STAT 215: Lab for Week 5

Tests and Intervals Using Analytic Approximation

Last revised September 29, 2017

The goal of this lab is to get some practice doing hypothesis tests and constructing confidence intervals using theoretical approximations to sampling distributions of a standardized statistic.

We will use the `mosaic` and `Lock5Data` packages in this lab.

```
library("mosaic")
library("Lock5Data")
```

1 Working With Distributions in R

We often need to calculate areas under the tails of a Normal curve (e.g., to compute P-values), as well as to find particular quantiles of the Normal (e.g., to find the endpoints for a confidence interval). You can use StatKey to do this by selecting "Theoretical Distributions: Normal" and then "Edit Parameters" to change the mean and standard deviation. After that the interface is like what we have used for bootstrap and randomization distributions.

Or you can use R. The main R functions to do this are `xpnorm()` and `xqnorm()`.

Both use the following pattern (elements in caps, other than `TRUE` and `FALSE`, should be replaced by values).

```
## returns area to the left of CUTOFF in a N(MEAN,SD) distribution
xpnorm(CUTOFF, mean = MEAN, sd = SD, lower.tail = TRUE)
## returns area to the right of CUTOFF
xpnorm(CUTOFF, mean = MEAN, sd = SD, lower.tail = FALSE)
## returns the Pth quantile of a N(MEAN, SD) distribution
## (i.e., the value with proportion P below it)
xqnorm(PROPORTION, mean = MEAN, sd = SD, lower.tail = TRUE)
## returns the (1 - P)th quantile of a N(MEAN, SD) distribution
## (i.e., the value with proportion P above it)
xqnorm(PROPORTION, mean = MEAN, sd = SD, lower.tail = FALSE)
```

1.1 Warmup Exercises

Use the `xpnorm()` or `xqnorm()` function to answer the following questions involving normal distributions. (You don't need to turn these in)

1. Find the specified areas for a normal density.
   (a) The area above 62 in a N(50,10) density.
   (b) The area below 8 in a N(10,2) density.
2. Find the endpoint, x on the given normal density curve with the given property.
   (a) The area to the right of x on a N(10,4) density curve is about 0.05.
   (b) The area to the left of x on a N(100,25) density curve is about 0.35.
3. Suppose weights of newborn babies in one community are normally distributed with a mean of 7.5 pounds and a standard deviation of 1.2 pounds.
   (a) Use the 95% rule to sketch a graph of this normal density curve. Include a scale with at least three values on the horizontal axis.
   (b) What percent of newborns weigh less than 5 pounds?
   (c) What percent of newborns weight more than 11 pounds?
   (d) If a newborn baby is at the 15th percentile for weight, what is the baby's weight?

2 Intervals and Tests for a Single Proportion

This section is provided for reference.

The key idea of this section is that the theoretical standard error of a distribution of sample proportions is

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

and so we can use this wherever we would otherwise need to get a standard error from a bootstrap or a randomization distribution consisting of sample proportions.

The choice of p in this formula is the center of whichever distribution we are constructing. Should the distribution be centered at the observed sample statistic, $\hat{p}$ as it is when we construct a confidence interval? Or should it be centered around the hypothetical parameter value stated by the null hypothesis, as when computing a P-value?

When building a confidence interval, we use

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

but when computing a P-value for a null hypothesis that states that the population proportion $p = p_0$, we use

$$SE = \sqrt{\frac{p_0(1-p_0)}{n}}$$

If the expected counts in the sample for both values of the binary response variable are at least 10 or so; that is, $np \geq 10$ *and* $n(1-p) \geq 10$, then we can use a Normal distribution when computing the P-value or the confidence interval. Here, as before, we use the center of the relevant distribution for p: $\hat{p}$ for a confidence interval and p_0 for a P-value.

2.1 Confidence Interval for a Proportion

When these conditions are satisfied, we can build a confidence interval by finding the values that surround the middle C% of the distribution. These values are the (100-C)/2 percentile and the 100 - (100-C)/2 percentile, since the remaining (100-C)% is split between the two extremes.

