Agenda

- Sample
- Point Estimation
- Sampling Distribution
- Interval estimation
- Hypothesis tests
- Practical Significance

Why do we need sampling?

Census → Collection of data from the entire population

- expensive
- unnecessary
- time consuming

Statistical Inference → Use sample data to draw conclusion about the population

Sampling

Parameter → A measurable factor that defines a characteristic of a population, process, or system

"Random sampling" is the method to collect a sample (n) **representative of the population (N).**

simple random sample - each possible sample of size *n* has the same probability of being selected. (*finite population*)

random sample - every element has the same probability of being selected in the sample. (in*finite population*)

Other methods:

- Available sample
- Volunteer sample
- Quota sample
- Referral sample
- Stratified sample

Point Estimation

Numerical value obtained for $\bar{x}$, s and $\bar{p}$ is called "point estimate"

When the expected value of a point estimator equals the population parameter, we say the point estimator is **unbiased.**

Sample mean $\bar{x} = \frac{\Sigma x_i}{n}$

Sample standard deviation $s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}}$

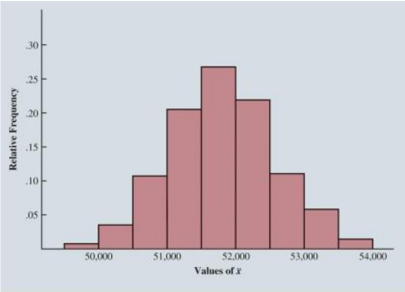
Sample proportion $\bar{p} = \frac{x}{n}$

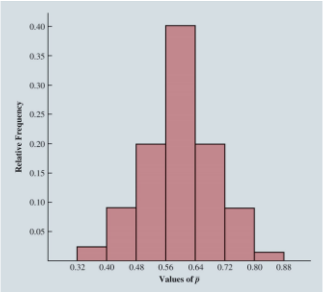
Population Parameter	Parameter Value	Point Estimator	Point Estimate
μ = Population mean annual salary	\$51,800	$\bar{x}$ = Sample mean annual salary	\$51,814
σ = Population standard deviation for annual salary	\$4,000	s = Sample standard deviation for annual salary	\$3,348
p = Population proportion completing the management training program	0.60	$\bar{p}$ = Sample proportion having completed the management training program	0.63

Sampling Distribution

Knowledge of the sample distribution and its properties enables us to make probability statements about how close the sample mean, $\overline{x}$ Is to the population mean μ.

Sample Number	Sample Mean ($\bar{x}$)	Sample Proportion ($\bar{p}$)
1	51,814	0.63
2	52,670	0.70
3	51,780	0.67
4	51,588	0.53
.	.	.
.	.	.
.	.	.
500	51,752	0.50





Sampling distribution – animation

Sampling Distribution

EXPECTED VALUE OF $\bar{x}$

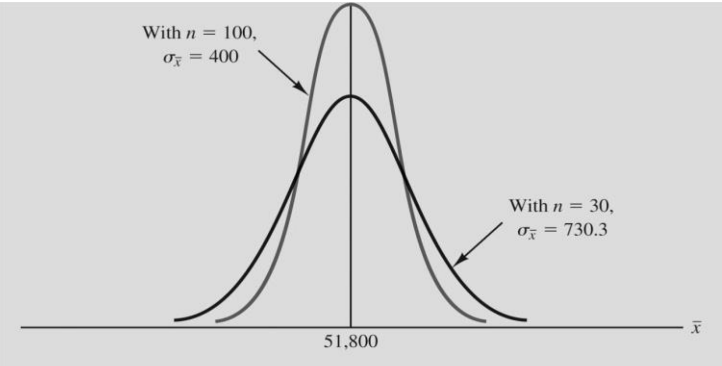
$$E(\bar{x}) = \mu \tag{6.1}$$

where

$E(\bar{x})$ = the expected value of $\bar{x}$
μ = the population mean

ESTIMATED STANDARD DEVIATION OF $\bar{x}$

Finite Population	*Infinite Population*	
(n/N > 0.05) $s_{\bar{x}} = \sqrt{\frac{N-n}{N-1}}\left(\frac{s}{\sqrt{n}}\right)$	$s_{\bar{x}} = \left(\frac{s}{\sqrt{n}}\right)$	(6.3)



