

## 1 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  $X$   $\{P_\theta\}_{\theta \in \Theta}$  is a family of probability distributions on  $E$ .

$\Theta$  is a parameter set, i.e. a set consisting of some possible values of  $\theta$ .

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

### 1.1 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:

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

## 2 Estimators

A **statistic** is any measurable function calculated with the data  $(\bar{X}_n, \max(X_i), \dots)$  (etc).

An **estimator**  $\hat{\theta}_n$  of  $\theta$  is any statistic which does not depend on  $\theta$ .

Estimators are random variables if they depend on the data (= realizations of random variables).

An estimator  $\hat{\theta}_n$  is **weakly consistent**

if:  $\lim_{n \rightarrow \infty} \hat{\theta}_n = \theta$  or  $\hat{\theta}_n \xrightarrow[n \rightarrow \infty]{P} \mathbb{E}[g(X)]$ .

If the convergence is almost surely it is **strongly consistent**.

**Asymptotic normality of an estimator:**

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

$\sigma^2$  is called the **Asymptotic Variance**

of the estimator  $\hat{\theta}_n$ . In the case of the sample mean it is the same variance as for the sample  $X_i$ .

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:**

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

**Quadratic risk of an estimator**

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

## 3 LLN and CLT

Let  $X_1, \dots, X_n \stackrel{iid}{\sim} P_\mu$ , where  $E(X_i) = \mu$  and  $\text{Var}(X_i) = \sigma^2$  for all  $i = 1, 2, \dots, n$  and  $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ .

**Law of large numbers:**

$$\bar{X}_n \xrightarrow[n \rightarrow \infty]{P.a.s.} \mu$$

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

**Central Limit Theorem for Mean:**

$$\sqrt{n} \frac{\bar{X}_n - \mu}{\sqrt{\sigma^2}} \xrightarrow[n \rightarrow \infty]{(d)} N(0, 1)$$

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow[n \rightarrow \infty]{(d)} N(0, \sigma^2)$$

**Central Limit Theorem for Sums:**

$$\sum_{i=1}^n X_i \xrightarrow[n \rightarrow \infty]{(d)} N(n\mu, \sqrt{n}(\sigma^2))$$

$$\frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}(\sigma^2)} \xrightarrow[n \rightarrow \infty]{(d)} N(0, 1)$$

**Variance of the Mean:**

$$\begin{aligned} \text{Var}(\bar{X}_n) &= \left(\frac{\sigma^2}{n}\right)^2 \text{Var}(X_1 + X_2, \dots, X_n) \\ &= \frac{\sigma^2}{n} \end{aligned}$$

**Expectation of the mean:**

$$\begin{aligned} E[\bar{X}_n] &= \frac{1}{n} E[X_1 + X_2, \dots, X_n] \\ &= \mu. \end{aligned}$$

## 4 Quantiles of a Distribution

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

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

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

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

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

If the distribution is **standard normal**  $X \sim N(0, 1)$ :

$$\begin{aligned} \mathbb{P}(|X| > q_\alpha) &= \alpha \\ &= 2\Phi(q_{\alpha/2}) \end{aligned}$$

Use **standardization** if a gaussian has unknown mean and variance  $X \sim N(\mu, \sigma^2)$  to get the quantiles by using  $Z$ -tables (standard normal tables).

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

$$= \Phi\left(\frac{t - \mu}{\sigma}\right)$$

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

$$q_\alpha = \frac{t - \mu}{\sigma}$$

## 5 Confidence intervals

Confidence Intervals follow the form:

(statistic)  $\pm$  (critical value)(estimated standard deviation of statistic)

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

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

Any random interval  $\mathcal{I}$ , depending on the sample  $X_1, \dots, X_n$  but not at  $\theta$  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  $\theta$ :

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

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

### 5.1 Two-sided asymptotic CI

Let  $X_1, \dots, X_n = \bar{X}$  and  $\bar{X} \stackrel{iid}{\sim} P_\theta$ . A two-sided CI is a function depending on  $\bar{X}$  giving an upper and lower bound in which the estimated parameter lies  $\mathcal{I} = [l(\bar{X}), u(\bar{X})]$  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  $\bar{X}$  it has a variance  $\text{Var}(\hat{\theta}_n)$  and a mean  $\mathbb{E}[\hat{\theta}_n]$ . Since the CLT is valid for every distribution standardizing the

distributions yields an asymptotic CI:

$$\begin{aligned} \mathcal{I} &= [\hat{\theta}_n - \frac{q_{\alpha/2} \sqrt{\text{Var}(\bar{X}_n)}}{\sqrt{n}}, \\ &\quad \hat{\theta}_n + \frac{q_{\alpha/2} \sqrt{\text{Var}(\bar{X}_n)}}{\sqrt{n}}] \end{aligned}$$

This expression depends on the real variance  $\text{Var}(X_i)$  of the r.v.s, the variance has to be estimated. Three possible methods: plugin (use sample mean or empirical variance), solve (solve quadratic inequality), conservative (use the theoretical maximum of the variance).

### 5.2 Sample Mean and Sample Variance

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

**Sample Mean:**

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$$

**Sample Variance:**

$$\begin{aligned} S_n &= \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \\ &= \frac{1}{n} \left( \sum_{i=1}^n X_i^2 - n\bar{X}_n^2 \right) \end{aligned}$$

**Unbiased estimator of sample variance:**

$$\begin{aligned} \hat{S}_n &= \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \\ &= \frac{n}{n-1} S_n \end{aligned}$$

### 5.3 Delta Method

To find the asymptotic CI if the estimator is a function of the mean. Goal is to find an expression that converges a function of the mean using the CLT.

Let  $Z_n$  be a sequence of r.v.  $\sqrt{n}(Z_n - \theta) \xrightarrow[n \rightarrow \infty]{(d)} N(0, \sigma^2)$  and let  $g: \mathbb{R} \rightarrow \mathbb{R}$  be continuously differentiable at  $\theta$ , then:

$$\begin{aligned} \sqrt{n}(g(Z_n) - g(\theta)) &\xrightarrow[n \rightarrow \infty]{(d)} \\ N(0, g'(\theta)^2 \sigma^2) \end{aligned}$$

**Example:** let  $X_1, \dots, X_n \exp(\lambda)$  where  $\lambda > 0$ . Let  $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$  denote the sample mean. By the CLT, we know

that  $\sqrt{n}(\bar{X}_n - \frac{1}{\lambda}) \xrightarrow[n \rightarrow \infty]{(d)} N(0, \sigma^2)$  for some value of  $\sigma^2$  that depends on  $\lambda$ . If we set  $g: \mathbb{R} \rightarrow \mathbb{R}$  and  $x \mapsto 1/x$ , then by the Delta method:

$$\begin{aligned} \sqrt{n}\left(g(\bar{X}_n) - g\left(\frac{1}{\lambda}\right)\right) &\xrightarrow[n \rightarrow \infty]{(d)} N(0, g'(\mathbb{E}[X])^2 \text{Var}(X)) \\ &\xrightarrow[n \rightarrow \infty]{(d)} N(0, g'\left(\frac{1}{\lambda}\right)^2 \frac{1}{\lambda^2}) \\ &\xrightarrow[n \rightarrow \infty]{(d)} N(0, \lambda^2) \end{aligned}$$

## 6 Asymptotic Hypothesis tests

Two hypotheses  $(\Theta_0)$  disjoint set from

$\Theta_1$ :  $\begin{cases} H_0: \theta \in \Theta_0 \\ H_1: \theta \in \Theta_1 \end{cases}$ . Goal is to reject

$H_0$  using a test statistic.

A test  $\psi$  has level  $\alpha$  if  $\alpha_\psi(\theta) \leq \alpha, \forall \theta \in \Theta_0$ . and **asymptotic level**  $\alpha$  if  $\lim_{n \rightarrow \infty} P_\theta(\psi = 1) \leq \alpha$ .

A **hypothesis-test** has the form

$$\psi = 1\{T_n \geq c\}$$

for some test statistic  $T_n$  and threshold  $c \in \mathbb{R}$ . Threshold  $c$  is usually  $q_{\alpha/2}$

**Rejection region:**

$$R_\psi = \{T_n > c\}$$

**Symmetric about zero and acceptance Region interval:**

$$\psi = 1\{|T_n| - c > 0\}.$$

**Power of the test:**

$$\pi_\psi = \inf_{\theta \in \Theta_1} (1 - \beta_\psi(\theta))$$

Where  $\beta_\psi$  is the probability of making a Type2 Error and  $\inf$  is the maximum.

**Two-sided test:**

$$\begin{aligned} H_1: \theta &\neq \Theta_0 \\ 1(|T_n| > q_{\alpha/2}) \end{aligned}$$

**One-sided tests:**

$$\begin{aligned} H_1: \theta &> \Theta_0 \\ 1(T_n < -q_\alpha) H_1 &: \theta < \Theta_0 \\ 1(T_n > q_\alpha) \end{aligned}$$

**Type1 Error:**

Test rejects null hypothesis  $\psi = 1$  but it is actually true  $H_0 = \text{TRUE}$  also known as the level of a test.

