1 Statistical models

$$E, \{\mathbb{P}_\theta\}_{\theta\in\Theta}$$

E is a sample space for X i.e. a set that contains all possible outcomes of X
$\{\mathbb{P}_\theta\}_{\theta\in\Theta}$ is a family of probability distributions on E.
Θ is a parameter set, i.e. a set consisting of some possible values of Θ.
θ is the true parameter and unknown.
In a parametric model we assume that $\Theta\subset\mathbb{R}^d$, for some $d\ge1$.

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 ($\overline{X_n}$, $max(X_i)$, etc).

An **estimator** $\hat\theta_n$ of θ is any statistic which does not depend on θ.

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

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

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

σ^2 is called the **Asymptotic Variance** of the estimator $\hat\theta_n$. In the case of the sample mean it is the same variance as as the single X_i.
If the estimator is a function of the sample mean the **Delta Method** is needed to compute the asymptotic variance.**Asymptotic Variance** $\neq$ Variance of an estimator.
Bias of an estimator:

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

Quadratic risk of an estimator

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

3 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 $i=1,2,...,n$ and $\overline{X_n}=\frac1n\sum_{i=1}^n X_i$.
Law of large numbers:

$$\overline{X_n}\xrightarrow[n\to\infty]{P,a.s.}\mu$$
$$\frac1n\sum_{i=1}^n g(X_i)\xrightarrow[n\to\infty]{P,a.s.}\mathbb{E}[g(X)]$$

Central Limit Theorem for Mean:

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

Central Limit Theorem for Sums:

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

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:

$$E[\overline{X_n}]=\frac1n E[X_1+X_2,...,X_n]$$
$$= \mu.$$

4 Quantiles of a Distribution

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

$$\mathbb{P}(X\le q_\alpha)=q_\alpha = 1-\alpha$$
$$\mathbb{P}(X\ge q_\alpha)=\alpha$$
$$F_X(q_\alpha)=1-\alpha$$
$$F_X^{-1}(1-\alpha)=\alpha$$

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

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

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).

$$\mathbf{P}(X\le t)=\mathbf{P}\left(Z\le\frac{t-\mu}{\sigma}\right)$$
$$= \Phi\left(\frac{t-\mu}{\sigma}\right)$$
$$Z=\frac{X-\mu}{\sigma}\sim N(0,1)$$
$$q_\alpha=\frac{t-\mu}{\sigma}$$

5 Confidence intervals

Confidence Intervals follow the form:

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

Let $(E,(\mathbb{P}_\theta)_{\theta\in\Theta})$ 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 θ:
Any random interval $\mathcal{I}$, depending on the sample $X_1,\dots X_n$ but not at θ and such that:
$\mathbb{P}_\theta[\mathcal{I}\ni\theta]\ge1-\alpha,\ \forall\theta\in\Theta$
Confidence interval of **asymptotic level** $1-\alpha$ for θ:
Any random interval $\mathcal{I}$ whose boundaries do not depend on θ and such that:
$\lim_{n\to\infty}\mathbb{P}_\theta[\mathcal{I}\ni\theta]\ge1-\alpha,\ \forall\theta\in\Theta$

5.1 Two-sided asymptotic CI

Let $X_1,\dots,X_n=\tilde X$ and $\tilde X\overset{iid}{\sim}P_\theta$. A two-sided CI is a function depending on $\tilde X$ giving an upper and lower bound in which the estimated parameter lies $\mathcal{I}=[l(\tilde X,u(\tilde X)]$ with a certain probability $\mathbb{P}(\theta\in\mathcal{I})\ge1-q_\alpha$ and conversely $\mathbb{P}(\theta\notin\mathcal{I})\le\alpha$
Since the estimator is a r.v. depending on $\tilde X$ it has a variance $Var(\hat\theta_n$ and a mean $\mathbb{E}[\hat\theta_n]$. Since the CLT is valid for every distribution standardizing the distributions and massaging the expression yields an asymptotic CI:

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

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

5.2 Sample Mean and Sample Variance

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

$$\overline{X}_n=\frac1n\sum_{i=1}^nX_i$$

Sample Variance:

$$S_n=\frac1n\sum_{i=1}^n(X_i-\overline{X}_n)^2$$
$$= \frac1n(\sum_{i=1}^nX_i^2)-\overline{X}_n^2$$

Unbiased estimator of sample variance:

$$\tilde S_n=\frac{1}{n-1}\sum_{i=1}^n\left(X_i-\overline{X}_n\right)^2$$
$$= \frac{n}{n-1}S_n$$

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. $\sqrt(n)(Z_n-\theta)\xrightarrow[n\to\infty]{(d)}N(0,\sigma^2)$ and let $g:R\longrightarrow R$ be continuously differentiable at θ, then:

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

Example: let $X_1,...,X_n$ $exp(\lambda)$ where $\lambda>0$. Let $\overline{X}_n=\frac1n\sum_{i=1}^nX_i$ denote the sample mean. By the CLT, we know that $\sqrt n\left(\overline{X}_n-\frac1\lambda\right)\xrightarrow[n\to\infty]{(d)}N(0,\sigma^2)$ for some value of σ^2 that depends on λ.
If we set $g:\mathbb{R}\to\mathbb{R}$ and $x\mapsto1/x$, then by the Delta method:

$$\sqrt n\left(g(\overline{X}_n)-g\left(\frac1\lambda\right)\right)$$
$$\xrightarrow[n\to\infty]{(d)} N(0,g'(E[X])^2\text{Var}X)$$
$$\xrightarrow[n\to\infty]{(d)} N(0,g'\left(\frac1\lambda\right)^2\frac{1}{\lambda^2})$$
$$\xrightarrow[n\to\infty]{(d)} N(0,\lambda^2)$$

6 Asymptotic Hypothesis tests

Two hypotheses (Θ_0 disjoint set from Θ_1): $\begin{cases}H_0:\theta\epsilon\Theta_0\\H_1:\theta\epsilon\Theta_1\end{cases}$. Goal is to reject H_0 using a test statistic.

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

A hypothesis-test has the form

$$\psi=\mathbf{1}\{T_n\ge c\}$$

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:

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

Power of the test:

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

Where β_ψ is the probability of making a Type2 Error and inf is the maximum.
Two-sided test:

$$H_1:\theta\neq\Theta_0$$
$$\mathbf{1}(|T_n|>q_{\alpha/2})$$

One-sided tests:

$$H_1:\theta>\Theta_0$$
$$\mathbf{1}(T_n<-q_\alpha)H_1 \qquad :\theta<\Theta_0$$
$$\mathbf{1}(T_n>q_\alpha)$$

Type1 Error:
Test rejects null hypothesis $\psi=1$ but it is actually true $H_0=TRUE$ also known as the level of a test.
Type2 Error:
Test does not reject null hypothesis $\psi=0$ but alternative hypothesis is true $H_1=TRUE$

Example: Let $X_1,\dots,X_n\overset{i.i.d.}{\sim}\text{Ber}(p^*)$. Question: is $p^*=1/2$.
$H_0:p^*=1/2; H_1:p^*\neq1/2$
If asymptotic level α then we need to standardize the estimated parameter $\hat p=\overline{X}_n$ first.

$$T_n=\sqrt n\frac{|\overline{X}_n-0.5|}{\sqrt{0.5(1-0.5)}}$$
$$\psi_n=\mathbf{1}(T_n>q_{\alpha/2})$$

where $q_{\alpha/2}$ denotes the $q_{\alpha/2}$ quantile of a standard Gaussian, and α is determined by the required level of ψ. 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 θ. Then, g is called a pivotal quantity or a pivot.
Example: let X be a random variable with mean μ and variance σ^2. Let $X_1,\dots,X_n$ be iid samples of X. Then,

$$g_n\triangleq\frac{\overline{X_n}-\mu}{\sigma}$$

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

6.1 P-Value

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

If $pvalue\le\alpha$, H_0 is rejected by ψ_α at the (asymptotic) level α

The smaller the p-value, the more confidently one can reject H_0.
Left-tailed p-values:

$$pvalue=\mathbb{P}(X\le x|H_0)$$
$$= \mathbf{P}(Z<T_{n,\theta_0}(\overline{X}_n)))$$
$$= \Phi(T_{n,\theta_0}(\overline{X}_n))$$
$$Z\sim\mathcal{N}(0,1)$$

