Version 1.0

**Prof. Dr. Florian Kauffeldt**

# Notation and Formulae Statistics

## Content

# 1 Notation (commonly used in Textbooks)

## 1.1 Population and Sample Elements

| **Symbol** | **Meaning** |
|---|---|
| $X, Y, Z,$ ...(capital letters) | Random variables (range = population elements) |
| $x, y, z,$ ...(lowercase letters) | A subset of the range of a random variable (e.g., a sample) |
| $x_i, y_i, z_i,$ ...(lowercase letters with subscript) | A single specific value (number) that a random variable may take on |
| $N$ | Population size |
| $n$ | Sample size |

## 1.2 Probabilities

| **Symbol** | **Meaning** |
|---|---|
| $P(x)$ | Short for $P(X \in x)$, i.e., the probability that event $x$ occurs.<br>Examples:<br>• Random variable takes on $x_i$ $(x = x_i)$: $P(x_i) \coloneqq P(X = x_i)$<br>• Random variable in interval $x = [x_i, x_j]$: $P([x_i, x_j]) \coloneqq P(x_i \leq X \leq x_j)$ |
| $P(x \mid y)$ | Short for $P(X \in x \mid Y \in y)$, i.e., the conditional probability that event $x$ occurs, given that event $y$ has occurred. |
| $P(x, y)$ | Short for $P(X \in x, Y \in y)$, i.e., the joint probability that $x$ and $y$ occur. |
| $f_X$ | Probability densitiy function (pdf) of a continuous random variable $X$ |
| $F_X$ | Cumulative distribution function (cdf) of a continuous random variable $X$ |

## 1.3 Parameters and Statistics

| **Paramter / Statistic** | **Notation Parameter** | **Notation Statistic** |
|---|---|---|
| Population / sample *mean* | $\mu$ | $\bar{x}$ |
| Population / sample *standard deviation and variance* | $\sigma \text{ or } sd(X)$<br>$\sigma^2 \text{ or } var(X)$ | $s \text{ or } sd(x)$<br>$s^2 \text{ or } var(x)$ |
| Population / sample *covariance* of the variables $X$ and $Y$ | $\sigma_{XY} \text{ or } cov(X,Y)$ | $s_{xy} \text{ or } cov(x,y)$ |
| Population / sample *Pearson correlation coefficient* of the variables $X$ and $Y$ | $\rho_{XY}$ | $r_{xy}$ |
| Population / sample *Pearson partial correlation coefficient* of the variables $X$ and $Y$ when controlling for $Z$ | $\rho_{XY\|Z}$ | $r_{xy\|z}$ |
| Population / sample Spearman rho correlation coefficient | $\rho_{XY}^{S}$ | $r_{xy}^{S}$ |

## 1.4 Special Symbols

| **Symbol** | **Meaning** |
|---|---|
| $\bar{X}$ | Random variable of the sample mean |
| $E[X], E[x]$ | Expected value of population / sample |
| $se$ | Standard error (= standard deviation) of the distribution (random variable) of a statistic |
| $\hat{\mu}$ | Estimate of the population mean |
| $\hat{\sigma}^2$ | Estimate of the population variance |
| $H_0, H_1$ | Null / Alternative Hypothesis |

# 2 Formulae for Variance, Covariance and Correlation

## 2.1 Expected Value

Let x be a sample of size $n$ with $v$ unique values, then:

$$E[x] = \frac{frequency\ x_1}{n} \cdot x_1 + \cdots + \frac{frequency\ x_v}{n} \cdot x_v = P(x_1) \cdot x_1 + \cdots + P(x_v) \cdot x_v,$$

i.e., the expected value is the probability-weighted sum of the unique values of a (sample) random variable.

## 2.2 Variance and Covariance

Let $(x, y)$ be a sample of size $n$, then:

$$var(x) = \frac{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n} = E[(x - \bar{x})^2]$$

$$cov(x, y) = \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + \cdots + (x_n - \bar{x})(y_n - \bar{y})}{n} = E[(x - \bar{x})(y - \bar{y})]$$

Observe that $cov(x, x) = var(x)$.