To construct a confidence interval, we can start with the critical Z-scores, which are the corresponding quantiles of a $\mathcal{N}(0,1)$ distribution, and then convert these back to the original scale with the inverse Z-transformation:

$$\text{endpoint} = \text{observed statistic} + Z^* \times SE$$

where, in this case, the observed statistic is $\hat{p}$ and the standard error is $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$:

$$\text{endpoint} = \hat{p} + Z^* \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

We can use R to calculate the critical values, which we can then convert back to the original scale.

```
### lower.proportion and upper.proportion are the proportions that
### we want to be to the left of the lower and upper cutoff
### in the standardized distribution.
### E.g. for 95% confidence, lower.proportion = 0.025
### and upper.proportion = 0.975
zstar.lower <- xqnorm(lower.proportion)
zstar.upper <- xqnorm(upper.proportion)
### Then, having computed p.hat and se elsewhere, we get a CI:
CI.lower <- p.hat + zstar.lower * se; CI.upper <- p.hat + zstar.upper * se
```

2.2 Hypothesis Test for a Proportion

Recall that when we do hypothesis tests, we build a distribution on the assumption that the null hypothesis is true. This has generally been a *randomization distribution*, and is centered at the null parameter value.

But, when the conditions for a Normal approximation hold, we can use a Normal distribution instead; computing a P-value by finding the area under the Normal which lies beyond the observed sample statistic.

As with a confidence interval, we can either use the Normal with the null mean and the standard error based on the null parameter and the sample size.

We can convert the observed statistic to a z-score, and locate the z-score in the

distribution of z-scores, which is a Standard Normal:

$$Z_{obs} = \frac{\text{observed stat} - \text{null parameter}}{\text{standard error}} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Having calculated the standardized test statistic, Z_{obs} we can compute the P-value in R:

```
## lower.tail = TRUE vs FALSE governs whether we get the right or left
## tail probability
P.value <- xpnorm(z.observed, mean = 0, sd = 1, lower.tail = ... )
## OR, for a two-tailed test, we can get the right tail proportion
## (forcing a positive z by taking the absolute value)
## and multiply it by 2 to take the other (symmetric) tail into
## account as well
two.tailed.P <-
    2 * xpnorm(abs(z.observed), mean = 0, sd = 1, lower.tail = FALSE)
```

2.3 Homework about Inferring a Single Proportion

1. In a sample of 120 soccer matches played in the Football Association (FA) premier league in Great Britain, the home team won 70 times. Create a randomization distribution of 1000 sample proportions to test whether is evidence for a home field advantage in this league. (You can use the `do() * rflip()` construction in R, or use StatKey. To see how to use `rflip()` you can either consult the lecture slides, or use the built in documentation (i.e., type `?rflip`)

   (a) Find a one and two-tailed P-value using the randomization distribution.

   (b) Consider using a normal distribution to model the sampling distribution of proportions under H_0 instead of the randomization distribution. What should the mean and standard deviation be? Calculate a P-value using this method.

   (c) Compare the answer from the normal distribution to what you found from the randomization distribution. Are the results similar?

   (d) Construct a 99% confidence interval for the "long run" home team win percentage using the Normal model (use critical z-scores to find the multiplier on the standard error).

3 Difference Between Two Proportions

If we want to compare the proportion of a particular outcome between two groups, the population parameter we are focused on is the difference between population parameters, $p_A - p_B$, and we use the difference of sample proportions, $\hat{p}_A - \hat{p}_B$ as our point estimate.

If we considered the two sample proportions separately, each would have standard error $\sqrt{\frac{p(1-p)}{n}}$, for the appropriate choice of p, and for n the sample size in the respective group. Since both of these are random, we can get even bigger differences by random chance. In general, the *variance* of a sum or difference of two random quantities that are varying independently of one another is the *sum* of the variances of each on its own. The *standard error* is the square root of the variance. Thus, we have

$$SE_{\hat{p}_A - \hat{p}_B} = \sqrt{\frac{p_A(1-p_A)}{n_A} + \frac{p_B(1-p_B)}{n_B}}$$

The conditions for Normality are the same as in the single proportion case, but need to be checked for *both* groups:

- $n_A p_A \geq 10$ and $n_A(1-p_A) \geq 10$
- $n_B p_B \geq 10$ and $n_B(1-p_B) \geq 10$

Confidence Interval We do not need to use a t-distribution here, since the population standard deviation is a function only of the proportions, so we can simply use

$$\begin{aligned} \text{endpoint} &= \text{observed statistic} + Z^* \times SE \\ &= (\hat{p}_A - \hat{p}_B) + Z^* \times \sqrt{\frac{\hat{p}_A(1-\hat{p}_A)}{n_A} + \frac{\hat{p}_B(1-\hat{p}_B)}{n_B}} \end{aligned}$$

for a confidence interval, where Z^* is the appropriate quantile from a Stanard Normal

Hypothesis Test The null hypothesis usually states that the two proportions are equal (and so their difference is zero), but does not specify what that value is. What value do we use when calculating the standard error?

