

## McNemar's Test and Introduction to ANOVA

---

Recall from the last lecture on nonparametric methods we used the example of reductions of forced vital capacity.

	FVC Reduc (ml)	
Subj	Placebo	Drug
1	224	213
2	80	95
3	75	33
4	541	440
5	74	-32
6	85	-28

At first blush, this looks like a two-sample problem. However, the FVC reductions under placebo and under drug are both observed for each subject. We have 6 pairs of measurements. Thus, by looking at the differences, placebo minus drug, we reduce it to a one-sample problem.

	FVC Reduc (ml)		
Subj	Placebo	Drug	Difference
1	224	213	11
2	80	95	-15
3	75	33	42
4	541	440	101
5	74	-32	106
6	85	-28	113

## McNemar's Test and Introduction to ANOVA

---

Assuming you had paired outcomes X and Y, like with the FVC reductions and that the means were roughly normally distributed, would it be wrong to use a two-sample t-test on the difference in means?

Unequal variance t-test:

$$T^* = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}}$$

If X and Y are not independent,

$$Var(\bar{X} - \bar{Y}) = \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m} - 2 \cdot Cov(\bar{X}, \bar{Y})$$

What will happen with the two-sample t-test when X and Y are not independent?

# McNemar's Test and Introduction to ANOVA

---

## An Illustrative Example in R:

```
set.seed(1117) # set to any fixed random number for reproducible results
options(scipen=20) # prevents scientific notation
```

```
c <- 0.75 # choose correlation
vx = 4*c^2 / (1 - c^2) # variance needed for desired correlation
```

```
PosCorPvals2sample <- c()
PosCorPvalsPaired <- c()
NegCorPvals2sample <- c()
NegCorPvalsPaired <- c()
sims <- 10^4
for(i in 1:sims){
  x <- rnorm(100, 0, sqrt(vx))
  y <- x + rnorm(100, 0, 2)
  z <- -x + rnorm(100, 0, 2)
  PosCorPvals2sample <- c(PosCorPvals2sample, t.test(x,y)$p.value)
  PosCorPvalsPaired <- c(PosCorPvalsPaired, t.test(x-y)$p.value)
  NegCorPvals2sample <- c(NegCorPvals2sample, t.test(x,z)$p.value)
  NegCorPvalsPaired <- c(NegCorPvalsPaired, t.test(x-z)$p.value)
}
```

```
plot(x, y)
sum( PosCorPvals2sample<0.05 )/sims
sum( PosCorPvalsPaired<0.05 )/sims
```

```
plot(x, z)
sum( NegCorPvals2sample<0.05 )/sims
sum( NegCorPvalsPaired<0.05 )/sims
```

## Type I Error Rates

Correlation	two-sample un-equal variances	paired one-sample test
-0.99	0.1685	0.0474
-0.75	0.1362	0.0505
-0.50	0.0973	0.0522
-0.25	0.0628	0.0502
0.00	0.0476	0.0476
0.25	0.0359	0.0493
0.50	0.0115	0.0493
0.75	0.0003	0.0493
0.99	0.0000	0.0493

### **McNemar's test for Paired Dichotomous data**

McNemar's test is analogous to the paired t-test, except that is applied to only dichotomous data.

We learned that when we want to compare the means in two groups of paired observations, the standard t-tests are not valid. This was because the variance of the difference in means was derived under the assumptions that the two groups are independent and so it ignores the covariance-correlation term (which can increase or decrease that term). So in this case, both the hypothesis tests and confidence intervals that are based on this assumption will not provide the 'correct' answer.

Side note: By correct, I mean that test will reject more or less often than the specified type I error. In statistics, this is what we mean by correct: that the test has the properties we designed it to.

To solve this problem we reduced the data from two dimensions down to one, by subtracting the observations within a pair and analyzing the differences with a one sample t-test and confidence interval on the differences. This approach avoids the correlation problem because the estimated variance on the differences accounts for the correlation.

## McNemar's Test and Introduction to ANOVA

---

With dichotomous data the problem is more complicated, but a similar approach exists. It is called 'McNemar's test for paired binomial data'.

