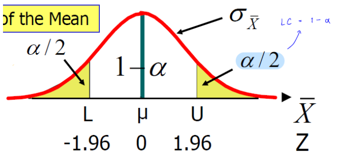
Confidence Interval (CI)
- Assuming data is normally distributed or if not use $n > 30$
- CI of 95% is that 95% of the population is in this CI meaning the population parameters (μ) also lies within this interval.
- Level of confidence (α), $\frac{\alpha}{2}$ for the proportion in the upper and lower tail
  - $95\%\ CI \rightarrow \alpha/2 = 0.025 \rightarrow Z = \pm 1.96$
  - $99\%\ CI \rightarrow \alpha/2 = 0.05 \rightarrow Z = \pm 2.58$
- For one side (Lower CI or Upper CI), just use $\boldsymbol{\alpha}$
  - $95\%\ Upper\ CI \rightarrow \alpha = 0.05 \rightarrow Z = 1.645$



CI – known σ
- Use Z-test.
- $\mu \le \bar{x} \pm Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \rightarrow \mu \in \left[\bar{x} - Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}},\ \bar{x} + Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right]$

  Lower confidence bound: $\mu \ge \bar{x} - Z_{\alpha} \cdot \frac{\sigma}{\sqrt{n}} \rightarrow [\mu,\ \infty)$

  Upper confidence bound: $\mu \le \bar{x} + Z_{\alpha} \cdot \frac{\sigma}{\sqrt{n}} \rightarrow (-\infty,\ \mu]$
- $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
- Determine sample size of mean. $n = \frac{Z^2_{\alpha/2} \cdot \sigma^2}{error^2}$

CI – unknown σ & small n ($n < 30$)
- Use T-test. $t = \frac{\bar{x}-\mu}{S/\sqrt{n}}$, with $DOF = (n-1)$
- $\mu \le \bar{x} \pm t_{\alpha/2, n-1} \cdot \frac{S}{\sqrt{n}}$
- T-distribution approach normal distribution when n is large.

CI – unknown σ & large n ($n \ge 30$)
- As n is large, we can assume that the population is normally distributed, so the Z-test is used. $Z = \frac{\bar{x}-\mu}{S/\sqrt{n}} \rightarrow noted$: when n is large, $\sigma \approx S = \sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}}$
- $\mu \le \bar{x} \pm Z_{\alpha/2} \cdot \frac{S}{\sqrt{n}}$

CI – population proportion
- Binary outcome (binomial distribution)
- Assumed normal distributed when $np \ge 5, n(1-p) \ge 5$, Z-test is used.
- $p \le p_s \pm Z_{\alpha/2} \cdot \sqrt{\frac{p_s(1-p_s)}{n}} \rightarrow p_s = \frac{n\ of\ success}{n\ total\ sample}$
- Determine sample size. $n = \frac{Z^2_{\alpha/2} \cdot p_s(1-p_s)}{error^2}$

Hypothesis Testing (1 sample)
- Idea is to try proving the null hypothesis (H_0) using a given sample.
- Test method.
  - Z-test:
    - known $\sigma \rightarrow Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$
    - proportion (p_s) $\rightarrow Z = \frac{p_s - p}{\sqrt{p(1-p)/n}} \rightarrow$ Noted: $\sigma_{ps} = \sqrt{\frac{p(1-p)}{n}}$, $\mu_{ps} = p$
  - T-test: unknown $\sigma \rightarrow t = \frac{\bar{x}-\mu}{S/\sqrt{n}}$
- If the value from test method (z, v) is not in the rejection zone created by the level of significance (α), H_0 is accepted or known as "There is not enough evidence to reject H_0"
- Critical value: value from level of significance (α) at the edge of rejection zone (~CI).
- Null hypothesis (H_0) is the hypothesis that contains "=". Ex. $\mu = 7, \mu \ge 9, \mu \le 12$. Alternate hypothesis ($H_1$) is the opposite of null hypothesis.
- Steps
  1. Set H_0
  2. **Cal for test statistics (z, t, f)**
  3. Determine region of rejection & critical value from α
  4. Compare value.
     a. Test stat: If test stat lies in region of rejection (RoR), H_0 is rejected.
     b. P-value: convert test stat into probability, if probability is less than α, H_0 is rejected. For two side, the P-value is doubled,

