Homework 1 Solution

Charlotte Wickham

2017-10-05

```
set.seed(8389824)

knitr::opts_chunk$set(message = FALSE, warning = FALSE,
  out.height = "2in",
  fig.width = 6, fig.height = 3)
```

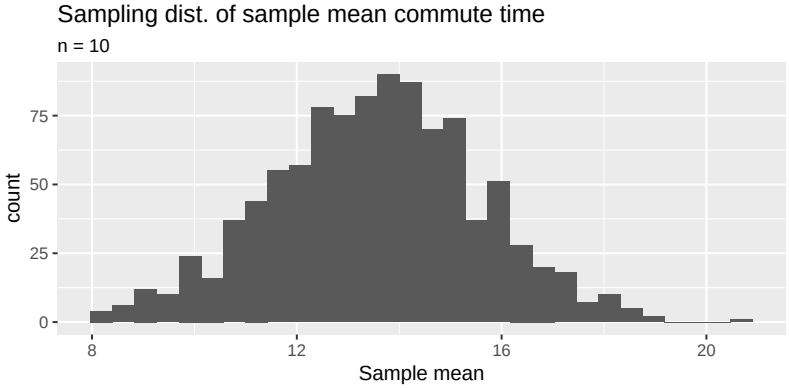
1. a) A population parameter is a one number summary of the population distribution, a statistic is a one number summary of a sample, i.e. the data at hand.

   b) The sampling distribution of a statistic is the distribution of the statistic's value over all possible random samples from the population.

2.
```
library(tidyverse)
# Read in class data
class_data <- read_csv("class_data.csv")

# Simulation settings
n_sim <- 1000
```

   a. Simulate the sampling distribution for the sample mean for samples of **size n = 10** from the commute times.

```
# Generate many samples
commute_samples_10 <- rerun(.n = n_sim,
  sample(class_data$commute_time, size = 10))

# Find sample mean of each sample
commute_sample_means <- map_dbl(commute_samples_10, ~ mean(.x))

ggplot() +
  geom_histogram(aes(x = commute_sample_means))+
  labs(x = "Sample mean",
    title = "Sampling dist. of sample mean commute time",
    subtitle = "n = 10")
```



   i) (1 pt) Describe how the distribution differs to that based on samples of size 5.

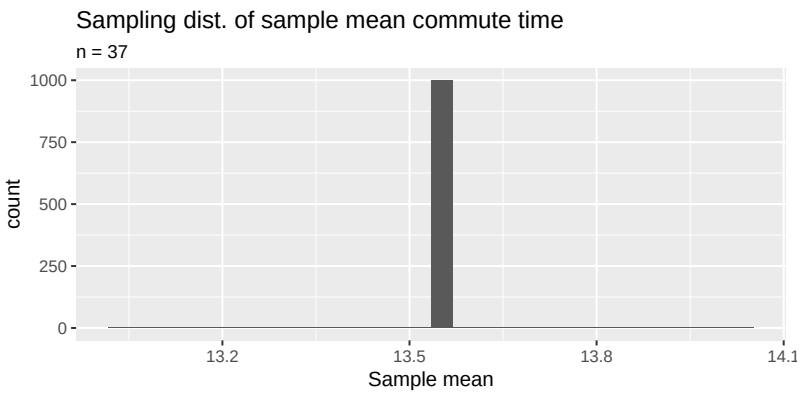
The primary difference is that the distribution is narrower.

ii) (2 pts) Justify your observations based on the properties we derived in Lecture on Fri Sep 29.

We saw, that the variance of the sample mean is σ^2/n. We would expect the variance of the sample mean with a sample size of 10 to be $1/2$ of that of the sample mean with a sample size of 5.

b. Simulate the sampling distribution for the sample mean for samples of **size n = 37** from the commute times.

```
# Generate many samples
commute_samples_37 <- rerun(.n = n_sim,
  sample(class_data$commute_time, size = 37))

# Find sample mean of each sample
commute_sample_means_37 <- map_dbl(commute_samples_37, ~ mean(.x))

ggplot() +
  geom_histogram(aes(x = commute_sample_means_37)) +
  labs(x = "Sample mean",
    title = "Sampling dist. of sample mean commute time",
    subtitle = "n = 37")
```



i) (1 pt) Examine the distribution. Are you surprised?

The distribution has no variation, the sample mean is always 13.54.