**Type2 Error:**

Test does not reject null hypothesis  $\psi = 0$  but alternative hypothesis is true  $H_1 = \text{TRUE}$

**Example:** Let  $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} \text{Ber}(p^*)$ .

Question: is  $p^* = 1/2$ .

$H_0: p^* = 1/2; H_1: p^* \neq 1/2$

If asymptotic level  $\alpha$  then we need to standardize the estimated parameter  $\hat{p} = \bar{X}_n$  first.

$$T_n = \sqrt{n} \frac{[\bar{X}_n - 0.5]}{\sqrt{0.5(1-0.5)}}$$

$$\psi_n = 1\{T_n > q_{\alpha/2}\}$$

where  $q_{\alpha/2}$  denotes the  $q_{\alpha/2}$  quantile of a standard Gaussian, and  $\alpha$  is determined by the required level of  $\psi$ . Note the absolute value in  $T_n$  for this two sided test.

**Pivot:**

Let  $T_n$  be a function of the random samples  $X_1, \dots, X_n, \theta$ . Let  $g(T_n)$  be a random variable whose distribution is the same for all  $\theta$ . Then,  $g$  is called a pivotal quantity or a pivot.

**Example:** let  $X$  be a random variable with mean  $\mu$  and variance  $\sigma^2$ . Let  $X_1, \dots, X_n$  be iid samples of  $X$ . Then,

$$g_n \triangleq \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}$$

is a pivot with  $\theta = [\mu, \sigma^2]^T$  being the parameter vector (not the same set of parameters that we use to define a statistical model).

### 6.1 P-Value

The (asymptotic) p-value of a test  $\psi_\alpha$  is the smallest (asymptotic) level  $\alpha$  at which  $\psi_\alpha$  rejects  $H_0$ . It is random since it depends on the sample. It can also be interpreted as the probability that the test-statistic  $T_n$  is realized given the null hypothesis.

If  $pvalue \leq \alpha$ ,  $H_0$  is rejected by  $\psi_\alpha$  at the (asymptotic) level  $\alpha$

The smaller the p-value, the more confidently one can reject  $H_0$ .

**Left-tailed p-values:**

$$\begin{aligned} pvalue &= \mathbb{P}(X \leq x|H_0) \\ &= \mathbb{P}(Z < T_{n,\theta_0}(\bar{X}_n)) \\ &= \Phi(T_{n,\theta_0}(\bar{X}_n)) \\ Z &\sim N(0, 1) \end{aligned}$$

**Right-tailed p-values:**

$$pvalue = \mathbb{P}(X \geq x|H_0)$$

**Two-sided p-values:** If asymptotic, create normalized  $T_n$  using parameters from  $H_0$ . Then use  $T_n$  to get to probabilities.

$$pvalue = 2\min\{\mathbb{P}(X \leq x|H_0), \mathbb{P}(X \geq x|H_0)\}$$

$$\mathbb{P}(|Z| > |T_{n,\theta_0}(\bar{X}_n)|) = 2(1 - \Phi(T_n))$$

$$Z \sim N(0, 1)$$

## 6.2 Comparisons of two proportions

Let  $X_1, \dots, X_n \stackrel{iid}{\sim} \text{Bern}(p_x)$  and  $Y_1, \dots, Y_n \stackrel{iid}{\sim} \text{Bern}(p_y)$  and be  $X$  independent of  $Y$ .  $\hat{p}_x = 1/n \sum_{i=1}^n X_i$  and  $\hat{p}_y = 1/n \sum_{i=1}^n Y_i$

$$H_0: p_x = p_y; H_1: p_x \neq p_y$$

To get the asymptotic Variance use multivariate Delta-method. Consider  $\hat{p}_x - \hat{p}_y = g(\hat{p}_x, \hat{p}_y); g(x, y) = x - y$ , then

$$\begin{aligned} \sqrt{n}(g(\hat{p}_x, \hat{p}_y) - g(p_x, p_y)) &\xrightarrow[n \rightarrow \infty]{(d)} \\ N(0, \nabla g(p_x - p_y)^T \Sigma \nabla g(p_x - p_y)) & \\ \Rightarrow N(0, p_x(1 - p_x) + p_y(1 - p_y)) & \end{aligned}$$

## 7 Non-asymptotic Hypothesis tests

### 7.1 Chi squared

The  $\chi_k^2$  distribution with  $d$  degrees of freedom is given by the distribution of  $Z_1^2 + Z_2^2 + \dots + Z_d^2$ , where  $Z_1, \dots, Z_d \stackrel{iid}{\sim} N(0, 1)$

If  $V \sim \chi_k^2$ :

$$\mathbb{E} = \mathbb{E}[Z_1^2] + \mathbb{E}[Z_2^2] + \dots + \mathbb{E}[Z_d^2] = d$$

$$\begin{aligned} \text{Var}(V) &= \text{Var}(Z_1^2) + \text{Var}(Z_2^2) + \dots + \\ \text{Var}(Z_d^2) &= 2d \end{aligned}$$

**Cochranes Theorem:**

If  $X_1, \dots, X_n \stackrel{iid}{\sim} N(\mu, \sigma^2)$ , then sample mean  $\bar{X}_n$  and the sample variance  $S_n$  are independent. The sum of squares of  $n$  variables follows a chi squared distribution with  $(n-1)$  degrees of freedom:

$$\frac{nS_n}{\sigma^2} \sim \chi_{n-1}^2$$

If formula for unbiased sample variance is used:

$$\frac{(n-1)S_n}{\sigma^2} \sim \chi_{n-1}^2$$

### 7.2 Student's T Test

Non-asymptotic hypothesis test for small samples (works on large samples too), data must be gaussian.

**Student's T distribution** with  $d$  degrees of freedom:  $t_d := \frac{Z}{\sqrt{V/n}}$  where  $Z \sim N(0, 1)$  and  $V \sim \chi_k^2$  are independent.

**Student's T test (one sample + two-sided):**

Let  $X_1, \dots, X_n \stackrel{iid}{\sim} N(\mu, \sigma^2)$  and suppose we want to test  $H_0: \mu = \mu_0 = 0$  vs.  $H_1: \mu \neq 0$ .

Test statistic follows Student's T distribution:

$$\begin{aligned} T_n &= \frac{\bar{X}_n - \mu_0}{\frac{s}{\sqrt{n}}} \\ &= \frac{\bar{X}_n - \mu_0}{\sqrt{\frac{s^2}{n}}} \\ &= \frac{\sqrt{n}(\bar{X}_n - \mu_0)}{\sqrt{\frac{s^2}{n}}} \\ &\sim \frac{N(0, 1)}{\sqrt{\frac{\chi_{n-1}^2}{n-1}}} \\ &\sim t_{n-1} \end{aligned}$$

Works bc. under  $H_0$  the numerator  $N(0, 1)$  and the denominator  $\frac{s^2}{\sigma^2} \sim \frac{1}{n-1} \chi_{n-1}^2$  are independent by Cochran's Theorem.

Student's T test at level  $\alpha$ :

$$\psi_\alpha = 1\{|T_n| > q_{\alpha/2}(t_{n-1})\}$$

**Student's T test (one sample, one-sided):**

$$\psi_\alpha = 1\{T_n > q_\alpha(t_{n-1})\}$$

**Student's T test (two samples, two-sided):**

Let  $X_1, \dots, X_n \stackrel{iid}{\sim} N(\mu_X, \sigma_X^2)$  and  $Y_1, \dots, Y_n \stackrel{iid}{\sim} N(\mu_Y, \sigma_Y^2)$ , suppose we want to test  $H_0: \mu_X = \mu_Y$  vs  $H_1: \mu_X \neq \mu_Y$ .

$$T_{n,m} = \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}$$

**Welch-Satterthwaite formula:**

When samples are different sizes we need to find the Student's T distribution of:  $T_{n,m} \sim t_N$

Calculate the degrees of freedom for  $t_N$  with:

$$N = \frac{\left(\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}\right)^2}{\frac{\sigma_X^2}{n^2(n-1)} + \frac{\sigma_Y^2}{m^2(m-1)}} \geq \min(n, m)$$

$N$  should be rounded down.

### 7.3 Walds Test

Squared distance of  $\hat{\theta}_n^{MLE}$  to true  $\theta_0$  using the fisher information  $I(\hat{\theta}_n^{MLE})$  as metric.

Let  $X_1, \dots, X_n \stackrel{iid}{\sim} P_{\theta^*}$  for some true parameter  $\theta^* \in \mathbb{R}^d$  and the maximum likelihood estimator  $\hat{\theta}_n^{MLE}$  for  $\theta^*$ .

Test  $H_0: \theta^* = 0$  vs  $H_1: \theta^* \neq 0$

Under  $H_0$ , the asymptotic normality of the MLE  $\hat{\theta}_n^{MLE}$  implies that:

$$\|\sqrt{n}I(0)^{1/2}(\hat{\theta}_n^{MLE} - 0)\|^2 \xrightarrow[n \rightarrow \infty]{(d)} \chi_d^2$$

**Test statistic:**

## 8.2 KL divergence

The KL divergence (aka relative entropy) KL between probability measures  $P$  and  $Q$  with the common sample space  $E$  and pmf/pdf functions  $f$  and  $g$  is defined as:

$$KL(P, Q) =$$

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

The KL divergence is not a distance measure! Always sum over the support of  $P$ !

