Inference for Proportions

Recipes for inference so far:

Confidence intervals (margin of error): - calculate a measure of variability for the sample estimate - Use a theoretical distribution to get a critical value - Generate an estimate and interval

$$estimate \pm criticalvalue * variability$$

Hypothesis testing: - articulate a null hypothesis and alternative hypothesis - choose an appropriate statistical test - generate a statistic - compare to a critical value or p value - reject or fail to reject the null hypothesis

Roadmap

Is my outcome continuous or categorical?

How many groups am I describing?

How many groups?	Independent?	parametric?	test
1	yes	yes	Z or one sample T
2	yes	yes	Two sample T
2	yes	no	Wilcox rank sum
2	no	yes	Paired T
2	no	no	Wilcox sign rank
3 or more	yes	yes	ANOVA
3 or more	yes	no	Kruskal Wallis

This lecture

Extending our recipe to categorical (binary, yes/no) outcomes

- Confidence interval for a single proportion
- Hypothesis tests for a single proportion
- Sample size estimates for a single proportion
- CI for the difference between two proportions
- Hypothesis tests for two proportions

Confidence intervals for a proportion

Conditions for inference about a proporiton

- Data are a simple random sample from the population
- The sample size n is large enough to ensure that the sampling distribution of $\hat{p}$ is close to normal

Recall the sampling distribution for $\hat{p}$

The sampling distribution for $\hat{p}$ is centered on p with a standard error of $\sqrt{\frac{p(1-p)}{n}}$

Normal approximation for binomial distributions (lecture 16)

Suppose that a count X has the binomial distribution with n observations and success probability p. When n is large, the distribution of X is approximately Normal. That is,

$$X \dot{\sim} N(\mu = np, \sigma = \sqrt{np(1-p)})$$

As a general rule, we will use the Normal approximation when n is so large that $np \geq 10$ and $n(1-p) \geq 10$.

It is most accurate for p close to 0.5, and least accurate for p closer to 0 or 1.

CI using our “recipe” from before

For a one sample t-test the CI looked like this:

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

If we follow the same format for the CI from previous chapters we would get:

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

This is indeed the large sample confidence interval for a population proportion

But... (there is a but)

- If we do not have a sampling distribution that approaches Normal, this confidence interval does not perform well, meaning that even if you think you should have 95% "confidence" that the CI contains the true value p, it is very often much lower.
- This means that if you were to repeat the procedure 100 times, fewer than 95 of the confidence intervals would contain the true value. This is not good!

Four types of CI for a proportion...

We will discuss four different ways to compute confidence intervals for proportions:

- Using the large sample method
- Using the "plus four" method (by hand)
- Using R's `prop.test`, which implements the "Wilson Score" method with continuity correction
- the exact or Clopper Pearson method

Plus four, an easy trick that to save the CI

- If you add 2 imaginary successes and 2 failures to the dataset (increasing the sample size by 4 imaginary trials), the interval performs well again.
- Let $\tilde{p} = \frac{\text{number of successes} + 2}{n+4}$
- Let $SE = \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}}$
- Then the CI is:

$$\tilde{p} \pm z^* \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}}$$

- This is called the plus four method
- Note we use z^* rather than t^*. This is because the standard error of the sampling distribution is completely determined by p and n, we don't need to estimate a second parameter. Because of this we stay in the land of z scores.
- Use this method when n is at least 10 and the confidence level is at least 90%

Why does the plus four method work?

- It is a simplification of a more complex method known as the Wilson Score Interval.
- You don't need to know why it works, just that it is better to use this "plus four" trick if you're making a confidence interval for a proportion by hand.

Example of the plus four method

A study examined a random sample of 75 SARS patients, of which 64 developed recurrent fever.

Therefore $\hat{p} = 64/75 = 85.33\%$

Using the plus 4 method: $\tilde{p} = \frac{64+2}{75+4} = 83.54\%$

$$SE = \sqrt{\frac{\tilde{p}(1-\tilde{p})}{75+4}} = \sqrt{\frac{.8354 \times (1 - 0.8354)}{79}} = 0.04172$$

Thus the plus four 95% CI is:

$$\tilde{p} \pm 1.96 \times SE = 0.8354 \pm 0.04172 = 79.37\% \text{ to } 87.71\%$$

Wilson score based Estimate for a proportion in R

- If you are using R, simply use `prop.test()`

The general syntax is:

prop.test(variable)

The default here will be a two sided test , you may change this by specifying "less", or "greater"

prop.test(variable, alternative=less)

What does the prop.test function in R use?

