Inferential Statistics II

Jafar Namdar
Postdoctoral Associate
MIT SCM & MIT Digital Supply Chain Transformation

Review

- Assignments are graded.

Two Sessions

SCM.250: Session I. Statistical Analysis
August 25th
8:30AM-11:30AM

Collect → Organize → Summarize → Present

SCM.250: Session II. Inferential Statistics
Today
1:00PM-4:00PM

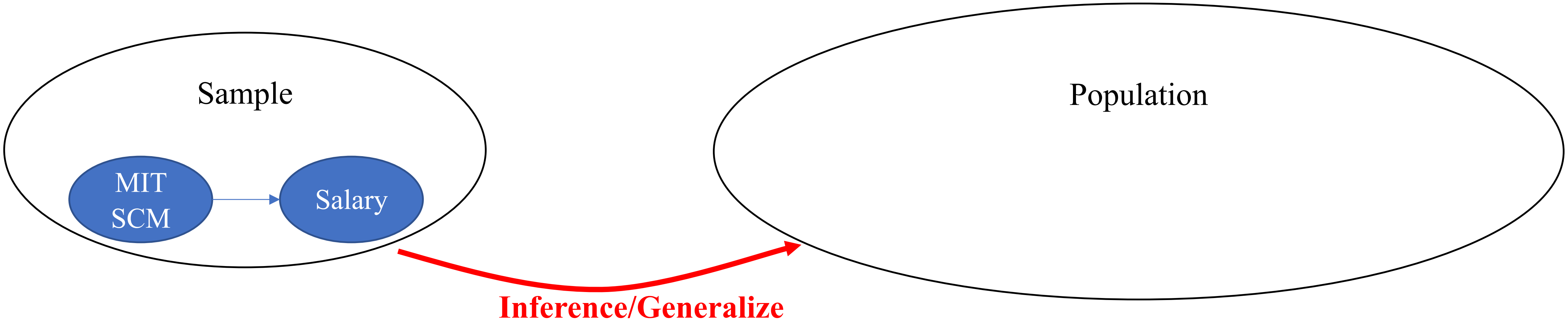
Transform the collected data into knowledge → Hypotheses testing → Make prediction → Generalize the findings from sample to population

Outline

- Population vs Sample
- Sampling Distribution
- Hypothesis Testing
- Multiple Linear Regression
- ANOVA (if time permits)

Inferential Statistics

- Inferential statistics consists of a set of procedures that enables us to draw conclusions about the characteristics of a whole population by studying the properties of a sample of a population.



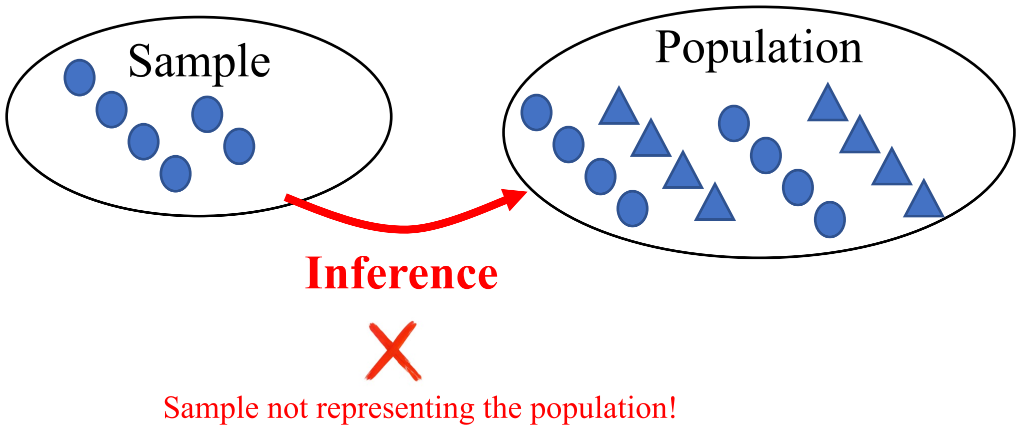
Theory of Estimation (Estimate Procedure)

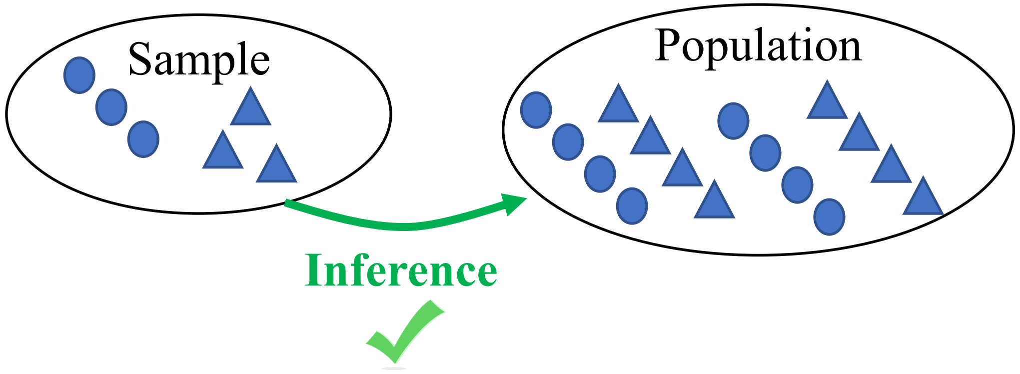
1. Estimation of some characteristics of the entire population
2. Point estimate
3. Confidence Intervals

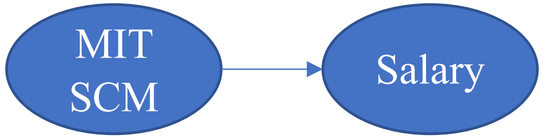
Hypothesis Testing

1. Testing the samples of the entire population meet the average.
2. Null hypothesis
3. Alternative hypothesis

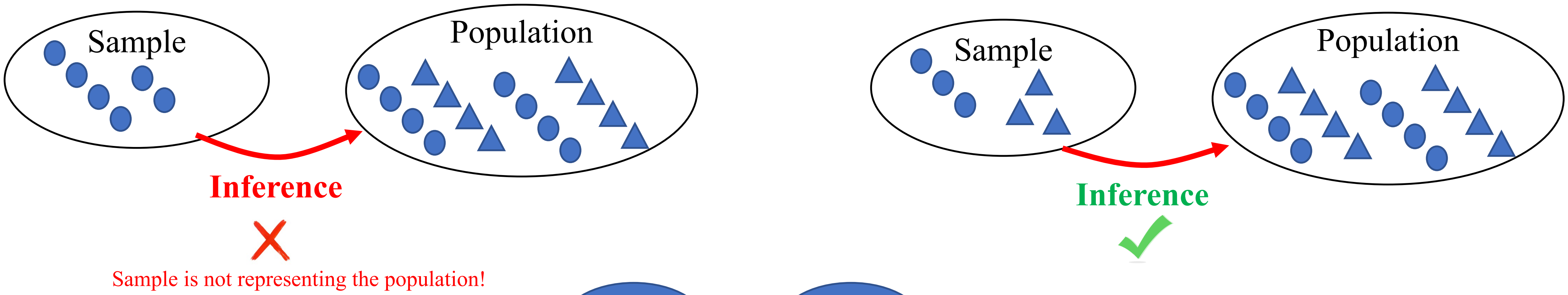
Population vs. Sample Size

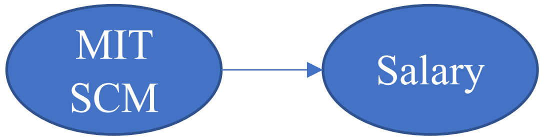






Population vs. Sample Size





- ✓ A population is the set or collection of all items which are analyzed in a specified purpose.
- ✓ A sample is a subset of population.
- ✓ A reasoning/inference from a sample to a population.
- ✓ Numbers that characterize a population are called population parameters (e.g., $\mu, \sigma, \beta, Y, \epsilon$, etc.). We will never know the exact or true values of the population parameters, we estimate them by using the sample.
- ✓ Numbers that characterize a sample of a population is called statistics (e.g., commonly used symbols $\overline{X}, S, \hat{\beta}, \hat{Y}, e$, etc.).