Asymmetric in general:  
 $KL(P, Q) \neq KL(Q, P)$   
Nonnegative:  $KL(P, Q) \geq 0$   
Definite: if  $P = Q$  then  $KL(P, Q) = 0$ .  
Does not satisfy triangle inequality in general:  
 $KL(P, V) \leq KL(P, Q) + KL(Q, V)$

### Estimator of KL divergence:

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

$$\widehat{KL}(P_{\theta}, P_{\theta}) = \text{const} - \frac{1}{n} \sum_{i=1}^n \log(p_{\theta}(X_i))$$

## 9 Maximum likelihood estimation

Let  $\{E, \{P_{\theta}\}_{\theta \in \Theta}\}$  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  $X_i \sim P_{\theta^*}$ .

The **likelihood** of the model is the product of the  $n$  samples of the pdf/pmf:

$$L_n(X_1, X_2, \dots, X_n, \theta) =$$

$$\begin{cases} \prod_{i=1}^n p_{\theta}(x_i) & \text{if } E \text{ is discrete} \\ \prod_{i=1}^n f_{\theta}(x_i) & \text{if } E \text{ is continuous} \end{cases}$$

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

$$\begin{aligned} \widehat{\theta}_n^{MLE} &= \arg\min_{\theta \in \Theta} \widehat{KL}_n(P_{\theta^*}, P_{\theta}) \\ &= \arg\max_{\theta \in \Theta} \sum_{i=1}^n \ln p_{\theta}(X_i) \\ &= \arg\max_{\theta \in \Theta} \ln \left( \prod_{i=1}^n p_{\theta}(X_i) \right) \end{aligned}$$

Since taking derivatives of products is hard but easy for sums and  $\exp()$  is very common in pdfs we usually take the log of the likelihood function before maximizing it.

$$\ell((X_1, X_2, \dots, X_n, \theta)) = \ln(L_n(X_1, X_2, \dots, X_n, \theta)) = \sum_{i=1}^n \ln(L_i(X_i, \theta))$$

Cookbook: set up the likelihood function, take log of likelihood function. Take the partial derivative of the log-likelihood function wrt. the parameter(s). Set the partial derivative(s) 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.

## 9.1 Fisher information

The Fisher information is the covariance matrix of the gradient of the loglikelihood function. It is equal to the negative expectation of the Hessian of the loglikelihood function and captures the negative of the expected curvature of the loglikelihood function.

Let  $\theta \in \Theta \subset \mathbb{R}^d$  and let  $(E, \{P_{\theta}\}_{\theta \in \Theta})$  be a statistical model. Let  $f_{\theta}(x)$  be the pdf of the distribution  $P_{\theta}$ . Then, the Fisher information of the statistical model is.

$$\begin{aligned} \mathcal{I}(\theta) &= \text{Cov}(\nabla \ell(\theta)) = \\ &= \mathbb{E}[\nabla \ell(\theta) \nabla \ell(\theta)^T] - \\ &= \mathbb{E}[\nabla \ell(\theta)] \mathbb{E}[\nabla \ell(\theta)]^T = \\ &= -\mathbb{E}[\mathcal{H} \ell(\theta)] \end{aligned}$$

Where  $\ell(\theta) = \ln f_{\theta}(X)$ . If  $\nabla \ell(\theta) \in \mathbb{R}^d$  it is a  $d \times d$  matrix. The definition when the distribution has a pmf  $p_{\theta}(x)$  is also the same, with the expectation taken with respect to the pmf.

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

Formula for the calculation of Fisher Information of  $X$ :

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

Models with one parameter (ie. Bernulli):

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

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

Models with multiple parameters (ie. Gaussians):

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

Cookbook:

Better to use 2nd derivative.

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

## 9.2 Asymptotic normality of the maximum likelihood estimator

Under certain conditions the MLE is asymptotically normal and consistent. This applies even if the MLE is not the sample average.

Let the true parameter  $\theta^* \in \Theta$ . Necessary assumptions:

- The parameter is identifiable
- For all  $\theta \in \Theta$ , the support  $P_{\theta}$  does not depend on  $\theta$  (e.g. like in  $Unif(0, \theta)$ );
- $\theta^*$  is not on the boundary of  $\Theta$ ;
- Fisher information  $\mathcal{I}(\theta)$  is invertible in the neighborhood of  $\theta^*$

- A few more technical conditions

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

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

## 10 Method of Moments

Let  $X_1, \dots, X_n \stackrel{iid}{\sim} P_{\theta^*}$  associated with model  $(\mathbb{R}, \{P_{\theta}\}_{\theta \in \Theta})$ , with  $\mathbb{E} \subseteq \mathbb{R}$  and  $\Theta \subseteq \mathbb{R}$ , for some  $d \geq 1$   
Population moments:

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

Empirical moments:

$$\widehat{m}_k(\theta) = \overline{X_k^k} = \frac{1}{n} \sum_{i=1}^n X_i^k$$

Convergence of empirical moments:

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

$$(\widehat{m}_1, \dots, \widehat{m}_d) \xrightarrow[n \rightarrow \infty]{P, a.s.} (m_1, \dots, m_d)$$

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  $\theta^*$ ). Find the moments (as many as parameters), set up system of equations, solve for parameters, use empirical moments to estimate.

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

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

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

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:

$$\sqrt{n}(\widehat{\theta}_n^{MAP} - \theta) \xrightarrow[n \rightarrow \infty]{(d)} N(0, \Gamma)$$

$$\begin{aligned} \text{where, } \Gamma(\theta) &= \\ &= \left[ \frac{\partial M^{-1}}{\partial \theta}(M(\theta)) \right]^T \Sigma(\theta) \left[ \frac{\partial M^{-1}}{\partial \theta}(M(\theta)) \right] \\ \Gamma(\theta) &= \nabla_{\theta}(M^{-1})^T \Sigma_{\theta}(M^{-1}) \end{aligned}$$

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

## 11 Bayesian Statistics

Bayesian inference conceptually amounts to weighting the likelihood  $L_n(\theta)$  by a prior knowledge we might have on  $\theta$ . Given a statistical model we technically model our parameter  $\theta$  as if it were a random variable. We therefore define the **prior distribution** (PDF):

$$\pi(\theta)$$

Let  $X_1, \dots, X_n$ . We note  $L_n(X_1, \dots, X_n | \theta)$  the joint probability distribution of  $X_1, \dots, X_n$  conditioned on  $\theta$  where  $\theta \sim \pi$ . This is exactly the likelihood from the frequentist approach.

### 11.1 Bayes' formula

The posterior distribution verifies:

$$\forall \theta \in \Theta, \pi(\theta | X_1, \dots, X_n) \propto$$

$$\pi(\theta) L_n(X_1, \dots, X_n | \theta)$$

The constant is the normalization factor to ensure the result is a proper distribution, and does not depend on  $\theta$ :

$$\pi(\theta | X_1, \dots, X_n) = \frac{\pi(\theta) L_n(X_1, \dots, X_n | \theta)}{\int \pi(\theta) L_n(X_1, \dots, X_n | \theta) d\theta}$$

We can often use a **prior**, let i.e. a prior that is not a proper probability distribution (whose integral diverges), and still get a proper posterior. For example, the improper prior  $\pi(\theta) = 1$  on  $\Theta$  gives the likelihood as a posterior.

## 11.2 Jeffreys Prior

$$\pi_J(\theta) \propto \sqrt{\det I(\theta)}$$

where  $I(\theta)$  is the Fisher information. This prior is **invariant by reparameterization**, which means that if we have  $\eta = \phi(\theta)$ , then the same prior gives us a probability distribution for  $\eta$  verifying:

$$\pi_J(\eta) \propto \sqrt{\det I(\eta)}$$

The change of parameter follows the following formula:

$$\pi_J(\eta) = \det(\nabla \phi^{-1}(\eta)) \pi_J(\phi^{-1}(\eta))$$

## 11.3 Bayesian confidence region

Let  $\alpha \in (0, 1)$ . A Bayesian confidence region with level  $\alpha^*$  is a random subset  $\mathcal{R} \subset \Theta$  depending on  $X_1, \dots, X_n$  (and the prior  $\pi$ ) such that:

$$P[\theta \in \mathcal{R} | X_1, \dots, X_n] \geq 1 - \alpha$$

Bayesian confidence region and confidence interval are **distinct notions**. The Bayesian framework can be used to estimate the true underlying parameter. In that case, it is used to build a new class of estimators, based on the posterior distribution.

## 11.4 Bayes estimator posterior mean:

$$\widehat{\theta}(\pi) = \int_{\Theta} \theta \pi(\theta | X_1, \dots, X_n) d\theta$$

**Maximum a posteriori estimator (MAP):**

$$\widehat{\theta}_{(\pi)}^{MAP} = \arg\max_{\theta \in \Theta} \pi(\theta | X_1, \dots, X_n)$$

The MAP is equivalent to the MLE, if the prior is uniform.