- In the R function `prop.test` (analogous to `t.test`) there are functions that calculate confidence intervals and hypothesis tests for binomial proportions.
- prop.test in R uses what is known as the "Wilson score interval with a continuity correction". Thus, when you use the `prop.test` function, you don't need to "plus 4", it will do it for you (and does an even better job because of the continuity correction.)

The exact method (Clopper Pearson)

- There is another method to compute confidence intervals for proportions that is often used called the Clopper Pearson method, or the "Exact method". It is implemented with R's `binom.test()`
- The exact method is statistically conservative, meaning that it gives better coverage than it suggests. That is, a 95% CI computed under this method includes the true proportion in the interval $> 95\%$ of the time.

further details

References (if you are interested in further reading - you will not be tested on this)

- Wilson, E.B. (1927). Probable inference, the law of succession, and statistical inference. Journal of the American Statistical Association, 22, 209–212.
- Newcombe R.G. (1998). Two-Sided Confidence Intervals for the Single Proportion: Comparison of Seven Methods. Statistics in Medicine, 17, 857–872.
- Newcombe R.G. (1998). Interval Estimation for the Difference Between Independent Proportions: Comparison of Eleven Methods. Statistics in Medicine, 17, 873–890.

Example using all four CI methods

Example applying all the methods

Suppose that 500 elderly individuals suffered hip fractures, of which 100 died within a year of their fracture. Compute the 95% CI for the proportion who died using:

- the large sample method,
- the plus for method (by hand),
- the Wilson Score method (using `prop.test`),
- the Clopper Pearson Exact method (using `binom.test`)

Example of large sample method to the CI for a proportion

```r
p.hat <- 100/500 # estimate proportion
se <- sqrt(p.hat*(1-p.hat)/500) # standard error
c(p.hat - 1.96*se, p.hat + 1.96*se) # CI
```

```
## [1] 0.1649385 0.2350615
```

Using the large sample method, the confidence interval is 16.5% to 23.5%.

Note that you could compute this by hand.

Example using the plus 4 method

```
p.tilde <- (100 + 2)/(500 + 4)
se <- sqrt(p.tilde*(1-p.tilde)/504) # standard error
c(p.tilde - 1.96*se, p.tilde + 1.96*se) # CI
```

```
## [1] 0.1673039 0.2374580
```

Using the plus 4 method, the confidence interval is 16.7% to 23.7%.

Example using the Wilson Score method to the CI for a proportion

```
prop.test(x = 100, n = 500, conf.level = 0.95)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  100 out of 500, null probability 0.5
## X-squared = 178.8, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.1663581 0.2383462
## sample estimates:
##   p
## 0.2
```

- The 95% confidence interval using the Wilson Score method is 16.6% to 23.8%.

Example using the Wilson Score method to the CI for a proportion

- Note that the `prop.test` function is also conducting a two-sided hypothesis test (where $H_0 : p_0 = 0.5$, unless otherwise specified). You can ignore the testing-related output and focus on the CI output when using the function to make a CI.

Example using the Clopper Pearson "Exact" method to the CI for a proportion

```
binom.test(x = 100, n = 500, conf.level = 0.95)
```

```
##
##  Exact binomial test
##
## data:  100 and 500
## number of successes = 100, number of trials = 500, p-value < 2.2e-16
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.1658001 0.2377918
## sample estimates:
## probability of success
##                    0.2
```

