Cheatsheet for 18.6501x by Blechturm

1 Algebra

Absolute Value Inequalities:
$|f(x)| < a \Rightarrow -a < f(x) < a$
$|f(x)| > a \Rightarrow f(x) > a$ or $f(x) < -a$

2 Calculus

Differentiation under the integral sign
$\frac{\mathrm{d}}{\mathrm{d}x}\left(\int_{a(x)}^{b(x)} f(x,t)\mathrm{d}t\right) = f(x,b(x))b'(x) - f(x,a(x))a'(x) + \int_{a(x)}^{b(x)} f_x(x,t)\mathrm{d}t.$

Concavity in 1 dimension

If $g: I \to \mathbb{R}$ is twice differentiable in the interval I:
concave:
if and only if $g''(x) \leq 0$ for all $x \in I$

strictly concave:
if $g''(x) < 0$ for all $x \in I$

convex:
if and only if $g''(x) \geq 0$ for all $x \in I$

strictly convex if:
$g''(x) > 0$ for all $x \in I$

Multivariate Calculus

The Gradient ∇ of a twice differntiable function f is defined as:
$\nabla f : \mathbb{R}^d \to \mathbb{R}^d$

$$\theta = \begin{pmatrix}\theta_1\\ \theta_2\\ \vdots\\ \theta_d\end{pmatrix} \mapsto \begin{pmatrix}\frac{\partial f}{\partial \theta_1}\\ \frac{\partial f}{\partial \theta_2}\\ \vdots\\ \frac{\partial f}{\partial \theta_d}\end{pmatrix}\Bigg|_{\theta}$$

Hessian

The Hessian of f is a symmetric matrix of second partial derivatives of f

$$\mathbf{H}h(\theta) = \nabla^2 h(\theta) = \begin{pmatrix}\frac{\partial^2 h}{\partial\theta_1\partial\theta_1}(\theta) & \cdots & \frac{\partial^2 h}{\partial\theta_1\partial\theta_d}(\theta)\\ & \vdots & \\ \frac{\partial^2 h}{\partial\theta_d\partial\theta_1}(\theta) & \cdots & \frac{\partial^2 h}{\partial\theta_d\partial\theta_d}(\theta)\end{pmatrix} \in \mathbb{R}^{d\times d}$$

A symmetric (real-valued) $d \times d$ matrix $\mathbf{A}$ is:

Positive semi-definite:
$\mathbf{x}^T\mathbf{A}\mathbf{x} \geq 0 \quad$ for all $\mathbf{x} \in \mathbb{R}^d$.

Positive definite:
$\mathbf{x}^T\mathbf{A}\mathbf{x} > 0$ for all non-zero vectors $\mathbf{x} \in \mathbb{R}^d$

Negative semi-definite (resp. negative definite):
$\mathbf{x}^T\mathbf{A}\mathbf{x}$ is negative for all $\mathbf{x} \in \mathbb{R}^d - \{0\}$.

Positive (or negative) definiteness implies positive (or negative) semi-definiteness.

If the Hessian is positive definite then f attains a local minimum at a (convex).

If the Hessian is negative definite at a, then f attains a local maximum at a (concave).

If the Hessian has both positive and negative eigenvalues then a is a saddle point for f.

3 Important probability distributions

Bernoulli

Parameter $p \in [0,1]$, discrete
$p_x(k) = \begin{cases} p, & \text{if k = 1}\\ (1-p), & \text{if k = 0}\end{cases}$
$\mathbb{E}[X] = p$
$Var(X) = p(1-p)$

Binomial

Parameters p and n, discrete. Describes the number of successes in n independent Bernoulli trials.

$p_x(k) = \binom{n}{k}p^k(1-p)^{n-k}, k = 1,\ldots,n$

$\mathbb{E}[X] = np$

$Var(X) = np(1-p)$

Multinomial

Parameters $n > 0$ and $p_1,\ldots,p_r$.
$p_x(x) = \frac{n!}{x_1!,\ldots,x_n!}p_1,\ldots,p_r$

$\mathbb{E}[X_i] = n * p_i$
$Var(X_i) = np_i(1-p_i)$

Poisson

Parameter λ. discrete, approximates the binomial PMF when n is large, p is small, and $\lambda = np$.

