Mathematical Statistics
KTH

Formulas and tables in mathematical statistics

Autumn 2016

1. Combinatorics

$\binom{n}{k} = \frac{n!}{k!(n-k)!}$. Interpretation: $\binom{n}{k}$ = number of subsets of size k formed from a set of n elements.

2. Random variables

$V(X) = E(X^2) - (E(X))^2$

$C(X,Y) = E\big((X - E(X))(Y - E(Y))\big) = E(XY) - E(X)E(Y)$

$\rho(X,Y) = \frac{C(X,Y)}{D(X)D(Y)}$

3. Discrete distributions

Binomial distribution

X is $\mathrm{Bin}(n,p)$ if $p_X(k) = \binom{n}{k}p^k(1-p)^{n-k}$, $k = 0, 1, \ldots, n$, where $0 < p < 1$.

$E(X) = np$, $V(X) = np(1-p)$

"For-the-first-time"-distribution

X is $\mathrm{fft}(p)$ if $p_X(k) = p(1-p)^{k-1}$, $k = 1, 2, 3, \ldots$, where $0 < p < 1$.

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

Hypergeometric distribution

X is $\mathrm{Hyp}(N,n,p)$ if $p_X(k) = \frac{\binom{Np}{k}\binom{N(1-p)}{n-k}}{\binom{N}{n}}$, $0 \le k \le Np$, $0 \le n-k \le N(1-p)$, where N, Np and n are positive integers and $N \ge 2$, $n < N$, $0 < p < 1$. $E(X) = np$, $V(X) = \frac{N-n}{N-1} \cdot np(1-p)$

Poisson distribution

X is $\mathrm{Po}(\mu)$ where $\mu > 0$ if $p_X(k) = \frac{\mu^k}{k!} \cdot e^{-\mu}$, $k = 0, 1, 2, \ldots$

$E(X) = \mu$, $V(X) = \mu$

4. Continuous distributions

Uniform distribution

X is $U(a,b)$ where $a < b$ if $f_X(x) = \begin{cases} \frac{1}{b-a} & \text{for } a < x < b \\ 0 & \text{otherwise} \end{cases}$

$E(X) = \frac{a+b}{2}$, $V(X) = \frac{(b-a)^2}{12}$

Exponential distribution

X is $\mathrm{Exp}(\lambda)$ where $\lambda > 0$ if $f_X(x) = \begin{cases} \lambda \cdot e^{-\lambda x} & \text{for } x > 0 \\ 0 & \text{otherwise} \end{cases}$

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

Normal distribution

X is $N(\mu, \sigma)$ if $f_X(x) = \dfrac{1}{\sqrt{2\pi} \cdot \sigma} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \; -\infty < x < \infty, \; \sigma > 0.$

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

X is $N(\mu, \sigma)$ if and only if $\dfrac{X - \mu}{\sigma}$ 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 $\varphi(x) = \dfrac{1}{\sqrt{2\pi}} \cdot e^{-x^2/2}, \; -\infty < x < \infty.$

A linear combination $\sum a_i X_i + b$ 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 μ 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

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

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

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

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

7. Chebychev's inequality

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

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

8. Statistical material

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

$$s^2 = \frac{1}{n-1} \sum_{j=1}^{n} (x_j - \overline{x})^2 = \frac{1}{n-1} \left[\sum_{j=1}^{n} x_j^2 - \frac{1}{n} \Big(\sum_{j=1}^{n} x_j \Big)^2 \right]$$

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 θ. The value θ^*_{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 θ.

9.2 Method of Least squares

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

$E(X_i) = \mu_i(\theta_1, \theta_2, \ldots, \theta_k)$ and $V(X_i) = \sigma^2$, 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

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

9.3 Mean error

An estimate of $D(\theta^*)$ is called *the mean error* of θ^* 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) $E(g(\theta^*_1, \ldots, \theta^*_n)) \approx g\big((\theta_1)^*_{\text{obs}}, \ldots, (\theta_n)^*_{\text{obs}}\big)$

$$V(g(\theta^*_1, \ldots, \theta^*_n)) \approx \sum_{i=1}^{n} \sum_{j=1}^{n} C(\theta^*_i, \theta^*_j) \cdot \left[\frac{\partial g}{\partial x_i} \cdot \frac{\partial g}{\partial x_j}\right]_{x_k = (\theta_k)^*_{\text{obs}}, k=1,\ldots,n}$$

10. Some common distributions in statistics

χ^2-distribution

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

$\sum_{k=1}^{f} X_k^2$ is $\chi^2(f)$-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 $\dfrac{X}{\sqrt{Y/f}}$ is $t(f)$-distributed.

11. Distributions for sample variables when the sample is normally distributed

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

Then we have:

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

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

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

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

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

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

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

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

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

d) $\dfrac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ is $t(n_1 + n_2 - 2)$

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

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

12. Confidence intervals

12.1 λ-method

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

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

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

12.2 t-method

Let θ^* be $N(\theta, D)$ where D and θ are unknown and D does not depend on θ. Let D^*_{obs} be a point estimate of D such that $\dfrac{\theta^* - \theta}{D^*}$ is $t(f)$. Then

$\theta^*_{\text{obs}} \pm D^*_{\text{obs}} \cdot t_{\alpha/2}(f)$

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

12.3 Approximative method

Let θ^* be approximatively $N(\theta, D)$.

Suppose that D^*_{obs} is a suitable point estimate of D. Then

$\theta^*_{\text{obs}} \pm D^*_{\text{obs}} \cdot \lambda_{\alpha/2}$ is a confidence interval

for θ with the *approximate* confidence level $1 - \alpha$.

12.4 Method based on χ^2-distribution

Let θ^*_{obs} be a point estimate of a parameter θ such that

$f \cdot \left(\dfrac{\theta^*}{\theta}\right)^2$ is $\chi^2(f)$. Then

$$\left(\theta^*_{\text{obs}}\sqrt{\frac{f}{\chi^2_{\alpha/2}(f)}}\,,\ \theta^*_{\text{obs}}\sqrt{\frac{f}{\chi^2_{1-\alpha/2}(f)}}\right)$$

is a confidence interval for θ 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) $\beta^* = \dfrac{\sum_{i=1}^n (x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^n (x_i - \overline{x})^2}$ is $N\left(\beta, \dfrac{\sigma}{\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}}\right)$

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

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

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

e) S^2 is independent of α^* and β^*

13.2 Confidence intervals

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

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

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

13.3 Computational aspects

$$S_{xy} = \sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^n (x_i - \overline{x}) y_i = \sum_{i=1}^n x_i (y_i - \overline{y}) = \sum_{i=1}^n x_i y_i - n\overline{x}\,\overline{y}$$

$$S_{xx} = \sum_{i=1}^n (x_i - \overline{x})^2 = \sum_{i=1}^n x_i^2 - n\overline{x}^2$$

$$S_{yy} = \sum_{i=1}^n (y_i - \overline{y})^2$$

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

14. Hypothesis testing

14.1 Definitions

The significance level (probability of error of first kind) α 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 θ is the correct parameter value.

14.2 Confidence method

Reject $H_0 : \theta = \theta_0$ on the level α if θ_0 does not fall within a suitably chosen confidence interval with the confidence level $1 - \alpha$.

14.3 χ^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

$Q' = \sum_{j=1}^{r} \frac{\left(x_j - np_j(\theta^*_{\text{obs}})\right)^2}{np_j(\theta^*_{\text{obs}})}$ is an outcome of an approximatively $\chi^2(r-k-1)$-distributed random variable.

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

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:

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

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.