Online Section 10: Chi-Square Tests and Difference of Means

Author: Ethan A. Fosse

Course: Stat 100/E-100

Date: Week of November 10th, 2014 IMPORTANT: Although this document contains text explaining various concepts and code, it is not a standalone document. You should use this document mainly as a supplement to this week's online section rather than as a substitute.

In this week's online section we will cover several topics: (1) Running a one-way Chi-square test for goodness of fit in R; (2) conducting a two-way Chi-square test for independence in R; (3) analyzing a difference two means (common in experimental designs) in R; (4) bootstrapping confidence intervals in R; understanding prediction and inferences with linear regression; (5) .

Before we begin, as always we load our R packages for this session.

```
## Warning: package 'mosaic' was built under R version 3.1.1
## Warning: package 'mosaicData' was built under R version 3.1.1
## Warning: package 'Lock5Data' was built under R version 3.1.1
```

```r
library(MASS) # load the MASS package
```

Topic 1: How do I concduct a Chi-Square Test (One-Way) for independence?

```
##
## Attaching package: 'MASS'
##
## The following object is masked from 'package:dplyr':
##
##     select
```

```r
levels(survey$Smoke)
```

```
## [1] "Heavy" "Never" "Occas" "Regul"
```

```r
#  creating a table
smoke.freq <- table(survey$Smoke)
print(smoke.freq)
```

```
##
## Heavy Never Occas Regul
##    11   189    19    17
```

```r
# suppose the campus smoking statistics is as below. Determine whether the sample data in survey suppor

smoke.prob <- c(.045, .795, .085, .075)
chisq.test(smoke.freq, p=smoke.prob)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  smoke.freq
## X-squared = 0.107, df = 3, p-value = 0.9909
```

As the p-value 0.991 is greater than the .05 significance level, we do not reject the null hypothesis that the sample data in survey supports the campus-wide smoking statistics.

Topic 2: How do I conduct a Chi-Square Test for Independence (Two-Way)? Two random variables x and y are called independent if the probability distribution of one variable is not affected by the presence of another.

The Smoke column records the students smoking habit, while the Exer column records their exercise level. The allowed values in Smoke are “Heavy”, “Regul” (regularly), “Occas” (occasionally) and “Never”. As for Exer, they are “Freq” (frequently), “Some” and “None.”"

We can tally the students smoking habit against the exercise level with the table function in R. The result is called the contingency table of the two variables. We want to test the hypothesis whether the students smoking habit is independent of their exercise level at .05 significance level.

```r
library(MASS) # load the MASS package
tbl <- table(survey$Smoke, survey$Exer)
tbl # the contingency table
```

```
##
##         Freq None Some
##   Heavy    7    1    3
##   Never   87   18   84
##   Occas   12    3    4
##   Regul    9    1    7
```

```r
chisq.test(tbl) # note: warning message is just telling us that we have a small sample size
```

```
## Warning: Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  tbl
## X-squared = 5.49, df = 6, p-value = 0.4828
```

As the p-value 0.4828 is greater than the .05 significance level, we do not reject the null hypothesis that the smoking habit is independent of the exercise level of the students.

Topic 3: How do I compare means between two matched samples? Two data samples are matched if they come from repeated observations of the same subject. Here, we assume that the data populations follow the normal distribution. Using the paired t-test, we can obtain an interval estimate of the difference of the population means.

In the built-in data set named immer, the barley yield in years 1931 and 1932 of the same field are recorded. The yield data are presented in the data frame columns Y1 and Y2.