$\mathbf{p_x}(k) = exp(-\lambda)\frac{\lambda^k}{k!}$ for $k = 0,1,\ldots,$

$\mathbb{E}[X] = \lambda$
$Var(X) = \lambda$

Exponential

Parameter λ, continuous
$f_x(x) = \begin{cases}\lambda exp(-\lambda x), & \text{if x >= 0}\\ 0, & \text{o.w.}\end{cases}$

$F_x(x) = \begin{cases}1 - exp(-\lambda x), & \text{if x >= 0}\\ 0, & \text{o.w.}\end{cases}$

$\mathbb{E}[X] = \frac{1}{\lambda}$
$Var(X) = \frac{1}{\lambda^2}$

Univariate Gaussians

Parameters μ and $\sigma^2 > 0$, continuous
$f(x) = \frac{1}{\sqrt{(2\pi\sigma)}}exp(-\frac{(x-\mu)^2}{2\sigma^2})$
$\mathbb{E}[X] = \mu$
$Var(X) = \sigma^2$

Invariant under affine transformation:

$aX + b \sim N(X + b, a^2\sigma^2)$

Symmetry:

If $X \sim N(0,\sigma^2)$, then $-X \sim N(0,\sigma^2)$

$\mathbb{P}(|X| > x) = 2\mathbb{P}(X > x)$

Standardization:

$Z = \frac{X-\mu}{\sigma} \sim N(0,1)$

$\mathbf{P}(X \leq t) = \mathbf{P}\left(Z \leq \frac{t-\mu}{\sigma}\right)$

Higher moments:

$\mathbb{E}[X^2] = \mu^2 + \sigma^2$
$\mathbb{E}[X^3] = \mu^3 + 3\mu\sigma^2$
$\mathbb{E}[X^4] = \mu^4 + 6\mu^2\sigma^2 + 3\sigma^4$

Uniform

Parameters a and b, continuous.
$\mathbf{f_x}(x) = \begin{cases}\frac{1}{b-a}, & \text{if a < x <b}\\ 0, & \text{o.w.}\end{cases}$

$\mathbb{E}[X] = \frac{a+b}{2}$
$Var(X) = \frac{(b-a)^2}{12}$

Maximum of n iid uniform r.v.

Minimum of n iid uniform r.v.

Cauchy

continuous, parameter m,
$f_m(x) = \frac{1}{\pi}\frac{1}{1+(x-m)^2}$

$\mathbb{E}[X] = notdefined!$
$Var(X) = notdefined!$

$\mathrm{med}(X) = P(X > M) = P(X < M)$
$= 1/2 = \int_{1/2}^{\infty}\frac{1}{\pi}\cdot\frac{1}{1+(x-m)^2}dx$

4 Quantiles of a Distribution

Let α in $(0,1)$. The quantile of order $1-\alpha$ of a random variable X is the number q_α such that:

$q_\alpha = \mathbb{P}(X \leq q_\alpha) = 1 - \alpha$

$\mathbb{P}(X \geq q_\alpha) = \alpha$

$F_X(q_\alpha) = 1 - \alpha$

$F_X^{-1}(1-\alpha) = \alpha$

If $X \sim N(0,1)$:

$\mathbb{P}(|X| > q_\alpha) = \alpha$

5 Expectation

$\mathbb{E}[X] = \int_{-inf}^{+inf} x \cdot f_X(x) \quad dx$

$\mathbb{E}[g(X)] = \int_{-inf}^{+inf} g(x) \cdot f_X(x)dx$

$\mathbb{E}[X|Y = y] = \int_{-inf}^{+inf} x \cdot f_{X|Y}(x|y) \quad dx$

Integration limits only have to be over the support of the pdf. Discrete r.v. same as continuous but with sums and pmfs.

Total expectation theorem:

$\mathbb{E}[X] = \int_{-inf}^{+inf} f_Y(y) \cdot \mathbb{E}[X|Y = y]dy$

Expectation of constant a:

$\mathbb{E}[a] = a$

Product of **independent** r.vs X and Y :

$\mathbb{E}[X \cdot Y] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$

Product of **dependent** r.vs X and Y :

$\mathbb{E}[X \cdot Y] \neq \mathbb{E}[X] \cdot \mathbb{E}[Y]$

$\mathbb{E}[X \cdot Y] = \mathbb{E}[\mathbb{E}[Y \cdot X|Y]] = \mathbb{E}[Y \cdot \mathbb{E}[X|Y]]$

Linearity of Expectation where a and c are given scalars:

$\mathbb{E}[aX + cY] = a\mathbb{E}[X] + c\mathbb{E}[Y]$

If Variance of X is known:

$\mathbb{E}[X^2] = var(X) - \mathbb{E}[X]$

6 Variance

Variance is the squared distance from the mean.

