STAT011 Statistical Methods I

Lecture 9 Sampling Distribution

Lu Chen
Swarthmore College
2/19/2019

Review

- Exploratory data analysis of one variable
- Exploratory data analysis of the relationship between two variables
- Association and causation
  - Lurking variable
  - Types of associations: *direct causation, mediation, common response, confounding.*
- Data collection
- Design of experiments: *comparative experiments, matched pairs design, block design*
- Sampling design: *simple random sample, stratified sample, multistage sample*

Outline

- Population, sample, parameter and statistic
- Statistical inference
- Sampling variability
  - Simulation
- Sampling distribution
- Bias and variability
  - Manage bias and variability
- Sampling distribution of a sample mean

Population and sample

Population: The entire group of individuals that we want information about.

Sample: A part of the population that we actually examine in order to gather information.

- Usually, it is difficult to gather information from every single individual in a population (exception: US census).
- However, it is much easier to gather information from part of the population. Most data we study are samples from populations.
- A population can be large or small. Its size depends on the question of interest.

Population and sample

Determine the population and sample of the following examples.

Do Swarthmore students like to study at the library?

- **Population**: all Swarthmore students
- **Sample**: part of the Swarthmore students who were given the survey

How much were the homes sold in Pennsylvania in 2018?

- **Population**: All PA homes sold in 2018
- **Sample**: part of the PA homes sold in 2018 that we study
- **Question**: Is the mean price of the sample PA homes the same as the mean price of all PA homes sold in 2018?
- Population mean: mean price of all PA homes sold in 2018
- Sample mean: mean price of the sample PA homes

Parameter and statistic

A **parameter** is a number that describes the population.

- Fixed
- Unknown (It will be known if population data is available)

A **statistic** is a number that describes a sample.

- Changes from sample to sample
- Used to estimate the unknown population parameter

- Mean price of all PA homes sold in 2018 is a **population parameter**.
- Mean price of the sample PA homes we study is a **sample statistic**.

Parameter and statistic

Suppose Variable X has mean μ and standard deviation σ. The mean and standard deviation of a sample from X are $\bar{x}$ and s.

	Mean	Standard deviation
Population Parameter	μ	σ
Sample Statistic	$\bar{x}$	s

One of major tasks of Statistics is, in fact, to **estimate population parameters using sample statistics**.

- How to obtain a good sample that is representative of the population?
- How do we know that the sample statistics are good estimates of the population parameters (since they change from sample to sample)?

Statistical inference

> **Statistical inference** uses a fact about a sample to estimate the truth about the whole population.

- The process of estimating population mean μ using the sample mean $\bar{x}$ is making inference about the **center** of population from the **center** of sample.
- Similarly, the process of estimating population SD σ using the sample SD s is making inference about the **spread** of population from the **spread** of sample.
- In this course, most of the time, we focus on the estimation of the population mean using sample mean and comparing population means between groups.

Samples are different

- Each sample, if randomly generated from a population, is different from one another.

```
set.seed(9)
# rnorm(): randomly generates data from a Normal distribution
rnorm(n = 5) # Generate 5 values from N(0, 1)
```

```
## [1] -0.7667960 -0.8164583 -0.1415352 -0.2776050  0.4363069
```

```
x1 <- rnorm(n = 5); x1 # Semicolon separates codes that are written in one line
```

```
## [1] -1.18687252  1.19198691 -0.01819034 -0.24808460 -0.36293689
```

```
x2 <- rnorm(n = 5); x2
```

```
## [1]  1.2775705 -0.4688971  0.0710541 -0.2660384  1.8452572
```

```
x3 <- rnorm(n = 5); x3
```

```
## [1] -0.83944966 -0.07744806 -2.61770553  0.88788403 -0.70749145
```

Sample statistics are different

- The value of a statistic calculated from each sample is also different from one another.

```
mean(x1)
```

```
## [1] -0.1248195
```

```
mean(x2)
```

```
## [1] 0.4917892
```

```
mean(x3)
```

```
## [1] -0.6708421
```

Sampling variability

> **Sampling variability**: The value of a statistic varies in repeated random sampling.

