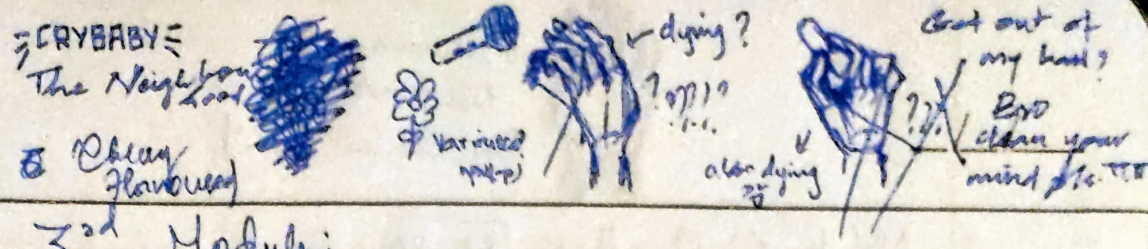


13 Feb.

3rd Module:

* Statistic Proportion:

p - represents the population's proportion.

$\hat{p}$ - represents the sample's proportion.

$\hat{p} = \frac{X}{n}$ → ~~# individuals in the sample with the characteristic.~~

n → # individuals in the sample.

$\hat{p} \rightarrow [0,1]$.

Mean of $\hat{p}$ = $E[\hat{p}] = E\left[\frac{X}{n}\right] = \frac{1}{n} E[X] = \frac{1}{n} \times np = p$. (→ binomial.)

Variance of $\hat{p}$ = $Var\left(\frac{X}{n}\right) = \frac{1}{n^2} Var(X) = \frac{1}{n^2} \times np(1-p)$

$= \frac{1}{n} p(1-p) = \frac{p(1-p)}{n}$.

Standard deviation of $\hat{p}$ = $\sqrt{\frac{p(1-p)}{n}}$.

Central limit Theorem: Condition $np \geq 15$. OR $n(1-p) \geq 15$.

n & p. Even when n is small, if $p = 0.5$, it gives a considerably normal distribution.

With n increasing, it draws closer to a normal distribution with smooth.

When n is small, $p = 0.04$, it gives → a left skewed graph. Only with increase in size n can make it normal.

Similarly, when $p = 0.96$, it gives → a right skewed graph. With increase in n, we get more normal dist.

→ Solomon's left hand?
[Not scary, just sad :-].

__/__/__

$N(\mu, \sigma) \Rightarrow N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$.

Use →

Normal distribution only for p too low or too high when the value of n is large.

Q4. In a certain Liberal precinct(?) across Toronto, 81% of the voters are registered democrats. What is the probability that, in a random sample of 100 voters from this precinct, no more than 80 of the voters would be registered democrats.

→ $n = 100$. $\hat{p} = \frac{80}{100} = 0.8$. $p = 81\% = 0.81$.

$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.81(0.19)}{100}} = 0.039$.

$z = \frac{\hat{p} - p}{\sigma_{\hat{p}}} = \frac{0.8 - 0.81}{0.039} = \frac{-0.01}{0.039} = -0.2564$

0.05 | 0.06

-0.2 0.401 0.397

$\therefore P(\hat{p} \leq 0.8) = P(z \leq -0.2599) = 0.3994$.

Confidence Interval for p:

$\hat{p} \pm z_{\alpha/2} \times SE\hat{p}$.

where standard error $SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

Qn. A study investigated the proportion of male births in Liverpool, England. In a sample of 5045 births to non-smoking parents, there were 2685 males and 2360 females born. Construct a confidence interval for p, i.e., pop proportion of male births.

$\hat{p} = \frac{2685}{2685+2360} = \frac{2685}{5045} = 0.5322$.

$SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.5322(0.9678)}{5045}} = \sqrt{0.00004935}$

$= 0.007025$.

$\therefore CI = \hat{p} \pm z_{\alpha/2}\ 0.007025 = 0.5322 \pm 1.96 \cdot 0.007025$

$= 0.5322 \pm 0.013769$.

$\rightarrow (0.518431, 0.545969)$.

Qn. A recent survey of 1000 married men revealed that 56% of them have been unfaithful atleast once. Form a 95% confidence interval to estimate the true proportion of married men who are unfaithful.

$\hat{p} = \frac{560}{1000} = 0.56$.

$SE(\hat{p}) = \sqrt{\frac{0.56(0.44)}{1000}} = 0.0157$.

