Formulas and tables
in mathematical statistics
1. Combinatorics
\[ \binom{n}{k} = \frac{n!}{k!(n-k)!} \]. Interpretation: \( \binom{n}{k} \) is the number of subsets of size \( k \) formed from a set of \( n \) elements.

2. Random variables
\[
\begin{align*}
V(X) &= E(X^2) - (E(X))^2 \\
C(X,Y) &= E((X - E(X))(Y - E(Y))) = E(XY) - E(X)E(Y) \\
\rho(X,Y) &= \frac{C(X,Y)}{D(X)D(Y)}
\end{align*}
\]

3. Discrete distributions

Binomial distribution
\( X \) is Bin\((n,p)\) if \( p_X(k) = \binom{n}{k} p^k (1-p)^{n-k} \), \( k = 0,1,\ldots,n \), where \( 0 < p < 1 \).
\[
E(X) = np, \quad V(X) = np(1-p)
\]

"For-the-first-time"-distribution
\( X \) is fft\((p)\) if \( p_X(k) = p(1-p)^{k-1} \), \( k = 1,2,3,\ldots \), where \( 0 < p < 1 \).
\[
E(X) = \frac{1}{p}, \quad V(X) = \frac{1-p}{p^2}
\]

Hypergeometric distribution
\( X \) is Hyp\((N,n,p)\) if \( p_X(k) = \frac{\binom{Np}{k} \binom{N-n}{n-k}}{\binom{N}{n}} \), \( 0 \leq k \leq Np \),
\[
0 \leq n - k \leq N(1-p), \text{ where } N, Np \text{ and } n \text{ are positive integers and } N \geq 2, \quad n < N, \quad 0 < p < 1. \quad E(X) = np, \quad V(X) = \frac{N-n}{N-1} \cdot np(1-p)
\]

Poisson distribution
\( X \) is Po\((\mu)\) where \( \mu > 0 \) if \( p_X(k) = \frac{\mu^k}{k!} e^{-\mu} \), \( k = 0,1,2,\ldots \)
\[
E(X) = \mu, \quad V(X) = \mu
\]

4. Continuous distributions

Uniform distribution
\( X \) is U\((a,b)\) where \( a < b \) if \( f_X(x) = \begin{cases} 
\frac{1}{b-a} & \text{for } a < x < b \\
0 & \text{otherwise}
\end{cases} \)
\[
E(X) = \frac{a+b}{2}, \quad V(X) = \frac{(b-a)^2}{12}
\]
Exponential distribution

$X$ is Exp($\lambda$) where $\lambda > 0$ if

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

$E(X) = \frac{1}{\lambda}$, $V(X) = \frac{1}{\lambda^2}$

Normal distribution

$X$ is $N(\mu, \sigma)$ if

$$f_X(x) = \frac{1}{\sqrt{2\pi}\cdot \sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty, \quad \sigma > 0.$$  

$E(X) = \mu$, $V(X) = \sigma^2$

$X$ is $N(\mu, \sigma)$ if and only if $\frac{X-\mu}{\sigma}$ is $N(0,1)$.

If $Z$ is $N(0,1)$ then $Z$ has the distribution function $\Phi(x)$ according to Table 1 and the density function $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$, $-\infty < x < \infty$.

A linear combination $\sum a_iX_i + b$ of independent, normally distributed random variables is normally distributed.

5. Central limit theorem

If $X_1, X_2, \ldots, X_n$ are independent identically distributed random variables with expectation $\mu$ and standard deviation $\sigma > 0$ then $Y_n = X_1 + \cdots + X_n$ is approximatively $N(n\mu, \sigma\sqrt{n})$ if $n$ is large.

6. Approximation

Hyp($N,n,p$) $\sim$ Bin($n,p$) if $\frac{n}{N} \leq 0.1$

Bin($n,p$) $\sim$ Po($np$) if $p \leq 0.1$

Bin($n,p$) $\sim$ $N(np, \sqrt{np(1-p)})$ if $np(1-p) \geq 10$

Po($\mu$) $\sim$ $N(\mu, \sqrt{\mu})$ if $\mu \geq 15$

7. Chebychev’s inequality

If $E(X) = \mu$ and $D(X) = \sigma > 0$ then for every $k > 0$

$$P(|X - \mu| > k\sigma) \leq \frac{1}{k^2}$$

8. Statistical material

$$\bar{x} = \frac{1}{n} \sum_{j=1}^{n} x_j$$

$$s^2 = \frac{1}{n-1} \sum_{j=1}^{n} (x_j - \bar{x})^2 = \frac{1}{n-1} \left[ \sum_{j=1}^{n} x_j^2 - \frac{1}{n} (\sum_{j=1}^{n} x_j)^2 \right]$$
9. Point estimation