Right-tailed p-values:

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

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

$$pvalue=2min\{\mathbb{P}(X\le x|H_0),\mathbb{P}(X\ge x|H_0)\}$$
$$\mathbb{P}(|Z|>|T_{n,\theta_0}(\overline{X}_n)|=2(1-\Phi(T_n))$$
$$Z\sim N(0,1)$$

6.2 Comparisons of two proportions

Let $X_1,\dots,X_n\overset{iid}{\sim}Bern(p_x)$ and $Y_1,\dots,Y_n\overset{iid}{\sim}Bern(p_y)$ and be X independent of Y. $\hat p_x=1/n\sum_{i=1}^nX_i$ and $\hat p_x=1/n\sum_{i=1}^nY_i$

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

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

$\sqrt(n)(g(\hat p_x,\hat p_y)-g(p_x-p_y))\xrightarrow[n\to\infty]{(d)}$
$N(0,\nabla g(p_x-p_y)^T\Sigma\nabla g(p_x-p_y))$

$$\Rightarrow N(0,p_x(1-px)+p_y(1-py))$$

7 Non-asymptotic Hypothesis tests

7.1 Chi squared

The χ^2_d distribution with d degrees of freedom is given by the distribution of $Z_1^2+Z_2^2+\dots+Z_d^2$, where $Z_1,\dots,Z_d\overset{iid}{\sim}\mathcal{N}(0,1)$
If $V\sim\chi^2_k$:

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

$Var(V)=Var(Z_1^2)+Var(Z_2^2)+\ldots+Var(Z_d^2)=2d$

Cochranes Theorem:
If $X_1,...,X_n\overset{iid}{\sim}N(\mu,\sigma^2)$, then sample mean $\overline{X}_n$ 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:

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

If formula for unbiased sample variance is used:

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

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 $Z\sim\mathcal{N}(0,1)$ and $V\sim\chi^2_k$ are independent.

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

Let $X_1,...,X_n\overset{iid}{\sim}N(\mu,\sigma^2)$ and suppose we want to test $H_0:\mu=\mu_0=0$ vs. $H_1:\mu\neq0$.

Test statistic follows Student's T distribution:

$$T_n=\frac{Z}{S}$$
$$= \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt n}}$$
$$= \frac{\sqrt n\frac{\overline{X}_n-\mu_0}{\sigma}}{\sqrt{\frac{\tilde S_n}{\sigma^2}}}$$
$$\sim \frac{N(0,1)}{\sqrt{\frac{\chi^2_{n-1}}{n-1}}}$$
$$\sim t_{n-1}$$

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

Student's T test at level α:

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

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

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

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

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

$$T_{n,m}=\frac{\overline{X}_n-\overline{Y}_m}{\sqrt{\frac{\hat\sigma^2{}_X}{n}+\frac{\hat\sigma^2{}_Y}{m}}}$$

Welch-Satterthwaite formula:

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

Calculate the degrees of freedom for t_N with:

$$N=\frac{\left(\frac{\hat\sigma^2{}_X}{n}+\frac{\hat\sigma^2{}_Y}{m}\right)^2}{\frac{\hat\sigma^2{}_X}{n^2(n-1)}+\frac{\hat\sigma^2{}_Y}{m^2(m-1)}}\ge\min(n,m)$$

N should be rounded down.