## 12 OLS

Given two random variables  $X$  and  $Y$ , how can we predict the values of  $Y$  given  $X$ ?

Let us consider  $(X_1, Y_1), \dots, (X_n, Y_n) \sim iid \mathbb{P}$  where  $\mathbb{P}$  is an unknown joint distribution.  $\mathbb{P}$  can be described entirely by:

$$g(X) = \int f(X, y) dy$$

$$h(Y | X = x) = \frac{f(x, Y)}{g(x)}$$

where  $f$  is the joint PDF,  $g$  the marginal density of  $X$  and  $h$  the conditional density. What we are interested in is  $h(Y | X)$ .

**Regression function:** For a partial description, we can consider instead the conditional expectation of  $Y$  given  $X = x$ :

$$x \mapsto f(x) = \mathbb{E}[Y | X = x] = \int y h(y | x) dy$$

We can also consider different descriptions of the distribution, like the median, quantiles or the variance.

**Linear regression:** trying to fit any function to  $\mathbb{E}[Y | X = x]$  is a nonparametric problem; therefore, we restrict the problem to the tractable one of linear function:

$$f: x \mapsto a + bx$$

**Theoretical linear regression:** let  $X, Y$  be two random variables with two moments such as  $\mathbb{V}[X] > 0$ . The theoretical linear regression of  $Y$  on  $X$  is the line  $a^* + b^*x$  where

$$(a^*, b^*) = \arg\min_{(a, b) \in \mathbb{R}^2} \mathbb{E}[(Y - a - bX)^2]$$

Which gives:

$$b^* = \frac{\text{Cov}(X, Y)}{\mathbb{V}[X]}, \quad a^* = \mathbb{E}[Y] - b^* \mathbb{E}[X]$$

**Noise:** we model the noise of  $Y$  around the regression line by a random variable  $\varepsilon = Y - a^* - b^*X$ , such as:

$$\mathbb{E}[\varepsilon] = 0, \quad \text{Cov}(X, \varepsilon) = 0$$

We have to estimate  $a^*$  and  $b^*$  from the data. We have  $n$  random pairs  $(X_1, Y_1), \dots, (X_n, Y_n) \sim iid (X, Y)$  such as:

$$Y_i = a^* + b^* X_i + \varepsilon_i$$

The **Least Squares Estimator (LSE)** of  $(a^*, b^*)$  is the minimizer of the squared sum:

$$(\hat{a}_n, \hat{b}_n) = \arg\min_{(a, b) \in \mathbb{R}^2} \sum_{i=1}^n (Y_i - a - bX_i)^2$$

The estimators are given by:

$$\hat{b}_n = \frac{\overline{XY} - \overline{X}\overline{Y}}{\overline{X^2} - \overline{X}^2}, \quad \hat{a}_n = \overline{Y} - \hat{b}_n \overline{X}$$

The **Multivariate Regression** is given by:

$$Y_i = \sum_{j=1}^p X_i^{(j)} \beta_j^* + \varepsilon_i = \underbrace{X_i^\top}_{1 \times p} \underbrace{\beta^*}_{p \times 1} + \varepsilon_i$$

We can assume that the  $X_i^{(1)}$  are 1 for the intercept.

- If  $\beta^* = (a^*, b^* \top)^\top$ ,  $\beta_1^* = a^*$  is the intercept.
- the  $\varepsilon_i$  is the noise, satisfying  $\text{Cov}(X_i, \varepsilon_i) = 0$

The **Multivariate Least Squares Estimator (LSE)** of  $\beta^*$  is the minimizer of the sum of square errors:

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^n (Y_i - X_i^\top \beta)^2$$

**Matrix form:** we can rewrite these expressions. Let  $Y = (Y_1, \dots, Y_n)^\top \in \mathbb{R}^n$ , and  $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^\top$ . Let

$$X = \begin{pmatrix} X_1^\top \\ \vdots \\ X_n^\top \end{pmatrix} \in \mathbb{R}^{n \times p}$$

$X$  is called the **design matrix**. The regression is given by:

$$Y = X\beta^* + \varepsilon$$

and the LSE is given by:

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \|Y - X\beta\|_2^2$$

Let us suppose  $n \geq p$  and  $\text{rank}(X) = p$ . If we write:

$$F(\beta) = \|Y - X\beta\|_2^2 = (Y - X\beta)^\top (Y - X\beta)$$

Then:

$$\nabla F(\beta) = 2X^\top (Y - X\beta)$$

**Least squares estimator:** setting  $\nabla F(\beta) = 0$  gives us the expression of  $\hat{\beta}$ :

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y$$

**Geometric interpretation:**  $X\hat{\beta}$  is the orthogonal projection of  $Y$  onto the subspace spanned by the columns of  $X$ :

$$X\hat{\beta} = PY$$

where  $P = X(X^\top X)^{-1} X^\top$  is the expression of the projector.

**Statistic inference:** let us suppose that:  
\* The design matrix  $X$  is deterministic and  $\text{rank}(X) = p$ .  
\* The model is **homoscedastic**:  $\varepsilon_1, \dots, \varepsilon_n$  are i.i.d.  
\* The noise is Gaussian:  $\varepsilon \sim N_n(0, \sigma^2 I_n)$ .  
We therefore have:

$$Y \sim N_n(X\beta^*, \sigma^2 I_n)$$

Properties of the LSE:

$$\hat{\beta} \sim N_p(\beta^*, \sigma^2 (X^\top X)^{-1})$$

The quadratic risk of  $\hat{\beta}$  is given by:

$$\mathbb{E}[\|\hat{\beta} - \beta^*\|_2^2] = \sigma^2 \text{Tr}(X^\top X)^{-1})$$

The prediction error is given by:

$$\mathbb{E}[\|Y - X\hat{\beta}\|_2^2] = \sigma^2(n - p)$$

The unbiased estimator of  $\sigma^2$  is:

$$\hat{\sigma}^2 = \frac{1}{n - p} \|Y - X\hat{\beta}\|_2^2 = \frac{1}{n - p} \sum_{i=1}^n \varepsilon_i^2$$

By **Cochran's Theorem**:

$$(n - p) \frac{\hat{\sigma}^2}{\sigma^2} \sim \chi_{n-p}^2, \quad \hat{\beta} \perp \hat{\sigma}^2$$

**Significance test:** let us test  $H_0: \beta_j = 0$  against  $H_1: \beta_j \neq 0$ . Let us call

$$\gamma_j = \left( (X^\top X)^{-1} \right)_{jj} > 0$$

then:

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 \gamma_j}} \sim t_{n-p}$$

We can define the test statistic for our test:

$$T_n^{(j)} = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 \gamma_j}}$$

The test with non-asymptotic level  $\alpha$  is given by:

$$\psi_\alpha^{(j)} = \mathbb{1}[\|T_n^{(j)}\| > q_{\alpha/2}(t_{n-p})]$$

**Bonferroni's test:** if we want to test the significance level of multiple tests at the same time, we cannot use the same level  $\alpha$  for each of them. We must use a stricter test for each of them. Let us consider  $S \subseteq \{1, \dots, p\}$ . Let us consider

The **Bonferroni's test** with significance level  $\alpha$  is given by:

$$H_0: \forall j \in S, \beta_j = 0, \quad H_1: \exists j \in S, \beta_j \neq 0$$

$$\psi_\alpha^{(S)} = \max_{j \in S} \psi_{\alpha/K}^{(j)}$$

where  $K = |S|$ . The rejection region therefore is the union of all rejection regions:

$$R_\alpha^{(S)} = \bigcup_{j \in S} R_{\alpha/K}^{(j)}$$

This test has nonasymptotic level at most  $\alpha$ :

$$\mathbb{1}_{H_0} \left[ R_\alpha^{(S)} \right] \leq \sum_{j \in S} \mathbb{1}_{H_0} \left[ R_{\alpha/K}^{(j)} \right] = \alpha$$

This test also works for implicit testing (for example,  $\beta_1 \geq \beta_2$ ).

## 13 Generalized Linear Models

We relax the assumption that  $\mu$  is linear. Instead, we assume that  $g \circ \mu$  is linear, for some function  $g$ :

$$g(\mu(x)) = x^\top \beta$$

The function  $g$  is assumed to be known, and is referred to as the link function. It maps the domain of the dependent variable to the entire real line. It has to be strictly increasing, it has to be continuously differentiable and its range is all of  $\mathbb{R}$ .