Assuming that the data in immer follows the normal distribution, find the 95% confidence interval estimate of the difference between the mean barley yields between years 1931 and 1932.

```r
library(MASS) # load the MASS package
head(immer)
```

```
##   Loc Var    Y1    Y2
## 1  UF   M  81.0  80.7
## 2  UF   S 105.4  82.3
## 3  UF   V 119.7  80.4
## 4  UF   T 109.7  87.2
## 5  UF   P  98.3  84.2
## 6   W   M 146.6 100.4
```

```r
# We apply the t.test function to compute the difference in means of the matched # samples. As it is a
t.test(immer$Y1, immer$Y2, paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  x and immer$Y2
## t = 3.32, df = 29, p-value = 0.002413
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   6.12 25.70
## sample estimates:
## mean of the differences
##                     15.9
```

Between years 1931 and 1932 in the data set immer, the 95% confidence interval of the difference in means of the barley yields is the interval between 6.122 and 25.705.

Topic 4: How do I bootstrap confidence intervals in R? Note: This is from Question 8 from homework #3

Do this in R. It involves the bootstrap. If you can do this problem, you are on your way to generating confidence intervals for anything you want. Save you R code for yourself; it will be a good reference. See the sample lecture code from Lecture 12 for a starting point.

In addition to the commute time (in minutes), the `CommuteAtlanta` dataset gives the distance for the commutes (in miles) for 500 workers sampled from the Atlanta metropolitan area.

a) *Find the mean and standard deviation of the commute distances in CommuteAtlanta.*

Answer: For the original sample the mean commute distance is 18.16 miles and the standard deviation is 13.8 miles.

```r
# loading the dataset
data(CommuteAtlanta)

# examining the dataset
head(CommuteAtlanta)
```

```
##      City Age Distance Time Sex
## 1 Atlanta  19       10   15   M
## 2 Atlanta  55       45   60   M
## 3 Atlanta  48       12   45   M
## 4 Atlanta  45        4   10   F
## 5 Atlanta  48       15   30   F
## 6 Atlanta  43       33   60   M
```

```r
# mean
mean(CommuteAtlanta$Distance)
```

```
## [1] 18.2
```

```r
# standard deviation
sd(CommuteAtlanta$Distance)
```

```
## [1] 13.8
```

b) *Use R to create a bootstrap distribution of the sample means of the distances. Describe the shape and center of the distribution.*

Answer: One bootstrap distribution of distance means is shown below. It is bell-shaped, centered around 18.2, and shows sample means ranging between about 16.5 and 20.5 miles.

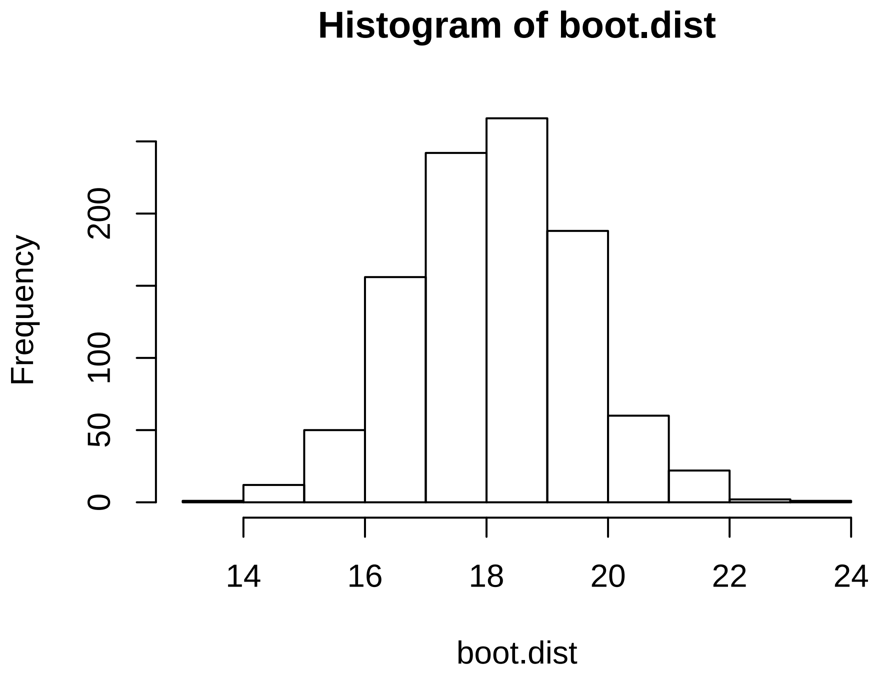
```r
# bootstrap distribution of sample means
dist <- CommuteAtlanta$Distance
boot.dist <- replicate(1000, {
  samp <- sample(dist, size=100, replace=TRUE)
  mean(samp)
  })

# to obtain the mean of the samples
mean(boot.dist)
```

```
## [1] 18.1
```

```r
# to get the standard error for the mean commute time
sd(boot.dist)
```

```
## [1] 1.41
```

```r
# creating a histogram
hist(boot.dist)
```