7.3 Walds Test

Squared distance of $\widehat\theta_n^{MLE}$ to true θ_0 using the fisher information $I(\widehat\theta_n^{MLE})$ as metric.

Let $X_1,\dots,X_n\overset{iid}{\sim}\mathbf{P}_{\theta^*}$ for some true parameter $\theta^*\in\mathbb{R}^d$ and the maximum likelihood estimator $\widehat\theta_n^{MLE}$ for θ^*.

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

Under H_0, the asymptotic normality of the MLE $\widehat\theta_n^{MLE}$ implies that:

$$\left\|\sqrt n\mathcal{I}(\mathbf{0})^{1/2}(\widehat\theta_n^{MLE}-\mathbf{0})\right\|^2\xrightarrow[n\to\infty]{(d)}\chi^2_d$$

Test statistic:

$$T_n=n(\widehat\theta_n^{MLE}-\theta_0)^\top I(\widehat\theta_n^{MLE})(\widehat\theta_n^{MLE}-\theta_0)$$
$$\xrightarrow[n\to\infty]{(d)}\chi^2_d$$

Wald test of level α:

$$\psi_\alpha=\mathbf{1}\{T_n>q_\alpha(\chi^2_d)\}$$

7.4 Likelihood Ratio Test

Parameter space $\Theta\subseteq\mathbb{R}^d$ and H_0 is that parameters θ_{r+1} through θ_d have values θ_c^{r+1} through θ_d^c leaving the other r unspecified. That is:
$H_0:(\theta_{r+1},...,\theta_d)^T=\theta_{r+1...d}=\theta_0$
Construct two estimators:

$$\widehat{\theta}_n^{MLE}=argmax_{\theta\in\Theta}(\ell_n(\theta))$$
$$\widehat{\theta}_n^c=argmax_{\theta\in\Theta_0}(\ell_n(\theta))$$

Test statistic:

$$T_n=2(\ell(X_1,..X_n|\widehat\theta_n^{MLE})-\ell(X_1,..X_n|\widehat\theta_n^c))$$

Wilk's Theorem: under H_0, if the MLE conditions are satisfied:

$$T_n\xrightarrow[n\to\infty]{(d)}\chi^2_{d-r}$$

Likelihood ratio test at level α:

$$\psi_\alpha=\mathbf{1}\{T_n>q_\alpha(\chi^2_{d-r})\}$$

7.5 Implicit Testing

Todo

7.6 Goodness of Fit Discrete Distributions

Let $X_1,...,X_n$ be iid samples from a categorical distribution. Test $H_0:p=p^0$ against $H_1:p\neq p^0$. Example: against the uniform distribution $p^0=(1/K,\dots,1/K)^\top$.

Test statistic under H_0:

$$T_n=n\sum_{k=1}^K\frac{(\hat p_k-p_k^0)^2}{p_k^0}\xrightarrow[n\to\infty]{(d)}\chi^2_{K-1}$$

Test at level alpha:

$$\psi_\alpha=\mathbb{1}\{T_n>q_\alpha(\chi^2_{K-1})\}$$

7.7 Kolmogorov-Smirnov test

7.8 Kolmogorov-Lilliefors test

7.9 QQ plots

Heavier tails: below > above the diagonal.
Lighter tails: above > below the diagonal.
Right-skewed: above > below > above the diagonal.
Left-skewed: below > above > below the diagonal.

8 Distances between distributions

8.1 Total variation distance

The total variation distance TV between the propability measures P and Q with a sample space E is defined as: $\text{TV}(\mathbf{P},\mathbf{Q})=\max_{A\subset E}|\mathbf{P}(A)-\mathbf{Q}(A)|$,
Calculation with f and g:

$$TV(\mathbf{P},\mathbf{Q})=$$
$$\begin{cases}\frac12\sum_{x\in E}|f(x)-g(x)|, discr\\ \frac12\int_{x\in E}|f(x)-g(x)|dx, cont\end{cases}$$