- In this example, $X \sim N(0, 1)$, population mean $\mu = 0$.
- We generated three samples X_1 , X_2 and X_3 from $X \sim N(0, 1)$.
- The sample mean for the first one is $\bar{x}_1 = -0.12$.
- The sample mean for the second one is $\bar{x}_2 = 0.49$.
- The sample mean for the third one is $\bar{x}_3 = -0.67$.
- The sample means are **different from one another**.
- In addition, the sample means are all **different from the population mean**.

Sampling variability

$$\mu = 0, \bar{x}_1 = -0.12, \bar{x}_2 = 0.49, \bar{x}_3 = -0.67$$

- Which sample is better in the sense of estimating the population mean?
- In practice, we do not know the value of the population mean and only have one sample, how do we know the sample is a good sample representing the population? How do we know the sample mean is a good estimate of the population mean?
- We will never know. But we can evaluate the sampling procedure (the way we generate the sample).
- If it is a good sampling procedure, we believe most of the times it will generate a good sample.

Sampling variability

> **All of statistical inference is based on one idea**: to see how trustworthy a procedure is, ask what would happen if we repeated it many times.

- This can hardly be done in real world. Therefore, we do it using **simulation**.
- The sampling procedure we evaluate here is the most common one: **simple random sampling** (each of the n individuals in the sample has an equal chance to be chosen from the population).

Simulation

> **Simulation**: We imitate taking many samples by using computer software or other tools to emulate chance behavior.

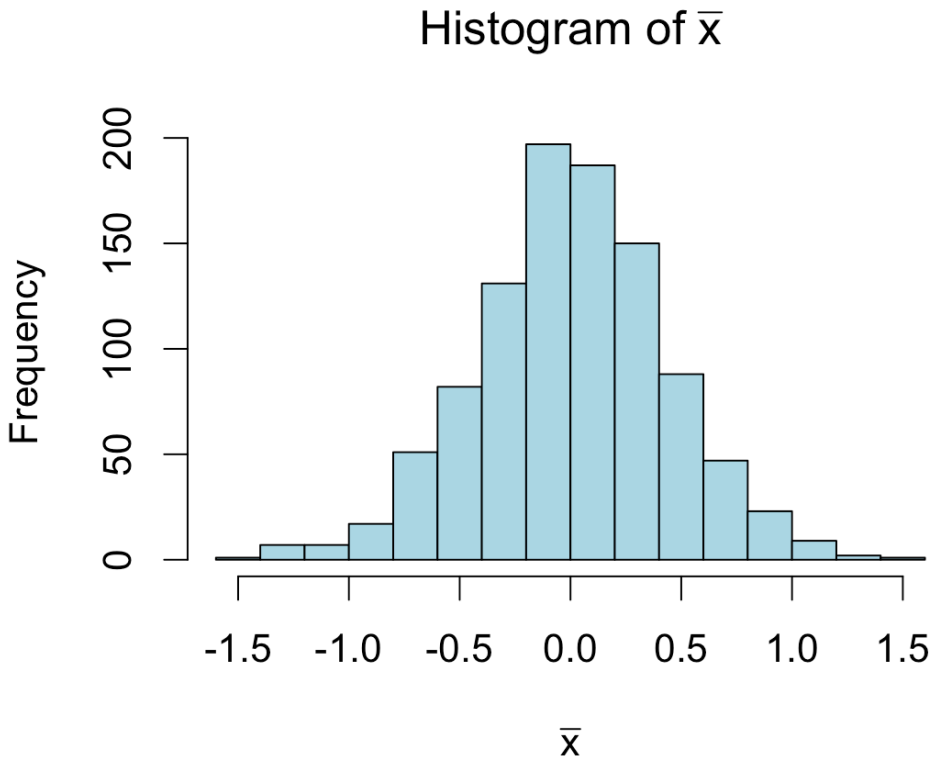
1. Determine the distribution of population — e.g., $X \sim N(0, 1)$.
2. Take an SRS of size n from the population — e.g., 5 values from $N(0, 1)$.
3. Calculate the statistic based on the sample — e.g., mean of the 5 values.
4. Repeat 2. and 3. many times — e.g., 1000 times.
5. Evaluate the distribution of these many statistics — e.g., distribution of 1000 means calculated from the 1000 samples.