c) *Use the bootstrap distribution to estimate the standard error for mean commute distance when using samples of size 500.*

Answer: The standard error of the means for this set of 1000 bootstrap samples is approximately 0.61 miles.

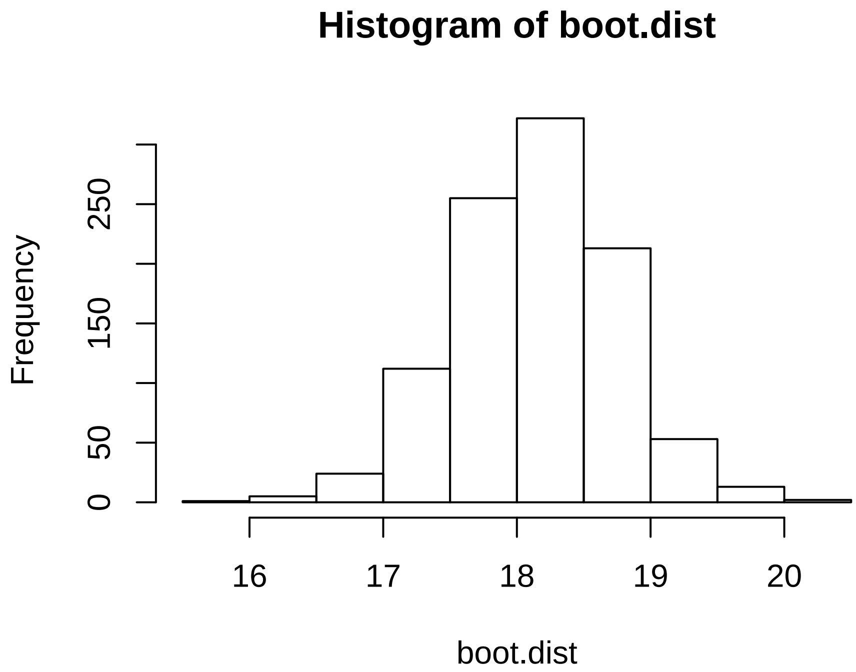
```
# bootstrap distribution of sample means
dist <- CommuteAtlanta$Distance
boot.dist <- replicate(1000, {
  samp <- sample(dist, size=500, replace=TRUE)
  mean(samp)
  })

# to obtain the mean of the samples
mean(boot.dist)
```

```
## [1] 18.1
```

```
# to get the standard error for the mean commute time
sd(boot.dist)
```

```
## [1] 0.609
```

```
# creating a histogram
hist(boot.dist)
```



d) *Use the standard error to find and interpret a 95% confidence interval for the mean commute distance of Atlanta workers.*

Answer: A 95% confidence interval is approximately given by the values of 16.94 to 19.38.

```
# our point estimate
pe <- mean(CommuteAtlanta$Distance)
print(pe)
```

```
## [1] 18.2
```

```
# calculating the standard error of the means
SE <- sd(boot.dist)
print(SE)
```

```
## [1] 0.609
```

```
# calculating our margin of error
moe <- 2*SE

# creating our confidence interval
print(pe-moe) # lower bound
```

```
## [1] 16.9
```

```
print(pe+moe) # upper bound
```

```
## [1] 19.4
```

```
# compare with creating a confidence interval using in-built mosaic commands
confint(t.test(~Distance, data = CommuteAtlanta))
```

```
## mean of x     lower     upper     level
##     18.16     16.94     19.37      0.95
```

