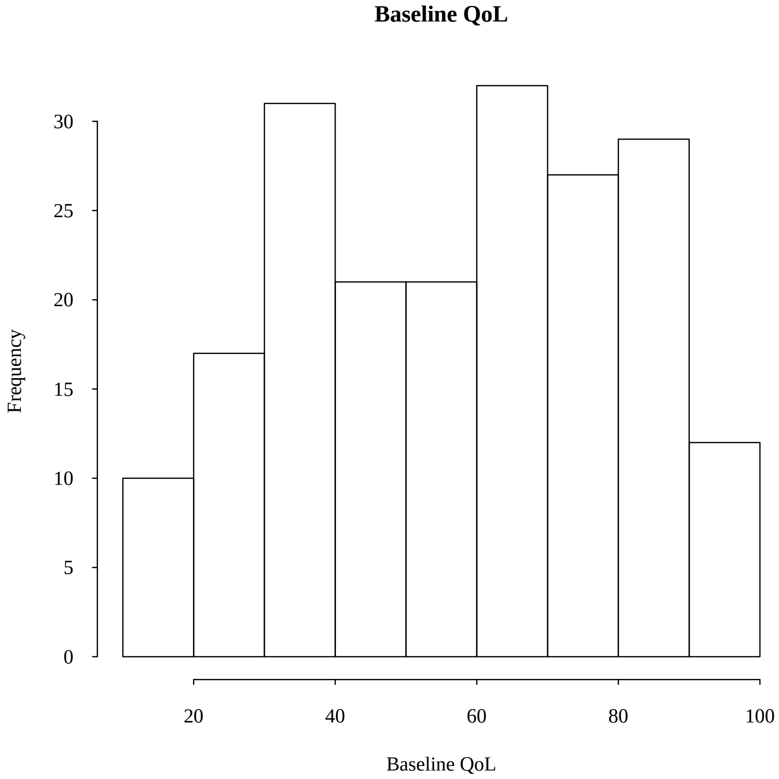
- Because we want to see if the mean is *different from* $55$. This is a $2$-sided test.

  $H_0 : \mu = 55$

  $H_1 : \mu \neq 55$

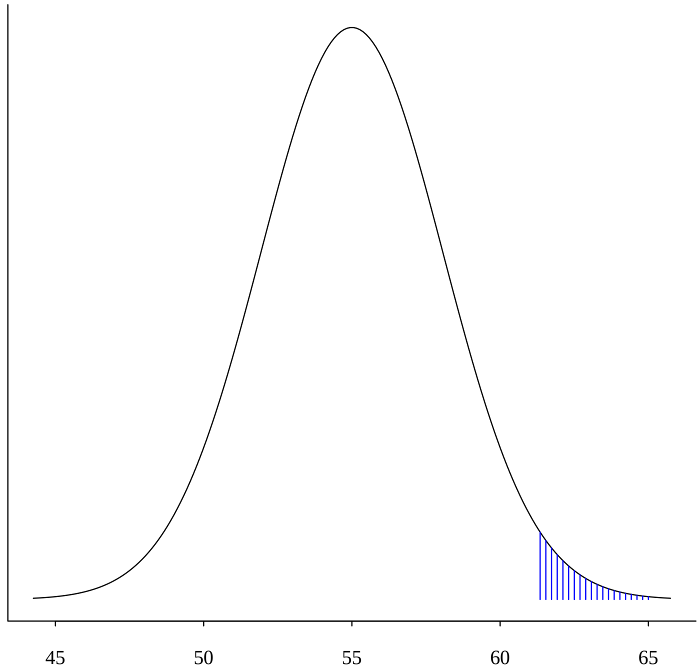
- We want to limit the type I error rate to $5\%$. But now this is a 2-sided test, we can make type I error rate on the upper side and lower side, ($H_0$ will be rejected if $\overline{X}$ is much bigger and much smaller than $55$.) and $5\%$ needs to be split into two. $2.5\%$ each. So if the p-value is less than $2.5\%$, we will reject $H_0$.

- Do we have a normally distributed data? Maybe not. But we are not that concerned because we are pretty sure that $\overline{X}$ is normally distributed.



```
(msq0 <- mean(surgery$QoL0))

[1] 61.35

(n <- nrow(surgery))

[1] 130
```

- The following is the distribution of $\overline{x}$ under $H_0$.



The shaded area is the p-value. How do we compute this?

We use *the* standard normal distribution, which is $Normal(0,1)$. We can convert any normal distribution to the standard normal by

$$z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}.$$

And we know $\overline{X} \sim Normal\left(55, \frac{35}{\sqrt{130}}\right)$, so

$$\begin{aligned} z &= \frac{61.35 - 55}{35/\sqrt{130}} \\ &= \frac{6.35}{3.07} \\ &= 2.068 \end{aligned}$$

Once we compute a $Z$ value, we can look up *the* table of standard normal probabilities. Or

```
(z.value <- (msq0 - 55)/(35/sqrt(n)))

[1] 2.068

pnorm(z.value, lower.tail = FALSE)

[1] 0.0193
```

Or we can use `pnorm` function without standardizing the value.

```
(pval <- pnorm(msq0, mean = 55, sd = 35/sqrt(n), lower.tail = FALSE))

[1] 0.0193
```

So the shaded area is $0.0193$, which is smaller than $2.5\%$. So we reject $H_0$ and conclude that the we have enough evidence to claim that the true mean is different from $55$. Can we say “bigger”? Probably. With this hypothesis testing, strictly speaking all we can claim is difference, but we can estimate the true mean and see if it is bigger than $55$.

**Confidence interval.** A $95\%$ confidence interval is an interval centered at the observed sample mean. Its width depends on the standard error ($\sigma/\sqrt{n}$).

$$\overline{X} \pm Z_{0.975} \times \frac{\sigma}{\sqrt{n}},$$

where $Z_{0.975}$ is the z value (from the standard normal distribution), and we know it is $1.96$.

```
qnorm(0.975)

[1] 1.96
```

If we want different value from $95\%$ (confidence limit), we can compute the corresponding z value using `qnorm` function. For example, for a $90\%$ confidence interval, we'd use $1.645$.

```
qnorm(0.95)

[1] 1.645
```

For the current problem, a $95\%$ confidence interval is

$$\begin{aligned}
&\overline{X} \pm Z_{0.975} \times \frac{\sigma}{\sqrt{n}} \\
&= 61.35 \pm 1.96 \times 3.07 \\
&= 61.35 \pm 6.02 \\
&= (55.33, 67.37)
\end{aligned}$$

So we think that the true mean is somewhere between $55.33$ and $67.37$.

## 6.2 Unknown variance

So far, we have assumed that the (unknown, true) population variance (and standard deviation) to be known. It is rare that we know the population standard deviation. When we don't know the population standard deviation, we substitute it with the sample standard deviation. Instead of

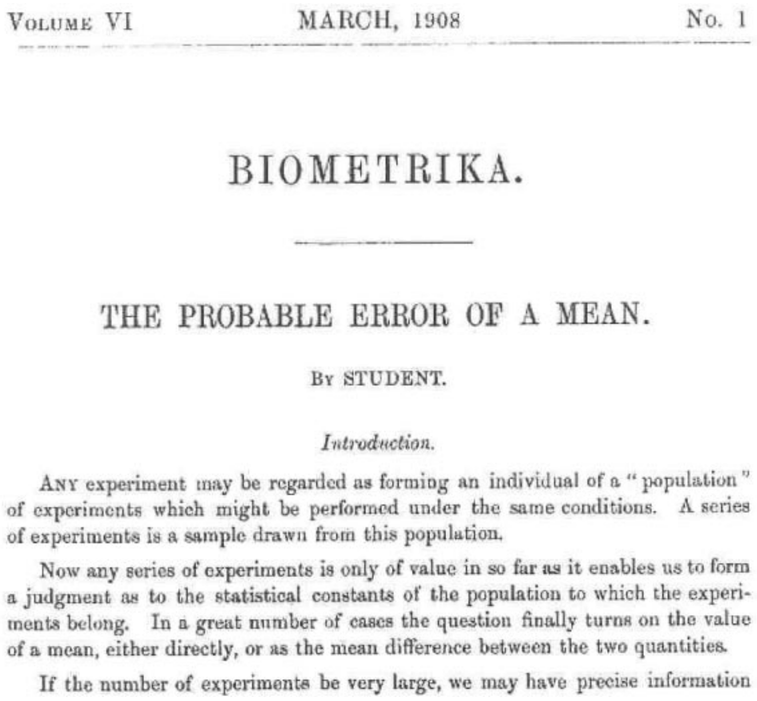
$$z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}},$$

we use

$$t = \frac{\overline{X} - \mu}{s/\sqrt{n}},$$

where $s$ is an estimate of the true standard deviation from the sample. $t$ does not have a normal distribution any more. Instead, its distribution is a Student's t distribution.
t distribution

- developed by William Gossett, who used the pen name "Student".



VOLUME VI MARCH, 1908 No. 1

BIOMETRIKA.

THE PROBABLE ERROR OF A MEAN.

BY STUDENT.

*Introduction.*

ANY experiment may be regarded as forming an individual of a "population" of experiments which might be performed under the same conditions. A series of experiments is a sample drawn from this population.

Now any series of experiments is only of value in so far as it enables us to form a judgment as to the statistical constants of the population to which the experiments belong. In a great number of cases the question finally turns on the value of a mean, either directly, or as the mean difference between the two quantities.

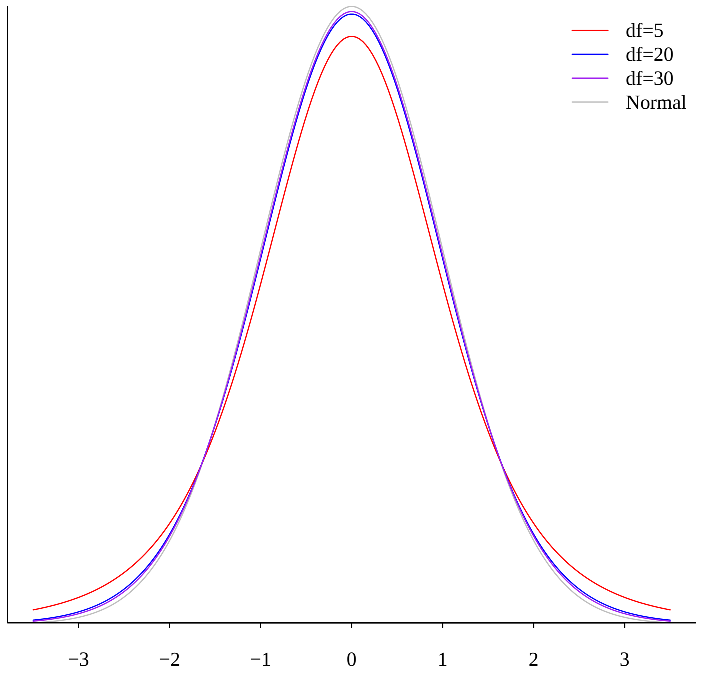
If the number of experiments be very large, we may have precise information

- It is centered at $0$.
- It is symmetric around $0$.
- It looks a lot like a normal distribution but has heavier tails.
- Its shape depends on the *degree of freedom*, which depends on the sample size.

  For a simple one-sample problem, the degree of freedom is $n-1$, which comes from the fact

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}.$$

- As the degree of freedom increases, t distributions start to look a lot like the standard normal distribution.



$97.5$ percentiles (to form a $95\%$ confidence interval)

```
qnorm(0.975)

[1] 1.96

qt(0.975, df = c(5, 10, 15, 20, 25, 30))

[1] 2.571 2.228 2.131 2.086 2.060 2.042
```

### 6.2.1 CEASAR example: revisited

Now we repeat the last example (beginning of Chapter 6) without assuming that the standard deviation is $35$. Sample standard deviation is

```
(s <- sd(surgery$QoL0))

[1] 22.78
```

$$
\begin{aligned}
t &= \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \\
&= \frac{61.35 - 55}{22.78/\sqrt{130}} \\
&= \frac{6.35}{2.00} \\
&= 3.178
\end{aligned}
$$

P-value is

```
(t.value <- (msq0 - 55)/(s/sqrt(n)))

[1] 3.178

pt(t.value, df = n - 1, lower = FALSE)

[1] 0.0009273
```

To compute a confidence interval, we replace $Z_{0.975}$ with $t_{0.975,df}$. For degree of freedom $129$ ($n = 130$), we have

```
(t975 <- qt(0.975, df = n - 1))

[1] 1.979
```

so that

$$
\begin{aligned}
&\overline{X} \pm t_{0.975,129} \times \frac{s}{\sqrt{n}} \\
&= 61.35 \pm 1.979 \times 2.00 \\
&= 61.35 \pm 3.95 \\
&= (57.40, 65.30)
\end{aligned}
$$

Using R.

```
t.test(surgery$QoL0, mu = 55, alternative = "two.sided")


One Sample t-test

data:  surgery$QoL0
t = 3.2, df = 129, p-value = 0.002
alternative hypothesis: true mean is not equal to 55
95 percent confidence interval:
 57.4 65.3
sample estimates:
mean of x
    61.35
```

- Why is this p-value different? My way is to divide $\alpha = 5\%$ by $2$ and compare the p-value to $2.5\%$. Their way is to multiply the p-value by $2$ and compare it with $\alpha = 5\%$.
- Why is this degree of freedom $130$ instead of $129$. After $df = 100$, `R` rounds `df` to the nearest $10$.

## 6.3 Paired t test

Paired data arise when two observations are made on the same individual. Or more generally, two *correlated* data are analyzed.

- Observations before and after treatment.
- Same individuals taking two drugs (cross-over clinical trial).
- Experiments involving siblings (one acting as a control).

The biggest advantage of an experiment utilizing paired data is reduced variance.

$$
\begin{aligned}
Var(X-Y) &= Var(X) + Var(Y) - 2Cov(X,Y) \\
&= Var(X) + Var(Y) - 2Cor(X,Y)\sqrt{Var(X)Var(Y)}
\end{aligned}
$$

So as long as the *correlation* is greater than $0$, variance of $X-Y$ is smaller than $Var(X)+Var(Y)$, which is the variance of $X-Y$ when they are uncorrelated.

Generally, we would like to test hypotheses about difference of the averages. We often write $\mu_1$ and $\mu_2$ to denote the population means at time $1$ and $2$, respectively, and use $\mu_d$ to denote $\mu_2-\mu_1$. So to test there is no difference between two time points (two drugs, two siblings), we write

$$H_0 : \mu_d = 0$$

$$H_1 : \mu_d > 0$$

if what we want to say is $\mu_2 > \mu_1$.