Simulation

```r
mean_x <- NULL # Define an object named mean_x and assign nothing to it
set.seed(9)
# This a "for" loop, where i takes values from 1 to 1000
for(i in 1:1000){
  mean_x[i] <- mean(rnorm(5)) # Calculate the mean of each sample and assign
                              # it to the i'th element of mean_x,
} # At the end, mean_x will contain 1000 means from the 1000 samples
head(mean_x) # Look at the first six values of mean_x
mean(mean_x) # Calculate the mean of the 1000 means
sd(mean_x) # Calculate the SD of the 1000 means
hist(mean_x, col="lightblue") # Plot the distribution of the 1000 means
```

- All the 1000 samples have size 5.
- All the 1000 samples are generated from the same population distribution $N(0, 1)$.
- The same statistic (mean) is calculated for all the 1000 samples.

Simulation




```
head(mean_x)
```

```
## [1] -0.3132175 -0.1248195  0.49178
```

```
mean(mean_x)
```

```
## [1] 0.01022393
```

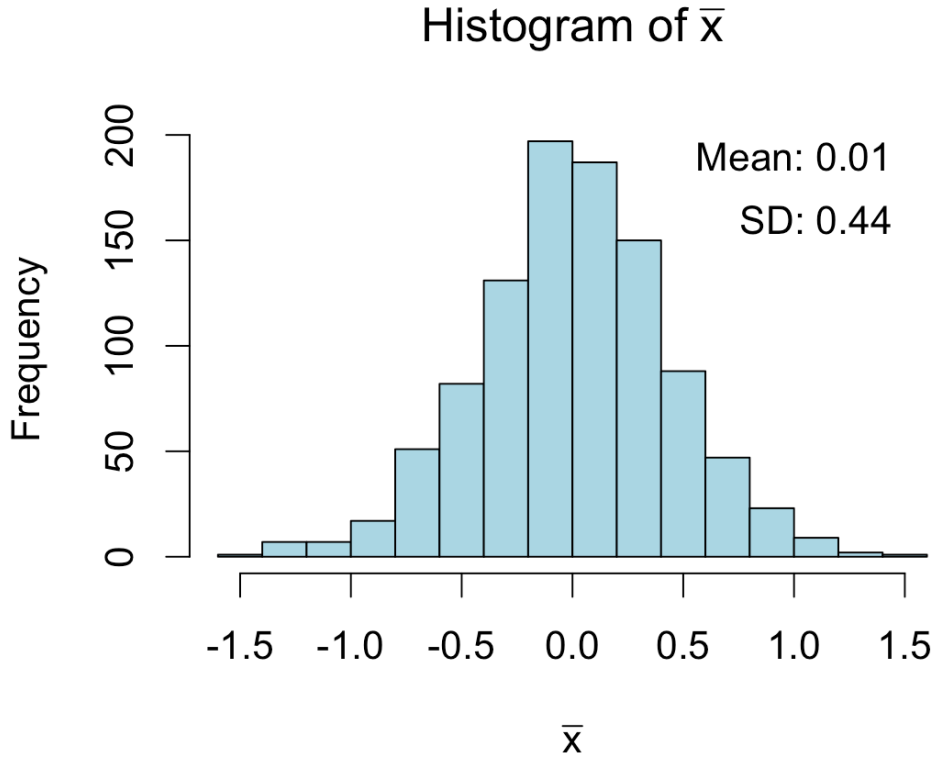
```
sd(mean_x)
```

```
## [1] 0.4356788
```

- The *shape* of the distribution looks **Normal**
- The *mean* of the 1000 sample means: **0.01**
- The *SD* of the 1000 sample means: **0.44**

