Agenda

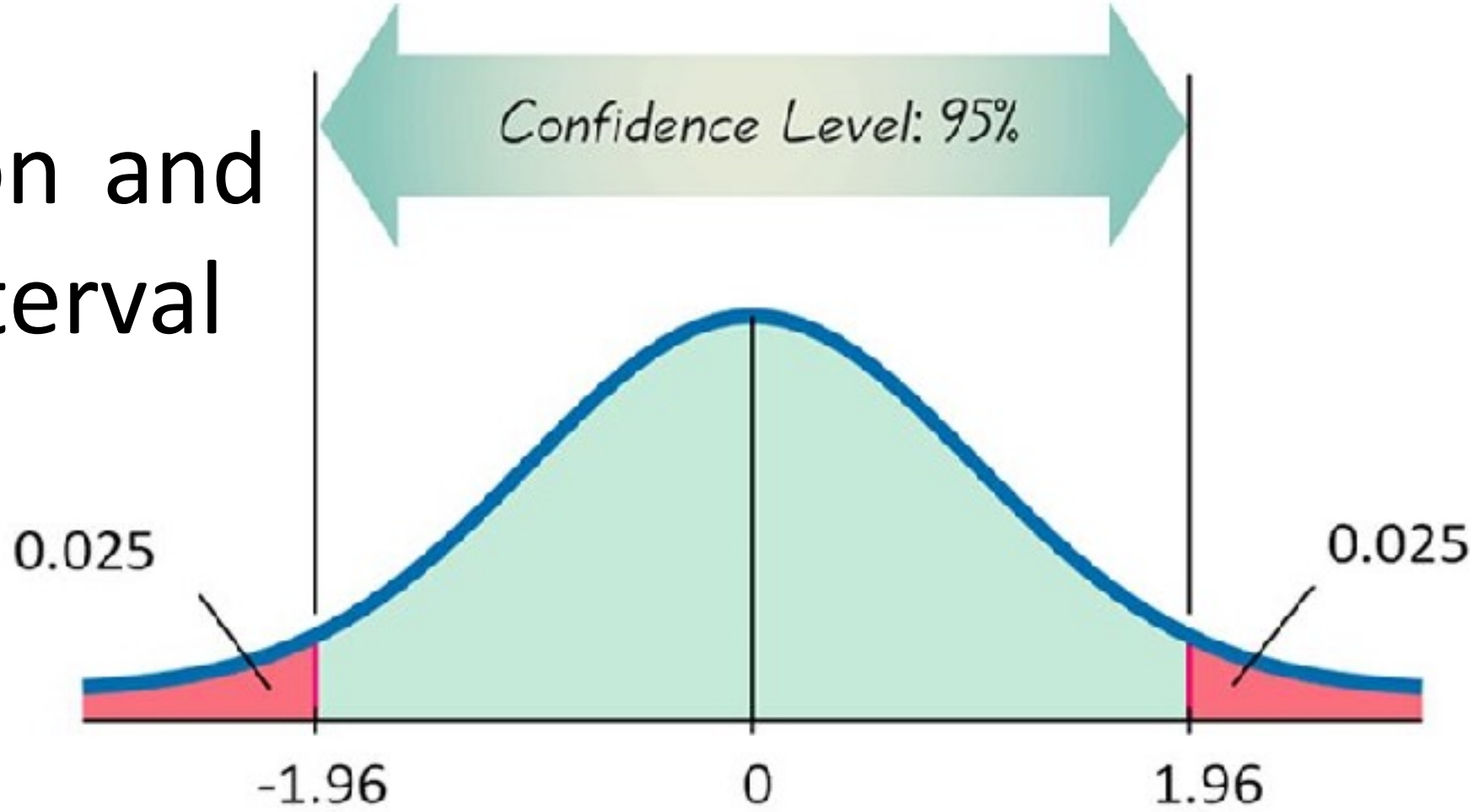
- Final review (L06-L10)

- Python tutorials
  - Probability
  - Confidence interval and Hypothesis Testing
  - Pandas dataframe and time series

- Quiz 6 (L10)

Final review (L06-L10)

L06: Estimation and Confidence Interval



The One-Sample z Confidence Interval for μ

The general formula for a confidence interval for a population mean μ is

$$\bar{x} \pm (z \text{ critical value})\left(\frac{\sigma}{\sqrt{n}}\right)$$

When

1. $\bar{x}$ is the sample mean from a **simple random sample,**
2. the **sample size *n* is large** (generally $n > 30$)
3. σ, **the population standard deviation, is known**

The three most used confidence levels, 90%, 95%, and 99%, use z critical values 1.645, 1.96, and 2.58, respectively.

Sample size

Problem: Find the **sample size** necessary in order to obtain a specified maximum error and level of confidence (assume the standard deviation is known).

The bound-on error of estimation associated with a 95% confidence interval is

$$B = 1.96\left(\frac{\sigma}{\sqrt{n}}\right)$$

Solve this expression for n:

$$n = \left(\frac{1.96\sigma}{B}\right)^2$$

If the desired confidence level is something other than 95%, 1.96 is replaced by the appropriate z critical value (for example, 2.58 for 99% confidence).

Confidence Interval When σ Is Unknown

The confidence interval just developed has an obvious drawback: To compute the interval endpoints, σ must be known.

$$\bar{x} \pm (z \text{ critical value})\left(\frac{\sigma}{\sqrt{n}}\right)$$

When σ is unknown, we use the sample standard deviation s to estimate σ.

In place of z-scores, we must use the following to t value:

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

Confidence Interval When σ Is Unknown

The general formula for a confidence interval for a population mean μ based on a sample of size n when is

$$\bar{x} \pm (t \text{ critical value})\left(\frac{s}{\sqrt{n}}\right)$$

When

1. $\bar{x}$ is the sample mean from a **simple random sample,**
2. the population distribution is normal**, or** the sample size n is large (generally n > 30)
3. σ, **the population standard deviation, is unknown**

where the *t* critical value is based on df = *n* - 1. t Table gives critical values appropriate for each of the confidence levels 90%, 95%, and 99%, as well as several other less frequently used confidence levels.

Choosing the Sample Size

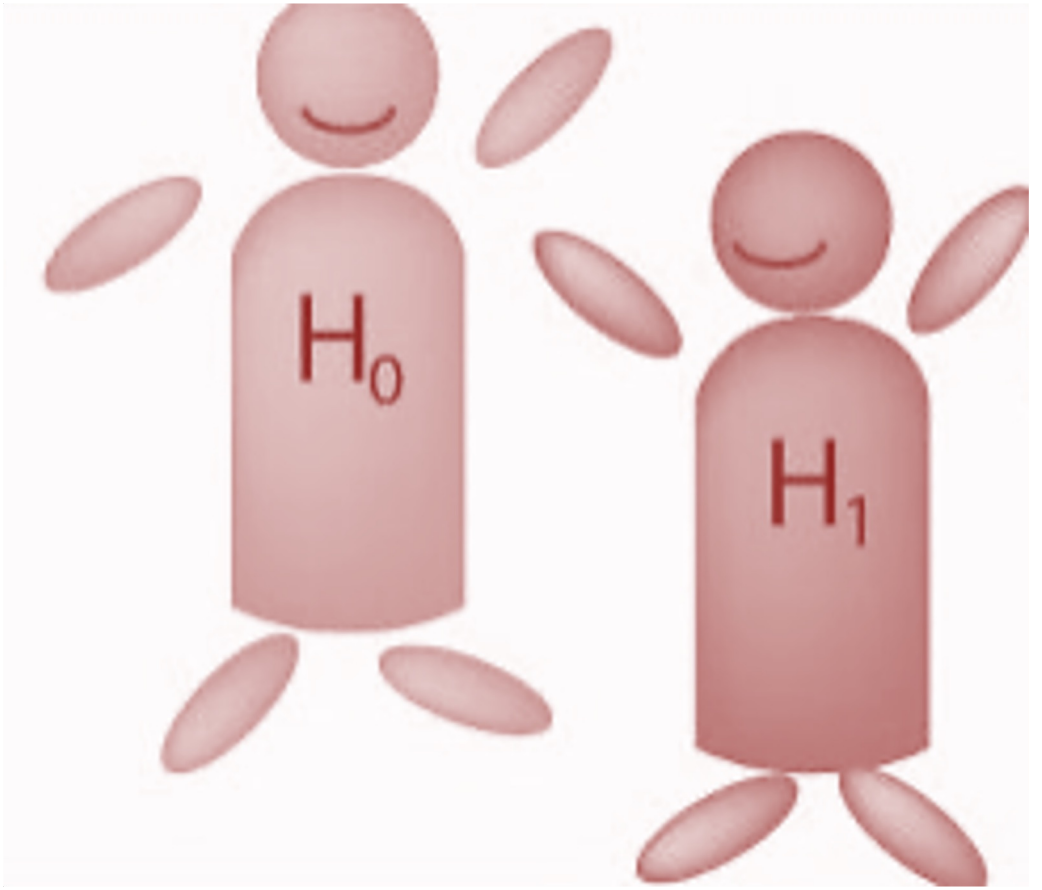
The sample size required to estimate a population mean μ to within an amount B with 95% confidence is

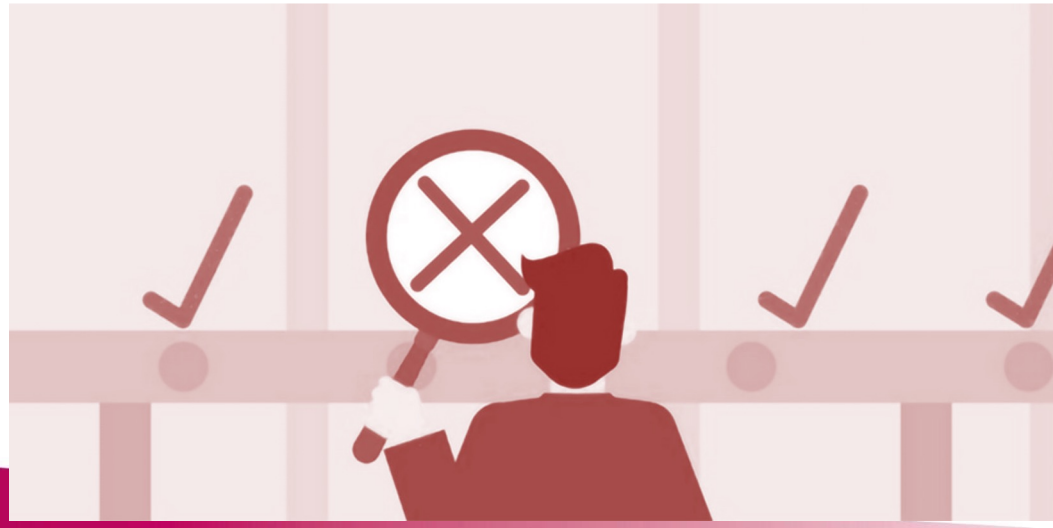
$$n = \left(\frac{1.96\sigma}{B}\right)^2$$

If σ is unknown, it may be estimated based on previous information or, for a population that is not too skewed, by using (range)/4.