$Var(X) = \mathbb{E}[(X - \mathbb{E}(X))^2]$

$Var(X) = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2$

Variance of a product with constant a:

$Var(aX) = a^2Var(X)$

Variance of sum of two **dependent** r.v.:

$Var(X + Y) = Var(X) + Var(Y) + 2Cov(X,Y)$

Variance of sum of two **independent** r.v.:

$Var(X + Y) = Var(X) + Var(Y)$

7 Covariance

The Covariance is a measure of how much the values of each of two correlated random variables determine each other

$Cov(X,Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)]$

$Cov(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$

$Cov(X,Y) = \mathbb{E}[(X)(Y - \mu_Y)]$

Possible notations:

$Cov(X,Y) = \sigma(X,Y) = \sigma_{(X,Y)}$

Covariance is commutative:

$Cov(X,Y) = Cov(Y,X)$

Covariance with of r.v. with itself is variance:

$Cov(X,X) = \mathbb{E}[(X - \mu_X)^2] = Var(X)$

Useful properties:

$Cov(aX + h, bY + c) = abCov(X,Y)$

$Cov(X, X + Y) = Var(X) + cov(X,Y)$

$Cov(aX + bY, Z) = aCov(X,Z) + bCov(Y,Z)$

If $Cov(X,Y) = 0$, we say that X and Y are uncorrelated. If X and Y are independent, their Covariance is zero. The converse is not always true. It is only true if X and Y form a gaussian vector, ie. any linear combination $\alpha X + \beta Y$ is gaussian for all $(\alpha,\beta) \in \mathbb{R}^2$ without $\{0,0\}$.

8 Law of large Numbers and Central Limit theorem univariate

Let $X_1,\ldots,X_n \overset{iid}{\sim} P_\mu$, where $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$ for all $i = 1,2,\ldots,n$ and $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Law of large numbers:

$\overline{X_n} \xrightarrow[n\to\infty]{P,a.s.} \mu$.

$\frac{1}{n}\sum_{i=1}^{n} g(X_i) \xrightarrow[n\to\infty]{P,a.s.} \mathbb{E}[g(X)]$

Central Limit Theorem:

$\sqrt{(n)}\frac{\overline{X_n}-\mu}{\sqrt{(\sigma^2)}} \xrightarrow[n\to\infty]{(d)} N(0,1)$

$\sqrt{(n)}(\overline{X_n} - \mu) \xrightarrow[n\to\infty]{(d)} N(0,\sigma^2)$

Variance of the Mean:

$Var(\overline{X_n}) = (\frac{\sigma^2}{n})^2 Var(X_1 + X_2, ..., X_n) = \frac{\sigma^2}{n}$.

Expectation of the mean:

$E[\overline{X_n}] = \frac{1}{n} E[X_1 + X_2, ..., X_n] = \mu$.

9 Statistical models

$E, \{P_\theta\}_{\theta \in \Theta}$

E is a sample space for X i.e. a set that contains all possible outcomes of X

$\{\mathbb{P}_\theta\}_{\theta \in \Theta}$ is a family of probability distributions on E.

Θ is a parameter set, i.e. a set consisting of some possible values of Θ.

θ is the true parameter and unknown. In a parametric model we assume that $\Theta \subset \mathbb{R}^d$, for some $d \geq 1$.

Identifiability:

$\theta \neq \theta' \Rightarrow \mathbb{P}_\theta \neq \mathbb{P}_{\theta'}$

$\mathbb{P}_\theta = \mathbb{P}_{\theta'} \Rightarrow \theta = \theta'$

A Model is well specified if:

$\exists \theta \ s.t.\ \mathbb{P} = \mathbb{P}_\theta$

10 Estimators

A statistic is any measurable functionof the sample, e.g. $\overline{X_n}, max(X_i)$, etc. An Estimator of θ is any statistic which does not depend on θ.

An estimator $\hat{\theta}_n$ is weakly consistent if: $\lim_{n\to\infty} \hat{\theta}_n = \theta$ or $\hat{\theta}_n \xrightarrow[n\to\infty]{P} \mathbb{E}[g(X)]$. If the convergence is almost surely it is strongly consistent.

Asymptotic normality of an estimator:

$\sqrt(n)(\hat{\theta}_n - \theta) \xrightarrow[n\to\infty]{(d)} N(0, \sigma^2)$

σ^2 is called the **Asymptotic Variance** of $\hat{\theta}_n$. In the case of the sample mean it the variance of a single X_i. If the estimator is a function of the sample mean the **Delta Method** is needed to compute the Asymptotic Variance. Asymptotic Variance $\neq$ Variance of an estimator.

Bias of an estimator:

$Bias(\hat{\theta}_n = \mathbb{E}[\hat{\theta_n}] - \theta$

Quadratic risk of an estimator:

$R(\hat{\theta}_n) = \mathbb{E}[(\hat{\theta}_n - \theta)^2] = Bias^2 + Variance$

11 Confidence intervals

Let $(E, (\mathbb{P}_\theta)_{\theta\in\Theta})$ be a statistical model based on observations $X_1, \ldots X_n$ and assume $\Theta \subseteq \mathbb{R}$. Let $\alpha \in (0,1)$.