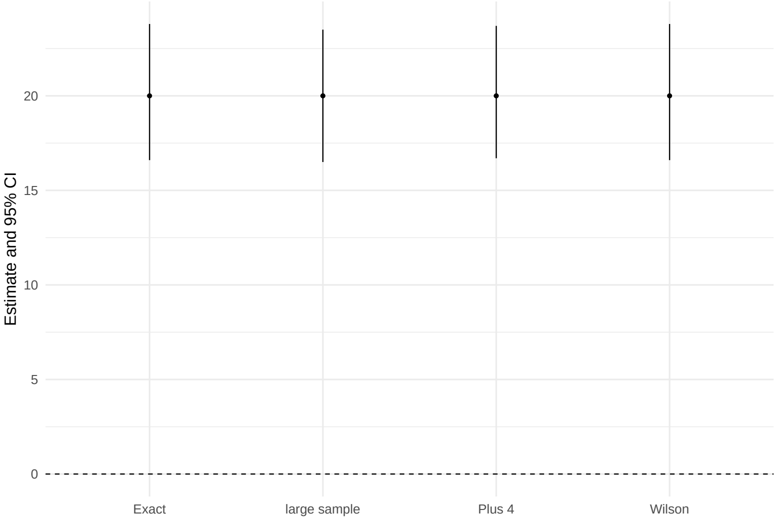
- The 95% confidence interval using the exact binomial test is 16.6% to 23.8%.

Example using the Clopper Pearson "Exact" method to the CI for a proportion

- Note that the `binom.test` function is also conducting a two-sided hypothesis test (where $H_0 : p_0 = 0.5$, unless otherwise specified). You can ignore the testing-related output and focus on the CI output when using the function to make a CI.

Comparison of the four methods

We can graphically compare the CIs :



Summary of the confidence intervals across the methods

Method	95% Confidence Interval	R Function
Large sample	16.5% to 23.5%	by hand
Plus four	16.7% to 23.7%	by hand
Wilson Score*	16.6% to 23.8%	`prop.test`
Exact	16.6% to 23.8%	`binom.test`

*with continuity correction

- Note that only the large sample method is symmetric around $\hat{p} = 20\%$. This is okay. There is no reason why we require a symmetric confidence interval.
- When the Normal approximation assumptions are satisfied, the methods give very similar results, as shown here.

Example 2

Suppose that there were 100 elderly individuals with falls observed, and 2 died.

Example 2

Large sample and plus four method calculations

```
p.hat <- 2/100 # estimate proportion
se <- sqrt(p.hat*(1-p.hat)/100) # standard error
c(p.hat - 1.96*se, p.hat + 1.96*se) # CI
```

```
## [1] -0.00744  0.04744
```

```
p.tilde <- (2 + 2)/(100 + 4)
se <- sqrt(p.tilde*(1-p.tilde)/104) # standard error
c(p.tilde - 1.96*se, p.tilde + 1.96*se) # CI
```

```
## [1] 0.00150119 0.07542189
```

Example 2

```
binom.test(x = 2, n = 100, p = 0.5, conf.level = 0.95)
```

```
##
##  Exact binomial test
##
## data:  2 and 100
## number of successes = 2, number of trials = 100, p-value < 2.2e-16
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.002431337 0.070383932
## sample estimates:
## probability of success
##                   0.02
```

Example 2

```r
prop.test(x = 2, n = 100, p = 0.5, conf.level = 0.95)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  2 out of 100, null probability 0.5
## X-squared = 90.25, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.003471713 0.077363988
## sample estimates:
##    p
## 0.02
```

Example 2 summary

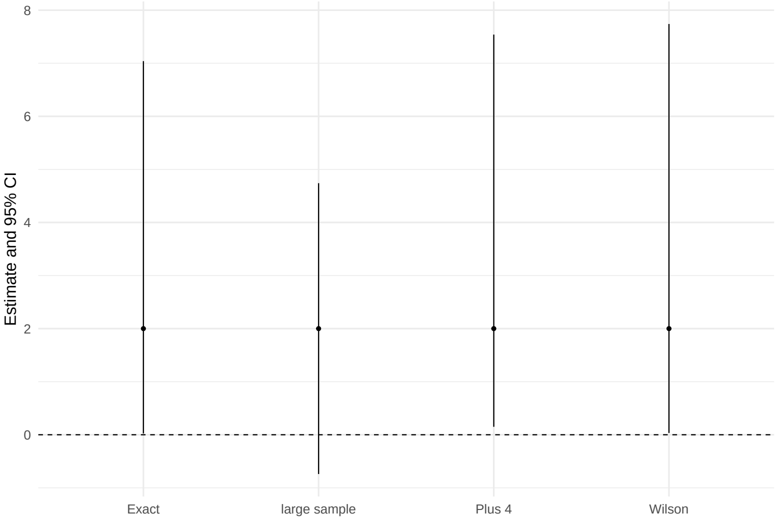
Here are the 95% CIs applying the four different methods:

