

Statistics formula sheet

Summarising data

Sample mean:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i.$$

Sample variance:

$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right).$$

Sample covariance:

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

Sample correlation:

$$r = \frac{g}{s_x s_y}.$$

Probability

Addition law:

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

Multiplication law:

$$P(A \cap B) = P(A)P(B|A) = P(B)P(A|B).$$

Partition law: For a partition B_1, B_2, \dots, B_k

$$P(A) = \sum_{i=1}^k P(A \cap B_i) = \sum_{i=1}^k P(A|B_i)P(B_i).$$

Bayes' formula:

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A)} = \frac{P(A|B_i)P(B_i)}{\sum_{i=1}^k P(A|B_i)P(B_i)}.$$

Discrete distributions

Mean value:

$$E(X) = \mu = \sum_{x_i \in S} x_i p(x_i).$$

Variance:

$$\text{Var}(X) = \sum_{x_i \in S} (x_i - \mu)^2 p(x_i) = \sum_{x_i \in S} x_i^2 p(x_i) - \mu^2.$$

The binomial distribution:

$$p(x) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \text{ for } x = 0, 1, \dots, n.$$

This has mean $n\theta$ and variance $n\theta(1-\theta)$.

The Poisson distribution:

$$p(x) = \frac{\lambda^x \exp(-\lambda)}{x!} \text{ for } x = 0, 1, 2, \dots$$

This has mean λ and variance λ .

Continuous distributions

Distribution function:

$$F(y) = P(X \leq y) = \int_{-\infty}^y f(x) dx.$$

Density function:

$$f(x) = \frac{d}{dx} F(x).$$

Evaluating probabilities:

$$P(a < X \leq b) = \int_a^b f(x) dx = F(b) - F(a).$$

Expected value:

$$E(X) = \mu = \int_{-\infty}^{\infty} x f(x) dx.$$

Variance:

$$\text{Var}(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx = \int_{-\infty}^{\infty} x^2 f(x) dx - \mu^2.$$

Hazard function:

$$h(t) = \frac{f(t)}{1 - F(t)}.$$

Normal density with mean μ and variance σ^2 :

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{1}{2} \left(\frac{x - \mu}{\sigma} \right)^2 \right\} \text{ for } x \in [-\infty, \infty].$$

Weibull density:

$$f(t) = \lambda \kappa t^{\kappa-1} \exp(-\lambda t^\kappa) \text{ for } t \geq 0.$$

Exponential density:

$$f(t) = \lambda \exp(-\lambda t) \text{ for } t \geq 0.$$

This has mean λ^{-1} and variance λ^{-2} .

Test for population mean

Data: Single sample of measurements x_1, \dots, x_n .

Hypothesis: $H : \mu = \mu_0$.

Method:

- Calculate \bar{x} , s^2 , and $t = |\bar{x} - \mu_0| \sqrt{n}/s$.
- Obtain critical value from t -tables, $df = n - 1$.

- **Reject** H at the $100p\%$ level of significance if $|t| > c$, where c is the tabulated value corresponding to column p .

Paired sample t -test

Data: Single sample of n measurements x_1, \dots, x_n which are the pairwise differences between the two original sets of measurements.

Hypothesis: $H : \mu = 0$.

Method:

- Calculate \bar{x} , s^2 and $t = \bar{x}\sqrt{n}/s$.
- Obtain critical value from t -tables, $df = n - 1$.
- **Reject** H at the $100p\%$ level of significance if $|t| > c$, where c is the tabulated value corresponding to column p .

Two sample t -test

Data: Two separate samples of measurements x_1, \dots, x_n and y_1, \dots, y_m .

Hypothesis: $H : \mu_x = \mu_y$.

Method:

- Calculate \bar{x} , s_x^2 , \bar{y} , and s_y^2 .
- Calculate

$$s^2 = \{(n-1)s_x^2 + (m-1)s_y^2\} / (n+m-2).$$

- Calculate $t = \frac{\bar{x} - \bar{y}}{\sqrt{s^2 \left(\frac{1}{n} + \frac{1}{m}\right)}}$.
- Obtain critical value from t -tables, $df = n + m - 2$.
- **Reject** H at the $100p\%$ level of significance if $|t| > c$, where c is the tabulated value corresponding to column p .

CI for population mean

Data: Sample of measurements x_1, \dots, x_n .

Method:

- Calculate \bar{x} , s_x^2 .
- Look in t -tables, $df = n - 1$, column p . Let the tabulated value be c say.
- $100(1-p)\%$ confidence interval for μ is $\bar{x} \pm cs_x / \sqrt{n}$.

CI for difference in population means

Data: Separate samples x_1, \dots, x_n and y_1, \dots, y_m .

Method:

- Calculate \bar{x} , s_x^2 , \bar{y} , s_y^2 .

- Calculate

$$s^2 = \{(n-1)s_x^2 + (m-1)s_y^2\} / (n+m-2).$$

- Look in t -tables, $df = n + m - 2$, column p . Let the tabulated value be c say.
- $100(1-p)\%$ confidence interval for the difference in **population** means i.e. $\mu_x - \mu_y$, is

$$(\bar{x} - \bar{y}) \pm c \left\{ \sqrt{s^2 \left(\frac{1}{n} + \frac{1}{m}\right)} \right\}.$$

Regression and correlation

The linear regression model:

$$y_i = \alpha + \beta x_i + z_i.$$

Least squares estimates of α and β :

$$\hat{\beta} = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{(n-1)s_x^2}, \quad \text{and } \hat{\alpha} = \bar{y} - \hat{\beta} \bar{x}.$$

Confidence interval for β

- Calculate $\hat{\beta}$ as given previously.
- Calculate $s_e^2 = s_y^2 - \hat{\beta}^2 s_x^2$.
- Calculate $SE(\hat{\beta}) = \sqrt{\frac{s_e^2}{(n-2)s_x^2}}$.
- Look in t -tables, $df = n - 2$, column p . Let the tabulated value be c .
- $100(1-p)\%$ confidence interval for β is $\hat{\beta} \pm c SE(\hat{\beta})$.

Test for $\rho = 0$

Hypothesis: $H : \rho = 0$.

- Calculate

$$t = r \left(\frac{n-2}{1-r^2} \right)^{1/2}.$$

- Obtain critical value from t -tables, $df = n - 2$.
- **Reject** H at $100p\%$ level of significance if $|t| > c$, where c is the tabulated value corresponding to column p .

Approximate CI for proportion θ

$$p \pm 1.96 \sqrt{\frac{p(1-p)}{n-1}}$$

where p is the observed proportion in the sample.

Test for a proportion

Hypothesis: $H : \theta = \theta_0$.

- Test statistic $z = \frac{p - \theta_0}{\sqrt{\frac{\theta_0(1 - \theta_0)}{n}}}$.
- Obtain critical value from normal tables.

Comparison of proportions

Hypothesis: $H : \theta_1 = \theta_2$.

- Calculate

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}.$$

- Calculate

$$z = \frac{p_1 - p_2}{\sqrt{p(1-p) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

- Obtain appropriate critical value from normal tables.

Goodness of fit

Test statistic

$$\chi^2 = \sum_{i=1}^m \frac{(o_i - e_i)^2}{e_i}$$

where m is the number of categories.

Hypothesis $H : F = F_0$.

- Calculate the expected class frequencies under F_0 .
- Calculate the χ^2 test statistic given above.
- Determine the degrees of freedom, ν say.
- Obtain critical value from χ^2 tables, $df = \nu$.
- Reject $H : F = F_0$ at the 100 p % level of significance if $\chi^2 > c$ where c is the tabulated critical value.