ECON 3640–001

Problem Set 4 – Solutions

Marcio Santetti

Spring 2022

Problem 1

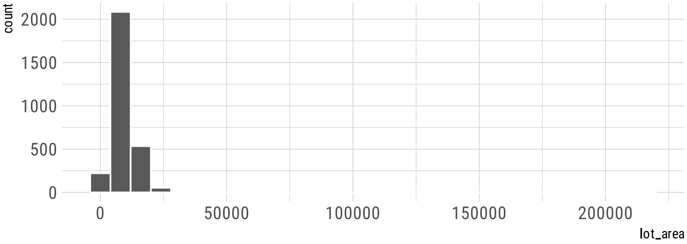
Using the `AmesHousing` package, do the following:

```
library(tidyverse)
library(AmesHousing)
library(janitor)
library(ggeasy)

ames_data <- ames_raw

ames_data <- ames_data %>%
  clean_names()
```

(a) Select the `lot_area` column, showing data on lot sizes (in square feet). Draw its histogram.

ANSWER:

```
ames_data %>%
  ggplot(aes(x = lot_area)) +
  geom_histogram(color = "white", binwidth = 8000)
```



(b) Given that this data set has information on **all** houses in Ames, Iowa between 2006 and 2010, we may assume that it brings the whole *population* data. Calculate the average lot size (μ) for this population.

ANSWER:

```
ames_data %>%
  summarize(mean_lot = mean(lot_area))
```

```
## # A tibble: 1 × 1
##   mean_lot
##      <dbl>
## 1   10148.
```

(c) Now, run a sampling procedure (using a `for` loop), extracting 5,000 samples of size `n = 50` from the population data on lot areas. Compute the sample means $\bar{x}$ of these samples and store them in an array called `samples_50`.

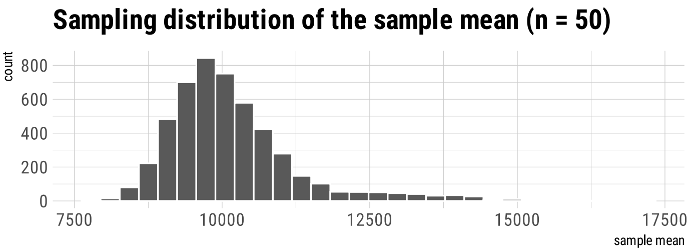
ANSWER:

```
lot <- ames_data %>%
  pull(lot_area)    ## pulling the data for the lot_area variable.


samples_50 <- rep(NA, 5000)

for(i in 1:5000){

  sampling_50 <- sample(lot, size = 50)

  samples_50[i] <- mean(sampling_50)

}
```

(d) Given that in part (c) you've computed a summary statistic from a population, plot a histogram of the sampling distribution of these $\bar{x}$ values. **Hint**: transform your `samples_50` array in a `tibble` first.

ANSWER:

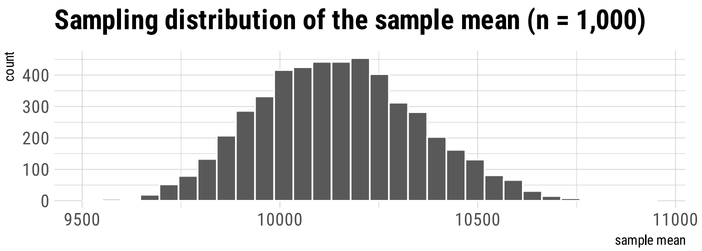
```
samples_50 <- samples_50 %>%
  as_tibble()


samples_50 %>%
  ggplot(aes(x = value)) +
  geom_histogram(color = "white") +
  labs(title  = "Sampling distribution of the sample mean (n = 50)",
       x = "sample mean")
```



(e) Now, run the same procedure as in part (c), this time increasing the sample size to `n = 1,000`. Plot a histogram reflecting the sampling distribution of x.