Method	95% Confidence Interval	R Function
Large sample	-0.74% to 4.74%	by hand
Exact	0.024% to 7.04%	`binom.test`
Wilson Score*	0.034% to 7.74%	`prop.test`
Plus four	0.15% to 7.54%	by hand

*with continuity correction

Example 2

We can graphically compare the CIs from the previous slide:



Example 2

Findings:

- Notice how different the intervals are, especially large sample vs. others.
- Notice that the large sample lower bound is non-sensical (i.e., we can't have negative proportions!)
- The large sample CI differs from the others because the Normal approximation assumptions are not satisfied.

Hypothesis testing for a proportion

Hypothesis tests of a proportion

When you only have one sample what is the null hypothesis? You're interested in knowing whether there is evidence against the null hypothesis that the population proportion p is equal to some specified value p_0. That is:

$$H_0 : p = p_0$$

Hypothesis tests of a proportion

Recall the sampling distribution for the proportion:

- Normally distributed
- Centered at p_0 under the null distribution
- Has a standard error of $\sqrt{\frac{p_0(1-p_0)}{n}}$

Hypothesis tests of a proportion

The test statistic for the null hypothesis is:

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

This is a z-test (not a t-test) so we compared to the standard Normal distribution and ask what is the probability of observing a z value of this magnitude (or more extreme).

Hypothesis tests of a proportion

One sided alternatives:

- $H_a : p > p_0$
- $H_a : p < p_0$

Two-sided alternative:

- $H_a : p \neq p_0$

When to use this test? Use this test when the expected number of successes and failures is ≥ 10. That is, when $np_0 \geq 10$ and $n(1 - p_0) \geq 10$.

Example of a hypothesis test for a proportion

Consider a SRS of 200 patients undergoing treatment to alleviate side-effects from a rigorous drug regimen at a particular hospital, where 33 patients experienced reduced or no side-effects.

$\hat{p} = 33/200 = 0.165 = 16.5\%$

Suppose that historically, the rate of patients with little or no side-effects is 10%. Does the new treatment increase the rate? That is:

$H_0 : p = 0.10$ vs. $H_a : p > 0.10$

Example of a hypothesis test for a proportion

Step 1: Calculate $\hat{p} = 16.5\%$ from previous slide.

Step 2: Calculate the standard error of the sampling distribution for p under the null hypothesis: $SE = \frac{\sqrt{p_0(1-p_0)}}{n} = \frac{\sqrt{0.1(1-0.1)}}{200} = 0.0212132$

Step 3: Calculate the z-test for the proportion:

$$z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.165-0.10}{0.0212132} = 3.06413$$

Example of a hypothesis test for a proportion

Step 4: Calculate the probability of seeing a z-value of this magnitude *or larger*:

```
pnorm(q = 3.06413, lower.tail = F)
```

```
## [1] 0.00109152
```

Step 5: Evaluate the evidence against the null hypothesis. Because the p-value is so small (0.1%), there is little chance of seeing a proportion equal to 16.5% or larger if the true proportion was actually 10%. Thus, there is evidence in favor of the alternative hypothesis, that the underlying proportion is larger than 10%.

Sample size for a proportion

How big should the sample be to estimate a proportion?

Suppose that you want to estimate a sample size for a proportion within a given margin of error. That is, you want to put a maximum bound on the width of the corresponding confidence interval.

Let m denote the desired margin of error. Then $m = z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

We can solve this equation for n, but we also need to plug in a value for p. To do that we make a guess for p denoted by p^*.

p^* is your best estimate for the underlying proportion. You might gather this estimate from a completed pilot study or based on previous studies published by someone else. If you have no best guess, you can use $p^* = 0.5$. This will produce the most conservative estimate of n. However if the true p is less than 0.3 or greater than 0.7, the sample size estimated may be much larger than you need.

How big should the sample be to estimate a proportion?