Symmetric: $TV(\mathbf{P},\mathbf{Q})=TV(\mathbf{Q},\mathbf{P})$
Positive: $TV(\mathbf{P},\mathbf{Q})\ge0$
Definite: $TV(\mathbf{P},\mathbf{Q})=0\iff\mathbf{P}=\mathbf{Q}$
Triangle inequality: $TV(\mathbf{P},\mathbf{V})\le TV(\mathbf{P},\mathbf{Q})+TV(\mathbf{Q},\mathbf{V})$

If the support of $\mathbf{P}$ and $\mathbf{Q}$ is disjoint:

$$TV(\mathbf{P},\mathbf{V})=1$$

TV between continuous and discrete r.v:

$$TV(\mathbf{P},\mathbf{V})=1$$

8.2 KL divergence

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

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

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

Asymetric in general: $KL(\mathbf{P},\mathbf{Q}) \neq KL(\mathbf{Q},\mathbf{P})$

Nonnegative: $KL(\mathbf{P},\mathbf{Q}) \geq 0$

Definite: if $\mathbf{P} = \mathbf{Q}$ then $KL(\mathbf{P},\mathbf{Q}) = 0$

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:

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

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

9 Maximum likelihood estimation

Let $\{E,(\mathbf{P}_\theta)_{\theta\in\Theta}\}$ 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 **likelihood** of the model is the product of the n samples of the pdf/pmf:

$$L_n(X_1, X_2, \ldots, X_n, \theta) = \begin{cases} \prod_{i=1}^n p_\theta(x_i) & \text{if } E \text{ is discrete} \\ \prod_{i=1}^n f_\theta(x_i) & \text{if } E \text{ is continous} \end{cases}$$

The maximum likelihood estimator is the (unique) θ that minimizes $\widehat{KL}(\mathbf{P}_{\theta^*},\mathbf{P}_\theta)$ 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.)

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

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.

$$\begin{aligned}\ell((X_1, X_2, \ldots, X_n, \theta)) &= ln(L_n(X_1, X_2, \ldots, X_n, \theta)) \\ &= \sum_{i=1}^n ln(L_i(X_i, \theta)\end{aligned}$$

Cookbook: set up the likelihood function, take log of likelihood function. Take the partial derivative of the loglikelihood 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 $(E,\{\mathbf{P}_\theta\}_{\theta\in\Theta})$ be a statistical model. Let $f_\theta(\mathbf{x})$ be the pdf of the distribution $\mathbf{P}_\theta$. Then, the Fisher information of the statistical model is.

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

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

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

Formula for the calculation of Fisher Information of X:

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

Models with one parameter (ie. Bernulli):

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

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

Models with multiple parameters (ie. Gaussians):

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

Cookbook:

Better to use 2nd derivative.

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

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 $\mathbb{P}_\theta$ does not depend on θ (e.g. like in $Unif(0,\theta)$);
- θ^* is not on the boundary of Θ;
- Fisher information $\mathcal{I}(\theta)$ is invertible in the neighborhood of θ^*
- A few more technical conditions

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

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

10 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:

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

Empirical moments:

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

Convergence of empirical moments:

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

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

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

11 Bayesian Statistics

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

$$\pi(\theta)$$

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

11.1 Bayes' formula

. The posterior distribution verifies:

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

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

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

We can often use an **improper prior**, 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 Θ gives the likelihood as a posterior.

11.2 Jeffreys Prior

$$\pi_J(\theta) \propto \sqrt{detI(\theta)}$$

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 η verifying:

$$\tilde{\pi}_J(\eta) \propto \sqrt{det\tilde{I}(\eta)}$$

The change of parameter follows the following formula:

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

11.3 Bayesian confidence region

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

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

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:

$$\hat{\theta}_{(\pi)} = \int_\Theta \theta\pi(\theta|X_1, \ldots, X_n)d\theta$$

Maximum a posteriori estimator (MAP):

$$\hat{\theta}_{(\pi)}^{MAP} = argmax_{\theta\in\Theta}\pi(\theta|X_1, \ldots, X_n)$$

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 $(X_1, Y_1), \ldots, (X_n, Y_n) \sim^{iid} \mathbb{P}$ 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 expection of Y given $X = x$:

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

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

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

Which gives:

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

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

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

We have to estimate a^* and b^* from the data. We have n random pairs $(X_1, Y_1), \ldots, (X_n, Y_n) \sim_{iid} (X,Y)$ 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:

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

The estimators are given by:

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

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 assuming that the $X_i^{(1)}$ are 1 for the intercept.