ANSWER:

```
samples_1000 <- rep(NA, 5000)

for(i in 1:5000){

  sampling_1000 <- sample(lot, size = 1000)

  samples_1000[i] <- mean(sampling_1000)

}

samples_1000 <- samples_1000 %>%
  as_tibble()


samples_1000 %>%
  ggplot(aes(x = value)) +
  geom_histogram(color = "white") +
  labs(title  = "Sampling distribution of the sample mean (n = 1,000)",
       x = "sample mean")
```



(f) Compare the population mean (μ) with the means of the two sampling distributions for $\bar{x}$ you've computed in parts (c) and (e).

ANSWER:

```
ames_data %>%
  summarize(mean_lot = mean(lot_area))
```

```
## # A tibble: 1 × 1
##   mean_lot
##      <dbl>
## 1   10148.
```

```
samples_50 %>%
  summarize(mean_sampling50 = mean(value))
```

```
## # A tibble: 1 × 1
##   mean_sampling50
##             <dbl>
## 1          10173.
```

```
samples_1000 %>%
  summarize(mean_sampling1000 = mean(value))
```

```
## # A tibble: 1 × 1
##   mean_sampling1000
##               <dbl>
## 1            10148.
```

We can see that, as the sample size increases, the sampling distribution of the average lot area converges to the "true" population mean value μ.

(g) Compare the population variance (σ^2) with the variances of the two sampling distributions for $\bar{x}$ you've computed in parts (c) and (e).

ANSWER:

```
samples_50 %>%
  summarize(var_sampling50 = var(value))
```

```
## # A tibble: 1 × 1
##   var_sampling50
##            <dbl>
## 1       1296148.
```

```
samples_1000 %>%
  summarize(var_sampling1000 = var(value))
```

```
## # A tibble: 1 × 1
##   var_sampling1000
##              <dbl>
## 1           41691.
```

As expected, the variance of the sampling distribution with larger sample size is small, relative to the one with *n = 50*.

(h) Run the same procedure as you've done in parts (c)—(g), this time with a different summary statistic: the population and sample **median**.

ANSWER:

```
# Now, the median:

ames_data %>%
  summarize(median_lot = median(lot_area))
```

```
## # A tibble: 1 × 1
##   median_lot
##        <dbl>
## 1      9436.
```

```
#-- sampling n = 50:

samples_50_median <- rep(NA, 5000)

for(i in 1:5000){

  sampling_50_median <- sample(lot, size = 50)

  samples_50_median[i] <- median(sampling_50_median)

}

samples_50_median <- samples_50_median %>%
  as_tibble()


#-- sampling n = 1,000

samples_1000_median <- rep(NA, 5000)

for(i in 1:5000){

  sampling_1000_median <- sample(lot, size = 1000)

  samples_1000_median[i] <- median(sampling_1000_median)

}
```

```
samples_1000_median <- samples_1000_median %>%
  as_tibble()


ames_data %>%
  summarize(median_lot = median(lot_area))
```

```
## # A tibble: 1 × 1
##   median_lot
##        <dbl>
## 1      9436.
```

```
samples_50_median %>%
  summarize(median_samples50 = median(value))
```

```
## # A tibble: 1 × 1
##   median_samples50
##              <dbl>
## 1             9426
```

```
samples_1000_median %>%
  summarize(median_samples1000 = median(value))
```

```
## # A tibble: 1 × 1
##   median_samples1000
##                <dbl>
## 1              9441.
```

We observe a similar converging behavior for the sample median when the sampling procedure happens with a larger sample size.

Problem 2

Answer the following questions using the `qnorm()` and `rnorm()` functions in `R`:

(a) What is the total area to the **left** of 1 for a random variable $X \sim \mathcal{N}(2, 1)$? In other words, what is *P(X < 1)*?

ANSWER:

```
pnorm(q = 1, mean = 2, sd = 1)
```

```
## [1] 0.1586553
```

(b) What is the quantile corresponding to a probability of 85% to its **left** for a random variable $X \sim \mathcal{N}(3, 1)$?