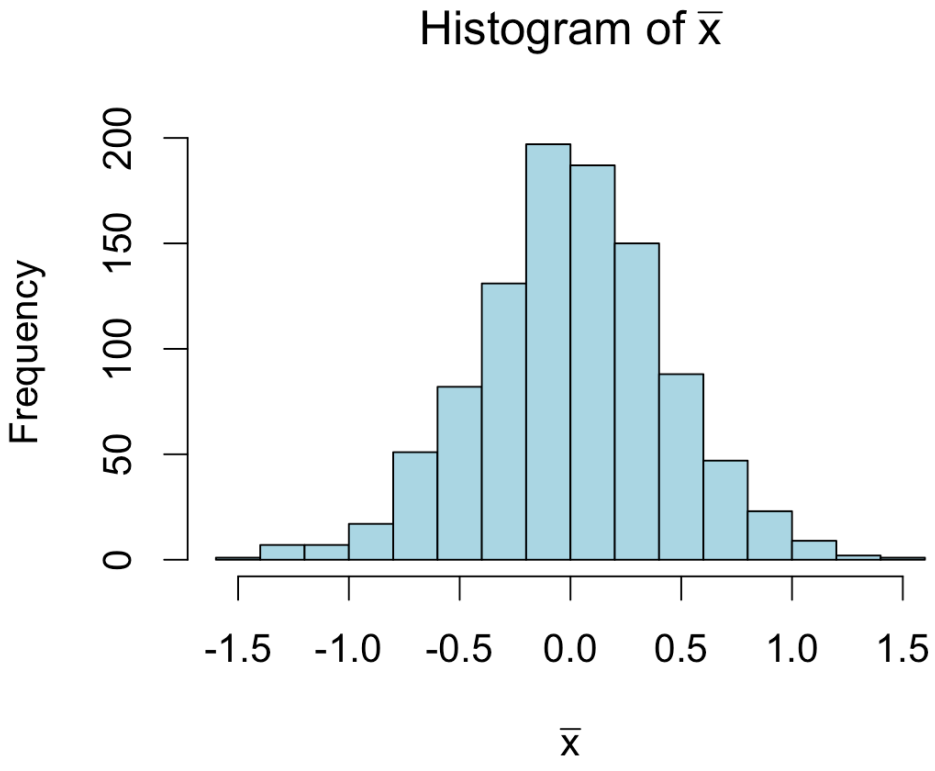
Sampling distribution

> The **sampling distribution** of **a statistic** is the distribution of values taken by the statistic in *all possible samples* of **the same size** from **the same population**.



- This is a **sampling distribution** of the **sample mean**.
- All the 1000 samples have the **same size** 5.
- All the 1000 samples are generated from the **same population distribution** $N(0, 1)$.

Sampling distribution



$$X \sim N(0, 1)$$

- Population mean: 0
- Mean of sample mean: 0.01
- Population SD: 1
- SD of sample mean: 0.44
- Is this sampling procedure (SRS) a good one?

Bias and variability

To evaluate whehter the sampling procedure works well or not, we use **bias** and **variability** to describe the sampling distribution of the sample mean.

