Topic 2: Basics of statistical inference - confidence intervals

Sonja Petrović

Created for ITMD/ITMS/STAT 514

Spring 2021.

Overview

So... you can use a sample statistic to 'guess' a population parameter value. How good is your guess?

Agenda for today:

- point estimates
- why some are better than others
- the true meaning of confidence intervals.
- we will start with a data set that comes with an R package, and we will explore it a bit.

A working example

```
#install.packages("AmesHousing")
library(AmesHousing)
ames <- make_ames()
```

- real estate data from the city of Ames, Iowa.
  - The details of every real estate transaction in Ames is recorded by the City Assessor's office.
  - all residential home sales in Ames between 2006 and 2010.
- This collection represents our population of interest.
- We would like to learn about these home sales by taking smaller samples from the full population.

A working example

```
#install.packages("AmesHousing")
library(AmesHousing)
ames <- make_ames()
```

- real estate data from the city of Ames, Iowa.
  - The details of every real estate transaction in Ames is recorded by the City Assessor's office.
  - all residential home sales in Ames between 2006 and 2010.
- This collection represents our population of interest.
- We would like to learn about these home sales by taking smaller samples from the full population.

How large is this dataset, anyway?

```
nrow(ames)
```

```
[1] 2930
```

```
ncol(ames)
```

```
[1] 81
```

What are the variables in the data set?

```
colnames(ames)
```

```
 [1] "MS_SubClass"        "MS_Zoning"          "Lot_Frontage"
 [4] "Lot_Area"           "Street"             "Alley"
 [7] "Lot_Shape"          "Land_Contour"       "Utilities"
[10] "Lot_Config"         "Land_Slope"         "Neighborhood"
[13] "Condition_1"        "Condition_2"        "Bldg_Type"
[16] "House_Style"        "Overall_Qual"       "Overall_Cond"
[19] "Year_Built"         "Year_Remod_Add"     "Roof_Style"
[22] "Roof_Matl"          "Exterior_1st"       "Exterior_2nd"
[25] "Mas_Vnr_Type"       "Mas_Vnr_Area"       "Exter_Qual"
[28] "Exter_Cond"         "Foundation"         "Bsmt_Qual"
[31] "Bsmt_Cond"          "Bsmt_Exposure"      "BsmtFin_Type_1"
[34] "BsmtFin_SF_1"       "BsmtFin_Type_2"     "BsmtFin_SF_2"
[37] "Bsmt_Unf_SF"        "Total_Bsmt_SF"      "Heating"
[40] "Heating_QC"         "Central_Air"        "Electrical"
[43] "First_Flr_SF"       "Second_Flr_SF"      "Low_Qual_Fin_SF"
[46] "Gr_Liv_Area"        "Bsmt_Full_Bath"     "Bsmt_Half_Bath"
[49] "Full_Bath"          "Half_Bath"          "Bedroom_AbvGr"
[52] "Kitchen_AbvGr"      "Kitchen_Qual"       "TotRms_AbvGrd"
[55] "Functional"         "Fireplaces"         "Fireplace_Qu"
[58] "Garage_Type"        "Garage_Finish"      "Garage_Cars"
[61] "Garage_Area"        "Garage_Qual"        "Garage_Cond"
[64] "Paved_Drive"        "Wood_Deck_SF"       "Open_Porch_SF"
[67] "Enclosed_Porch"     "Three_season_porch" "Screen_Porch"
[70] "Pool_Area"          "Pool_QC"            "Fence"
[73] "Misc_Feature"       "Misc_Val"           "Mo_Sold"
[76] "Year_Sold"          "Sale_Type"          "Sale_Condition"
[79] "Sale_Price"         "Longitude"          "Latitude"
```

So many variables! For today, let's focus on just a couple of variabls: sale price of the home (`Sale_Price`) and the above ground living area of the house in square feet (`Gr_Liv_Area`):

```
summary(ames$Sale_Price)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  12789  129500  160000  180796  213500  755000
```

```
area <- ames$Gr_Liv_Area
summary(area)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    334    1126    1442    1500    1743    5642
```

An old friend: Quantiles??

Remember the definition of, say, *the 25th percentile (Q1)* in the distribution of a r.v. x. Finding these values are useful for describing the distribution, as we can use them for descriptions like "the middle 50% of the homes have areas between such and such square feet".

Population & sample

- We have access to the entire population, but this is rarely the case in real life.
- Gathering information on an entire population is often extremely costly or impossible.
- Because of this, we often take a sample of the population and use that to understand the properties of the population.

