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 Algebra

3 Calculus

3.1 Concavity in 1 dimension

If $g : I \to \mathbb{R}$ is twice differentiable in the interval I, i.e. $g''(x)$ exists for all $x \in I$, then g is
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 $x \in I$;

3.2 Multivariate Calculus

Gradient

Let

$$f : \mathbb{R}^d \longrightarrow \mathbb{R}\theta = \theta = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_d \end{pmatrix} \mapsto f(\theta)$$

denote a twice differentiable function, the Gradient ∇ of 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 the matrix $\mathbf{H} : \mathbb{R}^d \to \mathbb{R}^{d \times d}$ whose entry in the i-th row and j-th column is defined by

$$(\mathbf{H}f)_{ij} := \frac{\partial^2}{\partial\theta_i \partial\theta_j} f, \quad 1 \leq i, j \leq d$$

Semi-Definiteness

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

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

Positive definite if inequality above is strict $\mathbf{x}^T \mathbf{A} \mathbf{x} > 0$ for all non-zero vectors $\mathbf{x} \in \mathbb{R}^d$

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

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

Concavity

4 Important probability distributions

Bernoulli

Parameter $p \in [0,1]$. Discrete, describes the success or failure in a single trial.

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

$E[X] = p$
$Var(X) = p(1-p)$

Exponential

Parameter λ. Continuous

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

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

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

Normal (Gaussian)

Parameters μ and $\sigma^2 > 0$. Continuous

$f(x) = \frac{1}{\sqrt{(2\pi\sigma)}} exp(-\frac{(x-\mu)^2}{2\sigma^2})$
$E[X] = \mu$
$Var(X) = \sigma^2$
Useful properties:

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 $k = 0, 1, \ldots,$

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

Uniform

5 Expectation and Variance

Expectation

Variance

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

Covariance

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

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

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

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

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

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

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

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

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 $\{0,0\}$.

Variance and expectation of mean of n iid random variables

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

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:

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

6 LLN and CLT

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

Weak and strong law of large numbers:

$\overline{X_n} = \frac{1}{n}\sum_{i=1}^n X_i \xrightarrow[n\to\infty]{P, a.s.} \mu$.

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

Central Limit Theorem:

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

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

7 Statistical models

8 Estimators

9 Confidence intervals

Onesided

Twosided

Delta Method

10 Hypothesis tests

Onesided

Twosided

P-Value

11 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:

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

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

12 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

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

Bernoulli

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

Loglikelihood 1 trial:
$log(L_1(p)) = xlog(p) + (1-x)log(1-p)$

Likelihood n trials:

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

$= L_n = \prod_{i=1}^n (x_i p + (1-x_i)(1-p))$

Loglikelihood n trials:

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:

$log(L_n(x_1, \ldots, x_n, p, n)) = \ln(nC_x \; p^x(1-p)^{n-x}) = \ln(nC_x) + x\ln(p) + (n-x)\ln(1-p)$

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:
$log(L(x_1 \ldots x_n; \lambda)) =$
$-n\lambda + log(\lambda)(\sum_{i=1}^n x_i)) - log(\prod_{i=1}^n x_i!)$

Gaussian

Likelihood:

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

Loglikelihood:

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

Exponential

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

Uniform

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

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

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

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:

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

13 Multivariate Random Variables

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

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

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

14 Fisher Information

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.

Loglikelihood of X: $\ell(\theta) = \ln L_1(X, \theta) = \ln f_\theta(X)$

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 Expression to use $-\mathbf{E}(\ell''(\theta))$

15 Covariance Matrix

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

The covariance matrix Σ is defined as the following matrix outer product: $\Sigma = \mathbb{E}[(\mathbf{X} - \mu_\mathbf{X})(\mathbf{X} - \mu_\mathbf{X})^T]$.

$\Sigma == \mathbb{E}[\mathbf{X}\mathbf{X}^T] - \mathbb{E}[\mathbf{X}]\mathbb{E}[\mathbf{X}]^T =$
$= \mathbb{E}[\mathbf{X}\mathbf{X}^T] - \mu_\mathbf{X}\mu_\mathbf{X}^T$.