Probability Rules

Set Properties
$S \cap (T \cup U) = (S \cap T) \cup (S \cup U)$
$S \cup (T \cap U) = (S \cup T) \cap (S \cup U)$

De Morgan's Law
$(S \cap T)^C = S^C \cup T^C$
$S^C \cap T^C = (S \cup T)^c$

Permutation – ordered arrangement of object ($N!$)

Combination – unordered arrangement of object $\binom{N}{n}$

Multinomial Probability
$$\frac{n!}{n_1!\,n_2!\ldots n_r!} p_1{}^{n_1} p_2{}^{n_2} \ldots p_r{}^{n_r}$$

Multiplication Rule
$$P_{X,Y,Z} = P_X(x)P_{X|Y}(x|y)P_{Z|X,Y}(z|x,y)$$

Marginal PDF
$$f_X(x) = \int f(x,y)dy$$

Conditional PDF
$$f_{X|Y} = \frac{f(x,y)}{f(y)}$$

Total Probability Theorem

Discrete
$$P_X(x) = \sum_y P_Y(y)P_{(X|Y)}(x|y)$$
Continuous
$$f_X(x) = \int_{-\infty}^{\infty} f_Y(y)f_{X|Y}(x|y)\,dy$$

Total Expectation Theorem

Discrete
$$E[X|Y=y] = \sum_x xP_{(X|Y)}(x|y)$$
$$E[X] = \sum_x p_Y(y)E[X|Y=y]$$
Continuous
$$E[X|Y=y] = \int_{-\infty}^{\infty} xf_{X|Y}(x|y)\,dx$$
$$E[X] = \int_{-\infty}^{\infty} f_Y(y)E[X|Y=y]\,dy$$

Law of Iterated Expectation
$E\big[E[Y|X]\big] = E[Y]$

Conditional Variance/ Law of Total Variance
$$Var(X) = E[Var(X|Y)] + Var(E[X|Y])$$

Mean and Variance of the Sum of a Random Number of Random Variables
$E[Y] = E[N]E[X]$
$$Var(Y) = E[N]Var(X) + E[X]^2Var(N)$$

Transformation of Random Variable

Linear
$$f_y(y) = \frac{1}{|a|}f_x\left(\frac{y-b}{a}\right)$$

General
$$f_y(y) = \frac{dx(y)}{dy}f_X(x(y))$$

Cookbook: For monotonic functions
Find the appropriate limits for Y
Find CDF $F_x(x(y))$
Differentiate with respect to y

Convolution
If X and Y are independent and $Z = X + Y$, then
$$f_z(z) = \int_{-\infty}^{\infty} f_x(z-y)f_y(y)dy$$

Linear Regression

Correlation Coefficient
$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X\sigma_Y}$$

Linear Least Mean squares
$$\widehat{\theta_L} = E[\theta] + \frac{Cov(X,Y)}{Var(X)}(X - E[X])$$
$$a = \frac{Cov(X,Y)}{Var(X)} = \rho\frac{\sigma_\theta}{\sigma_x} = \bar{Y} - \hat{b}\,\bar{X}$$
$$b = E[\theta] - \frac{Cov(X,Y)}{Var(X)}(E[X]) = \frac{\overline{XY} - \bar{X}\bar{Y}}{\overline{X^2} - \bar{X}^2}$$

MSE of the Linear Estimator
$$E\left[\left(\widehat{\theta_L} - \theta\right)^2\right] = (1-p^2)Var(\theta)$$

Multivariate Linear Regression Model
$$Y_i = \beta_i^{TX} + \epsilon_i = X_1^T\beta^* + \epsilon_1$$

β^*-intercept, Y_i - response/dependent variable, X_i –independent/explanatory variable

Least Square Estimator
$$\hat{\beta} = argmin\sum_{i=1}^{n}(Y_i - X_i^T\,\beta)^\wedge 2 = (X^TX)^{-1}X^TY$$
X- Design Matrix with rank p
$$\epsilon \sim Nu_n(0, \sigma^2 I_n)$$
$$\hat{\beta} \sim N_n(\beta^*, \sigma^2(X^TX)^{-1}$$
$$quadratic\ risk = E\left[\left\|\hat{\beta} - \beta\right\|_2^2\right] = \sigma^2 trace((X^TX)^{-1})$$
$$prediction\ error = \sigma^2(n-p)$$
$$\widehat{\sigma^2} = \frac{\left\|Y - X\,\hat{\beta}\right\|^2}{n-p} = \frac{\sum_{i=1}^{n}\widehat{\epsilon^2}}{n-p}$$