The basic idea is still to analyze the differences (because the difference in means, or in this case proportions, is still the average of the differences) and then use the variance of the differences instead of trying to estimate the correlation.

But rather than do this directly on the differences, we arrange everything in a 2x2 table. The catch is that we use a different test-statistic for this table.

### **Setup:**

N pairs of (zero or one) responses,  $(X_i, Y_i)$ ,  $i=1, \dots, N$   
 $X_i$  is a Bernoulli ( $\theta_1$ ) random variable  
 $Y_i$  is a Bernoulli ( $\theta_2$ ) random variable

Here  $\theta_1$  is the probability of success on the first observation within the pair and the  $\theta_2$  is the probability of success on the second observation within the pair.

An important note is that the probability of success,  $\theta$ , does not depend on  $i$ . And we want to test if  $H: \theta_1 = \theta_2$ , i.e. is the probability of success different on the first trial than on the second.

## McNemar's Test and Introduction to ANOVA

---

Example: A recent screening study of 49,528 women (the DMIST trial) compared two different imaging modalities (mammogram, digital mammogram) for detecting breast cancer. Both modalities were preformed on each woman. If both exams were negative the women were followed for one year to be sure cancer was not present and if either exam was positive a biopsy was preformed.

(This was designed and analyzed at Brown University's Center for Statistical Sciences; Reference is Pisano et. al., NJEM, 2005)

Out of the 49,528 enrolled women only 42,555 were eligible, completed all exams and had pathology or follow-up information (called reference standard information).

In studies of diagnostic tests, one should always separate the true positive cases and the true negative cases as determined by reference standard information, and analyze them separately. In this case there were 334 women with breast cancer and 42,221 women without breast cancer.

## McNemar's Test and Introduction to ANOVA

---

Below is the data from the 334 women with breast cancer. Screen film mammograms detected 136 women, digital mammograms detected 138, but only 84 of these women were detected by both modalities. The data are displayed in this 2x2 table:

Data on positive cases from DMIST trial.

		Screen Mammogram		
		Detected	Missed	
Digital Mammo	Detected	84	54	138
	Missed	52	144	196
		136	198	334

We want to see if the proportion of detected women differs between the modalities. It might be tempting to use a simply Chi-square test for this contingency table, but that would be wrong because that test is built to examine the assumption that the rows and columns are independent, which they are clearly not (because the same women are in both groups).

So the test we use is McNemar's test.

## McNemar's Test and Introduction to ANOVA

---

McNemar's test:

		Time 2		
		Success	Failure	
Time 1	Success	a	b	a+b
	Failure	c	d	c+d
		a+c	b+d	N

Let

$$\theta_1 = P(\text{Success} \mid \text{Time 1}) = (a+b)/N$$

$$\theta_2 = P(\text{Success} \mid \text{Time 2}) = (a+c)/N$$

The null hypothesis is  $H_0: \theta_1 = \theta_2$  which implies that  $E[(a+b)/N] = E[(a+c)/N]$  or  $E[b] = E[c]$ .

Another form of the null hypothesis is  $H_0: \theta_1 / \theta_2 = 1$  or  $H_0: \theta_1 / (1 - \theta_1) / \theta_2 / (1 - \theta_2) = 1$  (Odds ratio of detection for Digital over screen is one)

$$\chi^2 = \frac{(b - c)^2}{(b + c)} \text{ with df}=1$$

Notice that  $(\theta_1 - \theta_2)^2 = ((b-c)/N)^2$ , so we see that the difference in proportions is indeed the top of the test statistic, keeping the connection between the Chi-square test and the Z-test for difference in proportions.

So McNemar's test, in this form, is an approximate test that requires large samples to be valid.



