SDS → Unit 4

→ Chi-Square test →

→ Probabilities of k items are denoted by: $p_1, p_2, p_3, \ldots p_k$.

→ Null hypothesis is of the form

$H_0: p_1 = p_{01}, p_2 = p_{02}, \ldots p_k = p_{0k}$

→ $\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$

→ When expected values are all sufficiently large good approx, possible.

$(k-1)$ degrees of freedom denoted by:

$$\chi^2_{k-1} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

When each outcome has same probability each expected value is $\frac{\Sigma \text{Observed value}}{\text{Total no. of values}}$

→ For row & column data:

$$E_{ij} = \frac{R_i \times C_j}{N} \rightarrow \text{Column sum}$$

Row sum

Row sum + Column sum

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Degree of freedom $= (i-1) \times (j-1)$

Non-parametric tests

- Distribution-free tests
- Samples not req, to come from any specific dist.
- Need assumptions but they are not as restrictive as t-test.

- Wilcoxon Signed-Rank test
  - Can be used even with outliers

Slide 34

9.3, 9.0, 0.9, 21.7, 11.5, 13.9

$H_0: M \geq 12$, $H_1: M < 12$.

X	X-12	Rank	
9.3	-2.7	-3	
9.0	-3	-4	$S_+ = 7$
0.9	-11.1	-6	
21.7	9.7	(5)	P-value:
11.5	-0.5	-1	$P(S_+ < 7)$
13.9	1.9	(2)	

$P(S_+ < 4) = 0.1094$, $P(S_+ \geq 7) > 0.1094$

↓ can't reject H_0

when $n > 20$ → normal dist

$$Z = \frac{S_+ - \frac{n(n-1)}{4}}{\sqrt{\frac{(n+1)(2n+1)}{24}}}$$

- Ties replace with avg of their pos.

3, 4, 4, 5, 7 → 1, 2.5, 2.5, 4, 5

- 0 → dropped from sample & n reduced by 1.

Mann–Whitney Test

→ Wilcoxon Rank-Sum Test →

→ 2 assumptions →

→ First → pop, continuous

→ Second → prob density func must be iddy on shape & size.

→ 2 samples from 2 diff pops, same shape

Slide 45

Value	Pop	Rank
20	X	1
28	X	2
29	X	3
34	Y	4
35	Y	5
36	X	6
38	X	7
41	Y	8
46	Y	9
47	Y	10
49	Y	11

$H_0: M_X \geq M_Y$

$H_1: M_X < M_Y$

$W = 19$

$m = 5 \quad n = 6$

$p = 0.0260$ ↓ reject H_0

→ If $m, n > 8$ normal dist

$$Z = \frac{W - \frac{m(m+n+1)}{2}}{\sqrt{\frac{mn(m+n+1)}{12}}}$$

Page No.

- Fixed-level testing →
→ measures plausibility of null hypothesis by prod/ P-value.
→ ↓ P-value → ↓ plausible H_0.

Choose α such that $0 < \alpha < 1$.
If $P \leq \alpha$ → reject H_0, else do not reject H_0.

Slide 59

5% level, $\alpha = 0.05$, Inv Normal $\rightarrow z = -1.645$

$$z = \frac{\bar{x} - M}{s/\sqrt{n}}$$

$$\bar{x} = \frac{zs}{\sqrt{n}} + M$$

$$\bar{x} = \frac{-1.645 \times 0.6}{\sqrt{80}} + 2$$

$$\bar{x} = 1.89$$

H_0 reject if → $\bar{x} \leq 1.89$ → rej. region.

10% level, $\alpha = 0.1$, $z = -1.28$

$$\bar{x} = \frac{-1.28 \times 0.6}{\sqrt{80}} + 2$$

$$\bar{x} = 1.9141$$

$1.85 < 1.9141$

H_0 rej at 10% level

Page No.

→ Errors →
→ Type I error: Reject H_0 when true
→ Type 2 error: Fail to reject H_0 when false

Prob of type I error never greater than α.

