Statistical Formulas

Descriptive Statistics

Frequencies

Sample size n num of individuals in the sample.

Absolute frequency n_i. (num of x_i in the sample)

Relative frequency $f_i = n_i/n$.

Cumulative absolute freq $N_i = \sum_{k=0}^{i} n_i$.

Cumulative relative freq $F_i = N_i/n$.

Dispersion statistics

Interquartile range $IQR = Q_3 - Q_1$.

Variance $s^2 = \dfrac{\sum (x_i - \bar{x})^2 n_i}{n} = \dfrac{\sum x_i^2 n_i}{n} - \bar{x}^2$

Standard deviation $s = +\sqrt{s^2}$.

Coefficient of variation $cv = \dfrac{s}{|\bar{x}|}$.

Central tendency statistics

Mean $\bar{x} = \dfrac{\sum x_i n_i}{n}$.

Median me The value with cum.rel.freq. $F_{me} = 0.5$.

Mode mo The most frequent value.

Shape statistics

Coefficient of skewness $g_1 = \dfrac{\sum (x_i - \bar{x})^3 f_i}{s^3}$.

Coefficient of kurtosis $g_2 = \dfrac{\sum (x_i - \bar{x})^4 f_i}{s^4} - 3$.

Position statistics

Quartiles Q_1, Q_2, Q_3 divide the distribution into 4 equal parts. Their cum.rel.freqs. are $F_{Q_1} = 0.25$, $F_{Q_2} = 0.5$ and $F_{Q_3} = 0.75$.

Percentiles $P_1, P_2, \cdots, P_{99}$ divide the distribution into 100 equal parts.
The cum.rel.freq. is $F_{P_i} = i/100$.

Standarization

$$z = \frac{x - \bar{x}}{s_x}$$

Regression and correlation

Linear regression

Covariance $s_{xy} = \dfrac{\sum x_i y_j n_{ij}}{n} - \bar{x}\bar{y}$.

Regression lines :

y on x: $y = \bar{y} + \dfrac{s_{xy}}{s_x^2}(x - \bar{x})$

x on y: $x = \bar{x} + \dfrac{s_{xy}}{s_y^2}(y - \bar{y})$

Regression coefficients

$$(y \text{ on } x)\ b_{yx} = \frac{s_{xy}}{s_x^2} \qquad (x \text{ on } y)\ b_{xy} = \frac{s_{xy}}{s_y^2}$$

Coefficient of determination

$$r^2 = \frac{s_{xy}^2}{s_x^2 s_y^2} \qquad 0 \le r^2 \le 1$$

Correlation coefficient

$$r = \frac{s_{xy}}{s_x s_y}. \qquad -1 \le r \le 1$$

Non-linear regression

Exponential model $y = e^{a+bx}$
Apply the logarithm to the dependent variable and compute the line $\log y = a + bx$.

Logarithmic model $y = a + b \log x$
Apply the logarithm to the independent variable and compute the line $y = a + b \log x$.

Potential model $y = ax^b$
Apply the logarithm to both variables and compute the line $\log y = a + b \log x$.

Probability

Basic probability

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

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

Difference $P(A-B) = P(A) - P(A\cap B)$.

Contrary $P(\overline{A}) = 1 - P(A)$.

Conditional probability $P(A|B) = \dfrac{P(A\cap B)}{P(B)}$.

Independent events $P(A|B) = P(A)$.

Total prob. Theorem

$$P(B) = \sum_{i=1}^{n} P(A_i)P(B|A_i)$$

.

Bayes Theorem

$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{\sum_{i=1}^{n} P(A_i)P(B|A_i)}$$

.

Diagnostic tests

Prevalence Proportion of people with the disease $P(D)$.

Sensitivity $P(+|D)$.

Specificity $P(-|\overline{D})$.

Positive Predictive Value (PPV) $P(D|+)$.

Negative Predictive Value (NPV) $P(\overline{D}|-)$.

Positive Likelihood Ratio (LR+) $\dfrac{P(+|D)}{P(+|\overline{D})}$.

Negative Likelihood Ratio (LR-) $\dfrac{P(-|D)}{P(-|\overline{D})}$.

Random Variables

Discrete

Binomial probability function $B(n,p)$

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

Poisson probability function $P(\lambda)$

$$f(x) = e^{-\lambda}\frac{\lambda^x}{x!}$$

Law of rare events $B(n,p) \approx P(np)$ for $n \geq 30$ and $p \leq 0.1$.