1 General definitions

1.1 Basic

- Sample variance

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

- Sample correlation coefficient

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

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

- Mean Squared Error (MSE)

$$\mathrm{MSE}(\theta; T(X), g(\theta)) = \mathbb{E}_\theta\left(T(X) - g(\theta)\right)^2 \tag{1.3}$$

- Bias-variance decomposition

$$\mathrm{MSE}(\theta; T(X)) = \mathrm{var}_\theta T + \left(\mathbb{E}_\theta T(X) - g(\theta)\right)^2 \tag{1.4}$$

- Empirical distribution function

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

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.

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

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

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

1.3 Time Series

Time series below are assumed to be *weakly stationary* in the following sense:

- Stochastic time series $X_t(\omega)$ are called *weakly stationary*, if $\mathbb{E}X_t$ and $\mathbf{cov}(X_t, X_s)$ are inependent of time shifts; in particular, it first and second moments exist.

- For a weakly stationary time series X_t, the following functios are defined:

  - *Autocovariance funciton*:
$$\gamma(h) = \big(X_{t+h}, X_t\big) \tag{1.9}$$

  - *Autocorrelation function*
$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} \tag{1.10}$$

  - *Partial autocorrelation function* φh is defined as the coefficient of linear regression of X_{t+h} on X_t when controlled for constant and $X_{t+1}, \ldots X_{t+h-1}$ (see Proposition 3.18). In particular, $\varphi(0) = 1$ and $\varphi(1) = \rho(1)$

2 Important distributions

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

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

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

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

- Poisson distribution Poisson(λ), $\lambda > 0$
  - λ is the average number of events per interval
  - pdf
$$p_\lambda(k) = e^{-\lambda} \frac{\lambda^k}{k!} \tag{2.3}$$
- Geometric distribution $G(\theta)$, $0 \leq \theta \leq 1$
  - pdf
$$f_\theta(k) = (1-\theta)^{1-k}\theta \tag{2.4}$$
  - cdf
$$F_\theta(k) = 1 - (1-\theta)^k \tag{2.5}$$
- Exponential distribution $F(x;\lambda)$
  - pdf
$$f_\lambda(x) = \lambda e^{-\lambda x} \tag{2.6}$$
  - cdf
$$F_\lambda(x) = 1 - e^{-\lambda x} \tag{2.7}$$
  - $\mathbb{E}_\lambda X = 1/\lambda$
- Beta distribution $B(\alpha, \beta)$, $\alpha, \beta > 0$
  - pdf
$$f_{\alpha,\beta}(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)},\ B(\alpha,\beta) \equiv \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \tag{2.8}$$
- Weibull distribution
  - α and λ are the "shape" and 'inverse scale" parameters.
  - pdf
$$f_{\lambda,\alpha}(x) = \lambda^\alpha \alpha x^{\alpha-1} e^{-(\lambda x)^\alpha} \tag{2.9}$$
  - cdf
$$F_{\lambda,\alpha}(x) = 1 - e^{-(\lambda x)^\alpha} \tag{2.10}$$
- Gamma distribution $\Gamma(\alpha, \lambda)$, $\alpha > 0, \lambda > 0$

- α and λ are known as "shape" and "inverse scale" parameters.
- pdf
$$f_{\alpha,\lambda}(x) = \frac{x^{\alpha-1}\lambda^{\alpha}e^{-\lambda x}}{\Gamma(\alpha)} \tag{2.11}$$
- cdf (where $\gamma(s,x) = \int_0^x t^{s-1}e^{-t}dt$ — is the "incomplete gamma function")
$$F_{\alpha,\lambda}(x) = \frac{\gamma(\alpha, x\beta)}{\Gamma(\alpha)} \tag{2.12}$$

Definition 2.1. A family of probability densities p_θ that depends on a parameter θ is called a k-dimensional exponential family if there exist functions $c(\theta)$, $h(x)$, $Q_j(\theta)$, and $V_j(x)$ such that

$$p_\theta(x) = c(\theta)h(x)e^{\sum_{j=1}^{k} Q_j(\theta)V_j(x)}$$

3 Fundamental results

Theorem 3.1. *Let $X_1, \ldots X_n$ be an i.i.d. ramdom 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^2_{n-1}-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. $\|X\|^2 - n\bar{X}^2 = (n-1)S_X^2$