## 2.3 Pearson Correlation Coefficient

Sample Pearson correlation coefficient:

$$r_{xy} = \frac{cov(x, y)}{sd(x) \cdot sd(y)}$$

## 2.4 Pearson Partial Correlation Coefficient

Sample Pearson partial correlation coefficient for *one contol variable* ($z$):

$$r_{xy|z} = \frac{r_{xy} - r_{xz} \cdot r_{yz}}{\sqrt{1 - r_{xz}^2} \cdot \sqrt{1 - r_{yz}^2}}$$

Sample Pearson partial correlation coefficient for $k$ contol variables ($z_1, \dots, z_k$):

$$r_{xy|z_1 \dots z_k} = \frac{r_{xy} - (r_{xz_1} \cdot r_{yz_1} + \cdots + r_{xz_1} \cdot r_{yz_k})}{\sqrt{1 - (r_{xz_1}^2 + \cdots + r_{xz_k}^2)} \cdot \sqrt{1 - (r_{yz_1}^2 + \cdots + r_{yz_k}^2)}}$$

# 3 Formulae for Conditional Probability, Independence, and Standard Score

## 3.1 Conditional Probability

The probability that $x$ occurs, given $y$ has occured is:

$$P(x \mid y) = \frac{P(y \mid x) \cdot P(x)}{P(y)} = \frac{P(x, y)}{P(y)}$$

## 3.2 Stochastic Independence

Two random variables are stochastically independent if

$$P(x, y) = P(x) \cdot P(y) \; for \; all \; x, y$$

## 3.3 Standard Score (z Score)

The z score of a value $x_i$ of a random variable $X$ is:

$$z(x_i) = \frac{x_i - mean_X}{sd(X)}$$

If $X$ is normally distributed $Z_X$ is standard normally distributed.

# 4 Formulae for Inferential Statistics

## 4.1 Finite Sample Correction of Standard Error

If

(i). sampling is done without replacement from a finite population and

(ii). the sample size ($n$) is large relative to the population size ($N$),

the standard error ($se$) must be muliplied by $fpc$ (finite population correction):

$$se = se_{without\ fpc} \cdot fpc,$$

where $fpc = \sqrt{\frac{N-n}{N-1}}$.

## 4.2 Point Estimators

- The estimator for the population mean is the sample mean:

$$\hat{\mu} = \bar{x}$$

- The estimator for the population variance is the Bessel-corrected sample variance:

$$\hat{\sigma}^2 = s^2 \cdot \frac{n}{n-1}$$

- The estimator for a correlation coefficient is the sample correlation coefficient:

$$\hat{\rho} = r$$

## 4.3 Interval Estimators

- The general formula for an interval estimator is:

$$point\ estimate \pm moe,$$

  where the margin of error (moe) is some multiple of the standard error (se):

$$moe = m \times se$$

- Interval estimator of mean with confidence level $\alpha$ and known population variance:

$$\bar{x} \pm z_{(1-\alpha)} \cdot se$$

## 4.4 P-Value for z Tests

Let $ts$ be the test statistic, then

$$p = 1 - \Phi(|ts|) = P(X > |ts|),$$

where $\Phi$ is the cumulative distribution function of the standard normal distribution. Then:

$$One - tailed\ p - value = p$$

$$Two - tailed\ p - value = p \times 2$$

# 5 Test Statistics of Parametric Tests

## 5.1 One- and Two-Sample Z Test of Means

### 5.1.1 One-sample z test of means

Let $x$ be a sample of size $n$ (with sample mean $\bar{x}$) of random variable $X$ (with poulation standard deviation $\sigma$).

- The $standard\ error$ $(se)$ of the distribution of sample means ($\bar{X}$) is:

$$se = \frac{\sigma}{\sqrt{n}}$$

- The $test\ statistic$ of testing $X$ against a pre-specified level $\mu_0$ is:

$$z = \frac{\bar{x} - \mu_0}{se}$$

### 5.1.2 Two-sample z test of means

Let $x, y$ be samples of size $n_x, n_y$ (with sample means $\bar{x}, \bar{y}$) of random variable $X, Y$ (with poulation standard deviation $\sigma_X, \sigma_Y$).

Then, the random variable $D = X - Y$ has sample mean $\bar{d} = \bar{x} - \bar{y}$.