$CI = \hat{p} \pm z_{\alpha/2}\ 0.0151 = 0.56 \pm 1.96 \times 0.0151$

$= 0.56 \pm 0.030172$

$\therefore CI = (0.529828, 0.590172)$.

24 Feb.

Sampling Variability for CI and Required size:

Suppose we are about to draw a sample and we wish to estimate the population proportion p. We may wish to estimate p to within an amount m with 95% confidence.
How large of a sample size is required?

→ $1.96 \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq m$.

maximum when: $\hat{p} = 0.5$.

Q. Note from others.

Hypothesis Testing:

$H_0: p = p_0$ → Hypothesized value of p.
$H_1: p > p_0$ or $p < p_0$ or $p \neq p_0$.

$$Z = \frac{\hat{p} - p_0}{SE_0(\hat{p})}$$

where $SE_0(\hat{p}) = \sqrt{\frac{p_0(1-p_0)}{n}}$

Depends on H_1:
When $H_1: p < p_0$ — Left tailed.
$H_1: p > p_0$ — Right tailed.
$H_1: p \neq p_0$ — Two tailed.

If p value $\leq \alpha$, reject H_0.

Ex: In a random sample of 200 people, 18 had major defects. male births.

$$\hat{p} = \frac{18}{200} = 0.09.$$

Test the null hypothesis that the true proportion of male births is 0.5, against the alternative hypothesis that it differs from 0.5.

→ $\hat{p} = \frac{18}{200} = 0.09.$

$H_0 : p = p_0 = 0.5.$

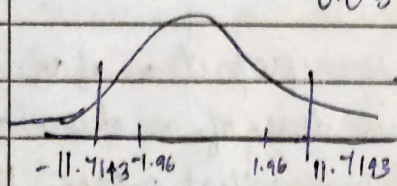
$H_1 : p \neq p_0 \neq 0.5.$ Two tailed - "differ".

$\alpha = 0.05.$ $\rightarrow \pm 1.96.$

$$Z = \frac{\hat{p} - p_0}{Sto(\hat{p})}.$$

$$Sto(\hat{p}) = \sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{0.5(1-0.5)}{200}} = \sqrt{\frac{0.25}{200}} = 0.035.$$

$$z = \frac{0.09 - 0.035 - 0.5}{0.035} = \frac{-0.41}{0.035} = -11.7143.$$



∴ Reject H_0.

(∵ def area $\leq \alpha$.) - with p value from z table (?)

Ex: Investors in New IT have a historical success rate of 65%. In a recent poll, a random sample of 250 New IT investors showed that only 130 were successful. At $\alpha = 0.01$, is there enough evidence to conclude that the success rate has increased from 65%?

$\hat{p} = \frac{180}{250} = 0.72 = 72\%$. Ratio > 65%.

$H_0: p = p_0 = 0.65$

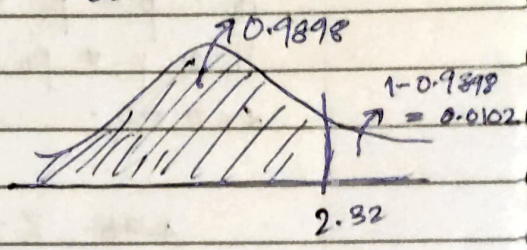
$H_1: p > p_0 > 0.65$. → one tailed .

$SE(\hat{p}) = \sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{0.65(1-0.65)}{250}}$

$= \sqrt{\frac{0.2275}{250}} = 0.0302$.

$z = \frac{\hat{p} - p_0}{SE(\hat{p})} = \frac{0.72 - 0.65}{0.0302} = \frac{0.07}{0.0302} = 2.31788$.

$z = 2.32$.



1-pvalue = 0.9898 .

~~0.9898 > 0.04, ∴ fail to reject Ho.~~

$p \geq 1 - 0.9898 = 0.0102$.

∴ ~~$p \leq 0.04$~~ $0.0102 \leq 0.01$, we reject Ho .

Ex. Suppose we wish to estimate the proportion of mice that would be killed by a certain dose of a chemotherapy drug. Previous studies have shown that appox 10% of mice are killed by this dose of the drug. Suppose we wish to estimate p to within 0.001 with 99% confidence, how large of a sample size is required?

$z \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq m$.