Even though there are two groups of data (time $1$ and time $2$), we compute their differences and treat the whole problem as one-sample problem. Paired t test is one-sample t test on differences! Let's see if the QoL scores decreased for the radiation group. Our hypotheses are:

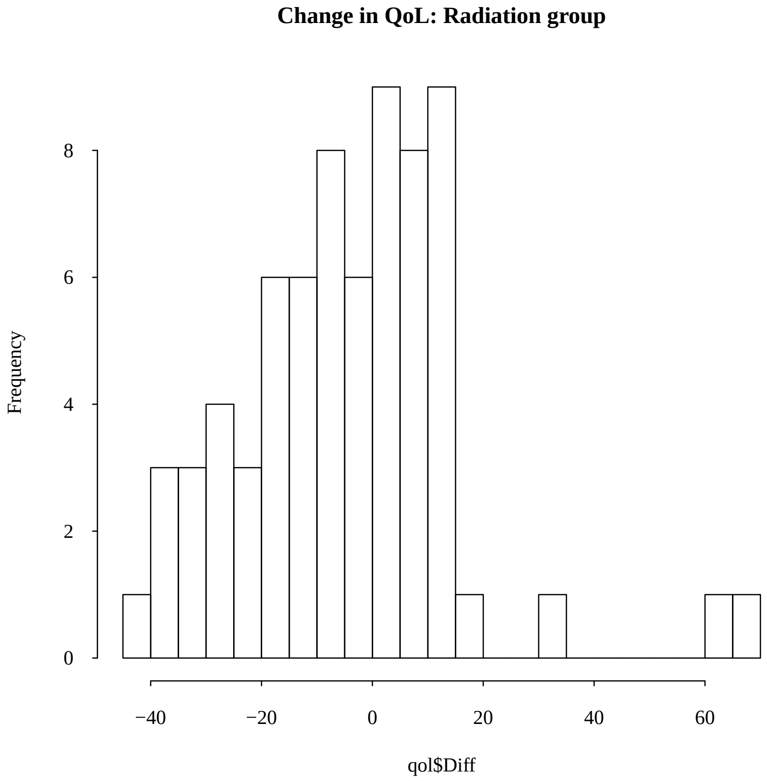
$$H_0 : \mu_d = 0$$

$$H_1 : \mu_d < 0$$

First we compute the difference of QoL for everyone in the radiation group.

```
radiation <- subset(d, Treatment == "Radiation")
qol <- subset(radiation, select = c("QoL0", "QoL6"))
qol$Diff <- qol$QoL6 - qol$QoL0

head(qol, 20)

   QoL0 QoL6  Diff
7  35.0 34.7  -0.3
12 79.4 93.6  14.2
15 60.0 71.8  11.8
16 28.0  3.6 -24.4
19 62.1 18.6 -43.5
21 67.0 80.7  13.7
22 78.3 76.3  -2.0
24 87.0 72.8 -14.2
25 63.6 53.9  -9.7
27 68.9 59.1  -9.8
28 31.7 13.6 -18.1
```

```
30 64.4 75.9  11.5
32 80.0 77.4  -2.6
34 33.2  0.8 -32.4
37 67.8 73.0   5.2
39 18.9 23.2   4.3
40 88.4 71.7 -16.7
51 38.6 20.6 -18.0
54 28.8 42.5  13.7
58 32.6 11.1 -21.5
```

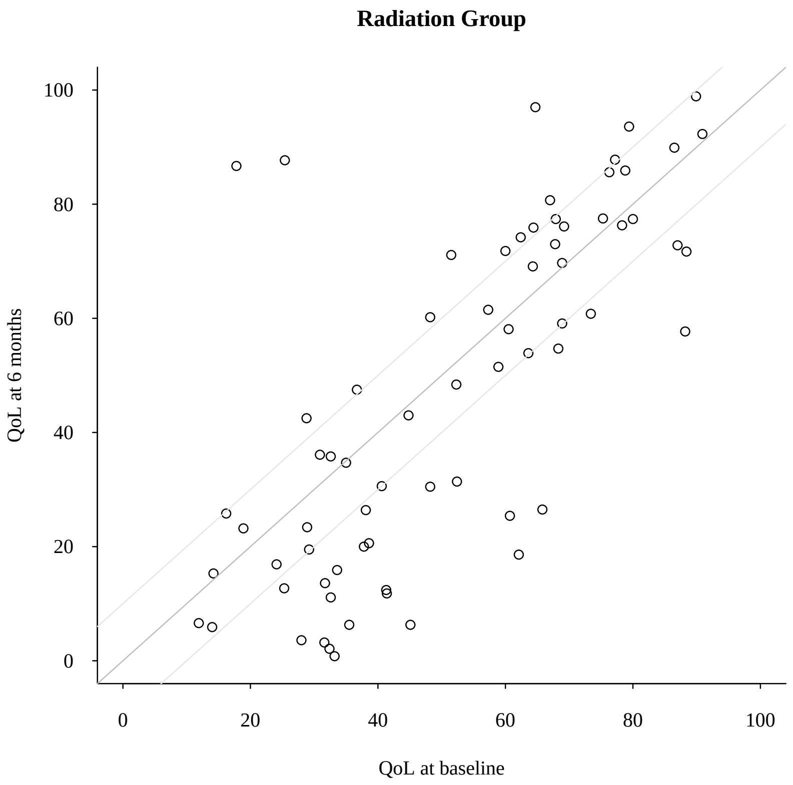


```
t.test(qol$Diff, mu = 0, alternative = "less", conf.level = 0.95)


One Sample t-test

data:  qol$Diff
t = -1.8, df = 69, p-value = 0.04
alternative hypothesis: true mean is less than 0
95 percent confidence interval:
    -Inf -0.4041
```

```
sample estimates:
mean of x
   -4.416

t.test(qol$QoL6, qol$QoL0, alternative = "less", paired = TRUE)


Paired t-test

data:  qol$QoL6 and qol$QoL0
t = -1.8, df = 69, p-value = 0.04
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
    -Inf -0.4041
sample estimates:
mean of the differences
                 -4.416
```

A one-sided confidence interval is given for a one-sided test. Sometimes, we want a two-sided confidence interval. To make a confidence interval and a one-sided test consistent, we can compute $90\%$ confidence interval for one-sided test with $\alpha = 5\%$.

```
t.test(qol$Diff, mu = 0, alternative = "two.sided", conf.level = 0.9)


One Sample t-test

data:  qol$Diff
t = -1.8, df = 69, p-value = 0.07
alternative hypothesis: true mean is not equal to 0
90 percent confidence interval:
 -8.4274 -0.4041
sample estimates:
mean of x
   -4.416
```



What happens if we cannot assume that $\overline{X}$ follows a normal distribution even approximately? This happens when the underlying distribution of $X$ is not normal and sample size is small. We'll have to use a *nonparametric method*, which does not require normality. (Later!)

## 6.4 Two sample data -Equal variance

Now we consider comparing two means from two groups of data (unpaired).

- There are two populations; one with mean $\mu_1$ and variance $\sigma_1^2$, and another one with mean $\mu_2$ and variance $\sigma_2^2$.

- First we assume $\sigma_1^2 = \sigma_2^2$ (Equal variance) and test

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

- Sample sizes are $n_1$ and $n_2$.
- Test statistic is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\text{se of numerator}}$$

- *If* we know the true variance $\sigma^2$, the variance of $\bar{x}_1$ - $\bar{x}_2$ is

$$\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \sigma^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)$$

- We need to estimate $\sigma_2$ from the two samples.

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 2)s_2^2}{n_1 + n_2 - 2}.$$

  This is called “pooled variance”.

- The true standard error of the difference of the sample means is

$$\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

  and its estimate is

$$s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

and

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

- Degree of freedom is the sum of the individual d.f., $n_1 - 1 + n_2 - 1 = n_1 + n_2 - 2$, which is the denominator of $s^2$. $-2$ comes from having to estimate two sample means.
- If $H_0$ is true, $t$ has the $t_{n_1+n_2-2}$ distribution.

### 6.4.1 CEASAR example: two-sample t test

We think that younger patients are more likely to receive surgery. Let's confirm this. Let $\mu_s$ and $\mu_r$ be the average ages of the surgery and radiation groups, respectively. We are to test

$$H_0 : \mu_s - \mu_r = 0$$
$$H_1 : \mu_s - \mu_r < 0.$$

Let's use $\alpha = 5\%$.
Here are some information about age of the patients.

```
mean(surgery$Age)

[1] 62.56

sd(surgery$Age)

[1] 8.242

mean(radiation$Age)

[1] 66.46

sd(radiation$Age)

[1] 8.217
```

The pooled variance is:

$$
\begin{aligned}
s_p^2 &= \frac{(n_s-1)s_s^2+(n_r-1)s_r^2}{n_s+n_r-2} \\
&= \frac{(129)8.24^2+(69)8.22^2}{198} \\
&= 67.795. \\
s_p &= \sqrt{67.795} \\
&= 8.234
\end{aligned}
$$

And t statistic is

$$
\begin{aligned}
t &= \frac{\overline{x}_t-\overline{x}_s}{s_p\sqrt{1/n_s+1/n_r}} \\
&= \frac{62.56-66.46}{8.23\sqrt{1/130+1/70}} \\
&= -3.19.
\end{aligned}
$$

And p-value is

```
pt(-3.19, df = 198, lower.tail = TRUE)

[1] 0.0008272
```

We can do this using `t.test` function.

```
t.test(surgery$Age, radiation$Age, alternative = "less", paired = FALSE, var.equal = TRUE)


Two Sample t-test

data:  surgery$Age and radiation$Age
t = -3.2, df = 198, p-value = 0.0008
alternative hypothesis: true difference in means is less than 0
```

```
95 percent confidence interval:
   -Inf -1.878
sample estimates:
mean of x mean of y
    62.56     66.46
```

A $95\%$ confidence interval for the difference of the means has the form:

$$\bar{x}_2 - \bar{x}_2 \pm t_{0.975,df} \times s_p \sqrt{1/n_1 + 1/n_2}$$

For the current problem (for $\mu_s - \mu_r$)

$$62.56 - 66.46 \pm t_{0.975,198} 8.23 \sqrt{0.022}$$

$t_{0.975,198}$ is $1.972$, and

$$-3.90 \pm 2.41 = (-6.303, -1.488)$$

Using `R`, we get the same answer:

```
t.test(surgery$Age, radiation$Age, alternative = "two.side", paired = FALSE, var.equal = TRUE,
    conf.level = 0.95)


Two Sample t-test

data:  surgery$Age and radiation$Age
t = -3.2, df = 198, p-value = 0.002
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -6.303 -1.488
sample estimates:
mean of x mean of y
    62.56     66.46
```

With such large sample sizes, the answers would not differ much if we used a normal distribution instead.

```
qt(0.975, df = 198)

[1] 1.972

qnorm(0.975)

[1] 1.96
```

## 6.5 Two sample data -Unequal variances

- When the true group variances are not assumed to be the same, we can not combine two samples to estimate the true variance. (We can not use the pooled variance.)

- This case the test statistic is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

```
(tt <- t.test(surgery$Age, radiation$Age, alternative = "less", paired = FALSE, var.equal = FALSE))


Welch Two Sample t-test

data:  surgery$Age and radiation$Age
t = -3.2, df = 142, p-value = 0.0009
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
   -Inf -1.876
sample estimates:
mean of x mean of y
    62.56     66.46
```

- Where does df $140$ come from? (Note: Newer version of $R$ gives a right df.)

- The Satterthwaite approximation is a formula to calculate an “effective” degrees of freedom in a two-sample t test.

$$
\begin{aligned}
df &= \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2} \\
&= 141.8.
\end{aligned}
$$

```
tt$parameter

   df
141.8
```

- In the old days, we used $df =$ smaller $n-1$.

# Chapter 7

# Nonparametric Methods

When the underlying distribution is not normal and we don't have large enough samples to apply the Central Limit Theorem, we cannot use the methods based on normal approximation (parametric methods). In such a case, a non-parametric method may be useful.

| Parametric test | Nonparametric test |
|---|---|
| 1 sample $t$ | Wilcoxon signed-rank |
| 2 sample $t$ | Wilcoxon rank sum |
| ANOVA | Kruskal-Wallis |
| Pearson's correlation | Spearman's rank correlation |

Example: Suppose that we have the following samples from Group A and Group B.

```
(A0 <- c(8, 7, 4, 9, 11, 12, 5, 13, 12))

[1]  8  7  4  9 11 12  5 13 12

(B <- c(3, 4, 2, 6, 9, 4, 2))

[1] 3 4 2 6 9 4 2

t.test(A0, B, paired = FALSE, var.equal = FALSE)


Welch Two Sample t-test

data:  A0 and B
```

```
t = 3.3, df = 14, p-value = 0.005
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.638 7.791
sample estimates:
mean of x mean of y
    9.000     4.286
```

With addition of just one large number, `t.test` looks very different.

```
(A1 <- c(8, 7, 4, 9, 11, 12, 5, 13, 12, 100))

 [1]   8   7   4   9  11  12   5  13  12 100

(B <- c(3, 4, 2, 6, 9, 4, 2))

[1] 3 4 2 6 9 4 2

t.test(A1, B, paired = FALSE, var.equal = FALSE)


Welch Two Sample t-test

data:  A1 and B
t = 1.5, df = 9.2, p-value = 0.2
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -6.931 34.560
sample estimates:
mean of x mean of y
   18.100     4.286
```

The non-significant result is due to increased standard deviation, which is particularly not robust to outliers.

```
sd(A0)

[1] 3.24

sd(A1)

[1] 28.94
```

- Non-parametric methods are not heavily influenced by outliers or skewness.

  They do not use the means or standard deviations. Instead they transform the data to ranks.

- What are the hypotheses about? (They are not about the means.)

  Loosely speaking these tests are about the population medians.

- The confidence interval and the test may not correspond to each other.