We can use the null assumption that there is only one big group, and use the *combined* sample proportion as our null proportion. Then our Z-statistic becomes

$$\begin{aligned} Z_{obs} &= \frac{\text{observed stat} - \text{null parameter}}{\text{standard error}} \\ &= \frac{\hat{p}_A - \hat{p}_B - 0}{\sqrt{\frac{\hat{p}_{combined}(1-\hat{p}_{combined})}{n_A} + \frac{\hat{p}_{combined}(1-\hat{p}_{combined})}{n_B}}} \\ &= \frac{\hat{p}_A - \hat{p}_B - 0}{\sqrt{\hat{p}_{combined}(1-\hat{p}_{combined})\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}} \end{aligned}$$

which we can locate in the Standard Normal for a P-value.

Example: Penguins The scientists who studied whether metal bands were harmful to penguin survival also examined whether they affected the penguins' breeding patterns. The metal-band penguins successfully bred in 32% of 122 breeding seasons (combined across penguins), whereas the controls had offpsring in 44% of 160 breeding seasons.

Let p_{metal} be the long run breeding success rate for penguins wearing metal bands, and $p_{control}$ be the corresponding rate for the controls.

Let's first construct a 99% confidence interval for the difference in these proportions: $p_{metal} - p_{control}$.

The observed statistic is

$$\hat{p}_{metal} - \hat{p}_{control} = 0.32 - 0.44 = -0.12$$

Let's store this in an R variable as well, along with the sample sizes:

```
n.m <- 122; n.c <- 160
phat.m <- 0.32; phat.c <- 0.44
phat.m.minus.phat.c <- phat.m - phat.c
```

Let's check the normality assumptions

```
n.m * phat.m
n.m * (1 - phat.m)
n.c * phat.c
n.c * (1 - phat.c)
```

All are greater than 10, so we are safe to use the Normal model.

The critical Z-scores come from the standard Normal

```
## mean= and sd= left out since we're working with a standard Normal
zstar.99 <- xqnorm(c(0.005, 0.995))
```

The standard error is computed using

$$SE_{\hat{p}_m - \hat{p}_c} = \sqrt{\frac{p_m(1-p_m)}{n_m} + \frac{p_c(1-p_c)}{n_c}}$$

```
se <- sqrt(phat.m * (1 - phat.m) / n.m + phat.c * (1 - phat.c) / n.c)
```

Finally, convert the critical Z values back to the difference-in-proportions scale:

```
(CI.99 <- phat.m.minus.phat.c + zstar.99 * se)
```

Now, the standard error is obtained by first finding the combined observed success proportion

```
## We have to figure out the success counts to get the combined proportion
p.combined <- (phat.m * n.m + phat.c * n.c) / (n.m + n.c)
se <- sqrt(p.combined * (1 - p.combined) * ( 1 / n.m + 1 / n.c))
```

Now, let's test whether the difference is significantly negative (a one-tailed test) using a 0.05 significance level.

$$H_0 : p_{metal} - p_{control} = 0$$
$$H_1 : p_{metal} - p_{control} < 0$$

For the standard error, we need to find $\hat{p}_{combined}$. First we compute the breeding counts in each group, so that we can sum them and divide by the combined sample size:

```
(p.combined <- (n.m * phat.m + n.c * phat.c) / (n.m + n.c))
```

Now we can compute the standard error using this combined proportion

```
(se.combined <- sqrt(p.combined * (1 - p.combined) * (1 / n.m + 1 / n.c)))
```

Now, find the observed z-score:

```
(z.obs <- (phat.m.minus.phat.c - 0) / se.combined)
```

and find the one-tailed P-value:

```
xpnorm(z.obs, lower.tail = TRUE)
```

so we can reject H_0 at the 0.05 level.

3.1 Homework about Two Proportions

4. In the dataset `ICUAdmissions` in the `Lock5Data` package, the variable `Status` indicates whether the ICU (Intensive Care Unit) patient lived (0) or died (1), while the variable `Infection` indicates whether the patient had an infection (1 for yes, 0 for no) at the time of admission to the ICU. Find a 95% confidence interval for the difference in the proportion who die between those with an infection and those without."
5. Test whether the above difference in proportions is significantly different from 0.
6. Check your results using the built in `prop.test()` function, with the form

   ```
   prop.test(Response ~ Explanatory, data = DataSet, conf.level = ...)
   ```