→ Power of test → prob of rejecting H_0 when it is false.
→ Power = $1 - P(\text{type II error}) = 1 - \beta$

→ Factors affecting power →
→ Sample size (n), $n \propto P$
→ Significance level (α), $\alpha \propto P$
→ Type of stat test
→ Effect size (True mean − Hypothesized mean)
→ Standard deviation (σ), $P \propto \frac{1}{\sigma}$

Power of 1 tailed test > 2 tailed test
greater effect size, greater power

μ close to H_0 → ↓ power
μ far from H_0 → ↑ power

	H_0 true	H_0 false
Reject H_0	Type I error	Correct decision
Fail to reject H_0	Correct decision	Type II error

Power ↑

→ In general, power > 0.80 or > 0.90 considered acceptable

→ Computing power →
→ Compute rej reg.
→ Compute prob that test stat falls in rej reg if alt hypothesis true.

→ Types of data → Univariate
→ Bivariate
→ Multivariate

→ Univariate → analysis using mean, median, mode, spread of data, histogram, piechart, etc.

→ Bivariate → analysis using correlation coeff, regression analysis, scatter plot.

find relation b/w two sets of values. (x, y) ordered pair
x → ind var, y → dep var

→ Scatter plot → math diagram that plots pairs of data on x-y graph to reveal relationship b/w data sets. Shows strength of relationship, direction of relation b/w var, & whether outliers exist.

corr coeff → only used when relation b/w var → linear

→ Correlation → strength of relation b/w two linear var.

→ Pearson correlation coeff →

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

$r = -0.5$ to 0.5 → weak
$r = \pm 0.8$ to ± 0.5 → moderate
$r = \pm 0.8$ to ± 1 → strong

→ correlation → +ve, -ve, no/poor

$r = \pm 1$ → perfect positive/negative corr
$0 < r < 1$ → +ve corr
$0 > r > -1$ → -ve corr
$r = 0$ → no corr

Page No.

→ Anscombe's Quartet → diff nos, but ≈ Summary stats. (Ex: mean, variance, etc.)

— Conclusions of Anscombe's Quartet →
→ Presence of outlier impacts evaluated value of r.
→ Basic stat, insuff, inadeq, for desc realistic datasets.
→ Imp to look at data visually before any kind of analysis.

→ Confounding var, → influences both ind & dep var causing spurious correlation.
Cause two major prob →
→ ↑ variance
→ Introduce bias
→ They act as extra ind var that have hidden effect on dep var.
→ Can be cause of correlation.

→ Controlled exp reduce the risk of confounding. Study rep no of times in variety of cond before drawing reliable conclusions.

→ Regression analysis → Study of set of data to make best guess or some sort of pred.
→ In stat modelling, it is used for est relation b/w dep var & 1 more ind var.

Regression Models → Simple → Linear
→ Non-Linear
→ Multiple → Linear
→ Non-Linear

- When two var, → linear relationship, scatter plot clustered along straight line → least fit line.

Least square line: $\hat{y}_i = \beta_0 + \beta_1 x_i$

β_0 → y intercept of least square line

where $\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$ → $\hat{\beta}_1$ → slope of least square line

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

- Residual → e_i = yobserved − ypredicted

- Least square line → sum of squared residuals → min

$\sum_{i=1}^{n} e_i^2$ → minimum

- Goodness of fit → how well model explains given set of data

Regression Sum of Square ← $\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$ + $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

→ Total sum of squares → error sum of squares

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

↓

coeff of determination → stat measure of how close obs data is to fitted regression line!

$r^2 = \frac{\text{Regression Sum of Squares}}{\text{Total sum of Squares}}$

↓ Value b/w 0 & 1. ↑ r^2 → better prediction

- If the vertical spread doesn't vary with fitted value → homoscedastic else, heteroscedastic
- Outlier that makes significant diff to least square line when removed → influential point.

~~Comments~~

- Weighted least squares Regression → points with ↓ variance → greater influence on computation of least squares line.