Homework 7 Key

Interval Estimation

Instructions

Please answer the following questions in the order in which they are posed. Add a few empty lines below each and write your answers there. **Focus on answering in complete sentences and show work whether we ask for it or not**. You will also need scratch paper/pen to work out the answers before typing it.

For help with formatting documents in RMarkdown, please consult R Markdown: The Definitive Guide. Another option is to search using Google.

Exercises

1. (Measurement error) Recall the pH-meter from Homework 6 which was known to give readings that were systematically higher or lower by a quantity δ_0. In order to estimate δ_0, six measurements $X_1, X_2, \ldots, X_6$ were made from a solution with pH **known** to be 4.84. In your previous homework, you were asked to come up with an estimator for δ_0. Let's call it $\hat{\delta}_0^{mom}$.

   Now, suppose four measurements - Y_1, Y_2, Y_3, Y_4 - are made from a solution with an unknown pH-level μ_0 resulting in 4.33, 4.22, 4.23, 4.37. As in the previous homework, the measurement error model is that Y_1, Y_2, Y_3, Y_4 is drawn independently from a distribution with mean $\mu_0 + \delta_0$ and variance σ_0^2.

   Consider the estimator

$$\hat{\mu}_0 = \bar{Y} - \hat{\delta}_0^{mom}$$

   for μ_0.

a. Show that $\hat{\mu}_0$ is an unbiased estimator of μ_0.

   Here we want to show that $E\left[\hat{\mu}_0\right] = \mu_0$. In homework 6, we proved that $\hat{\delta}_0^{mom}$ is an unbiased estimator of δ_0. So

$$\begin{aligned}
E\left[\hat{\mu}_0\right] &= E\left[\bar{Y} - \hat{\delta}_0^{mom}\right] \\
&= E\left[\bar{Y}\right] - E\left[\hat{\delta}_0^{mom}\right] \quad \text{linearity of expectation} && (1) \\
&= \mu_0 + \delta_0 - \delta_0 && (2) \\
&= \mu_0
\end{aligned}$$

   where 2 follows from 1 using Theorem 18.1 which states that the sample mean $\bar{Y}$ is an unbiased estimator of the population mean $\mu_0 + \delta_0$.

b. Give an expression for the standard error of $\hat{\mu}_0$. That is, find $\sqrt{Var\left(\hat{\mu}_0\right)}$. Show your work. (State any assumptions you need to make)

The variance of $\hat{\mu}_0$ is calculated below:

$$\begin{aligned}
Var\left[\hat{\mu}_0\right] &= Var\left[\bar{Y} - \hat{\delta}_0^{mom}\right] \\
&= Var\left[\bar{Y}\right] + Var\left[\hat{\delta}_0^{mom}\right], \quad \text{independence of the samples} \\
&= Var\left[\frac{1}{4}\left(Y_1 + Y_2 + Y_3 + Y_4\right)\right] + \frac{\sigma_0^2}{6} \quad \text{from HW 6} \\
&= \frac{1}{16}\left(Var(Y_1) + Var(Y_2) + Var(Y_3) + Var(Y_4)\right) + \frac{\sigma_0^2}{6} \quad \text{independence of } \; Y's \\
&= \frac{4\sigma_0^2}{16} + \frac{\sigma_0^2}{6} \\
&= \frac{5\,\sigma_0^2}{12}.
\end{aligned}$$

Therefore

$$SE\left(\hat{\mu}_0\right) = \sigma_0\,\sqrt{\frac{5}{12}}.$$

c. The variability in the pH measurements - σ_0 - is the same for both the X measurements and also the Y measurements. This makes sense since the variability in the readings is related to the meter, not the specific solution it is being used on.