Topic 5: How can I use bootstrapping to make inferences with linear regression? Note: This is from Question 9 from homework #3

Use R for this problem. The data on Table 1 shows the percent of children immunized with the DPT vaccine (for diphtheria, poliomyelitis, andtetanus) and the under age 5 mortality rate (number of children who do not survive past age 5 per 1,000 live births). The data is also on the web site in csv format, in the file `child_mortality.csv`. See prior homework or the R Cookbook for instructions on how to load a csv file.

The variable giving the percent immunized is `Percent.Imm`; under age 5 mortality is the variable `Mortality.Rate`.

Nation	% Immunized	Deaths per 1000 Live Births
Bolivia	40	165
Brazil	54	85
Canada	85	9
China	95	43
Egypt	81	94
Ethiopia	26	226
Finland	90	7
France	95	9
Greece	83	12
India	83	145
Italy	85	11
Japan	83	6
Mexico	65	51
Poland	98	18
Senegal	47	189
Turkey	74	90
United Kingdom	75	10
United States	97	12
USSR	79	35
Yugoslavia	91	27

Table 1: Immunization Data for "Immunization Effectiveness" problem

a) *Make a scatterplot of the relationship between these two variables and add the least-squares regression line to the scatterplot, using Percent.Imm as the predictor (horizontal axis) and `Mortality.Rate` as the response variable (vertical axis).*

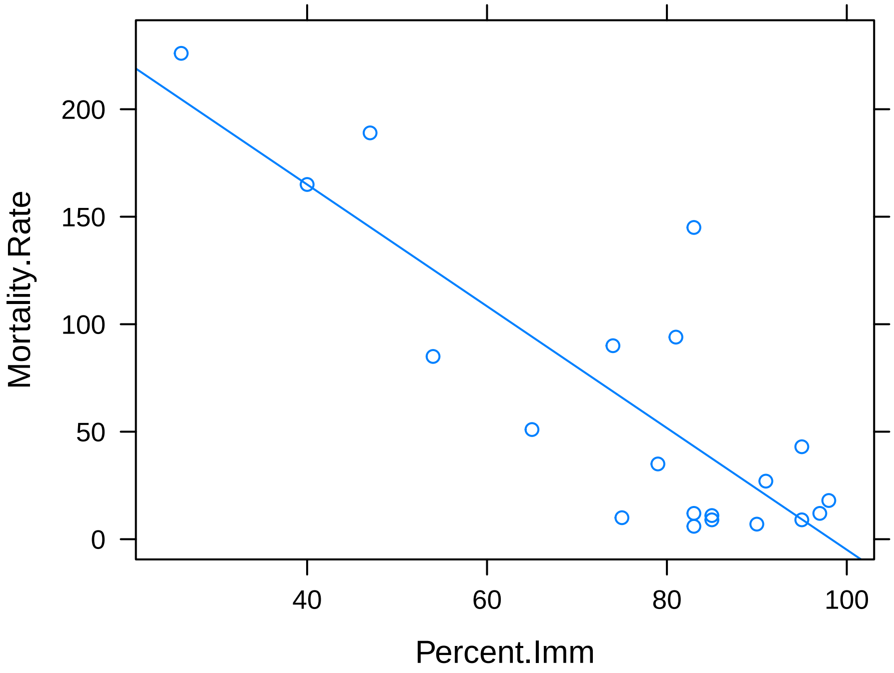
Answer: First we need to load the .csv file into R.

```
# package for loading a .csv file
library(foreign)
child.mort <- read.csv("C:/child_mortality.csv")
head(child.mort)
```

```
##     Nation Percent.Imm Mortality.Rate
## 1  Bolivia          40            165
## 2   Brazil          54             85
## 3   Canada          85              9
## 4    China          95             43
## 5    Egypt          81             94
## 6 Ethiopia          26            226
```