Population vs. Sample Size

Why do we use Sample instead of Population?

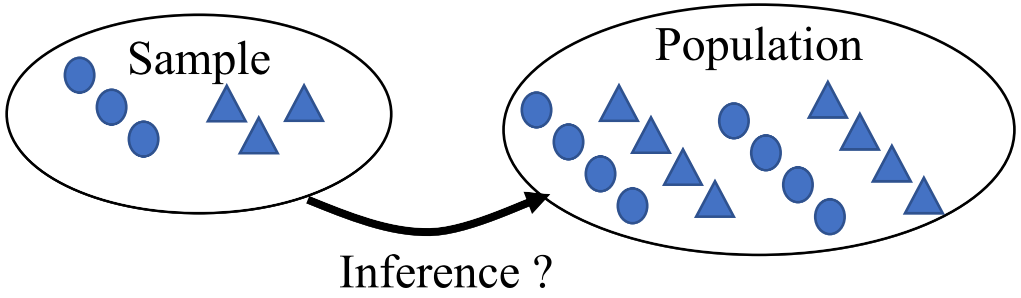
There are various reasons for taking a sample rather than the entire population:

1) Expense
2) Response time (speed)
3) Large population
4) Manageability

Population vs. Sample Size

- A sample may contain a number of **atypical** elements - which will ultimately produce inaccurate estimates of the population parameters.
- **Sampling bias** – a sample is biased if it is obtained by a method that favors the selection of individuals having particular characteristics.

- Example
  - Consider a situation where I have access to the MIT-SCM alumni file. In that file, 1,000 applicants voluntarily disclosed their starting salaries. The ratio of males to females is approximately one (see the following figure). Does my sample represent the population or it is biased?



Sampling Distribution

- To mitigate against bias, samples should be collected randomly selected using a probability distribution.

- Random variables are assumed to independent, and every variable has the same probability distribution:

$$X_1, X_2, \dots, X_n$$

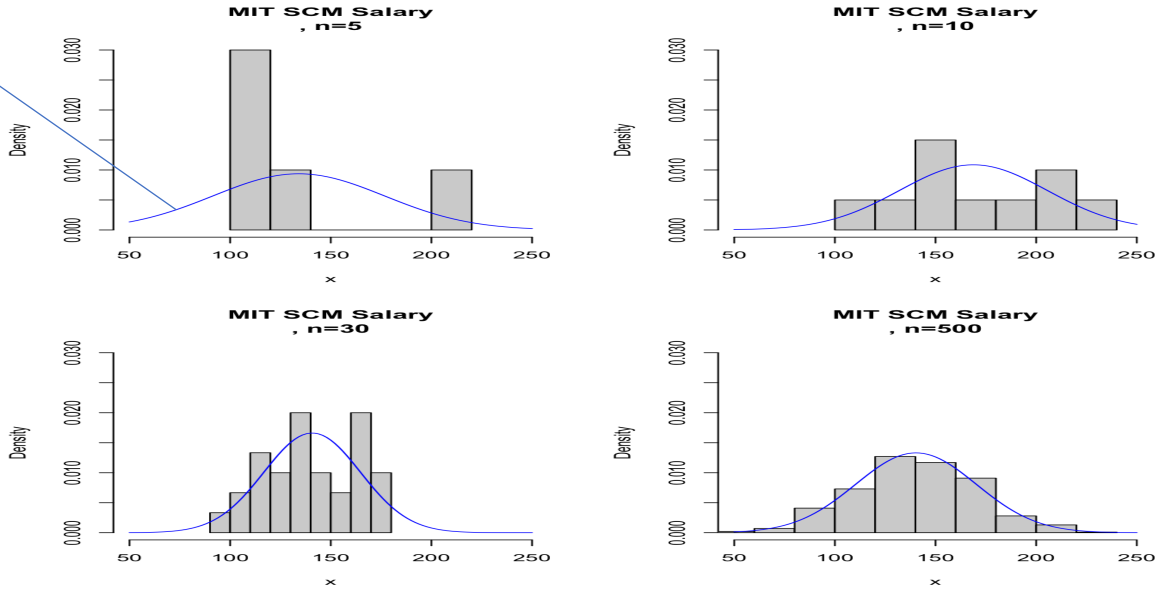
- The probability distribution of a statistic is called sampling distribution.

Population Parameters vs. Sample Statistics

- In addition to sample selection method, sampling distribution relies on the sample size and the population distribution.

Example	Population Parameters	Sample Statistics Estimated by
Mean	True mean, μ	$\overline{X} = \frac{\sum_i X_i}{n}$
Standard Deviation	True Standard Deviation, σ	$S = \sqrt{\frac{\sum_i (X_i - \overline{X})^2}{n-1}}$

Blue Graph: Population Distribution, God Knows this but we will never!



R Code

```
set.seed(7)
par(mfrow=c(2,2))
x <- rnorm(500, mean = 140, sd = 30) # your df$col
mean_x <- mean(x)
sd_x <- sd(x)
hist(x, freq = FALSE,xlim = c(50,250),ylim = c(0,.03),main = 'MIT SCM Salary\n, n=500')
curve(dnorm(x, mean = mean_x, sd = sd_x), add = TRUE, col = "blue")
```

Central Limit Theorem

- Definition: If $X_1, X_2, \dots, X_n$ are random samples drawn from a population with overall mean μ and finite variance σ^2, and if $\overline{X}_n$ is the sample mean of the first n samples, then the limiting form of the distribution, $Z = \lim_{n\to\infty}\left(\frac{\overline{X}_n-\mu}{\sigma_{\overline{X}}}\right)$, with $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$, is a standard normal distribution.*

In a simple language, the average of randomly selected sample follows normal distribution with the average of $\overline{X}_n$ and standard deviation of $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$

$$\overline{X}_n \sim Normal\ \left(\overline{X}_n, \frac{\sigma}{\sqrt{n}}\right)$$

Or

$$Z = \frac{\overline{X}_n-\mu}{\sigma_{\overline{X}}} \sim \text{Normal (0,1) aka Standard Normal}$$

*Montgomery, Douglas C.; Runger, George C. (2014). *Applied Statistics and Probability for Engineers* (6th ed.). Wiley. p. 241. ISBN 9781118539712.

Hypothesis Testing

- A statistical hypothesis is a statement about the parameters of one or more populations.
- Null hypothesis (H_0): A claim about a population characteristics that is initially assumed to be true.
  - It must always includes equality signs ($=, \geq, or \leq$)
- Alternative hypothesis (H_1): The competing claim which we favor to prove.
  - Basically anything against H_0; typical signs ($\neq, >, <$)

- There are two possible outcomes of hypothesis testing:

  1) Reject H_0 in favor of H_1
  2) Fail to reject the null hypothesis H_0

The choice of language matters here! Thus, **never** say we **accept** H_0 or H_1!

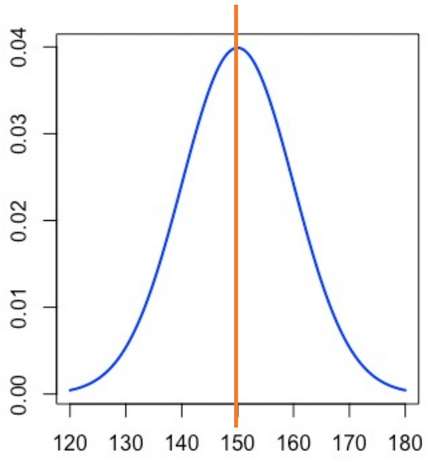
- Let's frame some hypotheses? Give me an example.