ANSWER:

```
qnorm(p = 0.85, mean = 3, sd = 1)
```

```
## [1] 4.036433
```

(c) For a Standard Normal distribution, what **quantile** corresponds to a probability of 2.5% to its left?

ANSWER:

```
qnorm(p = 0.025, mean = 0, sd = 1)
```

```
## [1] -1.959964
```

(d) For a Standard Normal distribution, what **quantile** corresponds to a probability of 5% to its left?

ANSWER:

```
qnorm(p = 0.05, mean = 0, sd = 1)
```

```
## [1] -1.644854
```

(e) For a Standard Normal distribution, what **quantile** corresponds to a probability of .5% to its left?

ANSWER:

```
qnorm(p = 0.005, mean = 0, sd = 1)
```

```
## [1] -2.575829
```

Problem 3

A statistician took a random sample of 50 observations from a population with a population standard deviation (σ) of 25 and computed the sample mean to be 100.

(a) Estimate the population mean with 90% confidence.

ANSWER:

$$\text{CI} = \bar{x} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

What is the value of $z_{\alpha/2}$ for a significance level (α) of 10%?

```
qnorm(p = 0.10/2, mean = 0, sd = 1)
```

```
## [1] -1.644854
```

Plugging in:

$$\text{CI} = \bar{x} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) = 100 \pm (-1.64)\left(25/\sqrt{50}\right) = [94.2; 105.8]$$

(b) Repeat part (a) using a 95% confidence level.

ANSWER:

What is the value of $z_{\alpha/2}$ for a significance level (α) of 5%?

```
qnorm(p = 0.05/2, mean = 0, sd = 1)
```

```
## [1] -1.959964
```

Using `R` as a calculator:

```
# Lower confidence limit:
100 + (-1.96) * ( 25 / sqrt(50) )
```

```
## [1] 93.07035
```

```
# Upper confidence limit:
100 - (-1.96) * ( 25 / sqrt(50) )
```

```
## [1] 106.9296
```

So the confidence interval for the population mean at 95% is [93.07; 106.93].

(c) Repeat part (a) using a 99% confidence level.

ANSWER:

What is the value of $z_{\alpha/2}$ for a significance level (α) of 1%?

```
qnorm(p = 0.01/2, mean = 0, sd = 1)
```

```
## [1] -2.575829
```

```
# Lower confidence limit:
100 + (-2.57) * ( 25 / sqrt(50) )
```

```
## [1] 90.91368
```

```
# Upper confidence limit:
100 - (-2.57) * ( 25 / sqrt(50) )
```

```
## [1] 109.0863
```

So the confidence interval for the population mean at 99% is [90.91; 109.08].

(d) Describe the effect on the confidence interval estimate of increasing the confidence level.

ANSWER:

As the confidence (significance) level increases (decreases), the *range* of our confidence interval increases. In other words, the *variance* of our estimation increases the more confident we want to be regarding our confidence interval.

Problem 4

The mean of a random sample of 25 observations from a normal population with a standard deviation (σ) of 50 is 200.

(a) Estimate the population mean with 95% confidence.

ANSWER:

```
# Lower confidence limit:
200 + (-1.96) * ( 50 / sqrt(25) )
```

```
## [1] 180.4
```

```
# Upper confidence limit:
200 - (-1.96) * ( 50 / sqrt(25) )
```

```
## [1] 219.6
```

(b) Repeat part (a) changing the population standard deviation to 25.

ANSWER:

```
# Lower confidence limit:
200 + (-1.96) * ( 25 / sqrt(25) )
```

```
## [1] 190.2
```

```
# Upper confidence limit:
200 - (-1.96) * ( 25 / sqrt(25) )
```

```
## [1] 209.8
```

(c) Repeat part (a) changing the population standard deviation to 10.