### One sample test

First let's see if the median of the population $A1$ is not equal to $6$. We use a Wilcoxon signed rank test.

```
wilcox.test(A1, mu = 6, alternative = "two.sided", conf.int = TRUE)

Warning in wilcox.test.default(A1, mu = 6, alternative = "two.sided", conf.int = TRUE): cannot
compute exact p-value with ties
Warning in wilcox.test.default(A1, mu = 6, alternative = "two.sided", conf.int = TRUE): cannot
compute exact confidence interval with ties


Wilcoxon signed rank test with continuity correction

data:  A1
V = 50, p-value = 0.02
alternative hypothesis: true location is not equal to 6
95 percent confidence interval:
  6.5 53.5
sample estimates:
(pseudo)median
            10
```

If we use a t test,

```
t.test(A1, mu = 6, alternative = "two.sided", conf.int = TRUE)


One Sample t-test

data:  A1
t = 1.3, df = 9, p-value = 0.2
alternative hypothesis: true mean is not equal to 6
95 percent confidence interval:
 -2.601 38.801
sample estimates:
mean of x
     18.1
```

### Two sample test

Now test if the medians of the population $A1$ and $B$ are the same. For this 2 sample test, we use a Wilcoxon rank-sum test (aka Mann-Whitney U test).

```
wilcox.test(A1, B, alternative = "two.sided", conf.int = TRUE)

Warning in wilcox.test.default(A1, B, alternative = "two.sided", conf.int = TRUE): cannot compute
exact p-value with ties
Warning in wilcox.test.default(A1, B, alternative = "two.sided", conf.int = TRUE): cannot compute
exact confidence intervals with ties


Wilcoxon rank sum test with continuity correction

data:  A1 and B
W = 62, p-value = 0.008
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 2 9
sample estimates:
difference in location
                  5.25
```

The corresponding t test is:

```
t.test(A1, B, alternative = "two.sided", conf.int = TRUE)


Welch Two Sample t-test

data:  A1 and B
t = 1.5, df = 9.2, p-value = 0.2
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -6.931 34.560
sample estimates:
mean of x mean of y
   18.100     4.286
```

Rank-based tests like Wilcoxon rank-sum are about the same as t test on ranks.

```
dx <- data.frame(y = c(A1, B), g = rep(c("A", "B"), c(length(A1), length(B))))
dx$r <- rank(dx$y)
dx
```

```
     y g    r
1    8 A 10.0
2    7 A  9.0
3    4 A  5.0
4    9 A 11.5
5   11 A 13.0
6   12 A 14.5
7    5 A  7.0
8   13 A 16.0
9   12 A 14.5
10 100 A 17.0
11   3 B  3.0
12   4 B  5.0
13   2 B  1.5
14   6 B  8.0
15   9 B 11.5
16   4 B  5.0
17   2 B  1.5

t.test(r ~ g, data = dx, alternative = "two.sided")


Welch Two Sample t-test

data:  r by g
t = 3.6, df = 14, p-value = 0.003
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  2.679 10.678
sample estimates:
mean in group A mean in group B
         11.750           5.071
```

Perhaps, these non-parametric tests should be our default choice...

# Chapter 8

# Proportions

## 8.1 One proportion

When outcome is binary (Yes/No; Success/Failure), we count the number of 'successes' ($X$) out of $n$. The number of successes has $Binomial(n, p)$ distribution, where $p$ is the true (unknown) probability of success. We are interested in estimating this $p$. (population proportion)

- Given that $X \sim Binomial(n, p)$, the mean is $np$ and variance is $np(1-p)$. And standard deviation is $\sqrt{np(1-p)}$.
- The sample proportion, $\hat{p} = X/n$, has the mean $p$ and standard deviation $p(1-p)/n$.
- To make an inference about $p$, we either use an approximate method (normal approximation) or an exact method.
- When sample size is large ($X \geq 30$), normal approximation is good.

Approximate $95\%$ confidence interval for $p$ is

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

To test $H_0 : p = p_0$, use

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

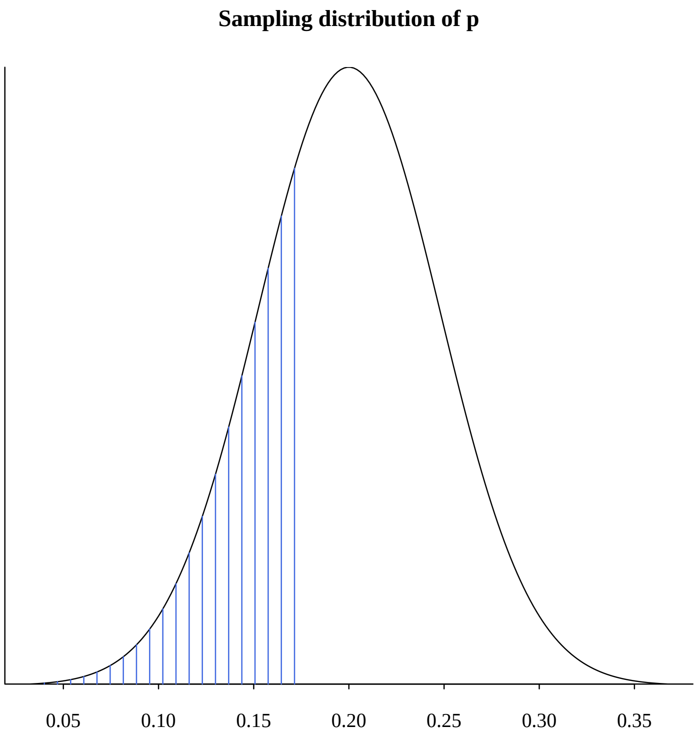
which is approximately $Normal(0,1)$ under $H_0$. When estimating $p$, we never use a t distribution.

Example: Using the CEASAR data, estimate the proportion of the radiation patients who had diabetes, and test if it is different from $20\%$, and give a $95\%$ confidence interval.

```
table(radiation$Diabetes)


 No Yes 
 58  12 

prop.table(table(radiation$Diabetes))


    No    Yes 
0.8286 0.1714 
```

From this, we get $\hat{p} = 0.1714$. The null value, $p_0 = 0.20$. So we have

$$\begin{aligned} z &= \frac{0.1714 - 0.20}{\sqrt{0.20(1-0.80)/70}} \\ &= -0.5976 \end{aligned}$$


Sampling distribution of p

P-value is

```
pnorm(-0.5976)

[1] 0.2751
```

and this is compared to $5\%/2$. We do not have enough evidence to reject $p = 0.20$.

To compute a $95\%$ confidence interval, we use

```
qnorm(0.975)

[1] 1.96
```

so that it is $\hat{p} \pm 1.96\sqrt{(\hat{p}(1-\hat{p}))/n}$

```
phat <- 12/70
margin <- qnorm(0.975) * sqrt(phat * (1 - phat)/70)
phat + c(-1, 1) * margin
```

```
[1] 0.08314 0.25972
```

Thus, a $95\%$ confidence interval is $(0.0831, 0.2597)$.

We can use `binconf` function in `Hmisc` library to get the same answer.

```
library(Hmisc)
binconf(x = 12, n = 70, method = "all", include.x = TRUE, include.n = TRUE, alpha = 0.05)

              X  N PointEst   Lower  Upper
Exact        12 70   0.1714 0.09184 0.2803
Wilson       12 70   0.1714 0.10088 0.2762
Asymptotic 12 70   0.1714 0.08314 0.2597
```

Because the “Asymptotic” method only works well as $n$ gets large, sometimes it does not give a confidence interval with a specified confidence level ($95\%$ in this example). Wilson score interval (1927) has an improved coverage probability and is often preferred.

## 8.2 Two proportions

Now we are concerned with the problem of comparing two proportions. Data are often tabulated in a 2 by 2 table.

Let’s test if the probabilities of success are the same for treatment 1 and 2 given the following data.

| | Success | Failure |
|---|---|---|
| Treatment 1 | 7 | 13 |
| Treatment 2 | 3 | 22 |

```
## Set up the data
trt <- rep(c("Trt.1", "Trt.2"), c(7 + 13, 3 + 22))
outcome <- rep(rep(c("Success", "Failure"), 2), c(7, 13, 3, 22))
outcome <- factor(outcome, levels = c("Success", "Failure"))
(dat <- data.frame(trt, outcome))

     trt outcome
1  Trt.1 Success
2  Trt.1 Success
```

```
3  Trt.1 Success
4  Trt.1 Success
5  Trt.1 Success
6  Trt.1 Success
7  Trt.1 Success
8  Trt.1 Failure
9  Trt.1 Failure
10 Trt.1 Failure
11 Trt.1 Failure
12 Trt.1 Failure
13 Trt.1 Failure
14 Trt.1 Failure
15 Trt.1 Failure
16 Trt.1 Failure
17 Trt.1 Failure
18 Trt.1 Failure
19 Trt.1 Failure
20 Trt.1 Failure
21 Trt.2 Success
22 Trt.2 Success
23 Trt.2 Success
24 Trt.2 Failure
25 Trt.2 Failure
26 Trt.2 Failure
27 Trt.2 Failure
28 Trt.2 Failure
29 Trt.2 Failure
30 Trt.2 Failure
31 Trt.2 Failure
32 Trt.2 Failure
33 Trt.2 Failure
34 Trt.2 Failure
35 Trt.2 Failure
36 Trt.2 Failure
37 Trt.2 Failure
38 Trt.2 Failure
39 Trt.2 Failure
40 Trt.2 Failure
41 Trt.2 Failure
42 Trt.2 Failure
43 Trt.2 Failure
44 Trt.2 Failure
45 Trt.2 Failure
```

### 8.2.1 Fisher's exact test

- First we fix both margins.

| | Success | Failure | Total |
|---|---|---|---|
| Treatment 1 | | | 20 |
| Treatment 2 | | | 25 |
| Total | 10 | 35 | 45 |

- Then if we fix 1 of the 4 cells (Treatment 1, Success), we have a complete table.

- We can determine which of the tables are "more extreme" than observed for each table. Below is an example of an "extreme" table under the null (no difference between treatments).

| | Success | Failure | Total |
|---|---|---|---|
| Treatment 1 | 10 | 10 | 20 |
| Treatment 2 | 0 | 25 | 25 |
| Total | 10 | 35 | 45 |

- We know how to compute the probability of each table (using a hypergeometric distribution), and we also have a way to order all the tables from "least extreme under null" to "most extreme".

- Sum of the probabilities that correspond to "more extreme than observed" is the p-value.

- This is called "exact" because we don't use large-sample approximation.

```
fisher.test(table(dat))


Fisher's Exact Test for Count Data

data:  table(dat)
p-value = 0.08
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  0.7172 26.9850
sample estimates:
odds ratio
     3.825
```

### 8.2.2 Odds ratio

- The summary statistic often used for a 2 by 2 table is an odds ratio.
- Odds of Success = # Success / # Failure
  Contrast it with Probability of Success = # Success / (# Success + # Failure)
- Equivalent to $p/(1-p)$ if $p$ is the probability of success.
- Odds range from $0$ to $\infty$.
- To compare two groups (treatments), we compute the odds for each group and compute the ratio. (Odds ratio)
- Odds ratio (OR) $= \frac{p_1}{1-p_1} / \frac{p_2}{1-p_2}$.
- Odds ratio ranges from $0$ to $\infty$, and $OR = 1$ means $p_1 = p_2$.

For the current example, odds ratio is:

```
(7/13)/(3/22)

[1] 3.949
```

It's slightly different from `fisher.test` output as it uses a slightly different method. We read the $R$ output as "Odds of success for Treatment 1 is $3.8$ times ($95\%$ CI: $0.72$ to $27.0$) as big as the odds for Treatment 2."

### 8.2.3 $\chi^2$ test

We can use a $\chi^2$ test to test the same hypothesis $p_1 = p_2$.

| | Success | Failure | Total |
|---|---|---|---|
| Treatment $1$ | | | $20$ |
| Treatment $2$ | | | $25$ |
| Total | $10$ | $35$ | $45$ |

- If we assume $p_1 = p_2$, then we'd expect $10/45$ $(22.2\%)$ of $20$ Treatment 1 data to be success.

- We can compute the expected cell count for each cell to get:

| | Success | Failure | Total |
|---|---|---|---|
| Treatment $1$ | $4.4$ | $15.6$ | $20$ |
| Treatment $2$ | $5.6$ | $19.4$ | $25$ |
| Total | $10$ | $35$ | $45$ |

- We use how far the actual data are from this expected cell counts to test the hypotheses.

- "$\sum$(Observed − Expected)$^2$/Expected" follows a $\chi^2$ distribution.

```
chisq.test(table(dat), correct = FALSE)

Warning in chisq.test(table(dat), correct = FALSE): Chi-squared approximation may be incorrect


Pearson's Chi-squared test

data:  table(dat)
X-squared = 3.4, df = 1, p-value = 0.07
```

- $\chi^2$ test uses asymptotic theory and works when $n$ is large.

- There is a popular belief that $\chi^2$ should be used only all the cell counts are $5$ or larger. (and Fisher's test should be used for small data set.)

- $\chi^2$ test actually works fine with small cell counts.

- We don't use Yates' continuity correction because it is an overly conservative test.

### 8.2.4 Normal theory

- Recall that the null hypothesis is $H_0 : p_1 = p_2$. This does not specify what value these two population proportions are equal to.

- Sample proportions are $\hat{p}_1$ and $\hat{p}_2$.
- The z value is

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\,(1/n_1 + 1/n_2)}}$$

- $\hat{p}$ are from combined data. $(x_1 + x_2)/(n_1 + n_2)$.
- We can combine the groups because $p_1 = p_2$ under the null.
- For a confidence interval, such an assumption is not applicable, and we use

$$\hat{p}_1 - \hat{p}_2 \pm z \times \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

  $z$ is $1.96$ for a $95\%$ confidence interval.

Back to example:

```
x1 <- 7
n1 <- 20
x2 <- 3
n2 <- 25
(p1 <- x1/n1)

[1] 0.35

(p2 <- x2/n2)

[1] 0.12

(p.hat <- (x1 + x2)/(n1 + n2))

[1] 0.2222

# Z-value
(z <- (p1 - p2)/sqrt(p.hat * (1 - p.hat) * (1/n1 + 1/n2)))

[1] 1.844

# P-value
pnorm(z, lower.tail = FALSE)
```

```
[1] 0.03258

# Confidence interval for p1-p2
p1 - p2 + c(-1, 1) * qnorm(0.975) * sqrt(p1 * (1 - p1)/n1 + p2 * (1 - p2)/n2)

[1] -0.01479  0.47479
```

There is not enough evidence that the two treatments' probabilities of success are different. Estimate of the difference of the probabilities (Treatment 1 $-$ Treatment 2) is $0.23$ and its confidence interval is $(-0.015, 0.47)$.

We can use the following function à la me to do the same test:

```
twoSamplePropTest <- function(x, n, confLimit = 0.95) {
    # This will test if proportions are equal.  x and n are vectors of length 2.
    pObs <- x/n
    phat0 <- sum(x)/sum(n)

    zValue <- -diff(pObs)/sqrt(phat0 * (1 - phat0) * sum(1/n))
    pValue <- pnorm(abs(zValue), lower.tail = FALSE)
    ci <- -diff(pObs) + c(-1, 1) * qnorm(1 - (1 - confLimit)/2) * sqrt(sum(pObs * (1 - pObs)/n))

    list(estimate = pObs, z = zValue, oneSided.p = pValue, conf.int = ci)
}

twoSamplePropTest(x = c(7, 3), n = c(20, 25), confLimit = 0.95)

$estimate
[1] 0.35 0.12

$z
[1] 1.844

$oneSided.p
[1] 0.03258

$conf.int
[1] -0.01479  0.47479
```

# Chapter 9

# Sample Size and Power

When we cannot reject $H_0$ (or when our confidence interval contains the null value), we are left with no conclusion. We can never say that we show that there is no difference. The only thing we can say is that we didn't plan our study well and the sample size was too small.

