Homework 6 Solution

Math 445, Spring 2017

Problem 1 Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with unknown mean μ and unknown variance σ^2.

(a) What distribution does $\dfrac{\sum_{i=1}^n (X_i - \overline{X})^2}{\sigma^2}$ have?

We know from theorem B.16 that $\dfrac{\sum_{i=1}^n (X_i - \overline{X})^2}{\sigma^2} = \dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$

(b) Why is $\dfrac{\sum_{i=1}^n (X_i - \overline{X})^2}{\sigma^2}$ a pivotal quantity?

$\dfrac{\sum_{i=1}^n (X_i - \overline{X})^2}{\sigma^2}$ is a pivotal quantity because it is a function of the data and parameter of interest whose distribution does not depend on the parameters.

(c) Describe a method for constructing a confidence interval for σ^2 with a specified confidence level $1 - \alpha$. Hint: Determine constants c_1 and c_2 such that

$$P\left(c_1 < \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{\sigma^2} < c_2\right) = 1 - \alpha$$

To find a $(1-\alpha) \cdot 100\%$ confidence interval for σ^2, first we find the endpoints finding c_1 and c_2 by considering two probability statements: $P\left(\dfrac{\sum_{i=1}^n (X_i - \overline{X})^2}{\sigma^2} / \theta \leq c_1\right) = \alpha/2$ and $P\left(\dfrac{\sum_{i=1}^n (X_i - \overline{X})^2}{\sigma^2} / \theta \leq c_2\right) = 1-\alpha/2$. That is, c_1 and c_2 are the $\alpha/2$ and $1-\alpha/2$ quantiles of the χ^2_{n-1} distribution, respectively.

Next, we rearrange the inequality to isolate σ^2:

$$P\left(\frac{\sum_{i=1}^n (X_i - \overline{X})^2}{c_2} < \sigma^2 < \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{c_1}\right) = 1 - \alpha$$

Problem 2 Let $Y_1, Y_2, \ldots, Y_n$ denote a random sample of size n from a population with uniform distribution on the interval $(0, \theta)$. Let $Y_{max} = \max(Y_1, Y_2, \ldots, Y_n)$ and $U = Y_{max}/\theta$.

(a) Show that U has CDF

$$F(u) = \begin{cases} 0 & u < 0 \\ u^n & 0 \leq u \leq 1 \\ 1 & u > 1 \end{cases}$$

We can establish this using the CDF method: $P(U \leq u) = P(Y_{max} \leq \theta u)$. Thus, the CDF of U is the CDF of Y_{max} evaluated at θu, which is easily found from probability theory:

$$P(Y_{max} \leq \theta u) = P(Y_1 \leq \theta u) P(Y_2 \leq \theta u) \cdots P(Y_n \leq \theta u) = \left(\frac{\theta u}{\theta}\right)^n = u^n, \ 0 < u < 1$$

(b) Why is U a pivotal quantity?

U is a pivotal quantity because it is a function of the data and parameter of interest whose distribution does not depend on the parameter θ.

(c) Find a 95% confidence interval for θ.

To find a 95% confidence interval for θ we find the endpoints finding a and b such that $P(Y_{max}/\theta \leq a) = 0.025$ and $P(Y_{max}/\theta \leq b) = 0.975$. Rearranging these we see that the lower bound will be Y_{max}/b and the upper bound will be Y_{max}/a. To complete the problem, we must find a and b, which are found by inverting the CDF:

$$F(a) = 0.025 \iff a^n = 0.025 \implies a = 0.025^{1/n}$$

Similarly, we find that $b = 0.975^{1/n}$.

Note: a lower confidence bound would make more sense for this problem, so you should think about how you would form this one-sided confidence interval.

Problem 3 An investigator is interested in determining if the average amount of sleep for Lawrentians is less than the recommended amount of at least 8 hours per night. For a random sample of $n = 91$ students in a survey given the first week of class, the average amount of sleep was calculated to be 7.319 hours per night on and the standard deviation in this sample was 1.029.

(a) Find the 95% confidence interval for the average amount of nightly sleep for all Lawrence students.

A 95% CI for μ is given by

$$\overline{x} \pm t_{n-1,.975} \cdot \frac{s}{\sqrt{n}} = 7.319 \pm 1.9867 \cdot \frac{1.029}{\sqrt{91}} = (7.105,\ 7.533)$$

Below is the relevant work in R:

```
qt(0.975, df = 90)

## [1] 1.986675

7.319 - qt(0.975,df=90) * 1.029/sqrt(91)

## [1] 7.1047

7.319 + qt(0.975,df=90) * 1.029/sqrt(91)

## [1] 7.5333
```

(b) Based on your confidence interval, is there sufficient evidence to show that the average amount of sleep is less than the recommended amount?