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

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

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

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

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:

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

and the LSE is given by:

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

Let us suppose $n \geq p$ and $rank(X) = p$. 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 $rank(X) = p$. * The model is **homoscedastic**: $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. * The noise is Gaussian: $\epsilon \sim N_n(0, \sigma^2 I_n)$. 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:

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

The prediction error is given by:

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

The unbiased estimator of σ^2 is:

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

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 α is given by:

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

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

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

The *Bonferroni's test* with significance level α is given by:

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

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 α:

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

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

13 Generalized Linear Models

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

$$g(\mu(\mathbf{x})) = \mathbf{x}^T\beta$$

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 $\{\mathbf{P}_\theta : \theta \in \Theta\}$, where the parameter space $\Theta \subset \mathbb{R}^k$ is -k dimensional, is called a k-parameter exponential family on $\mathbb{R}^1$ if the pmf or pdf $f_\theta : \mathbb{R}^q \to \mathbb{R}$ of $\mathbf{P}_\theta$ can be written in the form:

$$f_\theta(\mathbf{y}) = h(\mathbf{y})\exp(\boldsymbol{\eta}(\boldsymbol{\theta}) \cdot \mathbf{T}(\mathbf{y}) - B(\boldsymbol{\theta})) \quad \text{where}$$

$$\begin{cases} \boldsymbol{\eta}(\boldsymbol{\theta}) = \begin{pmatrix} \eta_1(\boldsymbol{\theta}) \\ \vdots \\ \eta_k(\boldsymbol{\theta}) \end{pmatrix} & : \mathbb{R}^k \to \mathbb{R}^k \\ \mathbf{T}(\mathbf{y}) = \begin{pmatrix} T_1(\mathbf{y}) \\ \vdots \\ T_k(\mathbf{y}) \end{pmatrix} & : \mathbb{R}^q \to \mathbb{R}^k \\ B(\boldsymbol{\theta}) & : \mathbb{R}^k \to \mathbb{R} \\ h(\mathbf{y}) & : \mathbb{R}^q \to \mathbb{R}. \end{cases}$$

if $k = 1$ it reduces to:

$$f_\theta(y) = h(y)\exp(\eta(\theta)T(y) - B(\theta))$$

14 Expectation

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

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

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

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

Total expectation theorem:

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

Law of iterated expectation:

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

Expectation of constant a:

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

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

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

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

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

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

Linearity of Expectation where a and c are given scalars:

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

If Variance of X is known:

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

15 Variance

Variance is the squared distance from the mean.

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

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

Variance of a product with constant a:

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

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

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

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

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

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

16 Covariance

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

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

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

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

Possible notations:

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

Covariance is commutative:

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

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

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

Useful properties:

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

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

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

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

17 correlation coefficient

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

18 Important probability distributions

Bernoulli

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

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

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

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

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:

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

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.

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

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

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

Likelihood:

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

Loglikelihood:

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

MLE:

Fisher Information:

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

Canonical exponential form:

$f_p(y) = exp(y\underbrace{(\ln(p)-\ln(1-p))}_{\theta} + \underbrace{n\ln(1-p)}_{-b(\theta)} + \underbrace{\ln(\binom{n}{y})}_{c(y,\phi)})$

Geometric

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

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

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

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

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 kth arrival.

