Equation Summary Ch. 2 → 12

All finals material

Chapter 2:

① Measure of central tendancy:

1 → mean → Most accurate ~ affected by outlies

2 → median → middle accuracy ~ not affected by outliers

3 → mode → least accurate ~ affected by mode in far ends.

1 → mean : $\bar{x} = \frac{\sum x}{n}$ $\bar{x} = \frac{\sum fx}{\sum f}$

2 → median = $Q_2 = n \times 50\% = n^{th}$ term

Imp → 4^{th} term is 4^{th} term, 4.5^{th} term is 4^{th} and 5^{th} term.

3 → Mode = Most repeated number

* coefficient of variation : $C.V = \frac{\sigma}{\mu}$

② Measure of spread:

1 → Range → max - min

2 → Inter-Quartile Range → $Q_3 - Q_1$

3 → Variance → S^2

4 → Standard deviation → S

⇒ $$\boxed{S^2 = \frac{\sum x^2}{n-1} - \frac{(\sum x)^2}{n(n-1)}}$$

③ Quantiles, percentiles and box-and-whisker:

$Q_1 = n \times 25\%$

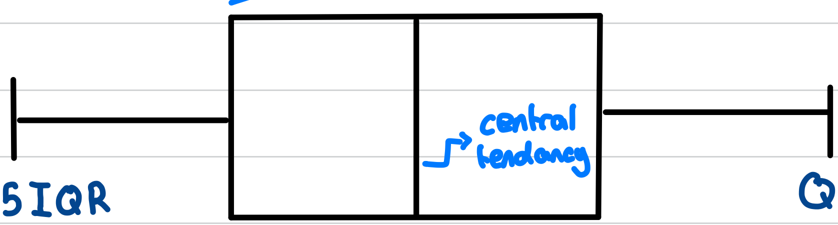
$Q_3 = n \times 50\%$

$P_k : (n) \times k\%$

If its a decimal, take next whole number. $7.15^{th} \rightarrow 8^{th}$

If whole, take that number and next one
$10^{th} \rightarrow \frac{10^{th} + 11^{th}}{2}$

measure of spread



$Q_1 - 1.5\,IQR$ $\quad$ $Q_3 + 1.5\,IQR$

Q_1	Q_2	Q_3
n 25%	n 50%	n 75%
↳ decimal → next	↳ decimal → next	↳ decimal → next
↳ whole → both	↳ whole → both	↳ whole → both

Outliers:

Upper extreme $Q_3 + 1.5\,IQR$

Lower extreme $Q_1 - 1.5\,IQR$

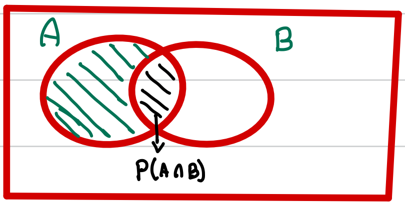
Chapter 3:

① $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}$ (conditional probability)

② $P(A \cap \bar{B}) = P(A) - P(A \cap B)$

basically prob. of A and outside B
because $\bar{B} = 1 - B$.

So:



$\therefore P(A \cap \bar{B})$
$\Rightarrow \underbrace{P(A)}_{\text{green}} - \underbrace{P(A \cap B)}_{\text{black}}$

③ Total probability rule:

$P(A) = P(A \mid B) \times P(B) + P(A \mid \bar{B}) \times P(\bar{B})$

Given that $A \cap B$ & $A \cap \bar{B}$ are mutually exclusive (disjoint)

④ Baye's rule:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A) \times P(A)}{P(B)}$$

$$= \frac{P(B|A) \times P(A)}{P(B|A) \times P(A) + P(B|\bar{A}) \times P(\bar{A})}$$