   A natural estimate for σ_0 is a pooled standard deviation s_p calculated from both samples. The formula for s_p is below:

$$s_p^2 = \frac{\sum_{i=1}^{6}(x_i - \bar{x})^2 + \sum_{j=1}^{4}(y_i - \bar{y})^2}{6 + 4 - 2}$$

   Calculate s_p, the pooled estimate of σ_0.

```r
#six measurements for solution with pH = 4.84 from homework 6
x<- c(4.71, 4.63, 4.69, 4.76, 4.58, 4.83)
#four measurements for solution with unknown pH from this homework
y<- c(4.33, 4.22, 4.23, 4.37)

n1 <- length(x)
n2 <- length(y)

sp <- sqrt( ( (n1-1)*var(x) + (n2-1)*var(y))/(n1+n2-2)  )
cat("Pooled SD", sp)
```

```
## Pooled SD 0.08402009
```

d. Calculate the estimated standard error of $\hat{\mu}_0$. Show your steps.

The estimated standard error of $\hat{\mu}_0$ is

$$\hat{SE}\left(\hat{\mu}_0\right) = s_p\,\sqrt{\frac{5}{12}} = 0.054.$$

2. (Force) A type of metal bar breaks when a force of size X is applied, where X has PDF

$$f(x) = 2\alpha_0 x\, e^{-\alpha_0\, x^2} \qquad x > 0$$

   where $\alpha_0 > 0$ is an unknown parameter. We observe a breaking force of 40. Find a 95% confidence interval for α_0.

Hint: We are looking for a random interval $[L, U]$ which contains α_0 with probability 95%. Construct the interval by "inverting" the probability statement

$$P\left(q_{0.025} \leq X \leq q_{0.975}\right) = 0.95$$

where $q_{0.025}$ and $q_{0.975}$ are the 2.5th and 97.5th percentiles of the distribution of X.

The *pth* percentile of a continuous random variable X is the number q such that

$$F(q) = p$$

where F is the CDF of X. In this case

$$\begin{aligned} F(q) &= \int_0^q f(x)\,dx \\ &= \int_0^q 2\,\alpha_0\,x\,e^{-\alpha_0\,x^2}\,dx \\ &= \int_0^{\alpha_0 q^2} e^{-u} du \qquad u = \alpha_0 x^2 \Rightarrow du = 2\alpha_0 x dx \\ &= \left[-e^{-u}\right]_0^{\alpha_0 q^2} \\ &= 1 - e^{-\alpha_0 q^2}. \end{aligned} \tag{3}$$

The 2.5th percentile - $q_{0.025}$ - is obtained by setting the expression in equation (3) to 0.025 and solving for q. Therefore

$$q_{0.025} = \sqrt{-\frac{1}{\alpha_0}\ln(0.975)}.$$

Similarly the 97.th percentile - $q_{0.975}$ - is

$$q_{0.975} = \sqrt{-\frac{1}{\alpha_0}\ln(0.025)}..$$

Therefore we have the probability statement:

$$P\left(q_{0.025} \leq X \leq q_{0.975}\right) = P\left(\sqrt{-\frac{1}{\alpha_0}\ln(0.975)} \leq X \leq \sqrt{-\frac{1}{\alpha_0}\ln(0.025)}\right) = 0.95$$

Inverting the left hand side of the event gives

$$\sqrt{-\frac{\ln(0.975)}{\alpha_0}} \leq X \Rightarrow \alpha_0 \geq \frac{-\ln(0.975)}{X^2}.$$

Inverting the right hand side of the event gives

$$\sqrt{-\frac{\ln(0.025)}{\alpha_0}} \geq X \Rightarrow \alpha_0 \leq \frac{-\ln(0.025)}{X^2}.$$

Hence

$$P\left(-\frac{\ln(0.975)}{X^2} \geq \alpha_0 \leq -\frac{\ln(0.025)}{X^2}\right) = 0.95$$

and

$$[-\frac{\ln(0.975)}{X^2}, -\frac{\ln(0.025)}{X^2}]$$

is a 95% confidence interval for α_0. When $x = 40$, the interval is $[0,\ 0.002]$.

value	0	1	2	3	4	5
frequency	13	18	23	15	6	8

3. (CLT) A sample of 83 observations for an integer-valued random variable Y is shown below:

   Use the Central Limit Theorem to find a 90% confidence interval for $\pi_0 = P(Y \geq 2)$. Show your work, develop your answer. We are grading on style.

