Tutorial 5: Associations between variables

Johannes Karreth

RPOS 517, Day 5

This tutorial shows you:

- how to conduct a difference-of-means (t-) test
- how to conduct a difference-of-proportions test

Note on copying & pasting code from the PDF version of this tutorial: Please note that you may run into trouble if you copy & paste code from the PDF version of this tutorial into your R script. When the PDF is created, some characters (for instance, quotation marks or indentations) are converted into non-text characters that R won't recognize. To use code from this tutorial, please type it yourself into your R script or you may copy & paste code from the *source file* for this tutorial which is posted on my website.

Associations between numerical variables

Background: North Carolina births

In 2004, the state of North Carolina released a large data set containing information on births recorded in this state. This data set is useful to researchers studying the relation between habits and practices of expectant mothers and the birth of their children. We will work with a random sample of observations from this data set.

Exploratory analysis

Load the `nc` data set into our workspace.

```
nc <- read.csv("http://www.jkarreth.net/files/nc.csv")
```

We have observations on 13 different variables, some categorical and some numerical. The meaning of each variable is as follows.

variable	description
`fage`	father's age in years.
`mage`	mother's age in years.
`mature`	maturity status of mother.
`weeks`	length of pregnancy in weeks.
`premie`	whether the birth was classified as premature (premie) or full-term.
`visits`	number of hospital visits during pregnancy.
`marital`	whether mother is `married` or `not married` at birth.
`gained`	weight gained by mother during pregnancy in pounds.
`weight`	weight of the baby at birth in pounds.
`lowbirthweight`	whether baby was classified as low birthweight (`low`) or not (`not low`).
`gender`	gender of the baby, `female` or `male`.
`habit`	status of the mother as a `nonsmoker` or a `smoker`.
`whitemom`	whether mom is `white` or `not white`.

1. What are the cases in this data set? How many cases are there in our sample?

As a first step in the analysis, we should consider summaries of the data. This can be done using the `summary` command:

```
summary(nc)
```

```
##       fage            mage            mature        weeks      
##  Min.   :14.00   Min.   :13   mature mom :133   Min.   :20.00  
##  1st Qu.:25.00   1st Qu.:22   younger mom:867   1st Qu.:37.00  
##  Median :30.00   Median :27                     Median :39.00  
##  Mean   :30.26   Mean   :27                     Mean   :38.33  
##  3rd Qu.:35.00   3rd Qu.:32                     3rd Qu.:40.00  
##  Max.   :55.00   Max.   :50                     Max.   :45.00  
##  NA's   :171                                    NA's   :2      
##        premie        visits            marital        gained      
##  full term:846   Min.   : 0.0   married    :386   Min.   : 0.00  
##  premie   :152   1st Qu.:10.0   not married:613   1st Qu.:20.00  
##  NA's     :  2   Median :12.0   NA's       :  1   Median :30.00  
##                  Mean   :12.1                     Mean   :30.33  
##                  3rd Qu.:15.0                     3rd Qu.:38.00  
##                  Max.   :30.0                     Max.   :85.00  
##                  NA's   :9                        NA's   :27     
##      weight       lowbirthweight    gender          habit       
##  Min.   : 1.000   low    :111    female:503   nonsmoker:873  
##  1st Qu.: 6.380   not low:889    male  :497   smoker   :126  
##  Median : 7.310                               NA's     :  1  
##  Mean   : 7.101                                              
##  3rd Qu.: 8.060                                              
##  Max.   :11.750                                              
##                                                              
##       whitemom    
##  not white:284  
##  white    :714  
##  NA's     :  2  
##                 
##                 
##                 
##
```

The first *new* thing you may notice in this summary is that for some variables, the summary contains an extra row "NA's". This row lists the number of missing observations for each variable: observations (in this case, births) for which the information on the respective variable was not coded. This has implications for your how you handle these data in your analysis; we will return to this later.

As you review the variable summaries, consider which variables are categorical and which are numerical. For numerical variables, are there outliers? If you aren't sure or want to take a closer look at the data, make a graph.

Consider the possible relationship between a mother's smoking habit and the weight of her baby. Plotting the data is a useful first step because it helps us quickly visualize trends, identify strong associations, and develop research questions.