- Tacking on new samples after concluding "sample size was too small" is also problematic because it inflates type I error rate.
- So it is critical that we start an experiment with the *right* sample size.
- In general, we compute the sample size so that the hypothesis test has enough power, which is the probability of rejecting $H_0$ under $H_1$.
- Example: We want to say that the baseline QoL is higher for the surgery group.
  $H_0 : \mu_s = \mu_r$
  $H_1 : \mu_s > \mu_r$.
  And we want to make sure that we reject $H_0$ when the true difference is at least $10$.
  We say, we want power to be $90\%$ when $\mu_s - \mu_r = 10$.
- Power is usually set at $90\%$ or $80\%$.

What's wrong with the sample size being too large?

Factors that affect sample size and power.

Everything being equal...

- Sample size ↑ ... Power ↑
- Type I error rate ($\alpha$) ↓ ... Power ↓
- Difference to detect ↑ ... Power ↑
- Standard deviation ↑ ... Power ↓

## 9.1 Continuous random variables

For continuous outcome (t test), we need to know the standard deviation to perform sample size computation! Let's compute the sample size needed to detect a difference of $10$ in the baseline QoL between two groups. We set $\alpha = 0.025$. (equivalent to $5\%$ two-sided test). Assume that $\sigma = 30$.

```
power.t.test(n = NULL, delta = 10, sd = 30, sig.level = 0.025, power = 0.9, alternative = "one")


     Two-sample t test power calculation

              n = 190.1
          delta = 10
             sd = 30
      sig.level = 0.025
          power = 0.9
    alternative = one.sided

NOTE: n is number in *each* group

power.t.test(n = NULL, delta = 10, sd = 30, sig.level = 0.025, power = 0.8, alternative = "one")


     Two-sample t test power calculation

              n = 142.2
          delta = 10
             sd = 30
      sig.level = 0.025
          power = 0.8
    alternative = one.sided

NOTE: n is number in *each* group
```

If we assume that $\sigma = 20$.

```
power.t.test(n = NULL, delta = 10, sd = 20, sig.level = 0.025, power = 0.9, alternative = "one")


     Two-sample t test power calculation

              n = 85.03
          delta = 10
             sd = 20
      sig.level = 0.025
          power = 0.9
    alternative = one.sided

NOTE: n is number in *each* group
```

```
power.t.test(n = NULL, delta = 10, sd = 20, sig.level = 0.025, power = 0.8, alternative = "one")


     Two-sample t test power calculation

              n = 63.77
          delta = 10
             sd = 20
      sig.level = 0.025
          power = 0.8
    alternative = one.sided

NOTE: n is number in *each* group
```

## 9.2 Binary random variables

To test $p_1 = p_2$ against $p_1 > p_2$ with a one-sided $\alpha = 0.025$, power=$90\%$ when $p_1 - p_2 = 0.10$.

- Even though the null and alternative hypotheses do not specify $p_1$ and $p_2$ separately, the power calculation requires setting $p_1$ and $p_2$.

```
require(Hmisc)
bsamsize(p1 = 0.1, p2 = 0.2, fraction = 0.5, alpha = 0.05, power = 0.8)

 n1  n2
199 199

bsamsize(p1 = 0.2, p2 = 0.3, fraction = 0.5, alpha = 0.05, power = 0.8)

   n1    n2
293.2 293.2

bsamsize(p1 = 0.3, p2 = 0.4, fraction = 0.5, alpha = 0.05, power = 0.8)

   n1    n2
355.9 355.9
```

This example shows, even though the detectable difference is the same $(10$ percentage difference), the sample sizes differ much depending on the location of $p_1$ and $p_2$. Larger sample size is required near $0.5$.

## 9.3 Additional topics on sample size and power

### 9.3.1 Continuous endpoint

- If $\sigma$ is doubled, $n$ is $4$ times bigger.

```
power.t.test(n = NULL, delta = 10, sd = 20, sig.level = 0.025, power = 0.9, alternative = "one")


     Two-sample t test power calculation

              n = 85.03
          delta = 10
             sd = 20
      sig.level = 0.025
          power = 0.9
    alternative = one.sided

NOTE: n is number in *each* group

power.t.test(n = NULL, delta = 10, sd = 40, sig.level = 0.025, power = 0.9, alternative = "one")
```

```
     Two-sample t test power calculation

              n = 337.2
          delta = 10
             sd = 40
      sig.level = 0.025
          power = 0.9
    alternative = one.sided

NOTE: n is number in *each* group
```

- $N$ for two-sided test with type I error $= \alpha$ is the same for $N$ for one-sided test with $\alpha/2$.

```
power.t.test(n = NULL, delta = 10, sd = 20, sig.level = 0.05, power = 0.9, alternative = "two")


     Two-sample t test power calculation

              n = 85.03
          delta = 10
             sd = 20
      sig.level = 0.05
          power = 0.9
    alternative = two.sided

NOTE: n is number in *each* group

power.t.test(n = NULL, delta = 10, sd = 20, sig.level = 0.025, power = 0.9, alternative = "one")


     Two-sample t test power calculation

              n = 85.03
          delta = 10
             sd = 20
      sig.level = 0.025
          power = 0.9
    alternative = one.sided

NOTE: n is number in *each* group
```

- $2$ sample tests require larger sample sizes than $2$ times $n$ for $1$ sample test.

```
power.t.test(n = NULL, delta = 10, sd = 20, sig.level = 0.05, power = 0.9, alternative = "two",
    type = "one")


     One-sample t test power calculation

              n = 44
          delta = 10
             sd = 20
      sig.level = 0.05
          power = 0.9
    alternative = two.sided

power.t.test(n = NULL, delta = 10, sd = 20, sig.level = 0.05, power = 0.9, alternative = "two",
    type = "two")


     Two-sample t test power calculation

              n = 85.03
          delta = 10
             sd = 20
      sig.level = 0.05
          power = 0.9
    alternative = two.sided

NOTE: n is number in *each* group
```

- The sample size given above is for *each group*.
- Variance of difference is large when the data are uncorrelated.

$$Var(X - Y) = Var(X) + Var(Y)$$

- For paired test, we need the standard deviation of differences.

```
## sd for difference is not 20 if sd for X and for Y are 20.
power.t.test(n = NULL, delta = 10, sd = 20, sig.level = 0.05, power = 0.9, alternative = "two",
    type = "paired")
```

```
     Paired t test power calculation

              n = 44
          delta = 10
             sd = 20
      sig.level = 0.05
          power = 0.9
    alternative = two.sided

NOTE: n is number of *pairs*, sd is std.dev. of *differences* within pairs
```

- We need an additional assumption about $\rho$, correlation of $X$ and $Y$. Then we can compute

$$Var(X-Y) = Var(X) + Var(Y) - 2\rho\sqrt{Var(X)Var(Y)}$$

```
(sdd <- sqrt(20^2 + 20^2 - 2 * 0.8 * 20 * 20))

[1] 12.65

power.t.test(n = NULL, delta = 10, sd = sdd, sig.level = 0.05, power = 0.9, alternative = "two",
    type = "paired")


     Paired t test power calculation

              n = 18.84
          delta = 10
             sd = 12.65
      sig.level = 0.05
          power = 0.9
    alternative = two.sided

NOTE: n is number of *pairs*, sd is std.dev. of *differences* within pairs
```

- Sample size computation for odds ratio uses the same formula for difference of proportions.

- Given $p_1$ and an odds ratio (to detect) $\psi$, we can compute $p_2$.

$$\psi = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$$
$$p_2 = \frac{\psi p_2}{\psi p_2 + 1 - p_2}$$

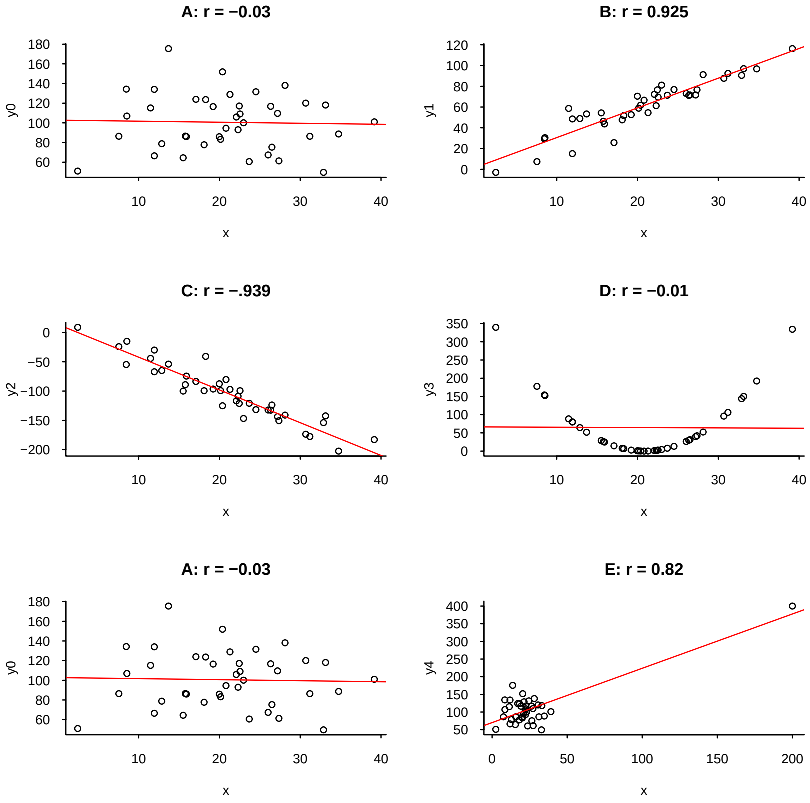
```
find.p2 <- function(p1, odds.ratio) {
    odds.ratio * p1/(odds.ratio * p1 + 1 - p1)
}

(p2x <- find.p2(0.3, 1.8))

[1] 0.4355

bpower(p1 = 0.3, p2 = p2x, n1 = 200, n2 = 200, alpha = 0.05)

 Power
0.8046

(p2y <- find.p2(0.1, 1.8))

[1] 0.1667

bpower(p1 = 0.1, p2 = p2y, n1 = 200, n2 = 200, alpha = 0.05)

 Power
0.5005
```

# Chapter 10

# Regression Analysis

## 10.1 Correlation

- Correlation is the measure of *linear* association between two (continuous) random variables. The most popular statistic is Pearson's correlation ($\rho$). It is a parametric statistic, and all the cautions as before apply, i.e., it doesn't work well with outliers and skewness.
- $\rho$ ranges from $-1$ to $1$. $\rho = 0$ means that the two random variables are *uncorrelated*. $\rho$ near $1$ is strong positive correlation, and near $-1$ is strong negative correlation.



- A simple linear regression would make sense in B and C.
- We can test for significant association by testing if the population correlation is $0$.

$$t = \frac{r\sqrt{n-2}}{1-r^2}$$

has $t$ distribution with df $= n-2$ under the null ($\rho = 0$).

- Confidence interval is a little complex...

```
cor.test(x, y0, alternative = "two.sided", method = "pearson", conf.level = 0.95)


Pearson's product-moment correlation

data:  x and y0
```

```
t = -0.19, df = 38, p-value = 0.9
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.3385  0.2840
sample estimates:
     cor
-0.03021
```

- A popular *nonparametric* correlation measure is Spearman's correlation ($\rho$). This is Pearson's correlation on ranks.

```
cor(x4, y4)  # Default is Pearson.

[1] 0.8258

cor(x4, y4, method = "spearman")

[1] 0.07711

cor(rank(x4), rank(y4), method = "pearson")

[1] 0.07711
```

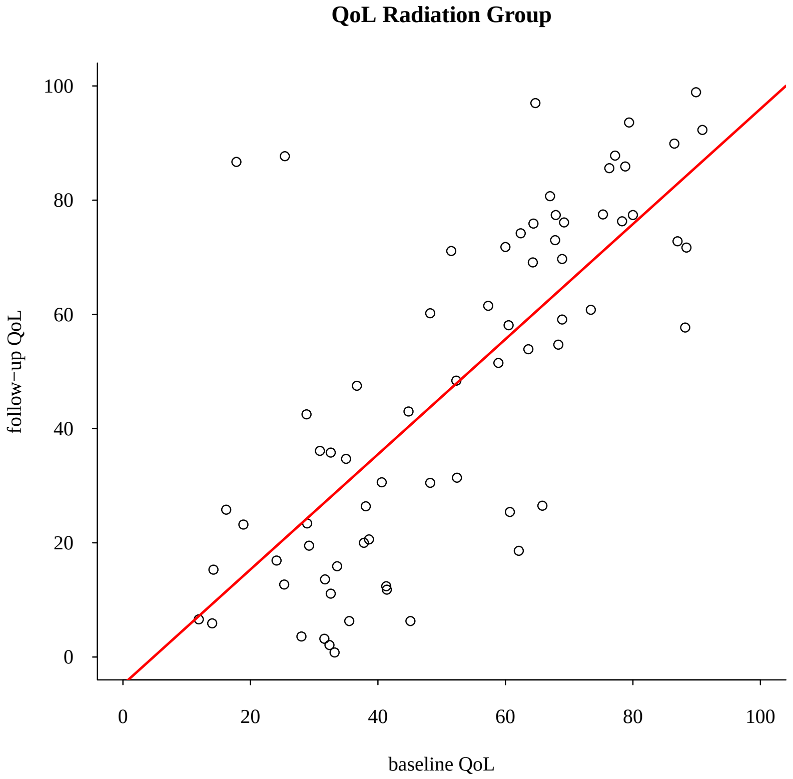
- Spearman's correlation is also a measure of *linear* association. It does not work with a non-linear association.

```
cor(x, y3, method = "pearson")

[1] -0.008873

cor(x, y3, method = "spearman")

[1] -0.001689
```

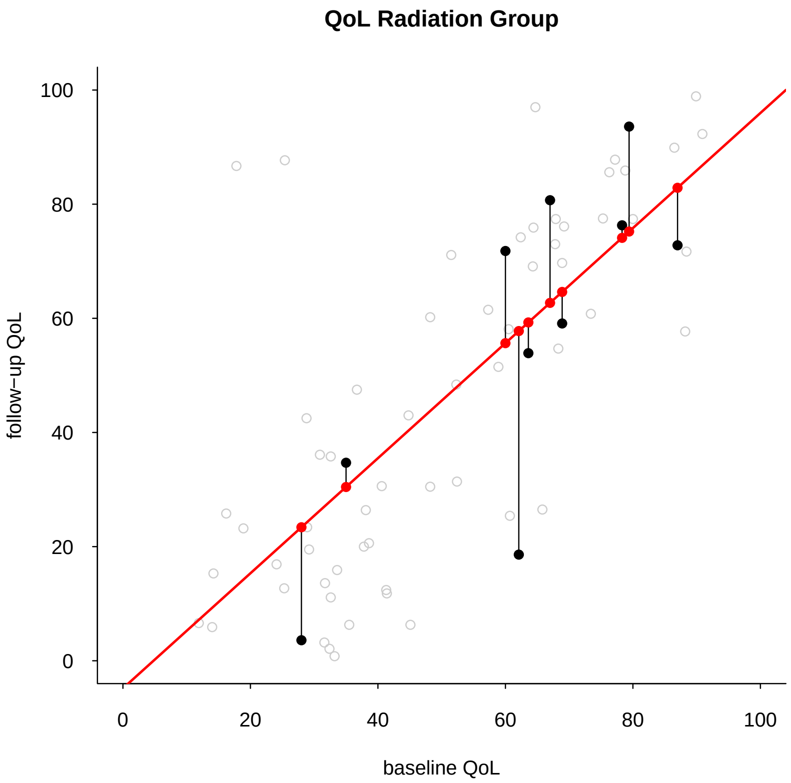
## 10.2 Simple regression

A regression analysis fits a line that describes the relationship between two continuous variables. Let's see the CEASAR example: The following scatter plot shows QoL0 (baseline) on the x-axis and QoL6 (6 months) on the y-axis for the radiation group.

```r
library(rms)
## To use regression functions in rms package.
ddist <- datadist(d)
options(datadist = "ddist")
```



This line is fit with the Ordinary Least Squares (OLS) method, and it is the best in the sense that sum of $(\text{vertical distance to the line})^2$ is minimized.