   Hint: You actually have 83 independent Bernoulli random variables - $X_1, X_2, \ldots, X_{83}$ - where each X_i is one if $Y \geq 2$ and zero otherwise. Therefore you can think of $X_1, X_2, \ldots, X_{83} \overset{i.i.d.}{\sim} Binom(1, \pi_0)$ and you wish to construct a confidence interval for the mean of the distribution - π_0 - using the CLT.

We have that for n iid RV, $X_1, ..., X_n$ where $E[X] = \mu$ and $SD(X) = \sigma_0$ then by CLT:

$$\frac{1}{n}\sum_{i=1}^{N} X_i = \bar{X} \sim N(\mu, \sigma_0/\sqrt{n})$$

Based on the hint, if we let $X_1, \ldots, X_{83}$ be Bernoulli RV, with probability, $\pi_0 = P(Y \geq 2)$, then we have that $E[X] = \pi_0$ and $SD(X) = \sqrt{\pi_0(1-\pi_0)}$. Thus by CLT we will have that:

$$\bar{X} \sim N(\pi_0, \frac{\sqrt{\pi_0(1-\pi_0)}}{\sqrt{n}})$$

We know that for a 90% CI we may consider the 95'th and 5'th percentile of a normal distribution. Denote these values $q_{.95}$ and $q_{.05}$. Thus we have by properties of a normal distribution that:

$$P(q_{.05} \leq \frac{\bar{X} - \pi_0}{\frac{\sqrt{\pi_0(1-\pi_0)}}{\sqrt{n}}} \leq q_{.95}) = .9$$

Looking at this formula, we see that the right hand side is equal to 90%, which is what we want for our CI. Inside the probability statement we see we have a π_0 we may "solve" for (get π_0 alone in the middle). Doing some algebra we get:

$$P(q_{.05} \leq \frac{\bar{X} - \pi_0}{\frac{\sqrt{\pi_0(1-\pi_0)}}{\sqrt{n}}} \leq q_{.95}) = P(\bar{X} - q_{.95}\frac{\sqrt{\pi_0(1-\pi_0)}}{\sqrt{n}} \leq \pi_0 \leq \bar{X} - q_{.05}\frac{\sqrt{\pi_0(1-\pi_0)}}{\sqrt{n}}) = .9$$

Noting that $q_{.05} = -q_{.95}$ since normal distributions are symmetric we may simplify this CI as:

$$[\bar{X} - q_{.95}\frac{\sqrt{\pi_0(1-\pi_0)}}{\sqrt{n}}, \bar{X} + q_{.95}\frac{\sqrt{\pi_0(1-\pi_0)}}{\sqrt{n}}]$$

We have in this case, each X_i is equal to 1 if Y was greater than or equal to 2, and 0 otherwise. Thus using our table, we see that 31 of the 83 values are below 2, and 52 of the values are for Y greater than or equal to 2. Thus we have that $\bar{X} = \frac{52}{83} = \hat{\pi}_0$. Using our estimator for π_0, we can also approximate the standard deviation, $\frac{\sqrt{\pi_0(1-\pi_0)}}{\sqrt{n}}$ as $\frac{\sqrt{\hat{\pi}_0(1-\hat{\pi}_0)}}{\sqrt{n}}$. Thus we can approximate this interval as:

$$[\bar{X} - q_{.95}\frac{\sqrt{\hat{\pi}_0(1-\hat{\pi}_0)}}{\sqrt{83}}, \bar{X} + q_{.95}\frac{\sqrt{\hat{\pi}_0(1-\hat{\pi}_0)}}{\sqrt{83}}]$$