## McNemar's Test and Introduction to ANOVA

---

Back to our DMIST example comparing sensitivity:

		Screen Mammogram		
		Detected	Missed	
Digital Mammo	Detected	84	54	138
	Missed	52	144	196
		136	198	334

Here is the *Stata* output for our data:

```
. mcci 84 54 52 144
```

Cases	Controls		Total
	Exposed	Unexposed	
Exposed	84	54	138
Unexposed	52	144	196
Total	136	198	334

```
McNemar's chi2(1) =      0.04      Prob > chi2 = 0.8460
Exact McNemar significance probability      = 0.9227
```

Proportion with factor

Cases	.4131737	[95% Conf. Interval]	
Controls	.4071856		
difference	.005988	-.0574189	.069395
ratio	1.014706	.8757299	1.175737
rel. diff.	.010101	-.0912974	.1114995
odds ratio	1.038462	.696309	1.549997

## McNemar's Test and Introduction to ANOVA

---

Three comments:

1) Some books, like Pagano and Gauvreau, suggest a slightly different version of this test to help when some cell counts are small. Continuity correction:

$$\chi^2 = \frac{(|b - c| - 1)^2}{(b + c)} \text{ with df}=1$$

Opinions differ on when to use this, but typically it is used when any cell counts are less than 5. Same reasoning applies for the continuity corrected version of the Chi-square test.

2) There is no exact analytical formula for the variance of the difference of paired proportions. So the easiest way to get a confidence interval is to simply perform a hypothesis test for every null hypothesis (difference is zero, difference is 0.01, 0.02, 0.03 etc.) and use the set of null hypotheses that DO NOT reject as your confidence interval. This procedure is known as "inverting the hypothesis test", and this is how *Stata* gets the confidence interval for the difference in paired proportions.

3) You should ignore the proportions for cases and controls that *Stata* provides. What you really are comparing here are 138/334 versus 136/334, which has a relative risk of  $138/136 = 1.014$  and risk difference of  $2/334 = .00598802$  (See output).

## McNemar's Test and Introduction to ANOVA

---

We can do a similar analysis for the negative cases:

Data on negative cases from DMIST trial.

		Screen Mammogram		
		Detected	Missed	
Digital Mammo	Detected	40409	780	41189
	Missed	894	139	1032
		41302	919	42221

```
. mcci 40409 780 893 139
```

Cases	Controls		Total
	Exposed	Unexposed	
Exposed	40409	780	41189
Unexposed	893	139	1032
Total	41302	919	42221

```
McNemar's chi2(1) =      7.63      Prob > chi2 = 0.0057
Exact McNemar significance probability      = 0.0062
```

Proportion with factor

Cases	.9755572		
Controls	.9782336	[95% Conf. Interval]	
-----			
difference	-.0026764	-.0045987	-.0007541
ratio	.9972641	.9953276	.9992043
rel. diff.	-.1229597	-.2154003	-.0305192
odds ratio	.8734602	.792443	.9626051

Notice that the p-value is significant (less than 0.05), but the estimated difference is tiny and of no clinical consequence (specificity is the same).

## McNemar's Test and Introduction to ANOVA

---

Based on these results, both screening modalities appear to have the same overall performance.

Interestingly, a similar analysis was done on the subgroup of women less than 50 years old: 72 of these women had breast cancer and 14,203 did not. The data for young positive women are as follows:

```
. mcci 18 17 7 30
```

Cases	Controls		Total
	Exposed	Unexposed	
Exposed	18	17	35
Unexposed	7	30	37
Total	25	47	72

```
McNemar's chi2(1) =      4.17      Prob > chi2 = 0.0412
Exact McNemar significance probability      = 0.0639
```

Proportion with factor

Cases	.4861111		
Controls	.3472222	[95% Conf. Interval]	
		-----	
difference	.1388889	-.0044424	.2822202
ratio	1.4	1.011942	1.93687
rel. diff.	.212766	.0315035	.3940284
odds ratio	2.428571	.957147	6.92694 (exact)

So it appears the sensitivity of these tests are different by about 14%.

Notice that the exact and approximate p-values are different and that the confidence interval for the odds ratio includes 1, but the confidence interval for the relative risk ('ratio') does not. Discuss!