Hypothesis Testing (Type I)

- Let's say person A says (or hypothesizes) the average salary of MIT-SCM is \$150K (This means that person is talking about the average of population or $\mu = \$150{,}000$). This is H_0.
  - $H_0: \quad \mu_0 = 150k$
- Alternatively, his/her argumentative friend argues that statement is not true. This is H_1.
  - $H_1: \mu_0 \neq 150k$
- How to resolve this conflict?
  - Statistics gives us tools to construct the Confidence Interval (CI) when H_0 is true.

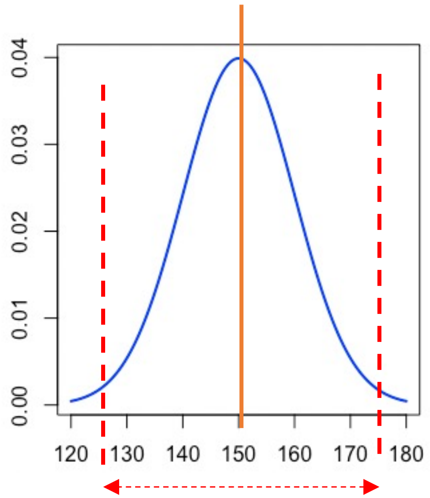
Hypothesis Testing (Type I)

- Let's say person A says (or hypothesizes) the average salary of MIT-SCM is \$150K (This means that person is talking about the average of population or $\mu = \$150{,}000$). This is H_0.
    - H_0: $\mu_0 = 150k$
- Alternatively, his/her argumentative friend argues that statement is not true. This is H_1.
    - $H_1: \mu_0 \neq 150k$
- How to resolve this conflict?
    - Statistics gives us tools to construct the Confidence Interval (CI) when H_0 is true.



$\mu_0 = 150K$

This is what person A claims.



$(1-\alpha)\%$ CI

Acceptance Region

Definition of $(1-\alpha)\%$ CI: If you draw samples 100 times, $(1-\alpha)\%$ times the sample mean or $\overline{X}$ falls into this interval

This statement means, we allow error for %5 of times!

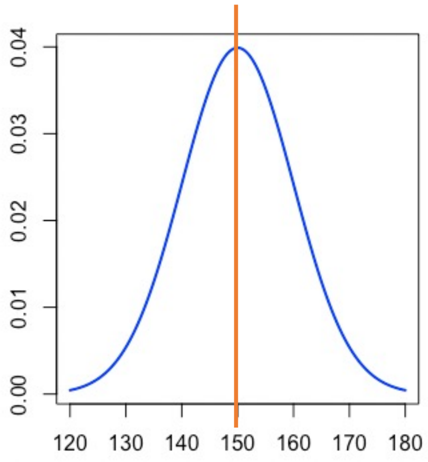
This is called α or Type I error

Typical Choice of $\alpha = 10\%, 5\%, 1\%$

1. If $\alpha = 10\%$, then $(1-\alpha)\%$ means 90% of times our sample falls into the acceptance region (Very Narrow Acceptance Region or CI).
2. If $\alpha = 5\%$, then $(1-\alpha)\%$ means 95% of times our sample falls into the acceptance region. (Commonly Used).
3. If $\alpha = 1\%$, then $(1-\alpha)\%$ means 99% of times our sample falls into the acceptance region (Very Wide Acceptance Region or CI).

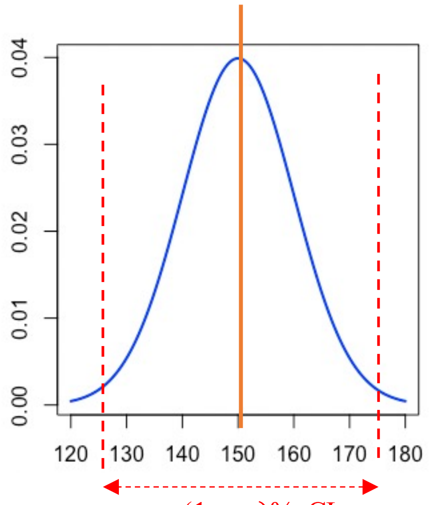
Hypothesis Testing (Type I)

- Let's say person A says (or hypothesizes) the average salary of MIT-SCM is \$150K (This means that person is talking about the average of population or $\mu = \$150{,}000$). This is H_0.
  - H_0: $\mu_0 = 150k$
- Alternatively, his/her argumentative friend argues that statement is not true. This is H_1.
  - $H_1: \mu_0 \neq 150k$
- How to resolve this conflict?
  - Statistics gives us tools to construct the confidence interval when H_0 is true.



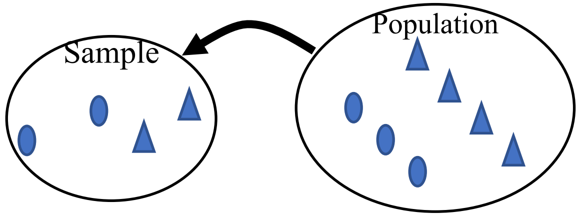
$\mu_0 = 150K$

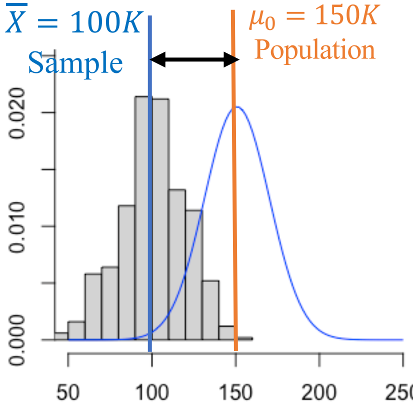
This is what person A claims.



$$\overline{X} = \frac{\sum_i X_i}{n}$$

$$S = \sqrt{\frac{\sum_i (X_i - \overline{X})^2}{n-1}}$$





Definition of $(1-\alpha)\%$ CI: If you draw samples 100 times, $(1-\alpha)\%$ times the sample mean or $\overline{X}$ falls into this interval

This statement means, we allow error for %5 of times!

This is called α or Type I error

So, what do you say?

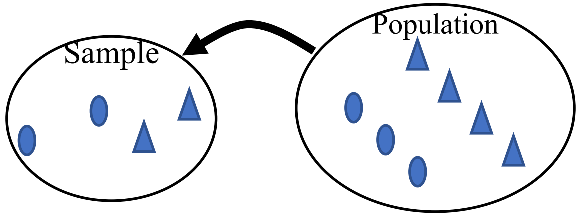
Do we Reject H_0 or we fail to reject H_0 ?