2. Make a side-by-side boxplot of `habit` and `weight`. What does the plot highlight about the relationship between these two variables?

The box plots show how the medians of the two distributions compare, but we can also compare the means of the distributions using the following function to split the `weight` variable into the `habit` groups, then take the mean of each using the `mean` function.

```
by(data = nc$weight, INDICES = nc$habit, FUN = mean)
```

```
## nc$habit: nonsmoker
## [1] 7.144273
## ------------------------------------------------------------
## nc$habit: smoker
## [1] 6.82873
```

There is an observed difference, but is this difference statistically significant? In order to answer this question we will conduct a hypothesis test .

Inference

3. Check if the conditions necessary for inference are satisfied. Note that you will need to obtain sample sizes to check the conditions. You can compute the group size using the same `by` command above but replacing `mean` with `length`.
4. Write the hypotheses for testing if the average weights of babies born to smoking and non-smoking mothers are different.

Next, we introduce a new function, `t.test`, that we will use for conducting hypothesis tests and constructing confidence intervals.

```
t.test(weight ~ habit, data = nc, alternative = "two.sided", conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  weight by habit
## t = 2.359, df = 171.325, p-value = 0.01945
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.05151165 0.57957328
## sample estimates:
## mean in group nonsmoker    mean in group smoker
##                7.144273                6.828730
```

Let's pause for a moment to go through the arguments of this base function. The first argument is `y`, which is the response variable that we are interested in: `nc$weight`. The second argument is the explanatory variable, `x`, which is the variable that splits the data into two groups, smokers and non-smokers: `nc$habit`. The two variables are separated by a tilde in the form `y ~ x`; we'll encounter this "language" again later. The third argument, `data`, is the name of the dataset in which the two variables can be found. Next we specify the `alternative` hypothesis of our test; it can be `"less"`, `"greater"`, or `"twosided"`. Lastly, the `conf.level` specifies the confidence level for this test. By default the function reports an interval for $(\mu_{nonsmoker} - \mu_{smoker})$.

On your own

- Conduct a hypothesis test evaluating whether the average weight gained by younger mothers is different than the average weight gained by mature mothers.
- Now, a non-inference task: Determine the age cutoff for younger and mature mothers. Use a method of your choice, and explain how your method works.
- Pick a pair of numerical and categorical variables and come up with a research question evaluating the relationship between these variables. Formulate the question in a way that it can be answered using a hypothesis test and/or a confidence interval. Answer your question using the appropriate function in R, report the statistical results, and also provide an explanation in plain language.

Associations between categorical variables

In August of 2012, news outlets ranging from the Washington Post to the Huffington Post ran a story about the rise of atheism in America. The source for the story was a poll that asked people, "Irrespective of whether you attend a place of worship or not, would you say you are a religious person, not a religious person or a convinced atheist?" This type of question, which asks people to classify themselves in one way or another, is common in polling and generates categorical data. In this lab we take a look at the atheism survey and explore what's at play when making inference about population proportions using categorical data.

The survey

To access the press release for the poll, conducted by WIN-Gallup International, click on the following link:

http://www.wingia.com/web/files/richeditor/filemanager/Global_INDEX_of_Religiosity_and_Atheism_PR___6.pdf

Take a moment to review the report then address the following questions.

1. In the first paragraph, several key findings are reported. Do these percentages appear to be *sample statistics* (derived from the data sample) or *population parameters*?
2. The title of the report is "Global Index of Religiosity and Atheism". To generalize the report's findings to the global human population, what must we assume about the sampling method? Does that seem like a reasonable assumption?

The data

Turn your attention to Table 6 (pages 15 and 16), which reports the sample size and response percentages for all 57 countries. While this is a useful format to summarize the data, we will base our analysis on the original data set of individual responses to the survey. Load this data set into R with the following command.

```
atheism <- read.csv("http://www.jkarreth.net/files/atheism.csv")
```

3. What does each row of Table 6 correspond to? What does each row of `atheism` correspond to?

To investigate the link between these two ways of organizing this data, take a look at the estimated proportion of atheists in the United States. Towards the bottom of Table 6, we see that this is 5%. We should be able to come to the same number using the `atheism` data.