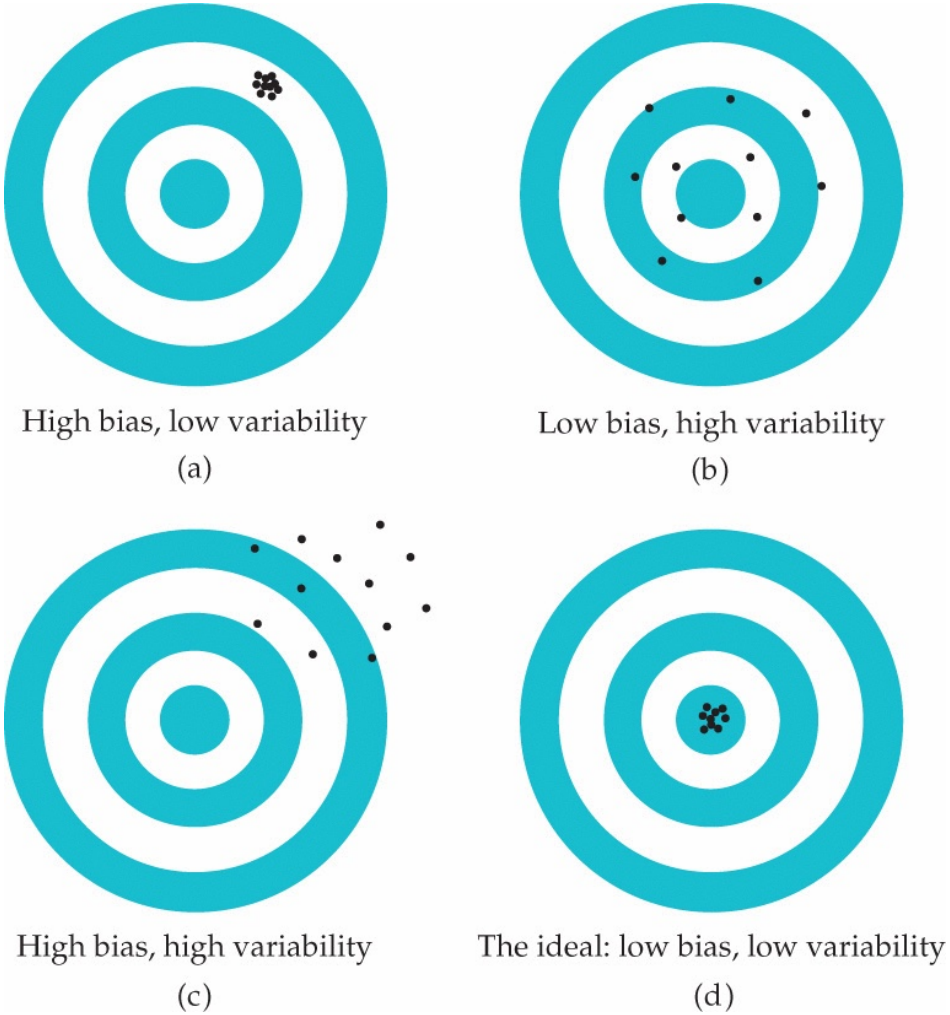
Bias concerns the center of the sampling distribution.

$$\textbf{Bias} = \textbf{Mean of the statistic} - \textbf{Population parameter}$$

A statistic used to estimate a parameter is an **unbiased estimator** if its mean is equal to the true value of the parameter being estimated.

The **variability** of a statistic is described by the spread of its sampling distribution. This spread is determined by the sampling design and the sample size n. Statistics from larger samples have smaller spreads.