- The vertical distance is $|\hat{y}_i - y_i|$, where $\hat{y}_i$ is the fitted value (predicted value) for the $i^{th}$ observation.
- The vertical distances are called *residuals*.
- This line is written as

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

where $y_i$ is follow-up QoL, $x_i$ is baseline QoL, $\beta_0$ is the intercept, $\beta_1$ is the slope, and $\varepsilon_i$ is the residuals. Because we *estimate* the slope and intercept using the observed sample, we write:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \varepsilon_i.$$

For the current example, we have:

$$\hat{y} = -4.835 + 1.008x.$$

- What do these *regression coefficients* mean?
  - $\beta_0$ is the intercept, which is the value of $y$ when $x = 0$. Sometimes, it does not have a good interpretation; e.g., when $x$ is BMI, height, or weight. In this example, because nobody had $QoL = 0$, $\hat{\beta}_0 = -4.835$ is probably meaningless. It is outside of range!
  - $\beta_1$ is the slope, which is increment in $y$ corresponding to a unit increment in $x$. QoL6 increases by $1.008$ when QoL0 is increased by $1$.

Let's look at the $R$ output.

```
(mod.rad <- ols(QoL6 ~ QoL0, data = radiation))

Linear Regression Model

 ols(formula = QoL6 ~ QoL0, data = radiation)

                 Model Likelihood     Discrimination
                    Ratio Test           Indexes
 Obs        70    LR chi2     55.93    R2        0.550
 sigma20.2780    d.f.            1    R2 adj    0.544
 d.f.       68    Pr(> chi2) 0.0000    g        25.750

 Residuals

     Min       1Q  Median      3Q     Max
 -39.171 -13.228   1.046  11.262  73.590


           Coef    S.E.   t     Pr(>|t|)
 Intercept -4.8350 6.1817 -0.78 0.4368
 QoL0       1.0082 0.1105  9.12 <0.0001
```

- Coefficients, standard errors, t values, and p-values.

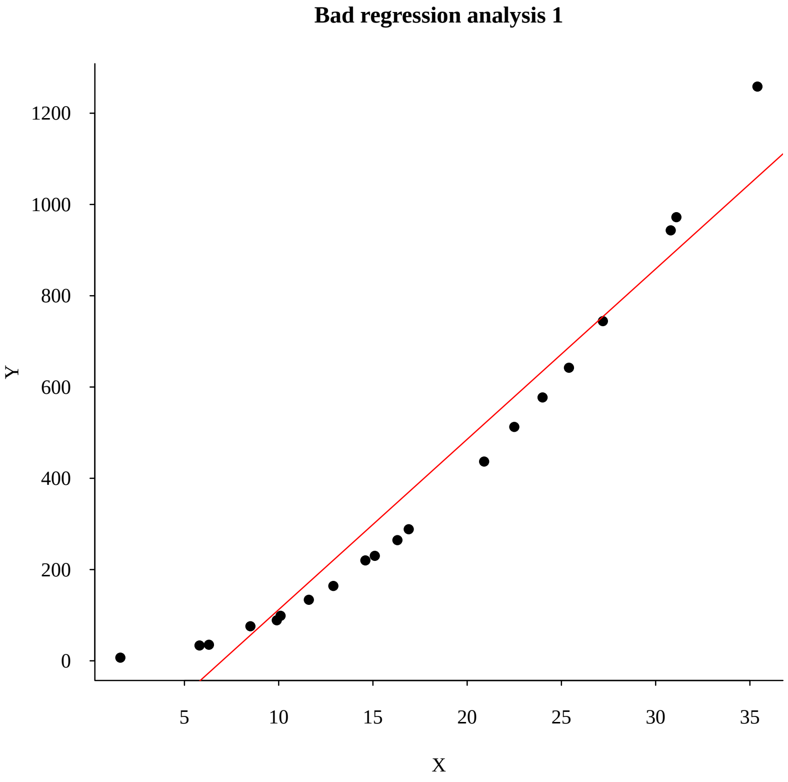
- $R^2$ (in this example, $0.550$)

  $55.0\%$ of variation in $Y$ is explained by $X$. $R^2$ is square of $r$ (Pearson's correlation of $X$ and $Y$)

```
cor(radiation$QoL0, radiation$QoL6)

[1] 0.7418

cor(radiation$QoL0, radiation$QoL6)^2

[1] 0.5502
```

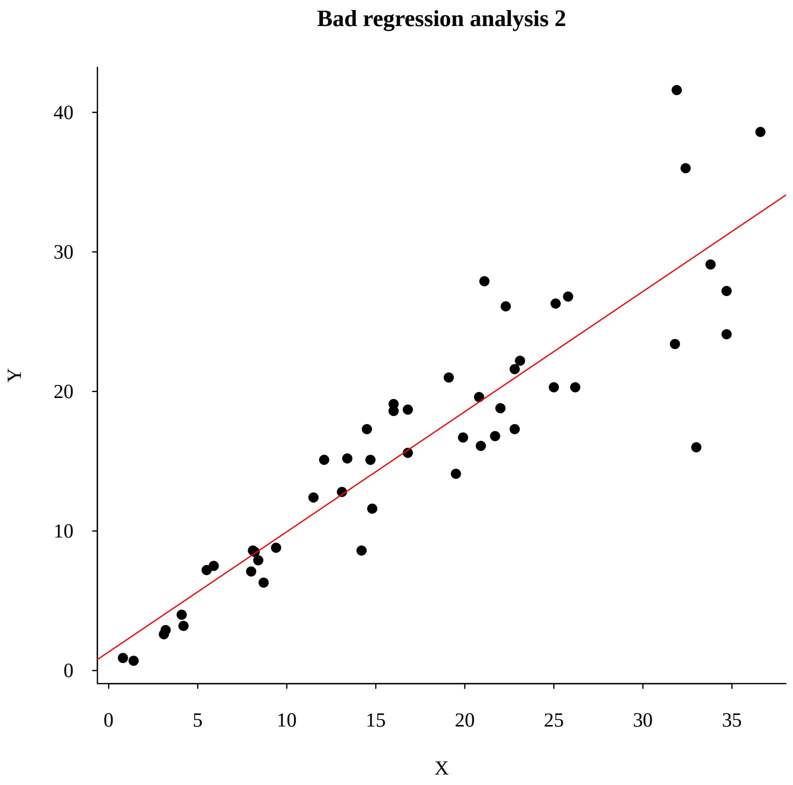
### 10.2.1 Model diagnostics

Not all regression models are good.



Bad regression analysis 1

```
mod.bad1

Linear Regression Model

 ols(formula = yb1 ~ xb1)

                 Model Likelihood     Discrimination
                    Ratio Test           Indexes
 Obs       20    LR chi2     56.92    R2       0.942
 sigma90.0104    d.f.            1    R2 adj   0.939
 d.f.      18    Pr(> chi2) 0.0000    g      413.140

 Residuals

    Min     1Q Median     3Q    Max
 -82.87 -64.41 -28.85  56.31 208.35


           Coef      S.E.    t     Pr(>|t|)
 Intercept -261.0750 42.9050 -6.08 <0.0001
 xb1         37.3278  2.1846 17.09 <0.0001
```



```
mod.bad2

Linear Regression Model

 ols(formula = yb2 ~ xb2)

                Model Likelihood     Discrimination
                   Ratio Test           Indexes
 Obs       50    LR chi2     81.54    R2        0.804
 sigma4.2336    d.f.            1    R2 adj    0.800
 d.f.      48    Pr(> chi2) 0.0000    g         9.815

 Residuals

      Min       1Q   Median       3Q      Max
 -13.7445  -2.3158  -0.4114   2.7599  12.8026


          Coef   S.E.   t     Pr(>|t|)
 Intercept 1.3329 1.2339  1.08 0.2855
 xb2       0.8610 0.0613 14.04 <0.0001
```

Assumptions for regression analysis

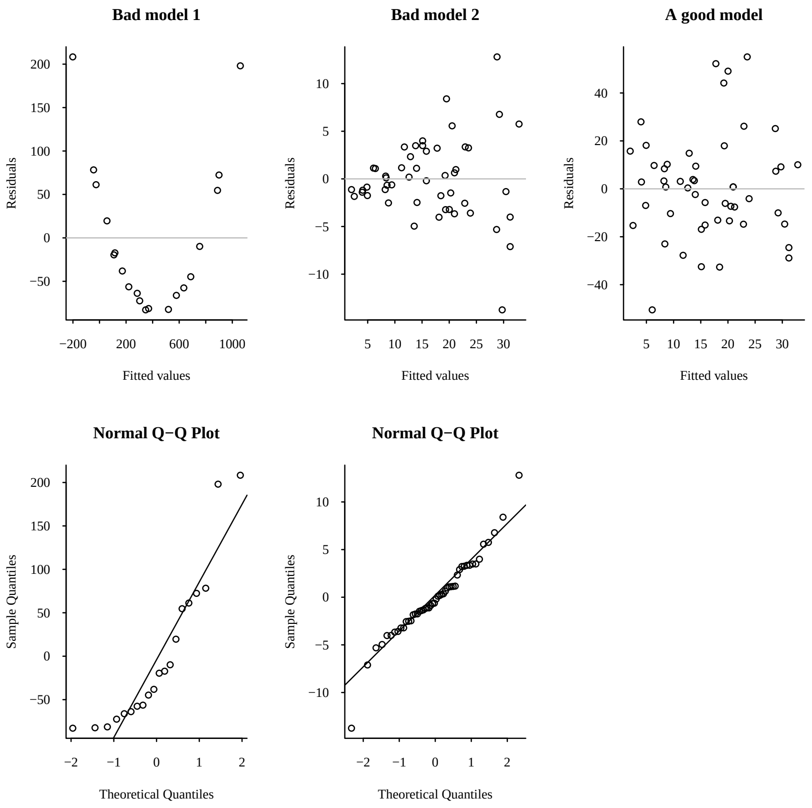
- Observations are independent. (Single observation from an individual)
- Residuals ($\varepsilon$) are $Normal(0, \sigma^2)$.

  Mean $0$ with a constant variance for all values of $X$.

  - Note that $y$ does not have to be normally distributed. (only $y - \hat{y}$)
  - Heteroskedasticity refers to the situation where variance increases as $x$ increases.
  - We use a residual plot to examine this assumption.

    A plot of residuals against $x$ or $\hat{y}$.

- When these assumptions are not met, inference (on coefficients and predicted values) will be wrong.



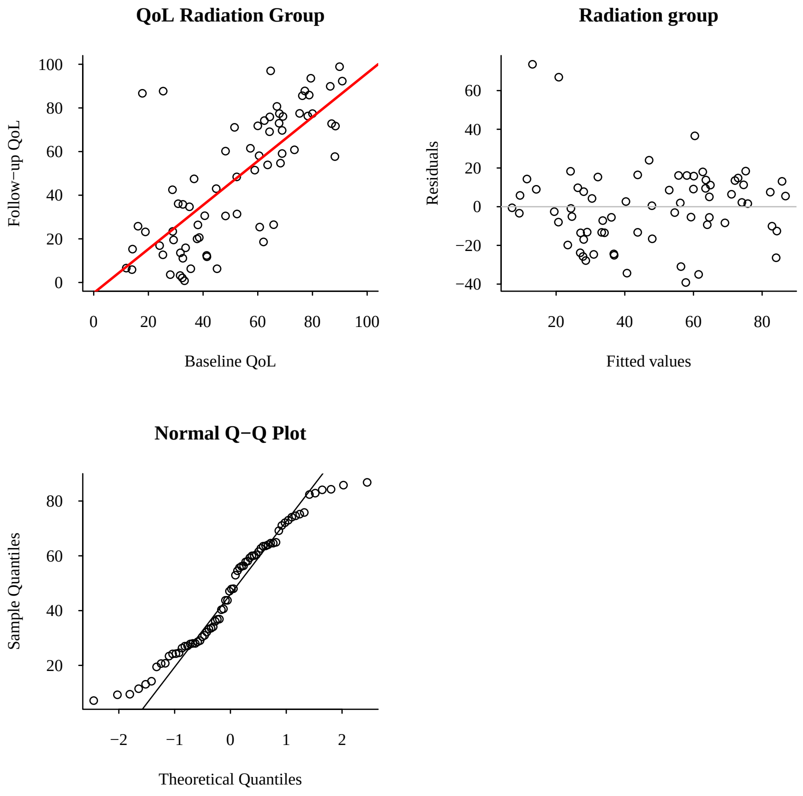
### 10.2.2 CEASAR example

QoL model for CEASAR radiation group.

```
(mod1 <- ols(QoL6 ~ QoL0, data = radiation, x = TRUE))

Linear Regression Model

 ols(formula = QoL6 ~ QoL0, data = radiation, x = TRUE)

                 Model Likelihood     Discrimination
                    Ratio Test           Indexes
 Obs       70    LR chi2     55.93    R2       0.550
 sigma20.2780    d.f.            1    R2 adj   0.544
 d.f.      68    Pr(> chi2) 0.0000    g       25.750

 Residuals

     Min      1Q  Median      3Q     Max
 -39.171 -13.228   1.046  11.262  73.590


           Coef    S.E.   t     Pr(>|t|)
 Intercept -4.8350 6.1817 -0.78 0.4368
 QoL0       1.0082 0.1105  9.12 <0.0001
```