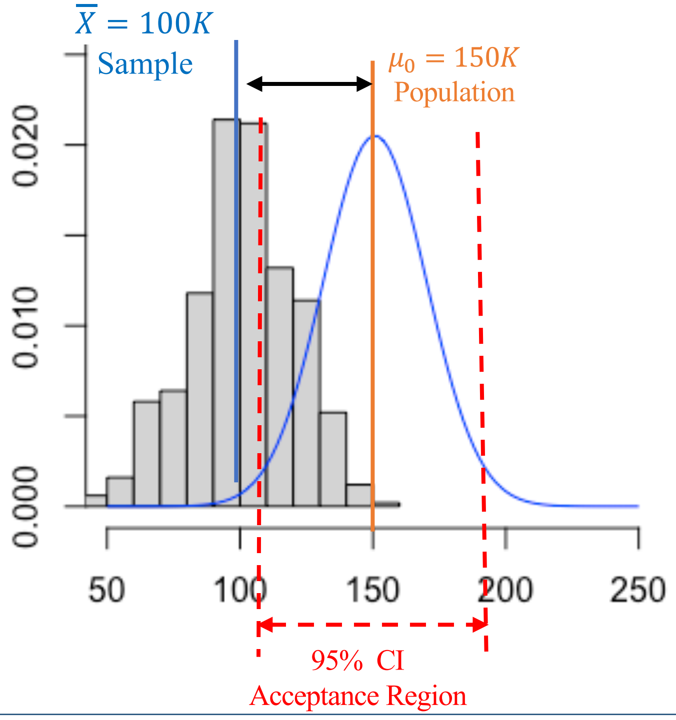
Hypothesis Testing

- To answer the Question "*Do we Reject H_0 or we fail to reject H_0?*"
  - There are two (kinda the same) methods to answer this question.
  1. If $\overline{X}$ falls into the confidence interval, then we have no evidence to reject (fail to reject), otherwise we reject.

$$\overline{X} = \frac{\sum_i X_i}{n}$$

$$S = \sqrt{\frac{\sum_i (X_i - \overline{X})^2}{n-1}}$$





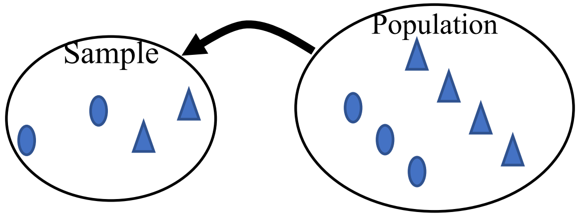
If we set $\alpha = 0.05$, then 95% CI is depicted by those dashed lines. Since $\overline{X}$ falls outside, then we have enough evidence to reject H_0 at 5 (or α%) percent level of significance.

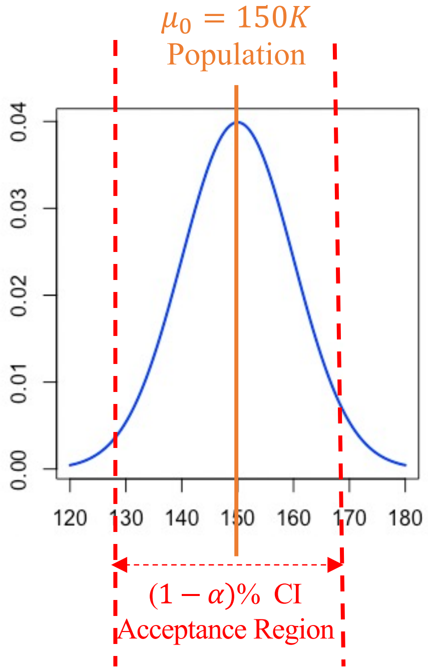
How do we calculate CI?

- To answer the Question “*Do we Reject H_0 or we fail to reject H_0?*”
  - There are two (kinda the same) methods to answer this question.
  1. If $\overline{X}$ falls into the confidence interval, then we have no evidence to reject (fail to reject), otherwise we reject.

$$\overline{X} = \frac{\sum_i X_i}{n}$$

$$S = \sqrt{\frac{\sum_i (X_i - \overline{X})^2}{n-1}}$$





$$LB = \mu_0 - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

$$UB = \mu_0 + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

Ideally

If we know σ, then we use z (normal distribution) to calculate CI.

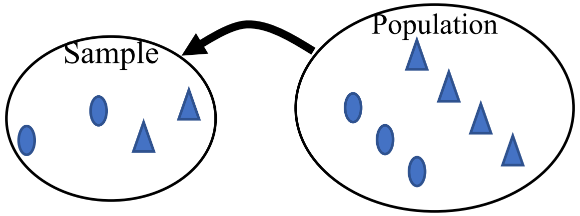
This hardly happens in real world.

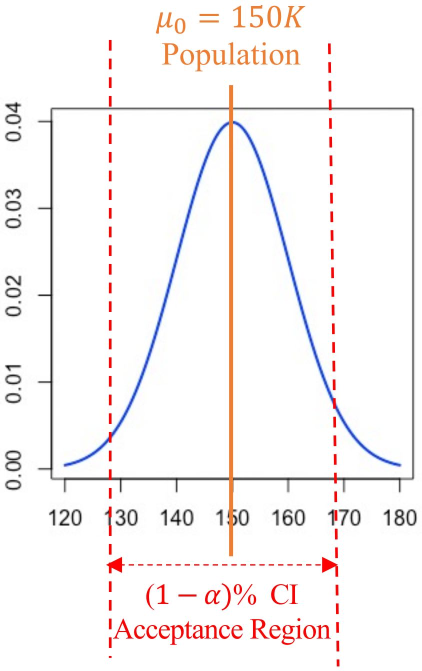
How do we calculate CI?

$$H_0: \mu = \mu_0$$
$$H_1: \mu \neq \mu_0$$

$$\overline{X} = \frac{\sum_i X_i}{n}$$

$$S = \sqrt{\frac{\sum_i (X_i - \overline{X})^2}{n-1}}$$





$$LB = \mu_0 - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \qquad UB = \mu_0 + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

Ideally

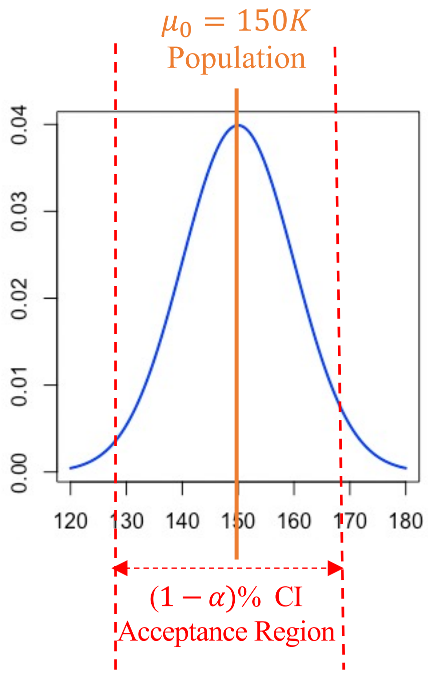
If we know σ, then we use z (normal distribution) to calculate CI.

This hardly happens in real world.

Alternatively (In Practice)

If we end up estimating σ by using s, then we use t distribution.

t distribution has an element of degree of freedom which is usually equals to $n-1$

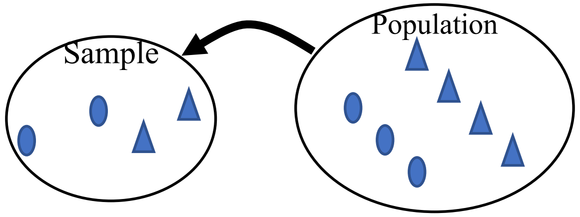


$$LB = \mu_0 - t_{\frac{\alpha}{2}, n-1} \frac{s}{\sqrt{n}} \qquad UB = \mu_0 + t_{\frac{\alpha}{2}, n-1} \frac{s}{\sqrt{n}}$$

How do we calculate CI?