## 13.1 The Exponential Family

A family of distribution  $\{P_{\theta} : \theta \in \Theta\}$ , where the parameter space  $\Theta \subset \mathbb{R}^k$  is  $-k$  dimensional, is called a  $k$ -parameter exponential family on  $\mathbb{R}^d$  if the pmf or pdf  $f_{\theta} : \mathbb{R}^d \rightarrow \mathbb{R}$  of  $P_{\theta}$  can be written in the form:

$$\begin{aligned} f_{\theta}(y) &= h(y) \exp(\eta(\theta) \cdot T(y) - B(\theta)) \quad \text{where} \\ \eta(\theta) &= \begin{pmatrix} \eta_1(\theta) \\ \vdots \\ \eta_k(\theta) \end{pmatrix} : \mathbb{R}^k \rightarrow \mathbb{R}^k \\ T(y) &= \begin{pmatrix} T_1(y) \\ \vdots \\ T_k(y) \end{pmatrix} : \mathbb{R}^d \rightarrow \mathbb{R}^k \\ B(\theta) &: \mathbb{R}^k \rightarrow \mathbb{R} \\ h(y) &: \mathbb{R}^d \rightarrow \mathbb{R} \end{aligned}$$

if  $k = 1$  it reduces to:



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

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

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:

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

## 15 Variance

Variance is the squared distance from the mean.

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

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

Variance of a product with constant  $a$ :

$$\text{Var}(aX) = a^2 \text{Var}(X)$$

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

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

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

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

$$\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y)$$

## 16 Covariance

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

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

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

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

Possible notations:

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

Covariance is commutative:

$$\text{Cov}(X, Y) = \text{Cov}(Y, X)$$

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

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

Useful properties:

$$\text{Cov}(aX + b, cY + d) = ac\text{Cov}(X, Y)$$

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

$$\text{Cov}(aX + b, cY + d) = a\text{Cov}(X, Y) + b\text{Cov}(Y, Z)$$

If  $\text{Cov}(X, Y) = 0$ , 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, i.e. any linear combination  $\alpha X + \beta Y$  is gaussian for all  $(\alpha, \beta) \in \mathbb{R}^2$  without  $[0, 0]$ .

## 17 correlation coefficient

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

## 18 Important probability distributions

### Bernoulli

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

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

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

$$\text{Var}(X) = p(1 - p)$$

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:

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

MLE:

$$\hat{p}_{MLE} = \frac{\sum_{i=1}^n X_i}{n}$$

Fisher Information:

$$I(p) = \frac{1}{p(1-p)}$$

Canonical exponential form:

$$f_\theta(y) = \exp\left(y\theta - \underbrace{\ln(1 + e^\theta)}_{b(\theta)} + \underbrace{0}_{c(y, \phi)}\right)$$

$$\theta = \ln\left(\frac{p}{1-p}\right)$$
$$\phi = 1$$

### Binomial

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

$$p_X(k) = \binom{n}{k} p^k (1 - p)^{n-k}, k = 0, \dots, n$$

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

$$\text{Var}(X) = np(1 - p)$$

Likelihood:

$$L_n(X_1, \dots, X_n, \theta) = \left(\prod_{i=1}^n \binom{n}{X_i}\right) \theta^{\sum_{i=1}^n X_i} (1 - \theta)^{nK - \sum_{i=1}^n X_i}$$

Loglikelihood:

$$\ell_n(\theta) = C + \left(\sum_{i=1}^n X_i\right) \log \theta + (nK - \sum_{i=1}^n X_i) \log(1 - \theta)$$

MLE:

Fisher Information:

$$I(p) = \frac{n}{p(1-p)}$$

Canonical exponential form:

$$f_\theta(y) = \exp\left(y \underbrace{(\ln(p) - \ln(1 - p))}_{\theta} + n \underbrace{\ln(1 - p)}_{-b(\theta)} + \ln\left(\binom{n}{y}\right) \underbrace{1}_{c(y, \phi)}\right) \underbrace{p^y (1 - p)^{n-y}}_{\phi(y, \phi)} = \begin{cases} \lambda \exp(-\lambda x), & \text{if } x \geq 0 \\ 0, & \text{o.w.} \end{cases}$$

### Geometric

Number of  $T$  trials up to (and including) the first success.

$$p_T(t) = (1 - p)^{t-1}, t = 1, 2, \dots$$

$$\mathbb{E}[T] = \frac{1}{p}$$

$$\text{var}(T) = \frac{1-p}{p^2}$$

### Pascal

The negative binomial or Pascal distribution is a generalization of the geometric distribution. It relates to the random experiment of repeated independent trials until observing  $m$  successes. I.e. the time of the  $k$ th arrival.

$$Y_k = T_1 + \dots + T_k$$

$T_i \sim \text{iid Geometric}(p)$

$$\mathbb{E}[Y_k] = \frac{k}{p}$$

$$\text{Var}(Y_k) = \frac{k(1-p)}{p^2}$$

$$p_{Y_k}(t) = \binom{t-1}{k-1} p^k (1 - p)^{t-k}$$
$$t = k, k + 1, \dots$$

### Multinomial

Parameters  $n > 0$  and  $p_1, \dots, p_r$ .

$$p_X(x) = \frac{n!}{x_1! \dots x_r!} p_1^{x_1} \dots p_r^{x_r}$$

$$\mathbb{E}[X_i] = n \cdot p_i$$

$$\text{Var}(X_i) = np_i(1 - p_i)$$

Likelihood:

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

Loglikelihood:

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

### Poisson

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

$$p_X(k) = \exp(-\lambda) \frac{\lambda^k}{k!} \text{ for } k = 0, 1, \dots$$

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

$$\text{Var}(X) = \lambda$$

Likelihood:

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

Loglikelihood:

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

MLE:

$$\hat{\lambda}_{MLE} = \frac{1}{n} \sum_{i=1}^n (X_i)$$

Fisher Information:

$$I(\lambda) = \frac{1}{\lambda}$$

Canonical exponential form:

$$f_\theta(y) = \exp\left(y\theta - \underbrace{e^\theta}_{b(\theta)} - \underbrace{\ln y!}_{c(y, \phi)}\right)$$

$$\theta = \ln \lambda$$

$$\phi = 1$$

Poisson process:

$k$  arrivals in  $t$  slots

$$p_X(k, t) = \mathbb{P}(N_t = k) = e^{-\lambda t} \frac{(\lambda t)^k}{k!}$$

$$\mathbb{E}[N_t] = \lambda t$$

$$\text{Var}(N_t) = \lambda t$$

### Exponential

Parameter  $\lambda$ , continuous

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

$$\mathbb{E}[X] = \frac{1}{\lambda}$$

$$\mathbb{E}[X^2] = \frac{2}{\lambda^2}$$

$$\text{Var}(X) = \frac{1}{\lambda^2}$$

Likelihood:

$$L(X_1 \dots X_n; \lambda) = \lambda^n \exp(-\lambda \sum_{i=1}^n X_i)$$

Loglikelihood:

$$\ell_n(\lambda) = n \ln(\lambda) - \lambda \sum_{i=1}^n (X_i)$$

MLE:

$$\hat{\lambda}_{MLE} = \frac{n}{\sum_{i=1}^n (X_i)}$$

Fisher Information:

$$I(\lambda) = \frac{1}{\lambda^2}$$

Canonical exponential form:

$$f_\theta(y) = \exp\left(y\theta - \underbrace{(-\ln(-\theta))}_{b(\theta)} + \underbrace{0}_{c(y, \phi)}\right)$$

$$\theta = -\lambda = -\frac{1}{\mu}$$

$$\phi = 1$$

### Shifted Exponential

Parameters  $\lambda, a \in \mathbb{R}$ , continuous

$$f_X(x) = \begin{cases} \lambda \exp(-\lambda(x - a)), & x \geq a \\ 0, & x < a \end{cases}$$

$$F_X(x) = \begin{cases} 1 - \exp(-\lambda(x - a)), & \text{if } x \geq a \\ 0, & x < a \end{cases}$$

$$\mathbb{E}[X] = a + \frac{1}{\lambda}$$

$$\text{Var}(X) = \frac{1}{\lambda^2}$$

Likelihood:

$$L(X_1 \dots X_n; \lambda, \theta) = \lambda^n \exp(-\lambda \sum_{i=1}^n (X_i - a)) \mathbf{1}_{\min_{i=1, \dots, n} (X_i) \geq a}$$

Loglikelihood:

$$\ell(\lambda, a) := n \ln \lambda - \lambda \sum_{i=1}^n (X_i - a) + n \ln a$$

MLE:

$$\hat{\lambda}_{MLE} = \frac{1}{\bar{X}_n - \hat{a}}$$

$$\hat{a}_{MLE} = \min_{i=1, \dots, n} (X_i)$$

### Univariate Gaussians

Parameters  $\mu$  and  $\sigma^2 > 0$ , continuous

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

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

$$\text{Var}(X) = \sigma^2$$

CDF of standard gaussian:

$$\Phi(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

Likelihood:

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

Loglikelihood:

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

MLE:

$$\hat{\mu}_{MLE} = \bar{X}_n$$
$$\hat{\sigma}_{MLE}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

Fisher Information:

$$I(\mu, \sigma^2) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^4} \end{pmatrix}$$

Canonical exponential form:

Gaussians are invariant under affine transformation:

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

Sum of independent gaussians:

$$\text{Let } X \sim N(\mu_X, \sigma_X^2) \text{ and } Y \sim N(\mu_Y, \sigma_Y^2)$$

$$\text{If } Y = X + Z, \text{ then } Y \sim N(\mu_X + \mu_Y, \sigma_X + \sigma_Y)$$

$$\text{If } U = X - Y, \text{ then } U \sim N(\mu_X - \mu_Y, \sigma_X + \sigma_Y)$$

Symmetry:

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

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

Standardization:

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

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

Higher moments:

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

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

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

Quantiles:

### Uniform

Parameters  $a$  and  $b$ , continuous.