The next code chunk shows a scatter plot of data points overlaid with a least-squares regresison line.

```
# scatterplot with least-squares regression line
# note: "p" refers to plotting points while "r" refers to plotting a simple
# regression line
xyplot(Mortality.Rate ~ Percent.Imm, data=child.mort, type=c("p","r"))
```



b) *Describe the relationship between your two variables and assess the legitimacy of a linear model here.*

Answer: The relationship is approximately linear, although we may want to run regression diagnostics to check for the presence of influential data points.

c) *What is the interpretation of the estimates of the intercept and slope in the specific context of this problem?*

Answer: We can interpret the model as follows. The slope tells us that for every one percentage point increase in child immunization, the mortality rate for children under 5 decreases by 2.83. The intercept gives us the mortality rate of children under 5 when the immunization rate is zero.

```r
# running the least-squares regression model
m <- lm(Mortality.Rate ~ Percent.Imm, data=child.mort)
summary(m)
```

```
##
## Call:
## lm(formula = Mortality.Rate ~ Percent.Imm, data = child.mort)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -55.88 -29.23  -0.12  21.31 101.77
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   278.26      35.46    7.85  3.2e-07 ***
## Percent.Imm    -2.83       0.45   -6.29  6.2e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 39.4 on 18 degrees of freedom
## Multiple R-squared:  0.687,  Adjusted R-squared:  0.67
## F-statistic: 39.6 on 1 and 18 DF,  p-value: 6.25e-06
```

When interpreting the intercept, it is often useful to look at the range of our inputs into the regression model. This is because the estimated intercept can lie far away from the central mass of data points.

```r
range(child.mort$Percent.Imm)
```

```
## [1] 26 98
```

The immunization rate ranges from 26 to 98 percent. No country in the sample has an immunization rate of zero, so we must be careful when interpreting the value of the intercept.

d) *Generate a 95% confidence interval for the slope using the bootstrap technique and the SE method.*

Answer: The following R code should take a single bootstrap sample and compute the slope. Note that you will have to uncomment the R code by removing the # symbol for each line.

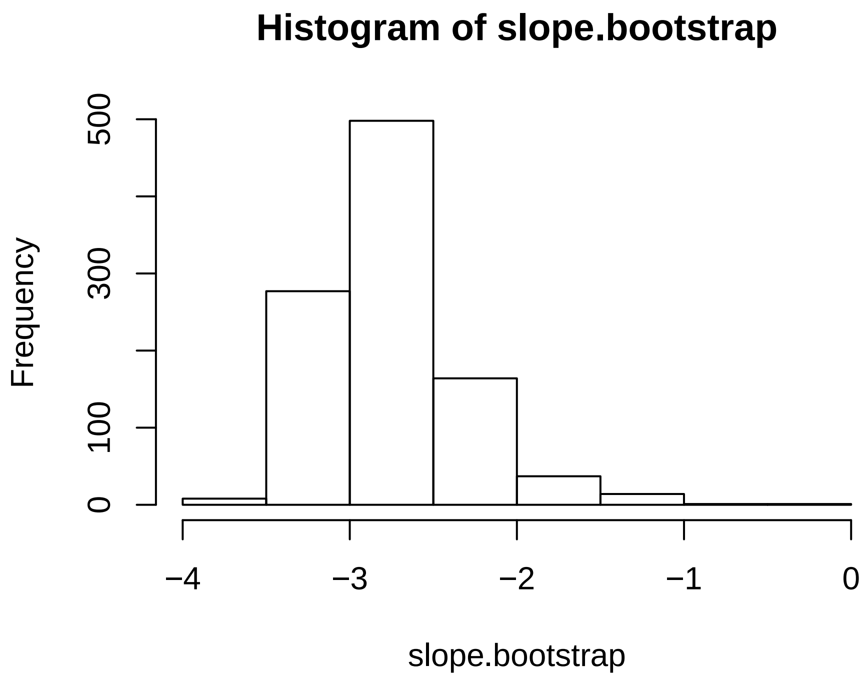
```r
# creating the distribution
slope.bootstrap <- replicate(1000, {
  boot.samp <- resample(child.mort, replace=TRUE)
  # or, alternatively: boot.samp <- sample(child.mort, size=20, replace=TRUE)
  boot.lm <- lm(Mortality.Rate ~ Percent.Imm, data=boot.samp)
  coef(boot.lm)[2]
  })

# our point estimate for the slope
pe <- mean(slope.bootstrap)
# our standard error
SE <- sd(slope.bootstrap)

# 95% confidence interval using SE method:
c(pe-2*SE, pe+2*SE)
```

```
## [1] -3.59 -1.94
```