CLT, Convergence, Inequalities, Delta Method

Central Limit Theorem
$$Z = \frac{\sqrt(n)\,|X - \mu|}{\sigma}$$

Covariance Matrix
$$\Sigma = \begin{matrix} Cov(X,X) & Cov(X,Y) \\ Cov(Y,X) & Cov(Y,Y) \end{matrix}$$

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

Multivariate CLT
$$\sqrt{n}(\overline{X_n} - \mu) \xrightarrow{distribution} N(0,\Sigma)$$
$$\sqrt{n}\Sigma^{-\frac{1}{2}}(\overline{X_n} - \mu) \xrightarrow{distribution} N(0, I_d)$$

Markov Inequality
$$P(X \geq a) \leq \frac{E[X]}{a}$$

Chebyshev Inequality
$$P(|X - \mu| \geq c) \leq \frac{\sigma^2}{c^2}$$
Chernoff Bound
$$P(|\mu_n - \mu| \geq a) \leq e^{-nh(a)}$$

Hoeffding's Inequality
$$P(|\overline{X_n} - \mu| \geq \epsilon) \leq 2e^{\frac{2n\epsilon^2}{(b-a)^2}}$$

Jensen's Inequality

If f(x) is convex then
$$E[f(x)] \geq f(E[x])$$

Convergence

Almost Surely
$$P\left(w: T_{n(w)} \xrightarrow[n\to\infty]{} T(w)\right) = 1$$

In Probability
$$P(|T_n - T| \geq \epsilon) \xrightarrow[n\to\infty]{} 0$$

In Distribution
$$E[f(T_n)] \xrightarrow[n\to\infty]{} E[f(t)]$$

Pointwise Convergence
$$\lim_{n\to\infty} g_n(x) = g(x)$$

Uniform Convergence
$$sup|g_n(x) - g(x)| = 0$$

Delta Method

If estimator Z_n is asymptotically normal and the function g is differentiable everywhere, then
$$\sqrt{n}\big(g(Z_n) - g(\theta)\big) \xrightarrow{distribution} N(0, g'(\theta)^2\sigma^2)$$

Multivariate Delta Method
$$\sqrt{n}\big(g(T_n) - g(\theta)\big) \xrightarrow{distribution} N(0, \nabla g(\theta)^T\Sigma\nabla g(\theta))$$

Estimator properties

Asymptotic Normality

$$\sqrt{\{n\}}(\hat{\theta\_n} - \theta) \xrightarrow{distribution} \mathrm{N}(0, \sigma^2)$$

Bias

$$bias(\widehat{\theta_n}) = E[\widehat{\theta_n}] - \theta$$

Efficiency

The smaller the variance the more efficient is the estimator

Consistency

As $n \to \infty$ the estimator goes towards a single point

Quadratic Risk

$$R(\widehat{\theta_n}) = E\left[\left|\widehat{\theta_n} - \theta\right|^2\right] = Var(\widehat{\theta_n}) + bias(\widehat{\theta_n})^2$$

Total Variation

$$TV(P_\theta, P_{\theta'}) = \frac{1}{2}\sum |P_\theta(X) - P_{\theta'}(X)|$$

same for continuous but with the integral

Kullback-Leibler Divergence(KL)

$$KL(P,Q) = \sum P(X)\ln(\frac{p(X)}{q(X)})$$

Hypothesis Tests

Fisher Information

$$I(\theta) = -E[H(l(\theta))] = Var(l'(\theta)) = \frac{1}{Var(\theta)}$$

Wald's Test

$$\sqrt{n}(I(\theta_0))^{\frac{1}{2}}(\widehat{\theta_n^{MLE}} - \theta_0) \xrightarrow{d} \mathrm{N}_d(0, I_d)$$

Single Parameter

$$T_n = \frac{n(\hat{\theta}^{MLE} - \theta_0)^2}{Var(\hat{\theta})} \sim \chi_d^2$$

Multiple Parameters

$$n(\widehat{\theta_n^{MLE}} - \theta_0)^T I(\theta_0)(\widehat{\theta_n^{MLE}} - \theta_0) \xrightarrow{d} \chi_d^2$$

