

# Statistics 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**  $N_i = \sum_{k=0}^i n_k$

**Cumulative relative freq**  $F_i = N_i/n$

### Central tendency statistics

**Mean**  $\bar{x} = \frac{\sum x_i n_i}{n}$

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

**Mode**  $mo$  The most frequent value.

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

### Dispersion statistics

**Interquartile range**  $IQR = Q_3 - Q_1$

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

**Standard deviation**  $s = \sqrt{s^2}$

**Coefficient of variation**  $cv = \frac{s}{|\bar{x}|}$

### Shape statistics

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

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

### Linear transformations

**Linear transformation**  $y = a + bx$

$$\bar{y} = a + b\bar{x}$$

$$s_y = bs_x$$

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

## Regression and correlation

### Linear regression

**Covariance**  $s_{xy} = \frac{\sum x_i y_i n_{ij}}{n} - \bar{x}\bar{y}$

**Regression lines** :

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

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

**Regression coefficients**

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

**Coefficient of determination**

$$r^2 = \frac{s_{xy}^2}{s_x^2 s_y^2} \quad 0 \leq r^2 \leq 1$$

**Correlation coefficient**

$$r = \frac{s_{xy}}{s_x s_y} \quad -1 \leq r \leq 1$$

### 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  $y = a + b \log x$ .

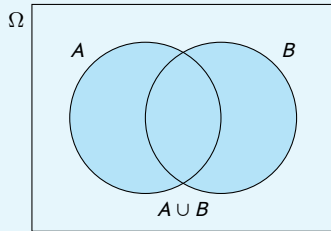
**Potential model**  $y = ax^b$

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

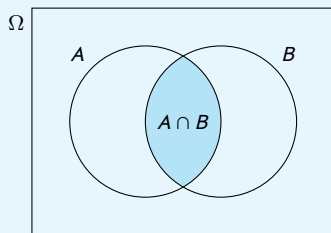
# Probability

## Event operations

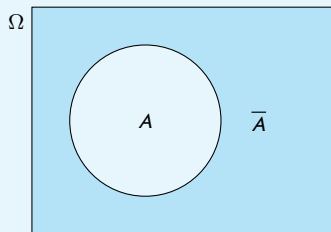
### Union



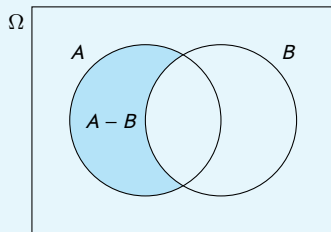
### Intersection



### Complement



### Difference



## Algebra of events

**Idempotency**  $A \cup A = A$ ,  $A \cap A = A$

**Commutative**  $A \cup B = B \cup A$ ,  $A \cap B = B \cap A$

**Associative**  $(A \cup B) \cup C = A \cup (B \cup C)$ ,  $(A \cap B) \cap C = A \cap (B \cap C)$

**Distributive**  $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$ ,  $(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$

**Neutral element**  $A \cup \emptyset = A$ ,  $A \cap \Omega = A$

**Absorbing element**  $A \cup \Omega = \Omega$ ,  $A \cap \emptyset = \emptyset$ .

**Complementary symmetric element**  $A \cup \bar{A} = \Omega$ ,  $A \cap \bar{A} = \emptyset$

**Double contrary**  $\bar{\bar{A}} = A$

**Morgan's laws**  $\overline{A \cup B} = \bar{A} \cap \bar{B}$ ,  $\overline{A \cap B} = \bar{A} \cup \bar{B}$

## 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**  $P(\bar{A}) = 1 - P(A)$

## Conditional probability

**Conditional probability**  $P(A|B) = \frac{P(A \cap B)}{P(B)}$

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

**Total probability Theorem**

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

**Bayes Theorem**

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

## Risks

	E	$\bar{E}$
Treatment	a	b
Control	c	d

**Prevalence** Proportion of individuals with E:  $P(E)$

**Incidence rate or absolute risk**  $R(E) = \frac{a}{a+b}$

**Odds**  $O(E) = \frac{a}{b}$

**Relative risk**  $RR(E) = \frac{a/(a+b)}{c/(c+d)}$

**Odds ratio**  $OR(E) = \frac{a/b}{c/d} = \frac{a \cdot d}{b \cdot c}$

## Diagnostic tests

	Disease D	No disease $\bar{D}$
Test +	VP	FP
Test -	FN	VN

**Sensitivity**  $P(+|D) = \frac{VP}{VP+FN}$

**Specificity**  $P(-|\bar{D}) = \frac{VN}{FP+VN}$

**Positive Predictive Value (PPV)**  $P(D|+) = \frac{VP}{VP+FP}$

**Negative Predictive Value (NPV)**  $P(\bar{D}|-) = \frac{VN}{FN+VN}$

**Positive Likelihood Ratio (LR+)**  $\frac{P(+|D)}{P(+|\bar{D})}$

**Negative Likelihood Ratio (LR-)**  $\frac{P(-|D)}{P(-|\bar{D})}$

## Random Variables

### Discrete

**Binomial probability function**  $B(n, p)$

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

**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  $p \leq 0.1$ .

### Continuous

**Normal**  $N(\mu, \sigma)$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

**Standard normal**  $N(0, 1)$

**Chi-square**  $\chi^2(n)$

$$X = Z_1^2 + \dots + Z_n^2,$$

where  $Z_i \sim N(0, 1)$ .

**Student's t**  $T(n)$

$$T = \frac{Z}{\sqrt{X/n}},$$

where  $Z \sim N(0, 1)$  and  $X \sim \chi^2(n)$ .

**Fisher's F**  $F(n, m)$

$$F = \frac{X/m}{Y/n},$$

where  $X \sim \chi^2(m)$  and  $Y \sim \chi^2(n)$ .