If the desired confidence level is something other than 95%, 1.96 is replaced by the appropriate z critical value (for example, 2.58 for 99% confidence).

L07:
Hypothesis Testing - One Population





Main Concept of Hypothesis Testing

1: Determine H_0 and H_1.

2: Under H_0, define a rare event –the event which happens with a very small probability in one experiment.

3: Collect data and compute the test statistics

4. Make decision: If data contradicts H_0, then reject H_0; otherwise, do NOT reject H_0.

Form of Hypothesis Testing

Null hypothesis

H_0: population characteristic = hypothesized value

Noted that the alternative hypothesis uses the same population characteristic and the same hypothesized value as the null hypothesis.

Alternative hypothesis

One-tailed test since you are interested in one direction

H_1: population characteristic > hypothesized value
H_1: population characteristic < hypothesized value

H_1: population characteristic ≠ hypothesized value

a **two-tailed** test

Type of Hypothesis Testing

According to the form of the alternative hypothesis, we can have the following

Four types of tests:

I) SIMPLE TEST

$$\begin{cases} H_0: \mu_X = \mu_0 \\ H_1: \mu_X = \mu_1 \end{cases}$$

II) ONE-SIDED RIGHT TEST

$$\begin{cases} H_0: \mu_X = \mu_0 \\ H_1: \mu_X > \mu_0 \end{cases}$$

III) ONE-SIDED LEFT TEST

$$\begin{cases} H_0: \mu_X = \mu_0 \\ H_1: \mu_X < \mu_0 \end{cases}$$

IV) TWO-SIDED TEST

$$\begin{cases} H_0: \mu_X = \mu_0 \\ H_1: \mu_X \neq \mu_0 \end{cases}$$

Unless we have enough information to do a simple test or a one-sided test, we would opt for the default which is a two-sided test.

Test Errors and Error Probabilities

- **Note that there is no perfect test statement. Each test statement must lead to the following two kinds of errors.**

	Not reject H_0	Reject H_0
If H_0 is true	No error	**TYPE I ERROR**
If H_0 is false	**TYPE II ERROR**	No error

TYPE I ERROR: the error of rejecting H_0 when it is in fact true.

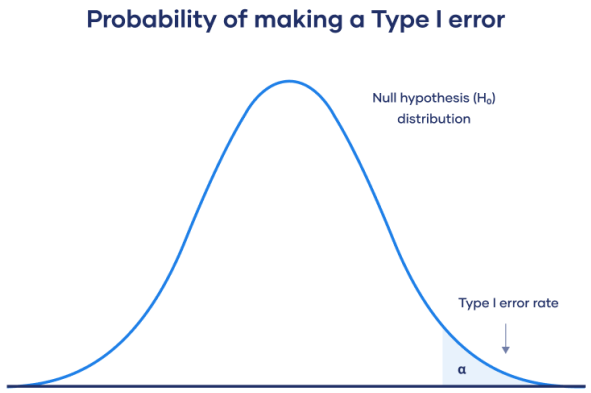
TYPE II ERROR: the error of not rejecting H_0 when it is in fact false.

Test Errors and Error Probabilities

Correspondingly, we have

α also called significance level

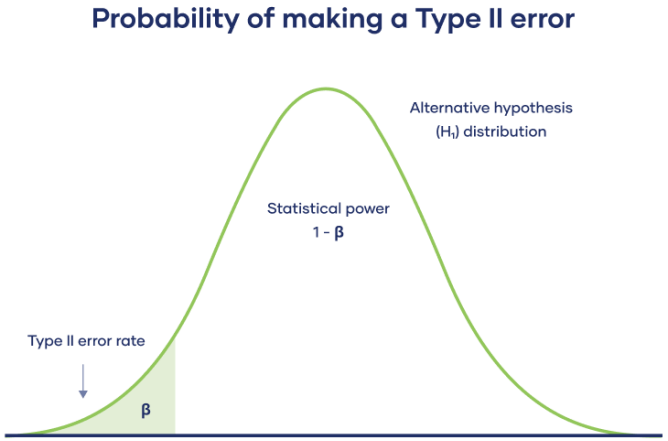
$$\alpha = P(Type\ I\ error) = P(reject\ H_0\ if\ H_0\ is\ true)$$



It is the probability of making a wrong decision to reject H_0.

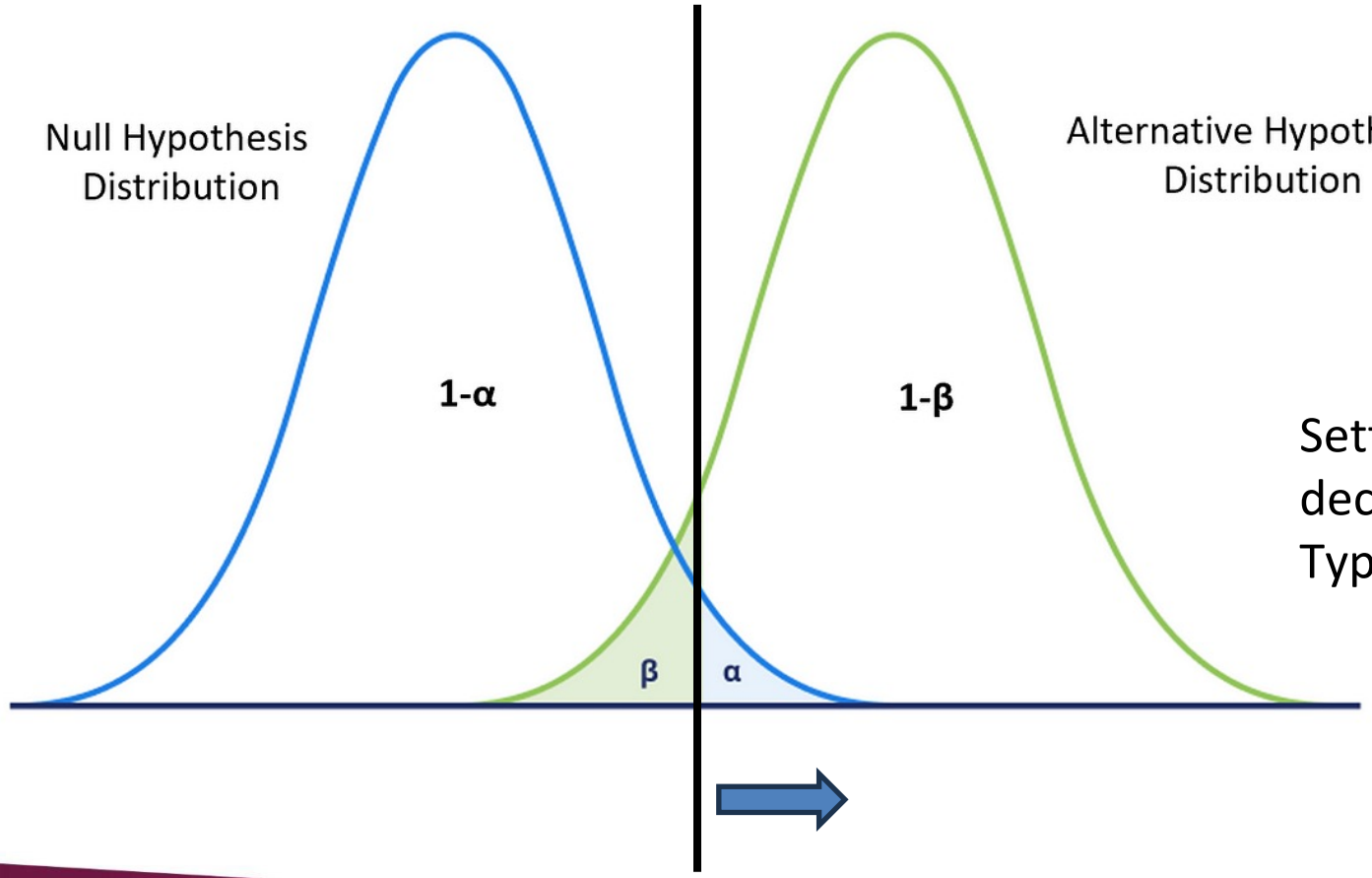
$$\beta = P(Type\ II\ error) = P(Not\ reject\ H_0\ if\ H_0\ is\ false)$$

It is the probability of making a wrong decision not to reject H_0.



Trade-off between Type I and Type II errors

The Type I and Type II error rates influence each other



The alpha level α (the significance level) represents the maximum probability of making a Type I error that the researcher is willing to accept.

Setting a lower significance level (α) decreases a Type I error risk; but increases a Type II error (β) risk. And vice versa

Determination of a Critical Value

So, in designing a test statement, we normally guarantee α in a desired low value (often choose 0.01, 0.05 or 0.1), and then find a test statement with β as small as possible.

How to design a test statement with this restriction of α?

A Probability-Value Approach

Hypothesis test:

1. A well-organized, step-by-step procedure used to make a decision.

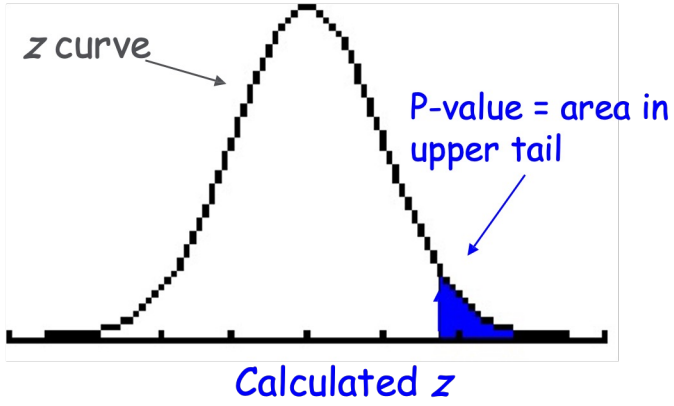
2. Probability-value approach (*p*-value approach)

What is P-value?

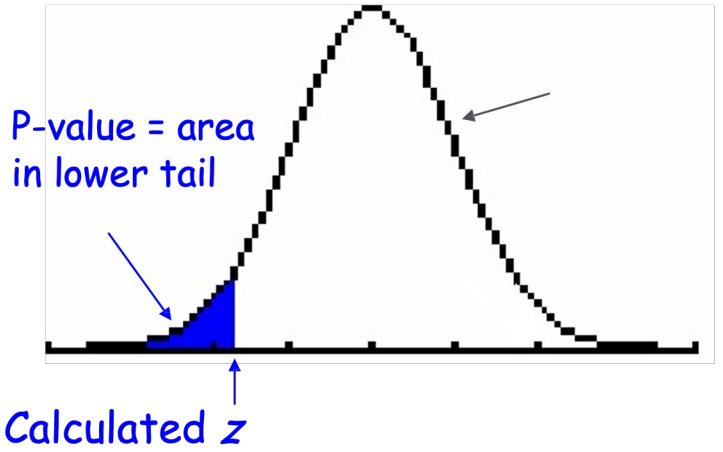
The ***P*-value** (also sometimes called the **observed significance level**) measures of the **strength of the evidence against the null hypothesis (H_0).** It's calculated from the observed data and represents the probability of obtaining results **at least as extreme as the observed results**, assuming that the null hypothesis is true.