Plugging in the values for $\hat{\pi}_0 = \bar{X} = 52/83$, and $q_{.95} = 1.6448536$ (using qnorm(.95)) we have our 90% confidence interval for $P(Y \geq 2)$ as:

$$[52/83 - 1.64 \times \frac{\sqrt{(52/83)(31/83)}}{\sqrt{83}}, 52/83 - 1.64 \times \frac{\sqrt{(52/83)(31/83)}}{\sqrt{83}}] = [0.539, 0.714]$$

4. (Airbnb) Read sections 18.3 and 19.2 in the Notes where I constructed a confidence interval for the mean (daily) price of 2 bedroom apartment rentals in Seattle. In this section you will repeat this calculation for a different subset of rentals: houses with 3 or more bedrooms where the entire home is for rent. The variables you will be filtering on and their values are shown below:
   - property_type: Houses
   - room_type: Entire home/apt
   - bedrooms: 3 or more

a. In this part, you will construct a large sample 95% confidence interval for the mean price of all such house rentals in Seattle. Be sure to

   - display the first five rows of the filtered data frame (showing just price)
   - make a histogram of `price` and
   - calculate and report a large sample 95% confidence interval for the mean daily price. (See section 18.3 from pages 206-208 for example code.)

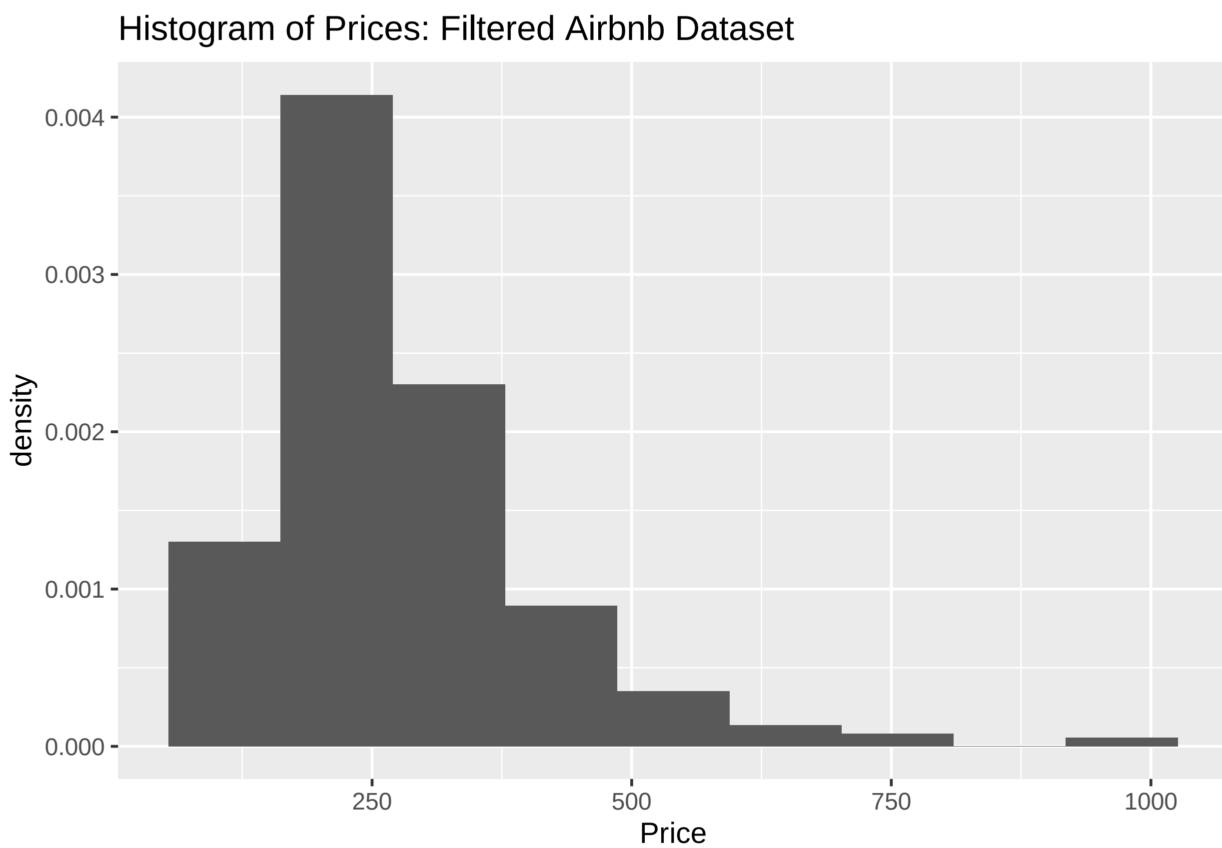
We can see the first five rows of the filtered data-frame are:

```
airbnb <- read_csv("listings.csv")
airbnb_new <- airbnb %>% filter(property_type == "House",
                                room_type == "Entire home/apt",
                                bedrooms >= 3) %>%
   mutate(price = parse_number(price)) %>%
   select(price)
airbnb_new %>% slice_head(n=5)
```

```
## # A tibble: 5 x 1
##   price
##   <dbl>
## 1   975
## 2   450
## 3   461
## 4   700
## 5   450
```