ANSWER:

```
# Lower confidence limit:
200 + (-1.96) * ( 10 / sqrt(25) )
```

```
## [1] 196.08
```

```
# Upper confidence limit:
200 - (-1.96) * ( 10 / sqrt(25) )
```

```
## [1] 203.92
```

(d) Describe what happens to the confidence interval estimate when the standard deviation is decreased.

ANSWER:

When σ decreases, we can see that the confidence interval, all else constant, becomes more precise. In other words, as σ goes down (we decrease uncertainty/variability), the lower confidence limit goes up, while upper limits of our CI go down.

Problem 5

A random sample of 25 was drawn from a normal distribution with a standard deviation (σ) of 5. The sample mean is 80.

(a) Determine the 95% confidence interval estimate of the population mean.

ANSWER:

```
# Lower confidence limit:
80 + (-1.96) * ( 5 / sqrt(25) )
```

```
## [1] 78.04
```

```
# Upper confidence limit:
80 - (-1.96) * ( 5 / sqrt(25) )
```

```
## [1] 81.96
```

(b) Repeat part (a) with a sample size of 100.

ANSWER:

```
# Lower confidence limit:
80 + (-1.96) * ( 5 / sqrt(100) )
```

```
## [1] 79.02
```

```
# Upper confidence limit:
80 - (-1.96) * ( 5 / sqrt(100) )
```

```
## [1] 80.98
```

(c) Repeat part (a) with a sample size of 400.

ANSWER:

```r
# Lower confidence limit:

80 + (-1.96) * ( 5 / sqrt(400) )
```

```
## [1] 79.51
```

```r
# Upper confidence limit:

80 - (-1.96) * ( 5 / sqrt(400) )
```

```
## [1] 80.49
```

(d) Describe what happens to the confidence interval estimate when the sample size increases.

ANSWER:

All else constant, when the sample size (n) increases, the CI becomes more precise. This happens because the variance is inversely related to the sample size: as the latter increases, the former decreases.

Problem 6

The following data represent a random sample of 9 marks (out of 10) on a statistics quiz. Estimate the population mean with 90% confidence.

$$7\ 9\ 7\ 5\ 4\ 8\ 3\ 10\ 9$$

ANSWER:

Now, we do not have the population standard deviation (σ) as known. Thus we work with its sample estimator, s.

Using `R`:

```r
data_7 ← tibble(
  marks = c(7, 9, 7, 5, 4, 8, 3, 10, 9)
  )

data_7 %>%
  summarize(mean_marks = mean(marks),
            sd_marks = sd(marks))
```

```
## # A tibble: 1 × 2
##   mean_marks sd_marks
##        <dbl>    <dbl>
## 1       6.89     2.42
```

The sample mean ($\bar{x}$) is 6.89 points and the sample standard deviation (s) is 2.42 points.

Then, we proceed with a 90% confidence interval:

$$\text{CI} = \bar{x} \pm t_{\alpha/2,\, \nu} \left(\frac{s}{\sqrt{n}} \right)$$

What is the value of $t_{\alpha/2,\, \nu}$ for a significance level (α) of 10%?

```
qt(p = 0.05, df = 9 - 1)  ## degrees of freedom (df) is the sample size (n) minus 1.
```

```
## [1] -1.859548
```

$$\text{CI} = \bar{x} \pm t_{\alpha/2,\, \nu} \left(\frac{s}{\sqrt{n}} \right) = 6.89 \pm (-1.85) \left(\frac{2.42}{\sqrt{9}} \right) = [5.39; 8.39]$$

Problem 7

For the following parts, calculate the value of the *test statistic*, set up the *rejection region*, determine the *p-value* and *interpret* the result.