Sampling Distribution - proportion

Sample proportion is the point estimator for the population proportion (p) $\cdots\cdots\blacktriangleright \bar{p} = \dfrac{x}{n}$

EXPECTED VALUE OF $\bar{p}$

$$E(\bar{p}) = p \qquad \textbf{(6.4)}$$

where

$E(\bar{p})$ = the expected value of $\bar{p}$
p = the population proportion

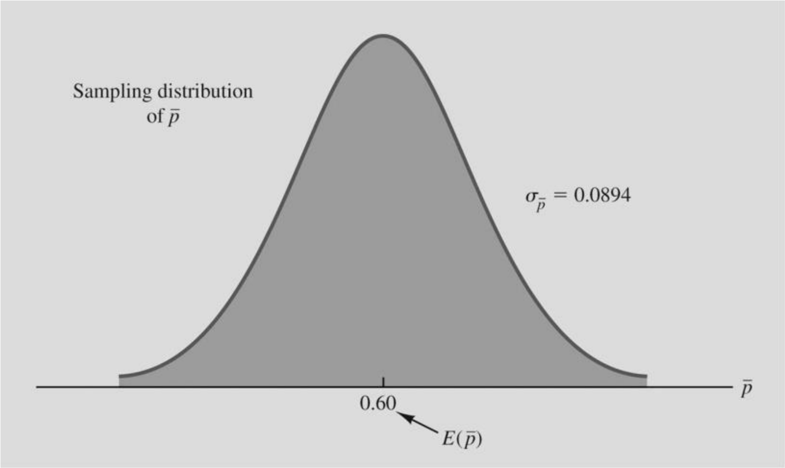
ESTIMATED STANDARD DEVIATION OF $\bar{p}$

Finite Population

$$s_{\bar{p}} = \sqrt{\frac{N-n}{N-1}}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

Infinite Population

$$s_{\bar{p}} = \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \qquad \textbf{(6.6)}$$



Hypothesis

Sample → **Conclusion**

Hypothesis - States the relationship between the variables involved or a phenomenon.

Testable, **defined terms** and **does not have to be correct**

Good or Bad hypothesis?

- Sales of Ford automobiles in America would be higher if Lexus did not exist.
- Students who do not have smartphones tend to have better grades.
- Clocks run clockwise because most people are right-handed.