#### Model diagnosis

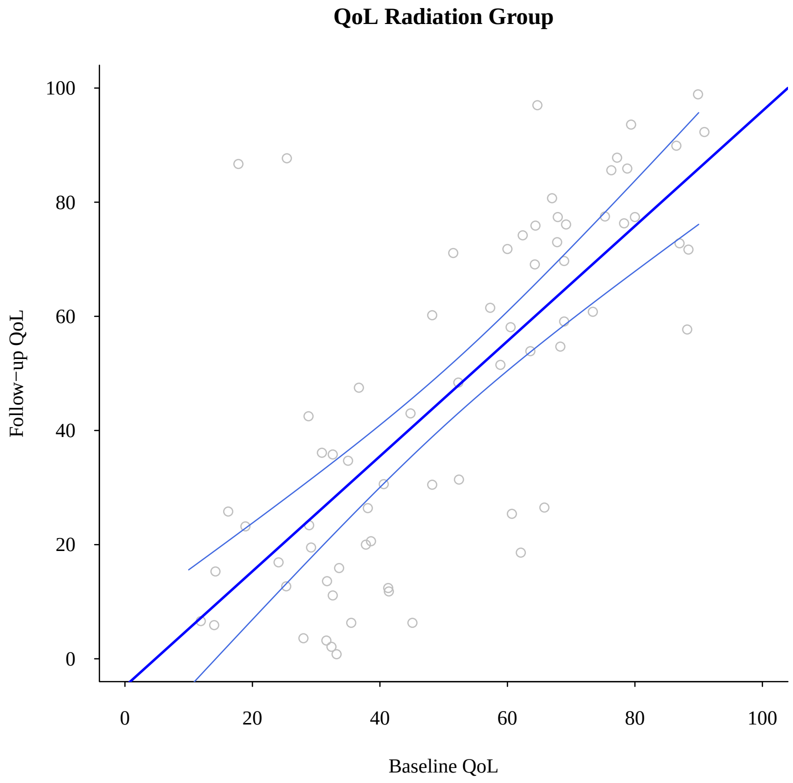
```
par(mfrow = c(2, 2), las = 1, family = "serif", bty = "L", tcl = -0.2)
plot(radiation$QoL0, radiation$QoL6, xlim = c(0, 100), ylim = c(0, 100), xlab = "Baseline QoL",
    ylab = "Follow-up QoL", main = "QoL Radiation Group")

abline(mod1, col = "red", lwd = 2)

plot(mod1$fitted, mod1$residuals, xlab = "Fitted values", ylab = "Residuals", main = "Radiation group")
abline(h = 0, col = 8)

qqnorm(mod1$fitted)
qqline(mod1$fitted)
```

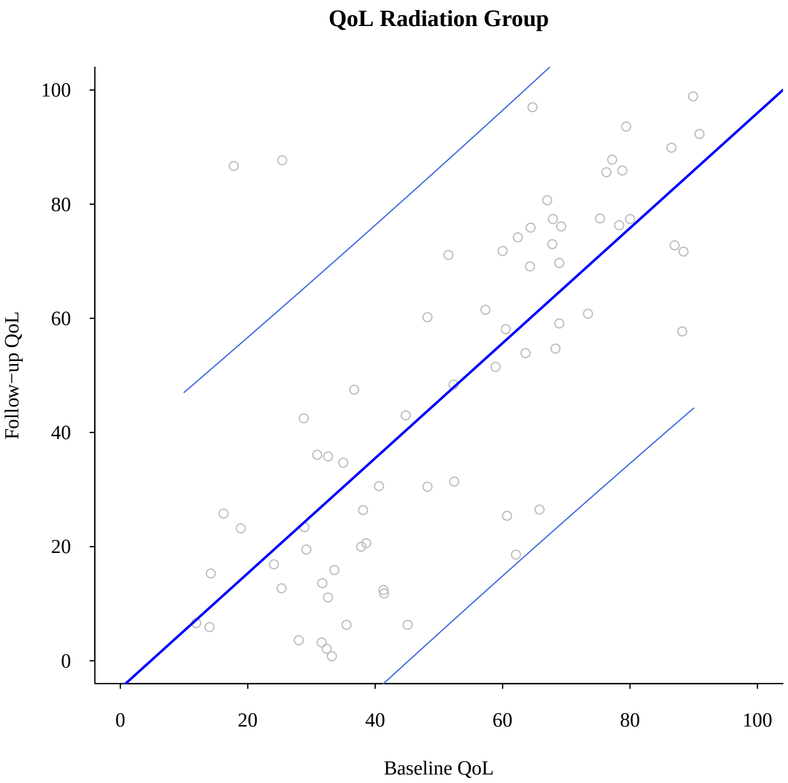


## Prediction intervals

```
newX <- seq(10, 90)
pred1 <- predict(mod1, newdata = data.frame(QoL0 = newX), conf.int = 0.95, conf.type = "mean")
pred2 <- predict(mod1, newdata = data.frame(QoL0 = newX), conf.int = 0.95, conf.type = "individual")

par(mfrow = c(1, 1), las = 1, family = "serif", bty = "L", tcl = -0.2)

plot(radiation$QoL0, radiation$QoL6, xlim = c(0, 100), ylim = c(0, 100), col = "grey", xlab = "Baseline
    ylab = "Follow-up QoL", main = "QoL Radiation Group")
abline(mod1, col = "blue", lwd = 2)

lines(newX, pred1$lower, col = "royalblue")
lines(newX, pred1$upper, col = "royalblue")

par(mfrow = c(1, 1), las = 1, family = "serif", bty = "L", tcl = -0.2)

plot(radiation$QoL0, radiation$QoL6, xlim = c(0, 100), ylim = c(0, 100), col = "grey", xlab = "Baseline
    ylab = "Follow-up QoL", main = "QoL Radiation Group")
abline(mod1, col = "blue", lwd = 2)
```

```
lines(newX, pred2$lower, col = "royalblue")
lines(newX, pred2$upper, col = "royalblue")
```





# END of BIOSTATISTICS I

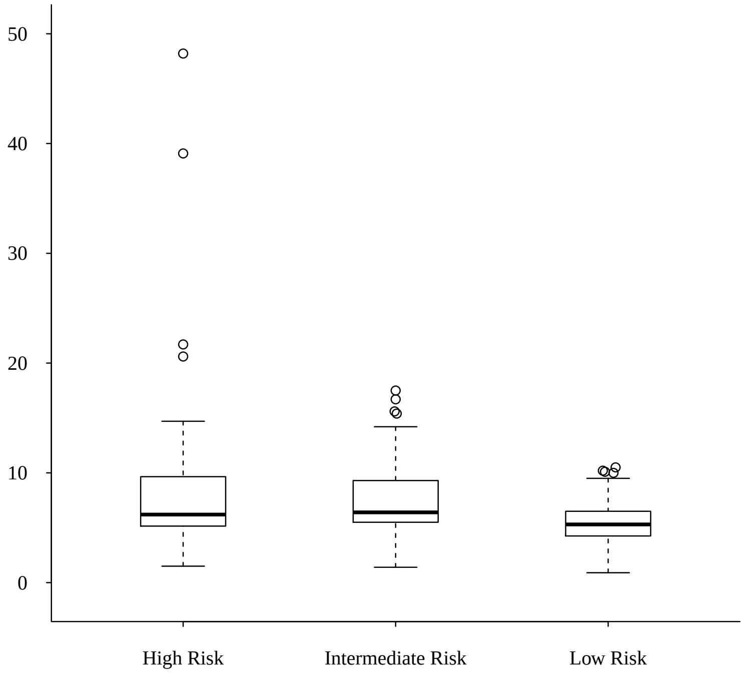
## 10.3 ANOVA as a regression analysis

- To compare $3$ or more groups on a continuous variable, analysis of variance (ANOVA) is used.
- A regression model with indicator variables (dummy variables) can accomplish the same goal.
- Indicator variable takes either $1$ or $0$.

Example: Is the baseline PSA different across different risk groups?

```
table(d$Risk)


        High Risk Intermediate Risk          Low Risk
               35                81                84

mean(d$PSA[d$Risk == "High Risk"])

[1] 9.697

mean(d$PSA[d$Risk == "Intermediate Risk"])

[1] 7.58

mean(d$PSA[d$Risk == "Low Risk"])

[1] 5.552
```

```
source("http://biostat.mc.vanderbilt.edu/wiki/pub/Main/TatsukiRcode/RFunctions0.R")

par(las = 1, family = "serif", bty = "L", tcl = -0.2)
tplot(PSA ~ Risk, data = d, type = "b", jit = 0.01)
```



```
(anv <- ols(PSA ~ Risk, data = d))

Linear Regression Model

 ols(formula = PSA ~ Risk, data = d)

                Model Likelihood     Discrimination
                   Ratio Test           Indexes
 Obs      200    LR chi2      19.67    R2        0.094
 sigma4.7322    d.f.             2    R2 adj    0.084
 d.f.     197    Pr(> chi2) 0.0001    g         1.607

 Residuals

     Min       1Q  Median       3Q      Max
 -8.1971 -2.2802 -0.8163  0.9517 38.5029


                        Coef    S.E.   t     Pr(>|t|)
 Intercept               9.6971 0.7999 12.12 <0.0001
 Risk=Intermediate Risk -2.1169 0.9572 -2.21 0.0282
 Risk=Low Risk          -4.1448 0.9521 -4.35 <0.0001
```

$$E[y] = \beta_0 + \beta_{Int} X_{Int} + \beta_{Low} X_{Low},$$

where

$$X_{Int} = 1 \text{ if Risk=Intermediate Risk, } 0 \text{ otherwise.}$$
$$X_{Low} = 1 \text{ if Risk=Low Risk, } 0 \text{ otherwise.}$$

so that

$$E[y] = \beta_0 \qquad \text{if Risk=High.}$$
$$E[y] = \beta_0 + \beta_{Int} \qquad \text{if Risk=Intermediate.}$$
$$E[y] = \beta_0 + \beta_{Low} \qquad \text{if Risk=Low.}$$

The High risk group's mean is $\hat{\beta}_0 = 9.70$. The Intermediate group's mean is $\hat{\beta}_0 + \hat{\beta}_1 = 7.58$. The Low risk group's mean is $\hat{\beta}_0 + \hat{\beta}_2 = 5.55$. Confidence intervals are:

```
confint(anv)

                         2.5 %  97.5 %
Intercept                8.120 11.2746
Risk=Intermediate Risk -4.005 -0.2292
Risk=Low Risk          -6.022 -2.2672
```

- Both $\beta_{Int}$ and $\beta_{Low}$ are significantly different from $0$, which means that the respective group averages are different from High-Risk group's average.
- How about Intermediate vs. Low?

  $H_0 : \beta_{Int} = \beta_{Low}$

  Another way of testing group differences are:

```
summary(anv, Risk = "High Risk")  ## only works with 'rms' functions. (ols)

             Effects              Response : PSA

 Factor                             Low High Diff. Effect S.E.   Lower 0.95 Upper 0.95
 Risk - Intermediate Risk:High Risk 1   2    NA    -2.117 0.9572 -4.005     -0.2292
 Risk - Low Risk:High Risk          1   3    NA    -4.145 0.9520 -6.022     -2.2672
```

```
summary(anv, Risk = "Low Risk")  ## Now reference group = 'Low Risk'

             Effects              Response : PSA

 Factor                            Low High Diff. Effect S.E.   Lower 0.95 Upper 0.95
 Risk - High Risk:Low Risk         3   1    NA    4.145  0.9520 2.2672     6.022
 Risk - Intermediate Risk:Low Risk 3   2    NA    2.028  0.7369 0.5746     3.481
```

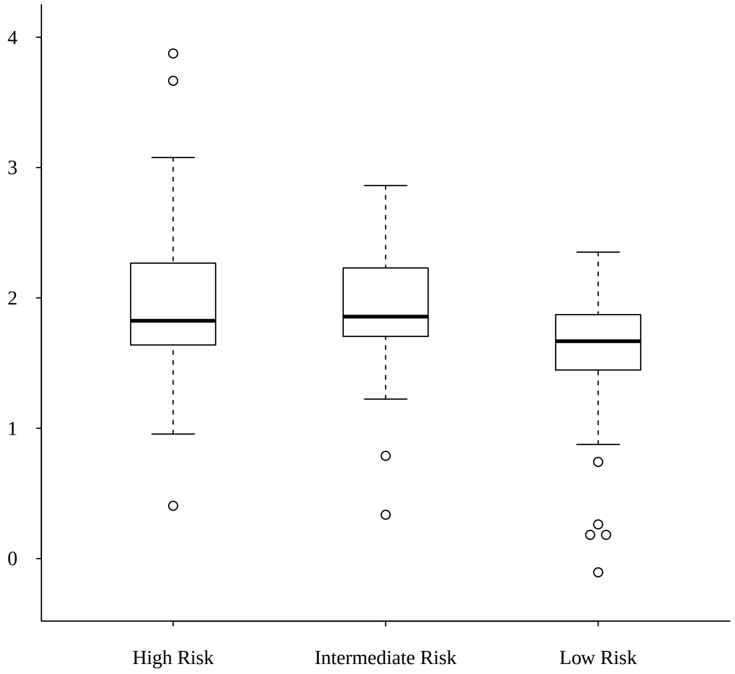
- How about the model assumptions, i.e., residuals are $Normal(0,\sigma)$.

  In ANOVA, this is equivalent to the *data* are normal with constant variance.

  For these data, the assumptions are probably not met.

