Chi - Square test.

This test looks at the variables on the side and top axes of a table and tests whether they are independant.

For example it can be used to show whether variations in political opinions depend on a respondent's age

H_0 No association between two variables
H_1 There is an association

Expected values: $\frac{\text{Row total} * \text{column total}}{\text{total no of observed}}$

χ^2 formula: $\sum \frac{\text{all cells } (\text{observed} - \text{expected value})^2}{\text{expected}}$

DF = (number of rows - 1) * (number of columns - 1)

Create confidence interval with DF

Example: Sample paper.

If p value (eg 0.004) is less than significance level (0.05) we can reject null hypothesis that there is an association between two variables yes/no and male/female

Using df create a confidence interval

eg.

$$-3.84 < x < 3.84$$

If chi square value is less than 3.84, not enough evidence against H_0

95% interval for difference in proportion

$$= (P_{dec} - P_{jun}) \pm t_{critical} * SE(P_{dec} - P_{jun})$$

$df = n_1 + n_2 - 2 = 1.96$

$$\pm 1.96 * SE$$

P_1 and P_2 are the sample proportion $\frac{\text{yes}}{\text{total}}$

$\frac{\text{yes Dec}}{\text{total dec}} - \frac{\text{yes June}}{\text{total jun}} = 0.02$

$$0.02 \pm SE = \sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2*(1-P_2)}{n_2}}$$

$$= -0.02 \text{ to } 0.06$$

neg and pos mean more in June / December

Regression

Scatterplot:
- Consists of x-axis and y axis and a series of dots
- Each dot represents one observation from a dataset
- Useful for determining the correlation between two variables
- Independent variable on x-axis
- Dependent variable of y axis

Example: Regression analysis: Yield grams (Vessels) Temperature (celcius)

Minitab output: Yield grams = 17.0 + 2.00 temp(c)

Predictor	Coef	SE Coef	T	P
constant	17.002	4.072	4.18	0.000
temp c	1.99517	0.05334	37.41	0.000

S = 4.01967 R-Sq = 98.4% R-Sq (adj) = 98.3%

Yield = equation made using coef rounded to nearest whole no

17.0 = yield when t=0
2.0 = increase in Y per unit increase in x.

Residual plot = Observed val - Predicted val
Actual value - value got using equation

SE coef - measure of variability of the slope from sample to sample
Rough 95% interval for every degree C extra: 2.00 ± 1.96*0.05 (1.99, 2.01) increase yield by (1.99 and 2.01)

dependent variable

P-values: Implicitly test hypothesis: -(constant)

H_0 Population slope = 0 (dependent)

H_1 Population slope ≠ 0

Test Statistic $\frac{\text{Coeff-0}}{SE(\text{coef})}$ $\frac{2.00-}{0.0533}$ = 37.4951 given by T

$p < 0.001$ we reject H_0

P value in independent variable: (temp c)

Test hypothesis H_0 Population intercept = 0

H_1 Population intercept ≠ 0

$p < 0.001$ we reject H_0

R^2:

- Measure of the fit of the model to the data
- 98.4% of the variance of "constant" accounted for.

S:

- For each observation we calculate the residual.
- S = 4.0187 0 the standard deviation of the residual

Assumptions:

- Residual should be normally distributed

- There should be no relationship between the residuals and the predicted value