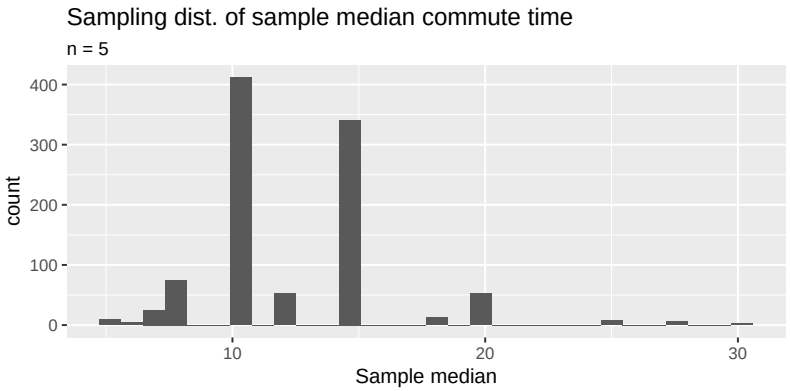
ii) (1 pt) Describe why the distribution looks the way it does.

Since the population is of size 37, a sample of size 37 without replacement is exactly the population. Every sample has the same mean, the population mean, and therefore has no variation.

c. Simulate the sampling distribution for the sample **median** for samples of size 5 from the commute times.

```
# Generate many samples
commute_samples_5 <- rerun(.n = n_sim,
  sample(class_data$commute_time, size = 5))

# Find sample median of each sample
commute_sample_medians <- map_dbl(commute_samples_5, ~ median(.x))

ggplot() +
  geom_histogram(aes(x = commute_sample_medians)) +
  labs(x = "Sample median",
```

```
    title = "Sampling dist. of sample median commute time",
    subtitle = "n = 5")
```



i) (1 pt) Describe the distribution.

The sample median only takes a discrete set of values (because there are a finite set of commute times, and the median will take only those values, or value halfway between two consecutive values). The sample median is most often 10 or 15, but we see values between 5 and 30.

ii) (1 pt) Use your simulated sampling distribution to estimate the probability, that a sample of size 5 from this population results in a median less than 10.

```
mean(commute_sample_medians < 10)
```

```
## [1] 0.113
```

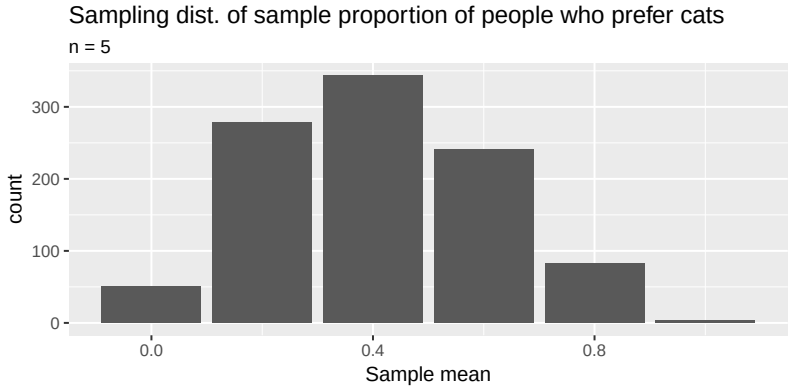
The probability the sample median commute time for a sample of size 5 is less than 10 is estimated to be 0.113.

iii) (1 pt) Do any of the properties of sampling distributions we derived in the lecture from Mon 25 to Fri 29 Sep justify the properties of this sampling distribution?

No, we only derived properties for the sample mean. (Although we did show that simulation of a sampling distribution should approach the true sampling distribution if we take large enough samples, so the observations made in this part are justified).

d. Simulate the sampling distribution for the sample **mean** for samples of size 5 from the `prefers_cats` column in `class_data`.

```
# Generate many samples
cats_samples_5 <- rerun(.n = n_sim,
  sample(class_data$prefers_cats, size = 5))

# Find sample mean of each sample
cats_sample_means <- map_dbl(cats_samples_5, ~ mean(.x))

ggplot() +
  geom_bar(aes(x = cats_sample_means))+
  labs(x = "Sample mean",
    title = "Sampling dist. of sample proportion of people who prefer cats",
    subtitle = "n = 5")
```



i) (1 pt) Describe the distribution. *(A histogram might not be the most appropriate plot here.)*

The sample mean of the binary `prefers_cats` variable only takes the values: 0, 0.2, 0.4, 0.6, 0.8, 1.0. These correspond to observing 0, 1, 2, 3, 4, 5 people who prefer cats in our sample of size 5. The distribution has highest density at 0.4, decreasing to the left and right.

ii) (1 pt) Do any of the properties of sampling distributions we derived in the lecture from Mon 25 to Fri 29 Sep justify the properties of this sampling distribution?