## McNemar's Test and Introduction to ANOVA

---

Looking at the proportion of positive cases that were detected compares the sensitivity. To examine specificity, we look at the group of negative cases. The data for young negative women are as follows:

```
. mcci 13535 285 331 52
```

Cases	Controls		Total
	Exposed	Unexposed	
Exposed	13535	285	13820
Unexposed	331	52	383
Total	13866	337	14203

```
McNemar's chi2(1) =      3.44      Prob > chi2 = 0.0638
Exact McNemar significance probability      = 0.0697
```

```
Proportion with factor
```

Cases	.9730339	[95% Conf. Interval]	
Controls	.9762726		
difference	-.0032388	-.0067337	.0002562
ratio	.9966825	.9931863	1.000191
rel. diff.	-.1364985	-.2903823	.0173853
odds ratio	.8610272	.7323123	1.011859

So it appears that the two tests differ slightly with respect to specificity, although the difference is small and likely uninteresting.

Digital mammograms appears to have better sensitivity (and the same specificity) for women under 50 and would therefore be a (slightly) better screening test.

### Exact p-values for McNemar's test

Because McNemar's test is based on the information only in the discordant pairs (the b and c off diagonal cells) the calculations of exact p-values is quite simple.

If the null hypothesis is true, then it implies that the b and c cells should be equal, very informally,  $H_0: b=c$ . These cells are just counts of people, so the underlying distribution has to be Binomial where  $\frac{1}{2}$  of the counts should be in each cell.

That is, under the null hypothesis

$$b \sim \text{Binomial}(b+c, 0.5)$$

So an exact p-value is  $P(b > b_{\text{obs}} \mid n=b+c \text{ and } \theta=1/2)$ .

Example: in the subgroup of women less than 50 years old 72 of these women had breast cancer. Here  $b=17$  and  $c=7$ , so

$$P(X \geq 17 \mid n = 24, \theta = 1/2) = \sum_{i=17}^{24} \binom{24}{i} 0.5^{24} = 0.03196$$

$$\text{two sided p - value} = 2(.03196) = .06391$$

```
. mcci 18 17 7 30
... (See page 10)
```

```
McNemar's chi2(1) =      4.17      Prob > chi2 = 0.0412
Exact McNemar significance probability      = 0.0639
```

### One Way Analysis of Variance (ANOVA)

We know how to test the equality of two normal means: Two samples, having  $n_1, \bar{x}_1, s_1$ , and  $n_2, \bar{x}_2, s_2$ .

Model (assumptions): Independent observations from two normal distributions with means  $\mu_1, \mu_2$  and *common variance*  $\sigma^2$ .

To test the hypothesis  $H_0 : \mu_1 = \mu_2$  (means are equal) vs.  $H_A : \mu_1 \neq \mu_2$  we use the test statistic

$$\frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{s_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

We reject  $H_0$  if the observed value of the test statistic exceeds the critical value found in t-table using  $n_1 + n_2 - 2$  degree of freedom.

Example: For  $n_1=10$  and  $n_2=82$ , the two-sided 5% critical value is 2.120.

## McNemar's Test and Introduction to ANOVA

---

This is the same thing if we checked to see if the square of the observed test statistic is bigger than  $(2.120)^2 = 4.4944$ .

Mathematically, we have

$$\begin{aligned} \frac{(\bar{x}_1 - \bar{x}_2)^2}{s_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)} &= \frac{n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2}{s_p^2} \\ &= \frac{n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2}{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} > (t_{\alpha/2})^2 \end{aligned}$$

where the sample mean without a subscript is the overall mean obtained from the combined sample. This is sometimes called the 'grand mean'.

⇒ This squared form of the test statistic is important because it shows us how to generalize the test for more than two groups.



## McNemar's Test and Introduction to ANOVA

---

For three groups  $n_1, \bar{x}_1, s_1$ ,  $n_2, \bar{x}_2, s_2$  and  $n_3, \bar{x}_3, s_3$ , we have the same assumptions: Independent observations from three normal distributions with means  $\mu_1, \mu_2, \mu_3$  and common variance  $\sigma^2$ .