We see our histogram for these new prices looks like:

```
ggplot(data=airbnb_new,
       mapping = aes(x=price,
                     y=..density..))+
   #sturges rule = diff(range(x))/log2(nrow(airbnb_new))
   geom_histogram(binwidth = 108)+
   labs(title = "Histogram of Prices: Filtered Airbnb Dataset",
        x = 'Price',
        y = 'density')
```



Now we may construct a 95% large sample confidence interval for these prices. We get:

```
options(pillar.sigfig = 6)
airbnb_new %>% summarise(xbar = mean(price),
                         s = sd(price),
                         n = n(),
                         se = s/sqrt(n),
                         lower = xbar - qnorm(.975)*se,
                         upper = xbar + qnorm(.975)*se)
```

```
## # A tibble: 1 x 6
##      xbar       s     n      se   lower   upper
##     <dbl>   <dbl> <int>   <dbl>   <dbl>   <dbl>
## 1 277.254 130.966   342 7.08182 263.374 291.135
```

b. In this part, you will construct a (non-parametric) bootstrap confidence interval for the mean price of houses with 3 or more bedrooms where the entire home is for rent. Be sure to

- display the bootstrap sampling distribution of the sample mean
- compare the bootstrap sampling distribution with the normal distribution
- calculate and report the standard bootstrap confidence limits (See section 19.2 on pages 219 - 222 for example code)

We see the bootstrap distribution and qqplot for 1000 resamples looks like:

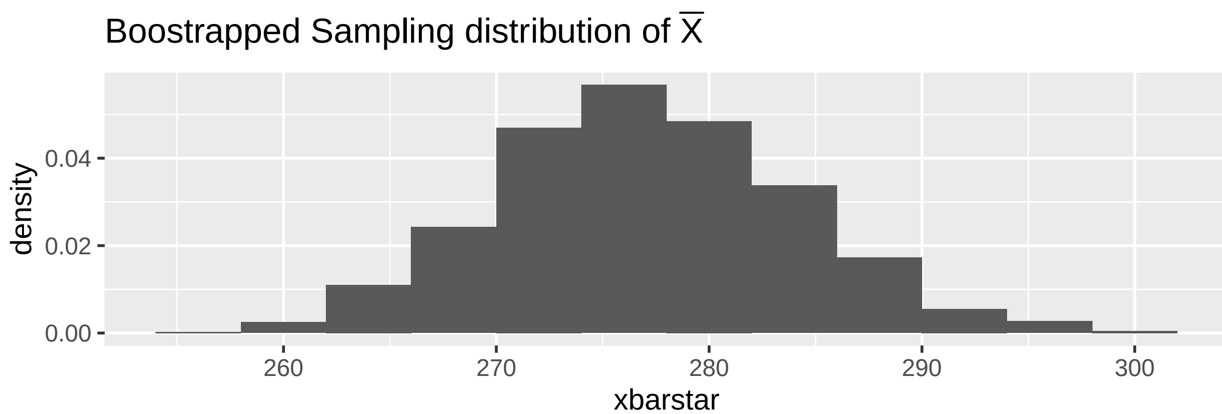
```
set.seed(14141)
B = 1000
boot_df <- tibble(
```

