

1 General definitions

1.1 Basic

- Sample variance

$$S_X^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \quad (1.1)$$

- Sample correlation coefficient

$$r_{X,Y} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)\sqrt{(S_X^2 S_Y^2)}} \quad (1.2)$$

- QQ-plot for cumulative distribution function F is the set of points $(q_F(\frac{i}{n+1}), x_{(i)})$, where $q_F(\cdot)$ is the quantile function for the distribution.
- Mean Squared Error (MSE)

$$\text{MSE}(\theta; T(X), g(\theta)) = \mathbb{E}_\theta (T(X) - g(\theta))^2 \quad (1.3)$$

- Bias-variance decomposition

$$\text{MSE}(\theta; T(X)) = \text{var}_\theta T + (\mathbb{E}_\theta T(X) - g(\theta))^2 \quad (1.4)$$

- Empirical distribution function

$$\hat{F}_n(x) = \sum_{i=1}^n \mathbb{I}(X_i \leq x) \quad (1.5)$$

1.2 k -th order statistic $X_{(k)}$

$X_{(k)}$ — k -th order statistic distribution for n i.i.d. variables from continuous distribution F .

$$f_{(k)}(x) = \frac{n!}{(k-1)!(n-k)!} F(x)^{k-1} (1-F(x))^{n-k} f(x) \quad (1.6)$$

$$F_{(k)}(x) = \sum_{j=k}^n \binom{n}{j} F(x)^j (1-F(x))^{n-j} \quad (1.7)$$

$$\mathbb{E}F(X_{(k)}) = \frac{k}{n+1} \quad (1.8)$$

2 Important distributions

- Student's t -distribution t_ν , $\nu \in \mathbb{R}_{>0}$

– pdf

$$\frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi\nu} \Gamma(\frac{\nu}{2})} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} \quad (2.1)$$

– cdf

$$\frac{1}{2} + x\Gamma\left(\frac{\nu+1}{2}\right) \frac{{}_2F_1\left(\frac{1}{2}, \frac{\nu+1}{2}; \frac{3}{2}, -\frac{x^2}{\nu}\right)}{\sqrt{\pi\nu} \Gamma(\frac{\nu}{2})} \quad (2.2)$$

– t -distribution with $n \in \mathbb{N}$ degrees of freedom arises from the ratio of independent $N(0, 1)$ - and χ_n^2 -distributions

- Poisson distribution $\text{Poisson}(\lambda)$, $\lambda > 0$

– λ is the average number of events per interval

– pdf

$$p_\lambda(k) = e^{-\lambda} \frac{\lambda^k}{k!} \quad (2.3)$$

- Geometric distribution $G(\theta)$, $0 \leq \theta \leq 1$

– pdf

$$f_\theta(k) = (1 - \theta)^{1-k} \theta \quad (2.4)$$

– cdf

$$F_\theta(k) = 1 - (1 - \theta)^k \quad (2.5)$$

- Exponential distribution $F(x; \lambda)$

– pdf

$$f_\lambda(x) = \lambda e^{-\lambda x} \quad (2.6)$$

– cdf

$$F_\lambda(x) = 1 - e^{-\lambda x} \quad (2.7)$$

– $\mathbb{E}_\lambda X = 1/\lambda$

- Beta distribution $B(\alpha, \beta)$, $\alpha, \beta > 0$

– pdf

$$f_{\alpha,\beta}(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}, \quad B(\alpha,\beta) \equiv \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \quad (2.8)$$

- Weibull distribution

– α and λ are the “shape” and “inverse scale” parameters.

– pdf

$$f_{\lambda,\alpha}(x) = \lambda^\alpha \alpha x^{\alpha-1} e^{-(\lambda x)^\alpha} \quad (2.9)$$

– cdf

$$F_{\lambda,\alpha}(x) = 1 - e^{-(\lambda x)^\alpha} \quad (2.10)$$

- Gamma distribution $\Gamma(\alpha, \lambda)$, $\alpha > 0, \lambda > 0$

– α and λ are known as “shape” and “inverse scale” parameters.

– pdf

$$f_{\alpha,\lambda}(x) = \frac{x^{\alpha-1} \lambda^\alpha e^{-\lambda x}}{\Gamma(\alpha)} \quad (2.11)$$

– cdf (where $\gamma(s, x) = \int_0^x t^{s-1} e^{-t} dt$ — is the “incomplete gamma function”)

$$F_{\alpha,\lambda}(x) = \frac{\gamma(\alpha, x\lambda)}{\Gamma(\alpha)} \quad (2.12)$$

3 Fundamental results

Theorem 3.1. *Let X_1, \dots, X_n be an i.i.d. random variables from the $N(\mu, \sigma^2)$ distribution, then*

1. \bar{X} is $N(\mu, \sigma^2/n)$ distributed;
2. $(n-1)S_X^2/\sigma^2$ is χ_{n-1}^2 -distributed (see 1.1);
3. \bar{X} and S_X^2 are independent;
4. $\sqrt{n}(\bar{X} - \mu)/\sqrt{S_X^2}$ has the t_{n-1} -distribution.