The calculation of the P-value depends on the form of the inequality in the alternative hypothesis.

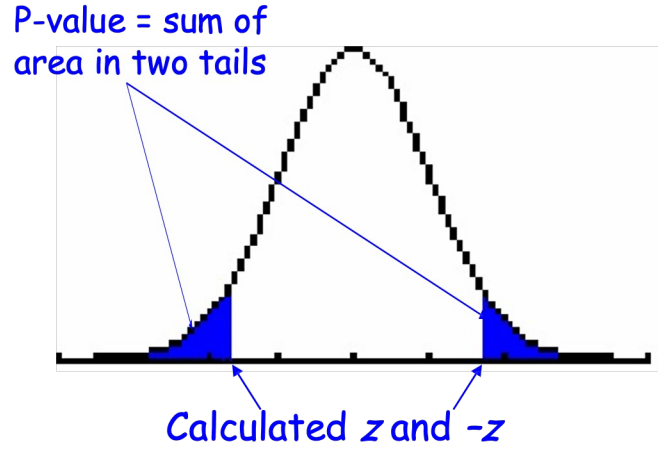
- H_1: *p* > *hypothesize value*



- H_1: *p* < *hypothesize value*



- H_a: *p* ≠ *hypothesize value*



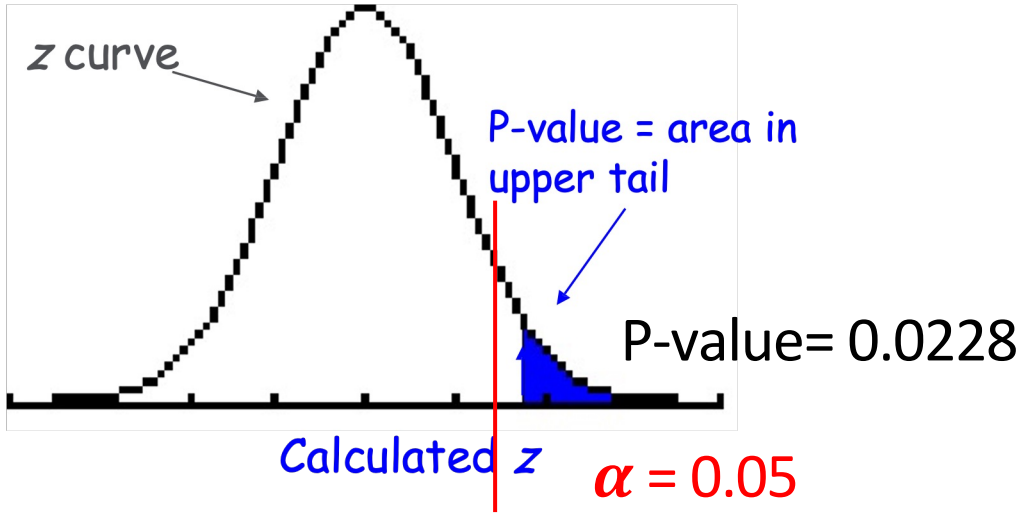
The smaller the p-value, the stronger the evidence against H_0 provided by the data.

Decision-making after computing the P-value

A decision about whether to reject or to fail to reject H_0 results from comparing the *P*-value to the chosen α:

H_0 should be rejected if *P*-value $\leq \alpha$.
H_0 should not be rejected if *P*-value $> \alpha$.

The *P*-value measures of the strength of the evidence against the null hypothesis (H_0).



$\alpha = 0.05$

P-value= 0.0228

P-value < α

Reject the null hypothesis (H_0)

The alpha α (significance level) represents the maximum probability of making a Type I error that the researcher is willing to accept.

Hypothesis Test of Mean μ (σ Known):

A Probability-Value Approach

1. The Set-Up:
   a. Describe the population parameter of concern.
   b. State the null hypothesis (H_0) and the alternative hypothesis (H_1).
2. The Hypothesis Test Criteria:
   a. Check the assumptions.
   b. Identify the probability distribution and the test statistic formula to be used.
   c. Determine the level of significance, α.
3. The Sample Evidence:
   a. Collect the sample information.
   b. Calculate the value of the test statistic.
4. The Probability Distribution:
   a. Calculate the *p*-value for the test statistic.
   b. Determine whether or not the *p*-value is smaller than α.
5. The Results:
   a. State the decision about H_0.
   b. State a conclusion about H_1.

6. Decision:

a) State the decision about H_0.

Decision about H_0 : Fail to reject H_0.

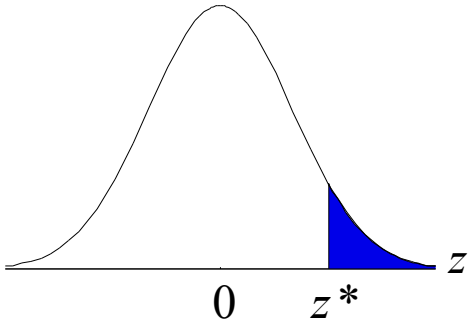
b) Write a conclusion about H_1.

There is not sufficient evidence at the 0.05 level of significance to show that the mean weight of cereal boxes is less than 24 ounces.

Finding *p*-values:

1. H_1 contains $>$ (Right tail)

p-value = $P(z > z^*)$



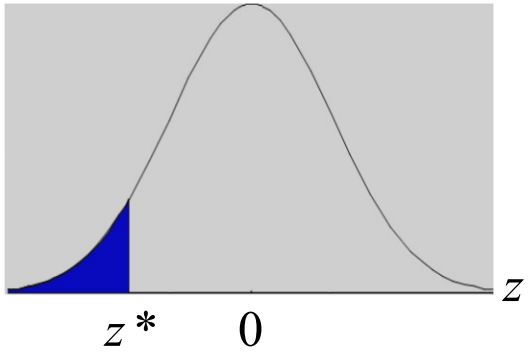
2. H_1 contains $<$ (Left tail)

p-value = $P(z < z^*)$

3. H_1 contains (Two-tailed)

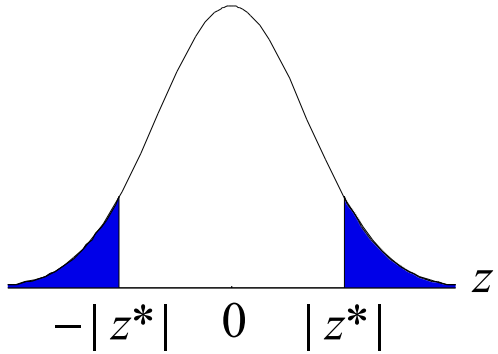
p-value = $P(z < -|z^*|) + P(z > |z^*|)$

$= 2 \times P(z > |z^*|)$



Note:

1. If we fail to reject H_0, there is no evidence to suggest the null hypothesis is false. This does not mean H_0 is true.
2. The p-value is the area, under the curve of the probability distribution for the test statistic, that is more extreme than the calculated value of the test statistic.
3. There are 3 separate cases for p-values. The direction (or sign) of the alternative hypothesis (H_1) is the key.



Another Approach for the Hypothesis Test of Mean μ (σ Known)

- Determine the critical region(s) and critical value(s).
- Determine the critical region(s) and critical value(s) calculated test statistic is in the critical region.

Decision Criteria

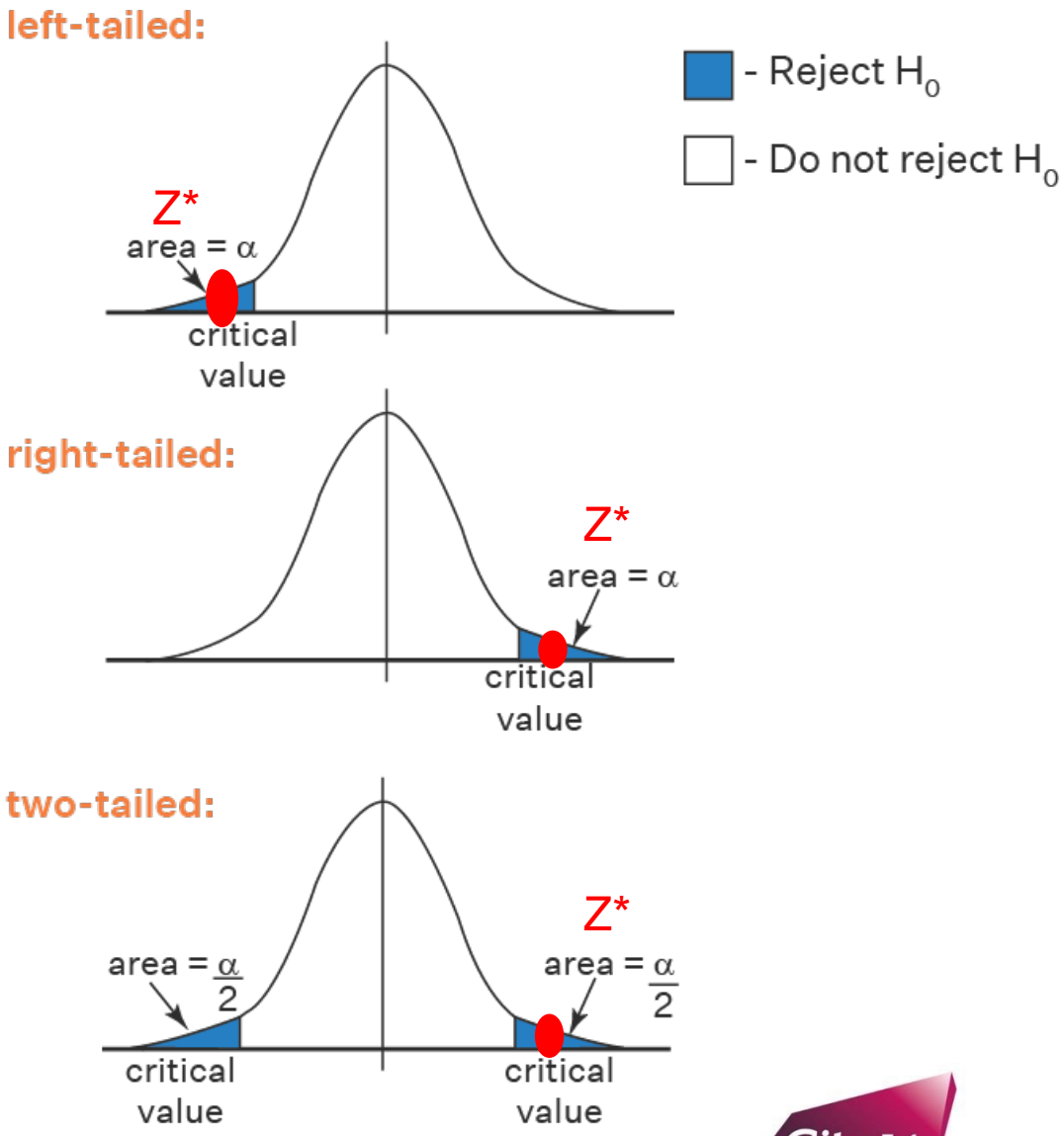
Test statistic $$z^* = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

- **Reject** the null hypothesis if test statistic < Z critical value (left-tailed hypothesis test) (data outside the acceptable region or in the rejection region).