Non asymptotic confidence interval of level $1-\alpha$ for θ:

Any random interval $\mathcal{I}$, depending on the sample $X_1, \ldots X_n$ but not at θ and such that:

$\mathbb{P}_\theta[\mathcal{I} \ni \theta] \geq 1 - \alpha, \ \ \forall \theta \in \Theta$

Confidence interval of **asymptotic level** $1-\alpha$ for θ:

Any random interval $\mathcal{I}$ whose boundaries do not depend on θ and such that:

$\lim_{n\to\infty} \mathbb{P}_\theta[\mathcal{I} \ni \theta] \geq 1 - \alpha, \ \ \forall \theta \in \Theta$

Two-sided asymptotic CI

Let $X_1, \ldots, X_n = \tilde{X}$ and $\tilde{X} \overset{iid}{\sim} P_\theta$. A two-sided CI is a function depending on $\tilde{X}$ giving an upper and lower bound in which the estimated parameter lies $\mathcal{I} = [l(\tilde{X}, u(\tilde{X})]$ with a certain probability $\mathbb{P}(\theta \in \mathcal{I}) \geq 1 - q_\alpha$ and conversely $\mathbb{P}(\theta \notin \mathcal{I}) \leq \alpha$

Since the estimator is a r.v. depending on $\tilde{X}$ it has a variance $Var(\hat{\theta}_n$ and a mean $\mathbb{E}[\hat{\theta}_n]$. After finding those it is possible to standardize the estimator using the CLT. This yields an asymptotic CI:

$\mathcal{I} = \hat{\theta}_n + [\frac{-q_{\alpha/2}\sqrt{Var(X_i)}}{\sqrt{n}}, \frac{q_{\alpha/2}\sqrt{Var(X_i)}}{\sqrt{n}}]$

This expression depends on the real variance $Var(X_i)$ of the r.vs, the variance has to be estimated. Three possible methods: plugin (use sample mean), solve (solve quadratic inequality), conservative (use the maximum of the variance).

Delta Method

If I take a function of the mean and want to make it converge to a function of the mean.

$\sqrt{n}(g(\widehat{m}_1) - g(m_1(\theta))) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, g'(m_1(\theta))^2\sigma^2)$

12 Hypothesis tests

Onesided

Twosided

P-Value

13 Random Vectors

A random vector $\mathbf{X} = \left(X^{(1)}, \ldots, X^{(d)}\right)^T$ of dimension $d \times 1$ is a vector-valued function from a probability space ω to $\mathbb{R}^d$:

$\mathbf{X} : \Omega \longrightarrow \mathbb{R}^d$

$$\omega \longrightarrow \begin{pmatrix} X^{(1)}(\omega) \\ X^{(2)}(\omega) \\ \vdots \\ X^{(d)}(\omega) \end{pmatrix}$$

where each $X^{(k)}$, is a (scalar) random variable on Ω.

PDF of $\mathbf{X}$: joint distribution of its components $X^{(1)}, \ldots, X^{(d)}$.

CDF of $\mathbf{X}$:

$\mathbb{R}^d \rightarrow [0,1]$

$\mathbf{x} \mapsto \mathbf{P}(X^{(1)} \leq x^{(1)}, \ldots, X^{(d)} \leq x^{(d)})$.

The sequence $\mathbf{X}_1, \mathbf{X}_2, \ldots$ converges in probability to $\mathbf{X}$ if and only if each component of the sequence $X_1^{(k)}, X_2^{(k)}, \ldots$ converges in probability to $X^{(k)}$.

Expectation of a random vector

The expectation of a random vector is the elementwise expectation. Let $\mathbf{X}$ be a random vector of dimension $d \times 1$.

$$\mathbb{E}[\mathbf{X}] = \begin{pmatrix} \mathbb{E}[X^{(1)}] \\ \vdots \\ \mathbb{E}[X^{(d)}] \end{pmatrix}.$$

The expectation of a random matrix is the expected value of each of its elements. Let $X = \{X_{ij}\}$ be an $n \times p$ random matrix. Then $\mathbb{E}[X]$, is the $n \times p$ matrix of numbers (if they exist):

$$\mathbb{E}[X] = \begin{bmatrix} \mathbb{E}[X_{11}] & \mathbb{E}[X_{12}] & \ldots & \mathbb{E}[X_{1p}] \\ \mathbb{E}[X_{21}] & \mathbb{E}[X_{22}] & \ldots & \mathbb{E}[X_{2p}] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}[X_{n1}] & \mathbb{E}[X_{n2}] & \ldots & \mathbb{E}[X_{np}] \end{bmatrix}$$

Let X and Y be random matrices of the same dimension, and let A and B be conformable matrices of constants.

$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$
$\mathbb{E}[AXB] = A\mathbb{E}[X]B$

Covariance Matrix

Let X be a random vector of dimension $d \times 1$ with expectation μ_X.