Definition 3.2. Let X be a random variable defined on probability space $(\Omega, \mathbb{P}_\theta)$, $\theta \in \Theta$. Suppose that the likelihood function $\theta \mapsto \ell_\theta \stackrel{\text{def}}{=} \log p_\theta$ is differentiable for all $x \in \Omega$. The gradient

$$\dot{\ell}_\theta(x) = \frac{\partial}{\partial\theta}\ell_\theta(x)$$

is called the *score function*. The *Fisher information* is defined as the matrix

$$i_\theta = \mathbb{V}_\theta \dot{\ell}_\theta(X)$$

3.1 Reminder on different convergence types

Definition 3.3. Let X_n be a sequence of random variables defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$:

- X_n is said to converge to X *almost surely* if $\mathbb{P}(\lim_{n\to\infty}(X_n = X) = 1$
- convergence in probability
- weak convergence
- L_p convergence

Theorem 3.4. *If a sequence of random variables converges almost surely then it converges in probability.*

Proposition 3.5. *Assume $\sum_{n\in\mathbb{Z}} \mathbb{E}|X_n| < \infty$, then $\sum_{n\in\mathbb{Z}} X_n$ is defined as an almost sure limit and:*

$$\mathbb{E}\sum_{n\in\mathbb{Z}} X_n = \sum_{n\in\mathbb{Z}} \mathbb{E}X_n$$

Proof. Established using Monotone and Dominated Convergence theorems applied to partial sums of random variables.

Proposition 3.6. *Let $(a_n)_{n\in\mathbb{Z}}$ be an element of $L^1(\mathbb{Z})$, and let $(Z_t)_{n\in\mathbb{Z}}$ be a sequence of random variables satisfying $\mathbb{E}|Z_t| < C_1 \forall t$ for some constant C_1. Then the convolution*

$$X_t = \sum_{n\in\mathbb{Z}} a_n Z_{t-n}$$

is defined almost surely $\forall t \in \mathbb{Z}$.

Moreover, if there exists a constant such that $\mathbb{E}Z_t^2 < C_2 \forall t$, then X_t is also a limit in L^2-norm and $\mathbb{E}X_t^2 \leq |a_n|_1^2 C_2$.

Theorem 3.7. *Let $(a_n)_{n\in\mathbb{Z}}$ be a an element of $l^2(\mathbb{Z})$. Let $Z_t = \sum_{n\in\mathbb{Z}} b_n e_{t-n}$, where b_n is absolutely summable and e_t is a sequence of uncorrelated $(0, \sigma^2)$ random variables.*

Then there exists a sequence of random variables X_t, such that:

1.

$$\lim_{n\to\infty} \mathbb{E}\left\{\left(X_t - \sum_{n\in\mathbb{Z}} b_n Z_{t-n}\right)^2\right\} = 0,$$

2.

$$\lim_{n\to\infty}\mathbb{E}\left\{\left(X_t-\sum_{n\in\mathbb{Z}}c_n e_{t-n}\right)^2\right\}=0,\ \ c_n=\sum_{j\in\mathbb{Z}}a_j b_{n-j},$$

3.

$$\mathbb{E}\left(X_t X_{t+h}\right)=\sum_{j\in\mathbb{Z}}c_j c_{j+h}\sigma^2$$

Corollary 3.8. *All statements of Theorem 3.7 are true if* $(a_n)_{n\in\mathbb{Z}}$ *is absolutely summable, and* $(b_n)_{n\in\mathbb{Z}}$ *is square summable.*

3.2 Estimator convergence via information matrix

Theorem 3.9. *Suppose that* Θ *is compact and convex and that* θ *is identifiable, and let* $\hat{\theta}_n$ *be the maximum likelihood estimator based on a sample of size* n *from the distribution with (marginal) probability density* p_θ*. Suppose, furthermore, that the map* $\vartheta\mapsto\log p_\vartheta(x)$ *is continuously differentiable for all* x*, with derivative* $\dot{\ell}_\vartheta(x)$*, such that* $\|\dot{\ell}_\vartheta(x)\|\le L(x)$ *for every* $\vartheta\in\Theta$*, where* $L(x)$ *is a function with* $\mathbb{E}_\theta L^2(X)<\infty$*. If* θ *is an interior point of* Θ *and the function* $\theta\mapsto i_\theta$ *is continuous and positive, then under* θ*,* $\sqrt{n}(\hat{\theta}_n-\theta)$ *converges in distribution to a normal distribution with expectation* 0 *and variance* i_θ^{-1}*. Therefore, under* θ*, as* $n\to\infty$*, we have*