$H_0 : \mu \le 40$

$H_1 : \mu > 40$

$\alpha = 0.05$

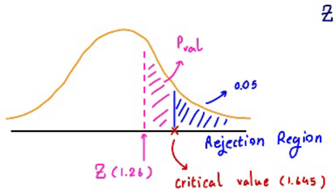
$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$

$= \frac{40.5 - 40}{1.25/\sqrt{10}}$

$= 1.26 \longrightarrow P_{val} = 1 - 0.8962$

$= 0.1038$

$P_{val} > 0.05$,
H_0 is not reject



$H_0 : \mu = 98.6$

$H_1 : \mu \ne 98.6$

$\alpha : 0.05$

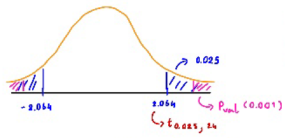
$t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$

$= \frac{98.264 - 98.6}{0.482/\sqrt{25}}$

$= -3.485$

$P_{val} = 2 \times 0.001$

$= 0.002$



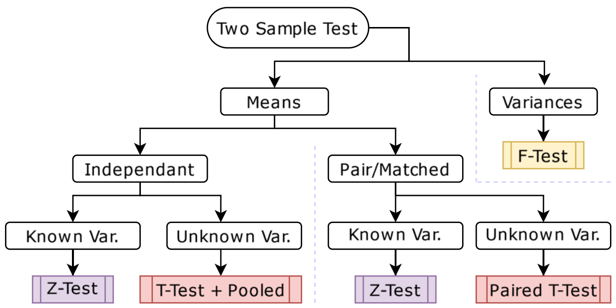
p = 0.002 which is lesser than 0.05, so H_0 is rejected.

$H_1 : \ne$, use two tails.

$H_1 : <$, one tail, RoR on left.

$H_1 : >$, one tail, RoR on right.

Hypothesis Testing (2 sample)

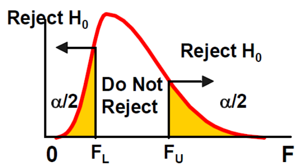


- **Criteria:** Assumed population is normally distributed or $n > 30$

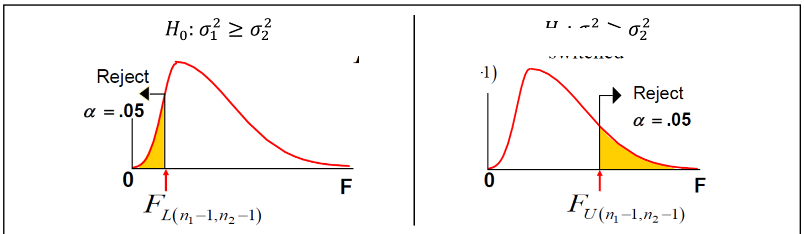
Mean – Independent Sample
- Variance known (Z-test) $\rightarrow Z = \frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$
- Variance unknown (Pooled T-test)
  - Population is normal or n > 30.
  - Assumed equal ($\sigma_1 = \sigma_2$)
  - Pooled sample variance $\rightarrow S_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{(n_1-1)+(n_2-1)}$
  - $t = \frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$, using $DOF = n_1 + n_2 - 2$

Mean – Related Sample (Paired/Match)
- Must observed paired (ex. Before & after)
- $n_1 = n_2 = n$ always
- Use the difference between pairs.
  - $D_i = x_{1,i} - x_{2,i}$
  - $\bar{D} = \frac{\sum D_i}{n}$
- Variance known $\rightarrow Z = \frac{\bar{D}-\mu_D}{\sigma_D/\sqrt{n}}$
- Variance unknown
  - $t = \frac{\bar{D}-\mu_D}{S_D/\sqrt{n}}$, using $DOF = n - 1$
  - $S_D = \sqrt{\frac{\sum(D_i-\bar{D})^2}{n-1}}$

Variance
- Both populations are normally distributed, this test is not robust to this violation.
- Use F-test: $F = \frac{S_1^2}{S_2^2}$
- $DOF_1 = n_1 - 1, DOF_2 = n_2 - 1$
- Critical value: $F_{L(DOF_1, DOF_2)} = \frac{1}{F_{U(DOF_2, DOF_1)}}$, $F_{U(DOF_1, DOF_2)}$
- If $F - value$ is between F_L & F_U, H_0 is not rejected.