(data is more extreme than the threshold)

- **Reject** the null hypothesis if test statistic > Z critical value (right-tailed hypothesis test) (data outside the acceptable region or in the rejection region).

- **Reject** the null hypothesis if the test statistic does not lie in the acceptance region/ in the rejection region (two-tailed hypothesis test).



Interference about Mean μ (σ unknown)

What about σ Unknown???

Hypothesis-Testing Procedure:

1. The t-statistic is used to complete a hypothesis test about a population mean μ.

2. The test statistic:

$$t^* = \frac{\bar{x} - \mu}{s/\sqrt{n}} \; with \; df = n - 1$$

3. The calculated t is the number of estimated standard errors of $\bar{x}$ from the hypothesized mean μ.

L08: Comparing Two Populations or Treatments

Before developing inferential procedures concerning $\mu 1 - \mu 2$, we must consider how the two samples, one from each population, are selected.

Dependent Sampling:

The same set of sources or related sets are used to obtain the data representing both populations.

Independent Sampling:

Two unrelated sets or sources are used, one set from each population.

Inferences concerning the Mean Difference using Two Dependent Samples

Paired difference: $d = x_1 - x_2$

Confidence Interval:

The confidence interval for estimating the mean difference μ_d is found using the formula:

$$\bar{d} \pm (t\ critical\ value)\frac{s_d}{\sqrt{n}}$$

Where $\bar{d}$ is the mean of the sample differences: $\bar{d} = \dfrac{\sum d}{n}$

and s_d is the standard deviation of the sample differences:

$$s_d = \sqrt{\frac{\sum d^2 - \left[\dfrac{\left(\sum d\right)^2}{n}\right]}{n-1}}$$

Hypothesis test for Comparing Two Population

Null Hypothesis: H_0: μ_d = hypothesized value

The hypothesized value is usually 0 – meaning that there is no difference.

Where μ_d is the mean of the differences in the paired observations

Test Statistic:

$$t^* = \frac{\bar{x}_d - \text{hypothesized value}}{s_d / \sqrt{n}}$$

Where n is the number of sample differences and x_d and s_d are the mean and standard deviation of the sample differences.

This test is based on df = $n - 1$.

Alternative Hypothesis:

H_a: $\mu_d >$ hypothesized value

H_a: $\mu_d <$ hypothesized value

H_a: $\mu_d \neq$ hypothesized value

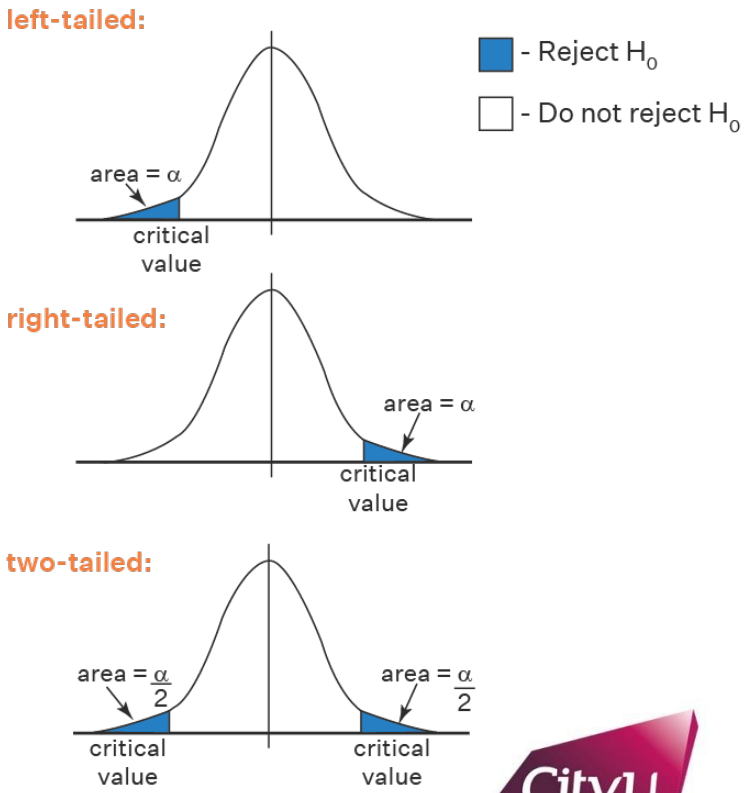
P-value:

Area to the right of calculated t

Area to the left of calculated t

2(area to the right of t) if $+t$ or

2(area to the left of t) if $-t$



Inferences concerning the Difference between Means using Two Independent Samples

- Inferences based on $\overline{x_1} - \overline{x_2}$

Varies from sample to sample

→ sampling distribution

Properties of the Sample Distribution of $\bar{x}_1$-$\bar{x}_2$

If the random samples on which $\bar{x}_1$ and $\bar{x}_2$ are based are selected independently of one another, then

1. $$\mu_{\bar{x}_1-\bar{x}_2} = \begin{pmatrix}\text{mean value}\\ \text{of } \bar{x}_1 - \bar{x}_2\end{pmatrix} = \mu_{\bar{x}_1} - \mu_{\bar{x}_2} = \mu_1 - \mu_2$$

   The sampling distribution of $\bar{x}_1 - \bar{x}_2$ is always centered at the value of $\mu_1 - \mu_2$, so $\bar{x}_1 - \bar{x}_2$ is an unbiased statistic for estimating $\mu_1 - \mu_2$.

2. $$\sigma^2_{\bar{x}_1-\bar{x}_2} = \begin{pmatrix}\text{variance of}\\ \bar{x}_1 - \bar{x}_2\end{pmatrix} = \sigma^2_{\bar{x}_1} + \sigma^2_{\bar{x}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

   and

   $$\sigma_{\bar{x}_1-\bar{x}_2} = \begin{pmatrix}\text{standard deviation}\\ \text{of } \bar{x}_1 - \bar{x}_2\end{pmatrix} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Properties of the Sample Distribution of $\bar{x}_1$-$\bar{x}_2$

3. If n_1 and n_2 are both large or the population distributions are (at least approximately) normal, $\bar{x}_1$ and $\bar{x}_2$ each have (at least approximately) a normal distribution.

This implies that the sampling distribution of $\bar{x}_1$ - $\bar{x}_2$ is also normal or approximately normal.

→can be standardized to obtain a variable with a sampling distribution that is approximately to the standard normal distribution.

$$z = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

When $\sigma 1\ and\ \sigma 2$ are known

$$t = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

When $\sigma 1\ and\ \sigma 2\ are\ un$known, we must estimate them using the corresponding sample variances, s_1^2 and s_2^2

The Two-Sample t Confidence Interval

The general formula for a confidence interval for : $\mu_1 - \mu_2$ when

1) The two samples are independently selected random samples from the populations of interest
2) The sample sizes are large (generally 30 or larger) or the population distributions are (at least approximately) normal.

is

$$(\bar{x}_1 - \bar{x}_2) \pm (t \text{ critical value})\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$\text{df} = \frac{(V_1 + V_2)^2}{\frac{V_1^2}{n_1 - 1} + \frac{V_2^2}{n_2 - 1}} \quad \text{where } V_1 = \frac{s_1^2}{n_1} \text{ and } V_2 = \frac{s_2^2}{n_2}$$

where df is the **smaller** of df_1 or df_2 when computing t^* without the aid of a computer or calculator.

Two-Sample *t* Test for Comparing Two Populations

Null Hypothesis: H_0: $\mu_1 - \mu_2$ = hypothesized value

Alternative Hypothesis:

The hypothesized value is often 0

H_a: $\mu_1 - \mu_2 >$ hypothesized value

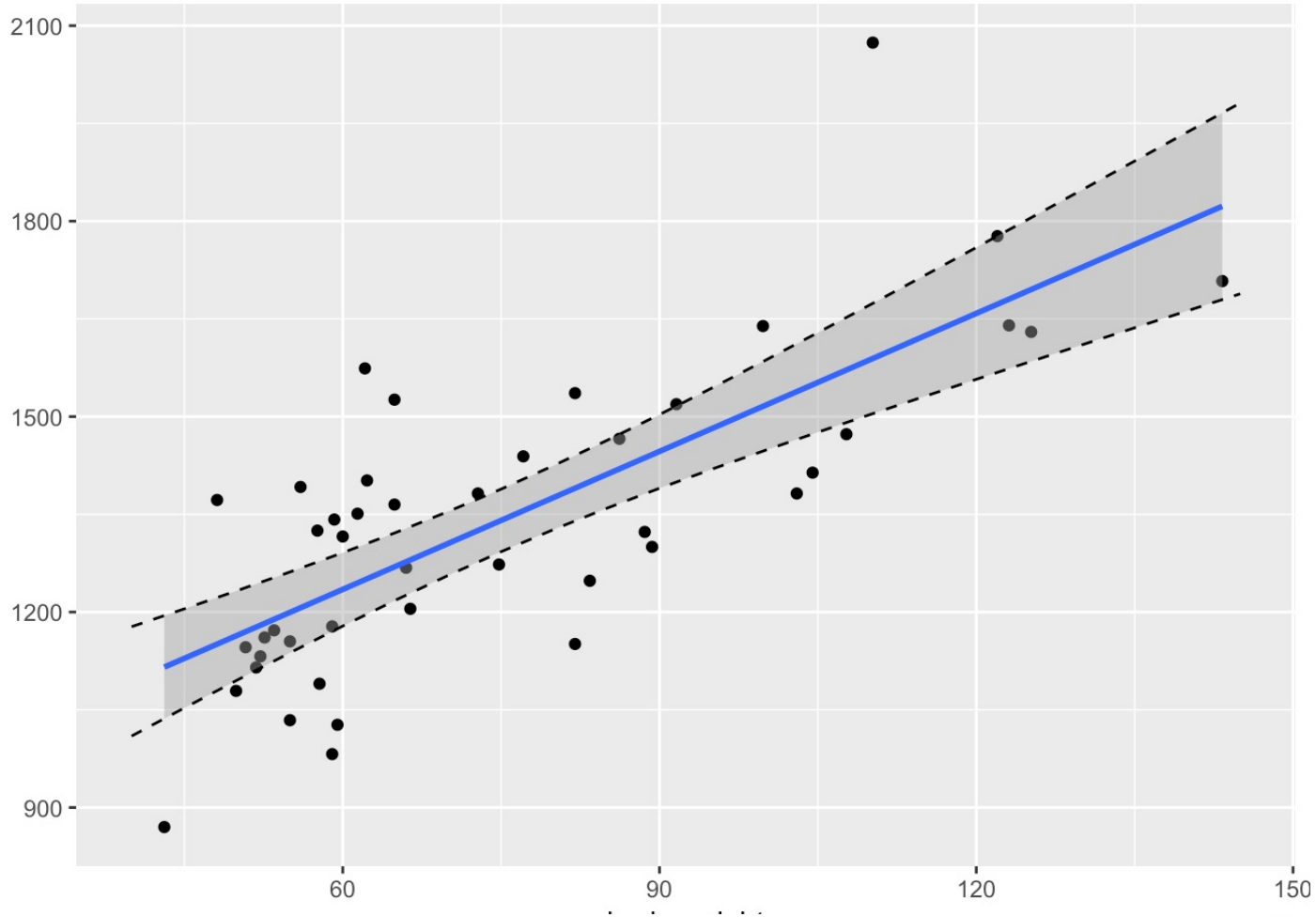
H_a : $\mu_1 - \mu_2 <$ hypothesized value

H_a: $\mu_1 - \mu_2 \neq$ hypothesized value

Test statistic: $$t = \frac{\bar{x}_1 - \bar{x}_2 - \text{hypothesized value}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

L09: Linear Regression

- Simple linear regression analysis



Estimating the Regression Line

Least Squares Approach

- "LEAST SQUARES approach" is the most commonly used method to find the straight line which is close to the data in statistics.