Null/Alternative Hypothesis

Null Hypothesis - hypothesis of no difference

Difference is due to **random error**

Alternative Hypothesis - Opposite of what is stated in the hypothesis

$H_0: \mu \geq \mu_0$

$H_a: \mu < \mu_0$

$H_0: \mu \leq \mu_0$

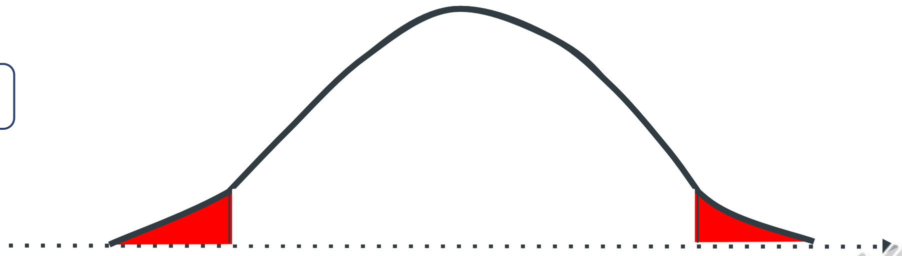
$H_a: \mu > \mu_0$

One-tailed

$H_0: \mu = \mu_0$

$H_a: \mu \neq \mu_0$

two-tailed



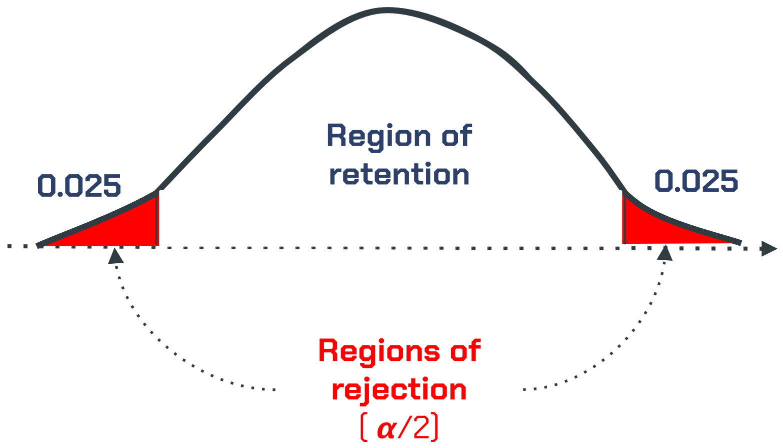
Significance level & Regions of rejection

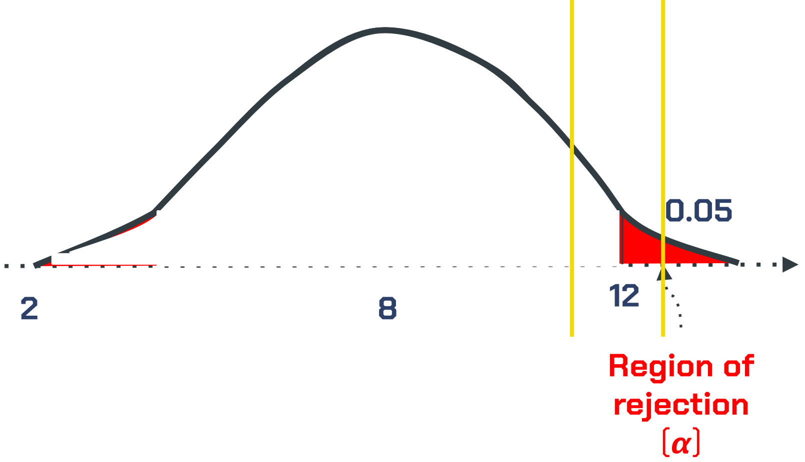
Significance level (α) - is the probability of rejecting the null hypothesis.

[0.01, 0.05]

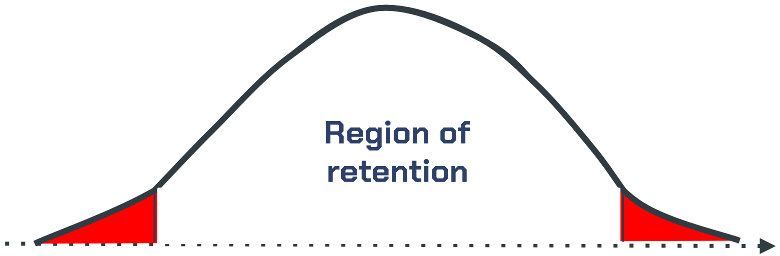
α = 0.05

Reject H_0 if p value $\leq \alpha$





Errors



	Reject H_0	Accept H_0
H_0 is true	Type I error	Correct
H_0 is false	Correct	Type II error