9.1 Method of Maximum likelihood

Let \( x_i \) be an observation of \( X_i \), \( i = 1, 2, \ldots, n \), where the distribution of \( X_i \) depends on an unknown parameter \( \theta \). The value \( \theta_{\text{obs}}^* \) which maximizes the \( L \)-function

\[
L(\theta) = \begin{cases} 
    p_{X_1, \ldots, X_n}(x_1, \ldots, x_n; \theta) = \text{(if independent)} = p_{X_1}(x_1; \theta) \cdots p_{X_n}(x_n; \theta) \\
    f_{X_1, \ldots, X_n}(x_1, \ldots, x_n; \theta) = \text{(if independent)} = f_{X_1}(x_1; \theta) \cdots f_{X_n}(x_n; \theta)
\end{cases}
\]

is called the Maximum likelihood estimate (ML estimate) of \( \theta \).

9.2 Method of Least squares

Let \( x_i \) be an observation of \( X_i \), \( i = 1, 2, \ldots, n \), and suppose that

\[
E(X_i) = \mu_i(\theta_1, \theta_2, \ldots, \theta_k) \quad \text{and} \quad V(X_i) = \sigma^2,
\]

where \( \theta_1, \theta_2, \ldots, \theta_k \) are unknown parameters and \( X_1, X_2, \ldots, X_n \) are independent.

The estimates of Least squares (LS estimates) of \( \theta_1, \theta_2, \ldots, \theta_k \) are the values \( (\theta_1)_{\text{obs}}^*, (\theta_2)_{\text{obs}}^*, \ldots, (\theta_k)_{\text{obs}}^* \) which minimize the sum of squares

\[
Q = Q(\theta_1, \theta_2, \ldots, \theta_k) = \sum_{i=1}^{n} (x_i - \mu_i(\theta_1, \theta_2, \ldots, \theta_k))^2.
\]

9.3 Mean error

An estimate of \( D(\theta^*) \) is called the mean error of \( \theta^* \) and is written \( d(\theta^*) \).

9.4 Error propagation

With notations and assumptions according to the text-book we have

a) \( E(g(\theta^*)) \approx g(\theta_{\text{obs}}^*) \)

\[
D(g(\theta^*)) \approx \left| g'(\theta_{\text{obs}}^*) \right| \cdot D(\theta^*)
\]

b) \( E(g(\theta_1^*, \ldots, \theta_n^*)) \approx g((\theta_1)_{\text{obs}}^*, \ldots, (\theta_n)_{\text{obs}}^*) \)

\[
V(g(\theta_1^*, \ldots, \theta_n^*)) \approx \sum_{i=1}^{n} \sum_{j=1}^{n} C(\theta_i^*, \theta_j^*) \cdot \left[ \frac{\partial g}{\partial x_i} \cdot \frac{\partial g}{\partial x_j} \right]_{x_k = (\theta_{\text{obs}}^*)_{k=1,\ldots,n}}
\]

10. Some common distributions in statistics

\( \chi^2 \)-distribution

If \( X_1, X_2, \ldots, X_f \) are independent and \( N(0, 1) \), we have that

\[
\sum_{k=1}^{f} X_k^2 \quad \text{is} \quad \chi^2(f) \quad \text{distributed}.
\]

\( t \)-distribution

If \( X \) is \( N(0, 1) \) and \( Y \) is \( \chi^2(f) \) and if \( X \) and \( Y \) are independent, we have that

\[
\frac{X}{\sqrt{Y/f}} \quad \text{is} \quad t(f) \quad \text{distributed}.
\]
11. Distributions for sample variables when the sample is normally distributed