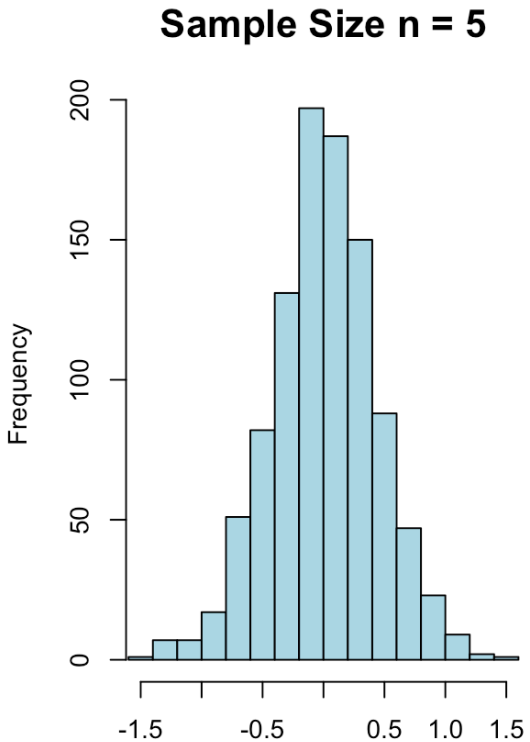
Bias and variability



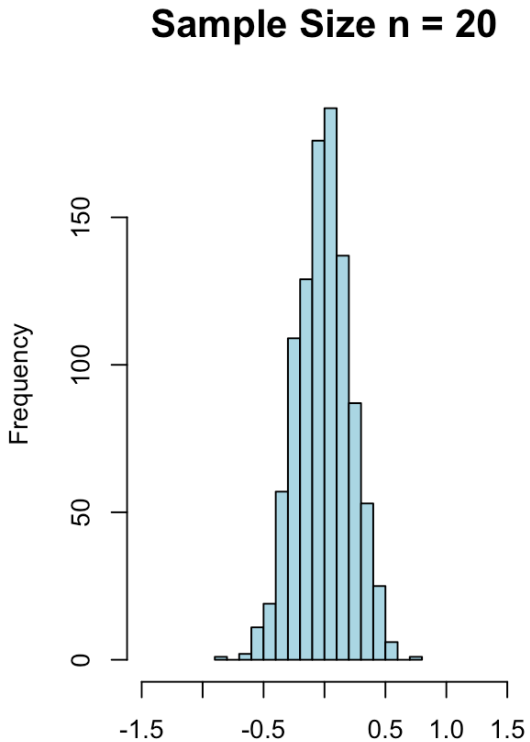
- **Bias** concerns the **center**. It is about **accuracy** of the estimation.
- **Variability** concerns the **spread**. It is about the **precision** and **consistency** of the estimation.
- The ideal case: low/no bias, low variability. In terms of the sampling distribution of the sample mean, the mean of the sampling distribution should be as **close** to population mean as possible; while the SD of the sampling distribution should be as **small** as possible.

Bias and variability

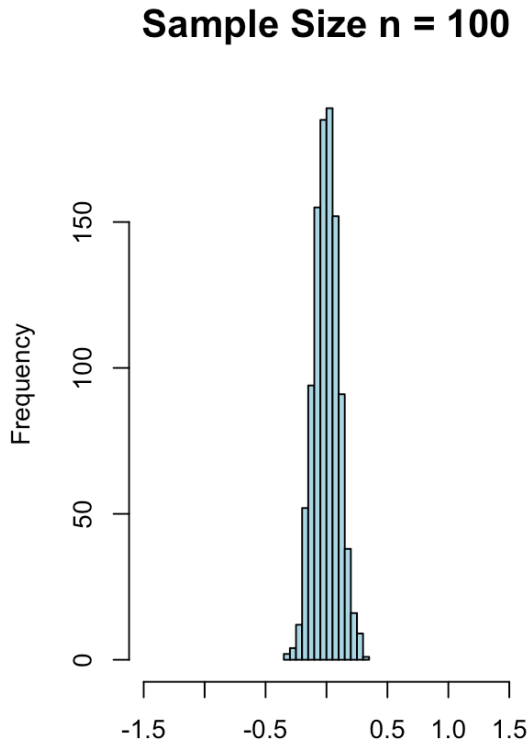
Simulation: 1000 samples



Mean: 0.010; SD: 0.436



Mean: -0.005; SD: 0.219



Mean: -0.001; SD: 0.101

Bias and variability - Simulation: 1000 samples

Sample size	Mean of mean $\bar{x}$	SD of mean $\bar{x}$
5	0.010	0.436
10	0.006	0.314
20	-0.005	0.219
50	-0.0017	0.144
80	-0.0018	0.114
100	-0.0006	0.101

- The mean of $\bar{x}$ is always close to the population mean (0) regardless of the sample size. As sample size increases, the mean of $\bar{x}$ gets closer and closer (but not always) to the population mean (this is called **Law of Large Numbers**).
- The SD of $\bar{x}$ is consistently getting smaller as sample size increases.

Manage bias and variability

For SRS,

- The mean of the statistic is always close to the population parameter. We reasonally infer that SRS is an unbiased sampling procedure.
- Larger sample size will always result in smaller spread of the statistic.