We estimate the true population regression line.

$$\hat{y} = a + bx$$

b = point estimate of β = $\dfrac{S_{xy}}{S_{xx}}$

a = point estimate of α = $\bar{y} - b\bar{x}$

$$b = \frac{\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)/n}{\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2/n} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{XY}}{S_{XX}}$$

$$a = \bar{y} - b\bar{x}$$

$$SXX = \Sigma\,(x - \bar{x})^2 = \Sigma\,x^2 - \frac{(\Sigma\,x)^2}{n} \qquad \text{(Sum Squares X)}$$

$$SYY = \Sigma(y - \bar{y})^2 = \Sigma\,y^2 - \frac{(\Sigma\,y)^2}{n} \qquad \text{(Sum Squares Y)}$$

$$SXY = \Sigma(x - \bar{x})(y - \bar{y}) = \Sigma\,xy - \frac{(\Sigma x)(\Sigma y)}{n} \qquad \text{(Sum Products X,Y)}$$

Quiz 5, Q1

Q1. The following table shows three pairs of observations on X and Y where Y is the observed percentage yield of a biological reaction at various centigrade temperatures, X. Obtain the least-squares line of regression of Y on X.

X (°C)	150	175	200
Y (W/m^2)	75.4	79.4	82.1

$$\sum_{i=1}^{3} X_i = 525, \quad \sum_{i=1}^{3} X_i^2 = 93125$$

$$\sum_{i=1}^{3} Y_i = 236.9, \quad \sum_{i=1}^{3} Y_i^2 = 18729.93$$

$$\sum_{i=1}^{3} X_i Y_i = 41625$$

And then by the least-squares approach we have

$$b = \frac{\sum_{i=1}^{n} X_i Y_i - (\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} Y_i)/n}{\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2/n} = \frac{41625 - (525)(236.9)/3}{93125 - (525)^2/3} = 0.134$$

$$a = \bar{Y} - b\bar{X} = \frac{236.9}{3} - 0.134\left(\frac{525}{3}\right) = 55.517$$

Finally, we can write down the fitted regression line

$$\hat{y} = 55.517 + 0.134x$$

Confidence Intervals for β_0 and β_1

Recall the general equation for the CI:

$$CI \text{ for } \mu = \bar{x} \pm t_{\alpha/2,df} \times SE_{\bar{x}}$$

- $\bar{x}$ is the sample mean,
- $t_{\alpha/2,df}$ is the t-value from the t-distribution for the desired confidence level and degrees of freedom (df),
- $SE_{\bar{x}}$ is the standard error of the mean, which is calculated as the sample standard deviation (s) divided by the square root of the sample size (n):

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

Similarly, The confidence interval for β_1, the slope in linear regression, can be determined

$$CI \text{ for } \beta_1 = \hat{\beta}_1 \pm t_{\alpha/2,df} \times SE(\hat{\beta}_1)$$

$$SE_{\beta 1} = \sqrt{\frac{MSE}{S_{XX}}} = \sqrt{\frac{S^2}{S_{XX}}}$$

MSE: Mean Squared Error

$$MSE = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Consequently, **the 100(1- α)% C.I. for β_1 is given by**

$$b \pm t_{n-2,\frac{\alpha}{2}}\sqrt{\frac{S^2}{S_{XX}}}$$

$$S^2 = \frac{\sum_{i=1}^{n} E_i^2}{n-2} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}$$

Confidence Intervals for β_0 and β_1

The confidence interval for β_0, the y-intercept in linear regression, can be determined using a similar approach to that of the slope.

$$CI \text{ for } \beta_0 = \hat{\beta}_0 \pm t_{\alpha/2,df} \times SE(\hat{\beta}_0)$$

$$SE_{\beta 0} = \sqrt{\frac{S^2 \sum_{i=1}^{n} x_i^2}{nS_{XX}}}$$

Consequently, **the 100(1- α)% C.I. for β_0 is given by**

$$(\bar{y} - b\bar{x}) \pm t_{n-2,\frac{\alpha}{2}} \sqrt{\frac{S^2 \sum_{i=1}^{n} x_i^2}{nS_{XX}}}$$

Example

An experiment to study the relationship between x = time spent exercises (minutes) and y = amount of oxygen consumed during the exercise period resulted in the following summary statistics.

$$n = 20, \sum x = 50, \sum y = 16705, \sum x^2 = 150, \sum y^2 = 14194231, \sum xy = 44194$$

a) Estimate the slope and y intercept of the population regression line

b) One sample observation on oxygen usage was 757 for a 2-minute exercise period. What amount of oxygen consumption would you predict for this exercise period, and what is the corresponding residual?

c) Compute a 99% confidence interval for the average change in oxygen consumption associated with 1-minute increase in exercise time.

c) $r = \dfrac{S_{XY}}{\sqrt{S_{XX}S_{YY}}}$

$$S_{XX} = \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2/n$$

$$S_{YY} = \sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2/n$$

$r^2 = 0.98$

$S^2 = 271.781$

$S = 16.486$

$$MSE = S^2 = \frac{S_{YY} - bS_{XY}}{n-2}$$

Solution

a)

$$S_{XY} = \sum_{i=1}^{n}(x_i y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)/n$$

$$= 44194 - \frac{50 \times 16705}{20} = 2431.5$$

$$S_{XX} = \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2/n = 150 - \frac{50 \times 50}{20} = 25$$

$$\bar{x} = 2.5;\ \bar{y} = 835.25$$

$$b = \frac{S_{XY}}{S_{XX}} = \frac{2431.5}{25} = 97.26$$

$$a = \bar{y} - b\bar{x} = 835.25 - 97.26 \times 2.5 = 592.1$$

b)

$$\hat{y} = bx + a = 97.26 \times 2 + 592.1 = 786.62$$

$$y - \hat{y} = 757 - 786.62 = -29.62$$

Calculation of the 99% confidence interval for β requires a t critical value based on df = n - 2 = 20 – 2 = 18 → 2.878

$$b \pm (t\ critical\ value) \times \frac{S}{\sqrt{S_{XX}}} = 97.26 \pm 2.878 \times \frac{16.486}{\sqrt{25}} = 106.749\ and\ 87.771$$

Two sides

t distribution critical values

	Upper-tail probability p								
df	.25	.20	.15	.10	.05	.025	.02	.01	.005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921
17	0.689	0.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898
18	0.688	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878

Hypothesis Testing for β_0 and β_1

- **For the slope β_1 ,**

1. One-sided right test:

Consider $H_0 : \beta_1 = b_1$.

$H_1 : \beta_1 > b_1$

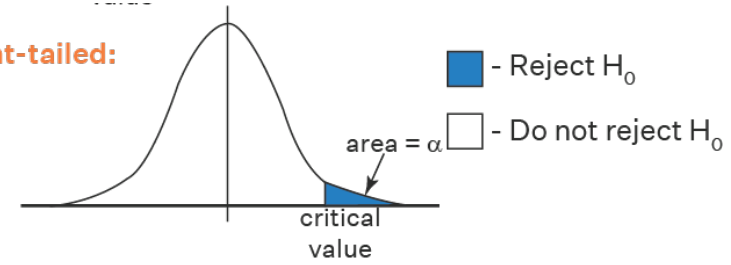
Test statistics

$$t = \frac{\hat{\beta}_j - b_1}{SE(\hat{\beta}_j)}$$

$$SE_{\beta 1} = \sqrt{\frac{MSE}{S_{XX}}} = \sqrt{\frac{S^2}{S_{XX}}}$$

Reject H_0 at a significance level α if

the t value $\frac{b - b_1}{\frac{s}{\sqrt{S_{XX}}}} > t_{n-2,\alpha}$



(when σ_x^2 is UNKNOWN)

Lies in the critical region → reject H_o

Hypothesis Testing for β_0 and β_1

2. One-sided left test:

Consider $H_0: \beta_1 = b_1$
$H_1: \beta_1 < b_1$.

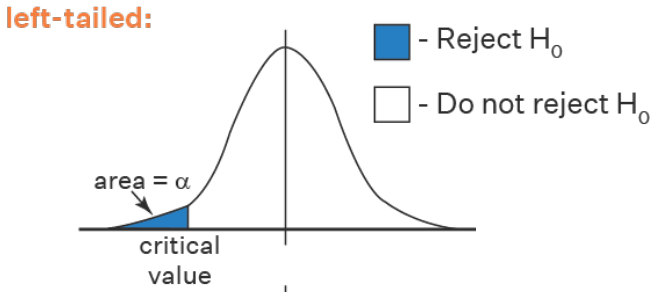
Test statistics

$$t = \frac{\hat{\beta}_j - b_1}{SE(\hat{\beta}_j)}$$

$$SE_{\beta 1} = \sqrt{\frac{MSE}{S_{XX}}} = \sqrt{\frac{S^2}{S_{XX}}}$$

Reject H_0 at a significance level α if

the t value $\dfrac{b - b_1}{\frac{s}{\sqrt{S_{XX}}}} < -t_{n-2,\alpha}$



(when σ_x^2 is UNKNOWN)

Lies in the critical region → reject H_o

Hypothesis Testing for β_0 and β_1

3. Two-sided test:

Consider $H_0: \beta_1 = b_1$
$H_1: \beta_1 \neq b_1$.

Test statistics

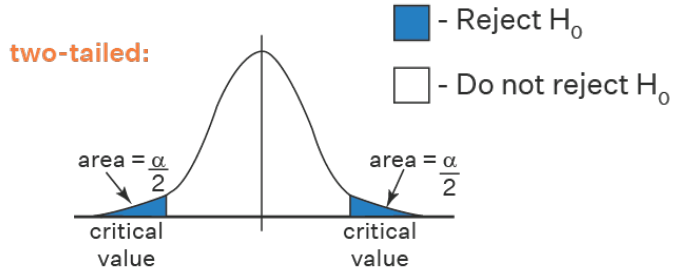
$$t = \frac{\hat{\beta}_j - b_1}{SE(\hat{\beta}_j)}$$

$$SE_{\beta 1} = \sqrt{\frac{MSE}{S_{XX}}} = \sqrt{\frac{S^2}{S_{XX}}}$$

Reject H_0 at a significance level α if

the absolute t value $\left|\frac{b-b_1}{\frac{s}{\sqrt{S_{XX}}}}\right| > t_{n-2,\frac{\alpha}{2}}$

(when σ_x^2 is UNKNOWN)



Lies in the critical region → reject H_o

Quiz 5, Q2

Q2. An article in The Journal of Clinical Endocrinology and Metabolism ["Simultaneous and Continuous 24-Hour Plasma and Cerebrospinal Fluid Leptin Measurements: Dissociation of Concentrations in Central and Peripheral Compartments" (2004, Vol. 89, pp. 258–265)] reported on a study of the demographics of simultaneous and continuous 24-hour plasma and cerebrospinal fluid leptin measurements. The data follow:

y=BMI (kg/m^2)	19.92	20.59	29.02	20.78	25.97	20.39	23.29	17.27	35.24
x=age (yr)	45.5	34.6	40.6	32.9	28.2	30.1	52.1	33.3	47.0

a) Estimate the slope and y intercept of the population regression line.

b) Please test whether linear relationship between age and BMI is significant under $\alpha = 0.01$.

