

Algorithms, Evidence, and Data Science Cookbook

Part I: Classic Statistical Inference

- * **Population:** the entire group
- * **Sample:** a subset of the population
- * **Mean:** μ is the mean of the population; \bar{x} is the mean of the sample

$$\frac{1}{n} \sum_{i=1}^n x_i$$

- * **Variance:** the dispersion around the mean

Variance of a population: Variance of a sample:

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2 \quad s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

- * **Standard Deviation:** square root of the variance
- * **Standard Error:** an estimate of the standard deviation of the sampling distribution

For a mean:

$$se(\bar{x}) = \sqrt{\frac{s^2}{n}}$$

For the difference between two means:

$$se(\bar{x}_1, \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Algorithms and Inference

- * **Algorithm:** set of data probability-steps to produce an estimator
- * **Inference:** measuring the uncertainty around the estimator *e.g.*: \bar{x} the algorithm, while $se(\bar{x})$ is the inference

A Regression Example

Linear Regression

any regression is a conditional mean $\hat{Y}_i = E(Y_i | X_i)$

- * Y : response variable
- * X : covariate/predictor/feature
- * $\hat{\beta}_0, \hat{\beta}_1$: regression coefficients

$$\hat{\beta}_0 = \hat{Y} - \hat{\beta}_1 \hat{X} \quad \hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

$$se(\hat{\beta}_0) = \hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right] \quad se(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

- * predicted values = fitted curve given x :

$$\hat{Y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

- * residuals $\hat{\epsilon}$:

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 + \hat{\beta}_1 X_i$$

- * residual sum of squares RSS

$$RSS(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^n \hat{\epsilon}_i^2$$

- * mean square error $\hat{\sigma}^2$

$$\hat{\sigma}^2 = \frac{RSS(\hat{\beta}_0, \hat{\beta}_1)}{n - 2}$$

LOWESS & LOESS

- * 1) specify the number of points within the range/window n
- * 2) neighbour weightings $w(x_k)$

$$w(x_k) = \begin{cases} \left(1 - \left|\frac{x_i - x_k}{d}\right|^3\right)^3 & \text{if } d \text{ is the distance between } x_i \text{ and the } k^{th} \text{ neighbouring point} \\ 0 & \text{otherwise} \end{cases}$$

- * 3) for each range, estimate a regression function

LOWESS: $\hat{y}_k = a + bx_k$

LOESS: $\hat{y}_k = a + bx_k + cx_k^2$

- * 4) robust weightings $G(x_k)$

$$G(x_k) = \begin{cases} \left(1 - \left(\frac{|y_i - \hat{y}_i|}{\delta \text{median}(|y_i - \hat{y}_i|)}\right)^2\right)^2 & \left| \frac{|y_i - \hat{y}_i|}{\delta \text{median}(|y_i - \hat{y}_i|)} \right| < 1 \\ 0 & \left| \frac{|y_i - \hat{y}_i|}{\delta \text{median}(|y_i - \hat{y}_i|)} \right| \geq 1 \end{cases}$$

LOWESS: $\hat{y}_k = \sum_k w(x_k) G(x_k) (a + bx_k)^2$

LOESS: $\hat{y}_k = \sum_k w(x_k) G(x_k) (a + bx_k + cx_k^2)^2$

- * 5) A series of new smoothed values is the result. The procedure can be repeated to get a more precise curve fitting.

Bootstrapping

- * bootstrap principle:

$\sigma_{\text{(sampling w/replacement)}} = \sigma_{\text{(across samples)}}$

- * bootstrap iterations: B

* original sample: $(x_i, y_i)_{i=1}^N$

* bootstrap samples: $(x_{j(b)}, y_{j(b)})_{j \in I}$ for $b = 1, \dots, B$, $I = \{1, \dots, N\}$, and j is the index that is randomly sampled from I

- * for each b , compute $y_{j(b)}$ using LOWESS or any other model

	b	1	2	3	...	B
j						
1		$y_{1(1)}$	$y_{1(2)}$	$y_{1(3)}$...	$y_{1(B)}$
2		$y_{2(1)}$	$y_{2(2)}$	$y_{2(3)}$...	$y_{2(B)}$
...	
N		$y_{N(1)}$	$y_{N(2)}$	$y_{N(3)}$...	$y_{N(B)}$

- * for each j row, the standard deviation σ_j^{boot} is

$$\sigma_j^{boot} = \sqrt{\frac{(\bar{y}_j - \bar{\bar{y}}_j)^2}{B - 1}}$$

- * sort $i(\bar{b})$ by value from min to max \rightarrow get the 5th and 95th values to get a 90% confidence interval

Hypothesis Testing

T-test, one-sample

- * null hypothesis $H_o : \mu = \mu_0$
- * alternative hypothesis $H_a : \mu \{=, > \text{ or } <\} \mu_0$
- * t - *statistic* standardizes the difference between \bar{x} and μ_0

$$t = \frac{\bar{x} - \mu_0}{se(\bar{x})}$$

degrees of freedom $df = n - 1$

- * p - *value*: probability that \bar{x} was obtained by chance given $\mu_0 = \mu$.