Rearranging the formula on the last slide for n, we get:

$$m = z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$\sqrt{n} m = z^* \sqrt{p(1-p)}$$

$$\sqrt{n} = \frac{z^*}{m} \sqrt{p(1-p)}$$

$$n = (\frac{z^*}{m})^2 p^* (1-p^*)$$

This last formula is the one we will use to estimate the required sample size.

Example of estimating sample size

Suppose after the midterm vote, you were interested in estimating the number of STEM undergraduate students who voted. First you need to decide what margin of error you desire. Suppose it is 4 percentage points or $m = 0.04$ for a 95% CI.

If you had no idea what proportion of STEM students voted then you let $p^* = 0.5$ and solve for n:

$n = (\frac{z^*}{m})^2 p^*(1 - p^*) = (\frac{1.96}{0.04})^2 \times 0.5 \times 0.5 = 600.25 = 601$

However, suppose you found a previous study that estimated the number of STEM students who voted to be 25%. Then what sample size would you need to detect this proportion?

$n = (\frac{z^*}{m})^2 p^*(1 - p^*) = (\frac{1.96}{0.04})^2 \times 0.25 \times 0.75 = 450.19 = 451$

Comparing two proportions (Chapter 20)

- Two SRS from independent populations

Notation:

Population	Population proportion	Sample size	Sample proportion
1	p_1	n_1	$\hat{p}_1$
2	p_2	n_2	$\hat{p}_2$

CI for the difference in two proportions

Large-sample confidence interval for the difference of two proportions

- Use when the number of observed successes and failures are > 10 for both samples

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

- Just like for the difference between two means, the SE of the difference is the square root of the sum of the variances.
- This large-sample interval often has a lower confidence level than the one specified. That is, if you repeated the method several times < 95 of the 100 created intervals would contain the true value for the difference between the proportions for a 95% CI.

Example using the large sample method

Patients in a randomized controlled trial were severely immobilized and randomly assigned to either Fragamin (to prevent blood clots) or to placebo. The number of patients experiencing deep vein thrombosis (DVT) was recorded

	DVT	no DVT	Total	$\hat{p}$
Fragamin	42	1476	1518	0.0277
Placebo	73	1400	1473	0.0496

- We can apply the large study method because the sample sizes are large and the number of observed successes and failures are > 10 (i.e., 42, 73, 1476, and 1400 all > 10).

Example using the large sample method

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

$$(0.0496 - 0.0277) \pm z^* \sqrt{\frac{0.0496_1(1-0.0496)}{1473} + \frac{0.0277(1-0.0277)}{1518}}$$

$$0.0219 \pm 1.96 \times 0.0071 = 0.008 \text{ to } 0.0358$$

Plus 4 method for the comparison of two proportions

- When the assumptions of the large sample method are not satisfied, we use the plus four method.
- When you have two samples this method says: add 4 observations, 1 success and 1 failure to each of the two samples.

$$\tilde{p}_1 = \frac{\text{no. of successes in pop1} + 1}{n_1+2} \quad \tilde{p}_2 = \frac{\text{no. of successes in pop2} + 1}{n_2+2}$$

$$(\tilde{p}_1 - \tilde{p}_2) \pm z^* \sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1+2} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_2+2}}$$

- Use when the sample size is at least five, with any counts of success and failure (can even use when number of successes or failures $= 0$)
- Much more accurate when the sample sizes are small
- May be conservative (giving a higher level of confidence than the one specified)

Example using the plus four method

	Flu	no Flu	Total	$\hat{p}$
Vaccine	4	96	100	0.04
Placebo	11	89	100	0.11

Here, we don't have 10 "successes" (flu) in both groups, so we cannot use the Normal approximation method.