(a) H_0: $\mu = 1,000$; H_a: $\mu \neq 1,000$

$\sigma = 200$, $n = 100$, $\bar{x} = 980$, $\alpha = 0.05$

ANSWER:

Test (*z*) statistic:

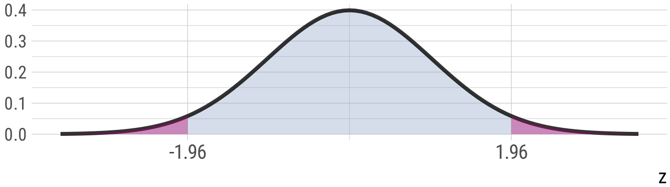
$$z = \frac{\bar{x} - \mu_{H_0}}{\sigma/\sqrt{n}} = \frac{980 - 1,000}{200/\sqrt{100}} = -1$$

This is a *two-tailed test*, given the *sign* of the alternative hypothesis. Thus, to obtain the critical (threshold) value, we divide α by 2.

```
qnorm(p = 0.05/2, mean = 0, sd = 1)
```

```
## [1] -1.959964
```

Visually:



Our test statistic of -1 falls oustide the rejection regions. Therefore, we **do not reject the null hypothesis**. There is not enough evidence to decide in favor of the alternative hypothesis, H_a.

Using the p-value method for a two-tailed test:

$$\text{p-value} = P(Z < -1) + P(Z > 1)$$

```
pnorm(q = -1, mean = 0, sd = 1) + 1 - pnorm(q = 1, mean = 0, sd = 1)
```

```
## [1] 0.3173105
```

This p-value is greater than the significance level. Thus, we do not reject the null hypothesis using either method.

(b) H_0: $\mu = 50$; H_a: $\mu > 50$

$\sigma = 5$, $n = 9$, $\bar{x} = 51$, $\alpha = 0.03$

ANSWER:

Test statistic:

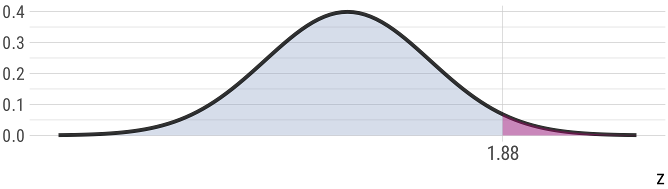
$$z = \frac{\bar{x} - \mu_{H_0}}{\sigma/\sqrt{n}} = \frac{51 - 50}{5/\sqrt{9}} = 0.6$$

Given the value of α and since this is a one-tailed test, our critical value is:

```
qnorm(p = 0.03, mean = 0, sd = 1, lower.tail = FALSE)
```

```
## [1] 1.880794
```

Visually:



Our test statistic falls oustide the rejection region. Thus, we also do not reject the null hypothesis.

Using the p-value method:

```
1 - pnorm(q = 1.88, mean = 0, sd = 1)  ## use the "1 - " since this is a right-tailed test.
```

```
## [1] 0.03005404
```

This p-value is greater than the significance level, so we do not reject the null hypothesis, regardless of the method.

Problem 8

Many Alpine ski centers base their projections of revenues and profits on the assumption that the average Alpine skier skis exactly four times per year. To investigate the validity of this assumption, a random sample of 63 skiers is drawn and each is asked to report the number of times he or she skied the previous year. If we assume that the standard deviation (σ) is 2, can we infer at the 10% significance level that the assumption is wrong? Assume the sample mean to be 5.3 times per year.

ANSWER:

Let's first *state the null and alternative hypotheses*:

H_0: $\mu = 4$

H_a: $\mu \neq 4$

Then, we compute the test statistic:

$$z = \frac{\bar{x} - \mu_{H_0}}{\sigma/\sqrt{n}} = \frac{5.3 - 4}{2/\sqrt{63}} = 5.159$$

Given that our level of significance is 10% and we have a two-tailed test, the critical values are:

```
qnorm(p = 0.05, mean = 0, sd = 1)
```

```
## [1] -1.644854
```

```
qnorm(p = 0.05, mean = 0, sd = 1, lower.tail = FALSE)
```

```
## [1] 1.644854
```

Our test statistic of 5.159 is way greater than the right-tail critical value of 1.64. Thus, by the rejection region method, we **reject** the null hypothesis in favor of the alternative.