e) *Check that the SE method is valid by examining the bootstrap sampling distribution. What are the three things you need to check? Is the SE method valid here?*

Answer: As a general rule of thumb, the SE method as outlined in WileyPlus is appropriate when: (1) the bootstrap distribution is symmetric; (2) the bootstrap distribution is bell-shaped; and (3) the distribution is continuous.

```r
# examining the bootstrap distribution
hist(slope.bootstrap)
```



```
# the SE method of computing the confidence interval:
c(pe-2*SE, pe+2*SE)
```

```
## [1] -3.59 -1.94
```

```
# the 95% confidence interval using percentile method:
quantile(slope.bootstrap, c(0.025, 0.975))
```

```
##  2.5% 97.5%
## -3.39 -1.70
```

The results give approximately the same results, although the percentile method is slightly narrow than the SE method. This difference is likely a result of the fact that the samples we are drawing are of size 20, which is fairly small.

f) *Interpret your 95% CI in the context of the problem.*

Answer: For the SE method, we are 95% confident that the true slope of immunization rate on child mortality is between approximately -3.57 and -1.94 (for the percentile method). For the percentile method, we are 95% confident that the true slope of immunization rate on child mortality is between approximately -3.38 and -1.65.

g) *Using your model, predict the change in the under-5 mortality rate for Bolivia if Bolivia were to increase the percentage of children immunized from 40% to 75%. Be sure to specify whether the change would be an increase or decrease.*

Answer: Based on our regression model, we predict that the under-5 mortality rate for Bolivia would decrease from about 164.98 to 65.86 (that is, it would decrease by 99.12).

```
# our regression model again
print(m)
```

```
##
## Call:
## lm(formula = Mortality.Rate ~ Percent.Imm, data = child.mort)
##
## Coefficients:
## (Intercept)  Percent.Imm
##      278.26        -2.83
```

```
# predicting a change in immunization
yhat1 <- 278.260 + (-2.832)*(40)
yhat2 <- 278.260 + (-2.832)*(75)
print(yhat1)
```

```
## [1] 165
```

```r
print(yhat2)
```

```
## [1] 65.9
```

```r
# predicted change in mortality rate
print(yhat1 - yhat2)
```

```
## [1] 99.1
```

h) *If Bolivia did implement such a program and achieved 75% immunization, how do you feel your prediction would actually do? Why?*

Answers will vary. There are reasons to be wary of such a prediction because immunization is related to many other factors that are related to lower levels of child mortality.

i) *The predicted change uses the slope of our regression line. We can thus generate a 95% confidence interval for the predicted change by using the confidence interval of the slope. Do this by using the using the lowest conceivable and highest conceivable slopes from you CI above to generate two different predicted changes. What is your final CI for predicted change?*