To test the hypothesis  $H_0 : \mu_1 = \mu_2 = \mu_3$  (all means equal, or no differences among means) versus  $H_A$ : not all means are equal, we use the following test statistic:

$$\frac{n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2 + n_3(\bar{x}_3 - \bar{x})^2}{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + (n_3 - 1)s_3^2}{n_1 + n_2 + n_3 - 3}} = \frac{s_b^2}{s_w^2}$$

where  $\bar{x}$  is the grand mean of all  $n = n_1 + n_2 + n_3$  observations:

$$\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2 + n_3 \bar{x}_3}{n_1 + n_2 + n_3}$$

Under  $H_0$  this statistic has an "F-distribution with 2 and n-3 degrees of freedom." The F-distribution is the square of the t-distribution just like the Chi-square is the square of the Z-distribution.

## McNemar's Test and Introduction to ANOVA

---

More generally, when there are  $k$  groups:

$$\begin{array}{c} n_1, \bar{x}_1, s_1^2 \\ n_2, \bar{x}_2, s_2^2 \\ \vdots \\ n_k, \bar{x}_k, s_k^2 \end{array}$$

The test statistic is

$$\begin{aligned} F_{k-1, n-k} &= \frac{\sum_{i=1}^k n_i (\bar{x}_i - \bar{\bar{x}})^2}{k-1} \div \frac{\sum_{i=1}^k (n_i - 1) s_i^2}{n-k} \\ &= s_b^2 / s_w^2 \end{aligned}$$

Clever statisticians have proved that  $E(s_w^2) = \sigma^2$ , and that when  $H_0$  is true,  $E(s_b^2) = \sigma^2$  as well. But when  $H_A$  is true,  $E(s_b^2) > \sigma^2$ .

The bigger the  $F$ -statistic, the stronger the evidence that the population means are not all equal.

Moreover when  $H_0$  is true the ratio  $s_b^2/s_w^2$  has an  $F$  probability distribution with  $k-1$  and  $n-k$  degrees freedom.

## McNemar's Test and Introduction to ANOVA

---

To test  $H_0$  at level 5%, find the critical value in the  $F$ -table and reject if the observed value  $s_b^2 / s_w^2$  exceeds the critical value. Or find the p-value in the table,  $p = P( F_{k-1, n-k} > F_{observed} )$ , and reject if it is smaller than 5%.

When  $k$  is two, the  $F_{1, n-2}$  is the same as  $(T_{n-2})^2$ . For example, we found that at 5% the critical value of  $T$  with 16 df is 2.120, so the critical value of  $T^2$  is 4.494.

Table 9 shows that this (4.49) is indeed the critical value of  $F$  with 1 and 16 df.

## McNemar's Test and Introduction to ANOVA

---

ANOVA is built from the same materials as the two sample t-test, and the same assumptions are required: normal distributions with equal variances.

As with the t-test, the ANOVA tests are robust (relatively insensitive) to failure of the normality assumption.

There is a nonparametric alternative test (the Kruskal-Wallis test) that is based on the ranks of the observations, and does not require that the underlying distributions be normal.

What do you do after the  $F$ -test rejects the hypothesis of no difference among the  $k$  population means, and you want to know which pairs of means are different?

There are various complicated approaches: The simplest is to test all of the possible pairs using two-sample t-tests (with  $s_w^2$  replacing  $s_p^2$ , so you have  $n-k$  df), performing all  $\binom{k}{2}$  tests at the level  $\alpha/\binom{k}{2}$ . This is known as a Bonferroni-adjusted  $\alpha$ -level.

### Adjusting the alpha ( $\alpha$ ) level

There are times when it is necessary to control the overall Type I error and keep it from inflating. For example, if you design a study of a new drug with two primary endpoints and you consider the test a success if the drug performs better on *either* endpoint. You may constrain the overall type I error to an  $\alpha$ -level by testing each endpoint at the  $\alpha/2$  level, so that the total chance of making a Type I error in this study would be  $\alpha$ .