Note: The F-test graph always starts from the right.



ANOVA

- Control one or more independent variables: treatment factors.
- Observe effects as dependent variables: response variable.
- Condition
  - Samples are randomly and independently drawn.
  - Population is normally distributed.
    - Less sensitive when n is the same.
  - Population variances are equal.
- $H_0: \mu_1 = \mu_2 = \cdots = \mu_n$
  - All μ is the same, no treatment effect → the factor has no effect on the response variable.
- H_1: not all μ are the same.
  - At least one μ is different, there are treatment effect.

- **Total variation** → $SST = \sum_{j=1}^{c}\sum_{i=1}^{n_j}\left(x_{ij} - \bar{\bar{x}}\right)^2$
  $\rightarrow (x_{11} - \bar{\bar{x}})^2 + (x_{12} - \bar{\bar{x}})^2 + \cdots + \left(x_{n_c c} - \bar{\bar{x}}\right)^2$
  - x_{ij}: data i in group j
  - n_j: number of observations in group j
  - c: number of groups.
  - $\bar{\bar{x}}$: overall mean $\left(\frac{sum\ of\ all\ x_{ij}}{total\ observation\ (n = c \cdot n_j)}\right)$
  - $SST = SSA + SSW$
  - $DOF = n - 1$

- **Among-Group variation** → $SSA = \sum_{j=1}^{c} n_j\left(\bar{x}_j - \bar{\bar{x}}\right)^2$
  $\rightarrow n_1(\bar{x}_1 - \bar{\bar{x}})^2 + n_2(\bar{x}_2 - \bar{\bar{x}})^2 + \cdots + n_c(\bar{x}_c - \bar{\bar{x}})^2$
  - $\bar{x}_j$: sample mean of group j.
  - $DOF = c - 1$
  - $MSA = \frac{SSA}{c-1}$

- **Within-Group variation** → $SSW = \sum_{j=1}^{c}\sum_{i=1}^{n_j}\left(x_{ij} - \bar{x}_j\right)^2$
  $\rightarrow (x_{11} - \bar{x}_1)^2 + (x_{12} - \bar{x}_1)^2 + \cdots + \left(x_{n_c c} - \bar{x}_c\right)^2$
  - $DOF = n - c$
  - $MSW = \frac{SSW}{n-c}$

- **F-test**: $f = \frac{MSA}{MSW}$, using $DOF_1 = c - 1, DOF_2 = n - c$
- Using f table if $f > F_{\alpha, DOF1, DOF2}$, reject H_0
- To find which μ is different, we use LSD.
  - Compare pairwise, find absolute mean difference (AMD) of each pair.
    Ex: $|\bar{x}_1 - \bar{x}_2|, |\bar{x}_1 - \bar{x}_3|, |\bar{x}_2 - \bar{x}_3|, \ldots$
  - Same sample size: $LSD = t_{\frac{\alpha}{2}, c(n_j - 1)} \cdot \sqrt{\frac{2MSW}{n_j}}$

    Different sample size: $LSD = t_{\frac{\alpha}{2}, n-c} \cdot \sqrt{MSW\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$
  - If AMD of a pair is larger than LSD, that pair have different μ.

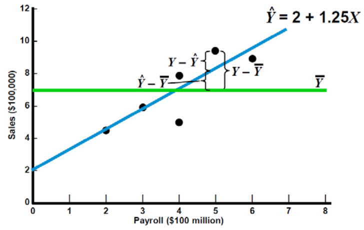
Two ways ANOVA

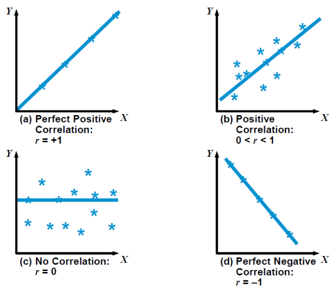
- Two factors affecting one response.

