Lab 11: Odds Ratios and Quiz 3 Review

April 4-5, 2017

New-ish Material (not on quiz 3)

A quick review of the categorical functions we learned last week:

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

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

```
# Chi-Squared Test
chisq.test(lister.table,correct = FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  lister.table
## X-squared = 8.4952, df = 1, p-value = 0.003561
```

```
# Fisher's Exact Test
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 the Fisher's Exact Test also gives you the odds ratio, which is calculated as $\frac{a/b}{c/d} = \frac{ad}{bc}$, with a, b, c, and d as in the following table.

```
##             Success Failure
## Thing       "a"     "b"
## Other Thing "c"     "d"
```

Interpretationally in this context, the easiest way to think of the odds ratio is to say that the odds of Success are 100*(OR-1) percent higher if we do the Thing than if we do the Other Thing.
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 the Thing instead of Success.

CI for OR

Calculating a confidence interval for an odds ratio is a tad more complicated than what we've done so far, since it's on a different scale. That being said, it's not too bad.

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

To be different but use the same data, let's find a CI for the OR for survival given non-sterile procedure. Now, our OR is going to be $\frac{b*c}{a*d}$.

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

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

1. Find the OR.

```
OR <- lister.table[1,2]*lister.table[2,1]/(lister.table[1,1]*lister.table[2,2])
print(OR)
```

```
## [1] 0.2095588
```

2. Find the log(OR)

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

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

3. Find the error term.

```
SElor<-sqrt(1/lister.table[1,2]+
              1/lister.table[2,1]+
              1/lister.table[1,1]+
              1/lister.table[2,2])
print(SElor)
```

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

4. Calculate the CI on this scale.

```
logCI <- lOR + qnorm(c(.025,.975)) * SElor
print(logCI)
```

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

5. Calculate the CI on the original scale.

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

