Central Limit Theorem, Samping, Standard Error

Simon J. Kiss

19/03/2020

Learning Outcomes

1. Students will understand to be able to provide a working definition of
    - *sample*
    - *population*
    - *probability based sample*
    - *central limit theorem*
    - *standard error*
    - *confidence interval*
    - *margin of error*

The problem

- Difference between a sample and a population
  - *Population is the universe of things you want to describe*
  - *Sample is the subset of the universe you have at hand*
- Different notations are used

Measure	**Sample**	**Population**
Average	$\bar{x}$	μ
Standard Deviation	s	σ

The Problem

- We want to know the population average, but we only get to measure a sample average?

Q: How well does the sample average reflect the population average?

A: Remarkably well when a sample is random (probabilistic) and large enough

The Problem

- Probabilistic (random) sample
    - *Ideal way to sample a population; all units of the population must have an equal chance to be selected into the sample*
- non-probabilistic sample
    - *Less-than-ideal, but still useful*
    - *Online panels of volunteers (public opinion research companies)*
    - *Snowball sampling (for hard to reach populations)*

The Central Limit Theorem

The distribution of sample means, drawn randomly, approximates a normal distributions, regardless of the distribution of the underlying population, as the sample gets larger.

- With a sufficiently large sample size, the average of a sample will reliably approach the average of the population

The Data

- Statistics Canada publishes Public Use Microdata File
- Available through the Laurier Libray, via ODESI
- It contains a *large* probabilistic (random) sample of the 2016 census
- individual level data on almost 1000000 Canadians

```
load(url("https://github.com/sjkiss/DMJN328/raw/master/Lecture_Notes/mar_23/data/census_2016.rdata"))
```

Example

- First we use the `look_for()` command in the `labelled` library

```
library(labelled)
look_for(census_2016, "wage")
```

```
##     variable                                      label
## 138    Wages Income: Wages, salaries and commissions
```

- So now we have the `Wages` variable

Example

- Check the average

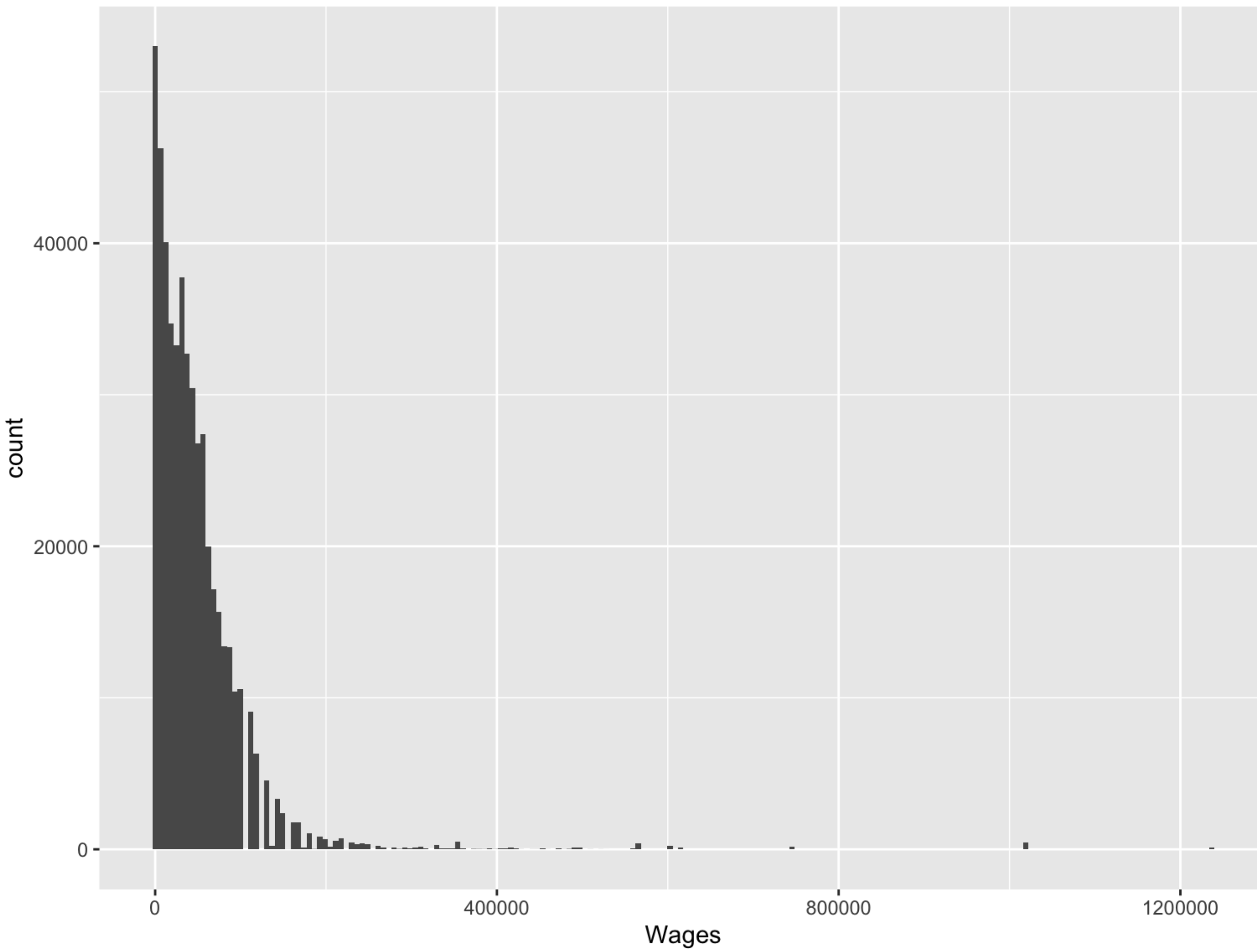
```
mean(census_2016$Wages, na.rm=T)
```

```
## [1] 47546.85
```