⑤ Total probability rule:

can't occur at same time

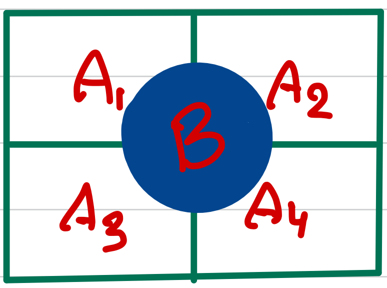
Given that events $A_1, \ldots\ldots, A_n$ are mutually exclusive

at least one event MUST occur

AND exhaustive

$$P(B) = \sum_{i=1}^{n} P(B|A_i)\, P(A_i)$$

$n = 4$



⑥ For any independent events:

$$P(A \cap B) = P(A) \times P(B)$$

⑦ For any two events:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

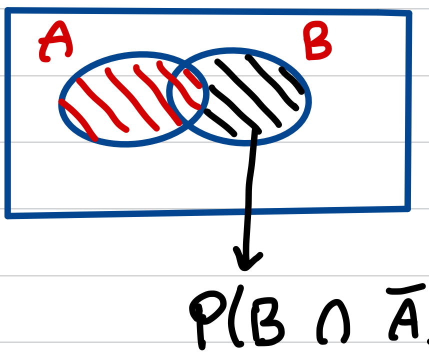
For mutually exclusive events:

$$P(A \cup B) = P(A) + P(B)$$

(because $P(A \cap B)$ does not exist in mutually exclusive events).

⑧ For two independent events:

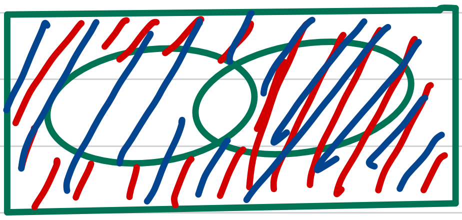
$$P(A \cup B) = P(A) + P(B) \times P(\bar{A})$$



$$P(A \cup B) = P(A) + P(B \cap \bar{A})$$

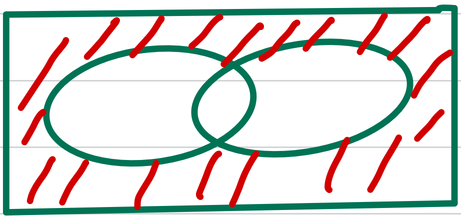
$$= P(A) + P(B) \times P(\bar{A})$$

9. $P(\bar{A} \cup \bar{B}) = P(\overline{A \cap B}) = 1 - P(A \cap B)$



only unshaded region is $P(A \cap B)$. $\therefore P(\bar{A} \cup \bar{B}) = P(\overline{A \cap B}) = 1 - P(A \cap B)$