Proof. $\|X\|^2 - n\bar{X}^2 = (n-1)S_X^2$

4 Estimators

4.1 Maximum of n uniformly distributed statistics

Set up: X_1, X_2, \dots, X_n i.i.d. drawn from $U[0, \theta]$, where θ is the parameter of interest.

- $\hat{\theta} = 2\bar{X}_n$
 - method of moments estimator
 - *unbiased*
 - $\text{MSE}(\theta, \hat{\theta}) = \frac{\theta^2}{3n}$, see (1.6)
- $X_{(n)}$ — n -th order statistic, i.e. maximum.
 - $\mathbb{E}_{\theta} X_{(n)} = \frac{n}{n+1}\theta$, see (1.6)
 - $\text{MSE}(\theta, X_{(n)}) = \frac{2\theta^2}{(n+2)(n+1)}$
- $\frac{n+2}{n+1}X_{(n)}$
 - best estimator of the form $cX_{(n)}$
 - $\text{MSE}(\theta, \frac{n+2}{n+1}X_{(n)}) = \frac{\theta^2}{(n+1)^2}$

4.2 Univariate normal distribution

- $(\hat{\mu}, \hat{\sigma}^2) = \left(\bar{X}_n, \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \right) = \left(\bar{X}_n, \frac{n-1}{n} S_X^2 \right)$
 - maximum likelihood estimator
 - method of moments estimator
 - $\hat{\mu}$ is *unbiased*
 - $\mathbb{E}_{(\mu, \sigma^2)} \hat{\sigma}^2 = \frac{n-1}{n} \sigma^2$

4.3 Empirical distribution function

Let X_1, \dots, X_n be an i.i.d. sample drawn from the distribution F .

- The empirical distribution function (ecdf) $\hat{F}(x) = \sum_{i=1}^n \mathbb{I}(X_i \leq x)$ (see 1.1)
 - *unbiased*
 - $\text{cov}_F(\hat{F}(u), \hat{F}(v)) = n^{-1}(F(\min(u, v)) - F(u)F(v))$ – positively correlated

5 Statistical tests

5.1 t -tests

5.1.1 One-sample t -test

Let X_1, X_2, \dots, X_n be an i.i.d. sample from the $N(\mu, \sigma^2)$ -distribution with μ and σ^2 unknown. Given $\mu_0 \in \mathbb{R}$ we test:

$$H_0 : \mu \leq \mu_0 \text{ against } H_1 : \mu > \mu_0 \quad (5.1)$$

Test statistic:

$$T = \sqrt{n} \frac{\bar{X}_n - \mu_0}{S_X} \quad (5.2)$$

By Theorem 3.1, under $\mu = \mu_0$ the statistic has Student's t_{n-1} distribution, consequently we can use.

$$\sup_{\mu \leq \mu_0} \mathbb{P}(T \geq t_{n-1, 1-\alpha}) \leq \alpha \quad (5.3)$$

5.1.2 t -Test for paired observations

Let $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ be the i.i.d. sample of paired observations. We assume that $Z_i \stackrel{\text{def}}{=} X_i - Y_i$ is $N(\Delta, \sigma^2)$ is normally distributed, and the ordinary One-sample t -test can be used to test the null hypotheses $H_0 : \Delta \geq 0$. Note that if X_i and Y_i are strongly correlated then variance of Z_i decreases and this improves the power of the t -test.

5.1.3 Two-sample t -test

Let X_1, X_2, \dots, X_n and Y_1, Y_2, \dots, Y_m be two mutually independent i.i.d samples from $N(\mu, \sigma^2)$ and $N(\nu, \sigma^2)$. The test checks

$$H_0 : \mu - \nu \leq 0 \text{ against } H_1 : \mu - \nu > 0 \quad (5.4)$$

Test statistic:

$$T = \frac{\bar{X} - \bar{Y}}{S_{X,Y} \sqrt{\frac{1}{n} + \frac{1}{m}}} \quad (5.5)$$

$$S_{X,Y}^2 = \frac{1}{m+n-2} \left(\sum_{i=1}^n (X_i - \bar{X}_n)^2 + \sum_{j=1}^m (Y_j - \bar{Y}_m)^2 \right) \quad (5.6)$$

Theorem 3.1 implies that $S_{X,Y}^2$ follows $\sigma^2 \cdot \chi_{m+n-2}^2$ distribution.

5.2 Kolmogorov-Smirnov test

Given and i.i.d. sample X_1, \dots, X_n from some unknown distribution F , we want to test:

$$H_0 : F = F_0 \text{ against } H_1 : F \neq F_0 \quad (5.7)$$

The test statistic is given by

$$T = \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)|, \quad (5.8)$$

where $\hat{F}_n(x)$ stands for the empirical distribution function (see 1.1). The distribution of T is the same for every continuous cdf F_0 . The following limit establishes the test:

$$\lim_{n \rightarrow \infty} \mathbb{P}_{F_0} \left(T > \frac{z}{n} \right) = 2 \sum_{j=1}^{\infty} (-1)^{j+1} e^{-j^2 z^2} \quad (5.9)$$