

Basic statistics formulas

Sets

De Morgan's Law

$$(A \cup B)^c = A^c \cap B^c \quad \& \quad (A \cap B)^c = A^c \cup B^c$$

Commutativity

$$A \cup B = B \cup A \quad \text{and} \quad A \cap B = B \cap A$$

Associativity

$$(A \cup B) \cup C = A \cup (B \cup C) \\ (A \cap B) \cap C = A \cap (B \cap C)$$

Distributivity

$$(A \cup B) \cap C = (A \cap C) \cup (B \cap C) \\ (A \cap B) \cup C = (A \cup C) \cap (B \cup C)$$

Probability

- $p(A) = 1 - p(A^c)$
- $p(A \cup B) = p(A) + p(B) - p(A \cap B)$
- If events A and B are mutually independent then $p(A \cap B) = p(A)p(B)$
- $p(A|B) = p(A \cap B)/p(B)$ so long as $p(B) > 0$
- $p(A \cap B) = p(A|B)p(B) = p(B|A)p(A)$ so long as $p(A) > 0$ and $p(B) > 0$
- If $\{B_1, B_2, \dots, B_k\}$ is a set of mutually exclusive and exhaustive events, then

$$p(A) = p(A|B_1)p(B_1) + p(A|B_2)p(B_2) + \dots + p(A|B_k)p(B_k)$$

Measures of Location

Sample mean

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

Median (for raw data i.e list of numbers ungrouped)

List the numbers in ascending order. Median is:

(n+1)/2 value if n is odd;

Mean of n/2th and (n+1)/2th values if n is even.

Measures of spread

Sample variance

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

Sample standard deviation, s

$$s = \sqrt{\text{variance}}$$

Range

Largest value – smallest value

Interquartile range (IQR)

Upper quartile – lower quartile

Coefficient of variation

$$s / \bar{x}$$

Expectation, variance, and covariance

If random variable X is discrete:

$$E(X) = \sum x_i p(x_i)$$

which is calculated over all possible values of X.

Let $g(X)$ denote a function of discrete X, then:

$$E(g(X)) = \sum g(x_i) p(x_i)$$

Expectation rules

Let X, Y, Z denote random variables; a, b denote constants

E1. $E(a) = a$

E2. $E(aX) = a E(X)$

E3. $E(X+Y) = E(X) + E(Y)$

Variance rules

V1. $\text{var}(a) = 0$

V2. $\text{var}(aX) = a^2 \text{var}(X)$

V3. $\text{var}(X \pm Y) = \text{var}(X) + \text{var}(Y) \pm 2 \text{Cov}(X, Y)$

Covariance rules

C1. $\text{Cov}(a, X) = 0$

C2. $\text{Cov}(aX, bY) = ab \text{Cov}(X, Y)$

C3. $\text{Cov}(X+Y, Z) = \text{Cov}(X, Z) + \text{Cov}(Y, Z)$

Normal density function

The normal density function (aka normal distribution) has 2 parameters: mean and variance. For the normal distribution:

90% of data falls within $\pm 1.65\sigma$

95% of data falls within $\pm 1.95\sigma$

Standardizing (Z-score)

$$z = \frac{x - \mu}{\sigma}$$

Sampling distribution of sample mean

Suppose $X \sim N(\mu, \sigma^2)$, then the distribution of the sample mean \bar{X} is

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

The standard error of \bar{X} is

$$\sqrt{\frac{\sigma^2}{n}}$$

This result is true if X does not follow the normal distribution but n is large (and then the result follows because of the Central Limit Theorem)

Confidence intervals for the mean

Parameter	Assumptions	Formula
Mean μ	Data normally distributed or n is large (n>30); σ^2 known	$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$
	Data normally distributed; n small; σ^2 unknown	$\bar{x} \pm t_{\alpha/2, n-1} \frac{s}{\sqrt{n}}$
Difference in means $\mu_X - \mu_Y$ Case of 2 independent distributions	Data are normally distributed; σ_X^2, σ_Y^2 are known	$(\bar{x} - \bar{y}) \pm z_{\alpha/2} \sqrt{\left(\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y} \right)}$
	Variances unknown; Large samples	$(\bar{x} - \bar{y}) \pm z_{\alpha/2} \sqrt{\left(\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y} \right)}$
	Data are normally distributed; σ_X^2, σ_Y^2 are unknown but $\sigma_X^2 = \sigma_Y^2$	$(\bar{x} - \bar{y}) \pm t_{\alpha/2, n_X+n_Y-2} \sqrt{s_p^2 \left(\frac{1}{n_X} + \frac{1}{n_Y} \right)}$ Where the estimate of the pooled variance is $s_p^2 = \frac{(n_X - 1)s_X^2 + (n_Y - 1)s_Y^2}{n_X + n_Y - 2}$

Hypothesis test for the mean

Hypothesis	Assumption	Test equation
Testing a single mean equals a value "a" $H_o : \mu = a$	Data normally distributed, or large sample; σ^2 known	$\frac{\bar{x} - a}{\sigma / \sqrt{n}}$ Z-table for critical value
	Data normally distributed σ^2 unknown	$\frac{\bar{x} - a}{s / \sqrt{n}}$ t-table with df = n-1 for critical value
Testing the difference between 2 means equals a number "a" (which includes the case of a=0 which is a test for no difference between means) $H_o : \mu_x - \mu_y = a$ Case of 2 independent distributions	Data normally distributed, or large sample; Independent samples; σ_x^2, σ_y^2 are known	$\frac{\bar{x} - \bar{y} - a}{\sqrt{\left(\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}\right)}}$ Z-table for critical value
	Data normally distributed, or large sample; Independent samples; σ_x^2, σ_y^2 are unknown	$\frac{\bar{x} - \bar{y} - a}{\sqrt{\left(\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}\right)}}$ Z-table for critical value
	Data are normally distributed; σ_x^2, σ_y^2 are unknown but $\sigma_x^2 = \sigma_y^2$	$\frac{\bar{x} - \bar{y} - a}{\sqrt{s_p^2 \left(\frac{1}{n_x} + \frac{1}{n_y}\right)}}$ Where the estimate of the pooled variance is $s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}$

Covariance and correlation

<u>Parameter</u>	<u>Population formula</u>	<u>Sample formula</u>
Covariance between variables X and Y	$COV(X,Y) = E(XY) - E(X)E(Y)$	$\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$
Pearson's correlation	$\frac{Cov(X, Y)}{\sigma_x \sigma_y}$	$\frac{\sum xy - n\bar{x}\bar{y}}{s_x s_y}$

Simple linear regression

Model: for $i = 1, 2, \dots, n$

$$y_i = \alpha + \beta x_i + u_i$$

OLS estimators

Intercept	Slope
$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$ $\text{var}(\hat{\alpha}) = \frac{\sigma^2 \sum x_i^2}{n(\sum x_i^2 - n\bar{x}^2)}$	$\hat{\beta} = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}$ $\text{var}(\hat{\beta}) = \frac{\sigma^2}{\sum x_i^2 - n\bar{x}^2}$

Estimator for variance of error term u is

$$\frac{\sum (y_i - \hat{y}_i)^2}{n-2}$$