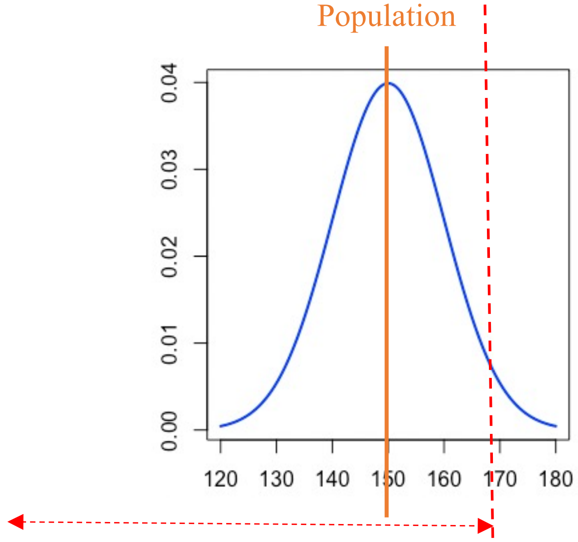
$$\overline{X} = \frac{\sum_i X_i}{n}$$

$$S = \sqrt{\frac{\sum_i (X_i - \overline{X})^2}{n-1}}$$



$H_0: \mu = \mu_0$
$H_1: \mu > \mu_0$

$\mu_0 = 150K$
Population



$(1-\alpha)\%$ CI
Acceptance Region

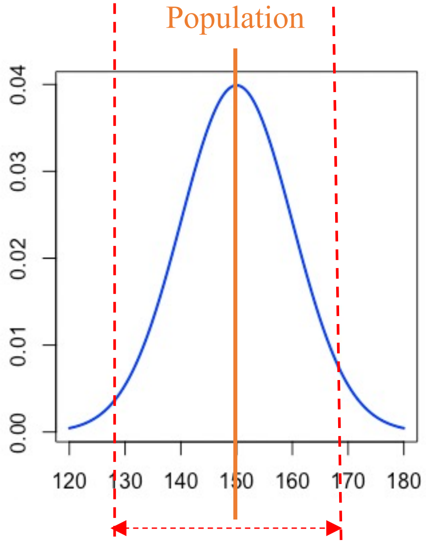
$$UB = \mu_0 + t_{\alpha,n-1}\frac{s}{\sqrt{n}}$$

One-tailed tests

Pay attention!

$H_0: \mu = \mu_0$
$H_1: \mu \neq \mu_0$

$\mu_0 = 150K$
Population



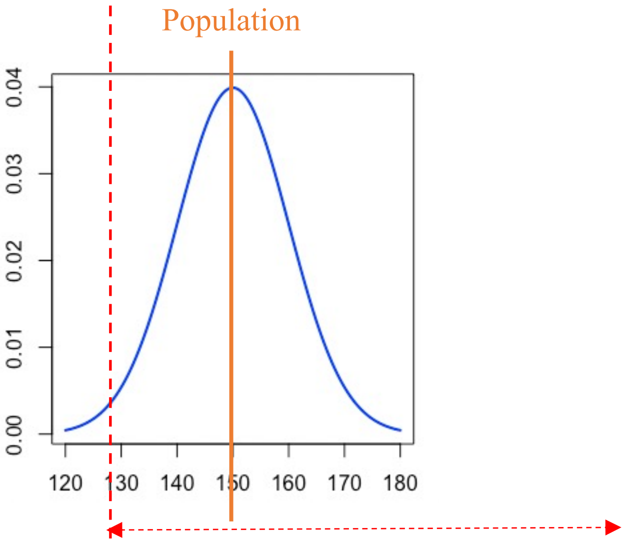
$(1-\alpha)\%$ CI
Acceptance Region

$$LB = \mu_0 - t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}} \qquad UB = \mu_0 + t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}$$

Two-tailed tests

$H_0: \mu = \mu_0$
$H_1: \mu < \mu_0$

$\mu_0 = 150K$
Population



$(1-\alpha)\%$ CI
Acceptance Region

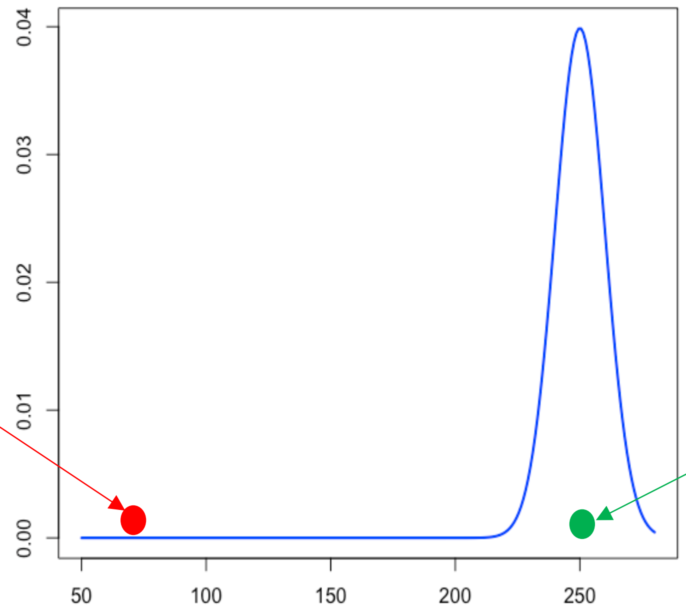
$$LB = \mu_0 - t_{\alpha,n-1}\frac{s}{\sqrt{n}}$$

One-tailed tests

Hypothesis Testing (P-value)

- To answer the Question "*Do we Reject H_0 or we fail to reject H_0?*"
  - There are two (kinda the same) methods to answer this question.
  1. If $\overline{X}$ falls into the confidence interval, then we have no evidence to reject (fail to reject), otherwise we reject.
  2. If P-value is larger than α (or $\frac{\alpha}{2}$ in a two-tailed test), then we have no evidence to reject (fail to reject), otherwise we reject.

- Is there any possibility that this observation belong to this distribution? Yes
- What is the probability to see observation like this or even more extreme than this? Extremely Low

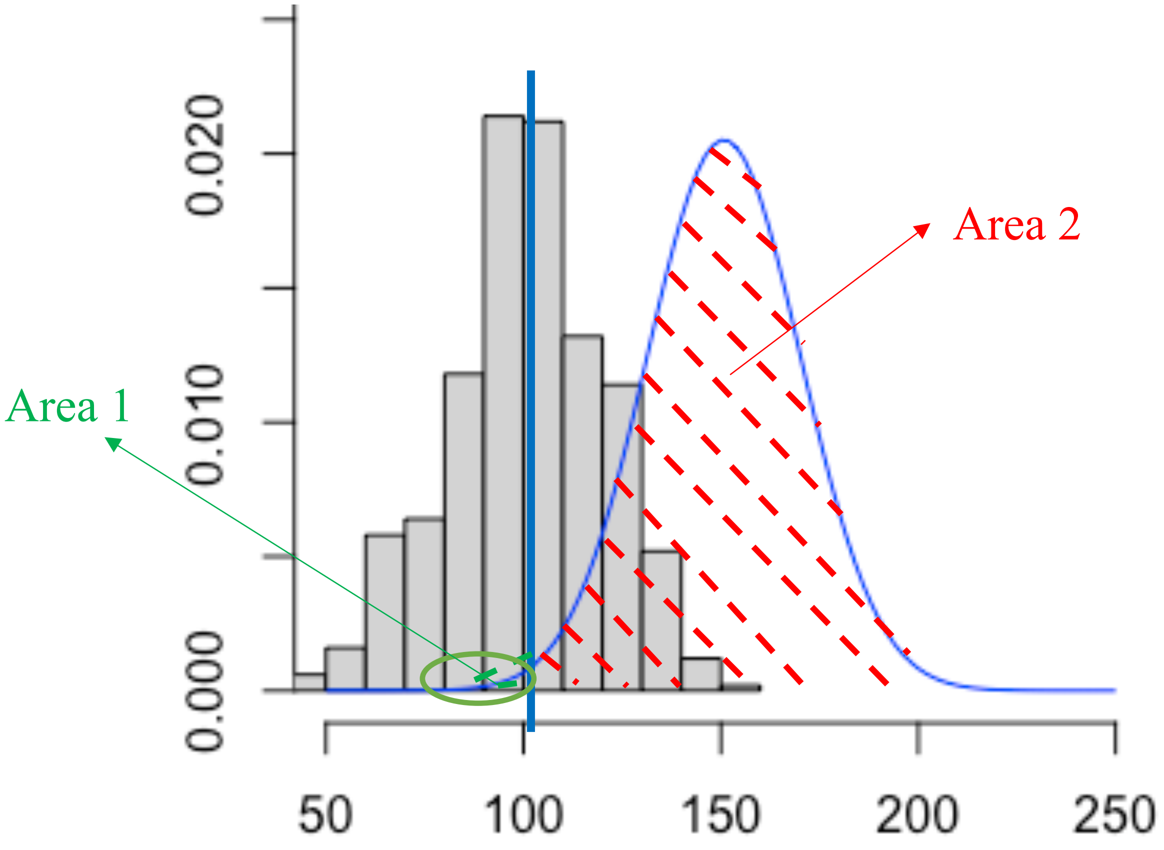


- Is there any possibility that this observation belong to this distribution? Yes
- What is the probability to see observation like this or even more extreme than this? Very High

Definition: The p-value is the probability of obtaining test results at least as extreme as the result actually observed, under the assumption that the null hypothesis is correct.