Matrix outer products!

$$\Sigma = \mathbb{E}[(X - \mu_X)(X - \mu_X)^T] = \mathbb{E}\left(\begin{bmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \\ \ldots \\ X_d - \mu_d \end{bmatrix} [X_1 - \mu_1, X_2 - \mu_2, \ldots, X_d - \mu_d]\right)$$

$$\Sigma = Cov(X) = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \ldots & \sigma_{1d} \\ \sigma_{21} & \sigma_{22} & \ldots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \ldots & \sigma_{dd} \end{bmatrix}$$

The covariance matrix Σ is a $d \times d$ matrix. It is a table of the pairwise covariances of the elemtents of the random vector. Its diagonal elements are the variances of the elements of the random vector, the off-diagonal elements are its covariances. Note that the covariance is commutative e.g. $\sigma_{12} = \sigma_{21}$

Alternative forms:

$\Sigma = \mathbb{E}[XX^T] - \mathbb{E}[X]\mathbb{E}[X]^T =$
$= \mathbb{E}[XX^T] - \mu_X\mu_X^T$

Let the random vector $X \in \mathbb{R}^d$ and A and B be conformable matrices of constants.

$Cov(AX + B) = Cov(AX) = ACov(X)A^T = A\Sigma A^T$

Every Covariance matrix is positive definite.

$\Sigma < 0$

Gaussian Random Vectors

A random vector $\mathbf{X} = (X^{(1)}, \ldots, X^{(d)})^T$ is a Gaussian vector, or multivariate Gaussian or normal variable, if any linear combination of its components is a (univariate) Gaussian variable or a constant (a "Gaussian"variable with zero variance), i.e., if $\alpha^T\mathbf{X}$ is (univariate) Gaussian or constant for any constant non-zero vector $\alpha \in \mathbb{R}^d$.

Multivariate Gaussians

The distribution of, X the d-dimensional Gaussian or normal distribution, is completely specified by the vector mean $\mu = \mathbb{E}[\mathbf{X}] = (\mathbb{E}[X^{(1)}], \ldots, \mathbb{E}[X^{(d)}])^T$ and the $d \times d$ covariance matrix Σ. If Σ is invertible, then the pdf of X is:

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)},$$

$\mathbf{x} \in \mathbb{R}^d$

Where $\det(\Sigma)$ is the determinant of Σ, which is positive when Σ is invertible.

If $\mu = 0$ and Σ is the identity matrix, then X is called a standard normal random vector .

If the covariant matrix Σ is diagonal, the pdf factors into pdfs of univariate Gaussians, and hence the components are independent.

The linear transform of a gaussian $X \sim N_d(\mu, \Sigma)$ with conformable matrices A and B is a gaussian:

$AX + B = N_d(A\mu + b, A\Sigma A^T)$

Multivariate CLT

Let $X_1, \ldots, X_d \in \mathbb{R}^d$ be independent copies of a random vector X such that $\mathbb{E}[x] = \mu$ ($d \times 1$ vector of expectations) and $Cov(X) = \Sigma$

$\sqrt(n)(\overline{X_n} - \mu) \xrightarrow[n\to\infty]{(d)} N(0, \Sigma)$

$\sqrt(n)\Sigma^{-1/2}\overline{X_n} - \mu \xrightarrow[n\to\infty]{(d)} N(0, I_d)$

Where $\Sigma^{-1/2}$ is the $d \times d$ matrix such that $\Sigma^{-1/2}\Sigma^{-1/2} = \Sigma^1$ and I_d is the identity matrix.

Multivariate Delta Method

Gradient Matrix of a Vector Function:

Given a vector-valued function $f : \mathbb{R}^d \rightarrow \mathbb{R}^k$, the gradient or the gradient matrix of f, denoted by ∇f , is the $d \times k$ matrix:

$$\nabla f = \left(\begin{matrix} | & | & | & | \\ \nabla f_1 & \nabla f_2 & \ldots & \nabla f_k \\ | & | & | & | \end{matrix}\right) = \left(\begin{matrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_k}{\partial x_1} \\ \vdots & \ldots & \vdots \\ \frac{\partial f_1}{\partial x_d} & \cdots & \frac{\partial f_k}{\partial x_d} \end{matrix}\right).$$

This is also the transpose of what is known as the Jacobian matrix $\mathbf{J}_f$ of f.

General statement, given

- $(\mathbf{T}_n)_{n\geq 1}$ a sequence of random vectors
- satisfying $\sqrt{n}\left(\mathbf{T}_n - \vec{\theta}\right) \xrightarrow[n\to\infty]{(d)} \mathbf{T}$,
- a function $\mathbf{g} : \mathbb{R}^d \rightarrow \mathbb{R}^k$ that is continuously differentiable at $\vec{\theta}$,

then

$\sqrt{n}\left(\mathbf{g}(\mathbf{T}_n) - \mathbf{g}(\vec{\theta})\right) \xrightarrow[n\to\infty]{(d)} \nabla\mathbf{g}(\vec{\theta})^T\,\mathbf{T}$

With multivariate Gaussians and Sample mean:

Let $\mathbf{T}_n = \overline{\mathbf{X}}_n$ where $\overline{\mathbf{X}}_n$ is the sample average of $\mathbf{X}_1, \ldots, \mathbf{X}_n \overset{iid}{\sim} \mathbf{X}$, and