There are a variety of opinions about whether this makes sense. The basic conflict is in figuring out which error you want to control: the error for an individual endpoint or the overall error for a study.

Controlling the overall error can lead to weird results because now rejection of one endpoint depends on how many other endpoints you decide to test.

For example, one endpoint might yield a p-value of 0.04, which would reject the null at the 5% level and conclude the drug works. But if you have two endpoints and constrain the overall error to 5% by testing each endpoint a 2.5% level, then you would fail to reject and conclude the drug does not work.

## McNemar's Test and Introduction to ANOVA

---

To make matters worse, both procedures have an overall 5% error rate. So you can only claim rejection at the 5% level in either case.

Procedures like this make people wonder if statisticians really have their head screwed on straight: The drug works if you did not test the other endpoint, but it does not work if you did.

This is a fascinating, but endless debate. The problem is that modern statistics uses one quantity (the tail area, either as a p-value or type I error) to do two things: (1) measure the strength of evidence against the null hypothesis and (2) tell me how often I make a mistake.

And it makes sense to adjust #2, but not #1. The only way this can be resolved is to trash this approach and try something new (like using a likelihood ratio to measure the strength of evidence and calculating its Type I and Type II error).

Blume and Peipert "What your statistician never told you about p-values" (2003) has a nice discussion of this point.

*"In fact, as a matter of principle, the infrequency with which, in particular circumstances, decisive evidence is obtained, should not be confused with the force, or cogency, of such evidence."* [Fisher, 1959]

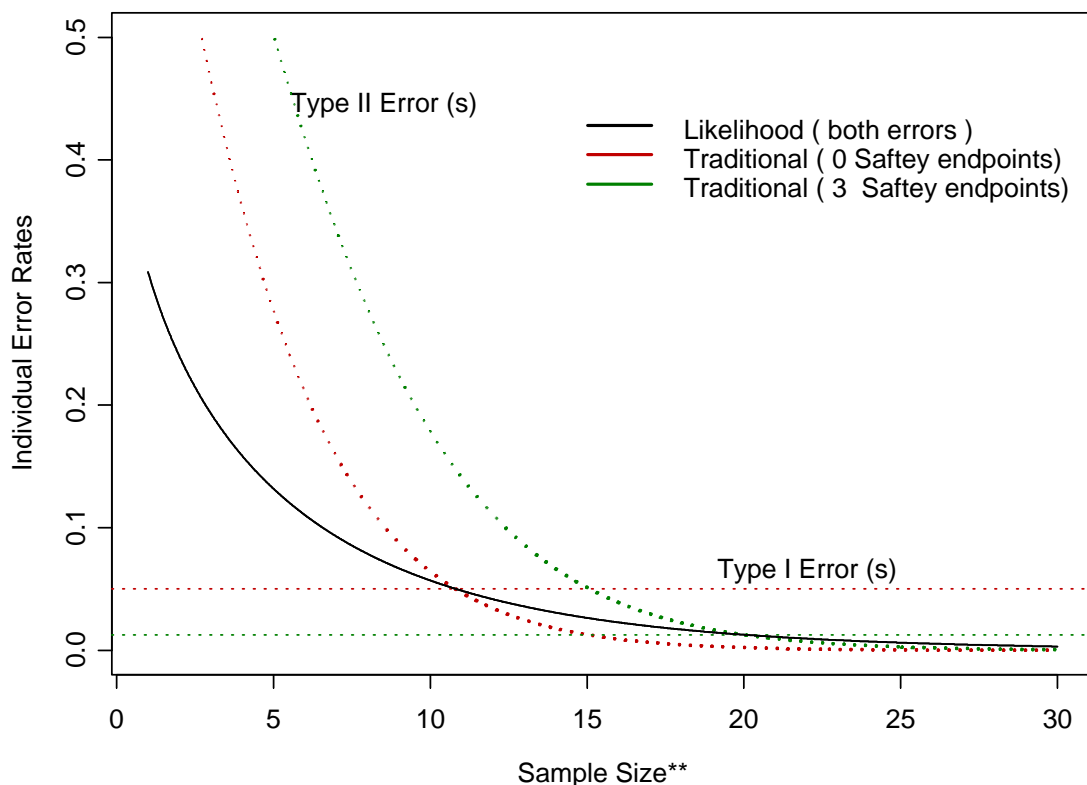
## McNemar's Test and Introduction to ANOVA

---

Single endpoint:

Frequentist error rates (Type I and Type II; reject when in the tails) along with likelihood error rates (reject when the likelihood ratio is greater than 1).  
Adjusting the Type I error to keep it at 5%.

