Class 24

DATA1220-55, Fall 2024

Sarah E. Grabinski

2024-10-30

Population Parameters versus Sample Statistics

Table 1: Sample statistics are used to estimate unknowable population parameters

Measure	Sample Statistic	Population Parameter
Mean	$\bar{x}$	μ
Proportion	$\hat{p}$	p
Difference in Means	$\bar{x}_1 - \bar{x}_2$	$\mu_1 - \mu_2$
Difference in Proportions	$\hat{p}_1 - \hat{p}_2$	$p_1 - p_2$
Standard Deviation	s	σ

Assumptions

- Data is ***reliable*** and ***valid***
  - $\bar{x}_{\text{observed}} \approx \bar{x}_{\text{expected}}$ or $\hat{p}_{\text{observed}} \approx \hat{p}_{\text{expected}}$ when data is ***reliable***
  - $\bar{x}_{\text{observed}} \approx \mu$ or $\hat{p}_{\text{observed}} \approx p$ when data is ***valid***

Assumptions

- Data is ***reliable*** and ***valid***
  - $\bar{x}_{\text{observed}} \approx \bar{x}_{\text{expected}}$ or $\hat{p}_{\text{observed}} \approx \hat{p}_{\text{expected}}$ when data is ***reliable***
  - $\bar{x}_{\text{observed}} \approx \mu$ or $\hat{p}_{\text{observed}} \approx p$ when data is ***valid***
- Observations are independent and identically distributed

Assumptions

- Data is ***reliable*** and ***valid***
  - $\bar{x}_{\text{observed}} \approx \bar{x}_{\text{expected}}$ or $\hat{p}_{\text{observed}} \approx \hat{p}_{\text{expected}}$ when data is ***reliable***
  - $\bar{x}_{\text{observed}} \approx \mu$ or $\hat{p}_{\text{observed}} \approx p$ when data is ***valid***
- Observations are independent and identically distributed
- Sufficient sample size
  - $n \geq 30$ for $\bar{x}$ (means)
  - $n \geq 20$, $n_{x=1} \geq 10$, & $n_{x=0} \geq 10$ for $\hat{p}$ (proportions)

Assumptions

- Data is ***reliable*** and ***valid***
  - $\bar{x}_{\text{observed}} \approx \bar{x}_{\text{expected}}$ or $\hat{p}_{\text{observed}} \approx \hat{p}_{\text{expected}}$ when data is ***reliable***
  - $\bar{x}_{\text{observed}} \approx \mu$ or $\hat{p}_{\text{observed}} \approx p$ when data is ***valid***
- Observations are independent and identically distributed
- Sufficient sample size
  - $n \geq 30$ for $\bar{x}$ (means)
  - $n \geq 20$, $n_{x=1} \geq 10$, & $n_{x=0} \geq 10$ for $\hat{p}$ (proportions)
- For means, the observed distribution in the sample approximates a normal distribution (less strict as $n \to \infty$)

The Central Limit Theorem

The distribution of the sample statistic $\bar{x}$ or $\hat{p}$ approximates the normal distribution $N\,(\text{population parameter}, \text{standard error})$ as $n \to \infty$.

The Central Limit Theorem

The distribution of the sample statistic $\bar{x}$ or $\hat{p}$ approximates the normal distribution $N\,(\text{population parameter}, \text{standard error})$ as $n \to \infty$.

- $\bar{x} \sim N\,(\mu, \frac{\sigma}{n}))$
- $\hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$

The Central Limit Theorem

The distribution of the sample statistic $\bar{x}$ or $\hat{p}$ approximates the normal distribution $N\left(\text{population parameter, standard error}\right)$ as $n \to \infty$.

- $\bar{x} \sim N\left(\mu, \frac{\sigma}{n}\right))$
- $\hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$

The ***sampling distribution*** is normal with $\mu =$ sample statistic and $\sigma = \text{standard error}$.

Standard Error of Sample Mean $\bar{x}$

The standard deviation of the sampling distribution of $\bar{x}$ is the population standard deviation σ divided by the square root of the size of the sample n.

$$SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Standard Error of Sample Mean $\bar{x}$

The standard deviation of the sampling distribution of $\bar{x}$ is the population standard deviation σ divided by the square root of the size of the sample n.