$\vec{\theta} = \mathbb{E}[\mathbf{X}]$. The (multivariate) CLT then gives $\mathbf{T} \sim \mathcal{N}(\mathbf{0}, \Sigma_{\mathbf{X}})$ where $\Sigma_{\mathbf{X}}$ is the covariance of $\mathbf{X}$. In this case, we have:

$$\sqrt{n}\left(\mathbf{g}(\mathbf{T}_n) - \mathbf{g}(\vec{\theta})\right) \xrightarrow[n\to\infty]{(d)} \nabla \mathbf{g}(\vec{\theta})^T \mathbf{T}$$

$$\nabla \mathbf{g}(\vec{\theta})^T \mathbf{T} \sim \mathcal{N}\left(0, \nabla \mathbf{g}(\vec{\theta})^T \Sigma_{\mathbf{X}} \nabla \mathbf{g}(\vec{\theta})\right)$$

$(\mathbf{T} \sim \mathcal{N}(\mathbf{0}, \Sigma_{\mathbf{X}}))$

14 Distance between distributions

Total variation

The total variation distance TV between the propability measures P and Q with a sample space E is defined as:

$$\mathrm{TV}(\mathbf{P},\mathbf{Q}) = \max_{A \subset E} |\mathbf{P}(A) - \mathbf{Q}(A)|,$$

Calculation with f and g:

$$\mathrm{TV}(\mathbf{P},\mathbf{Q}) = \begin{cases} \frac{1}{2}\sum_{x\in E}|f(x)-g(x)|, & \text{discr} \\ \frac{1}{2}\int_{x\in E}|f(x)-g(x)|dx, & \text{cont} \end{cases}$$

Symmetry:
$d(\mathbf{P},\mathbf{Q}) = d(\mathbf{Q},\mathbf{P})$
nonnegative:
$d(\mathbf{P},\mathbf{Q}) \geq 0$
definite:
$d(\mathbf{P},\mathbf{Q}) = 0 \iff \mathbf{P} = \mathbf{Q}$
triangle inequality:
$d(\mathbf{P},\mathbf{V}) \leq d(\mathbf{P},\mathbf{Q}) + d(\mathbf{Q},\mathbf{V})$
If the support of $\mathbf{P}$ and $\mathbf{Q}$ is disjoint:
$d(\mathbf{P},\mathbf{V}) = 1$
TV between continuous and discrete r.v:
$d(\mathbf{P},\mathbf{V}) = 1$

KL divergence

the KL divergence (also known as relative entropy) KL between between the propability measures P and Q with the common sample space E and pmf/pdf functions f and g is defined as:

$$\mathrm{KL}(\mathbf{P},\mathbf{Q}) = \begin{cases} \sum_{x\in E} p(x)\ln\left(\frac{p(x)}{q(x)}\right), & \text{discr} \\ \int_{x\in E} p(x)\ln\left(\frac{p(x)}{q(x)}\right)dx, & \text{cont} \end{cases}$$

Not a distance!
Sum over support of P!
Asymetric in general:
$\mathrm{KL}(\mathbf{P},\mathbf{Q}) \neq \mathrm{KL}(\mathbf{Q},\mathbf{P})$
Nonnegative:
$\mathrm{KL}(\mathbf{P},\mathbf{Q}) \geq 0$
Definite:
if $\mathbf{P} = \mathbf{Q}$ then $\mathrm{KL}(\mathbf{P},\mathbf{Q}) = 0$
Does not satisfy triangle inequality in general:
$KL(\mathbf{P},\mathbf{V}) \nleq KL(\mathbf{P},\mathbf{Q}) + KL(\mathbf{Q},\mathbf{V})$

Estimator of KL divergence:

$$\mathrm{KL}(\mathbf{P}_{\theta^*},\mathbf{P}_\theta) = \mathbb{E}_{\theta^*}\left[\ln\left(\frac{p_{\theta^*}(X)}{p_\theta(X)}\right)\right],$$

$$\widehat{KL}(\mathbf{P}_{\theta_*},\mathbf{P}_\theta) = const - \frac{1}{n}\sum_{i=1}^n log(p_\theta(X_i))$$

15 Likelihood

Let $(E, \{P_\theta\}_{\theta\in\Theta})$ denote a discrete or continuous statistical model. Let p_θ denote the pmf or pdf of P_θ. Let $X_1,\dots,X_n \overset{iid}{\sim} P_{\theta^*}$ where the parameter θ^* is unknown. Then the likelihood is the function

$L_n : E^n \times \Theta$
$L_n(x_1,\dots,x_n,\theta) = \prod_{i=1}^n P_\theta[X_i = x_i]$

Loglikelihood:
$\ell_n(\theta) = \ln(L(x_1,\dots,x_n\theta)) =$
$= ln(\prod_{i=1}^n f_\theta(x_i)) =$
$= \sum_{i=1}^n ln(f_\theta(x_i))$