Example:

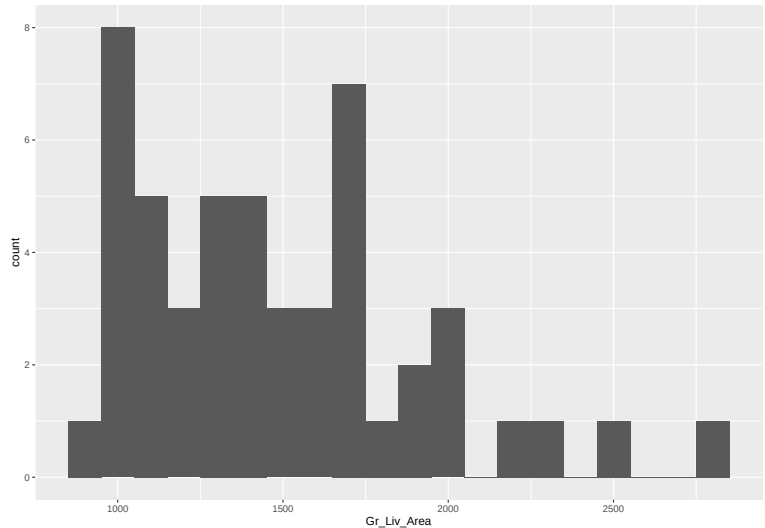
If we were interested in estimating the mean living area in Ames based on a sample, we can use the following command to survey the population:

```
# library("dplyr")  # need f or using sample_n  function belo
sampl <- ames %>%
  sample_n(size = 50)
```

This is like going into the City Assessor's database and pulling up the files on 50 random home sales. Working with these 50 files would be considerably simpler than working with all 2930 home sales.

Sampling distribution of area

```
# library("ggplot2")
ggplot(data = sampl, aes(x = Gr_Liv_Area)) +
  geom_histogram(binwidth = 100)
```



- What's your best guess, based only on this single sample, of an estimate of the average living area of houses sold in Ames?

Point estimators (we won't use much)

- What's your best guess, based only on this single sample, of an estimate of the average living area of houses sold in Ames?
  - $\overline{X}$ sample mean, or sample median:

```
sampl<-as_tibble(sampl)
mean(sampl$Gr_Liv_Area)
```

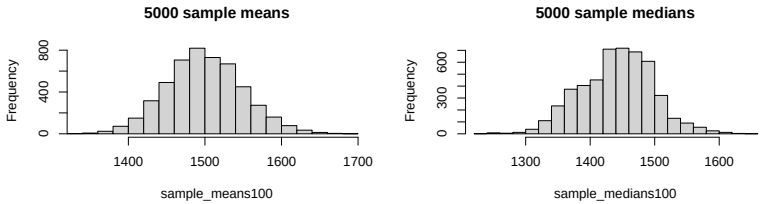
```
[1] 1481.86
```

```
median(sampl$Gr_Liv_Area)
```

```
[1] 1436.5
```

*this is called a **point estimator**.*

Interlude: which point estimator is 'better'?




```
# mean of sample mean:  (!)
mean(sample_means100)
```

```
[1] 1499.426
```

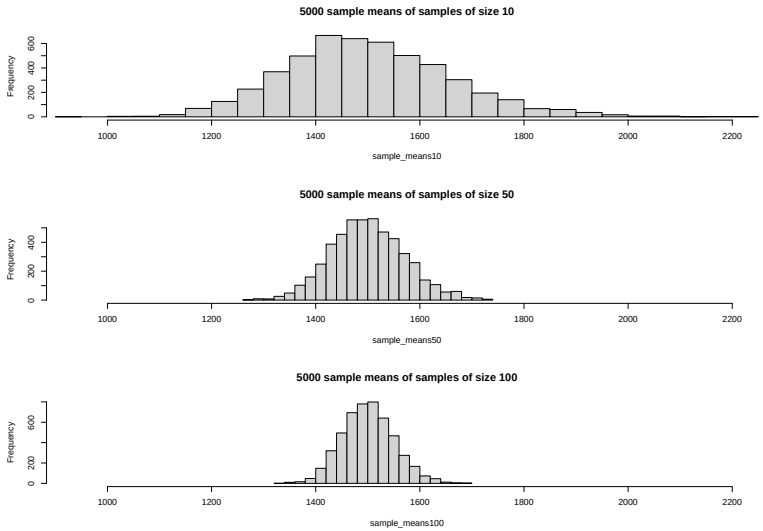
```
# mean of sample median:
mean(sample_medians100)
```

```
[1] 1441.614
```

```
# population mean:
mean(area)
```

```
[1] 1499.69
```

Effect of sample size, revisited



Intervals!

That was a *point estimate*. Let's get a better understanding of the average living area of houses sold in Ames.

> *Remember, you usually do not know the population, so you are 'throwing darts in the dark' to get a feel for this!*

An interval estimate

a random quantity, computed from a sample, that has some pre-set probability of containing the true population parameter.