10. $P(\bar{A} \cap \bar{B}) = P(\overline{A \cup B}) = 1 - P(A \cup B)$



$P(A \cup B)$ is not shaded

$\therefore P(\bar{A} \cap \bar{B}) = P(\overline{A \cup B}) = 1 - P(A \cup B)$

(11) **Sensitivity** = $P(T^+ \mid D)$

(12) **Specificity** = $P(T^- \mid \bar{D})$

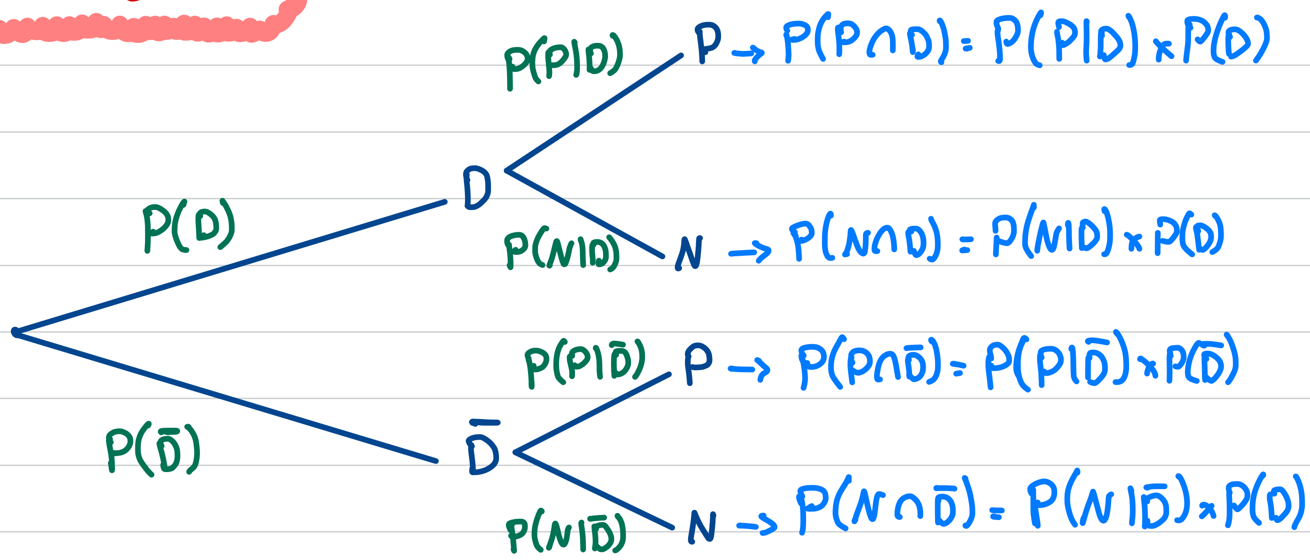
(13) **PV^+** = $P(D \mid T^+)$

(14) **PV^-** = $P(\bar{D} \mid T^-)$

(15) $P(T^+ \mid D) = 1 - P(T^- \mid D)$

(16) $P(T^- \mid \bar{D}) = 1 - P(T^+ \mid \bar{D})$

Cheat code:



Remember that first branch is $D / \bar{D}$ and second branch is P/N where its probability is conditional.

Chapter 4:

① Conditions for Binomial distribution:

→ A sample of independant trials n.

→ Only two possible outcomes, success p or failure q.

→ p and q are constant for each trial.

② $X \sim B(n, p) \Rightarrow P(X=a)$

$$\boxed{\therefore P(X=a) = {}^{n}C_{a} \times p^{a} \times q^{n-a}}$$

$(q = 1 - p)$

③ Variance and standard deviation:

$Var(x) = \sigma^2$

$Sd(x) = \sigma$

$$\Rightarrow \boxed{\begin{aligned} \sigma^2 &= npq \\ \sigma &= \sqrt{npq} \end{aligned}}$$

④ Discrete Random variable:

x	1	2	3
$P(X=x)$	a	b	c

1 → $a + b + c = 1$

2 → $E(x) = 1 \times a + 2 \times b + 3 \times c$

3 → $Var(x) = 1^2 \times a + 2^2 \times b + 3^2 \times c$

Chapter 5:

① $X \sim N(\mu, \sigma^2)$

$\Rightarrow \boxed{Z = \frac{X-\mu}{\sigma}}$ $\left\{ P(X > a) = P\left(Z > \frac{a-\mu}{\sigma}\right) \right.$

② $P(Z > -a) = P(Z < a)$

$P(Z < -a) = P(Z > a)$

$P(a < Z < b) = P(Z < b) - P(Z < a)$

Chapter 6:

Questions: They ask whats the probability of a sample mean ($\to \bar{x}$) or a sample proportion ($\to \hat{p}$) to be greater than a.

Or calculate confidence interval for sample mean or sample proportion.

① Standard error of the mean:

$Var(\bar{x}) = \frac{\sigma^2}{n}$

$$\boxed{S.d(\bar{x}) = \frac{\sigma}{\sqrt{n}}}$$

$\frac{\sigma}{\sqrt{n}}$ is estimated by $\frac{S}{\sqrt{n}}$

note: $Var(\bar{x})$ & $sd(\bar{x})$ are NOT the same as S^2 & S, they are not sample variance and s.d but they are s.d and variance of the set of sample means.