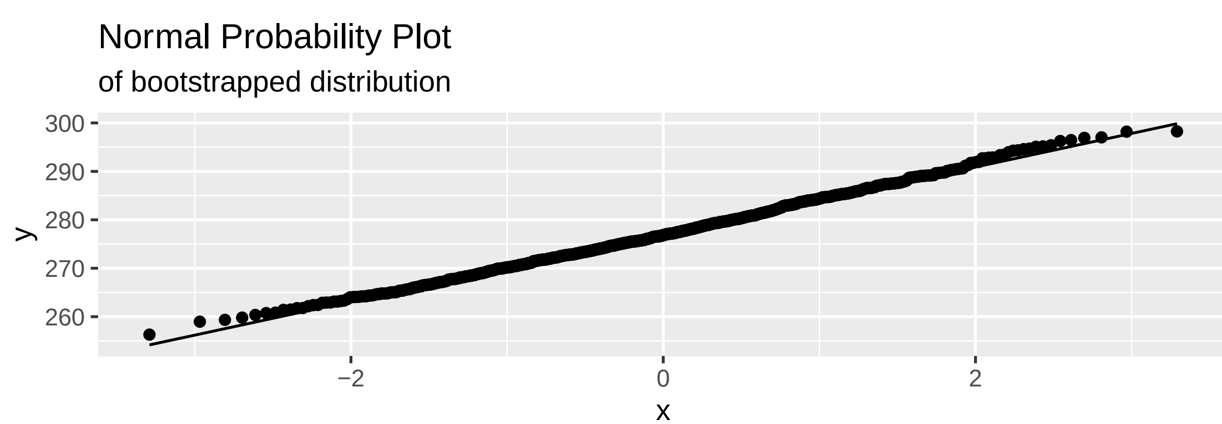
```
  xbarstar = replicate(n = B,
                       expr = mean(sample(x = airbnb_new$price,
                                          size = nrow(airbnb_new),
                                          replace = TRUE))))

p1 <- ggplot(data = boot_df,
          mapping = aes(x=xbarstar,
                        y=..density..))+
   #sturges rule: diff(range(xbarstar))/log2(B)
  geom_histogram(binwidth = 4) +
  labs(title = expression(paste("Boostrapped Sampling distribution of ",bar(X))))

p2 <- ggplot(data = boot_df,
          mapping = aes(sample = xbarstar))+
  stat_qq(distribution  = qnorm)+
  stat_qq_line(distribution = qnorm)+
  labs(title = "Normal Probability Plot",
       subtitle = "of bootstrapped distribution")

library(gridExtra)
grid.arrange(p1,p2)
```





We see from our relatively symmetric histogram and well fit qq line that a normal distribution is a good fit for these bootstrapped samples. Thus we may do the standard boostrap 95% interval of:

$$[\bar{X} - qnorm(.975) * se_{bootstrap}, \bar{X} + qnorm(.975) * se_{bootstrap}]$$

We see that this formula yields a 95% CI of:

```
round(mean(airbnb_new$price) + c(-1,1)*qnorm(.975)*sd(boot_df$xbarstar),3)
```

```
## [1] 263.488 291.021
```

We see this is quite similar to the large sample confidence interval in part a, which shows the strength of the CLT.