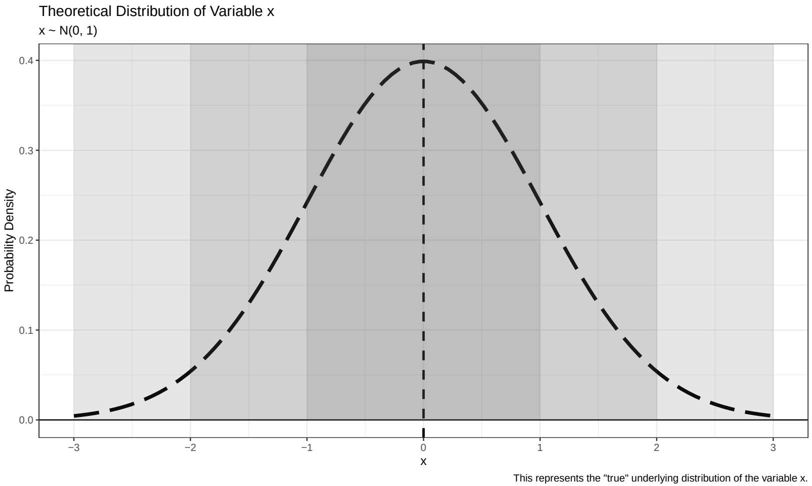
$$SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Because we don't have access to the "true" value of σ, we substitute the observed standard deviation in the sample s for inference and hypothesis testing.

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$

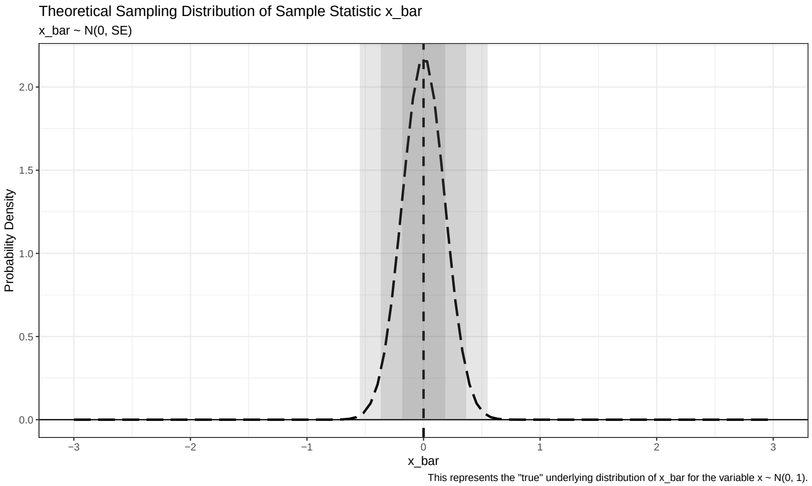
Population Distribution of x

The population in this figure has the "true" parameters of mean $\mu = 0$ and standard deviation $\sigma = 1$.

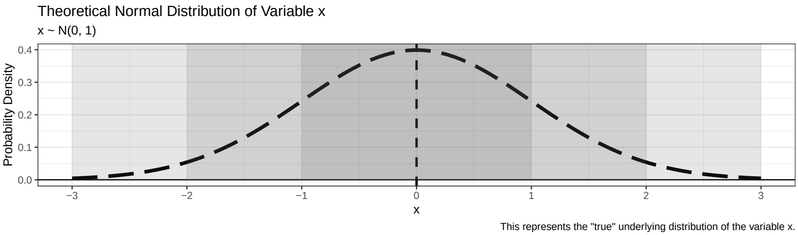


Sampling Distribution for $\bar{x}$

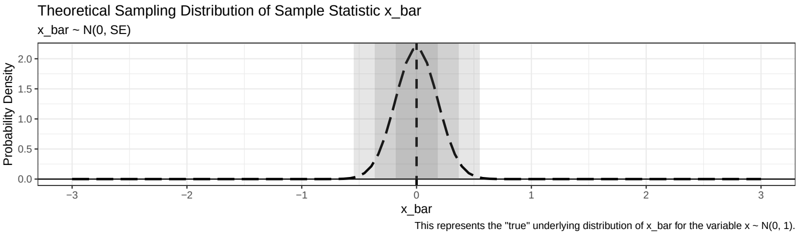
The ***sampling distribution*** is the distribution of sample statistics $\bar{x}$ from samples with size n taken from the population $x \sim N(0, 1)$, were you to sample infinite times.



Side-By-Side



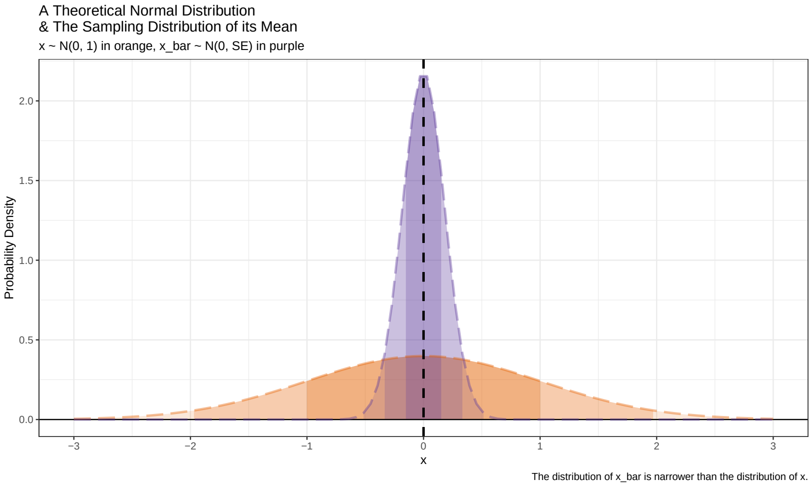
Theoretical Normal Distribution of Variable x
x ~ N(0, 1)
Probability Density
0.4
0.3
0.2
0.1
0.0
−3
−2
−1
0
1
2
3
x
This represents the "true" underlying distribution of the variable x.