$$\sqrt{n}(\hat{\theta}_n-\theta)\to N(0,i_\theta^{-1})$$

Theorem 3.10. (Cramer-Rao) *Suppose* $\theta\mapsto p_\theta(x)$ *is differentiable for every* x*. Then under certain regularity conditions any unbiased estimator* T *for* $g(\theta)$ *satisfies:*

$$\mathbb{V}_\theta(T)\ge g'(\theta)I_\theta^{-1}g'(\theta)^T,$$

where I_θ *denotes the* full *information matrix.*

3.3 Sufficient Statistics and UMVU estimators

Definition 3.11. For a statistical model $(\Omega,\mathbb{P}_\theta)$, $\theta\in\Theta$, a statistic $V(x)$ is called *sufficient* (for r.v. X) if conditional distribution $f(x|V=v)$ is independent of V.

Theorem 3.12. *A statistic $V(x)$ is sufficient if there exist functions $h(x)$ and $g(v, \theta)$ such that*

$$p_\theta(x) = h(x)g(V(x), \theta)$$

Theorem 3.13. (Rao-Blackwell) *Let $V = V(X)$ be a sufficient statistic, and let $T = T(X)$ be an arbitrary real-valued estimator for $g(\theta)$. Then there exists an estimator $T^* = T^*(V)$ for $g(\theta)$ that depends only on V, such that $\mathbb{E}_\theta T^* = \mathbb{E}_\theta T$ and $\mathbb{V}_\theta T^* \leq \mathbb{V}_\theta T$ for all θ. In particular, we have $MSE(\theta; T^*) \leq MSE(\theta; T)$. This inequality is strict unless $\mathbb{P}_\theta(T^* = T) = 1$.*

Definition 3.14. For a statistical model $(\Omega, \mathbb{P}_\theta)$, $\theta \in \Theta$, a statistic $V(x)$ is called complete if $\mathbb{E}_\theta(f(V)) = 0$, $\forall \theta \in \Theta$ implies $f(V) = 0$ a.s.

Theorem 3.15. *Let $V(x)$ be sufficient and complete, and $T(V)$ be an unbiased estimator for $g(\theta)$. Then $T(V)$ is UMVU estimator (i.e. has smallest variance among all unbiased estimators $\forall \theta \in \Theta$).*

Theorem 3.16. *Suppose that for a k-dimensional exponential family (2.1) the set below contains an interior point:*

$$(Q_1(\theta), \ldots Q_k(\theta)), \ \theta \in \Theta$$

Then the random vector $(V_1(x), \ldots V_n(x))$ is sufficient and complete.

3.4 Time Series

Theorem 3.17. *The real valued function $\rho(h)$ is the correlation function of a real valued stationary time series $X_t(\omega)$ with index set $t \in \mathbb{Z}$ if and only if it is representable in the form*

$$\rho(h) = \int_{-\pi}^{\pi} e^{ihx} dG(x),$$

where $G(x)$ is a symmetric distribution function.

Proposition 3.18. *Let X_t be weakly stationary time series:*

1. *The partial autocorrelation coefficient $\varphi(h)$ equals θ_{hh} in the linear regression:*

$$X_{t+h} = \theta_{0h} + \theta_{1h} X_{t+h-1} + \ldots + \theta_{hh} X_t + a_{ht}$$

2. *Let $\rho_{t+h,t\cdot(t+1,\dots t+h-1)}$ denote the partial correlation of X_{t+h} and X_t when controlled for $X_{t+1}, \dots X_{t+h-1}$. The squared norm of the residual term in the regression above equals:*

$$\mathbb{E}(a_{ht}^2) = \gamma(0) \prod_{i=1}^{h} (1 - \rho^2_{t+h,t+i-1\cdot(t+i,\dots t+h-1)})$$

Proof. These statements follow from basic Eucledian geometry. Denote by P the projection onto the subspace spanned by $X_{t+i}, \dots X_{t+h-1}$. Now consider sequentially orthogonal decompositions $X_{t+h} = P(X_{t+h}) + (1-P)(X_{t+h})$ and look at the component of the second summand along $(1-P)(X_{t+i})$.