Since the upper bound does not include 8, we reject it as a plausible value. Further, since the upper bound is less than 8, there is sufficient evidence that $\mu < 8$.

Problem 4 Financial statisticians are interested in the volatility of a market. When volatility is low, prices change very little over time. When volatility is high, prices change a lot. The direction of the change (up or down) is not important. All that matters is the magnitude of the change. The following data are a small part of a larger study on volatility of stock prices.

There are three major stock exchanges in the US. This small study concerns the volatility on two exchanges (NYSE and NASDAQ) during the week of 12-16 Feb 2007. On each exchange, 40 stocks were randomly selected from all stocks listed on the exchange that week. The volatility was calculated for each stock as the absolute value of the proportional change in stock price. A value of 0% indicates no change in stock price. An exchange with an average change of 5% is less volatile than one with an average change of 10%.

Summary statistics for each exchange are:

Exchange	n	average	s.d.
NYSE	40	2.91%	2.23%
NASDAQ	40	4.39%	3.37%

(a) Estimate the pooled standard deviation, s_p.

The pooled standard deviation is

$$s_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}} = \sqrt{\frac{(40-1)(2.23)^2+(40-1)(3.37)^2}{40+40-2}} = 2.85\%$$

(b) How many degrees of freedom do we have?

$df = 40 + 40 - 2 = 78.$

(c) Calculate and interpret a 95% confidence interval for the difference in mean volatility (NASDAQ - NYSE).

$$\overline{x}_1 - \overline{x}_2 \pm t_{78,\,0.975} s_p \sqrt{\frac{1}{n_1}+\frac{1}{n_2}} = 1.48\% \pm 1.99 \cdot 2.85\sqrt{\frac{1}{40}+\frac{1}{40}} = (0.21\%,\ 2.75\%)$$

```
qt(0.975, df = 78)

## [1] 1.990847
```

We are 95% confident that the difference in mean volatility between the NYSE and NASDAQ is between 0.21% and 2.75%.

(d) Assume that the sample size was 20 in each group, but the observed mean and s.d. are unchanged. Calculate a 95% confidence interval for the difference in mean volatility (NASDAQ - NYSE).

$$\overline{x}_1 - \overline{x}_2 \pm t_{38,\,0.975} s_p \sqrt{\frac{1}{n_1}+\frac{1}{n_2}} = 1.48\% \pm 2.02 \cdot 2.85\sqrt{\frac{1}{20}+\frac{1}{20}} = (-0.35\%,\ 3.31\%)$$

```
qt(0.975, df = 38)

## [1] 2.024394
```

(e) Assume that the sample size was 100 in each group, but the observed mean and s.d. are unchanged. Calculate a 95% confidence interval for the difference in mean volatility (NASDAQ - NYSE).

$$\overline{x}_1 - \overline{x}_2 \pm t_{198,\,0.975}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 1.48\% \pm 1.97 \cdot 2.85 \sqrt{\frac{1}{100} + \frac{1}{100}} = (0.69\%,\ 2.27\%)$$

```
qt(0.975, df = 198)

## [1] 1.972017
```

(f) How does your interval change as the sample size increases? Explain (briefly) why this is reasonable.

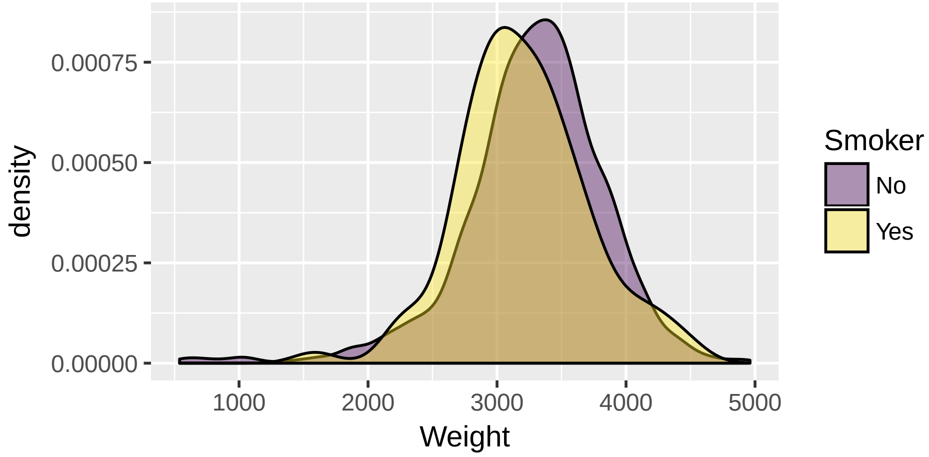
The confidence interval becomes wider as the sample size decreases. Because the variability decreases as the sample size increases, we expect the confidence interval to be more precise and narrower with large sample.

