Data Analysis for Public Policy Formulas – Autumn 2025

Mean of population (μ): $\mu = \frac{\sum y_i}{n}$

Mean of sample ($\bar{y}$): $\bar{y} = \frac{\sum y_i}{n}$

Expected value of probability distribution = $\mu = \sum yP(y)$

Mean of difference (2 dependent samples) : $\bar{y}_d = \frac{\sum(y_{2_i} - y_{1_i})}{n}$

Sum of squared deviations (SSD): $SSD = \sum(y_i - \bar{y})^2$

Standard deviation of population dataset (σ) : $\sigma = \sqrt{\frac{SSD}{n}}$

Standard deviation of sample dataset (s) : $s = \sqrt{\frac{SSD}{n-1}}$

Standard deviation of probability distribution : $\sigma = \sqrt{\sum(y - \bar{y})^2 \quad P(y)}$

Standard deviation two independent mean samples when pooling the variance (s_p) : $s_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$

Standard deviation of difference in 2 dependent samples (s_d) : $s_d = \sqrt{\frac{\sum(\bar{y}_{d_i} - \bar{y}_d)^2}{n-1}}$

Standard deviation of population probability distribution: $\sigma_\pi = \sqrt{\pi(1 - \pi)}$

Standard deviation of sample probability distribution : $s_{\hat{\pi}} = \sqrt{\hat{\pi}(1-\hat{\pi})}$

Standard error of population mean ($\sigma_{\bar{y}}$): $\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$

Estimated standard error of sample mean ($se_{\bar{y}}$): $se_{\bar{y}} = \frac{s}{\sqrt{n}}$

Estimated Standard error of the difference in 2 dependent samples: $se_d = \frac{s_d}{\sqrt{n}}$

Standard error of 1-sample probability, Score test - where the null hypothesis is assumed to be true (se_0): $se_0 = \sqrt{\frac{\pi_0(1-\pi_0)}{n}}$

Standard error of 1-sample probability, Wald-test ($se_{\hat{\pi}}$): $se_{\hat{\pi}} = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$

Standard error of 2-sample probability (se_0): $se_0 = \sqrt{\hat{\pi}(1-\hat{\pi})(\frac{1}{n_1}+\frac{1}{n_2})}$

Where $\hat{\pi} = \frac{\#\ success\ in\ \hat{\pi}_1 + \#\ success\ in\ \hat{\pi}_2}{\#\ tries\ in\ \hat{\pi}_1 + \#\ tries\ in\ \hat{\pi}_2}$

z-score for population mean or observation: $z = \frac{y-\mu}{\sigma}$

z-score for one sample proportion: $z = \frac{\hat{\pi}-\pi}{se_0}$

z-score for two sample comparison of proportions: $z = \frac{\hat{\pi}_2-\hat{\pi}_1-0}{se_0}$,

t-score for one sample mean : $t = \frac{\bar{y}-\mu_o}{se_{\bar{y}}}$

t-score for two sample comparison of means (unequal variances): $t = \frac{\bar{y}_2-\bar{y}_1-\mu_o}{se}$,

where $se = \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$

t-score for two sample comparison of means (equal/pooled variances):

$$t = \frac{\bar{y}_2 - \bar{y}_1 - \mu_o}{se_p},$$

where $se_p = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$

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

t-score for two sample comparison of dependent samples: $t = \frac{\bar{y}_d - \mu_o}{se_d}$

Confidence interval : CI = point estimate $CI = point\ estimate\ \pm Critical\ value * se$

Sample size necessary for a given margin of error (M)
when estimating population mean: $n = \sigma^2 (\frac{z}{M})^2$

Sample size necessary for a given margin of error (M)
when estimating proportion: $n = \pi(1-\pi)\ (\frac{z}{M})^2$

Expected value of cell (f_e) : $f_e = \frac{(row\ total * column\ total)}{sum\ total}$

Chi2 test of independence (χ^2) : $\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$, $df = (r-1)(c-1)$

Generalized Linear equation : $y = \alpha + \beta x$

Slope (β) : $\beta = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$

Intercept (α) : $\alpha = \bar{y} - \beta\bar{x}$

Sum of squared residuals (SSE) : $SSE = \sum(y - \hat{y})^2$

Pearson's r (r) : $r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{[\sum(x_i - \bar{x})^2][\sum(y_i - \bar{y})^2]}} = \left(\frac{s_x}{s_y}\right) b$