Theoretical Sampling Distribution of Sample Statistic x_bar
x_bar ~ N(0, SE)
Probability Density
2.0
1.5
1.0
0.5
0.0
−3
−2
−1
0
1
2
3
x_bar
This represents the "true" underlying distribution of x_bar for the variable x ~ N(0, 1).

Distribution of $\bar{x}$ versus x

Observed values of x are more variable than observed values of $\bar{x}$.



Example

- The "true" distribution in your population is normal with mean $\mu = 0$ and standard deviation $\sigma = 1$ ($x \sim N(0, 1)$)

Example

- The "true" distribution in your population is normal with mean $\mu = 0$ and standard deviation $\sigma = 1$ ($x \sim N(0, 1)$)
- Take repeated samples of $n = 50$ from the population $x \sim N(0, 1)$.

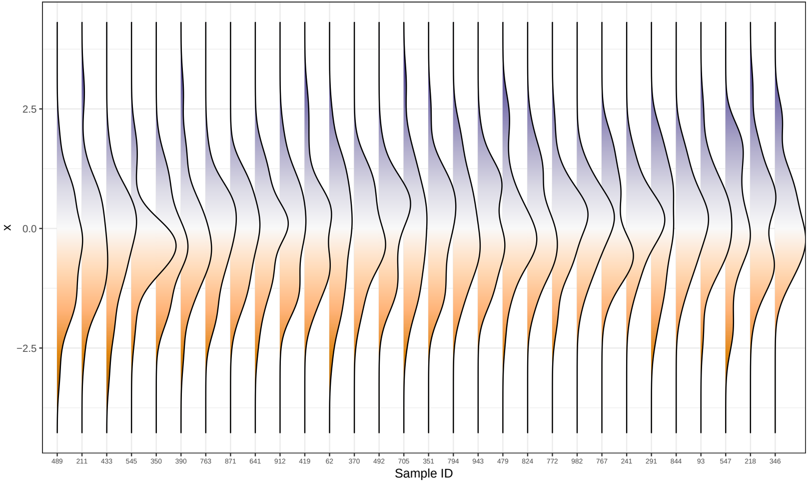
Example

- The "true" distribution in your population is normal with mean $\mu = 0$ and standard deviation $\sigma = 1$ ($x \sim N(0, 1)$)
- Take repeated samples of $n = 50$ from the population $x \sim N(0, 1)$.
- Calculate $\bar{x}_i$ for each sample of size n.

Example

- The “true” distribution in your population is normal with mean $\mu = 0$ and standard deviation $\sigma = 1$ ($x \sim N(0,1)$)
- Take repeated samples of $n = 50$ from the population $x \sim N(0,1)$.
- Calculate $\bar{x}_i$ for each sample of size n.
- Compare the observed distribution of $\bar{x}_i$ to the expected distribution $\bar{x} \sim N\left(0, \frac{1}{\sqrt{n}}\right)$.