② Central limit theorem: $\bar{x} \sim N(\mu, (\sigma/\sqrt{n})^2)$

$P(\bar{x} > a)$

$\downarrow$

$P(z > \frac{a-\mu}{\sigma/\sqrt{n}})$

1→ σ is known: $\boxed{z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}}$

2→ σ is unknown. AND $n \leq 30$: $\boxed{t = \frac{\bar{x}-\mu}{S/\sqrt{n}}}$ degree of freedom $= n-1$

Using S.

3 → σ is unknown. AND $n > 30$:
Using S.

$$\boxed{z = \frac{\bar{x} - \mu}{S/\sqrt{n}}}$$

③ Confidence Interval of the mean:

$$\mu = \bar{x} \pm E$$

E :

1 → σ is known: $\boxed{E = z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}}$

2 → σ is unknown AND $n \leq 30$:
Using S

$$\boxed{E = t_{\frac{\alpha}{2}} * \frac{S}{\sqrt{n}}}$$

3 → σ is unknown AND $n > 30$:
Using S

$$\boxed{E = z_{\frac{\alpha}{2}} * \frac{S}{\sqrt{n}}}$$

notes:

↳ To find $z_{\frac{\alpha}{2}}$. Use C.I % for D. Or $\frac{\alpha}{2}$ value for B.

↳ To find $t_{\frac{\alpha}{2}}$: 1 − C.I % => α => $\frac{\alpha}{2}$ => $\frac{\alpha}{2}$ + C.I % => $t_{\frac{\alpha}{2}}$

Ex: 95% C.I => 5% α => 2.5% $\frac{\alpha}{2}$ => 97.5% → look t-dist. → $t_{\frac{\alpha}{2}} = 2.776$
d.f = 4 (example)

↳ The bigger the C.I the more the error (smaller rejection region)

↳ The bigger the n, the less the error.

④ Finding sample size n:

$$n = \left(Z_{\frac{\alpha}{2}} * \frac{\sigma}{E} \right)^2$$ → equation derived from E.

↳ or $t_{\frac{\alpha}{2}}$ depending on conditions

⑤ Point estimation of p:

$$\hat{p} \overset{\sim}{\sim} N\left(p, \sqrt{\frac{pq}{n}}^{\,2}\right)$$

$\hat{p}$ => sample proportion.

p => population proportion.

$$\boxed{\therefore \quad Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}}$$

(always z-dist)

⑥ Standard Error:

$S.E: \sqrt{\frac{pq}{n}}$ Estimated by $\sqrt{\frac{\hat{p}\hat{q}}{n}}$

↳ S.E point estimate

⑦ Interval Estimate of p: (confidence interval)

$$\boxed{p = \hat{p} \pm E}$$

$$\boxed{E = Z_{\frac{\alpha}{2}} * \sqrt{\frac{\hat{p}\hat{q}}{n}}}$$

→ For C.I, we use $\sqrt{\frac{\hat{p}\hat{q}}{n}}$ → point estimator. For C.L.T we use $\sqrt{\frac{pq}{n}}$

Chapter 7:

Questions: Testing for one sample mean or population

① Type 1 & Type 2 errors:

Mnemonic :

* Type 1 error => α => Reject True H_0 αRT
(art) α reject true H_0

* Type 2 error => β => Accept False H_0 βAF
(Baf) β accept false H_0

② Testing of the hypothesis:

numerical value

1-> $H_0: \theta = \theta_0$ VS
* $H_1: \theta > \theta_0$ => Right tailed
* $H_1: \theta < \theta_0$ => Left tailed
* $H_1: \theta \neq \theta_0$ => Two tailed

2-> Test Statistic

3-> p-value method OR rejection region method

③ Test statistic for μ:

1 → σ is known: $\boxed{z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}}$

2 → σ is unknown. <u>AND</u> $n \leq 30$:
Using S. $\boxed{t = \dfrac{\bar{x} - \mu}{S/\sqrt{n}}}$ degree of freedom $= n - 1$

3 → σ is unknown. AND $n > 30$:
Using S. $\boxed{z = \dfrac{\bar{x} - \mu}{S/\sqrt{n}}}$

④ Test statistic for p:

$$\boxed{\therefore \quad z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}}$$

$$\boxed{\therefore \quad z_{corr} = \frac{|\hat{p} - p| - \frac{1}{2n}}{\sqrt{\frac{pq}{n}}}}$$

⑤ p-value method:

Use test stat to find p-value

* p-value > α Accept H_0 & Reject H_1
* p-value < α Reject H_0 & Accept H_1

	0.001		0.01		0.05	p-values
ery highly significant		Highly Significant		Significant		Not significant

* Reject H_0 ⇒ Due statistically significant

⑥ Rejection region method:

Use α to find critical value. Use test statistic to find z-value. If z-value in rejection region then reject H_0 else accept H_0.

Chapter 8:

Questions: Two samples. Test hypothesis. for means only.

① Two Dependent Samples: (same n)

1 -> d = after - before

2 -> $\bar{d} = \frac{\sum d}{n}$ $\quad d.f = n-1$

3 -> $S_d^2 = \frac{\sum d^2}{n-1} - \frac{(\sum d)^2}{n(n-1)}$

Test Statistic :

$$\boxed{t = \frac{\bar{d}}{S_d / \sqrt{n}} \quad n \leq 30 \qquad z = \frac{\bar{d}}{S_d / \sqrt{n}} \quad n > 30}$$

note that test stat is similiar to one sample mean, but $d = 0$ so we dont write it.

C.I => $\boxed{d = \bar{d} \pm E \quad \Big| \quad E = t_{\frac{\alpha}{2}} * \frac{S_d}{\sqrt{n}}}$

② Two independent samples:

Test Statistic: $$\boxed{t = \frac{\bar{x}_1 - \bar{x}_2}{S_p \cdot \sqrt{\frac{1}{n} + \frac{1}{m}}}}$$

degree of freedom $= n + m - 2$

🦅🇺🇸🛢️

C.I: $$\mu_1 - \mu_2 = (\bar{x}_1 - \bar{x}_2) \pm E$$

$$E = t_{\frac{\alpha}{2}} * S_p * \sqrt{\frac{1}{n} + \frac{1}{m}}$$

Pooled Variance, $S_p \Rightarrow \sqrt{\frac{S_1^2(n-1) + S_2^2(m-1)}{n + m - 2}}$

Chapter 10:

$\hat{P}_1 , \hat{P}_2$

Questions: Testing a hypothesis for two-sample proportion.

① Normal Theory method, z :

Conditions: * $n\,p^* q^* > 5$ \& $m\,p^* q^* > 5$

* Sample size is large, samples discrete and independant

two-tailed

① $H_0: P_1 - P_2 = 0$ Vs $H_1: P_1 - P_2 \neq 0$ (can also be right-tailed or left tailed but χ^2 is always right tailed)

② Test stat$_1$:

no correction

$$\boxed{Z = \frac{(\hat{P}_1 - \hat{P}_2) - \overbrace{(P_1 - P_2)}^{\text{Is always zero.}}}{\sqrt{p^* \underbrace{q^*}_{q^* = 1 - p^*}} * \sqrt{\frac{1}{n} + \frac{1}{m}}}}$$