Likelihood Ratio Test

$T_n = 2(ln(\widehat{\theta_n^{MLE}}) - ln(\widehat{\theta_n^c}))$

$T_n \to \chi_{d-r}^2$, by **Wilk's Theorem**

Chi Squared Test

$$\boldsymbol{n}\sum_{j=1}^{k}\frac{(\widehat{\boldsymbol{p}_j} - \boldsymbol{p}_j^0)^2}{\boldsymbol{p}_j^0} \xrightarrow{d} \chi_{k-1}^2$$

Where k is the number of categories

Chi Squared Test for Distribution

$$n\sum_{j=0}^{k}\frac{\left(\frac{N_j}{n} - f_{\hat{\theta}}(j)\right)^2}{f_{\hat{\theta}}(j)} \xrightarrow{d} \chi_{k-d-1}^2$$

Where d is the number of parameters of the distribution

One Sided Two Sample Test

$$\frac{\{\overline{X_n} - \overline{Y_n} - (\mu_1 - \mu_2)\}}{\sqrt{\frac{\widehat{\sigma_X^2}}{n} + \frac{\widehat{\sigma_Y^2}}{m}}}$$

Welch-Satterthwaite Formula

Calculates the degrees of freedom for multiple sample test

$$N = \frac{\left(\frac{\widehat{\sigma_d^2}}{n} + \frac{\widehat{\sigma_c^2}}{m}\right)^2}{\frac{\widehat{\sigma_d^4}}{n^2(n-1)} + \frac{\widehat{\sigma_c^4}}{m^2(m-1)}} \geq \min(n, m)$$

Kolmogorov-Smirnov Test

$$T_n = sup\sqrt{n}|F_n(t) - F^0(t)| = \sqrt{n}\max\{\max\left(\left|\frac{i-1}{n} - F^0(X_i)\right|\right)\}$$

Cameron Von Mises

$$d^2(F_n, F) = \int [F_n(t) - F(t)]^2 dF(t)$$

Anderson-Darling

$$d^2(F_n, F) = \int \frac{[F_n(t) - F(t)]^2}{F(t)(1 - F(t))} dF(t)$$

Kolmogorov-Lilliefors test

$\sup |F_n(t) - \Phi_{\hat{\mu}, \widehat{\sigma^2}}(t)|$

Linear Regression Significance Test

$\frac{\widehat{\beta_j} - \beta_j}{\sqrt{\widehat{\sigma^2}\gamma_j}} \sim t_{n-p}$,

where γ_j is the jth diagonal coefficient of $(X^T X)^{-1}$

$$T_n^j = \frac{\widehat{\beta_j}}{\sqrt{\widehat{\sigma^2}\gamma_j}}$$

Bonferroni's Test

$$R_{s,\alpha} = \cup R_{j,\frac{\alpha}{k}}$$

So in order to test that a group of explanatory variables affect the response variable we have to make the level α much more stringent by dividing it by the number of explanatory variables

Required Sample Size

$$N = \frac{\left(\Phi^{-1}(1-\beta) + \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right)^2}{\frac{\tau^2}{\sigma^2}\gamma(1-\gamma)}$$

Where γ is the proportion of people assigned to treatment and τ is a chosen parameter

Confidence Intervals

Methods

Conservative bound

We use the biggest variance

Quadratic Equation

$$(\theta - \overline{R_n})^2 \leq \frac{q_{\alpha/2}^2\,\sigma^2}{n}$$

Plug-In

Just plug in estimated mean and asymptotic variance into Gaussian CI

CI Solve Method

$$|\hat{\theta} - \theta| \leq \frac{q_{\alpha/2}\sigma}{\sqrt{n}}$$

Gaussian CI

$$\widehat{\boldsymbol{\theta}_n} \mp \frac{\boldsymbol{Z_{value}\sigma}}{\sqrt{n}}$$

Exponential CI

$$\widehat{\boldsymbol{\theta}_n} \mp \frac{Z_{value}\lambda}{\sqrt{n}}$$

Power of a Test

$$\Pi_\psi = \inf(1 - \mathrm{B}_\psi(\theta))$$

Inference

	H_0 True	H_1 True
Reject	Type 1 Error (α) (False Positive)	
Fail to Reject		Type 2 Error (β) (False Negative)

Bayesian Statistics

Bayes Formula for Pdfs

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

Conjugate Prior happens when prior and posterior distributions are the same but with different parameters.