Regression

- Dependent variable (Y): response variable
  Independent variable (X): explanatory or predictor variable
- Idea: try to find a straight line that best fits the given data, having smallest error when calculated between the actual and predicted values.
- **To determine the model** → $\hat{y} = b_0 + b_1 x$
  - $\hat{y}$: predicted value (response)
  - $\bar{y}$: average y → $\bar{y} = \frac{\sum y_i}{n}$
  - y: actual value of y from dataset
  - x: actual value of x from dataset (predictor)
  - b_1: slope of the regression → $b_1 = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2} = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2}$
  - b_0: y interception (AKA. bias) → $b_0 = \bar{y} - b_1\bar{x}$
- Error (residual): $e = y - \hat{y}$
- **To determine the fit of the model**, we calculate Coefficient of determination (r^2)
  - $r^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$
  - It explains how much the variability of $\boldsymbol{y}$ can be explained by the model based on $\boldsymbol{x}$.
    Ex. $r^2 = 0.69$ → 69% of the data can be explained by x based model.
- SST: Total variability about the mean → $SST = \sum(y - \bar{y})^2$
  SSE: Variability about the regression line → $SSE = \sum(y - \hat{y})^2$
  SSR: Total variability that explain the model → $SSR = \sum(\hat{y} - \bar{y})^2$

  $\Big\}\ SST = SSR + SSE$





- Coefficient of correlation (r) is the expression of the strength of linear relationship.
  - Value between -1 to $+1$
- **To determine the significance of the model**, we use F-test.
- $H_0: b_0 = 0$ → there is no relationship between x and y.
- $MSR = \frac{SSR}{k}$, $MSE = \frac{SSE}{n-k-1}$ → MSE is also known as σ^2
  - k: number of independent variables
  - n: number of observations
- $f = \frac{MSR}{MSE}$, using $DOF_1 = k, DOF_2 = n - k - 1$
  - Probability of f is known as $Significance\ F$ (F')
  - lower F' means the model is useful in predicting y.
  - $F' < \alpha$, reject H_0, there are relationship between x and y.

Multiple Regression analysis

- Study the model with multiple independent variables.
- $\hat{y} = b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k + \varepsilon$
  - ε: random error
- The largest b_i show that x_i has the highest impact to $\hat{y}$
- r^2 explain how much $\hat{y}$ can be explain by the model based on $x_1, x_2, \ldots, x_k$
  The rest $(1 - r^2)$ explain how the remaining $\hat{y}$ can be explained by other variables that aren't included in the model.

	A	B	C	D	E	F	G	H	I
16									
17	SUMMARY OUTPUT								
18									
19	*Regression Statistics*								
20	Multiple R	0.81968							
21	R Square	0.67188							
22	Adjusted R S	0.61222							
23	Standard Err	24312.6							
24	Observations	14							
25									
26	ANOVA								
27		*df*	*SS*	*MS*	*F*	*Significance F*			
28	Regression	2	13313936968	6.7E+09	11.2619				
29	Residual	11	6502131603	5.9E+08					
30	Total	13	19816068571						
31									
32		*Coefficient*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
33	Intercept	146631	25482.0829	5.7543	0.0001	90545.2073	202716.5798	90545.2073	202716.5798
34	SF	43.8194	10.2810	4.2622	0.0013	21.1911	66.4476	21.1911	66.4476
35	AGE	-2898.69	796.5649	-3.6390	0.0039	-4651.9139	-1145.4586	-4651.9139	-1145.4586

The coefficient of determination (r^2) is 0.67.

can use x_1 & x_2 to explain 67% of data } other 33% use other variable which we exclude from the model

A low significance level for *F* proves a relationship exists.

The regression coefficients are found here.

The *P*-values are used to test the individual variables for significance.

$$\hat{Y} = 146631 + 44X_1 - 2899X_2$$