Therefore,

> To **reduce bias**, use random sampling. When we start with the entire population, simple random sampling produces unbiased estimates – the values of a statistic computed from an SRS neither consistently overestimate nor consistently underestimate the value of the population parameter.
>
> To **reduce the variability** of a statistic from an SRS, use a larger sample. You can make the variability as small as you want by taking a large enough sample.

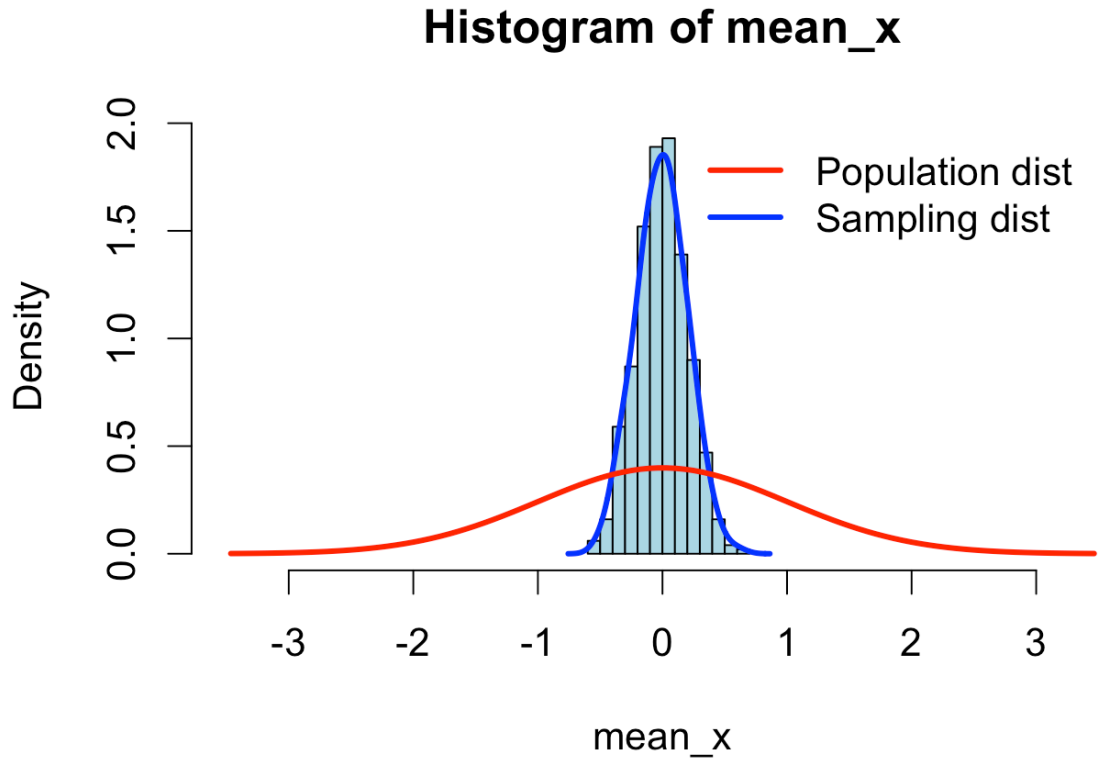
Sampling distribution of a sample mean

> The **population distribution** of **a variable** is the distribution of its values for all members of the population.
>
> The **sampling distribution** of **a statistic** is the distribution of values taken by the statistic in all possible samples of **the same size** from **the same population**.

- Distribution of X is a population distribution.
- Distribution of $\bar{x}$ is a sampling distribution, where $\bar{x}$ is the sample mean.

Sampling distribution of a sample mean

$X \sim N(0, 1)$. The following is the distribution of $\bar{x}$ for 1000 simulated samples with sample size $n = 25$.



- Mean of $\bar{x}$: -0.01
- SD of $\bar{x}$: 0.20.
- What is the distribution of $\bar{x}$?
- What is the relationship between the SD of X ($\sigma = 1$) and the SD of $\bar{x}$ (0.20)?