$m = 101 = 0.01$? $z = 2.575$

$= 2.575 \sqrt{\frac{0.5(1-0.5)}{n}} = 0.01$

$\frac{1.2875}{\sqrt{n}} = 0.01$

$n = \left(\frac{1.2875}{0.01}\right)^2 = 16.5$ ✗ not sure if correct

25 Feb

1 Hypothesis Test for comparing two proportions:

Ex: A study investigated a possible effect of magnetic pulse on the ability of homing pigeons to navigate back to the home loft. Pigeons were randomly ~~selected~~ divided into a magnetic pulse group and a control group. The pigeons were then released from a location 106 km from the home loft.
22 out of 38 control group pigeons returned to the home loft. 21 out of the 39 pigeons that received a magnetic pulse returned to the home loft.
- Construct a confidence interval for $p_1 - p_2$
- Test H_0: $p_1 = p_2$

→ CI $= \hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2} \times SE(\hat{p}_1 - \hat{p}_2)$.

where $SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$.

$\hat{p}_1 = \frac{x_1}{n_1} = \frac{22}{38} = 0.58$.

$\hat{p}_2 = \frac{x_2}{n_2} = \frac{21}{39} = 0.54$.

$z_{\alpha/2} = 1.96$.

$$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} = 0.112.$$

$$CI = 0.04 \pm 1.96(0.113)$$
$$= 0.04 \pm 0.22148$$
$$= (-0.18148, 0.26148)$$

To test $H_0: p_1 = p_2$:

$$SE_0(\hat{p}_1 - \hat{p}_2)$$

$H_0: p_1 = p_2$ or $p_1 - p_2 = 0$.

$H_1: p_1 \neq p_2$ or $p_1 - p_2 \neq 0$.

$$z = \frac{\hat{p}_1 - \hat{p}_2}{SE_0(\hat{p}_1 - \hat{p}_2)}$$

where $$SE_0(\hat{p}_1 - \hat{p}_2) = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

where $$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

pooled sample proportion since our assumption is $p_1 = p_2 = \hat{p}$.

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{22 + 21}{38 + 39} = 0.558$$

$$SE_0(\hat{p}_1 - \hat{p}_2) = \sqrt{0.558(1-0.558)\left(\frac{1}{38} + \frac{1}{39}\right)}$$

$$= 0.1132$$

$$\therefore z = \frac{\hat{p}_1 - \hat{p}_2}{SE_0(\hat{p}_1 - \hat{p}_2)} = \frac{0.58 - 0.54}{0.1132} = 0.353$$

Two tailed. $\alpha = 0.05$.

1-Pvalue = $1 - 0.6368 = 0.3632$.

Pvalue = $0.3632 \times 2 = 0.7264$.

→ Since $p > \alpha$, we fail to reject H_0.

Semester Assignments — January ☹ No No ☹ Febraury ☹ Yes Yes ☹
Many Many assignments. — Corrected by Jeslin
Wo death. the destroyer of assignments.
__/__/____

Qn. A machine puts out 16 imperfect articles in a sample of 500. After the machine is overhauled, it puts out 3 imperfect articles in a batch of 100. Has the machine improved?

$H_0: p_1 = p_2$ — One tailed.
$H_1: p_1 < p_2$? — Left tailed ?

$$\hat{p}_1 = \frac{x_1}{n_1} = \frac{16}{500} = 0.032.$$

$$\hat{p}_2 = \frac{x_2}{n_2} = \frac{3}{100} = 0.03.$$

$CI = \hat{p}_1 - \hat{p}_2 \pm Z_{\alpha/2}$?

27 Feb.

* Parametric & Non parametric Tests:

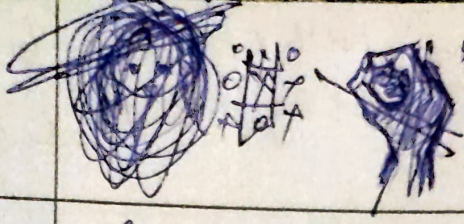
Non parametric:

(1) Chi-square Test: (χ^2).
For goodness of fit.

$$\chi^2 = \frac{\Sigma (O-E)^2}{E} = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}.$$

~~$\chi^2 = \sum \frac{(\text{Predicted} - \text{Actual})^2}{\text{Actual}}$~~. $\chi^2 = \sum \frac{(\text{Actual} - \text{Predicted})^2}{\text{Predicted}}$.

~~If~~ χ^2 Assume expected/actual values > 5.



χ^2 Critical $(\alpha, df_1) \leq \chi^2_{stat}$. → Reject H_0.

