

BIOS: 4120 Lab 11

April 14-15, 2020

In today's lab we will briefly revisit Fisher's Exact Test, discuss confidence intervals for odds ratios, and review for quiz 3.

Note: The material in the first part of this lab will not be on quiz 3.

Remember that Fisher's Exact Test can be performed using the `fisher.test()` function. In lab last week using the `lister` dataset, we saw that this function not only gives you the p-value but also the odds ratio and a 95% CI for the odds ratio.

```
library(DescTools)
lister <- read.delim("http://myweb.uiowa.edu/pbreheny/data/lister.txt")
lister.table <- Rev(table(lister))
print(lister.table)
```

```
##           Outcome
## Group      Survived Died
## Sterile           34    6
## Control           19   16
```

```
fisher.test(lister.table)
```

```
##
## Fisher's Exact Test for Count Data
##
## data: lister.table
## p-value = 0.005018
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  1.437621 17.166416
## sample estimates:
## odds ratio
##  4.666849
```

Note that this Fisher's Exact Test odds ratio is calculated using the formula:

$\frac{a/b}{c/d} = \frac{ad}{bc}$, where a, b, c, and d are defined by their location according to the following table:

```
##           Success Failure
## Option 1 a           b
## Option 2 c           d
```

The easiest way to think of this odds ratio is to say that the odds of Success are $100 \times (\text{OR} - 1)$ percent higher if we do Option 1 than if we do Option 2.

Be careful when calculating the OR that you interpret what you calculated. For instance, if I were to switch my Success and Failure column names, I'd be calculating the OR for Failure given Option 1 instead of Success.

CI for OR

Calculating a confidence interval for an odds ratio is slightly more complicated than what we've done so far because it's on a different scale. Here are the steps:

1. Find the odds ratio: $\frac{ad}{bc}$
2. Find the log odds ratio: $\log(OR)$
3. Find the error term: $SE_{\log OR} = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$
4. Calculate the CI on this scale: $CI_{\log OR} = \log(OR) \pm z_{\alpha/2} * SE_{\log OR}$
5. Calculate the CI on the original scale: $CI = \exp(CI_{\log OR})$

Let's find a CI for the OR for survival given sterile procedure by hand.

```
##           Outcome
## Group      Survived Died
## Sterile           34    6
## Control           19   16
```

```
##           Outcome
## Group      Survived Died
## Sterile a           b
## Control c           d
```

1. Find the OR.

```
OR <- (34*16)/(19*6)
print(OR)
```

```
## [1] 4.77193
```

2. Find the log(OR)

```
logOR <- log(OR)
print(logOR)
```

```
## [1] 1.562751
```

3. Find the error term.

```
SElogOR<-sqrt((1/19)+(1/6)+(1/16)+(1/34))
print(SElogOR)
```

```
## [1] 0.557862
```

4. Calculate the CI on this scale.

```
logCI <- logOR + qnorm(c(.025,.975)) * SElogOR
print(logCI)
```

```
## [1] 0.4693614 2.6561402
```

5. Calculate the CI on the original scale.

```
CI <- exp(logCI)
print(CI)
```

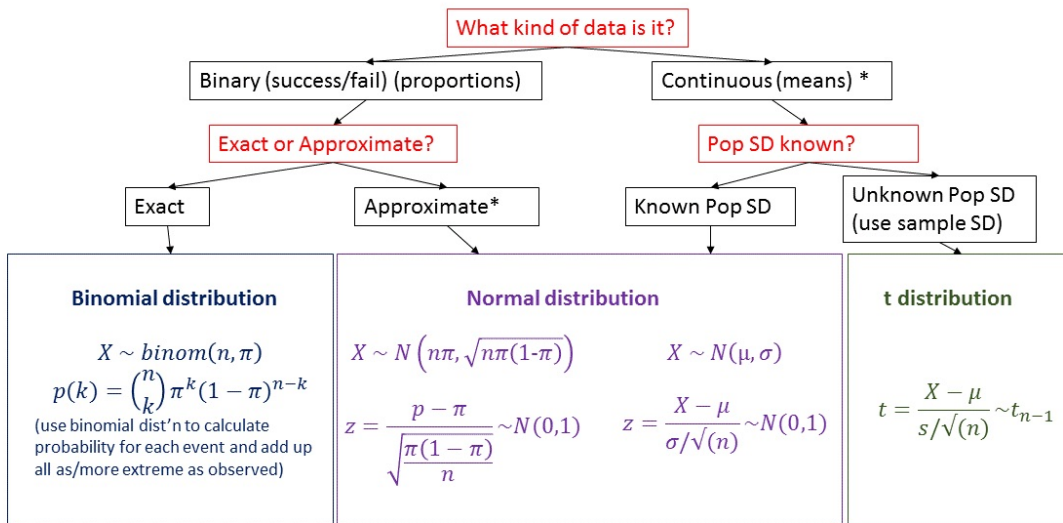
```
## [1] 1.598973 14.241215
```

Interpretation: We can say with 95% confidence that the true odds ratio for survival with sterile procedure relative to the control procedure is (1.599, 14.241), indicating significantly increased odds of survival on the sterile procedure compared to the non-sterile procedure. (Significant because the CI does not include the value 1.)

Note that this is the same as the relative odds of dying with the control surgery. This happens because the OR is symmetric.

