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 Matrixalgebra

3 Calculus

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 $\mathbf{x} \in \mathbb{R}^d - \{\mathbf{0}\}$.

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.

4 Important probability distributions

Bernoulli

Parameter $p \in [0,1]$, discrete

$p_x(k) = \begin{cases} p, & \text{if k = 1} \\ (1-p), & \text{if k = 0} \end{cases}$

$\mathbb{E}[X] = p$
$Var(X) = p(1-p)$

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 $k = 0, 1, \dots,$

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

Exponential

Parameter λ, continuous

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

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

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

Normal (Gaussian)

Univariate:
Parameters μ and $\sigma^2 > 0$, continuous

$f(x) = \frac{1}{\sqrt{(2\pi\sigma)}} exp(-\frac{(x-\mu)^2}{2\sigma^2})$

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

Linearity:

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

Symmetry:

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

Standardization:

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

$\mathbf{P}(X \leq t) = \mathbf{P}\Big(Z \leq \frac{t-\mu}{\,}$

Quantile: $q_\alpha = \mathbf{P}(X \leq q_\alpha) = 1 - \alpha$

Normal tables:

Moments:

Multivariate gaussians:

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.

5 Multivariate Random Variables

A random vector $\mathbf{X} = \left(X^{(1)}, \dots, X^{(d)}\right)^T$ 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.
The probability distribution of a random vector $\mathbf{X}$ is the joint distribution of its components $X^{(1)}, \dots, X^{(d)}$.
The cumulative distribution function (cdf) of a random vector $mathbf X$ is defined as
$F: \mathbb{R}^d \to [0,1]$

$\mathbf{x} \mapsto \mathbf{P}(X^{(1)} \leq x^{(1)}, \dots, X^{(d)} \leq x^{(d)})$.
Convergence in Probability in Higher Dimension

In other words, the sequence $\mathbf{X}_1, \mathbf{X}_2, \dots$ converges in probability to $\mathbf{X}$ if and only if each component sequence $X_1^{(k)}, X_2^{(k)}, \dots$ converges in probability to $X^{(k)}$.

6 Expectation and Variance

Expectation

Expectation of a random vector is the elementwise expectation. Let $\mathbf{X}$ be a random vector of dimension $d \times 1$ with expectation $\mu_\mathbf{X}$.
Let $\mu \triangleq \mathbb{E}[\mathbf{X}]$ denote the entry-wise mean,

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

Variance

$Var(X+Y) = Var(X) + Var(Y) + 2\mathsf{Cov}(X,Y)$

Covariance

The Covariance is a measure of how much the values of each of two correlated random variables determines the other
$\mathsf{Cov}(X,Y) = \sigma(X,Y) = \sigma_{(X,Y)}$

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

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

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

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

$\mathsf{Cov}(aX+h, bY+c) = ab\mathsf{Cov}(X,Y)$

$\mathsf{Cov}(X, X+Y) = Var(X) + cov(X,Y)$

$\mathsf{Cov}(aX+bY, Z) = a\mathsf{Cov}(X,Z) + b\mathsf{Cov}(Y,Z)$

If $Cov(X,Y) = 0$, we say that X and Y are uncorrelated. If X and Y are independent, they are uncorrelated. 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 $\{0,0\}$.

7 Covariance Matrix

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

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

8 Law of large Numbers and Central Limit theorem

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

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

Expectation of the mean:

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

9 Statistical models

10 Estimators

11 Confidence intervals

Onesided

Twosided

Delta Method

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

12 Hypothesis tests

Onesided

Twosided

P-Value

13 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:

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

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

14 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 $X_1,\ldots,X_n \overset{iid}{\sim} P_{\theta^*}$ where the parameter θ^* is unknown. Then the likelihood is the function

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

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

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:

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

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:

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

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.

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

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

Gaussian
Likelihood:

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

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

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

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 $X_1, X_2, \ldots, X_n$. 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:

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

14.1 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 to use with $-\mathbf{E}(\ell''(\theta))$ or $-\mathbb{E}[\mathbf{H}\ell(\theta)]$

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.

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

15 Method of Moments

Let $X_1,\ldots,X_n \overset{iid}{\sim} \mathbf{P}_{\theta^*}$ 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:

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

Empirical moments:

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

Convergence of empirical moments:

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

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

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.

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

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

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

The MOM estimator uses the empirical moments:

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

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 $(X_1^1, X_1^2 \ldots, X_1^d)$.