4. Using the command below, create a new dataframe called `us12` that contains only the rows in `atheism` associated with respondents to the 2012 survey from the United States. Next, calculate the proportion of atheist responses. Does it agree with the percentage in Table 6? If not, why?

```
us12 <- subset(atheism, nationality == "United States" & year == "2012")
```

Inference on proportions

As was hinted at in Exercise 1, Table 6 provides *statistics*, that is, calculations made from the sample of 51,927 people. What we'd like, though, is insight into the population *parameters*. You answer the question, "What proportion of people in your sample reported being atheists?" with a statistic; while the question "What proportion of people on earth would report being atheists" 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.

5. Write out the conditions for inference to construct a 95% confidence interval for the proportion of atheists in the United States in 2012. 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 `prop.test` function to do it for us. This function takes as its argument a `table` of the variable we are interested in. First, have a look at what the `table` command creates, then conduct the test:

```
table(us12$response)
```

```
##
##     atheist non-atheist
##          50         952
```

```
prop.test(table(us12$response))
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  table(us12$response), null probability 0.5
## X-squared = 810.1806, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.03761982 0.06574456
## sample estimates:
##         p
## 0.0499002
```

Note that since the goal is to construct an interval estimate for a proportion, it's necessary to think of one of the two outcomes as a "success", which here is a response of `"atheist"`.

Although formal confidence intervals and hypothesis tests don't show up in the report, suggestions of inference appear at the bottom of page 7: "In general, the error margin for surveys of this kind is $\pm$ 3-5% at 95% confidence".

6. Based on the R output, what is the margin of error for the estimate of the proportion of atheists in US in 2012?

7. Using the `prop.test` function, calculate confidence intervals for the proportion of atheists in 2012 in two other countries of your choice, and report the associated margins of error. Be sure to note whether the conditions for inference are met. It may be helpful to create new data sets for each of the two countries first, and then use these data sets in the `prop.test` function to construct the confidence intervals.

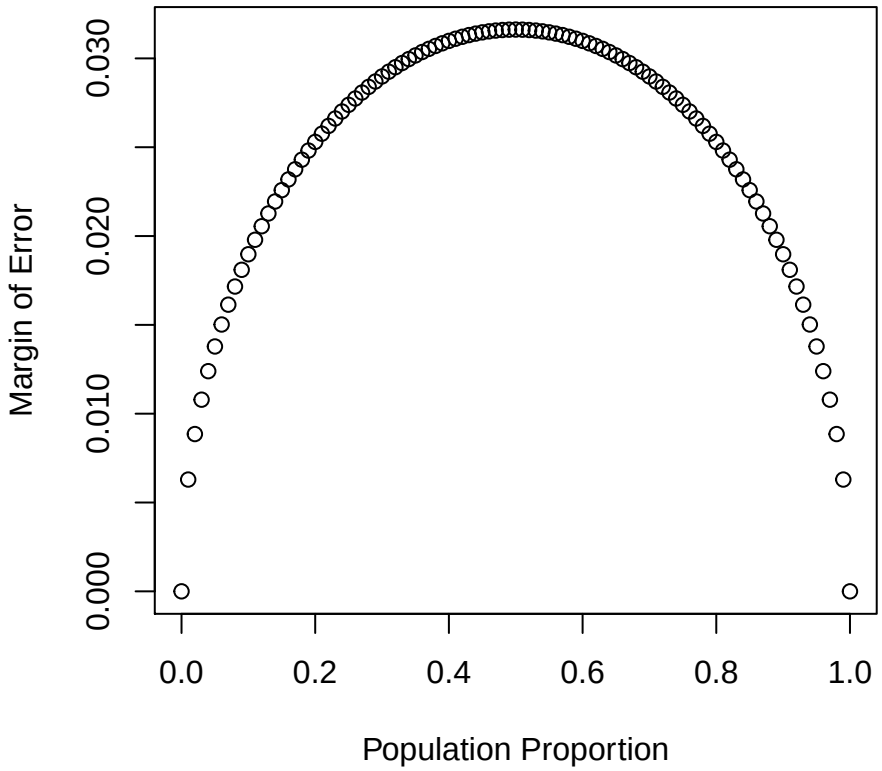
How does the proportion affect the margin of error?