Test Stat$_2$:

with correction

$$\boxed{Z_{corr} = \frac{|\hat{P}_1 - \hat{P}_2| - \left(\frac{1}{2n} + \frac{1}{2m}\right)}{\sqrt{p^* q^*} * \sqrt{\frac{1}{n} + \frac{1}{m}}}}$$

pooled proportion, $\boxed{p^* = \frac{X + Y}{n + m}}$

$$\hat{P}_1 = \frac{X}{n} \qquad \hat{P}_2 = \frac{Y}{m}$$

③ Either p-value method or rejection region method

② Contingency Table:

$$\boxed{* \ \chi^2 = \sum \frac{(O-E)^2}{F}}$$

$$\boxed{* \ \chi^2_{corr} = \sum \frac{(|O-E| - 0.5)^2}{E}}$$

note that we mainly use χ^2_{corr} for 2×2.

$$\chi^2 = \frac{(O_{11} - E_{11})^2}{E} + \frac{(O_{12} - E_{12})^2}{E} + \frac{(O_{21} - E_{21})^2}{E} + \frac{(O_{22} + E_{22})^2}{E}$$

Observed (given):

	1	2	
A	x	y	$x \overset{R_1}{+} y$
B	z	u	$z \overset{R_2}{+} u$
	$\underset{C_1}{x + z}$	$\underset{C_2}{y + u}$	$\underbrace{(x+y)+(z+u) \text{ or } (x+z)+(y+u)}_{T}$

Expected

	1	2
A	$\frac{C_1 \times R_1}{T}$	$\frac{C_2 \times R_1}{T}$
B	$\frac{C_1 \times R_2}{T}$	$\frac{C_2 \times R_2}{T}$

$$\boxed{\therefore \quad E = \frac{R \times C}{T}}$$

no of rows → no of columns

Degree of freedom: $(r-1) \times (c-1)$

notes:

* H_0 : independant variables. Two variables NOT associated

* H_1 : dependent variables. Two variables ARE associated.

* For 2×2 table, all E must be greater than 5

* For R×C table,
Always the expected value is greater than 1.
For every 5 cells, only one value of E less than 5 is allowed.

Ex:

5	9	121
1	17	2

Doesn't Satisfy

5	9	121
7	17	2

Satisfy

* Always take right tailed test

③ Chi-squared goodness of fit test:

Probability × Observed total value = Expected value

1 → Continuity correction

2 → Find probability for cell

3 → multiply probability with O_t, get expected

4 → find expected for all cells.

5 → Chi-squared test. Accept or reject H_0.

↳ dont forget continuity correction

$$\boxed{* \quad P(x < a) \times O_t}$$

Degree of freedom for Chi-squared goodness-of-fit:

$$\boxed{d.f = g - k - 1}$$

g: number of groups/categories

↳ If $E < 5$ then for this category we join it with another category and it counts as one.

k: number of estimated parameters

↳ point estimates ⇒ $\bar{x}$, s, $\hat{p}$

Chapter 11:

Questions: When they talk about Correlation / correlation coefficient

① Covariance: (not very important)

$$Cov(x, y) = E((x-\mu_x)(y-\mu_y)) = E(xy) - \mu_x \mu_y$$

② Correlation coefficient

population correlation coefficient $= \rho$

$$\rho = \frac{Cov(x,y)}{\sigma_x \sigma_y}$$ (equation not very important)

Sample correlation coefficient $= r$

formula sheet

$$\boxed{r = \frac{L_{xy}}{\sqrt{L_{xx} L_{yy}}} = \frac{S_{xy}}{S_x \cdot S_y}}$$

$r \Rightarrow \rho$ point estimator

$L_{xx}: \sum x^2 - \frac{(\sum x)^2}{n}$

$L_{yy} = \sum y^2 - \frac{(\sum y)^2}{n}$

$L_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n}$

formula sheet

* $-1 \leq r \leq 1$

→ closer to 1/-1 then stronger correlation

→ 1 => +ve correlation
-1 => -ve correlation

③ Statistical inference for hypothesis testing: Part 1

$H_0: \rho = 0$ Vs $H_1: \rho \neq 0$

Test Stat = $\boxed{t = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}}}$ In formula sheet.