Bernoulli

Likelihood 1 trial:
$L_1(p) = p^x(1-p)^{1-x}$

Loglikelihood 1 trial:
$\ell_1(p) = xlog(p) + (1-x)log(1-p)$

Likelihood n trials:

$L_n(x_1,\dots,x_n,p) =$
$= p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}$

Loglikelihood n trials:

$\ell_n(p) =$
$= \sum_{i=1}^n x_i \ln(p) + \left(n - \sum_{i=1}^n x_i\right)\ln(1-p)$

Binomial

Likelihood:

$L_n(x_1,\dots,x_n,p,n) =$
$= nC_x\, p^x(1-p)^{n-x} = p^{x_i}(1-p)^{1-x_i}$

Loglikelihood:

$\ell_n(p,n) =$
$= \ln(nC_x) + x\ln(p) + (n-x)\ln(1-p)$

C is a constant from n choose k, disappears after differentiating.

Multinomial

Parameters $n > 0$ and $p_1,\dots,p_r$. Sample space= $E = 1,2,3,\dots,j$

Likelihood:

$p_x(x) = \prod_{j=1}^n p_j{}^{T_j}$, where $T^j = \mathbb{1}(X_i = j)$ is the count how often an outcome is seen in trials.

Loglikelihood:
$\ell_n = \sum_{j=2}^n T_j \ln\left(p_j\right)$

Poisson

Likelihood:
$L_n(x_1,\dots,x_n,\lambda) = \prod_{i=1}^n \frac{\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!} e^{n\lambda}$

Loglikelihood:
$\ell_n(\lambda) =$
$= -n\lambda + log(\lambda)(\sum_{i=1}^n x_i)) - log(\prod_{i=1}^n x_i!)$

Gaussian

Likelihood:

$L(x_1 \dots x_n; \mu, \sigma^2) =$
$= \frac{1}{\left(\sigma\sqrt{2\pi}\right)^n}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2\right)$

Loglikelihood:

$\ell_n(\mu, \sigma^2) =$
$= -nlog(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2$

Exponential

Likelihood:
$L(x_1 \dots x_n; \lambda) = \lambda^n \exp\left(-\lambda\sum_{i=1}^n x_i\right)$

Loglikelihood:

Uniform

Likelihood:
$L(x_1 \dots x_n; b) = \frac{1(\max_i(x_i \leq b))}{b^n}$

Loglikelihood:

Maximum likelihood estimation

Cookbook: take the log of the likelihood function. Take the partial derivative of the loglikelihood function with respect to the parameter. Set the partial derivative to zero and solve for the parameter.

If an indicator function on the pdf/pmf does not depend on the parameter, it can be ignored. If it depends on the parameter it can't be ignored because there is an discontinuity in the loglikelihood function. The maximum/minimum of the X_i is then the maximum likelihood estimator.

Maximum likelihood estimator:

Let $\left\{E, (\mathbf{P}_\theta)_{\theta\in\Theta}\right\}$ be a statistical model associated with a sample of i.i.d. random variables $X_1, X_2, \dots, X_n$. Assume that there exists $\theta^* \in \Theta$ such that $X_i \sim \mathbf{P}_{\theta^*}$.

The maximum likelihood estimator is the (unique) θ that minimizes $\widehat{\mathrm{KL}}(\mathbf{P}_{\theta^*}, \mathbf{P}_\theta)$ over the parameter space. (The minimizer of the KL divergence is unique due to it being strictly convex in the space of distributions once is fixed.)

$\widehat{\theta}_n^{MLE} =$
$\mathrm{argmin}_{\theta\in\Theta}\widehat{\mathrm{KL}}_n(\mathbf{P}_{\theta^*}, \mathbf{P}_\theta) =$
$\mathrm{argmax}_{\theta\in\Theta}\sum_{i=1}^n \ln p_\theta(X_i) =$
$\mathrm{argmax}_{\theta\in\Theta}\ln\left(\prod_{i=1}^n p_\theta(X_i)\right)$

Gaussian Maximum-loglikelihood estimators:

MLE estimator for $\sigma^2 = \tau$:
$\hat{\tau}_n^{MLE} = \frac{1}{n}\sum_{i=1}^n X_i^2$

MLE estimators:

$\hat{\mu}_n^{MLE} = \frac{1}{n}\sum_{i=1}(x_i)$

15.1 Fisher Information

The Fisher information, captures the negative of the expected curvature of the loglikelihood function.

Let $(\mathbb{R}, \{\mathbf{P}_\theta\}_{\theta\in\mathbb{R}})$ denote a continuous statistical model. Let $f_\theta(x)$ denote the pdf (probability density function) of the continuous distribution $\mathbf{P}_\theta$. Assume that $f_\theta(x)$ is twice-differentiable as a function of the parameter θ.