Solutions:

a) To fit a linear regression model compute the following quantities. Note that $n = 9$.

$$\sum_{i=1}^{n} x_i = 344.3$$

$$\sum_{i=1}^{n} y_i = 212.47$$

$$\bar{x} = 38.2556$$

$$\bar{y} = 23.6078$$

$$\sum_{i=1}^{n} x_i^2 = 13731.7$$

$$\sum_{i=1}^{n} y_i^2 = 5267.45$$

$$\sum_{i=1}^{n} x_i y_i = 8271.52$$

Now, find S_{xx} and S_{xy}

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n} = 13731.7 - \frac{344.3^2}{9} = 560.3122$$

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n} = 8271.52 - \frac{344.3 \times 212.47}{9} = 143.3621$$

Find the least squares estimates of the slope and intercept.

$$b = \frac{S_{xy}}{S_{xx}} = \frac{143.3621}{560.3122} = 0.2559$$

$$a = \bar{y} - b\bar{x} = 23.6078 - (0.2559)(38.2556) = 13.8182$$

b) We will use a significance level of 0.01 to carry out this test.

To test whether linear relationship between age and BMI is significant:

H_0: $\beta_1 = 0$

H_1: $\beta_1 \neq 0$

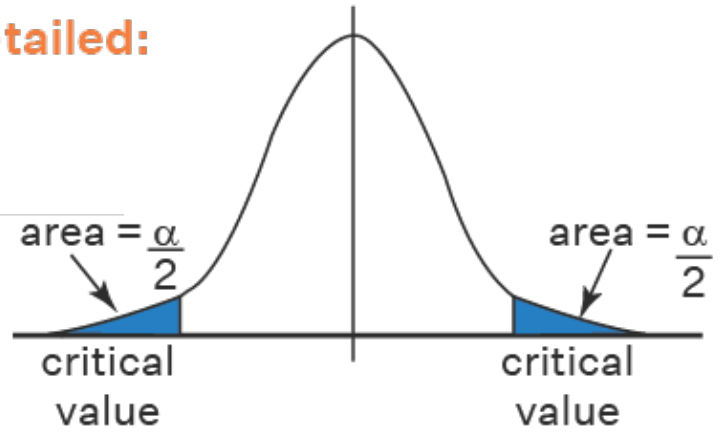
$\alpha = 0.01$

$$S_{YY} = \sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2/n = 5267.45 - \frac{212.47^2}{9} = 251.51$$

$$MSE = S^2 = \frac{S_{YY} - bS_{XY}}{n-2} = \frac{251.51 - 0.2559(143.3621)}{7} = 30.69$$

Test statistic: $t^* = \frac{b-b_1}{\frac{s}{\sqrt{S_{XX}}}} = \frac{0.2559-0}{\frac{5.54}{\sqrt{560.3122}}} = 1.09$

two-tailed:



From *t*-statistical table, we get:

Critical value: $t(7,0.005)=3.499$

Since t^* does not lie in the critical region, fail to reject H_0 at a significance level $\alpha =$ 0.01. We can thus conclude that there is no strong evidence to say that linear relationship between age and BMI is significant.

Hypothesis Testing for β_0 and β_1

- **Similarly, for the intercept β_0 ,**

1. One-sided right test:

Consider $H_0: \beta_0 = b_0$
$H_1: \beta_0 > b_0$.

Test statistics

$$t = \frac{\hat{\beta}_j - b_0}{SE(\hat{\beta}_j)}$$

$$SE_{\beta 0} = \sqrt{\frac{S^2 \sum_{i=1}^{n} x_i^2}{nS_{XX}}}$$

Reject H_0 at a significance level α if

$$\text{the } t \text{ value } \frac{a - b_0}{s\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{nS_{XX}}}} > t_{n-2,\alpha}$$

(when σ_x^2 is UNKNOWN)

Lies in the critical region → reject H_o

Hypothesis Testing for β_0 and β_1

2. One-sided left test:

Consider $H_0: \beta_0 = b_0$
$H_1: \beta_0 < b_0$

Test statistics

$$t = \frac{\hat{\beta}_j - b_0}{SE(\hat{\beta}_j)}$$

$$SE_{\beta 0} = \sqrt{\frac{S^2 \sum_{i=1}^{n} x_i^2}{nS_{XX}}}$$

Reject H_0 at a significance level α if

$$\text{the } t \text{ value } \frac{a - b_0}{s\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{nS_{XX}}}} < -t_{n-2,\alpha}$$

(when σ_x^2 is UNKNOWN)

Lies in the critical region → reject H_o

Hypothesis Testing for β_0 and β_1

3. Two-sided test:

Consider $H_0: \beta_0 = b_0$

$H_1: \beta_0 \neq b_0$

Test statistics

$$t = \frac{\hat{\beta}_j - b_0}{SE(\hat{\beta}_j)}$$

$$SE_{\beta 0} = \sqrt{\frac{S^2 \sum_{i=1}^{n} x_i^2}{nS_{XX}}}$$

Reject H_0 at a significance level α if

the absolute t value $\left| \frac{a - b_0}{s\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{nS_{XX}}}} \right| > t_{n-2,\frac{\alpha}{2}}$

(when σ_x^2 is UNKNOWN)

Lies in the critical region → reject H_o

Statistical relationship between two variables

Statistical relationship between two variables

Pearson correlation coefficient, which measures the strength and direction of the linear relationship between two continuous variables.

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

$$r = \frac{S_{XY}}{\sqrt{S_{XX} S_{YY}}}$$

- r=1: A perfect positive linear relationship.
- r=−1: A perfect negative linear relationship.
- r=0: No linear relationship.

- 0<r<1: A positive linear relationship.
- −1<r<0: A negative linear relationship.

Short Summary

1. S_{XX} is the sum of the squares of the difference between each *x* and the mean *x* value.

$$S_{XX} = \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2/n$$

2. S_{YY} is the sum of the squares of the difference between each *y* and the mean *y* value.

$$S_{YY} = \sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2/n$$

3. S_{xy} is sum of the product of the difference between *x* its means and the difference between *y* and its mean.

$$S_{XY} = \sum_{i=1}^{n}(x_i y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)/n$$

4. **Mean squared error (MSE)** measures the average of the squares of the errors—that is, the average squared difference between the estimated values and what is estimated.

$$MSE = S^2 = \frac{S_{YY} - bS_{XY}}{n-2}$$

5. ***r*** means a statistical relationship between two variables.

$$r = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}}$$

6. *a* and *b* are called the least-squares ESTIMATES of the unknown true values of β_0 (α) and β_1 (β), respectively.

$$b = \frac{S_{XY}}{S_{XX}}$$

$$a = \bar{y} - b\bar{x}$$

L10: Analysis of Variance

An ANOVA, short for “Analysis of Variance”, is used to determine whether or not there is a statistically significant difference between the means of three or more independent groups.

One-way ANOVA Tests:

We can summarize the ANOVA test as follows:

H0: $\mu_1 = \mu_2 = \cdots = \mu_k$, $(i.e.,\ no\ group\ effect)$ vs
H1: at least two of the means are not equal.

Briefly, the mathematical procedure behind the ANOVA test is as follow:

1. Compute the **variance between group means**.
2. Compute the **within-group variance**, also known as **residual variance**.

3.Produce the F-statistic as the ratio of variance.between.groups/variance.within.groups

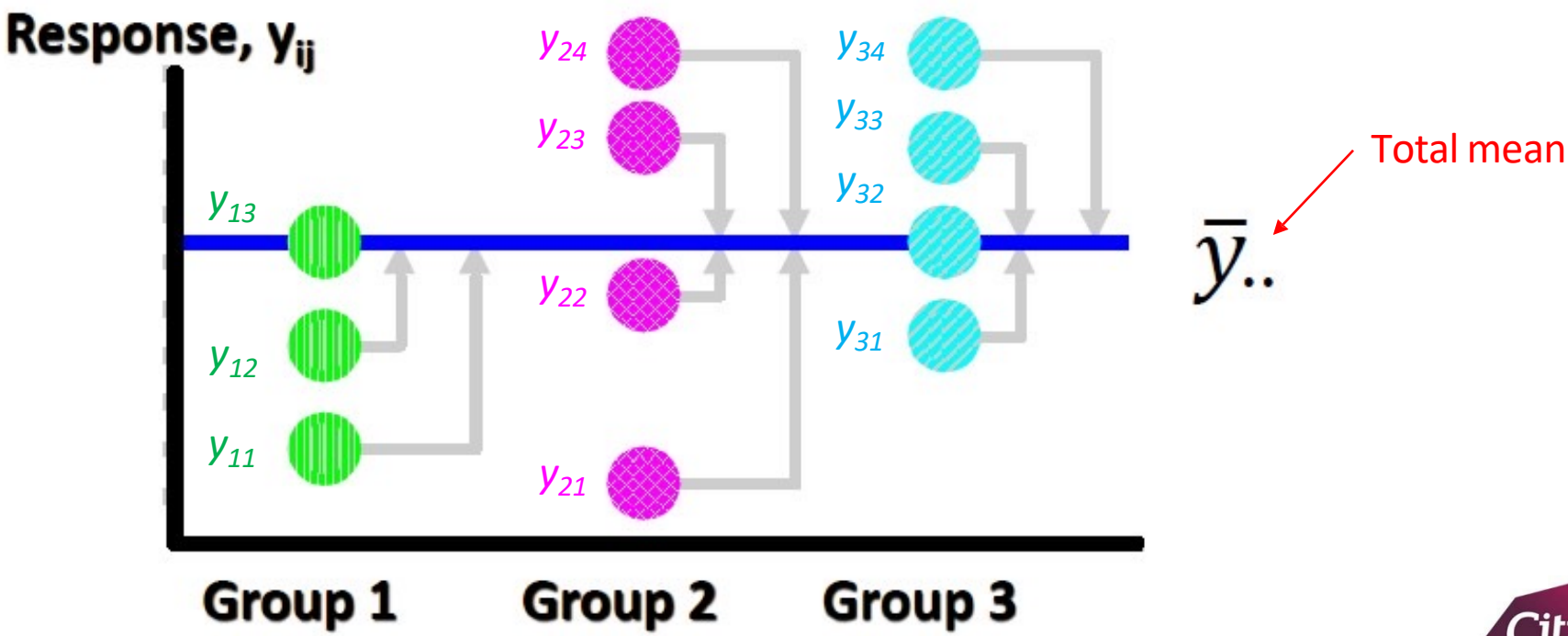
Total Variation

Total Sum of Squares

$$\text{SST} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2$$

For $k = 3$, $n_1 = 3$, $n_2 = 4$ and $n_3 = 4$.