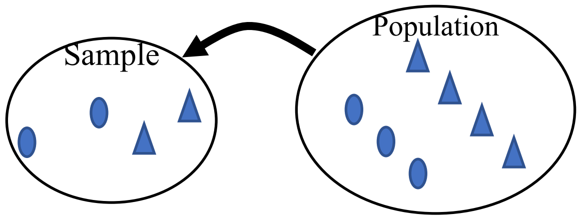
Hypothesis Testing (P-value)

- To answer the Question "*Do we Reject H_0 or we fail to reject H_0?*"
  - There are two (kinda the same) methods to answer this question.
  1. If $\overline{X}$ falls into the confidence interval, then we have no evidence to reject (fail to reject), otherwise we reject.
  2. If P-value is larger than α (or $\frac{\alpha}{2}$ in a two-tailed test), then we have no evidence to reject (fail to reject), otherwise we reject.



$$\overline{X} = \frac{\sum_i X_i}{n}$$

$$S = \sqrt{\frac{\sum_i (X_i - \overline{X})^2}{n-1}}$$



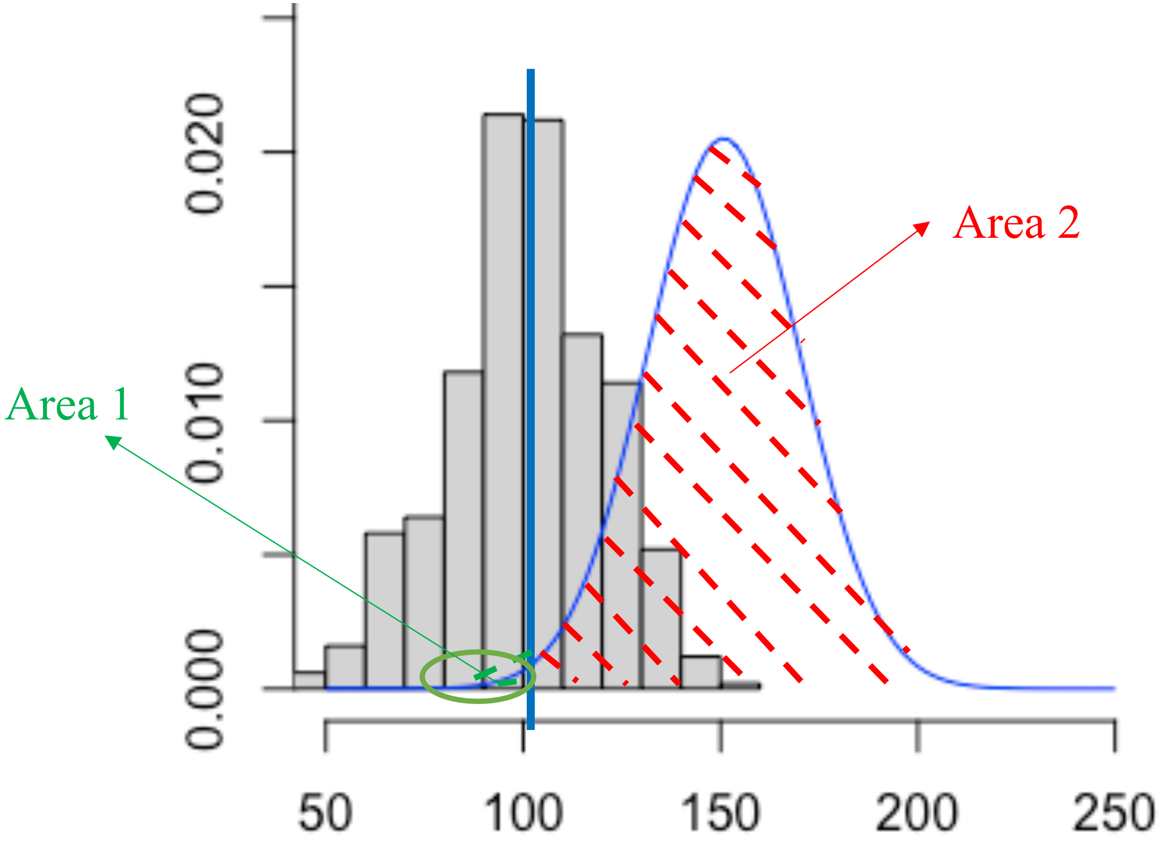
P-value=min(Area 1, Area 2)=Area 1

The p-value is the probability of obtaining test results at least as extreme as the result actually observed, under the assumption that the null hypothesis is correct.

In a simple language, p-value means to what extent the observed data belong to the population.

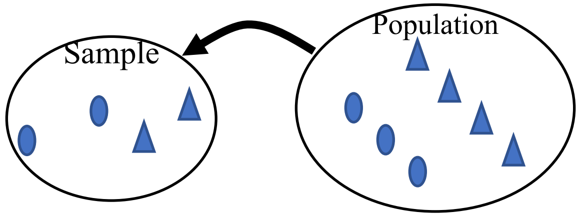
Hypothesis Testing (P-value)

- To answer the Question "*Do we Reject H_0 or we fail to reject H_0?*"
  - There are two (kinda the same) methods to answer this question.
  1. If $\overline{X}$ falls into the confidence interval, then we have no evidence to reject (fail to reject), otherwise we reject.
  2. If P-value is larger than α (or $\frac{\alpha}{2}$ in a two-tailed test), then we have no evidence to reject (fail to reject), otherwise we reject.



$$\overline{X} = \frac{\sum_i X_i}{n}$$

$$S = \sqrt{\frac{\sum_i (X_i - \overline{X})^2}{n-1}}$$



One-tailed

P-value=min(Area 1, Area 2)=Area 1 $\begin{cases} P\_value < \alpha & \text{Reject} \\ P\_value > \alpha & \text{Fail to reject} \end{cases}$

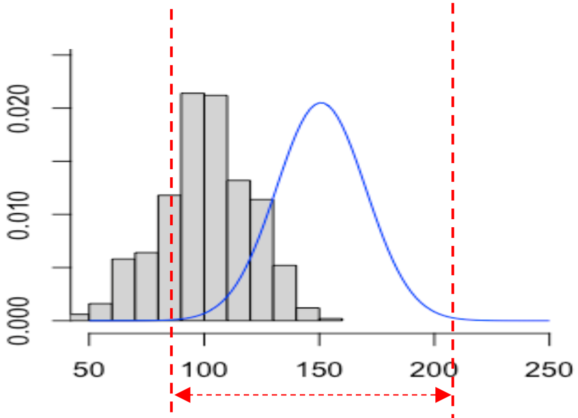
Two-tailed

P-value=min(Area 1, Area 2)=Area 1 $\begin{cases} P\_value < \alpha/2 & \text{Reject} \\ P\_value > \alpha/2 & \text{Fail to reject} \end{cases}$

Hypothesis Testing (Type I & II errors)

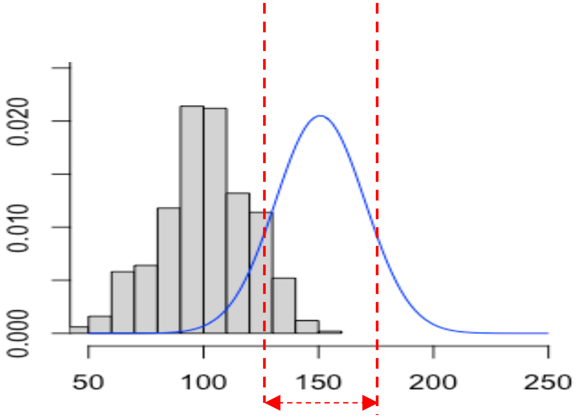
1. Type I error or α

- $\alpha = P(Type\ I\ Error) = P(Reject\ H_0\ when\ H_0\ is\ True)$
- What would happen if I increase α?