- So this could be considered the average salary for all of Canadians, i.e. the *population* average

Visualize Population Wages

```
library(tidyverse)
census_2016 %>%
  ggplot(., aes(x=Wages))+geom_histogram(bins=200)
```



Sampling from a Population

- Select only the wages variable

```
census_2016 %>%
  select(Wages)-> df
```

Sampling from a Population

- Take 100 random samples of size 5

```
#100 times
100 %>%
  #rerun the next command; sample_n takes random samples
  #na.omit deletes missing values, 5 is the size of the sample
  rerun(sample_n(na.omit(df), 5)) %>%
  #calculate the average of each sample
  map_df(~summarize(., avg=mean(Wages))) ->n5
#print the results
```

Sampling from a Population

Automated Example

- Take 100 random samples of size 5

```
print(n5)
```

```
## # A tibble: 100 x 1
##        avg
##      <dbl>
##   1 38600.
##   2 43800
##   3 46948.
##   4 26000
##   5 92600.
##   6 39400
##   7 26000
##   8 54800
##   9 51800
##  10 71200
## # … with 90 more rows
```

Sampling from a Population

Automated Example

- Take 100 random samples of size 10

```
#100 times
100 %>%
  #rerun the next command; sample_n takes random samples
  #na.omit deletes missing values, 5 is the size of the sample
  rerun(sample_n(na.omit(df), 10)) %>%
  #calculate the average of each sample
  map_df(~summarize(., avg=mean(Wages))) ->n10
```

Sampling from a Population

Automated Example

- Take 100 random samples of size 10

```
print(n10)
```

```
## # A tibble: 100 x 1
##       avg
##     <dbl>
##   1 67600
##   2 40700
##   3 30700
##   4 51400
##   5 62600
##   6 44100
##   7 40300.
##   8 57900
##   9 31500.
##  10 49600
## # … with 90 more rows
```

Sampling from a Population

- Take 100 random samples of size 100

```
#100 times
100 %>%
  #rerun the next command; sample_n takes random samples
  #na.omit deletes missing values, 5 is the size of the sample
  rerun(sample_n(na.omit(df), 100)) %>%
  #calculate the average of each sample
  map_df(~summarize(., avg=mean(Wages))) ->n100
```

Sampling from a Population

- Take 100 random samples of size 100

```
print(n100)
```

```
## # A tibble: 100 x 1
##       avg
##     <dbl>
##  1 48700.
##  2 47556.
##  3 43166.
##  4 45390.
##  5 50750.
##  6 48066.
##  7 54475.
##  8 49132.
##  9 63247.
## 10 43547.
## # … with 90 more rows
```

Sampling from a Population

- Take 100 random samples of size 500

```
#100 times
100 %>%
  #rerun the next command; sample_n takes random samples
  #na.omit deletes missing values, 5 is the size of the sample
  rerun(sample_n(na.omit(df), 500)) %>%
  #calculate the average of each sample
  map_df(~summarize(., avg=mean(Wages))) ->n500
#print the results
```