Formula for the calculation of Fisher Information of X:

$$\mathcal{I}(\theta) = \int_{-\infty}^{\infty}\frac{\left(\frac{\partial f_\theta(x)}{\partial\theta}\right)^2}{f_\theta(x)}dx$$

Models with one parameter (ie. Bernulli):

$\mathcal{I}(\theta) = \mathsf{Var}(\ell'(\theta))$

$\mathcal{I}(\theta) = -\mathbf{E}(\ell''(\theta))$

Models with multiple parameters (ie. Gaussians):

$\mathcal{I}(\theta) = -\mathbb{E}[\mathbf{H}\ell(\theta)]$

Cookbook:

Better to use 2nd derivative.

- Find loglikelihood
- Take second derivative (=Hessian if multivariate)
- Massage second derivative or Hessian (isolate functions of X_i to use with $-\mathbf{E}(\ell''(\theta))$ or $-\mathbb{E}[\mathbf{H}\ell(\theta)]$.
- Find the expectation of the functions of X_i and subsitute them back into the Hessian or the second derivative. Be extra careful to subsitute the right power back. $\mathbb{E}[X_i] \neq \mathbb{E}[X_i^2]$.
- Don't forget the minus sign!

Asymptotic normality of the maximum likelihood estimator

Under certain conditions (see slides) the MLE is asymptotically normal. This applies even if the MLE is not the sample average.

The asymptotic variance of the MLE is the inverse of the fisher information.

$$\sqrt(n)(\widehat{\theta}_n^{\mathrm{MLE}} - \theta^*) \xrightarrow[n\to\infty]{(d)} N_d(0, \mathcal{I}(\theta^*)^{-1})$$

16 Method of Moments

Let $X_1,\dots,X_n \overset{iid}{\sim} \mathbf{P}_{\theta^*}$ associated with model $(\mathbb{E}, \{\mathbf{P}_\theta\}_{\theta\in\Theta})$, with $\mathbb{E} \subseteq \mathbb{R}$ and $\Theta \subseteq \mathbb{R}$, for some d ≥ 1

Population moments:

$m_k(\theta) = \mathbb{E}_\theta[X_1^k], 1 \leq k \leq d$

Empirical moments:

$\widehat{m_k}(\theta) = \overline{X_n^k} = \frac{1}{n}\sum_{i=1}^n X_i^k$

Convergence of empirical moments:

$\widehat{m_k} \xrightarrow[n\to\infty]{P,a.s.} m_k$

$(\widehat{m_1},\dots,\widehat{m_d}) \xrightarrow[n\to\infty]{P,a.s.} (m_1,\dots,m_d)$

MOM Estimator M is a map from the parameters of a model to the moments of its distribution. This map is invertible, (ie. it results into a system of equations that can be solved for the true parameter vector θ^*). Find the moments (as many as parameters), set up system of equations, solve for parameters, use empirical moments to estimate.

$\psi : \Theta \to \mathbb{R}^d$

$\theta \mapsto (m_1(\theta), m_2(\theta), \dots, m_d(\theta))$

$M^{-1}(m_1(\theta^*), m_2(\theta^*), \dots, m_d(\theta^*))$

The MOM estimator uses the empirical moments:

$M^{-1}\left(\frac{1}{n}\sum_{i=1}^n X_i, \frac{1}{n}\sum_{i=1}^n X_i^2, \dots, \frac{1}{n}\sum_{i=1}^n X_i^d\right.$

Assuming M^{-1} is continuously differentiable at $M(0)$, the asymptotical variance of the MOM estimator is:

$$\sqrt(n)(\widehat{\theta_n^{MM}} - \theta) \xrightarrow[n\to\infty]{(d)} N(0, \Gamma)$$

where,

$\Gamma(\theta) = \left[\frac{\partial M^{-1}}{\partial\theta}(M(\theta))\right]^T \Sigma(\theta)\left[\frac{\partial M^{-1}}{\partial\theta}(M(\theta))\right]$

$\Gamma(\theta) = \nabla_\theta(M^{-1})^T\Sigma\nabla_\theta(M^{-1})$

Σ_θ is the covariance matrix of the random vector of the moments $(X_1^1, X_1^2 \dots, X_1^d)$.

17 M-estimation

Generalization of maximum likelihood estimation. No statistical model needs to be assumed to perform M-estimation.

Median

18 Hubert loss

$$h_\delta(x) = \begin{cases} \frac{x^2}{2} & \text{if } |x| < \delta \\ \delta(|x| - \delta/2) & \text{if } |x| > \delta \end{cases}.$$

the derivative of Huber's loss is the clip function :

$$\mathrm{clip}_\delta(x) := \frac{d}{dx} h_\delta(x) = \begin{cases} \delta & \text{if } x > \delta \\ x & \text{if } -\delta \le x \le \delta \end{cases}$$