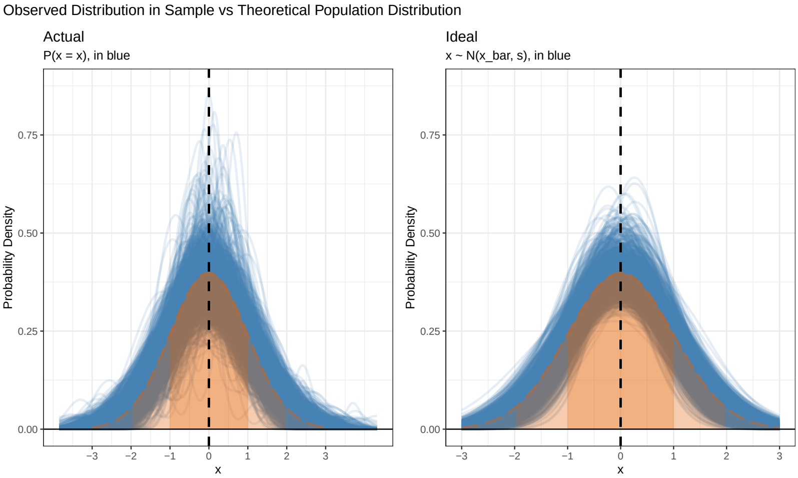
Observed Distributions of x in Samples



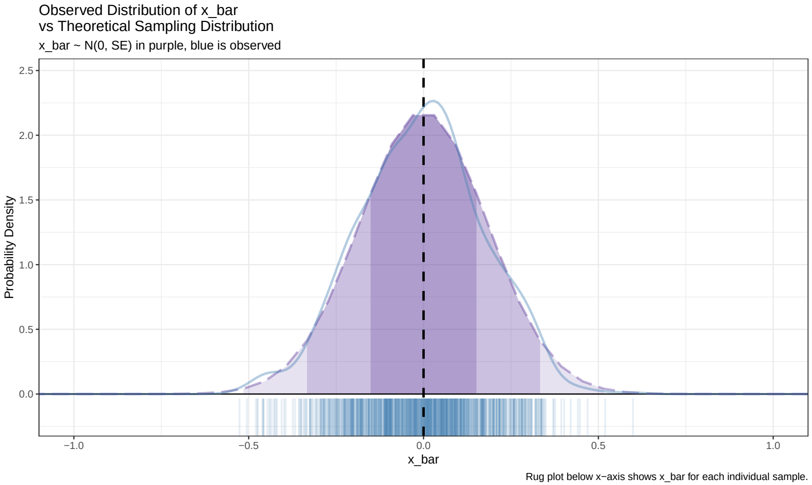
Observed Sample Means

Sample ID	x_bar	s
1	-0.047	0.981
2	0.178	0.835
3	0.024	0.870
4	-0.094	0.907
5	-0.184	1.148
6	0.154	0.942
7	0.015	0.967

Observed Distributions vs Expected Distribution



Observed Distributions vs Expected Distribution



Standard Error of Sample Proportion $\hat{p}$

The standard deviation of the sampling distribution of $\hat{p}$ for sample size n is...

$$SE_{\bar{x}} = \sqrt{\frac{p(1-p)}{n}}$$

Standard Error of Sample Proportion $\hat{p}$

The standard deviation of the sampling distribution of $\hat{p}$ for sample size n is...

$$SE_{\bar{x}} = \sqrt{\frac{p(1-p)}{n}}$$

Because we don't have access to the "true" value of p, we substitute the observed statistic in the sample $\hat{p}$ for inference and hypothesis testing.

$$SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Example

- The “true” distribution in your population is $p = 0.5$

Example

- The "true" distribution in your population is $p = 0.5$
- Take repeated samples of $n = 50$ from the population $p = 0.5$

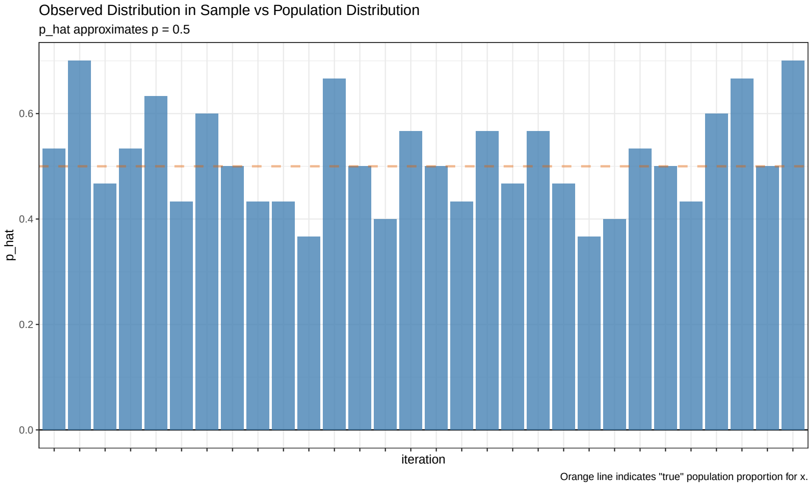
Example

- The "true" distribution in your population is $p = 0.5$
- Take repeated samples of $n = 50$ from the population $p = 0.5$
- Calculate $\hat{p}_i$ for each sample of size n.

Example

- The "true" distribution in your population is $p = 0.5$
- Take repeated samples of $n = 50$ from the population $p = 0.5$
- Calculate $\hat{p}_i$ for each sample of size n.
- Compare the observed distribution of $\hat{p}_i$ to the expected distribution $\hat{p} \sim N\left(0.5, \sqrt{\frac{0.5(1-0.5)}{n}}\right)$.

Observed Distributions of $\hat{p}$ in Samples



Observed Sample Proportions

Sample ID	p_hat	SE
1	0.533	0.091
2	0.667	0.086
3	0.400	0.089
4	0.500	0.091
5	0.433	0.090
6	0.600	0.089
7	0.667	0.086

Observed Distributions vs Expected Distribution