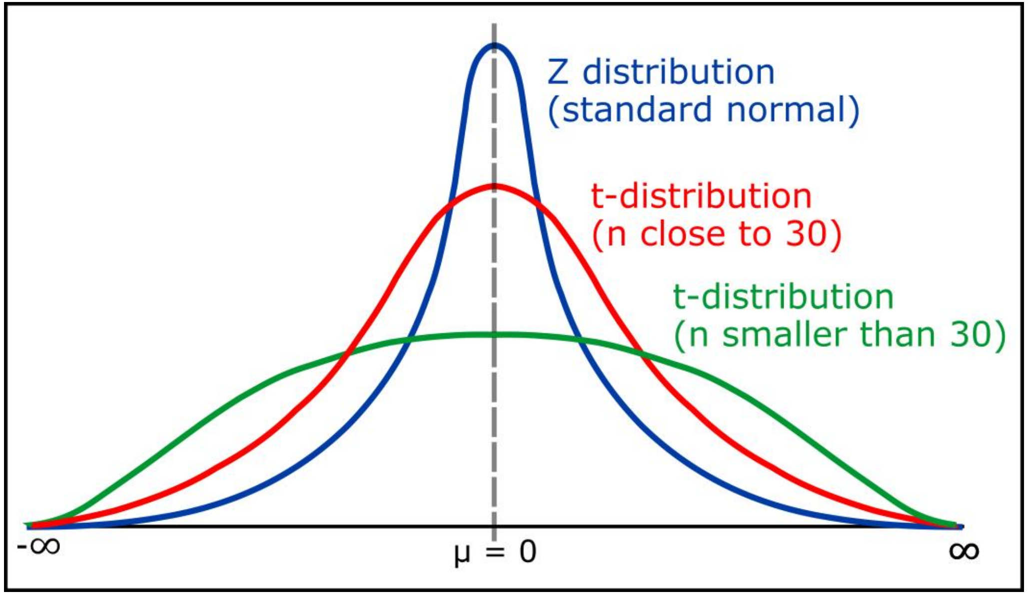
$$H_0: \mu = \mu_0$$
$$H_a: \mu \neq \mu_0$$

- Type I:
  - *Rejection of H_0,* which should be accepted
  - Decrease α, increase confidence level $(1 - \alpha)$
- Type II:
  - *Accepting H_0,* which should be rejected
  - Increase sample size

Hypothesis tests - Summary

Number of groups	*Quantitative*	*Nominal*
1 group	**T-test, z-test**	**chi-square test**
2 independent groups	**Independent samples t-test**	**chi-square test**
2 dependent groups	**Paired t-test**	**McNemar test**
>2 independent groups	**ANOVA**	**chi-square test**

Z-distribution vs t-distribution



https://www.geeksforgeeks.org/students-t-distribution-in-statistics/

Degrees of freedom

Number of independent values needed for calculation.

- Qualitative variable = k - 1
- Quantitative variable = n - 1 (if mean is known)

https://www.geeksforgeeks.org/students-t-distribution-in-statistics/

Single t-test, z-test

- Comparing sample with a population
  - Small sample size (<30), unknown variance → t-test
    - t-statistic, df, level of significance --> p-value
  - Large sample size (>30) → z-test
    - Z-statistic, level of significance --> p-value

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

$$z = (\bar{x} - \mu)/\sigma$$

Independent t-test

Comparing 2 independent samples

- One member cannot be part of both groups.
- Small sample size (<30), unknown variance
- t-statistic, df, significance

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} \qquad s_p = \frac{s_1^2(df_1) + s_2^2(df_2)}{df_1 + df_2}$$

Paired t-test

Comparing 2 dependent samples

- 2 set of observation from the same group.
- t-statistic, df, significance

$$t = \frac{\bar{x}_d}{S_d / \sqrt{n}}$$

Statistical tests – qualitative variables

- Chi-square test
  - Compare observed values to expected values.
  - Degrees of freedom – (m-1) × (n-1)

$$\chi_c^2 = \frac{(O_i - E_i)^2}{E_i}$$

- Mcnemer test
  - Paired nominal data.
  - 2 categories (yes/no). How many switched?

$$\chi^2 = \frac{(b - c)^2}{b + c}$$

Interval Estimation

- Because a point estimator cannot be expected to provide the exact value of a population parameter, interval estimation is frequently used to generate an estimate of the value of a population parameter.

- The general form of an interval estimate is:

Point estimate $\pm$ Margin of error

$\bar{x} \pm$ Margin of error

$\bar{p} \pm$ Margin of error

INTERVAL ESTIMATE OF A POPULATION MEAN

$$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}, \qquad \textbf{(6.7)}$$

where s is the sample standard deviation, α is the level of significance, and $t_{\alpha/2}$ is the t value providing an area of $\alpha/2$ in the upper tail of the t distribution with $n - 1$ degrees of freedom.

Summary

Point Estimation

Sampling Distribution

Interval estimation

Hypothesis tests