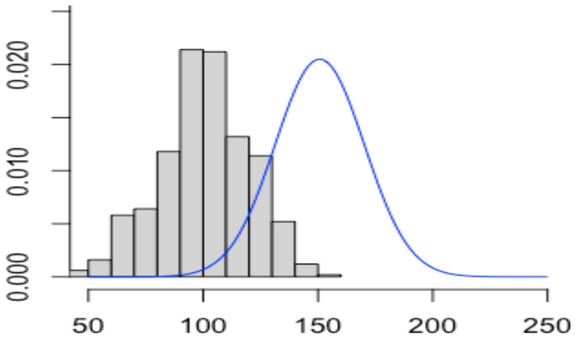


$(1-\alpha_1)\%$ CI

$\alpha_1 < \alpha_2$



$(1-\alpha_2)\%$ CI

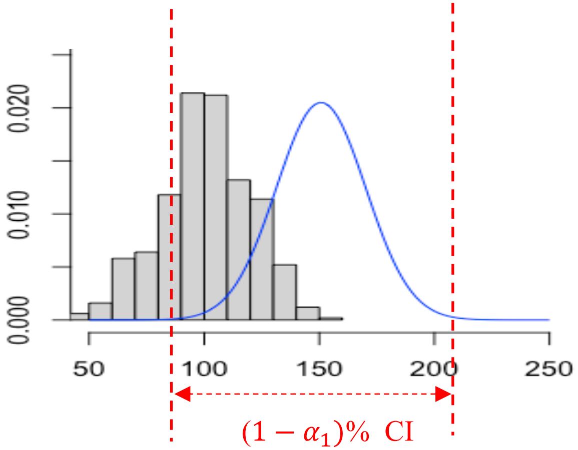


Maybe this sample was very odd!
Because we allow error for α% of times.

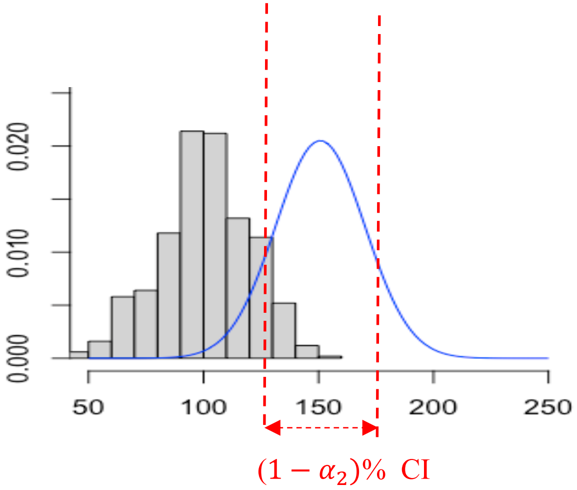
Hypothesis Testing (Type I & II errors)

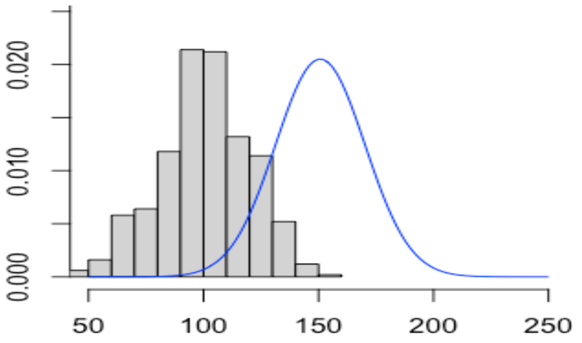
1. Type I error or α

- $\alpha = P(Type\ I\ Error) = P(Reject\ H_0\ when\ H_0\ is\ True)$
- What would happen if I increase α?



$\alpha_1 < \alpha_2$

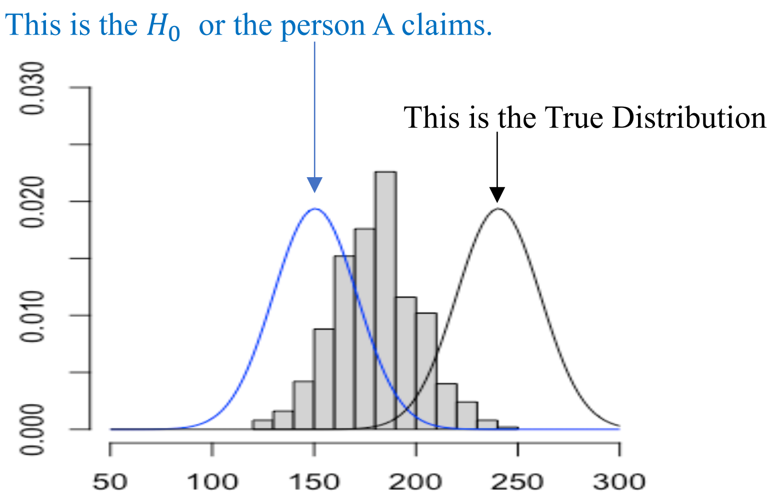




Maybe this sample was very odd!
Because we allow error for α% of times.

2. Type II error

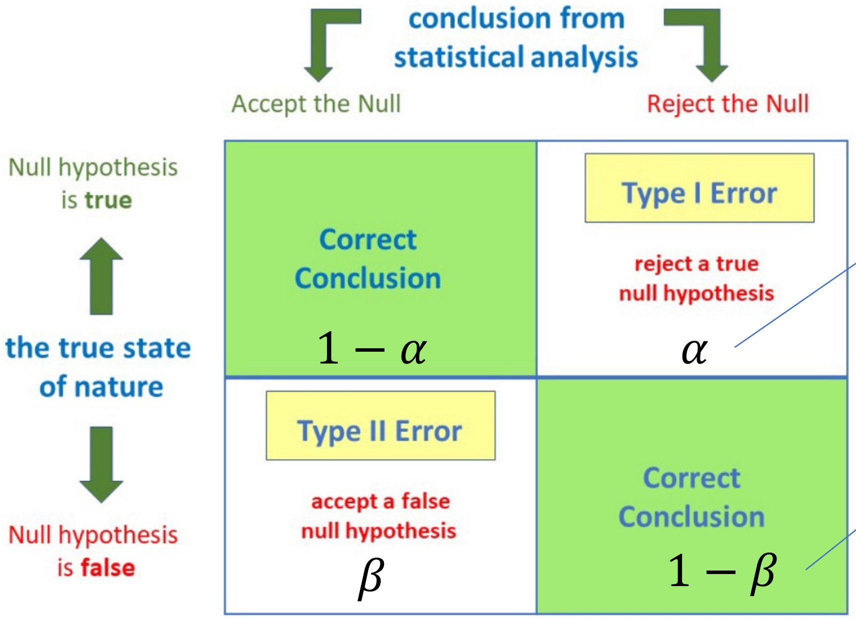
- $\beta = P(Type\ II\ Error) = P(Fail\ to\ reject\ H_0\ when\ H_0\ is\ False)$



Clearly, H_0 or the claim of person A is wrong, but we collect data and unfortunately we fail to reject it since our sample overlaps a lot with the blue graph.