To alleviate some problems with data distribution, a simple transformation (e.g., $log(\cdot)$ and $\sqrt{\cdot}$) sometimes works.

```
par(las = 1, family = "serif", tcl = -0.2, bty = "L")
tplot(log(PSA) ~ Risk, data = d, type = "b")
```



```
(anv2 <- ols(log(PSA) ~ Risk, data = d))

Linear Regression Model

 ols(formula = log(PSA) ~ Risk, data = d)

                Model Likelihood     Discrimination
                   Ratio Test           Indexes
 Obs      200    LR chi2     20.96    R2        0.099
 sigma0.5038    d.f.             2    R2 adj    0.090
 d.f.     197    Pr(> chi2) 0.0000    g         0.172

 Residuals

      Min        1Q   Median       3Q      Max
 -1.72588 -0.27281 -0.04169  0.27133  1.87864


                        Coef    S.E.   t     Pr(>|t|)
 Intercept               1.9967 0.0852 23.45 <0.0001
 Risk=Intermediate Risk -0.0604 0.1019 -0.59 0.5538
 Risk=Low Risk          -0.3762 0.1014 -3.71 0.0003


confint(anv2)

                          2.5 %  97.5 %
Intercept                1.8288  2.1647
Risk=Intermediate Risk -0.2614  0.1405
Risk=Low Risk          -0.5761 -0.1763
```

## 10.4 Multivariable regression models

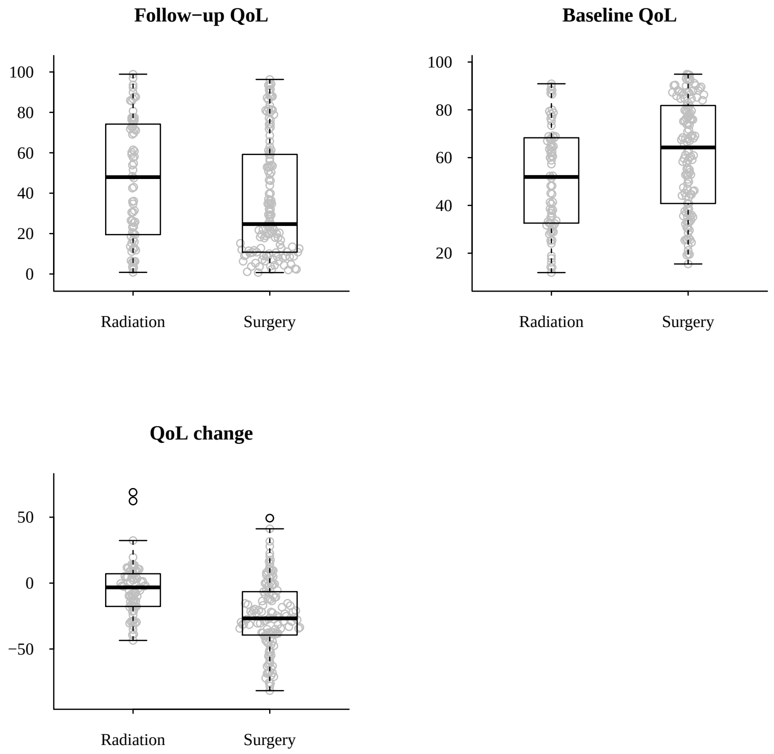
### 10.4.1 Confounding

- Recall that a confounder is a variable that is not of major interest but may be associated with response and/or independent variables.
- When there are confounders, the true association between $X$ (Explanatory variable) and $Y$ (Response) may not be analyzed properly without accounting for the confounders.
- A multivariable (multiple) regression model is a very effective method to account for confounders.
- We will be able to say, “effect of $X$ on $Y$ is this adjusting for this, that, and that.”

CEASAR Example: Comparing the follow-up QoL between the treatments.

- We would like to compare the follow-up QoL score between Surgery and Radiation group.
- We think (know by now) that many baseline characteristics are different between groups, so not adjusting for them will lead to biased conclusions.

```
par(mfrow = c(2, 2), las = 1, family = "serif", bty = "l", tcl = -0.2)
tplot(QoL6 ~ Treatment, data = d, type = "bd", jit = 0.01, main = "Follow-up QoL")
tplot(QoL0 ~ Treatment, data = d, type = "bd", jit = 0.01, main = "Baseline QoL")
tplot(I(QoL6 - QoL0) ~ Treatment, data = d, type = "bd", jit = 0.01, main = "QoL change")
```



```
(q6 <- aggregate(d$QoL6, by = list(d$Treatment), FUN = summary))

    Group.1 x.Min. x.1st Qu. x.Median x.Mean x.3rd Qu. x.Max.
1 Radiation   0.80     19.62    47.95  47.03     73.90  98.90
2   Surgery   0.70     10.80    24.70  36.30     59.20  96.30

(q0 <- aggregate(d$QoL0, by = list(d$Treatment), FUN = summary))

    Group.1 x.Min. x.1st Qu. x.Median x.Mean x.3rd Qu. x.Max.
1 Radiation  11.90     32.60    51.90  51.44     68.20  90.90
2   Surgery  15.50     41.47    64.25  61.35     81.77  94.90

(qd <- aggregate(d$QoL6 - d$QoL0, by = list(d$Treatment), FUN = summary))

    Group.1  x.Min. x.1st Qu. x.Median  x.Mean x.3rd Qu.  x.Max.
1 Radiation -43.500   -17.700   -3.250  -4.416     7.050  68.900
2   Surgery -81.600   -39.250  -26.750 -25.050    -6.775  49.300
```

### 10.4.2 Multivariable linear regression

aka Multiple linear regression.

Don't say "multivariate".

- $p$ independent variables, $x_1$, $x_2$, $\cdots$, $x_p$.
- Model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_p x_p + \varepsilon$
- Estimated equation: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots \hat{\beta}_p x_p$
- Assume that the difference variables act in an additive fashion (later *interaction*).
- Each $\beta$ represents an effect of increasing a variable by one unit, holding all others constant.

Example:

$$Y = \beta_0 + \beta_1 X_{age} + \beta_2 X_{sex} + \varepsilon,$$

where $X_{sex} = 0$ if male, $1$ if female.

- $\beta_1$ represents the change in the mean of $Y$ for males when ages increases by $1$ year. It is also the change in the mean of $Y$ for females.
- $\beta_2$ is the "female effect". The difference in the means of $Y$ for two subjects of the same age. (Any age).

$$E[Y|x_1, x_2] = \beta_0 + \beta_1 x_1 \text{ for male.}$$
$$E[Y|x_1, x_2] = \beta_0 + \beta_2 + \beta_1 x_1 \text{ for female.}$$

CEASAR example:

- We will analyze association between QoL6 (Y) and Treatment (X).

- We will adjust for the following baseline variables.

  QoL0, Age, PSA, Race, HeartDisease, Hypertension, Diabetes.

### Variables to include

- The number of variables that can be included in a multivariable regression model depends on sample size.
- The rule of thumb is $n = 15$ is required for $1$ degree of freedom.
  - A continuous variable uses $1$ degree of freedom.
  - A categorical variable uses $C-1$ degree of freedom, where $C$ is number of categories.
  - Flexible form (spline) and interactions need additional df.
- **Important**: We need to specify which variables to include in the model a priori.
  - We cannot look at a "Table 1" to decide which variables to account for. (univariate screening)
  - Doing so will inflate type I error rate and will result in overstating the association between $X$ and $Y$.

```
latex(summary(Treatment ~ Age + Race + MaritalStatus + Education + Income + PSA + Gleason +
    HeartDisease + Hypertension + Diabetes + QoL0 + QoL6, method = "reverse", overall = !TRUE,
    test = !FALSE, data = d), file = "")

Warning in chisq.test(tab, correct = FALSE): Chi-squared approximation may be incorrect
```

```
(qolM <- ols(QoL6 ~ Treatment + QoL0 + Age + Race + PSA + HeartDisease + Hypertension + Diabetes,
    data = d))

Linear Regression Model

 ols(formula = QoL6 ~ Treatment + QoL0 + Age + Race + PSA + HeartDisease +
     Hypertension + Diabetes, data = d)
```

Table 10.1: Descriptive Statistics by Treatment

| | N | Radiation | Surgery | Test Statistic |
|---|---|---|---|---|
| | | $N = 70$ | $N = 130$ | |
| Age | 200 | 63.0 67.0 73.0 | 56.0 63.5 68.0 | $F_{1,198}$=11.81, P<0.001[1] |
| Race : White | 200 | 69% (48) | 75% (97) | $\chi^2_2$=1.18, P=0.555[2] |
| Black | | 10% ( 7) | 10% (13) | |
| Other | | 21% (15) | 15% (20) | |
| MaritalStatus : Not married | 200 | 27% ( 19) | 10% ( 13) | $\chi^2_1$=9.95, P=0.002[2] |
| Education : High school | 200 | 47% (33) | 25% (33) | $\chi^2_3$=11.62, P=0.009[2] |
| Some college | | 21% (15) | 23% (30) | |
| College graduate | | 20% (14) | 26% (34) | |
| Graduate school | | 11% ( 8) | 25% (33) | |
| Income : - 30K | 195 | 34% (23) | 15% (19) | $\chi^2_3$=13.22, P=0.004[2] |
| 30K - 50K | | 24% (16) | 20% (26) | |
| 50K - 100K | | 25% (17) | 31% (40) | |
| 100K - | | 16% (11) | 34% (43) | |
| PSA | 200 | 5.00 6.30 9.05 | 4.70 5.80 7.45 | $F_{1,198}$=2.69, P=0.103[1] |
| Gleason : 6 or less | 200 | 46% (32) | 47% (61) | $\chi^2_3$=3.84, P=0.279[2] |
| 3 + 4 | | 44% (31) | 34% (44) | |
| 4 + 3 | | 4% ( 3) | 7% ( 9) | |
| 8,9,10 | | 6% ( 4) | 12% (16) | |
| HeartDisease : Yes | 200 | 9% ( 6) | 15% ( 19) | $\chi^2_1$=1.52, P=0.218[2] |
| Hypertension : Yes | 200 | 67% (47) | 52% (68) | $\chi^2_1$=4.1, P=0.043[2] |
| Diabetes : Yes | 200 | 17% ( 12) | 18% ( 24) | $\chi^2_1$=0.05, P=0.817[2] |
| QoL0 | 200 | 32.60 51.90 68.20 | 41.47 64.25 81.77 | $F_{1,198}$=8.98, P=0.003[1] |
| QoL6 | 200 | 19.62 47.95 73.90 | 10.80 24.70 59.20 | $F_{1,198}$=5.82, P=0.017[1] |

$a$ $b$ $c$ 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]Wilcoxon test; [2]Pearson test

```
                 Model Likelihood     Discrimination
                    Ratio Test           Indexes
Obs       200    LR chi2    101.45    R2       0.398
sigma23.7485    d.f.            9    R2 adj   0.369
d.f.      190    Pr(> chi2) 0.0000    g       21.676

Residuals

    Min       1Q  Median       3Q      Max
-51.839 -15.005  -3.115  17.023  73.087


                   Coef     S.E.    t     Pr(>|t|)
Intercept           32.1713 17.0368  1.89 0.0605
Treatment=Surgery -18.2629  3.7140 -4.92 <0.0001
QoL0                0.7244  0.0794  9.13 <0.0001
Age                -0.3901  0.2215 -1.76 0.0798
Race=Black          9.4428  5.7491  1.64 0.1021
Race=Other         -2.7699  4.5562 -0.61 0.5440
PSA                 0.5166  0.3516  1.47 0.1434
HeartDisease=Yes   -7.3865  5.1523 -1.43 0.1533
Hypertension=Yes    1.2876  3.5585  0.36 0.7179
Diabetes=Yes       -6.5694  4.4583 -1.47 0.1423
```

- Overall significance?

- $R^2$

- Significant Treatment effect?

- Other significant covariates?

- On average, QoL6 is higher for Radiation group by ...?

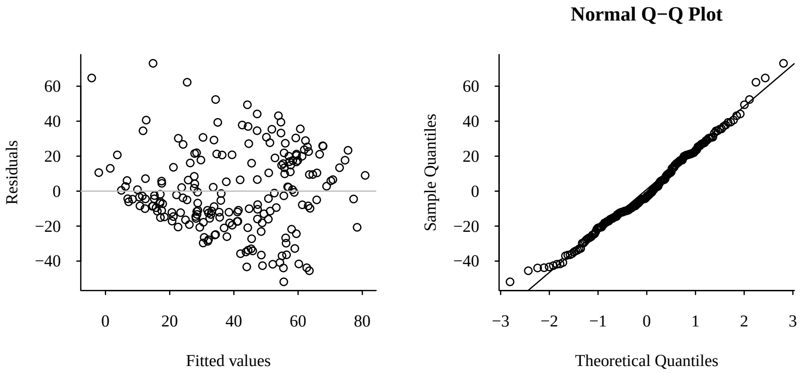
- Compared to a white person, a black person’s QoL6 is higher by ....?

- Compared to a $75$ years old, an $85$ years old’s QoL6 is ...?

```
confint(qolM)
```

```
                      2.5 %    97.5 %
Intercept            -1.4343  65.77687
Treatment=Surgery   -25.5890 -10.93689
QoL0                  0.5679   0.88098
Age                  -0.8271   0.04682
Race=Black           -1.8975  20.78313
Race=Other          -11.7572   6.21740
PSA                  -0.1769   1.21021
HeartDisease=Yes    -17.5495   2.77652
Hypertension=Yes     -5.7317   8.30690
Diabetes=Yes        -15.3635   2.22469
```

“QoL6 for Surgery group is, on average, $18.3$ points lower ($95\%$ CI: $10.9$ to $25.6$) compared to Radiation group.”



```
summary(qolM)

             Effects              Response : QoL6
```

```
Factor                            Low   High  Diff. Effect S.E.  Lower 0.95 Upper 0.95
QoL0                              36.62 78.22 41.6  30.136 3.302  23.6230   36.6490
Age                               58.00 70.00 12.0  -4.681 2.658  -9.9249    0.5619
PSA                                4.80  8.10  3.3   1.705 1.160  -0.5839    3.9937
Treatment - Radiation:Surgery      2.00  1.00   NA  18.263 3.714  10.9370   25.5890
Race - Black:White                 1.00  2.00   NA   9.443 5.749  -1.8975   20.7830
Race - Other:White                 1.00  3.00   NA  -2.770 4.556 -11.7570    6.2174
HeartDisease - Yes:No              1.00  2.00   NA  -7.386 5.152 -17.5500    2.7765
Hypertension - No:Yes              2.00  1.00   NA  -1.288 3.558  -8.3069    5.7317
Diabetes - Yes:No                  1.00  2.00   NA  -6.569 4.458 -15.3640    2.2247
```

```
summary(qolM, est.all = FALSE, QoL0 = c(30, 60, 80), Age = c(60, 65, 70), Treatment = "Radiation",
    Race = "White", HeartDisease = "No", Hypertension = "No", Diabetes = "No")

             Effects              Response : QoL6

 Factor                            Low High Diff. Effect  S.E.  Lower 0.95 Upper 0.95
 QoL0                              30  80   50     36.221 3.969  28.393     44.0490
 Age                               60  70   10     -3.901 2.215  -8.271      0.4682
 Treatment - Surgery:Radiation      1   2   NA    -18.263 3.714 -25.589    -10.9370
 Race - Black:White                 1   2   NA      9.443 5.749  -1.897     20.7830
 Race - Other:White                 1   3   NA     -2.770 4.556 -11.757      6.2174
 HeartDisease - Yes:No              1   2   NA     -7.386 5.152 -17.550      2.7765
 Hypertension - Yes:No              1   2   NA      1.288 3.558  -5.732      8.3069
 Diabetes - Yes:No                  1   2   NA     -6.569 4.458 -15.364      2.2247
```

“QoL6 is $36.2$ $(28.4, 44.0)$ points higher for someone with $QoL0 = 80$ than someone with $QoL0 = 30$.

```
qolM$coefficients["QoL0"]

  QoL0
0.7244

qolM$coefficients["QoL0"] * 50

 QoL0
36.22
```

### 10.4.3 Interactions

- In the last model, we assumed that the effect of Treatment is the same for everybody regardless of their baseline characteristics.