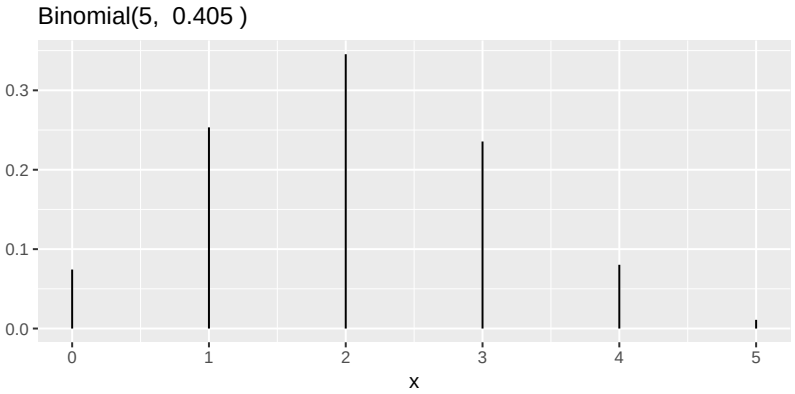
Yes. First we derived (well we saw) the exact sampling distribution for the sample mean in a Bernoulli population is a scaled Binomial that depends on the population proportion, p. Since we have access to the population we can calculate p:

```
n <- 5
(p <- mean(class_data$prefers_cats))
```

```
## [1] 0.4054054
```

and find the population proportion is 0.405. We can see the correspondence with the Binomial(n, p):

```
x <- 0:5
ggplot() +
  geom_linerange(aes(x = x, ymin = 0, ymax = dbinom(x, size = n, prob = p))) +
  labs(title = paste("Binomial(5, ", round(p, 3), ")"))
```



Second, since this is the sample mean we also know its mean is the same as the population mean, $E(\overline{Y}) = \mu = p = 0.405$, and its variance is the population variance scaled by n, $Var(\overline{Y}) = \sigma^2/n = p(1-p)/n = 0.048$.

e. Instead of sampling from `class_data`, you could sample from a population with a standard Normal distribution, by replacing `sample(class_data$commute_time, size = n)`, with `rnorm(n)`.

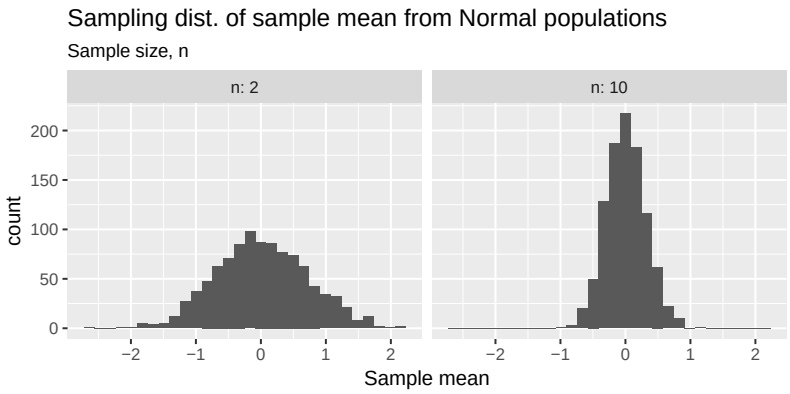
i) (2 pts) **Derive** the sampling distribution of the sample mean for samples of size n $= 2$ and n $= 10$ from a standard Normal population.

From class we saw if $Y \sim N(\mu, \sigma^2)$, then $\overline{Y} \sim N(\mu, \sigma^2/n)$.

So for $n = 2$, $\overline{Y} \sim N(0, 1/2)$
So for $n = 10$, $\overline{Y} \sim N(0, 1/10)$

ii) (2 pts) **Simulate** the sampling distribution of the sample mean for samples of size n $= 2$ and n $= 10$ from a standard Normal population. Are your simulated distributions consistent with your derived distributions?

```
norm_df <- data_frame(n = c(2, 10),
  samples = list(rerun(.n = n_sim, rnorm(2)),
                 rerun(.n = n_sim, rnorm(10))))

# Find sample mean of each sample
norm_df <- norm_df %>%
  mutate(mean = map(samples, ~ map_dbl(.x, ~ mean(.x)))) %>%
  unnest(mean)

ggplot(norm_df) +
  geom_histogram(aes(x = mean)) +
  facet_wrap(~ n, labeller = "label_both") +
  labs(x = "Sample mean",
    title = "Sampling dist. of sample mean from Normal populations",
    subtitle = "Sample size, n")
```




```
norm_df %>%
  group_by(n) %>%
  summarise(mean_ybar = mean(mean),
    var_ybar = var(mean))
```

```
## # A tibble: 2 x 3
##       n   mean_ybar   var_ybar
##   <dbl>       <dbl>      <dbl>
## 1     2 0.017192915 0.50741582
## 2    10 0.008682429 0.09055921
```

Yes, histograms look roughly Normal centered at zero and with variances about 0.5 and 0.1