χ^2 critical $(\alpha, df_1) > \chi^2$ stat. → fail to reject H_0.

$df_1 = k-1$.

Ex. Sample of 100. Allergy in 2015 - Yes: 53%, No 47%.
2020 - Yes 49%, No 51%.

H_0: $(p_1, p_2) = (53\%, 47\%)$.

H_1: $(p_1, p_2) \neq (53\%, 47\%)$.

Sample - observed; population - expected.

O_1: 49 E_1: 53

O_2: 51 E_2: 47

$$\chi^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} = \frac{(49-53)^2}{53} + \frac{(51-47)^2}{47}$$

$$= 0.302 + 0.34 = \underline{0.61}.$$

$\chi^2_{critical}(\alpha, df_1) = \chi^2(0.05, 21) = \chi^2(0.05, 1) = \underline{3.841}$.

Since ~~χ^2_{stat}~~ $\chi^2_{critical} > \chi^2_{stat}$, we fail to reject H_0.

pvalue: At $df_1 = 1$, $\alpha = 0.9$ $\alpha = 0.1$

$df_1 = 1$: 0.016 2.706

↓

0.64

pvalue = 0.42.

~~Since 0.42 > 0.05, Correct~~

Since $0.42 > 0.05$, fail to reject H_0.

Qn. Some research shows higher number of flight tickets are bought by males in comparison to females with ratio of 2:1. Out of 150 tickets, 88 were bought by males, 62 by females. We need to find out if the experimental manipulation causes the change in the results, or if we are observing a chance variation.

2:1 → 150.

→ 100 : 50. O_1 88.

E_1 : E_2. O_2 62.

$$\chi^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} = \frac{(88-100)^2}{100} + \frac{(62-50)^2}{50}$$

$$= 1.44 + 2.88 = 4.32.$$

$df_1 = 2 - 1 = 1.$

$\alpha = 0.05.$

$\chi^2_{critical} = (0.05, 1) = 3.841.$

Since $\chi^2_{critical}$ (3.841) $\leq \chi^2_{stat}$ (4.32), we reject H_0.

* Chi-Square test of Independence:

Qn. $n = 8100.$

	Cancer	No cancer	Total
Smokers	60 O_{11}	800 O_{12}	860
No smokers	40 O_{21}	7200 O_{22}	7240
% smokers	60%	10%	
Total	100	8000	8100

$E_{11} = \frac{Row_1}{N} \times \frac{Col_1}{N} \times N = \frac{860}{8100} \times \frac{100}{8100} \times 8100 = 10.62$.

$E_{12} = \frac{Row_1}{N} \times \frac{Col_2}{N} \times N = \frac{860}{8100} \times \frac{8000}{8100} \times 8100 = 849.4$.

$E_{21} = \frac{Row_2}{N} \times \frac{Col_1}{N} \times N = \frac{7240}{8100} \times \frac{100}{8100} \times 8100 = 89.4$.

$E_{22} = \frac{Row_2}{N} \times \frac{Col_2}{N} \times N = \frac{7240}{8100} \times \frac{8000}{8100} \times 8100 = 7150.62$.

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

$$= \frac{(60-10.62)^2}{10.62} + \frac{(800-849.4)^2}{849.4} + \frac{(40-89.4)^2}{89.4} + \frac{(7200-7150.62)^2}{7150.62}$$

$$= 229.6 + 2.873 + 27.3 + 0.341$$

$$= 260.114.$$

$df = (row-1)(col-1) = (2-1)(2-1) = 1 \times 1 = 1$.

$\chi^2_{critical}(0.05, 1) = 3.841$.

Since $\chi^2_{critical}(3.841) \leq \chi^2_{stat}(260.114)$, we reject H_0.

28 Feb.

* K-S test (Kolmogorov - Smirnov):
To compare sample & population distribution.
Observed & theoretical.

Ex. In a study done from various streams of a college of 60 students, with equal no. of students drawn from each stream, the intention of the students to join the Adventure club of college noted after the interviewed are listed below.

	BSc	BA	BCom	MA	Mcom.
#in class	5	9	11	16	19

It was expected that 12 students from each class would join the Adventure club. Using K-S, find if there is any difference among student classes with regards to their intention of joining the club.

→ H_0: $F_0(x) = F_n(x)$. → no difference.
H_1: $F_0(x) \neq F_n(x)$. → Yes difference.