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

$$F_X(x) = \begin{cases} 0, & \text{for } x \leq a \\ \frac{x-a}{b-a}, & x \in [a, b] \\ 1, & x \geq b \end{cases}$$

$$\mathbb{E}[X] = \frac{a+b}{2}$$

$$\text{Var}(X) = \frac{(b-a)^2}{12}$$

Likelihood:

$$L(X_1 \dots X_n; b) = \frac{1}{b^n} \mathbf{1}_{(\max_i X_i \leq b)}$$

Loglikelihood:

### Cauchy

continuous, parameter  $m$ ,

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

$$\mathbb{E}[X] = \text{not defined!}$$

$$\text{Var}(X) = \text{not defined!}$$

$$\text{med}(X) = P(X > M) = P(X < M)$$

$$= 1/2 = \int_{1/2}^{\infty} \frac{1}{1 + (x-m)^2} dx$$

### Chi squared

The  $\chi_d^2$  distribution with  $d$  degrees of freedom is given by the distribution

$$\text{of } Z_1^2 + Z_2^2 + \dots + Z_d^2, \text{ where } Z_1, \dots, Z_d \stackrel{\text{iid}}{\sim} \mathcal{N}(0, 1)$$

If  $V \sim \chi_k^2$ :

$$\mathbb{E} = \mathbb{E}[Z_1^2] + \mathbb{E}[Z_2^2] + \dots + \mathbb{E}[Z_d^2] = d$$

$$\text{Var}(V) = \text{Var}(Z_1^2) + \text{Var}(Z_2^2) + \dots + \text{Var}(Z_d^2) = 2d$$

### Student's T Distribution

$T_n := \frac{Z}{\sqrt{V/n}}$  where  $Z \sim \mathcal{N}(0, 1)$ , and  $Z$

and  $V$  are independent

### 18.1 Useful to know

#### 18.1.1 Min of iid exponential r.v

Let  $X_1, \dots, X_n$  be i.i.d.  $\text{Exp}(\lambda)$  random variables.

Distribution of  $\min_i(X_i)$

$$\mathbb{P}(\min_i(X_i) \leq t) = 1 - \mathbb{P}(\min_i(X_i) \geq t)$$

$$= 1 - (\mathbb{P}(X_1 \geq t))(\mathbb{P}(X_2 \geq t)) \dots (\mathbb{P}(X_n \geq t))$$

$$= 1 - (1 - F_X(t))^n = 1 - e^{-n\lambda t}$$

Differentiate w.r.t  $x$  to get the pdf of  $\min_i(X_i)$ :

$$f_{\min}(x) = (n\lambda)e^{-(n\lambda)x}$$

#### 18.1.2 Counting Committees

Out of  $2n$  people, we want to choose a committee of  $n$  people, one of whom will be its chair. In how many different ways can this be done?"

$$\binom{2n}{n} = 2n \binom{2n-1}{n-1}$$

"In a group of  $2n$  people, consisting of  $n$  boys and  $n$  girls, we want to select a committee of  $n$  people. In how many ways can this be done?"

$$\binom{2n}{n} = \sum_{i=0}^n \binom{n}{i} \binom{n}{n-i}$$

"How many subsets does a set with  $2n$  elements have?"

$$\binom{2n}{n} = \sum_{i$$



## 21 Matrixalgebra

$$\|Ax\|^2 = (Ax)^T(Ax) = x^T A^T Ax = x^T A^T Ax$$

## 22 Calculus

Differentiation under the integral sign

$$\frac{d}{dx} \left( \int_{a(x)}^{b(x)} f(x,t) dt \right) = f(x,b(x))b'(x) - f(x,a(x))a'(x) + \int_{a(x)}^{b(x)} f_x(x,t) dt$$

### Concavity in 1 dimension

If  $g: I \rightarrow \mathbb{R}$  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  $\nabla$  of a twice differentiable function  $f$  is defined as:

$$\nabla f: \mathbb{R}^d \rightarrow \mathbb{R}^d$$
$$\theta = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_d \end{pmatrix} \mapsto \begin{pmatrix} \frac{\partial f}{\partial \theta_1} \\ \frac{\partial f}{\partial \theta_2} \\ \vdots \\ \frac{\partial f}{\partial \theta_d} \end{pmatrix}_\theta$$

### Hessian

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

$$Hh(\theta) = \nabla^2 h(\theta) =$$

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

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

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

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

Negative semi-definite (resp. negative definite):

$x^T A x$  is negative for all  $x \in \mathbb{R}^d - \{0\}$ .

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).

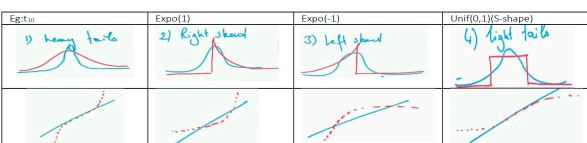
If the Hessian has both positive and negative eigenvalues then  $a$  is a saddle point for  $f$ .

### 23 Covariance Matrix

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

Matrix outer products!

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



Sample space:	Probability axioms:	Countable additivity axiom:
<ul style="list-style-type: none"><li>Mutually exclusive</li><li>Collectively exhaustive</li><li>Right Granularity</li></ul>	<ul style="list-style-type: none"><li><math>P(A) \geq 0</math></li><li><math>P(\Omega) = 1</math></li><li><math>P(A \cup B) = P(A) + P(B)</math> if A and B are disjoint</li></ul>	$P(A_1 \cup A_2 \cup \dots) = P(A_1) + P(A_2) + \dots$ If $A_1, A_2, \dots$ is an infinite sequence of disjoint events (holds only for countable sets) <b>Union Bound</b> : $P(A \cup B) \leq P(A) + P(B)$

### Inclusion-exclusion formula:

$$P(\bigcup_{k=1}^n A_k) = \sum_i P(A_i) - \sum_{i < j} P(A_i \cap A_j) + \sum_{i < j < k} P(A_i \cap A_j \cap A_k) - \dots + (-1)^{n-1} P(\bigcap_{k=1}^n A_k)$$

Conditional Probability for EVENTS:	Conditional PMF, given an event:	Probability Mass Function(PMF)(discrete r.v.)	Probability Density function(PDF)(continuous r.v.)
$P(A B) = \frac{P(A \cap B)}{P(B)}$ where $P(B) > 0$ Conditional probabilities share same properties like ordinary probabilities (follow probability axioms) Multiplicative rule: $P(A \cap B) = P(A) * P(B A) = P(B) * P(A B)$ General form: $P(A_1 \cap A_2 \cap \dots A_n) = P(A_1) * \prod_{i=2}^n P(A_i   A_1 \cap A_2 \cap \dots \cap A_{i-1})$ Total Probability theorem: $P(B) = \sum_i P(A_i) * P(B A_i)$ Independence: $P(B A) = P(B)$ occurrence of A provides no new information about B $P(A \cap B) = P(A) * P(B)$ then A and B are independent <ul style="list-style-type: none"><li>If A and B are independent then A and <math>B^c</math> are independent</li><li>If A and B are independent then <math>A^c</math> and <math>B^c</math> are independent</li><li>If A and B are independent then B and <math>A^c</math> are independent</li></ul> Pairwise independence: A, B and C are pairwise independent means: $P(A \cap B) = P(A) * P(B)$ $P(B \cap C) = P(B) * P(C)$ $P(A \cap C) = P(A) * P(C)$ It does not imply A, B and C are independent i.e.: $P(A \cap B \cap C) \neq P(A) * P(B) * P(C)$ General form: $P(A_1 \cap A_2 \cap \dots A_n) = P(A_1) * P(A_2) * \dots * P(A_n)$ then $A_1, A_2, \dots, A_n$ are Independent <b>Conditional Independence:</b> $P(A \cap B   C) = P(A C) * P(B C)$ If A and B are independent it doesn't imply conditional independence	$P(X A)(x) = P(X = x A)$ where $P(A) > 0$ $\sum_x P(X A)(x) = 1$ Total Probability theorem: $P(X) = \sum_i P(A_i) * P(X A_i)$ Expectation: $\sum_x g(x) * P(X A)(x)$ $E[X] = \sum_i P(A_i) * E[X A_i]$ Independence: $P_{X A}(x) = P(X=x)$ for all x $P(X=x \text{ and } A) = P(X=x) * P(A)$ for all x <b>Conditional PDF, given an event:</b> $f(X A)(x) = \begin{cases} \frac{f_X(x)}{P(A)} & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases}$ For small interval $\delta$ : $f_X(x) \approx P(x \in [x, x+\delta]   A) / \delta$ where $P(A) > 0$ $\int f_X(x) dx = 1$ Total Probability theorem: $f_X(x) = \sum_i P(A_i) * f_{X A_i}(x)$ Expectation: $E[X A] = \int x * f_X(x) dx$ $E[g(X) A] = \int g(x) * f_X(x) dx$ $E[X] = \sum_i P(A_i) * E[X A_i]$	$P_X(x) = P(X = x)$ $\sum_x P_X(x) = 1$ Properties: <ul style="list-style-type: none"><li><math>P_X(x) \geq 0</math></li><li><math>\sum_x P_X(x) = 1</math></li></ul> Cumulative distribution function: $F_X(x) = P(X \leq x) = \sum_{k \leq x} P_X(k)$ The graph of $F_X(x)$ is a step(staircase) function $P(X=x) = F_X(x+1) - F_X(x)$ (Size of the jump or increase in height of the stair) Expectation/mean: $E[X] = \sum_x x * P_X(x)$ exists only when $\sum_x  x  * P_X(x) < \infty$ Variance: $Var(X) = V(X) = E[(X - \mu)^2]$ $Var(X) = E[X^2] - [E[X]]^2$ Standard deviation $\sigma_X = \sqrt{Var(X)}$	$P(a \leq X \leq b) = \int_a^b f_X(x) dx$ For small interval $\delta$ : $P(a \leq X \leq a + \delta) \approx f_X(a) * \delta$ PDF are not probabilities but probabilities per unit length i.e. densities Properties: <ul style="list-style-type: none"><li><math>f_X(x) \geq 0</math></li><li><math>\int_{-\infty}^{\infty} f_X(x) dx = 1</math></li><li><math>P(X=x) = 0</math> since area at point is zero</li><li><math>P(a \leq X \leq b) = P(a &lt; X &lt; b)</math></li></ul> Cumulative distribution function: $F_X(x) = P(X \leq x) = \int_{-\infty}^x f_X(x) dx$ $\frac{dF_X(x)}{dx} = f_X(x)$ this derivative might be a bit difficult to evaluate at edge points where $f_X(x)$ is not differentiable Expectation/mean: $E[X] = \int_{-\infty}^{\infty} x * f_X(x) dx$ exists only when $\int_{-\infty}^{\infty}  x  * f_X(x) dx < \infty$ Variance: $Var(X) = V(X) = E[(X - \mu)^2] = E[X^2] - [E[X]]^2$ Standard deviation $\sigma_X = \sqrt{Var(X)}$

