Lab 6: Inference for Categorical Data

North Carolina births

This lab will utilize the same North Caroline birth data used in the previous inference lab (lab 5: Inference for numerical data). To refresh your memory, lets load the data and take a look at the names and formats of the different variables.

```
download.file("http://www.openintro.org/stat/data/nc.RData", destfile = "nc.RData")

load("nc.RData")

str(nc)
```

Exercise 1 In this data, what does the variable **lowbirthweight** describe? What about the vairable **habit**?

Exercise 2 Using the command below, create a new dataframe called **smokers** that contains only the rows in **nc** associated with moms who are smokers. Next, calculate proportion of smokers who had babies with low birth weights.

```
smokers <- subset(nc, nc$habit == "smoker")
```

Inference on proportions

Above we calculated the proportion of individuals in our sample who smoke and had newborns who had low birth weights. This proportion that we calculated is a statistic. What we'd like, though, is insight into the population *parameters*. You answer the question, "What proportion of people in your smoker sample had babies with low birth weights?" with a statistic; while the question "What proportion of mothers who smoke on earth would have babies with low birth weights" is answered with an estimate of the parameter.

The inferential tools for estimating population proportion are analogous to those used for means in the last chapter: the confidence interval and the hypothesis test.

Exercise 3 Write out the conditions for inference to construct a 95% confidence interval for the proportion of smokers who had babies with low birth weights in North Carolina. Are you confident all conditions are met?

If the conditions for inference are reasonable, we can either calculate the standard error and construct the interval by hand, or allow the **inference** function to do it for us.

```
inference(y = smokers$lowbirthweight, est = "proportion", type = "ci", method = "theoretical",
          success = "low")
```

Note that since the goal is to construct an interval estimate for a proportion, it's necessary to specify what constitutes a "success", which here is an entry in the field **lowbirthweight** of *low*.

Exercise 4 Based on the R output, what is the margin of error for the estimate of the proportion of mom smokers who had babies with a low birth weight?
Hint: the formula for the margin of error for a 95% confidence interval is ME $= 1.96\times$ (Standard Error).

Exercise 5 Using the **inference** function, calculate confidence intervals for the proportion of non-smokers who had babies with a low birth weight, and report the associated margins of error. Be sure to note whether the conditions for inference are met. It may be helpful to create a new data set for nonsmokers first, and then use this data set in the **inference** function to construct the confidence intervals.

How does the proportion affect the margin of error?

Imagine you've set out to survey 1000 people on two questions: are you female? and are you left-handed? Since both of these sample proportions were calculated from the same sample size, they should have the same margin of error, right? Wrong! While the margin of error does change with sample size, it is also affected by the proportion.

Think back to the formula for the standard error: $SE = \sqrt{p(1-p)/n}$. This is then used in the formula for the margin of error for a 95% confidence interval: $ME = 1.96 \times SE = 1.96 \times \sqrt{p(1-p)/n}$. Since the population proportion p is in this ME formula, it should make sense that the margin of error is in some way dependent on the population proportion. We can visualize this relationship by creating a plot of ME vs. p.

The first step is to make a vector **p** that is a sequence from 0 to 1 with each number separated by 0.01. We can then create a vector of the margin of error (**me**) associated with each of these values of **p** using the familiar approximate formula ($ME = 2 \times SE$). Lastly, we plot the two vectors against each other to reveal their relationship.

```
n <- 1000
p <- seq(0, 1, 0.01)
me <- 2 * sqrt(p * (1 - p)/n)
plot(me ~ p)
```

Exercise 6 Describe the relationship between **p** and **me**.

Success-failure condition

As we have done in the last two labs, it is important that you always check conditions before making inference. For inference on proportions, the sample proportion can be assumed to be nearly normal if it is based upon a random sample of independent observations and if both $np \geq 10$ and $n(1-p) \geq 10$. This rule of thumb is easy enough to follow, but it makes one wonder: what's so special about the number 10? The short answer is: nothing. You could argue that we would be fine with 9 or that we really should be using 11. What is the "best" value for such a rule of thumb is, at least to some degree, arbitrary.

We can investigate the interplay between n and p and the shape of the sampling distribution by using simulations. To start off, we simulate the process of drawing 5000 samples of size 1040 from a smoker population with a true low birth weight proportion of 0.1. For each of the 5000 samples we compute $\hat{p}$ and then plot a histogram to visualize their distribution.

```
p <- 0.1
n <- 1040
p_hats <- rep(0, 5000)
```

```
for (i in 1:5000) {
    samp <- sample(c("low", "not low"), n, replace = TRUE, prob = c(p, 1 - p))
    p_hats[i] <- sum(samp == "low")/n
}

hist(p_hats, main = "p = 0.1, n = 1040", xlim = c(0, 0.18))
```

These commands build up the sampling distribution of **p_hats** using the familiar **for** loop. You can read the sampling procedure for the first line of code inside the **for** loop as, "take a sample of size n with replacement from the choices of low and not low birth weights with probabilities p and $1 - p$, respectively." The second line in the loop says, "calculate the proportion of low birth weights in this sample and record this value." The loop allows us to repeat this process 5,000 times to build a good representation of the sampling distribution.

Exercise 7 Describe the sampling distribution of sample proportions at $n = 1040$ and $p = 0.1$. Be sure to note the center, spread, and shape.
Hint: Remember that R has functions such as **mean** to calculate summary statistics.

Exercise 8 Replicate the above simulation three more times but with modified sample sizes and proportions: for $n = 400$ and $p = 0.1$, $n = 1040$ and $p = 0.02$, and $n = 400$ and $p = 0.02$. Plot all four histograms together by running the **par(**mfrow = **c(2,2))** command before creating the histograms. You may need to expand the plot window to accommodate the larger two-by-two plot. Describe the three new sampling distributions. Based on these limited plots, how does n appear to affect the distribution of $\hat{p}$? How does p affect the sampling distribution?

Once you're done, you can reset the layout of the plotting window by using the command **par(**mfrow = **c(1,1))** or clicking on "Clear All" above the plotting window (if using RStudio). Note that the latter will get rid of all your previous plots.

On your own

1. Answer the following two questions using the **inference** function. As always, write out the hypotheses for any tests you conduct and outline the status of the conditions for inference.

   (a) Is there convincing evidence that the proportion of smokers who had babies with low birth weights differs from the proportion of non-smokers who had babies with low birth weights?
   Hint: Use **lowbirthweight** as the first input on the **inference**, and use **habit** as the grouping variable.

   (b) Is there convincing evidence that the proportion of married mothers with babies with low birth weights differs from that of single mothers?

2. Suppose you're hired by the local government to estimate the proportion of mothers that visit the the doctor's office at least 12 times before they give birth. According to the guidelines, the estimate must have a margin of error no greater than 1% with 95% confidence. You have no idea what to expect for p. How many people would you have to sample to ensure that you are within the guidelines?
   Hint: Refer to your plot of the relationship between p and margin of error. Do not use the data set to answer this question.