BIOS: 4120 Lab 9

March 20-21, 2018

Prior to break, we covered one-sample categorical data, and in today's lecture we discussed one-sample continuous data. In this lab, we will be conducting hypothesis tests and creating confidence intervals for both categorical and continuous data.

Note: A hat on a Greek letter indicates an estimator, so for example, when you see $\hat{\mu}$, this is the same thing as $\bar{x}$.

Example 1: a z-test for categorical data

Suppose the incidence rate of myocardial infarction per year was 0.005 among males age 45-54 in 1970. For 1 year starting in 1980, 5000 males age 45-54 were followed, and 15 new myocardial infarction cases were observed.

From the central limit theorem, we know that the sample proportion approximately follows a normal distribution (if the sample size is reasonably large), so we can perform a z-test on this data.

Conduct a hypothesis test to determine if true myocardial infarction rate changed from 1970 to 1980. *How would you interpret the result?*

$H_0 : \pi = 0.005$

$H_A : \pi \neq 0.005$

$\pi = 0.005$
$\hat{\pi} = \frac{15}{5000} = 0.003$
$n = 5000$

$SE = \sqrt{\frac{\pi(1-\pi)}{n}}$

$SE = \sqrt{\frac{0.005(1-0.005)}{5000}}$

$SE = 0.000997$

Remember that to compute a test statistic we use:

$z = \frac{\hat{\pi}-\pi}{SE}$

$z = \frac{0.003-0.005}{0.000997}$

$z = -2.01$

Find 2-tailed probability by looking up this z-score on the z-table:

$p = 2(0.022) = 0.044$

Interpretation: Based on this data, there is significant evidence to suggest that the true myocardial infarction rate changed from 1970 to 1980 (p = 0.044).

Using R:

We can use the 'pnorm' function to calculate this p-value in R.

```r
round(2*pnorm(2.01,mean=0,sd=1,lower.tail=FALSE),5)
```

```
## [1] 0.04443
```

```r
# OR
round(2*(1-pnorm(2.01,mean=0,sd=1)),5)
```

```
## [1] 0.04443
```

We can compare this to what we would get use the exact test using binom.test().

```r
binom.test(15, 5000, p = 0.005)
```

```
##
##  Exact binomial test
##
## data:  15 and 5000
## number of successes = 15, number of trials = 5000, p-value =
## 0.04422
## alternative hypothesis: true probability of success is not equal to 0.005
## 95 percent confidence interval:
##  0.001680019 0.004943224
## sample estimates:
## probability of success
##                  0.003
```

From the p-value that's given (p = 0.04422)we are able to see that normal approximation is virtually identical to the exact binomial test. *Why do you think this is especially when p is so close to 0?*

Creating a confidence interval (z)

Now we want to create a 95% confidence interval for p. *Interpret the interval.*

Remember that now standard error is based on $\hat{\pi}$ and becomes:

$$SE = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

$$SE = \sqrt{\frac{0.003(1-0.003)}{5000}}$$

$$SE = 0.000773$$

We will have to find $z_{\alpha/2}$ using the z-table. *What is our α for a 95% confidence interval?*

$z_{\alpha/2} = 1.96$ (from table)

Remember that the equation for the confidence interval is:

$$\hat{\pi} \pm z_{\alpha/2} * SE$$
$$0.003 \pm 1.96 * 0.000773$$
$$(0.0015, 0.0045)$$

Interpretation: We can say with 95% confidence that this interval contains the true myocardial infarction rate in 1980.

Interpretation Note: Remember that when we say "95% confidence" about an interval, this does NOT mean that there is a 95% probability of the true parameter being in the interval. It means that if we were to repeat this experiment a bunch of times, 95% of the intevals constructed in this manner would contain the true parameter. It's a bit of a touchy subject, so overall just be careful to not say "probability" when you're interpreting confidence intervals.

Using R:

To calculate a confidence interval in R, we use the 'qnorm' function.

```
0.003 + qnorm(0.025)* sqrt((0.003*(1-0.003))/5000)
```

```
## [1] 0.001484097
```

```
0.003 + qnorm(0.975)* sqrt((0.003*(1-0.003))/5000)
```

```
## [1] 0.004515903
```

This can also be found in one step using a vector as shown below:

```
0.003 + qnorm(c(0.025,0.975))* sqrt((0.003*(1-0.003))/5000)
```

```
## [1] 0.001484097 0.004515903
```

Notice that this confidence interval varies a bit from the confidence interval created using 'binom.test'(0.00168, 0.00494). *What may be the cause of this?*

Now that we've seen an example of categorical data, let's look at a continuous data example.

Example 2: a t-test

The distribution of weights for the population of males in the United States is approximately normal. We believe the mean $\mu = 172.2$. We conduct an experiment with a sample size of 50, and we find our sample mean to be 180 and the sample standard deviation to be 30. Conduct a hypothesis test to determine if the true mean is 172.2 based on our data. *How would you interpret the result?*

$H_0 : \mu = 172.2$

$H_A : \mu \neq 172.2$

$\mu = 172.2$
$\hat{\mu} = 180$
$s = 30$
$n = 50$
$df = n - 1 = 49$

To compute a test statistic we use:

$t = \frac{\hat{\mu} - \mu}{s/\sqrt{n}}$

$t = \frac{180 - 172.2}{30/\sqrt{50}}$

$t = 1.84$

Find 2-tailed probability using this test statistic and Student's t-table:

$0.05 < p < 0.1$

Interpretation: There is borderline (but not significant) evidence to suggest that the true mean weight of males in the United States is greater than 172.2, based on this data (0.05 < p < 0.1).

Using R:

We can use the 'pt' function to calculate this p-value in R.

```r
mu <- 172.2
mu.hat <- 180
s <- 30
n <- 50

t <- (mu.hat-mu)/(s/sqrt(n))

2*pt(1.84, df=n-1,lower.tail=FALSE)
```

```
## [1] 0.07182936
```

Notice that this p-value (while more precise) fits with what we were able to calculate by hand.

Constructing a confidence interval (t)

Now we want to create a 95% confidence interval for μ. *Interpret the interval.*

$\hat{\mu} = 180$
$s = 30$
$n = 50$

$SE = \frac{30}{\sqrt{50}}$

Remember that the equation for creating a confidence interval is:

$\hat{\mu} \pm z_{\alpha/2} * SE$

We can then find $t_{\alpha/2}$, plug in our given values, and calculate the interval.

$t_{\alpha/2} = 2.01$ (from table)

$180 \pm 2.01 * \frac{30}{\sqrt{50}}$

(171.4, 188.5)

Interpretation: We can say with 95% confidence that this interval contains the true mean weight of males in the US.

Using R:

```r
mu.hat + qt(c(.025,.975), n-1)*s/sqrt(n)
```

```
## [1] 171.4741 188.5259
```

```r
# which is the same as
180 + qt(c(.025,.975), 49)*30/sqrt(50)
```

```
## [1] 171.4741 188.5259
```

Practice Problem:

Suppose that the average IQ is 100. Perform a test to see if the children in the lead-IQ dataset have an average IQ. Also, create a 95% confidence interval for the mean IQ based on this data.