$Y_k = T_1 + \dots T_k$

$T_i \sim iid Geometric(p)$

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

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

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

$t = k, k+1, \dots$

Multinomial

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

$p_x(x) = \frac{n!}{x_1!,\dots,x_n!}p_1,\dots,p_r$

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

$Var(X_i) = np_i(1-p_i)$

Likelihood:

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

Loglikelihood:

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

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 $k = 0,1,\dots,$

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

$Var(X) = \lambda$

Likelihood:

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

Loglikelihood:

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

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

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

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

$Var(N_t) = \lambda t$

Exponential

Parameter λ, continuous

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

$P(X>a) = exp(-\lambda a)$

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

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

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

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

Likelihood:

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

Loglikelihood:

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

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

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

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

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

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

Likelihood:

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

Loglikelihood:

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

MLE:

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

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

Univariate Gaussians

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

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

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

$Var(X) = \sigma^2$

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:

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

MLE:

$\hat{\mu}_MLE = \overline{X}_n$

$\hat{\sigma^2}_MLE = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2$

Fisher Information:

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

Canonical exponential form:

Gaussians are invariant under affine transformation:

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

Sum of independent gaussians:

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

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

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

Symmetry:

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

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

Standardization:

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

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

Higher moments:

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

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

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

Quantiles:

Uniform

Parameters a and b, continuous.

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

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

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

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

Likelihood:

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

Loglikelihood:

Cauchy

continuous, parameter m,

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

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

$Var(X) = notdefined!$

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

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

Chi squared

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

If $V \sim \chi_k^2$:

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

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

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 $X_1,\dots,X_n n$ be i.i.d. $Exp(\lambda)$ random variables.
Distribution of $min_i(Xi)$

$\mathbf{P}(\min_i(X_i) \leq t) =$

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

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

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

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

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

18.1.2 Counting Commitees

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?"

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

"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?"

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

"Out of n people, we want to form a committee consisting of a chair and other members. We allow the committee size to be any integer in the range $1,2,\dots,n$. How many choices do we have in selecting a committee-chair combination?"

$$n2^{n-1} = \sum_{i=0}^n \binom{n}{i}i.$$

18.2 Finding Joint PDFS

$f_{X,Y}(x,y) = f_X(x)f_{Y|X}(y\mid x)$

19 Random Vectors

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.

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

CDF of $\mathbf{X}$:

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

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

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

Expectation of a random vector

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

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

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

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

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

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

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

Covariance Matrix

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

Matrix outer products!

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

$\mathbb{E}\left(\begin{bmatrix} X_1-\mu_1 \\ X_2-\mu_2 \\ \dots \\ X_d-\mu_d \end{bmatrix}[X_1-\mu_1, X_2-\mu_2,\dots,X_d-\mu_d]\right)$

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

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

Alternative forms:

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

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

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

Every Covariance matrix is positive definite.

$\Sigma < 0$

Gaussian Random Vectors

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

Multivariate Gaussians

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

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

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

Where $\det(\Sigma)$ is the determinant of Σ, which is positive when Σ is invertible.
If $\mu = 0$ and Σ is the identity matrix, then X is called a standard normal random vector .
If the covariant matrix Σ is diagonal, the pdf factors into pdfs of univariate Gaussians, and hence the components are independent.

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

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

Multivariate CLT

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

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

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

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

Multivariate Delta Method

20 Algebra

Absolute Value Inequalities:

$|f(x)| < a \Rightarrow -a < f(x) < a$

$|f(x)| > a \Rightarrow f(x) > a$ or $f(x) < -a$

21 Matrixalgebra

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

22 Calculus

Differentiation under the integral sign

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

Concavity in 1 dimension

If $g : I \to \mathbb{R}$ is twice differentiable in the interval I:

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

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

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

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

Multivariate Calculus

The Gradient ∇ of a twice differntiable function f is defined as:

$\nabla f : \mathbb{R}^d \to \mathbb{R}^d$

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

Hessian

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

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

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

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

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

Negative semi-definite (resp. negative definite):

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

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.

23 Covariance Matrix

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

Matrix outer products!

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