	Actual/Obsv	Theoretical	$F_0(x)$ CDF	$F_n(x)$ CDF	$D = \lvert F_0(x) - F_n(x) \rvert$
BSc	5	12	5/60	12/60	7/60
BA.	9	12	14/60	24/60	10/60
Bcom.	11	12	25/60	36/60	→ 11/60
MA	16	12	41/60	48/60	7/60
Mcom	19	12	60/60	60/60.	0
Total	60	60			

$D_{stat} = \max \lvert F_0(x) - F_n(x) \rvert = 11/60 = 0.183.$

$n > 40 \rightarrow \alpha = 0.05 \Rightarrow D_{critical} = \frac{1.36}{\sqrt{60}} = 0.175.$

Since $D_{critical}(0.175) < D_{stat}(0.183)$,
reject H_0.

__/__/__

Qn. Consider the data points

1.41 0.26 1.91 0.33 0.55 0.77 1.46 ~~1.91~~ 1.12.

Is there any evidence to suggest that the data were not randomly sample from a uniform (0,2) distribution?

→ H_0: $F(x) = F_0(x)$.

H_1: $F(x) \neq F_0(x)$.

* K-S - Two sample test:
Comparing distribution of 2 samples.

$D = \max |F_{n_1}(x) - F_{n_2}(x)|$.

Small sample : when n_1 & $n_2 < 40$.
Large sample : when n_1 and/or $n_2 > 40$.

Ex. consider a survey on two different universities of the PG students on the topic of their willingness to join the Research Funding Project on AI. The following is the results obtained:

Uni 1	3	2	3	5	8	9	8	8
Uni 2	2	8	2	4	4	3	6	

Determine whether the samples for uni 1 & 2 come from the same distribution.

→ H_0: $F_1(x) = F_2(x)$ → Same distribution.
H_1: $F_1(x) \neq F_2(x)$. → Different distribution.

Items	Frequency Uni 1	Uni 2	CDF $F_1(x)$	CDF $F_2(x)$	$D = \lvert F_1 - F_2 \rvert$
2	1	2	1/8	2/7	•0.161.
3	2	1	3/8	3/7	•0.0536.
4	0	2	3/8	5/7	0.34
5	1	0	4/8	5/7	0.214
6	0	1	4/8	6/7	→ 0.36.
8	3	1	7/8	7/7	0.125
9	1	0	8/8	7/7	0
Total:	8	7			

$\therefore D = 0.36$. $D_{critical}(0.05, 8, 7) = 40/56 = 0.7143$. $D_{test} < D_{critical}$, ∴ fail to reject.

* **Sign Test:**

- Non parametric test.
- Using population median η.
- If $T_{stat} > T_c$, fail to reject H_0; $T_{stat} < T_c$, reject H_0.

H_0: $\eta = \eta_0$.

H_1: $\eta \neq \eta_0$ ← Two tailed.

H_1: $\eta < \eta_0$ ← left tailed.

H_1: $\eta > \eta_0$ ← right tailed.

Ex. Following are the responses to the question "How many hours does they study before a major statistic test, mi'lord?".

6 5 1 2 2 5 7 5 8 7 4 7.

Use the sign test to test the hypothesis at 5% sig lvl. that the median # hours a student studies before a test is 3. Given that the critical value of sign test for $n = 11$ at 0.05 α for two tailed is 1.

→ $n \leq 25$, ∴ small sample.

H_0: $\eta = 3$. $T_c = 1$. $n = 11$.

H_1: $\eta \neq 3$ Two-tailed. $\eta_0 = 3$.

6-3	5-3	1-3	2-3	2-3	5-3	7-3	5-5	8-3	7-3	4-3	7-3.
3	2	-2	-1	-1	2	4	2	0	4	1	4.
+	+	-	-	-	+	+	+	0	+	+	+

$T_+ = 8$. ; $T_- = 3$. ;

$\Rightarrow T = \min(T^+, T^-) = \min(8, 3) = 3$.

Since $T(3) > T_c(1)$, we fail to reject.

Reject H_0 T_c

6 March 5th Module:

* Statistical Decision Theory:

Complete Class of Decision rules:
Set of decision rules with less error.

Minimal Complete Class of Decision rules:
'just enough' sort of set, no removal of rules needed, no addition either. just enough.

Loss function; Risk function.

Minimal CCDR: Smallest subset of the complete class that contains all the decision rules that are optimal with respect(;) to the loss function