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(x = p, y = me, ylab = "Margin of Error", xlab = "Population Proportion")
```



8. Describe the relationship between `p` and `me`.

Success-failure condition

The textbook emphasizes that you must 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?

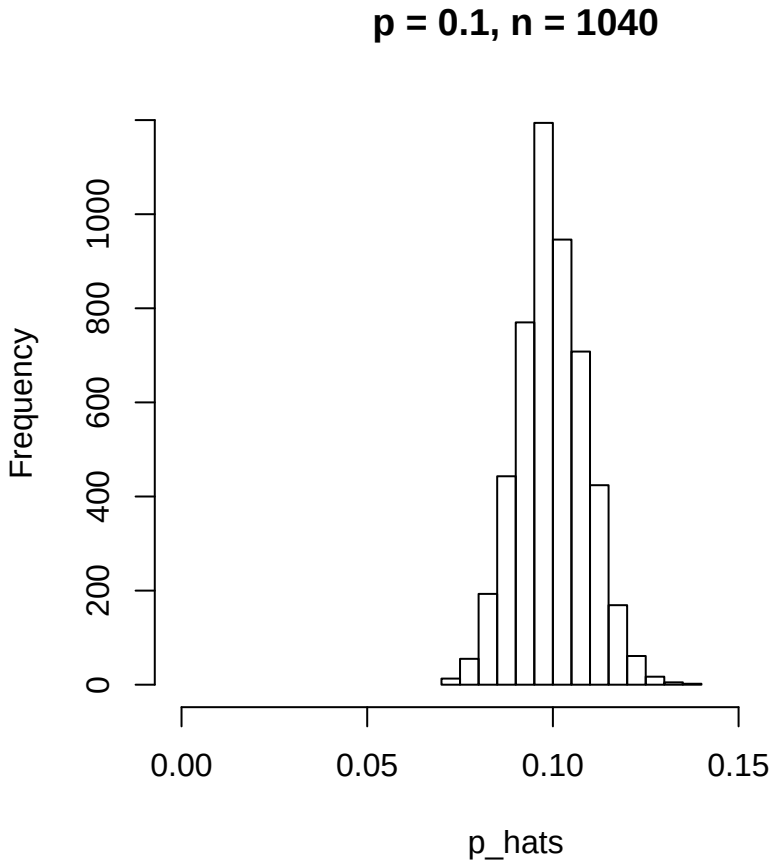
The short answer is: nothing. You could argue that we would be fine with 9 or that we really should be using 11. What is the "best" value for such a rule of thumb is, at least to some degree, arbitrary. However, when np and $n(1-p)$ reaches 10 the sampling distribution is sufficiently normal to use confidence intervals and hypothesis tests that are based on that approximation.

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 population with a true atheist 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("atheist", "non_atheist"), n, replace = TRUE, prob = c(p, 1-p))
  p_hats[i] <- sum(samp == "atheist")/n
}

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



These commands build up the sampling distribution of $\hat{p}$ 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 atheist and non-atheist with probabilities p and $1 - p$, respectively." The second line in the loop says, "calculate the proportion of atheists 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.

9. 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.

10. 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))` command.

11. If you refer to Table 6, you'll find that Australia has a sample proportion of 0.1 on a sample size of 1040, and that Ecuador has a sample proportion of 0.02 on 400 subjects. Let's suppose for this exercise that these point estimates are actually the truth. Then given the shape of their respective sampling distributions, do you think it is sensible to proceed with inference and report margin of errors, as the reports does?

On your own

The question of atheism was asked by WIN-Gallup International in a similar survey that was conducted in 2005. (We assume here that sample sizes have remained the same.) Table 4 on page 13 of the report summarizes survey results from 2005 and 2012 for 39 countries.

- Answer the following two questions using the appropriate function in R. 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 Spain has seen a change in its atheism index between 2005 and 2012?
  Hint: Create a new data set for respondents from Spain. Form confidence intervals for the true proportion of athiests in both years, and determine whether they overlap.

  b. Is there convincing evidence that the United States has seen a change in its atheism index between 2005 and 2012?

- If in fact there has been no change in the atheism index in the countries listed in Table 4, in how many of those countries would you expect to detect a change (at a significance level of 0.05) simply by chance?
  Hint: Look in the textbook index under Type 1 error.

- Suppose you're hired by the local government to estimate the proportion of residents that attend a religious service on a weekly basis. 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.