```
## [1] 0.07021873 0.62540154
```

And finally, interpretation: We can say with 95% confidence that the true odds ratio for survival with non-sterile procedure is (0.07, 0.625), indicating significantly reduced odds of survival on the non-sterile procedure compared to the sterile procedure. (Significant because the CI does not include 1)

Notice that this is the same as the relative odds of dying on the sterile surgery (at the end of the 4/4 lecture notes). This happens because the OR is symmetric, and is actually a pretty cool result.

Quiz 3 Review

Firstly, a flowchart for when to use each distribution:

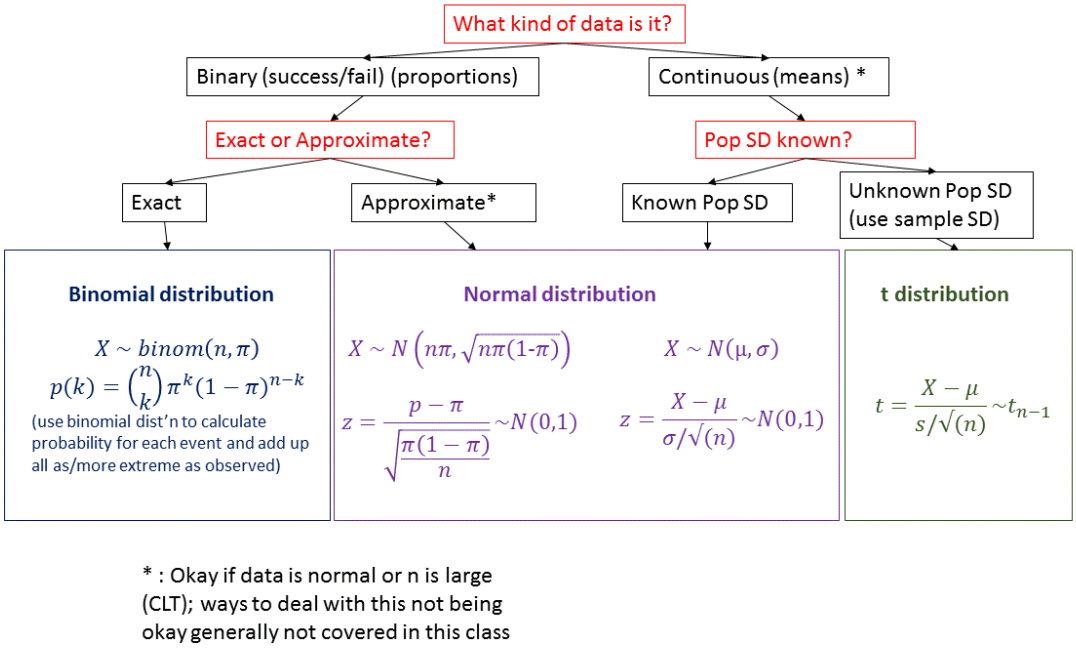

Figure 1: Fancy Schmancy Flowchart

Handy Guide for Answering Stats Problems

Stats problems can be confusing. Here's my process on how to approach them:

Step 1: Figure out what I'm given

I literally write out the values of $n, \bar{x}, s$, etc. for everything I'm given.

Step 2: Figure out what the problem is asking

What is the population I'm looking at? What kinds of distributions can I use with the information I have? Which is most appropriate in this situation? (I have a flow chart for this.) Do I want a test or an interval? Most importantly, what is the question being asked?

Step 2.5: Write out null and alternative hypotheses

(Skip this if I'm constructing an interval.)
Remember, the null case is the boring one, and the alternative is what I want to show. The hypotheses are written about a population parameter, so remember they should be Greek letters.

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

In this class, most statistics will be of the form $z = \frac{(\hat{X} - X_0)}{SE_{\hat{X}}}$, and most confidence intervals will look kind of like a deconstructed version of that: $\hat{X} - z^* * SE_{\hat{X}} < X < \hat{X} + z^* * SE_{\hat{X}}$.
(We derived this in class.)
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 numbers

Actually calculate the statistic or interval.

Step 4.5: Find p

If I'm running a test, compare my statistic to the distribution I picked in step 2 and find the probability of being as or more extreme than the value I observed.

Step 5: Interpret in context of the problem

This might very well be the most important step, and it's also probably the easiest. (Yay!)

Step 5a: Confidence intervals

Fill in the blanks in order to answer the question identified in step 2:
Based on this data, we can say with (confidence level) confidence that the true (parameter) lies with the interval (lower bound, upper bound).

Important note: The confidence level is a statement about the reliability of the process. If we repeat the experiment 100 times, we would expect 100*(confidence level) of the intervals we construct to contain the true parameter. We CANNOT make probability statements about the true parameter being within the interval. Don't do it. Please.

Step 5b: Hypothesis tests

Fill in the blanks in order to answer the question identified in step 2:
Based on this data, the (thing we're studying) is (modifier based on p) significantly (higher than/lower than/different from) (the null value) (p = [whatever p equals]).
Potentially add a statement on effect size.

Modifiers based on p (mostly taken from slide 12 of 1/24 notes):

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

A note on effect size

Recall the Nexium/Prilosec: 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 REALLY close to being the same. For practical purposes, we don't care. But the p value doesn't tell us that the actual difference is little, just that it is significant. So keep an eye out for when this happens and make a note of it when it does.

A note on 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 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.

Solutions

```
## Problem 1
print("Problem 1")

mu<-77
xbar<-80
s<-11
n<-37

# Don't have sigma, so running t-test
# H0: mu = 77
# HA: mu != 77

t<-(xbar-mu)/(s/sqrt(n))
2*pt(abs(t),n-1,lower.tail=FALSE)

# We do not have significant evidence to conclude that the at-risk mean sugar
#  intake is different from the general population mean sugar intake (p = 0.11).


## Problem 2
print("Problem 2")

mu<-90
sigma<-8

## Part 2a
print("Part 2a")

# p(x>95)
p<-pnorm(95,mu,sigma,lower.tail=FALSE)
#alternatively... (gives same answer)
x<-(95-90)/sigma
1-pnorm(x)

## Part 2b
print("Part 2b")

dbinom(3,10,p)
choose(10,3)*p^3*(1-p)^(10-3)

## Part 2c
print("Part 2c")

n<-10
z<-(95-90)/(sigma/sqrt(n))
pnorm(z,lower.tail=FALSE)
p<-pnorm(z,lower.tail=FALSE)


## Part 2d
print("Part 2d")
```

```r
1-dbinom(0,5,p)
1-choose(5,0)*p^0*(1-p)^(5-0)
```

```
## [1] "Problem 1"
## [1] 0.1058177
## [1] "Problem 2"
## [1] "Part 2a"
## [1] 0.2659855
## [1] "Part 2b"
## [1] 0.2592314
## [1] 0.2592314
## [1] "Part 2c"
## [1] 0.02405341
## [1] "Part 2d"
## [1] 0.1146189
## [1] 0.1146189
```

Problem 3
Part a: Power increases
Part b: Power decreases
Part c: Power increases