11.1 Let $X_1, \ldots, X_n$ be independent random variables which are all $N(\mu, \sigma)$.

Then we have:

a) $\bar{X}$ is $N\left( \mu, \frac{\sigma}{\sqrt{n}} \right)$

b) $\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2}$ is $\chi^2(n - 1)$

c) $\bar{X}$ and $S^2$ are independent

d) $\frac{\bar{X} - \mu}{S/\sqrt{n}}$ is $t(n - 1)$

11.2 Let $X_1, \ldots, X_{n_1}$ be $N(\mu_1, \sigma)$ and $Y_1, \ldots, Y_{n_2}$ be $N(\mu_2, \sigma)$ and all random variables are supposed to be independent. Then we have:

a) $\bar{X} - \bar{Y}$ is $N\left( \mu_1 - \mu_2, \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right)$

b) $\frac{(n_1 + n_2 - 2)S^2}{\sigma^2}$ is $\chi^2(n_1 + n_2 - 2)$ where $S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$

$$S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \bar{X})^2$$ and $$S_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2$$

c) $\bar{X} - \bar{Y}$ and $S^2$ are independent

d) $\frac{\bar{X} - \mu - (\mu_1 - \mu_2)}{S/\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ is $t(n_1 + n_2 - 2)$

11.3 Let $X_1, \ldots, X_{n_1}$ be $N(\mu_1, \sigma_1)$ and $Y_1, \ldots, Y_{n_2}$ be $N(\mu_2, \sigma_2)$ and all random variables are supposed to be independent. Then we have:

$\bar{X} - \bar{Y}$ is $N\left( \mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right)$

12. Confidence intervals

12.1 $\lambda$-method

Let $\theta^*$ be $N(\theta, D)$ where $D$ is known and $\theta$ unknown. Then

$\theta_{\text{obs}}^* \pm D \cdot \lambda_{\alpha/2}$

is a confidence interval for $\theta$ with the confidence level $1 - \alpha$.
12.2 $t$-method

Let $\theta^*$ be $N(\theta, D)$ where $D$ and $\theta$ are unknown and $D$ does not depend on $\theta$. Let $D_{\text{obs}}^*$ be a point estimate of $D$ such that $\frac{\theta^* - \theta}{D^*}$ is $t(f)$. Then $\theta_{\text{obs}}^* \pm D_{\text{obs}}^* \cdot \frac{t(\alpha/2)}{f}$ is a confidence interval for $\theta$ with the confidence level $1 - \alpha$.

12.3 Approximative method

Let $\theta^*$ be approximatively $N(\theta, D)$. Suppose that $D_{\text{obs}}^*$ is a suitable point estimate of $D$. Then $\theta_{\text{obs}}^* \pm D_{\text{obs}}^* \cdot \frac{\lambda(\alpha/2)}{f}$ is a confidence interval for $\theta$ with the approximate confidence level $1 - \alpha$.

12.4 Method based on $\chi^2$-distribution

Let $\theta_{\text{obs}}^*$ be a point estimate of a parameter $\theta$ such that $f \left( \frac{\theta^*}{\theta} \right)^2$ is $\chi^2(f)$. Then

$$\left( \frac{\theta_{\text{obs}}^*}{\frac{f}{\lambda_{\alpha/2}(f)}}, \frac{\theta_{\text{obs}}^*}{\frac{f}{\lambda_{1-\alpha/2}(f)}} \right)$$

is a confidence interval for $\theta$ with the confidence level $1 - \alpha$.

13. Linear regression

13.1 Distributions

Let $Y_i$ be $N(\alpha + \beta x_i, \sigma)$, $i = 1, 2, \ldots, n$, and independent. Then we have:

a) $\beta^* = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$ is $N\left( \beta, \frac{\sigma}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \right)$

b) $\alpha^* = \bar{Y} - \beta^* \bar{x}$ is $N\left( \alpha, \sigma \sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \right)$

c) $\alpha^* + \beta^* x_0$ is $N\left( \alpha + \beta x_0, \sigma \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \right)$

d) $\frac{(n-2)S^2}{\sigma^2}$ is $\chi^2(n-2)$ where $S^2 = \frac{1}{n-2} \sum_{i=1}^{n} (Y_i - \alpha^* - \beta^* x_i)^2$

e) $S^2$ is independent of $\alpha^*$ and $\beta^*$
13.2 Confidence intervals

\[ I_\alpha : \alpha^*_{\text{obs}} \pm t_{p/2}(n-2) s \sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \]

\[ I_\beta : \beta^*_{\text{obs}} \pm t_{p/2}(n-2) s \sqrt{\frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \]

\[ I_{\alpha + \beta x_0} : \alpha^*_{\text{obs}} + \beta^*_{\text{obs}} x_0 \pm t_{p/2}(n-2) s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \]