BAYES THEOREM:	Conditional PMF:	Conditional PDF:
$P(A_i B) = \frac{P(A_i) * P(B A_i)}{\sum_j P(A_j) * P(B A_j)}$	$P(X Y)(x y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}$ where $p_Y(y) > 0$ $\sum_x p(X Y)(x y) = 1$ Total Probability theorem: $p_X(x) = \sum_y p_Y(y) * p(X Y)(x y)$ Expectation: $E(X Y) = \sum_x x * p(X Y)(x y)$ $E[g(X) Y] = \sum_x g(x) * p(X Y)(x y)$ $E[X] = \sum_y p_Y(y) * E(X Y = y)$ Independence: $p_{X Y}(x) = p_X(x)$ for all x,y $p_{X,Y}(x,y) = p_X(x) * p_Y(y)$ for all x,y If X,Y are independent then $g(X)$ and $h(Y)$ are also independent and: $E[g(X) * h(Y)] = E[g(X)] * E[h(Y)]$ If X and Y are independent then $V(X+Y) = V(X) + V(Y)$	$f(X Y)(x y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$ where $f_Y(y) > 0$ For small interval $\delta$ : $P(a \leq X \leq a + \delta   y) \approx f_{X,Y}(x,y) * \delta$ The conditional PDF is slice of joint PDF Total Probability theorem: $f_X(x) = \int_y f_Y(y) * f(X Y)(x y) dy$ Expectation: $E(X Y = y) = \int x * f_{X Y}(x y) dx$ $E[g(X) Y] = \int g(x) * f_{X Y}(x y) dx$ $E[X] = \int_y f_Y(y) * E[X Y = y] dy$ Independence: $f_{X,Y}(x,y) = f_X(x) * f_Y(y)$ for all x,y Also: $f_{X Y}(x y) = f_X(x)$ $f_{X,Y}(x,y) = F_X(x) * F_Y(y)$ If X,Y are independent then $g(X)$ and $h(Y)$ are also independent and: $E[g(X) * h(Y)] = E[g(X)] * E[h(Y)]$ If X and Y are independent then $V(X+Y) = V(X) + V(Y)$
<b>More than two r.v. joint pmf:</b> The same goes for joint PDF just replace summation by integration $p_{X,Y,Z}(x,y,z) = P(X = x \text{ and } Y = y \text{ and } Z = z)$ $\sum_x \sum_y \sum_z p_{X,Y,Z}(x,y,z) = 1$ $p_X(x) = \sum_y \sum_z p_{X,Y,Z}(x,y,z)$ $p_{X,Y}(x,y) = \sum_z p_{X,Y,Z}(x,y,z)$ $p_{X,Z}(x,z) = \sum_y p_{X,Y,Z}(x,y,z)$ $p_{Y,Z}(y,z) = \sum_x p_{X,Y,Z}(x,y,z)$		

Parametric test	Walds test	T-test	Two sample Wald test	Two sample t-test
Test Statistics:	$T_n = \frac{\hat{\theta} - \theta_0}{\sqrt{\text{var}(\hat{\theta})}} = \frac{nI(\hat{\theta})^{1/2}(\hat{\theta} - \theta_0)}{\sqrt{I(\theta_0)^{-1}}} \xrightarrow{(d)} N(0,1)$ <small>only guarantees asymptotic level of significance</small>	$T_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} \xrightarrow{d} N_{n-1}$ <small>Applied even when n is small and the parameter for testing is normally distributed</small>	$T_n = \frac{\bar{X}_n - \bar{\theta}_0}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{m}}} \xrightarrow{(d)} N(0,1)$ <small>where: <math>\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i</math>, <math>s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2</math></small>	$\bar{X}_n - \bar{Y}_n - (\mu_1 - \mu_2) \sim I_N$ $\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{m}}$ where: $N = \frac{(\frac{s_1^2}{n} + \frac{s_2^2}{m})}{\frac{s_1^2}{n} + \frac{s_2^2}{m}} \xrightarrow{d} \min(n,m)$
<b>Bonferroni inequality:</b> $P(A_1 \cap A_2) \geq P(A_1) + P(A_2) - 1$ General form: $P(A_1 \cap A_2 \cap \dots A_n) \geq P(A_1) + P(A_2) + \dots + P(A_n) - (n-1)$		<b>De Morgan's Law:</b> 1. $(\bigcup_{i=1}^n S_n)^c = \bigcap_{i=1}^n S_n^c$ 2. $(\bigcap_{i=1}^n S_n)^c = \bigcup_{i=1}^n S_n^c$		<b>Combination:</b> $\binom{n}{k}$ = no of k - elements subsets from a given n - element set/Binomial coefficient $\sum_{k=0}^n \binom{n}{k} = 2^n$ $\frac{n!}{n_1! n_2! \dots n_r!} =$ Dividing a n - element set into r partition each with $n_i$ distinct elements/multinomial coefficient
<b>Reliability:</b> Given n components and probability of components working is p		Series: P(System is up)= $p^n$	Parallel: P(System is up)= $1 - \{1-p\}^n$	
<b>Discrete r.v.:</b> <b>PMF:</b> $P_X(x) = P(X = x)$ $= P(\{\omega \in \Omega \text{ such that } X(\omega) = x\})$  Properties: <ul style="list-style-type: none"><li><math>P_X(x) \geq 0</math></li><li><math>\sum_x P_X(x) = 1</math></li></ul> Cumulative distribution function: $F_X(x) = P(X \leq x) = \sum_{k \leq x} P_X(k)$ The graph of $F_X(x)$ is a step(staircase) function $P(X=x) = F_X(x + 1) - F_X(x)$ (Size of the jump or increase in height of the stair)  Expectation/mean: $E[X] = \sum_x x * P_X(x)$ exists only when $\sum_x  x  * P_X(x) < \infty$  Variance: $Var(X) = V(X) = E[(X - \mu)^2]$ $Var(X) = E[X^2] - [E[X]]^2$ Standard deviation= $\sigma_X = \sqrt{Var(X)}$	<b>Continuous r.v.:</b> <b>PDF:</b> $f_X(x) = \frac{dP_X(x)}{dx}$ For small interval $\delta$ : $P(a \leq X \leq a + \delta) \approx f_X(a) * \delta$ PDF are not probabilities but probabilities per unit length i.e. densities  Properties: <ul style="list-style-type: none"><li><math>f_X(x) \geq 0</math></li><li><math>\int_x f_X(x) dx = 1</math></li><li><math>P(X=x)=0</math> since area at point is zero</li><li><math>P(a \leq X \leq b) = P(a &lt; X &lt; b)</math></li></ul> Cumulative distribution function: $F_X(x) = P(X \leq x) = \int_{-\infty}^x f_X(x) dx$ $\frac{dF_X(x)}{dx} = f_X(x)$ this derivative might be a bit difficult to evaluate at edge points where $f_X(x)$ is not differentiable  Expectation/mean: $E[X] = \int_x x * f_X(x) dx$ exists only when $\int_x  x  * f_X(x) dx < \infty$  Variance: $Var(X) = V(X) = E[(X - \mu)^2] = E[X^2] - [E[X]]^2$ Standard deviation= $\sigma_X = \sqrt{Var(X)}$			
<b>Joint PMF:</b> $p_{X,Y}(x,y) = P(X = x \text{ and } Y = y)$ $\sum_x \sum_y p_{X,Y}(x,y) = 1$ $p_{X,Y}(x,y) \geq 0$ $p_X(x) = \sum_y p_{X,Y}(x,y)$ $p_Y(y) = \sum_x p_{X,Y}(x,y)$ Expected value rule: $E[g(X,Y)] = \sum_x \sum_y g(x,y) * p_{X,Y}(x,y)$	<b>Joint PDF:</b> $P(a \leq X \leq a + \delta, c \leq Y \leq c + \delta) \approx f_{X,Y}(x,y) * \delta^2$ Joint PDF are not probabilities but probabilities per unit area $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y) dx dy = 1$ $f_{X,Y}(x,y) \geq 0$ $f_X(x) = \int_y f_{X,Y}(x,y) dy$ $f_Y(y) = \int_x f_{X,Y}(x,y) dx$ Expected value rule: $E[g(X,Y)] = \int \int g(x,y) * f_{X,Y}(x,y) dx dy$			
<b>Joint CDF:</b> $F_{X,Y}(x,y) dx dy = P(X \leq x, Y \leq y) = \int_{-\infty}^x \int_{-\infty}^y f_{X,Y}(x,y) dx dy$ $f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}}{\partial x \partial y}(x,y)$		<b>Covariance:</b> It tells us if X and Y are moving in same directions or not $Cov(X,Y) = E[(X - E[X]) * (Y - E[Y])] = E[XY] - E[X] * E[Y]$ If X,Y are independent then $Cov(X,Y) = 0$ converse not true Properties: <ul style="list-style-type: none"><li><math>Cov(X,X) = Var(X)</math></li><li><math>Cov(a * X + b, Y) = a * Cov(X,Y)</math></li><li><math>Cov(X, Y+Z) = Cov(X,Y) + Cov(X,Z)</math></li></ul>		
<b>Correlation coefficient:</b> $\rho(X,Y) = E \left[ \left( \frac{X - E[X]}{\sigma_X} \right) * \left( \frac{Y - E[Y]}{\sigma_Y} \right) \right] = \frac{Cov(X,Y)}{\sigma_X * \sigma_Y}$ Properties: <ul style="list-style-type: none"><li><math>-1 \leq \rho \leq 1</math> i.e. <math> \rho  \leq 1</math></li><li>If X,Y are independent then <math>\rho = 0</math> is true converse is not true</li><li>If <math> \rho  = 1</math> then <math>(X - E[X]) = c * (Y - E[Y])</math> (X and Y are linearly related)</li><li><math>\rho(a * X + b, Y) = \text{sign}(a) * \rho(X,Y)</math></li><li>When <math>\rho = \frac{1}{2}</math> this means that X and Y have a underlying, common, hidden factor (e.g. <math>X = V + W</math>, <math>Y = V + Z</math> where V is common hidden factor)</li></ul>				