Example using the plus four method

$$\tilde{p}_1 = \frac{\text{no. of successes in pop1} + 1}{n_1 + 2} = \frac{5}{102}$$

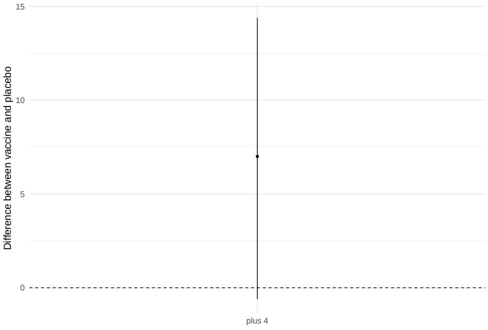
$$\tilde{p}_2 = \frac{\text{no. of successes in pop2} + 1}{n_2 + 2} = \frac{12}{102}$$

$$(\tilde{p}_1 - \tilde{p}_2) \pm z^* \sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1+2} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_2+2}}$$

$$\left(\frac{12}{102} - \frac{5}{102}\right) \pm 1.96 \times 0.0384 = \text{-}0.6\% \text{ to } 14.4\%$$

Example using the plus four method (continued)

The 95% CI of the difference ranged from -0.6 percentage points to 14.4% percentage points. While this CI contains 0 (the null hypothesized value for no difference) most of the values contained within it are positive, perhaps suggesting support for the alternative hypothesis. In this case, we might want to collect more data to create a more precise CI.



Hypothesis testing - two samples binary data

Hypothesis testing when you have two samples and binary data

$H_0 : p_1 = p_2$ or $p_1 - p_2 = 0$

H_a :

- $p_1 \neq p_2$ or $p_1 - p_2 \neq 0$ (two-sided)
- $p_1 > p_2$ or $p_1 - p_2 > 0$ (one sided upper tail)
- $p_1 < p_2$ or $p_1 - p_2 < 0$ (one sided lower tail)

What does it mean to assume the null is true?

- If the null hypothesis is true, then p_1 is truly equal to p_2. In this case, our best estimate of the underlying proportion that they are both equal to is

$$\hat{p} = \frac{\text{no. successes in both samples}}{\text{no. individuals in both samples}}$$

- Also, our best guess at the SE for $\hat{p}$ is:

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}$$

$$\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}$$

This is the formula for the SE for the difference between two proportions but we have substituted $\hat{p}$ for p_1 and p_2.

Hypothesis testing when you have two samples and binary data

Using the information from the previous slide, we can create the z-test for the difference between two proportions as:

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

Use this test when the counts of successes and failures are ≥ 5 in both samples

Example of hypothesis testing when you have two samples and binary data

Recall the RCT data on the occurrence of DVT between Fragamin vs. placebo groups:

	DVT	no DVT	Total	$\hat{p}$
Fragamin	42	1476	1518	0.0277
Placebo	73	1400	1473	0.0496

$H_0 : p_1 = p_2$, or that the rate of DVT is the same between Fragamin and placebo groups.

Suppose you're interested in knowing whether these two groups had different rates of DVT. Then, $H_a : p_1 \neq p_2$

Example of hypothesis testing when you have two samples and binary data

1. Compute $\hat{p} = \frac{42+73}{1518+1473} = \frac{115}{2991} = 0.03844868$
2. Compute the SE: $\sqrt{0.0384(1-0.0384)(\frac{1}{1518}+\frac{1}{1473})} = 0.007032308$
3. Compute the test statistic:

$$z = \frac{0.04955872 - 0.02766798}{0.007032308} = 3.11$$

4. Calculate the p-value

```
pnorm(q = 3.112881, lower.tail = F)*2
```

```
## [1] 0.001852707
```

Two sample hypothesis testing in R

Example of hypothesis testing when you have two samples and binary data

```
prop.test(x = c(42, 73), # x is a vector of number of successes
          n = c(1518, 1473)) # n is a vector of sample sizes
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(42, 73) out of c(1518, 1473)
## X-squared = 9.107, df = 1, p-value = 0.002546
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.036376917 -0.007404562
## sample estimates:
##     prop 1     prop 2
## 0.02766798 0.04955872
```

Example of hypothesis testing when you have two samples and binary data

- R gives a slightly different p-value because it has a continuity correction.
- This is okay. If you want to use R to check your hand calculation, you need to add the argument `correct = F` to the calculation.

Next up

The next lectures we will introduce the Chi-squared distribution and two tests - Goodness of fit - Chi-squared test of assiciation

This will be the end of new material for the semester. Bonus material on bootstrapping and permutations will be posted as notes, but has been dropped from the required material for the exam. It may be included as extra credit.

Parting Humor