13.3 Computational aspects

\[ S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} (x_i - \bar{x})y_i = \sum_{i=1}^{n} x_i(y_i - \bar{y}) = \sum_{i=1}^{n} x_iy_i - n \bar{x} \bar{y} \]

\[ S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n \bar{x}^2 \]

\[ S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \]

\[ (n-2)s^2 = S_{yy} - (\beta^*_{\text{obs}})^2 S_{xx} = S_{yy} - \beta^*_{\text{obs}} \cdot S_{xy} = \min_{\alpha, \beta} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 \]

14. Hypothesis testing

14.1 Definitions

The significance level (probability of error of first kind) \( \alpha \) is

(the maximal value of) \( P(\text{reject } H_0) \) when the hypothesis \( H_0 \) is true.

The power function \( h(\theta) = P(\text{reject } H_0) \) when \( \theta \) is the correct parameter value.

14.2 Confidence method

Reject \( H_0 : \theta = \theta_0 \) on the level \( \alpha \) if \( \theta_0 \) does not fall within a suitably chosen confidence interval with the confidence level \( 1 - \alpha \).

14.3 \( \chi^2 \)-test

We make \( n \) independent repetitions of an experiment which gives one of the results \( A_1, A_2, \ldots, A_r \) with respective probabilities \( P(A_1), P(A_2), \ldots, P(A_r) \).

Let for \( j = 1, 2, \ldots, r \) the random variable \( X_j \) denote the number of trials which give the result \( A_j \).
Test of given distribution: We want to test $H_0: P(A_1) = p_1, P(A_2) = p_2, \ldots, P(A_r) = p_r$ for given probabilities $p_1, p_2, \ldots, p_r$. Then

\[ Q = \sum_{j=1}^{r} \frac{(x_j - np_j)^2}{np_j} \]

is an outcome of an approximatively $\chi^2(r-1)$-distributed random variable if $H_0$ is true and $np_j \geq 5$, $j = 1, 2, \ldots, r$.

If we estimate $k$ parameters out of our data, $\theta = (\theta_1, \ldots, \theta_k)$, in order to estimate $p_1, p_2, \ldots, p_r$ with $p_1(\theta^*_\text{obs}), p_2(\theta^*_\text{obs}), \ldots, p_r(\theta^*_\text{obs})$ then

\[ Q' = \sum_{j=1}^{r} \frac{(x_j - np_j(\theta^*_\text{obs}))^2}{np_j(\theta^*_\text{obs})} \]

is an outcome of an approximatively $\chi^2(r-k-1)$-distributed random variable.

Computational aspect: $Q = \sum_{j=1}^{r} \frac{x_j^2}{np_j} - n$, $Q' = \sum_{j=1}^{r} \frac{x_j^2(\theta^*_\text{obs})}{np_j(\theta^*_\text{obs})} - n$

Homogeneity test: We want to test if the probabilities for the results $A_1, A_2, \ldots, A_r$ are the same in $s$ series of trials. Introduce notation according to the following table:

<table>
<thead>
<tr>
<th>Series</th>
<th>Number of observations of</th>
<th>Number of trials</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$A_1$ $A_2$ $A_3$ $\ldots$ $A_r$</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>$x_{11}$ $x_{12}$ $x_{13}$ $\ldots$ $x_{1r}$</td>
<td>$n_1$</td>
</tr>
<tr>
<td>2</td>
<td>$x_{21}$ $x_{22}$ $x_{23}$ $\ldots$ $x_{2r}$</td>
<td>$n_2$</td>
</tr>
<tr>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
</tr>
<tr>
<td>$s$</td>
<td>$x_{s1}$ $x_{s2}$ $x_{s3}$ $\ldots$ $x_{sr}$</td>
<td>$n_s$</td>
</tr>
<tr>
<td>Column sum</td>
<td>$m_1$ $m_2$ $m_3$ $\ldots$ $m_r$</td>
<td>$N$</td>
</tr>
</tbody>
</table>

Compute $Q = \sum_{i=1}^{s} \sum_{j=1}^{r} \frac{\left(x_{ij} - \frac{n_i m_j}{N}\right)^2}{\frac{n_i m_j}{N}}$.

$Q$ is an outcome of an approximatively $\chi^2((r-1)(s-1))$-distributed random variable.

Contingency table (test av independence between rows and columns):

The same test variable and distribution as above.