Quiz 3 Review

A flowchart for when to use each distribution:



* : Okay if data is normal or n is large (CLT); ways to deal with this not being okay generally not covered in this class

Figure 1: Flowchart

General approach to statistics problems:

Step 1: Figure out and list the given information.

Write out the values of n, \bar{x}, s , etc. for everything given in the prompt.

Step 2: Figure out what the problem is asking. Write the null hypothesis (if applicable).

What is the population we're looking at? What kind of distributions can be used with the information given? Which is most appropriate in this situation? (see chart) Do we want a test or an interval? Most importantly,

what is the question being asked?

If Defining Hypotheses: The null case is the one based on no change, difference, or improvement. The alternative is what we want to show or what we are looking to prove. Remember that the hypotheses describe and apply to a population parameter.

Step 3: Write out the equation for the test statistic or interval.

In this class, most test statistics will similar to the form: $z = \frac{(\hat{X}-X_0)}{SE_{\hat{X}}}$, and most confidence intervals will look similar to this: $\hat{X} - z^* * SE_{\hat{X}} < X < \hat{X} + z^* * SE_{\hat{X}}$,

where: X is the parameter of interest

\hat{X} is the estimate of the parameter

X_0 is the value of the parameter under the null hypothesis

$SE_{\hat{X}}$ is the standard error of the estimate

z^* is the critical value for the interval

Step 4: Plug in the values.

Calculate the statistic or interval using the given values. Find the p-value (if applicable). This means that if we're conducting a test, compare the test statistic to the distribution we picked in step 2, and find the probability of being as or more extreme than the value we observed.

Step 5: Interpret in the context of the problem.

This might very well be the most important step. Our interpretation is dependent on our study, so it varies. Below are some examples.

Situation: Paired study looking at the effect of oat bran consumption compared to cornflakes levels where the calculated p-value was 0.005

“There is strong evidence that eating oat bran as opposed to corn flakes leads to lower cholesterol levels.”

Situation: Weights of U.S. Males (mean believed to be 172.2) where observed mean was 180, and the calculated p-value was 0.09

“This study provides is borderline evidence to suggest that the true mean weight of males in the United States is greater than 172.2 pounds.” **Modifiers based on p-value:**

p	Evidence against null
> .1	not
.1	borderline
.05	moderately
.01	strongly
.001	overwhelmingly

Other Notes:

Effect size:

Recall the Nexium/PriLOSEC example. If we have a large enough n, we can find statistical significance in the smallest true difference in means. Here, the difference, while real, was only 3%. That's effectively REALLY close to being the same. For practical purposes, this difference doesn't matter, but the p-value doesn't tell us

that the actual difference is little, just that it is significant.

Paired data:

Paired data can be analyzed using the binomial distribution (proportion of patients that saw any improvement) or using continuous methods (using 1-sample methods on a set of the differences between the two).

Practice Problems

1. We are interested in testing whether a certain at-risk population for diabetes has a daily sugar intake that is equal to the general population, which is equal to 77 grams/day. A sample of size 37 was taken from this at-risk population, and we obtained a sample mean of 80 and sample standard deviation of 11 grams.

Perform a hypothesis test to test whether this population has a significantly different mean sugar intake from 77 grams.

2. The distribution of LDL cholesterol levels in a certain population is approximately normal with mean 90 mg/dl and standard deviation 8 mg/dl.

- (a) What is the probability an individual will have a LDL cholesterol level above 95 mg/dl?
- (b) Suppose we have a sample of 10 people from this population. What is the probability of exactly 3 of them being above 95 mg/dl?
- (c) Take the sample of size 10, as in part b. What is the probability that the sample mean will be above 95 mg/dl?
- (d) Suppose we take 5 samples of size 10 from the population. What is the probability that at least one of the sample means will be greater than 95 mg/dl?

3.

In the city of Chicago, 235 baseball fans were sampled and asked if they cheer for the White Sox or the Cubs. 155 of the questioned people preferred the Cubs, while the remaining 80 fans preferred the White Sox. Use normal approximation to test the hypothesis that the Cubs and White Sox have an equal number of fans in the city at the $\alpha = .05$ level. Construct a 95% Confidence Interval for the true proportion of Chicago fans who cheer for the Cubs, still using normal approximation.

4.

An article in the New England Journal of Medicine reported that among adults living in the United States, the average level of albumin in cerebrospinal fluid is 29.5 mg/dl, with a standard deviation of 9.25 mg/dl. We are going to select a sample of size 20 from this population. Assume albumin levels in cerebrospinal fluid in U.S. adults are normally distributed.

- a) How does the variability of our sample mean compare with the variability of albumin levels in the population?
- b) What is the probability that our sample mean will lie between 29 and 31 mg/dl?
- c) What two values will contain the middle 50% of our sample means?
- d) Now assume we don't know the standard deviation or mean of the population and our sample of 20 has a mean albumin level of 30.1 mg/dl and a standard deviation of 8.95 mg/dl. Construct a 95% confidence interval for the population mean and interpret it.
- e) Why is the confidence interval constructed in (d) wider than the confidence interval that would be constructed if we used the normal distribution? What variables affect the width of the confidence interval?

5. In the following scenarios, identify what will happen to the power of a hypothesis test:

(a) We increase the sample size.

(b) The standard deviation of the sample is larger than what we expected.

(c) Our effect size moves from 5 units to 10 units.