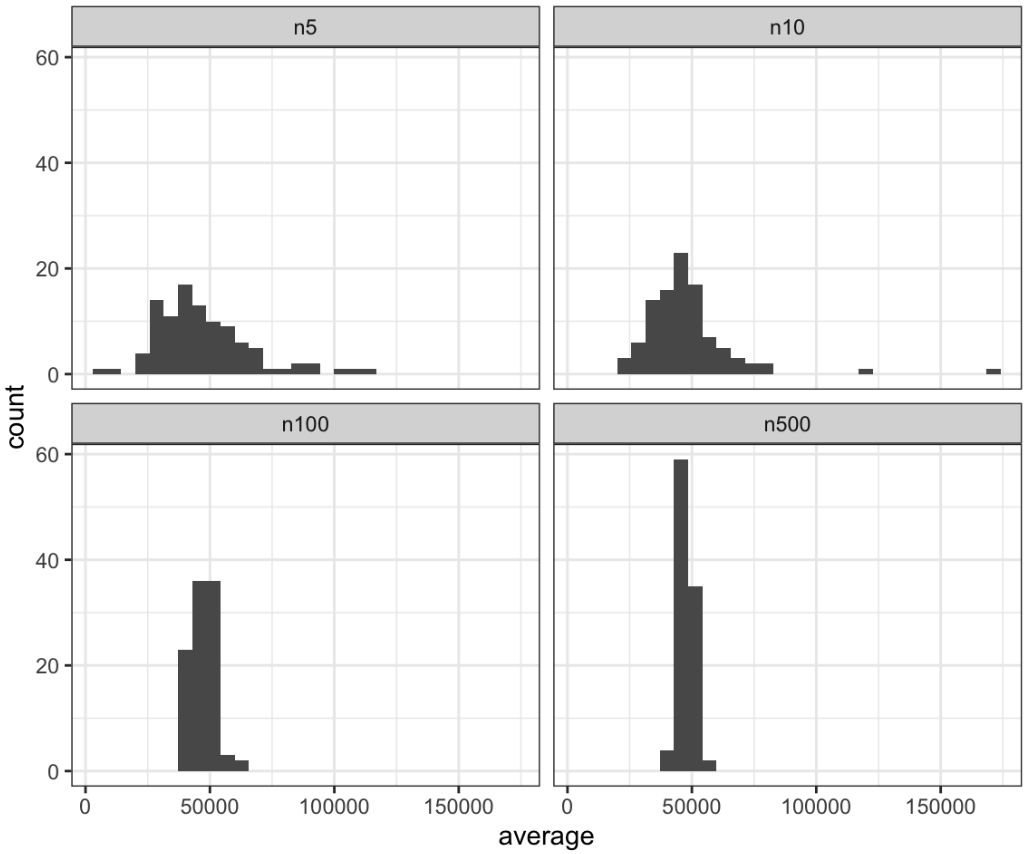
Sampling from a Population

```
print(n500)
```

```
## # A tibble: 100 x 1
##       avg
##     <dbl>
##  1 49830
##  2 44589.
##  3 46528.
##  4 47686.
##  5 44644.
##  6 46943.
##  7 52566.
##  8 47343.
##  9 49863.
## 10 51316.
## # … with 90 more rows
```

Sampling from a Population

- Show the distribution of all the averages from the four different sample sizes
- What do you notice about the samples of different sizes?



Standard Error

- the width of the distribution of sample means can be measured with a measure of dispersion
- Standard Error

$$SE = \frac{s}{\sqrt{n}}$$

- the larger the Standard Error, the larger the sampling distribution.

Relationship between sample size and standard error

- First we take samples of different sizes
- start with 1 sample of size 5

```
#Take 1 sample of size 5 from df$Wages
sample_n(na.omit(df), 5) %>%
  #summrize that sample calculating the averge wage, the standard deviation of the wages and how
        large the sample his
  summarize(avg=mean(Wages), sd=sd(Wages), n=n())-> sd5
```

Relationship between sample size and standard error

- First we take samples of different sizes
- start with 1 sample of size 10

```
#Take 1 sample of size 10 from df$Wages
sample_n(na.omit(df), 10) %>%
  #summrize that sample calculating the averge wage, the standard deviation of the wages and how
        large the sample his
  summarize(avg=mean(Wages), sd=sd(Wages), n=n())-> sd10
```

Relationship between sample size and standard error

- First we take samples of different sizes
- start with 1 sample of size 100

```
#Take 1 sample of size 5 from df$Wages
sample_n(na.omit(df), 100) %>%
  #summrize that sample calculating the averge wage, the standard deviation of the wages and how
        large the sample his
  summarize(avg=mean(Wages), sd=sd(Wages), n=n())-> sd100
```