$$SST = (y_{11} - \bar{y}_{..})^2 + (y_{12} - \bar{y}_{..})^2 + (y_{13} - \bar{y}_{..})^2 + (y_{21} - \bar{y}_{..})^2 + \cdots + (y_{33} - \bar{y}_{..})^2 + (y_{34} - \bar{y}_{..})^2$$



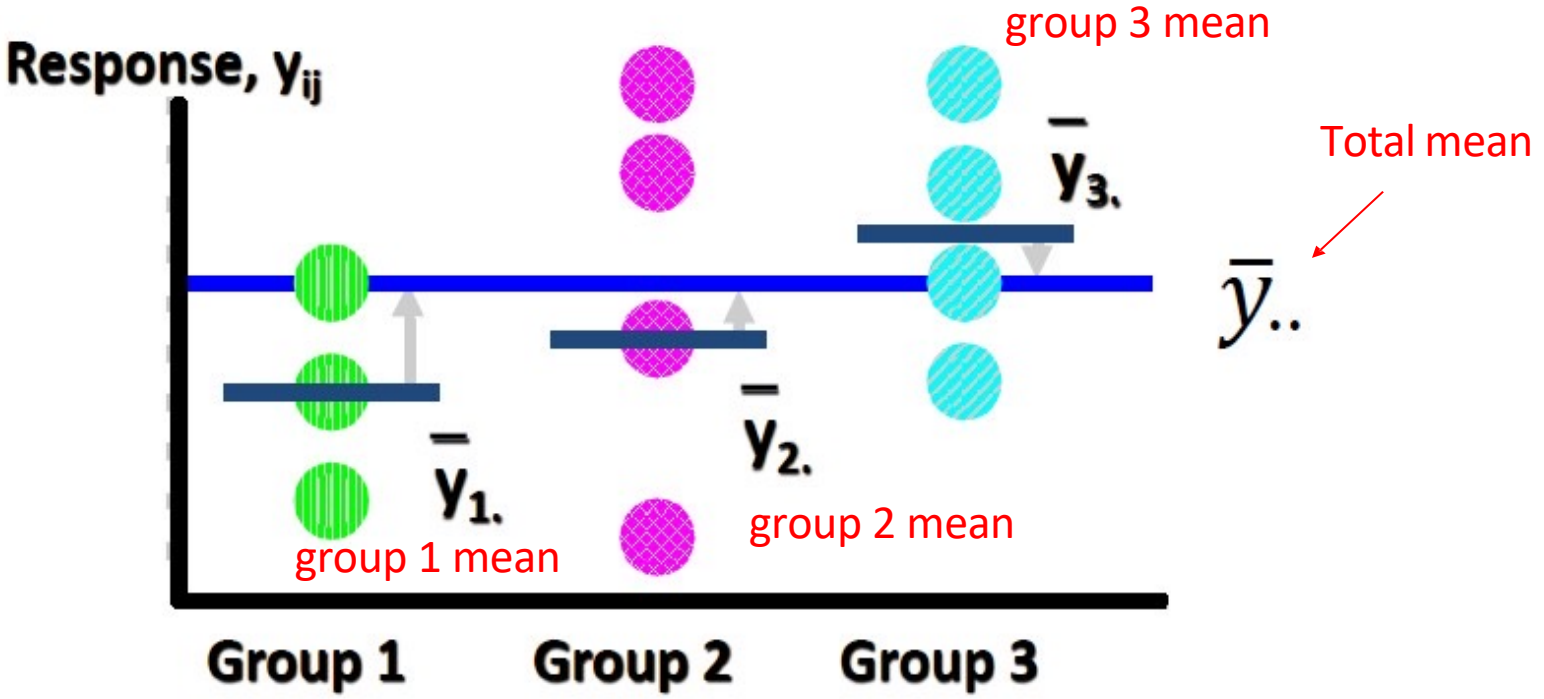
Total mean

$\bar{y}_{..}$

MAIN idea of ANOVA

- Between group variation (or the treatment sum of squares),

$$SS_{Treat} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 = \sum_{i=1}^{k} n_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$$



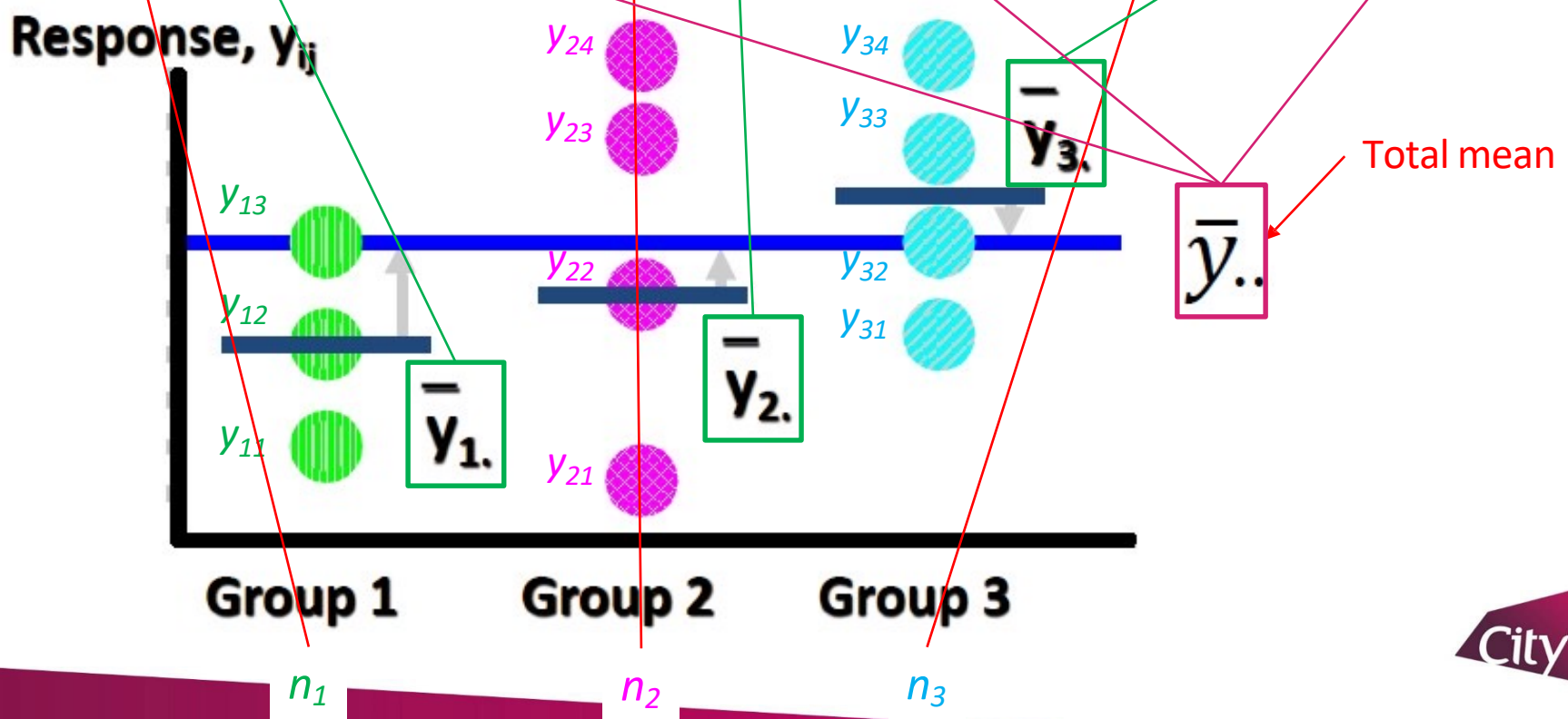
Between-group Variation

Treatment Sum of Squares

$$SS_{Treat} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 = \sum_{i=1}^{k} n_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$$

For k = 3, n_1= 3, n_2= 4 and n_3= 4.

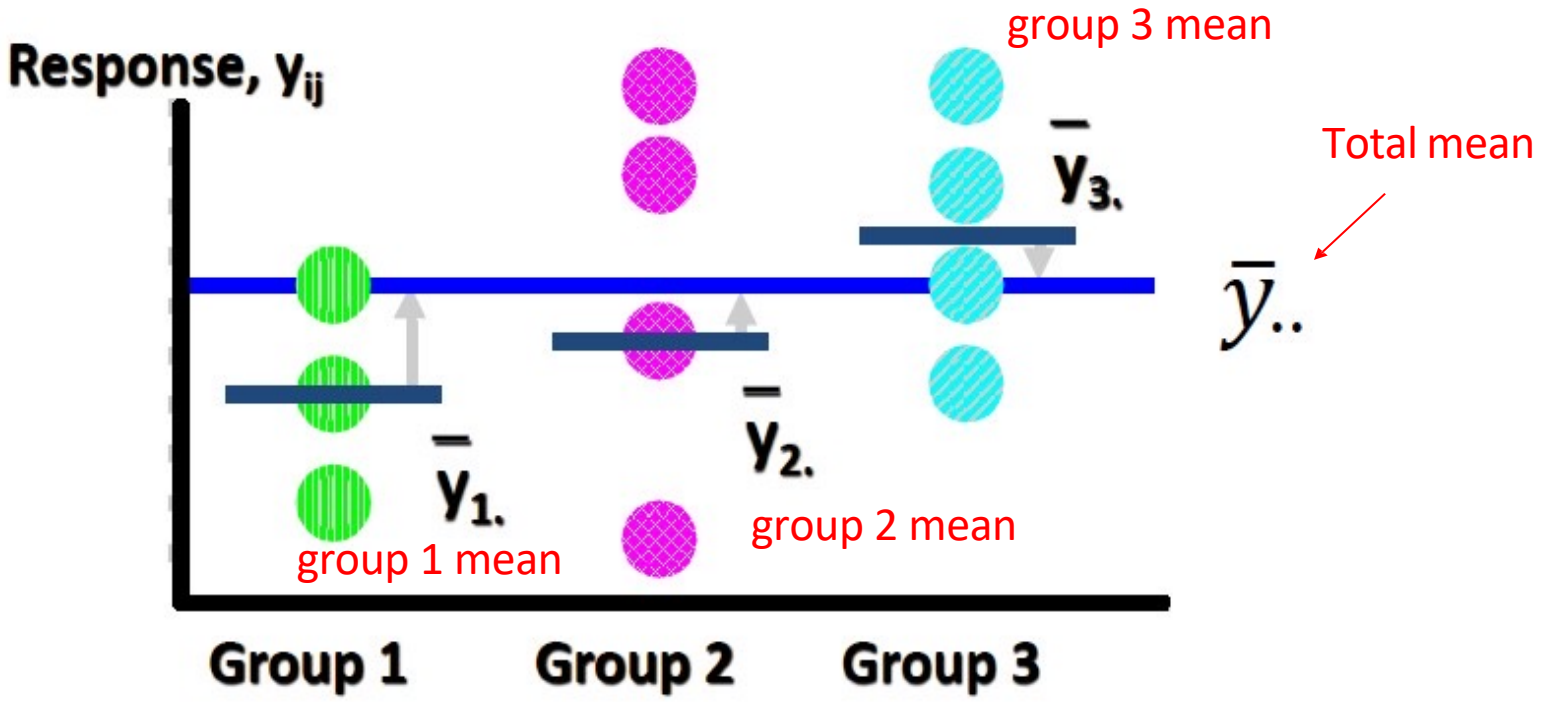
$$SS_{Treat} = n_1(\bar{y}_{1\cdot} - \bar{y}_{\cdot\cdot})^2 + n_2(\bar{y}_{2\cdot} - \bar{y}_{\cdot\cdot})^2 + n_3(\bar{y}_{3\cdot} - \bar{y}_{\cdot\cdot})^2$$