Frequentist	Bayesian
Parameter is not a r.v. but a fixed value	Parameter is a r.v. with certain prior distribution
$P(a<\theta<b)=0$ or 1 depending upon if the true parameter lies in the interval	

Definition:

of  $k$  – elements subsets from a given  $n$  – element set/Binomial coefficient

$$\sum_{k=0}^n \binom{n}{k} = 2^n$$

= Dividing a  $n$  – element set into  $r$  partition each with  $n_i$  distinct elements/multinomial coefficient

**Properties of expectation which is true for both continuous and discrete r.v.:**

- If  $X \geq 0$ , then  $E[X] \geq 0$
- If  $a \leq X \leq b$  then  $a \leq E[X] \leq b$
- $E[a] = a$  where  $a$  is a constant.  $E[a * X + b] = a * E[X] + b$  where  $a$  and  $b$  are constants
- $E[X+Y] = E[X] + E[Y]$  more generally  $E[X_1+X_2+...+X_n] = E[X_1] + E[X_2] + ... + E[X_n]$  (this is applicable even if the variables are not independent)
- If  $X$  and  $Y$  are independent  $E[XY] = E[X] * E[Y]$
- Discrete r.v.:  $E[g(X)] = \sum_x g(x) * P_X(x)$
- Continuous r.v.:  $E[g(x)] = \int_{-\infty}^{\infty} g(x) * f_X(x)$
- $E[g(X)] \neq g(E[X])$  But if  $g(x)$  is linear function  $g(x) = a * X + b$  then only  $E[g(X)] = g(E[X])$
- When the PMF is symmetric about a point then that point is the expected value of the PMF
- When the PMF is not symmetric the expected value is interpreted as center of gravity of the PMF
- Discrete r.v.: if  $X$  takes non negative integer value  $E[X] = \sum_{x=1}^{\infty} P(X \geq x)$
- Continuous r.v.: if  $X$  takes non negative integer value  $E[X] = \int_0^{\infty} 1 - F_X(x) dx = \int_0^{\infty} P(X \geq x) dx$

**Properties of variance which is true for both continuous and discrete r.v.:**

- $V(aX+b) = a^2 * V(X)$
- $V(X) \geq 0$  (ALWAYS)
- If  $X$  and  $Y$  are independent then  $V(X+Y) = V(X) + V(Y)$  or else it is  $V(X+Y) = V(X) + V(Y) + 2 * Cov(X,Y)$   
General form:  $V(X_1+X_2+...+X_n) = \sum_{i=1}^n V(X_i) + \sum_{i \neq j} Cov(X_i, X_j)$  (there are  $n^2 - n$  where  $i \neq j$ )

**Properties of CDF which is true for both continuous and discrete r.v.:**

- It's always a monotonically increasing function i.e.:  $y \geq x$  then  $F_X(y) \geq F_X(x)$
- $F_X(x)$  tends to 1 as  $x \rightarrow \infty$  i.e.:  $F_X(\infty) = 1$
- $F_X(x)$  tends to 0 as  $x \rightarrow -\infty$  i.e.:  $F_X(-\infty) = 0$
- $F(x)$  is right-continuous; that is, for every number  $x_0$ ,  $\lim_{x \rightarrow x_0^+} F(x) = F(x_0)$

**Mixed distribution:**

$X = \begin{cases} Y \text{ with probability } p \\ Z \text{ with probability } 1 - p \end{cases}$

Where  $Y$  is discrete r.v. and  $Z$  is a continuous r.v.

$X$  is neither continuous nor discrete it is a mixed r.v.

$F_X(x) = p * F_Y(x) + (1 - p) * F_Z(x)$

$E[X] = p * E[Y] + (1 - p) * E[Z]$

**Indicator r.v.:**

If  $X$  is the indicator of an event  $A$

$X = I_A$  where  $I_A = 1$  iff  $A$  occur  $E[I_A] = P(A)$

**Random Variable:**

A function from  $\Omega$  to the real numbers

**Sum of a random number of independent r.v.:**

$Y = X_1 + X_2 + ... + X_N$ , where  $N$  is a non-negative integer r.v.,  $X_i$ 's are i.i.d and independent of  $N$

$E[Y] = E[N] * E[X]$      $Var[Y] = E[N] * Var(X) + (E[X])^2 * Var(N)$

**Derived distribution-the discrete case:**

$Y = g(X)$

$p_Y(y) = \sum_{x:g(x)=y} p_X(x)$

$Y = a * X + b$

$p_Y(y) = p_X(\frac{y - b}{a})$

For  $g(X)$  being strictly increasing or decreasing we get:

$Y = g(X)$

$X = g^{-1}(Y) = h(Y)$

$f_Y(y) = |\frac{dh(y)}{dy}| * f_X(h(y))$

**Derived distribution of multiple r.v.'s:**

$Z = g(X,Y) = X + Y$

Discrete case:

$p_Z(z) = \sum_x p_X(x) * p_Y(z - x)$

Given that  $X$  and  $Y$  are independent

**Derived distribution-the continuous case:**

$Y = g(X)$

Two step procedure:

- Find CDF of  $Y = P(Y \leq y) = P(g(X) \leq y)$
- $f_Y(y) = \frac{dP_Y(y)}{dy}$

$Y = a * X + b$

$f_Y(y) = \frac{1}{|a|} * f_X(\frac{y - b}{a})$

Same two step procedure: Find CDF of  $Z$  and then differentiate

$f_Z(z) = \int f_X(x) * f_Y(z - x) dx$

Given that  $X$  and  $Y$  are independent

$f_Z(z) = f_X(z) * f_Y(z - x)$      $f_{X,Z}(x, z) = f_X(x) f_Y(z - x)$