Thus, we end up not rejecting a claim which is wrong, aka Type II error.

Hypothesis Testing (Type I & II errors)

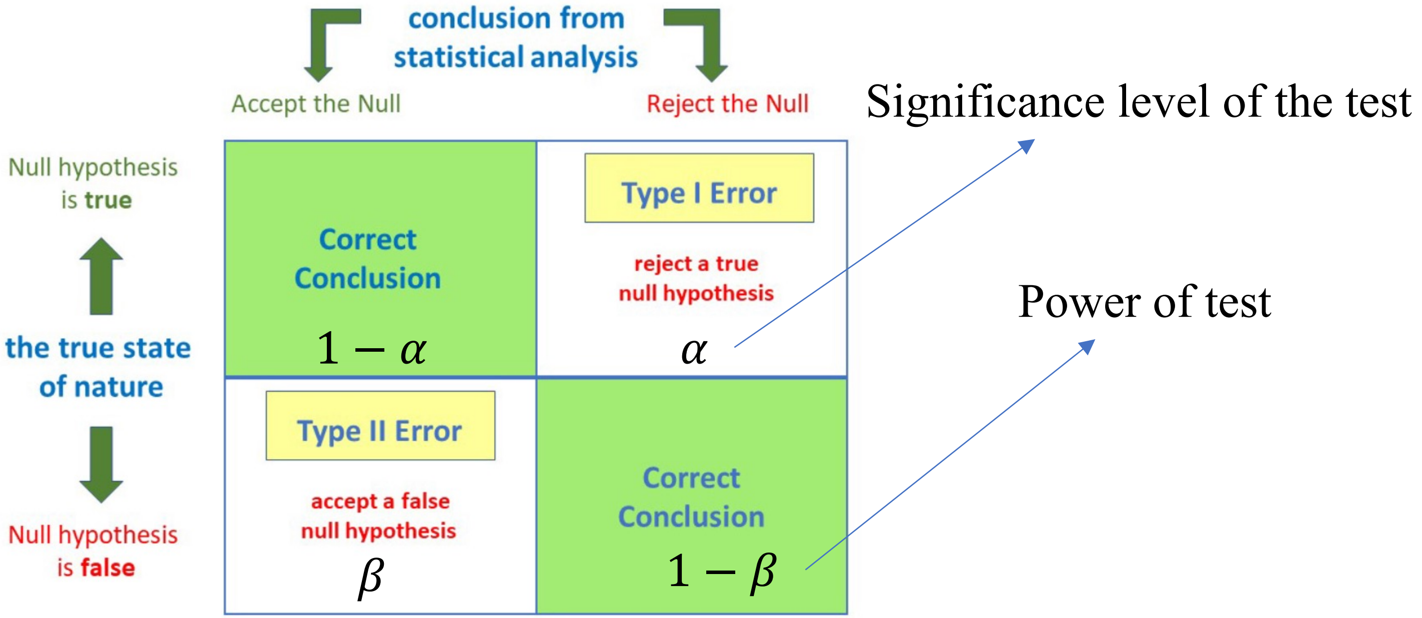


Significance level of the test

Power of test

https://www.simplypsychology.org/type_I_and_type_II_errors.html

Hypothesis Testing (Type I & II errors)



Sometimes we select the minimum sample size by these two measures.

1. First we set α at the maximum allowable Type I error rate (typically 1%,5% or 10%).
2. Next, we calculate Type II error or $(1-\beta)$.

$$1-\beta = f(n)$$

3. Sometime $f(n)$ is too complicated, so we use simulation to find out the minimum sample size.

https://www.simplypsychology.org/type_I_and_type_II_errors.html

Example (Type I & II Errors)

- Suppose there is a test for a particular disease.
  - ➢ If the disease really exists and is diagnosed early, it can be successfully treated
  - ➢ If it is not diagnosed and treated, the person will become severely disabled
  - ➢ If a person is erroneously diagnosed as having the disease and treated, no physical damage is done.

- First, clearly state your null hypothesis, then,
- What is Type I error?
- What is Type II error?
- Which error do you want to minimize here?

* Depending the way you frame your hypotheses, Type I and II might change.

Example (Type I & II Errors)

Suppose you are a manager of a company.

- One of your consultant tells you that you have to construct new production facility because the current production can handle up to 200% increase in demand. The consultant hypothesize that the demand is going increase 300%. The new facility cost 10 times more than annual lost sales & reputation cost.

What is null hypothesis?

Type I error?

Type II error?

Which error do you want to minimize the most?

Hypothesis Testing

- **General Procedure in Hypothesis Tests:**
  1) Identify parameters of interest
  2) State null and alternative hypotheses
  3) Select a significance level (α) and the critical values
  4) Determine the appropriate test statistics
  5) Calculate the test statistics
  6) Calculate rejection criteria and compare p-value
  7) Make statistical decision and interpret results

Simple Linear Regression

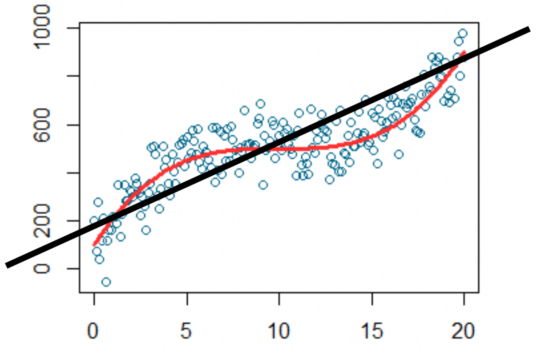
- A regression model is used to model and explore relationships between variables that are related nondeterministic manner.
- Simple Linear Regression: Only one independent variable x (regressor or predictor) and one dependent variable Y (response).

$$E(Y|x) = \mu_{Y|x} = \beta_0 + \beta_1 x$$

- The mean of Y is a linear function of x. However, the actual observed value y does not have exact linear relationship.
- The **fitted or estimated regression line:**

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x$$ Estimated

$$y = \beta + \beta_1 x + \epsilon$$ True relationship



Significance of Regression

What percentage of variation is explained by the model.

OLS Regression Results

Dep. Variable:	MPG	R-squared:	0.201
Model:	OLS	Adj. R-squared:	0.157
Method:	Least Squares	F-statistic:	4.532
Date:	Mon, 29 Aug 2022	Prob (F-statistic):	0.0473
Time:	18:24:56	Log-Likelihood:	-53.272
No. Observations:	20	AIC:	110.5
Df Residuals:	18	BIC:	112.5
Df Model:	1		
Covariance Type:	nonrobust		

$H_0: all\ Coefficients\ are\ zero$
$\beta_0 = \beta_1 = 0$

$H_1: Otherwise$

P-value

95% CI for each β

		coef	std err	t	P>|t|	[0.025	0.975]
β_0	const	33.5348	2.614	12.829	0.000	28.043	39.027
β_1	EngDis	-0.0354	0.017	-2.129	0.047	-0.070	-0.000

$H_0: \beta_0 = 0$
$H_1: \beta_0 \neq 0$

$H_0: \beta_1 = 0$
$H_1: \beta_1 \neq 0$

$$y = \beta_0 + \beta_1 x = 33.53 - 0.035x$$

What would be regression if the p-value of β_1 is equal to 0.2?

Supply Chain MANAGEMENT