Using the p-value method:

```
pnorm(q = - 5.159, mean = 0, sd = 1) + 1 - pnorm(q = 5.159, mean = 0, sd = 1)
```

```
## [1] 2.482723e-07
```

The p-value is way lower tha the significance level. So, also using the p-value method, we reject the null hypothesis. Therefore, the assumption that the average Alpine skier skis exactly four times per year is incorrect.

Problem 9

University bookstores order books that instructors adopt for their courses. The number of copies ordered matches the projected demand. However, at the end of the semester, the bookstore has too many copies on hand and must return them to the publisher. A bookstore has a policy that the proportion of books returned should be kept as small as possible. The average is supposed to be less than 10%. To see whether the policy is working, a random sample of book titles was drawn, and the fraction of the total originally ordered that are returned is recorded and listed here. Can we infer at the 10% significance level that the mean proportion of returns is less than 10%?

4 15 11 7 5 9 4 3 5 8

ANSWER:

Let's first *state the null and alternative hypotheses*:

H_0: μ = 10
H_a: μ < 10

We don't know the population standard deviation (σ) here. So, we need to compute both the sample mean and sample standard deviation.

```r
books <- tibble(
  prop = c(4, 15, 11, 7, 5, 9, 4, 3, 5, 8)
)

books %>%
  summarize(mean_prop = mean(prop),
            sd_prop = sd(prop))
```

```
## # A tibble: 1 × 2
##   mean_prop sd_prop
##       <dbl>   <dbl>
## 1       7.1    3.75
```

Then, we compute the test statistic:

$$t = \frac{\bar{x} - \mu_{H_0}}{s/\sqrt{n}} = \frac{7.1 - 10}{3.75/\sqrt{10}} = -2.445$$

We have $n - 1 = 9$ degrees of freedom. So the critical value is:

```r
qt(p = 0.10, df = 9)
```

```
## [1] -1.383029
```

Our test statistic lies within the rejection region. Thus, we **reject the null hypothesis** in favor of the alternative. So we can conclude that the bookstore's policy is working.

Using the p-value method:

```
pt(q = -2.445, df = 9)
```

```
## [1] 0.01852972
```

This p-value is way lower than the significance level of 10%. Thus, using this method we also reject the null hypothesis.

Problem 10

Companies that sell groceries over the Internet are called e-grocers. Customers enter their orders, pay by credit card, and receive delivery by truck. A potential e-grocer analyzed the market and determined that the average order would have to exceed $85 if the e-grocer were to be profitable. To determine whether an e-grocery would be profitable in one large city, she offered the service and recorded the size of the order for a random sample of customers. Can we infer from these data that an e-grocery will be profitable in this city?

100 120 75 40 89 51 200 96 31

ANSWER:

Let's first *state the null and alternative hypotheses*:

H_0: μ = \$85

H_a: μ > \$85

```
grocers <- tibble(
  order = c(100, 120, 75, 40, 89, 51, 200, 96, 31)
)

grocers %>%
  summarize(mean_order = mean(order),
            sd_order = sd(order))
```

```
## # A tibble: 1 × 2
##   mean_order sd_order
##        <dbl>    <dbl>
## 1       89.1     51.1
```

Computing the test statistic:

```
# t-test using R as calculator:

(89.1 - 85) / (51.1 / sqrt(9))
```

```
## [1] 0.2407045
```

Now, the critical value:

```r
qt(p = .05, df = 9 - 1, lower.tail = FALSE)  ## assuming 5% of significance.
```

```
## [1] 1.859548
```

Our test statistic is not included in the rejection region. Therefore, we **do not reject the null hypothesis**. In summary, the e-grocery will not be profitable in this city, given the sample data available.

Using the p-value method:

```r
1 - pt(q = 1.85, df = 9 - 1)
```

```
## [1] 0.05073832
```

This p-value is greater than our assumed significance level of 5%. Thus, this method also does not recommend rejecting the null hypothesis.