④ Statistical inference for hypothesis testing: Part 2

$H_0: \rho = \rho_0$ Vs $\rho \neq \rho_0$

↳ A non-zero number

Test stat:

1 → $z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$ Formula sheet

2 → $z = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)$ Get ρ from statistical test

3 → $\lambda = (z - z_0)\sqrt{n-3}$ Formula sheet

note that the value of λ is the value that you must compare with critical value on a z-distribution table.

⑤ Confidence Interval for population correlation coefficient and fisher's z-transformation:

① Find (z_1, z_2)

↳ fisher's

② Find (ρ_1, ρ_2)

↳ pop. corr. coeff.

① $Z \pm E$

↳ $Z = \frac{1}{2} \ln\left(\frac{1+r}{1-r}\right)$

↳ $E = Z_{\frac{\alpha}{2}} * \frac{1}{\sqrt{n-3}}$

$$\boxed{\therefore \; Z_{(1,2)} = Z \pm Z_{\frac{\alpha}{2}} * \frac{1}{\sqrt{n-3}}}$$

② $\rho = \frac{e^{2z} - 1}{e^{2z} + 1}$

* $\rho_1 = \frac{e^{2z_1} - 1}{e^{2z_1} + 1}$

* $\rho_2 = \frac{e^{2z_2} - 1}{e^{2z_2} + 1}$

C.I $= (\rho_1, \rho_2)$

Chapter 12:

Questions : When they ask for testing of 3 or more samples.

① Hypothesis testing of more than 2 samples using one way Anova Model. :

1 → $H_0: \mu_1 = \mu_2 = \dots\dots \mu_k$ vs $H_1: \mu_1 \neq \mu_2 \neq \dots\dots \mu_k$

↳ at least one pair not equal.

2 → Test Statistic = $\boxed{F = \dfrac{MS_B}{MS_W}}$

Formula sheet

⇓

* $\boxed{MS_B = \dfrac{SS_B}{k-1}}$ → $\boxed{SS_B = \sum n\bar{y}^2 - \dfrac{(\sum ny)^2}{N}}$

↳ N is total sample size

Degree of freedom ⇒ k-1

// ⇓ //

* $\boxed{MS_W = \dfrac{SS_W}{N-k}}$ → $\boxed{SS_W = \sum (n-1)s^2}$

Degree of freedom ⇒ N-k

② ANOVA notes:

* D.F numerator = $k-1$

$$\frac{MS_B}{MS_W}$$

* D.F denominator = $N-k$

* $\boxed{SS_T = SS_B + SS_W}$ ⇒ didn't see any questions in this.

③ Least Significant Difference test (LSD):

↳ I need some LSD after this exam 😲🤔

↳ LSD is used to see which means are not equal to each other

↳ We compare two means individually, so we use the equation from chapter 8 for two independent samples.

Test Stat:

$$\boxed{t = \frac{\bar{y}_1 - \bar{y}_2}{S_p \cdot \sqrt{\frac{1}{n} + \frac{1}{m}}}}$$

↱ the MS_w is used for all groups throughout so S_p is the same for any two means.

$$\boxed{S_p = \sqrt{MS_w}}$$

d.f = $N-k$

$$MS_W = \frac{SS_W}{N-k} = \frac{\sum (n-1) \cdot s^2}{N-k}$$

Example :

$H_0: \mu_1 = \mu_2$ Vs $H_1: \mu_1 \neq \mu_2$

② Test stat $\Rightarrow$ $t = \dfrac{\bar{y}_1 - \bar{y}_2}{S.p\sqrt{\frac{1}{n} + \frac{1}{m}}}$ $\quad S_p = \sqrt{MS_\omega}$

③ Use df and α to find critical values.

Thus either reject or accept H_1.

" لَا يُكَلِّفُ اللَّهُ نَفْسًا إِلَّا وُسْعَهَا "