Problem 5 The data set `TXBirths` in the `resampledata` package contains data on babies born in Texas in 2004 (see Case Study 1.2 in your textbook for more information).

(a) Create plots to investigate the distribution of weights of babies in each group. Comment on the shape of the distributions.

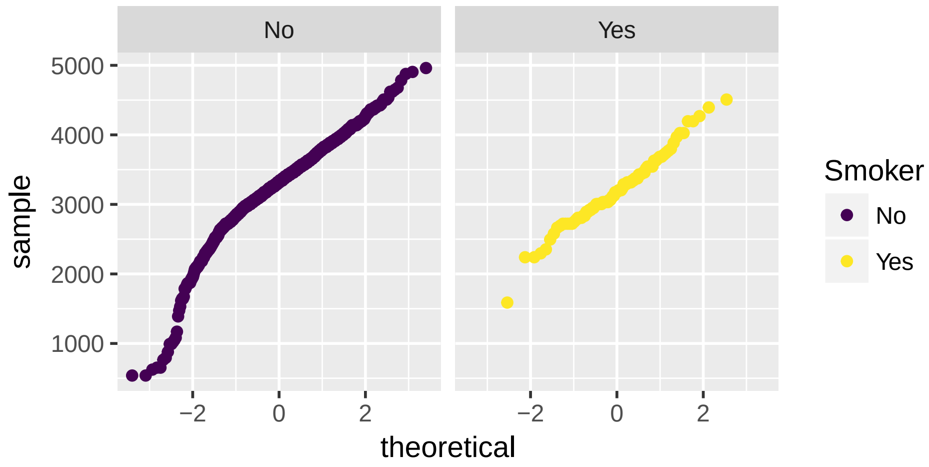
```
library(resampledata)
library(ggplot2)
library(viridis)

ggplot(TXBirths2004, aes(x = Weight, fill = Smoker)) +
  geom_density(alpha = 0.4) +
  scale_fill_viridis(discrete = TRUE)
```



Overlayed density plots reveal that babies whose mothers smoked weigh slightly less (the center of the distribution is shifted slightly left). The shapes of the two distributions are similar (unimodal, left skewed).

```
ggplot(TXBirths2004, aes(sample = Weight, color = Smoker)) +
  stat_qq() +
  facet_wrap(~Smoker) +
  scale_color_viridis(discrete = TRUE)
```



The normal Q-Q plots reveal that the distribution of weights for babies whose mothers are smokers are more nearly symmetric than babies whose mothers did not smoke. This is seen by the heavier lower tail.

(b) Calculate and interpret a 90% confidence interval for the difference in mean weight between the groups (No - Yes).

To calculate a 90% confidence interval, we can use the `t.test()` function:

```
smoker <- subset(TXBirths2004, select = Weight, subset = Smoker == "Yes", drop = TRUE)
nonsmoker <- subset(TXBirths2004, select = Weight, subset = Smoker == "No", drop = TRUE)
t.test(x = nonsmoker, y = smoker, conf.level = 0.9)
```

```
##
##  Welch Two Sample t-test
##
## data:  nonsmoker and smoker
## t = 1.4806, df = 102.38, p-value = 0.1418
## alternative hypothesis: true difference in means is not equal to 0
## 90 percent confidence interval:
##   -9.87141 172.88094
## sample estimates:
## mean of x mean of y
##  3287.494  3205.989
```

We are 90% confident that babies born to smokers weigh between 9.8 grams less and 172.9 grams more than babies born to nonsmokers, on average.

(c) Calculate a 90% bootstrap percentile interval for the difference in mean weight between the groups (No - Yes).

```
B <- 10e4
meandiffs <- numeric(B)
for(i in 1:B) {
  x <- sample(nonsmoker, replace = TRUE)
  y <- sample(smoker, replace = TRUE)
  meandiffs[i] <- mean(x) - mean(y)
}

quantile(meandiffs, probs = c(0.05, 0.95))
```

```
##         5%        95%
##  -8.621365 170.776919
```

(d) How do the confidence intervals from parts (b) and (c) compare?

The 90% bootstrap confidence interval is slightly narrower than the 90% t-based confidence interval, but they are very similar.