4 Estimators

4.1 Maximum of n uniformally distributed statistics

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

- $\hat{\theta} = 2\bar{X}_n$
  - method of moments estimator
  - *unbiased*
  - $\mathrm{MSE}(\theta, \hat{\theta}) = \frac{\theta^2}{3n}$, 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)
  - $\mathrm{MSE}(\theta, X_{(n)}) = \frac{2\theta^2}{(n+2)(n+1)}$
- $\frac{n+2}{n+1} X_{(n)}$
  - best estimator of the form $cX_{(n)}$
  - $\mathrm{MSE}(\theta, \frac{n+2}{n+1} X_{(n)}) = \frac{\theta^2}{(n+1)^2}$

4.2 Univariate normal distribution

- $(\hat{\mu}, \hat{\sigma}^2) = \left(\bar{X}_n, \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)\right) = \left(\bar{X}_n, \frac{n-1}{n}S_X^2\right)$
  - 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 $X_1, \ldots X_n$ 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*
  - $\mathbf{cov}_F\left(\hat{F}(u), \hat{F}(v)\right) = n^{-1}(F(\min(u,v)) - F(u)F(v))$ – positively correlated

4.4 Linear Regression

Theorem 4.1. (Ordinary Least Squares)
(i) In one-factor setting, maximum likelihood estimators for slope, intercept and variance are given by (see (1.1, 1.2)):

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

(ii) If the design matrix X in a multiple linear regression has full rank, then the maximum likelihood estimators are given by:

$$\hat{\beta} = (X^T X)^{-1}(X^T Y), \quad \hat{\sigma}^2 = \frac{\|Y - X\hat{\beta}\|^2}{n}$$

Theorem 4.2. (Weighted Least Squares/Heteroscedacity) *(i) Assume that error terms ε_i are have variance as $\sigma_i^2 \equiv z_i\sigma^2$ for known constants z_i. Let $w_i \stackrel{\text{def}}{=} (z_i\sigma^2)^{-1}$, then maximum likelihood estimators for slope, intercept and variance are given by (see (1.1, 1.2)):*

$$\tilde{\beta} = \frac{\sum w_i(x-\tilde{x})(y-\tilde{y})}{\sum w_i(x-\tilde{x})^2} = \frac{\sum w_i \sum w_i x_i y_i - \sum w_i x_i \sum w_i y_i}{\sum w_i \sum w_i x_i^2 - \left(\sum w_i x_i\right)^2}$$

$$\tilde{\alpha} = \tilde{y} - \tilde{\beta}\tilde{x}$$

$$\hat{\sigma}^2 = n^{-1} \sum \frac{1}{z_i}(y_i - \tilde{\beta}x_i - \tilde{\alpha})^2$$

(ii) For the multi-factor model, maximum likelihood estimators can be written in the form:

$$\tilde{\beta} = (X^T W X)^{-1}(X^T W Y)$$

Theorem 4.3. *Let $V \stackrel{\text{def}}{=} \operatorname{span}(X)$, and $V_0 \subset V$. Denote the projection onto V by P_V.*

1. *The likelihood ratio statistic for $H_0 : X\beta_0 \in V_0$ equals*

$$2\log\lambda_n(X,Y) = n\log\frac{\|(E-P_{V_0})Y\|^2}{\|(E-P_V)Y\|^2},$$

2. *Under the null hypothesis, the following quantity has $F_{n-p,p-p_0}$ distribution:*

$$\frac{\|(P_V - P_{V_0})Y\|^2/(p-p_0)}{\|(E-P_V)Y\|^2/(n-p)}$$

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:

$$H_0 : \mu \le \mu_0 \text{ against } H_1 : \mu > \mu_0 \tag{5.1}$$

Test statistic:

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

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

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

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

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

Test statistic:

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

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

Theorem 3.1 implies that $S^2_{X,Y}$ follows $\sigma^2 \cdot \chi^2_{m+n-2}$ 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:

$$H_0 : F = F_0 \text{ against } H_1 : F \neq F_0 \tag{5.7}$$

The test statistic is given by

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

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:

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