Frequentists properties('Error' rates)



\*Likelihood is not affected by multiple endpoints

\*\*Sample Size is relative to this example; but ordering holds in general

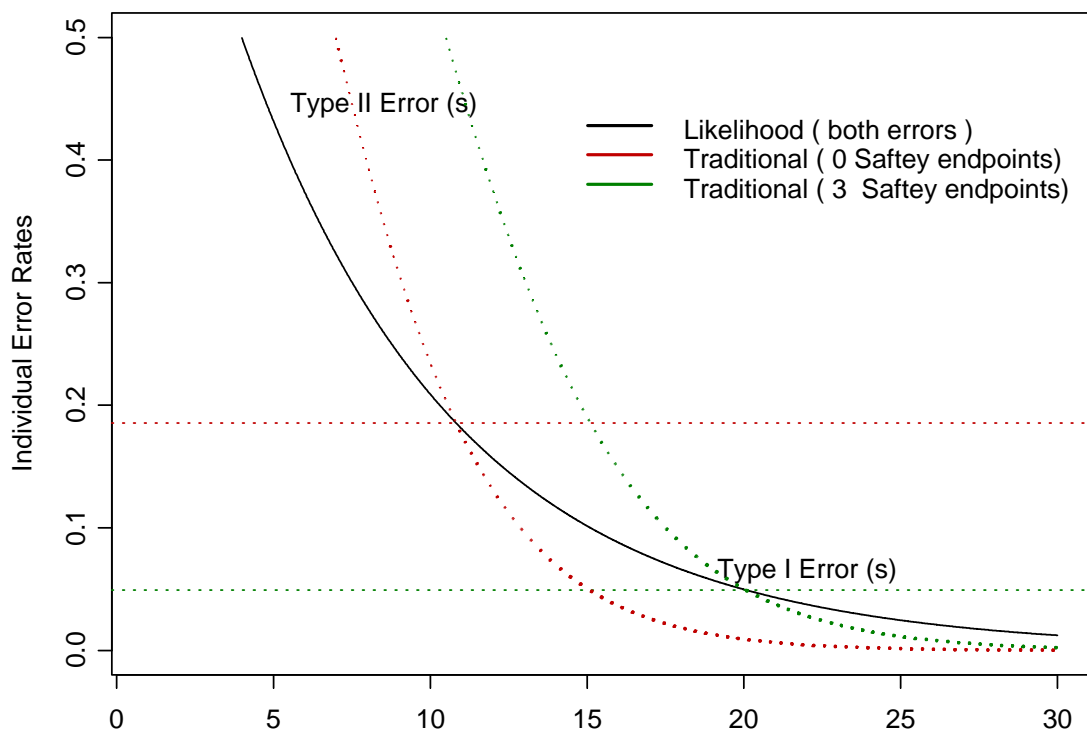
## McNemar's Test and Introduction to ANOVA

---

Overall endpoints:

Frequentist error rates (Type I and Type II; reject when in the tails) along with likelihood error rates (reject when the likelihood ratio is greater than 1). Adjusting the Type I error to keep it at 5%.

Frequentists properties('Error' rates)



Sample Size\*\*  
\*Likelihood is not affected by multiple endpoints

\*\*Sample Size is relative to this example; but ordering holds in general



## McNemar's Test and Introduction to ANOVA

---

Plots of the average error rates  $((\text{type I} + \text{type II})/2)$  for hypothesis testing and likelihood inference. Multiple endpoints are included along with the adjusted and not adjusted results for hypothesis testing.

