UNIVERSITY OF STIRLING
Computing Science & Mathematics

MATU9D2 : PRACTICAL STATISTICS : FORMULA SHEET

1. One Sample Summary Statistics

$$\bar{x} = \frac{\sum x_i}{n}$$

$$s^2 = \frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1}$$

2. Two Sample Summary Statistics

$$\bar{x}_1 = \frac{\sum x_{1i}}{n_1} \qquad s_1^2 = \frac{\sum x_{1i}^2 - \frac{\left(\sum x_{1i}\right)^2}{n_1}}{n_1 - 1}$$

$$\bar{x}_2 = \frac{\sum x_{2i}}{n_2} \qquad s_2^2 = \frac{\sum x_{2i}^2 - \frac{\left(\sum x_{2i}\right)^2}{n_2}}{n_2 - 1}$$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

3. Correlation and Regression

$$S_{XY} = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n}$$

$$S_{XX} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} \qquad S_{YY} = \sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}$$

$$r = \frac{S_{XY}}{\sqrt{S_{XX}\, S_{YY}}} \qquad R^2 = \frac{S_{XY}^2}{S_{XX}\, S_{YY}}$$

The least squares line is estimated using

$$\hat{\beta} = \frac{S_{XY}}{S_{XX}}, \qquad \hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x}$$

4. Relationships

The Chi-squared test has test statistic

$$X^2 = \sum_{all\ cells} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Using a significance level of 0.05, this is compared with χ^2(df; 0.05), where df= (r-1)(c-1).

The test of the Null hypothesis $H_0 : \rho = 0$ against the alternative $H_1 : \rho \neq 0$ has test Statistic

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Using significance level of 0.05, this is compared with $\pm$ t(n-2; 0.025).

5. Probability

A random variable X following a Binomial probability model with n trials and probability of success θ has probability distribution

$$\Pr[X = x] = \frac{n!}{x!(n-x)!}\theta^x (1-\theta)^{n-x}, \qquad x = 0, \ldots\ldots n.$$

and mean and variance given by

$$E(X) = n\theta, \qquad Var(X) = n\theta(1 - \theta).$$

If X follows a Normal probability model with mean μ and variance σ^2, then $Z = \frac{(X-\mu)}{\sigma}$ follows a N(0,1) distribution.

6. One Sample Confidence Intervals

An interval estimate for a population *proportion*, having approximate confidence 0.95, is

$$\hat{\theta} \quad \pm \quad 1.96 \times \sqrt{\frac{\hat{\theta}\left(1-\hat{\theta}\right)}{n}}$$

An interval estimate for a population *mean*, having confidence 0.95, is

$$\bar{x} \quad \pm \quad t(n-1; 0.025) \times \frac{s}{\sqrt{n}}$$

An interval estimate for a population *variance*, having confidence 0.95, is

$$\left(\frac{(n-1)s^2}{\chi^2(n-1;0.025)}, \frac{(n-1)s^2}{\chi^2(n-1;0.975)}\right).$$

7. Two Sample Tests

The test statistics for the test of the Null Hypothesis $H_0 : \mu_1 = \mu_2$ against the alternative $H_1 : \mu_1 \neq \mu_2$ is

$$t \quad = \quad \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

For a test having significance level 0.05, we compare this with $\pm$ t($n_1 + n_2 - 2$; 0.025).

The test statistic for the test of the Null Hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$ against the alternative $H_1 : \sigma_1^2 \neq \sigma_2^2$ is

$$\frac{s_1^2}{s_2^2} \quad \left(\frac{larger}{smaller}\right)$$

For a test having significance level 0.05, we compare this with

$$F(n_1 - 1, n_2 - 1; 0.025).$$

8. Inference in Simple Linear Regression

The residual variance σ^2 is estimated by

$$\hat{\sigma}^2 = \frac{S_{YY} - \dfrac{S_{XY}^2}{S_{XX}}}{n-2}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x}$$

$$\hat{\beta} = \frac{S_{XY}}{S_{XX}}$$

A test of the Null Hypothesis $H_0 : \beta = 0$ against the alternative $H_1 : \beta \neq 0$ has the test statistic

$$T = \frac{\hat{\beta}}{\sqrt{\dfrac{\hat{\sigma}^2}{S_{XX}}}}$$

Using a significance level of 0.05, this is compared with $\pm$ t(n -2; 0.025).

Parameter	Estimate	Estimated Standard Error
α	$\hat{\alpha}$	$\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}\right)}$
β	$\hat{\beta}$	$\sqrt{\frac{\hat{\sigma}^2}{S_{XX}}}$
$\alpha + \beta x$	$\hat{\alpha} + \hat{\beta}x$	$\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{XX}}\right)}$

In each case interval estimates, having confidence 0.95, are given by

$$Estimate \quad \pm \quad t(n-2;0.025) \times Estimated\ Standard\ Error$$

The formula for a 95% prediction interval for an <u>individual observation</u> is

$$\hat{\alpha} + \hat{\beta}x \quad \pm t(n-2;0.025)\sqrt{\hat{\sigma}^2\left(1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{XX}}\right)}$$

9. **TO COME**