MAIN idea of ANOVA

- Within group variation (or the error sum of squares),

$$\boxed{\text{SSE}} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\cdot})^2.$$



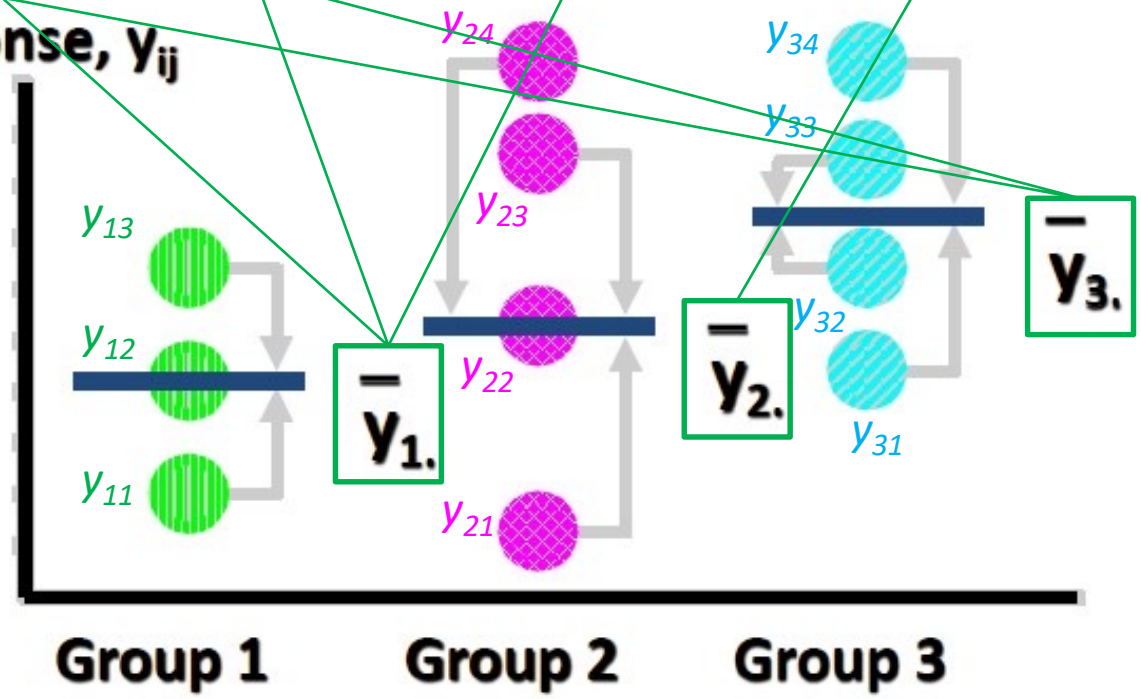
Within-group Variation

Error Sum of Squares

$$SSE = (n_1-1)s_1^2+(n_2-1)s_2^2+(n_3-1)s_3^2+\ldots+(n_k-1)s_k^2$$

For k = 3, n_1= 3, n_2= 4 and n_3= 4.

$$SSE = (y_{11}-\bar{y}_{1\cdot})^2+(y_{12}-\bar{y}_{1\cdot})^2+(y_{13}-\bar{y}_{1\cdot})^2+(y_{21}-\bar{y}_{2\cdot})^2+\cdots+(y_{33}-\bar{y}_{3\cdot})^2+(y_{34}-\bar{y}_{3\cdot})^2$$



MAIN idea of ANOVA

Other form of abbreviation:
SST=SSTo
$SS_{Treat} = SST_r$

The relationship between SST, SS_{Treat} , and SSE

$$SST = SS_{Treat} + SSE$$

Then, we can construct the following ANOVA table:

Source	d.f.	Sum of Squares (SS)	Mean sum of squares (MS)	F-value
Treatment/Between/Model	k-1	SS_{Treat}	$MS_{Treat} = SS_{Treat}/(k-1)$	$F = MS_{Treat}/MSE$
Error/Within	n-k	SSE	$MSE = SSE/(n-k)$	
Total	n-1	SST		

where

$$MS_{Treat} = \frac{SS_{Treat}}{k-1}$$

mean square treatment

and

$$MSE = \frac{SSE}{n-k}.$$

mean square error

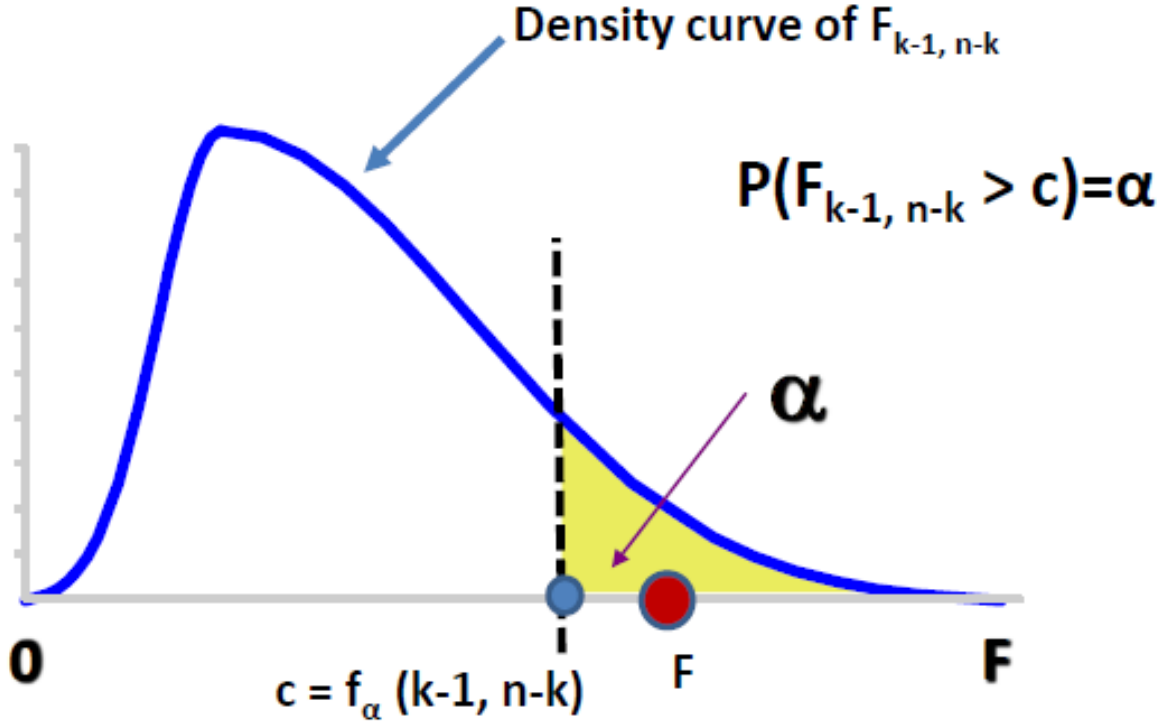
are two estimates of the common population variance σ^2

H0: $\mu_1 = \mu_2 = \cdots = \mu_k$, $(i.e.,\ no\ group\ effect)$ vs
H1: at least two of the means are not equal.

Under H_0 (i.e. all means are equal),

Reject H_0 at a significance level α if we get a large F, *say* $F > c$.

$$F = \frac{MS_{Treat}}{MSE}$$



The F-Distribution

$$F = \frac{MS_{Treat}}{MSE}$$

Each different combination of df_1 and df_2 produces a different F distribution.

$$MS_{Treat} = \frac{SS_{Treat}}{k-1} \quad \text{and} \quad MSE = \frac{SSE}{n-k}$$

df_1 = mean square treatment (Between group)

df_2 = mean square error (Within group)

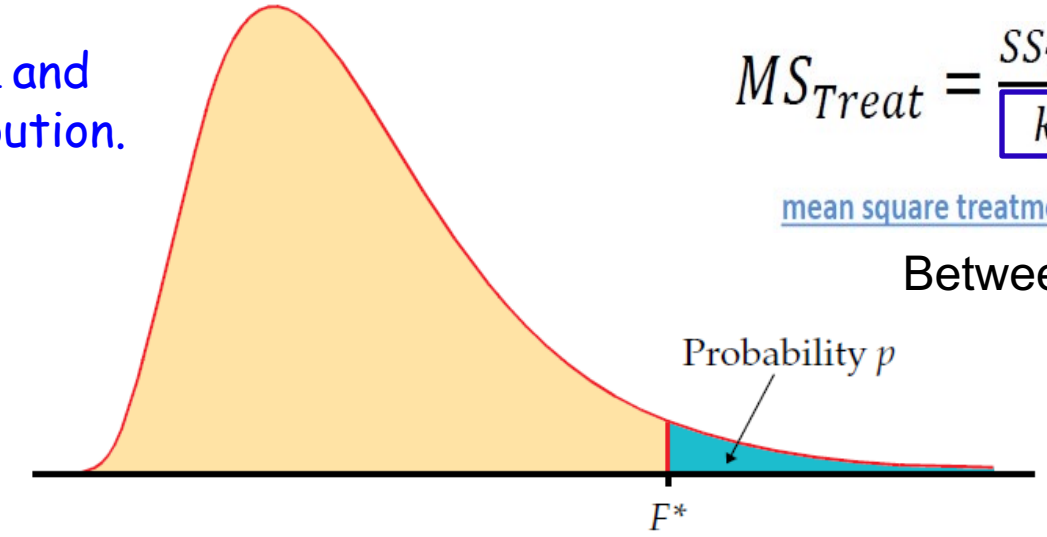


Table entry for p is the critical value F^* with probability p lying to its right.

K: number of groups

n: total number of observations is $n = n_1 + n_2 + \cdots + n_k$.

TABLE E

F critical values

		df_1 Degrees of freedom in the numerator								
	p	1	2	3	4	5	6	7	8	9
df_2	.100	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86
	.050	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54
1	.025	647.79	799.50	864.16	899.58	921.85	937.11	948.22	956.66	963.28
	.010	4052.2	4999.5	5403.4	5624.6	5763.6	5859.0	5928.4	5981.1	6022.5
	.001	405284	500000	540379	562500	576405	585937	592873	598144	602284