Example:

The interval $(1345.987, 1575.213)$ contains the true population mean with probability 95%.

- how was this calculated?
- what does the "95% probability" mean?

A formal definition and interpretation of confidence interval

[in the notes]

In pictures: confidence intervals

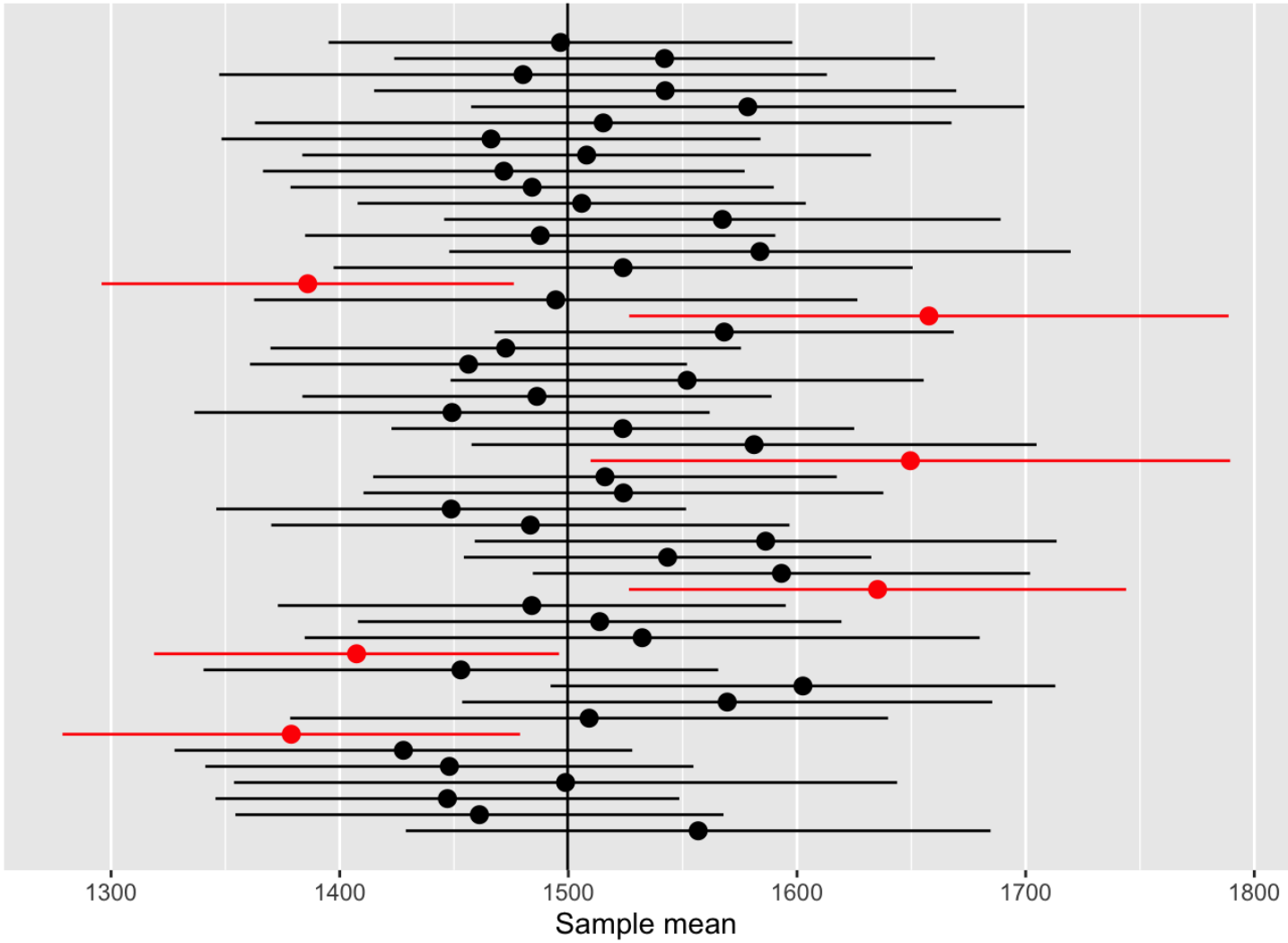


Figure 1: 100 confidence intervals for Ames housing

Now what??

What's next:

We will learn:

- how to construct confidence intervals for all of the statistics whose sampling distributions we studied
  - mean
  - diff in means
  - variance
- how to do the same thing for some **discrete** distributions:
  - sample proportion
  - diff in proporitions

Ultimate goal

- compute all of this by, essentially, one-liners in R and Python
- understand the output, the meaning, and able to communicate
- get a level of confidence yourself; be able to quantify the uncertainty behind the outputs!

License

This document is created for ITMD/ITMS/STAT 514, Spring 2021, at Illinois Tech.