Improper prior on θ is a measurable nonnegative function $\Pi()$ defined on θ that is not integrable

Jeffrey's Prior

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

$$iff\ \eta = \phi(\theta)$$

$$I(\eta) = I(\phi^{-1}(\eta)) * \frac{d\theta}{d\eta}\ and\ \Pi_j(\eta) \propto \sqrt{\det(\hat{I}(\eta))}$$

Conditional Quantile/ Median

$$Median = \int_{-\infty}^{m(x)} h(y|x)dy = \frac{1}{2}$$

Bayes Confidence Region

We need to take the inverse of the posterior likelihood function with a chosen q_α

Distributions, Likelihoods

Exponential Family

$$f_\theta(y) = exp\left(\sum_{i=1}^{k} \eta_i(\theta)T_i(y) - B(\theta)\right)h(y)$$

Canonical Exponential Family

$$f_\theta(y) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + C(y,\phi)\right)$$

$$E[x] = b'(\theta)$$

$$Var(x) = \phi b''(\theta)$$

ϕ is called a **dispersion parameter**

Link Function

$If\ g\big(m(x)\big) = X^T\beta + \epsilon$

Then g is the link function

Canonical Link

$g(m) = (b')^{-1}(m)$

d-dimensional Gaussian distribution

$$\frac{\exp\left(-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu)\right)}{\sqrt{(2\Pi)^d \det(\Sigma)}}$$

First Order Statistic

$f_{y^1} = n\big(1 - F_x(y)\big)^{n-1} f_x(y)$

Nth Order Statistic

$f_{y^n} = n\big(F_x(y)\big)^{n-1} f_x(y)$

LOGIT

$\log\left(-\frac{m(x)}{1-m(x)}\right) = X^T\beta$

PROBIT

$\Phi^{-1}\big(m(x)\big) = X^T\beta$

Cauchy Distribution

$$\frac{1}{\Pi(1 + (X-\mu)^2)}$$

Laplace Distribution

$$\frac{1}{2}e^{-|X-\mu|}$$

Pascal Distribution

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

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

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

Huber's Loss

$$\begin{cases} \frac{a^2}{2}, if\ |a| \le \delta \\ \delta\left(|a| - \frac{\delta}{2}\right), otherwise \end{cases}$$

Kurtosis

$$E[\left(\frac{X - E[X]}{\sigma_d(X)}\right)^4 - 3]$$

It measures how spread out the tails are, if kurtosis is positive we know that the tails are heavier than Gaussian and vice versa

Likelihoods

Uniform([0,b])

$$\frac{1}{b^n}$$

Categorical Likelihood

$L_n(X_1, \dots, X_n : P_1, \dots, P_k) = P_1^{N_1} \dots P_k^{N_k}$

It's a generalised Bernoulli model

Gaussian

$$\frac{1}{\sigma^n (2\Pi)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2\right)$$

Bernoulli

$$p\hat{}(\sum_{i=1}^{n} X_i)(1-p)\hat{}(n - \sum_{i=1}^{n} X_i)$$

Poisson

$$\frac{e^{-n\lambda}\lambda\hat{}(\sum_{i=1}^{n} X_i)}{X_{1!} \dots X_{n!}}$$

Exponential

$\lambda^n e^{-\lambda \sum_{i=1}^{n} X_i}$

Theorems

Slutsky's Theorem

$$T_n + U_n \xrightarrow{distribution} T + U$$

Same for other operations

Continuous Mapping Theorem

If there is convergence in probability for T, then

$$f(T_n) \xrightarrow{distribution} f(T)$$

Cochran's Theorem

For $X\_1, \dots, X\_n$ random variables with $X_n \sim N(\mu, \sigma^2)$, if S_n is the sample variance, then $\frac{nS_n}{\sigma^2} \sim \chi^2_{n-1}$ and variance is $S_n = \frac{\sum_{i=1}^{n}((X_i - \bar{X})^2)}{n-1}$, multivariate variance is $S_n = \frac{\sum_{i=1}^{n} X_i^2}{n} - \bar{X}_n{}^2$

Glivenko-Cantelli Theorem

$$sup|F_n(t) - F(t)| \xrightarrow[n\to\infty]{a.s.} 0,$$

$$where\ F_n(t)\ is\ the\ empirical\ CDF$$

Cox Proportional Hazard Model

$$f(t) = 1 - \exp\left(-\int_0^t h(u)\, e^{\beta^T X} du\right)$$

Data Analysis

Belmont Principles

1. Respect for Persons
2. Beneficence
3. Justice

Histogram

Trade off in the binsize of histogram is between variance and bias (detail and noise)