- The $pooled\ standard\ error$ $(se)$ of the distribution of sample means ($\bar{D}$) is:

$$se_{\bar{D}} = \sqrt{se_X^2 + se_Y^2},$$

  where $se_X = \sigma_X/\sqrt{n_x}$ and $se_Y = \sigma_Y/\sqrt{n_y}$.

- The $test\ statistic$ of testing $D$ against a pre-specified level $\mu_0$ (= 0, usually) is:

$$z = \frac{\bar{d} - \mu_0}{se_D}$$

## 5.2 Correlation Tests

### 5.2.1 Pearson Correlation Test

Let $r_{xy}$ be the sample correlation coefficient of a sample $(x, y)$ of size $n$.

- The $standard\ error$ $(se_R)$ of the distribution of correlation coeffients ($R$) is

$$se_R = \sqrt{\frac{1 - r_{xy}^2}{n - 2}}.$$

- The $test\ statistic$ of $R$ against a pre-specified level $\rho_0$ (= 0, usually) is:

$$t_{n-2} = \frac{r_{xy} - \rho_0}{se_R},$$

  which follows a t-distribution with n-2 degrees of freedom.

### 5.2.2 Pearson Partial Correlation Test

Let $r_{xy|z_1 \dots z_k}$ be the sample partial correlation coefficient of a sample $(x, y, z_1, \dots, z_k)$ of size $n$.

- The $standard\ error$ $(se_R)$ of the distribution of correlation coeffients ($R$) is

$$se_R = \sqrt{\frac{1 - r_{xy}^2}{n - k - 3}},$$

  where $k$ denotes the number of control variables.

- The $test\ statistic$ of $R$ against a pre-specified level $\rho_0$ (= 0, usually) is:

$$t_{n-k-3} = \frac{r_{xy|z_1 \dots z_k} - \rho_0}{se_R},$$

  which follows a t-distribution with $n - k - 3$ degrees of freedom.

### 5.2.3 Fisher Transformation

- The $Fisher\ transformation$ of a correlation coefficient $r$ is:

$$r^f = \frac{1}{2} \ln \left( \frac{1 + r}{1 - r} \right) = \text{artanh}\,(r)$$

  and has the following standard error ($se_f$):

$$se_f = \sqrt{\frac{1}{n - 3}}$$

- The Fisher test statistic is then:

$$z_{fisher} = \frac{r_f}{se_f}$$

  and follows a standard normal distribution if $n$ is sufficiently large.

A Fisher z Test based on the Fisher test statistic can be used to test any of the aforementioned correlations plus Spearman rank correlation.

# 6 Test Statistics of Non-Parametric Tests

## 6.1 Mann Whitney U Test

- The $U$ statistic is:

$$U = n_{max} \cdot n_{min} + \frac{n_{max}(n_{max}+1)}{2} - R_{max},$$

where $n_{max}, n_{min}$ are the sample sizes of the samples with the highest / lowest rank sum and $R_{max}$ ist he highest rank sum.

- If there are no tied ranks, the standard error of U is:

$$se_u \sqrt{\frac{n_{max} \cdot n_{min}(n_{max}+n_{min}+1)}{12}}.$$

- The normal approximated $test\ statistic$ is:

$$z_U = \frac{U - \frac{n_{max} \times n_{min}}{2}}{se_U}.$$

## 6.2 Chi2 Test

- The $chi^2$ statistic is:

$$\chi^2 = \sum \frac{(O-E)^2}{E},$$

where $O$ are the observed and $E$ are the expected frequencies.

- The normal approximated $test\ statistic$ is:

$$z_{\chi^2} = \frac{\chi^2 - dof}{\sqrt{2 \times dof}},$$

where dof = degrees of freedom.

# 7 $\chi^2$ – and T-Distribution

Let $Z_1, \dots, Z_k$ be independent standard normally distributed random variables, then the sum of their squares follows a $\chi^2$ –distribution with $k$ degrees of freedom:

$$\chi_k^2 = Z_1^2 + \cdots + Z_k^2.$$

Let $Z$ be a standard normally distributed and $\chi_k^2$ a $\chi^2$ –distributed random variable, then the following is t-distributed with $k$ degrees of freedom:

$$T_k = \frac{Z}{\sqrt{\chi_k^2/k}}.$$