- A model with an interaction allows an effect of one $X$ vary depending the value of another $X$.
- Subgroup analysis can be conducted using interactions.

A simple example.

$$Y = \beta_0 + \beta_1 X_{age} + \beta_2 X_{sex} + \beta_{12} X_{age} X_{sex} + \varepsilon,$$

where $X_{sex} = 0$ if male, $1$ if female.

- For male ($X_{sex} = 0$), we have $Y = \beta_0 + \beta_1 X_{age} + \varepsilon$.
- For female ($X_{sex} = 1$), we have $Y = \beta_0 + \beta_2 + (\beta_1 + \beta_{12}) X_{age} + \varepsilon$.
- Thus, $\beta_1$ is the age effect for male.
- $\beta_{12}$ is the additional age effect for female.
- $\beta_0$ is average $Y$ when $X_{age} = 0$ for male.
- $\beta_2$ is difference of average $Y$'s when $X_{age} = 0$ between male and female.
- We can test whether the age effect is the same for male and female by testing $\beta_{12} = 0$.

A new CEASAR model.

Now we consider the interactions between Treatment and the following variables: QoL0, Age, Race.

The reasons for including these interactions is:

- We think that the Treatment effect on QoL6 is different according to the values of QoL0, Age, and Race.
- We would like to analyze (investigate) the Treatment effect for, e.g., blacks and whites separately.

Do we have enough data to include all these interactions? Degrees of freedoms spent:

Treatment (1), QoL0 (1), Age (1), Race (2), PSA (1), HeartDisease (1), Hypertension (1), Diabetes (1)

Treatment - QoL0 (1), Treatment - Age (1), Treatment -Race (2) So the total df is $13$. We probably need about $n = 13 \times 15 = 195$.

```
(qolM.int <- ols(QoL6 ~ Treatment * (QoL0 + Age + Race) + PSA + HeartDisease + Hypertension +
    Diabetes, data = d))

Linear Regression Model

 ols(formula = QoL6 ~ Treatment * (QoL0 + Age + Race) + PSA +
     HeartDisease + Hypertension + Diabetes, data = d)

                 Model Likelihood     Discrimination
                    Ratio Test           Indexes
 Obs      200    LR chi2    112.84    R2       0.431
 sigma23.3286    d.f.           13    R2 adj   0.391
 d.f.     186    Pr(> chi2) 0.0000    g       22.464

 Residuals

     Min      1Q  Median      3Q     Max
 -52.253 -14.464  -3.292  15.222  72.142


                                Coef    S.E.    t     Pr(>|t|)
 Intercept                      -5.0609 24.9680 -0.20 0.8396
 Treatment=Surgery              56.4236 32.4887  1.74 0.0841
 QoL0                            1.0042  0.1334  7.53 <0.0001
 Age                            -0.0031  0.3478 -0.01 0.9929
 Race=Black                      4.8575  9.7740  0.50 0.6198
 Race=Other                     -1.7489  7.0026 -0.25 0.8031
 PSA                             0.4048  0.3489  1.16 0.2474
 HeartDisease=Yes               -8.7150  5.1224 -1.70 0.0905
 Hypertension=Yes               -1.6532  3.6611 -0.45 0.6521
 Diabetes=Yes                   -4.6510  4.4403 -1.05 0.2963
 Treatment=Surgery * QoL0       -0.4977  0.1716 -2.90 0.0042
 Treatment=Surgery * Age        -0.7457  0.4439 -1.68 0.0946
 Treatment=Surgery * Race=Black  9.5978 12.1310  0.79 0.4298
 Treatment=Surgery * Race=Other -1.1382  9.1542 -0.12 0.9012
```

- With an interaction, interpreting the coefficients gets complicated.

- Even though changing models based on the results of model fit is not recommended (it would increase type I error rate in inference.), we can check if the interaction terms were necessary.

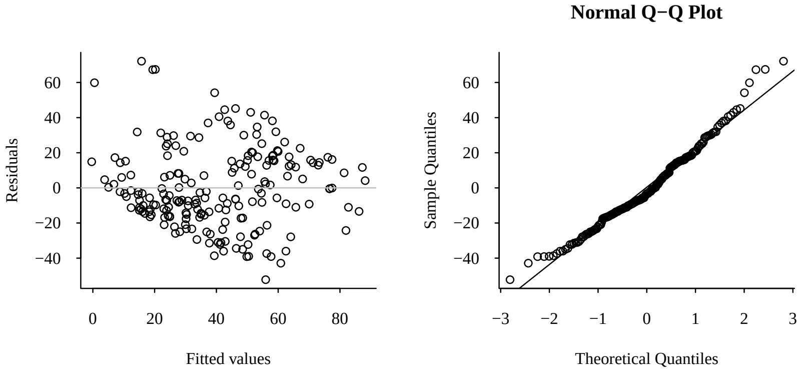
```
anova(qolM.int)

                Analysis of Variance          Response: QoL6

 Factor                                                d.f. Partial SS MS      F     P
 Treatment  (Factor+Higher Order Factors)                5   19569.7     3913.9  7.19 <.0001
  All Interactions                                       4    5932.6     1483.2  2.73 0.0308
 QoL0  (Factor+Higher Order Factors)                     2   45435.6    22717.8 41.74 <.0001
  All Interactions                                       1    4579.7     4579.7  8.42 0.0042
 Age  (Factor+Higher Order Factors)                      2    3842.2     1921.1  3.53 0.0313
  All Interactions                                       1    1535.9     1535.9  2.82 0.0946
 Race  (Factor+Higher Order Factors)                     4    2807.0      701.8  1.29 0.2757
  All Interactions                                       2     375.1      187.6  0.34 0.7089
 PSA                                                     1     732.7      732.7  1.35 0.2474
 HeartDisease                                            1    1575.3     1575.3  2.89 0.0905
 Hypertension                                            1     111.0      111.0  0.20 0.6521
 Diabetes                                                1     597.1      597.1  1.10 0.2963
 Treatment * QoL0  (Factor+Higher Order Factors)         1    4579.7     4579.7  8.42 0.0042
 Treatment * Age  (Factor+Higher Order Factors)          1    1535.9     1535.9  2.82 0.0946
 Treatment * Race  (Factor+Higher Order Factors)         2     375.1      187.6  0.34 0.7089
 TOTAL INTERACTION                                       4    5932.6     1483.2  2.73 0.0308
 REGRESSION                                             13   76736.7     5902.8 10.85 <.0001
 ERROR                                                 186  101225.4      544.2
```

Also *a likelihood ratio test* can be used to see if a bigger model is better than a smaller (nested) model significantly.

```
lrtest(qolM, qolM.int)


Model 1: QoL6 ~ Treatment + QoL0 + Age + Race + PSA + HeartDisease + Hypertension +
    Diabetes
Model 2: QoL6 ~ Treatment * (QoL0 + Age + Race) + PSA + HeartDisease +
    Hypertension + Diabetes

L.R. Chisq       d.f.          P
   11.3909     4.0000     0.0225
```

It looks like including these interactions are not a bad idea.

```
resid <- qolM.int$residuals
fitted <- qolM.int$fitted
par(mfrow = c(2, 2), las = 1, family = "serif", bty = "L", tcl = -0.2)
plot(fitted, resid, main = "", xlab = "Fitted values", ylab = "Residuals")
abline(h = 0, col = 8)

qqnorm(qolM.int$residuals)
qqline(qolM.int$residuals)
```



How to read / report this model.

- The variables not involved in any interactions can be interpreted in the same way as before: e.g.,
  - "On average, QoL6 increases by $0.4048$ with unit increment in PSA."
  - "On average, QoL6 is lower by $8.72$ for a patient with HeartDisease at baseline compared with a patient without.

- Use `summary()`

For Surgery group:

```
summary(qolM.int, Treatment = "Surgery", QoL0 = c(30, 60, 80), Age = c(60, 65, 70))

             Effects              Response : QoL6

 Factor                         Low  High Diff. Effect S.E.  Lower 0.95 Upper 0.95
 QoL0                           30.0 80.0 50.0  25.322 5.162  15.1370   35.506
 Age                            60.0 70.0 10.0  -7.488 2.818 -13.0480   -1.928
 PSA                             4.8  8.1  3.3   1.336 1.151  -0.9354    3.607
 Treatment - Radiation:Surgery   2.0  1.0   NA  21.908 4.326  13.3740   30.442
 Race - Black:White              1.0  2.0   NA  14.455 7.112   0.4240   28.487
 Race - Other:White              1.0  3.0   NA  -2.887 5.862 -14.4520    8.678
 HeartDisease - Yes:No           1.0  2.0   NA  -8.715 5.122 -18.8200    1.390
 Hypertension - No:Yes           2.0  1.0   NA   1.653 3.661  -5.5693    8.876
 Diabetes - Yes:No               1.0  2.0   NA  -4.651 4.440 -13.4110    4.109

Adjusted to: Treatment=Surgery QoL0=60 Age=65 Race=White
```

For Radiation group:

```
summary(qolM.int, Treatment = "Radiation", QoL0 = c(30, 60, 80), Age = c(60, 65, 70))

             Effects              Response : QoL6

 Factor                         Low  High Diff. Effect    S.E.  Lower 0.95 Upper 0.95
 QoL0                           30.0 80.0 50.0   50.20800 6.669  37.0510    63.366
 Age                            60.0 70.0 10.0   -0.03119 3.478  -6.8920     6.830
 PSA                             4.8  8.1  3.3    1.33580 1.151  -0.9354     3.607
 Treatment - Surgery:Radiation   1.0  2.0   NA  -21.90800 4.326 -30.4420   -13.374
 Race - Black:White              1.0  2.0   NA    4.85750 9.774 -14.4250    24.140
 Race - Other:White              1.0  3.0   NA   -1.74890 7.003 -15.5640    12.066
 HeartDisease - Yes:No           1.0  2.0   NA   -8.71500 5.122 -18.8200     1.390
 Hypertension - No:Yes           2.0  1.0   NA    1.65320 3.661  -5.5693     8.876
 Diabetes - Yes:No               1.0  2.0   NA   -4.65100 4.440 -13.4110     4.109

Adjusted to: Treatment=Radiation QoL0=60 Age=65 Race=White
```

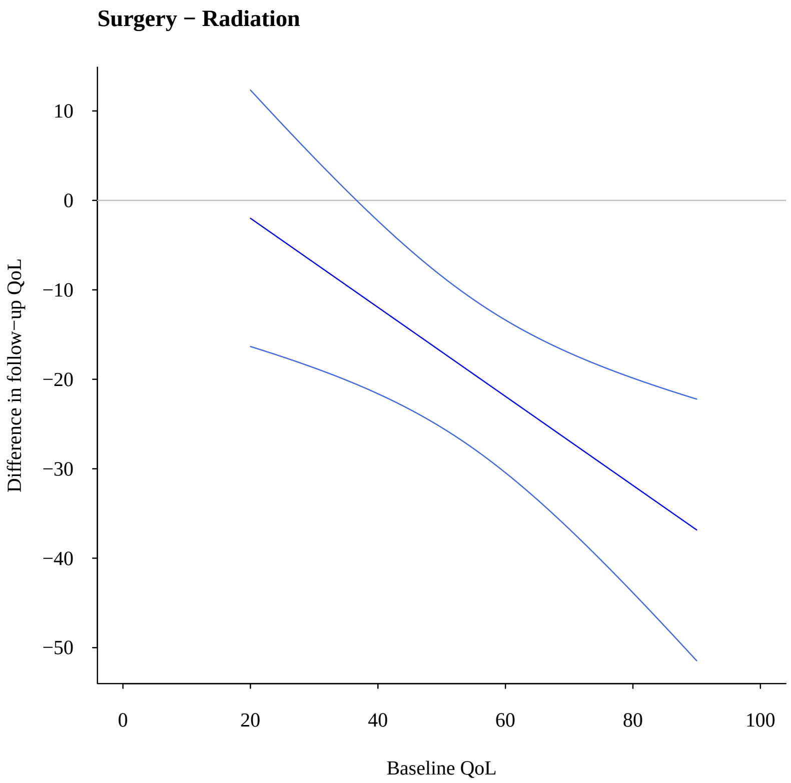
The Surgery effect is $-21.9\,(-30.4, -13.4)$ for a $65$-year old white patient with $QoL0 = 60$.

```
summary(qolM.int, Treatment = "Radiation", QoL0 = c(30, 40, 80), Age = c(60, 75, 80), est.all = FALSE)
```

```
             Effects               Response : QoL6

 Factor                         Low High Diff. Effect    S.E.  Lower 0.95 Upper 0.95
 QoL0                           30  80   50     50.20800 6.669  37.05     63.366
 Age                            60  80   20     -0.06238 6.955 -13.78     13.659
 Treatment - Surgery:Radiation   1   2   NA    -19.41000 6.113 -31.47     -7.351

Adjusted to: Treatment=Radiation QoL0=40 Age=75
```

The Surgery effect is $-19.4\,(-31.5,\,-7.4)$ for a $75$-year old white patient with $QoL0 = 40$. And finally, we can estimate the treatment effects for different values of age / baseline QoL / Race and show them in a plot.

```
extract.effect<- function(qol0, age, race='White', m=qolM.int){
    s <- summary(m, QoL0=qol0, Age=age, Race=race, Treatment='Radiation')
    out <- s[ grep('Treatment', row.names(s)), c(4,6,7)]
    }

Qol0 <- seq(20,90) ; L <- length(Qol0)
Age <- 65

out <- matrix(0, ncol=3, nrow=L)
for(i in 1:L){
    out[i,] <- extract.effect(qol0=Qol0[i], age=Age)
    }

par(las=1, family='serif', bty='L', tcl=-0.2)
plot(0,0, type='n', xlim=c(0,100), ylim=range(out), xlab='Baseline QoL',
        ylab='Difference in follow-up QoL', main='')
title(main='Surgery - Radiation', adj=0)
abline(h=0, col=8)

lines(Qol0, out[,1], col='blue')
lines(Qol0, out[,2], col='royalblue')
lines(Qol0, out[,3], col='royalblue')
```


Surgery − Radiation
Difference in follow−up QoL
10
0
−10
−20
−30
−40
−50
0
20
40
60
80
100
Baseline QoL