Answer: Responses will vary. Based on this approach, we estimate that the 95% confidence interval for predicted change is approximately 68 to 125.

```r
# 95% confidence interval using SE method:
c(pe-2*SE, pe+2*SE)
```

```
## [1] -3.59 -1.94
```

```r
# predicting a change in immunization for lower bound of CI
yhat1 <- 278.260 + (pe-2*SE)*(40)
yhat2 <- 278.260 + (pe-2*SE)*(75)
print(yhat1 - yhat2)
```

```
## [1] 126
```

```r
# predicting a change in immunization for upper bound of CI
yhat1 <- 278.260 + (pe+2*SE)*(40)
yhat2 <- 278.260 + (pe+2*SE)*(75)
print(yhat1 - yhat2)
```

```
## [1] 67.8
```

j) *Explain in your own words why generating a CI this way is valid.*

Answer: Responses will vary. This approach is valid because we are using the bootstrap distribution to calculate the lower and upper bounds of our predictions.

k) *The regression line shows an association. This is not necessarily causal, so your prediction, above, is not necessarily what would actually happen. Now we have a confidence interval. Does this fix the problem? Explain why or why not.*

Answer: No this dos not fix the problem. The confidence interval is just an interval estimate rather than a point estimate of the association.

l) *Now say we are trying to predict what the mortality rate would be, on average, for countries with 90% immunization. What is our best guess?*

Answer: Our estimated mortality rate is about 23.38.

```
yhat <- 278.260 + (-2.832)*(90)
print(yhat)
```

```
## [1] 23.4
```

m) *We need a SE for our prediction. Obtain one by modifying your bootstrapping code for the slope so that it generates a prediction for 90% immunization each iteration and then takes the standard deviation of those values.*

Answer: To have R predict the mortality rate for 90% immunization, try the following R code (making sure to uncomment the code by removing the # symbols for each line):

```
# creating the distribution
slope.bootstrap <- replicate(1000, {
  boot.samp <- resample(child.mort, replace=TRUE)
  # or, alternatively: boot.samp <- sample(child.mort, size=20, replace=TRUE)
  mlm <- lm(Mortality.Rate ~ Percent.Imm, data=boot.samp)
  coef(mlm)[1]  +  90*coef(mlm)[2]
  })

# our point estimate for our prediction
pe <- mean(slope.bootstrap)
# our standard error
SE <- sd(slope.bootstrap)
```

n) *Use a normal approximation to generate an 80% confidence interval for the 90% prediction.*

Answer: Responses will vary, but in general we expect the range to be narrower than for a 95% confidence interval. The 80% confidence interval for the prediction using the SE method is about 11.82 to 35.59, while for the percentile method it is approximately 12.36 to 35.92.

```
# 80% confidence interval using SE method:
c(pe-1.28*SE, pe+1.28*SE)
```

```
## [1] 10.7 35.4
```

```
# 80% confidence interval using percentile method:
quantile(slope.bootstrap, c(0.10, 0.90))
```

```
##  10%  90%
## 10.9 35.8
```

o) *Say we wanted to do generate a confidence interval as above for 20% immunization. Would this be a good idea or not? Why or why not?*

Answer: When generating a prediction from a regression model, it is important to examine the range of our inputs into the regression model.

```
# examining summary statistics of our dataset
summary(child.mort)
```

```
##       Nation    Percent.Imm   Mortality.Rate
##  Bolivia : 1   Min.   :26.0   Min.   :  6.0
##  Brazil  : 1   1st Qu.:71.8   1st Qu.: 10.8
##  Canada  : 1   Median :83.0   Median : 31.0
##  China   : 1   Mean   :76.3   Mean   : 62.2
##  Egypt   : 1   3rd Qu.:90.2   3rd Qu.: 91.0
##  Ethiopia: 1   Max.   :98.0   Max.   :226.0
##  (Other) :14
```

We can see that the immunization rate ranges from 26% to 98%. Although we can construct a prediction (and accordingly a confidence interval) for a country wiht 20% immunization, we would be extrapolating beyond the range of the data.

That’s it for this week!