Sampling distribution of a sample mean

Let $\bar{x}$ be the mean of an SRS of size n from a population having Normal distribution with mean μ and standard deviation σ. The mean and standard deviation of $\bar{x}$ are

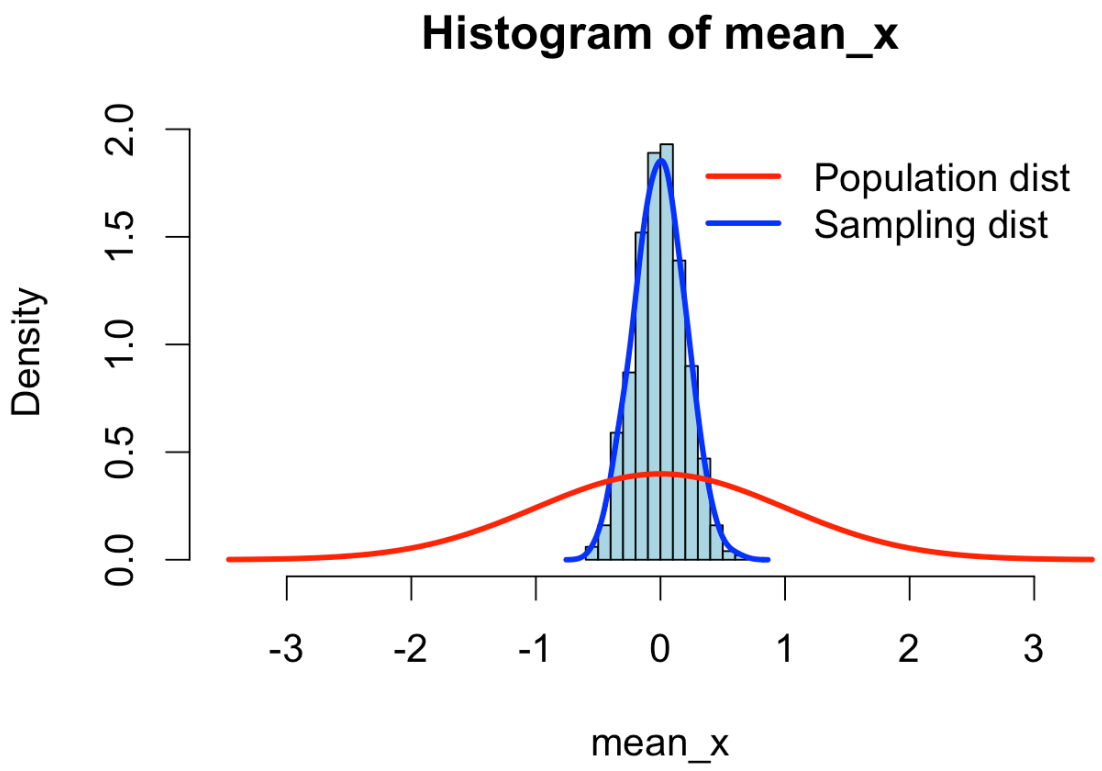
$$\mu_{\bar{x}} = \mu,$$
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}.$$

And $\bar{x}$ has the $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ distribution.

This says that if $X \sim N(\mu, \sigma)$, then

$$\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Sampling distribution of a sample mean



- Population distribution: $X \sim N(0, 1)$
- Sampling distribution of $\bar{x}$ (by formula):
  $\bar{x} \sim N\left(0, \frac{1}{\sqrt{25}}\right) = N(0, 0.2)$
  - 68% of $\bar{x}$ fall between [-0.2, 0.2]
  - 95% of $\bar{x}$ fall between [-0.4, 0.4]
  - 99.7% of $\bar{x}$ fall between [-0.6, 0.6]

Summary

- Population, sample, parameter and statistic
- Statistical inference
- Sampling variability
  - Simulation
- Sampling distribution
- Bias and variability
  - Manage bias and variability
- Sampling distribution of a sample mean