Relationship between sample size and standard error

- First we take samples of different sizes
- start with 1 sample of size 500

```
#Take 1 sample of size 5 from df$Wages
sample_n(na.omit(df), 500) %>%
  #summrize that sample calculating the averge wage, the standard deviation of the wages and how
        large the sample his
  summarize(avg=mean(Wages), sd=sd(Wages), n=n())-> sd500
```

Relationship between sample size and standard error

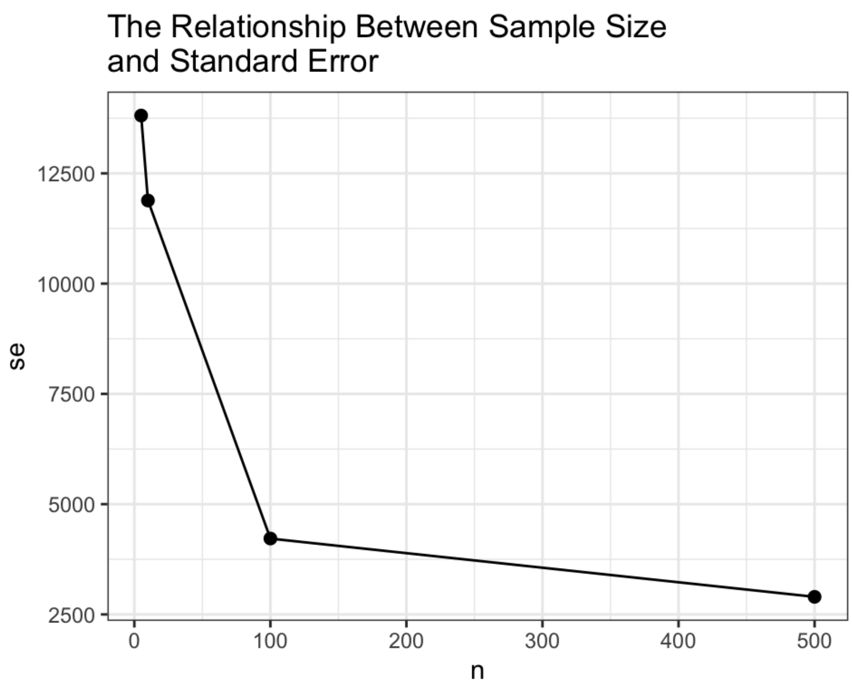
- Combine all into one and calculate the standard error.

$$SE = \frac{s}{\sqrt{n}}$$

```
#Combine all in one
#the bind_rows() function works when data frames have exactly the same variable names.
combined<-bind_rows(sd5, sd10, sd100, sd500)
# Calculate the standard error
combined<-mutate(combined, se=sd/sqrt(n))
```

Relationship between sample size and standard error

```
ggplot(combined, aes(x=n, y=se))+
  geom_point(size=2)+
  geom_line()+
  theme_bw()+
  labs(title="The Relationship Between Sample Size\nand Standard Error")
```



Confidence Intervals

- Because of the CLT, all possible samples are normally distributed
- Because of the Empirical rule (68-95-99), we know that 68% of samples are within one standard deviation (standard error), and 95% of samples are within two standard deviations.
- we can use this to quanitfy the uncertainty associted with any survey measurement.

Confidence Intervals

- Take our last sample of 500
- The average is 47277.01 with a standard error of 2898.9

```
print(combined)
```

```
## # A tibble: 4 x 4
##       avg     sd     n     se
##     <dbl>  <dbl> <int>  <dbl>
## 1 44800. 30882.     5 13811.
## 2 36000. 37582.    10 11885.
## 3 46526  42210.   100  4221.
## 4 47277. 64821.   500  2899.
```

Confidence Intervals

Q: What is the 95% confidence interval for our measurement?

Hint: How many standard deviations above and below a variable's average are 95%

Confidence Intervals

What does the range mean? - Technically, it means that 95% of random samples with this size and this standard deviation would produce an average Wage within these values - Non-technically, it means that we can be 95% confident that the Canadian average wage lies between these two values

Confidence Intervals

Margin of Error in Polls

from The Hill Times

> *Just more than 1 in 3 Americans — 37 percent — said in a new poll that they have a good amount or a great deal of trust in the information they hear about coronavirus from President Trump. … The survey of 835 adults was conducted on Friday and Saturday. It has an overall margin of error of 4.8 percentage points. Among 784 registered voters, the margin of error is 4.9 percentage points.*

Source: The Hill Times