* **algorithm:** read the t-distribution critical values (chart) for the p - *value* using t and df

if $(p - \text{value} < \alpha)$ { reject H_o and accept H_a }

else { can't reject H_o }

- * α is the predetermined value of significance (usually 0.05)

- * if (t is of the 'wrong' sign) $p - \text{value} = 1 - p - \text{value}_{chart}$

paired two-sample t-test

each value of one group corresponds to a value in the other group

- * **algorithm:** subtract the values for each sample to get one set of values and use μ_0 to perform a one-sample t-test

unpaired two-sample t-test

the two populations are independent

- * $H_o : \mu_1 = \mu_2$
- * $H_a : \mu_1 \{=, > \text{ or } <\} \mu_2$
- * t - *statistic*

$$t = \frac{\bar{x}_1 - \bar{x}_2}{se(\bar{x}_1, \bar{x}_2)}$$

degrees of freedom $df = (n_1 - 1) + (n_2 - 1)$

- * **algorithm:** same as in one-sample t-test

- * double the p - *value* for $H_a : \mu_1 \neq \mu_2$

- * **Type I error** α : probability of rejecting a true H_o

- * **Type II error** β : probability of failing to reject a false H_o

Notes

- * the OLS confidence intervals work asymptotically \rightarrow they assume the number of available observations is infinite, but it assumes normality

- * in LOWESS, n is not infinite, but it does not assume any distribution

Frequentist Inference

- * assumes the observed data comes from a probability distribution F
- * $x = (x_1, \dots, x_n)$ is the data vector (aka. *the sample's values*)
- * $X = (X_1, \dots, X_n)$ is the vector of random variables (aka. *a sample, individual draws of F*)
- * the expectation property $\theta = E_F(X_i)$ (aka. the true expectation value of any draw X_i)
- * $\hat{\theta}$ is the best estimate of θ

usually,

$$\hat{\theta} = t(x) \quad t(x) = \bar{x}$$

where $t(x)$ is the algorithm

- * $\hat{\theta}$ is sample specific, is a realization of $\hat{\Theta} = t(x)$. Typically,

$$E_F(\hat{\Theta}) = \mu \quad \mu \text{ is the expected value of producing an estimate using } t(x) \text{ when } x \text{ comes from } F$$

- * **Bias-Variance Trade-Off:** models with lower bias will have higher variance and vice versa.
- * **Bias:** error from incorrect assumptions to make target function easier to learn (high bias \rightarrow missing relevant relations or under-fitting)
- * **Variance:** error from sensitivity to fluctuations in the dataset, or how much the target estimate would differ if different training data was used (high variance \rightarrow modelling noise or over-fitting)

$$bias = \mu - \theta$$

$$(aka. \text{ expected } - \text{ true values}) \quad var = E_F\{(\hat{\Theta} - \mu)^2\}$$

Frequentist principles

- * usually defines parameters with infinite sequence of trials \rightarrow hypothetical data sets $X^{(1)}, X^{(2)}, \dots$ generate infinite samples $\hat{\Theta}^{(1)}, \hat{\Theta}^{(2)}, \dots$
- * 1) Plug-in principle: relate the sample $se(\bar{x})$ with the true variance.

$$var_F(x) = var_F = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

$$se(\bar{x}) = \left[\frac{var_F(x)}{n} \right]^{\frac{1}{2}}$$

- * 2) Taylor series approximations: relate $t(x)$ by local linear approximations (aka. compute $\bar{se}(x)$ of the transformed estimator)

$$se(\hat{\theta}) = se(\bar{x}) \frac{d\hat{\theta}}{d\bar{x}} = se(\bar{x}) \frac{dt(x)}{d\bar{x}}$$

- * 3.1) Parametric Families: given $x = (x_1, \dots, x_n)$, the Likelihood Function $L(x)$ (aka. the probability to observe x) is given by:
e.g. $\hat{\theta} = \mu$ for a normal distribution

$$P(x|N(\mu, \sigma^2)) = P(x_1|N(\mu, \sigma^2)) \dots P(x_n|N(\mu, \sigma^2))$$

$$P(x|N(\mu, \sigma^2)) = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \prod_{i=1}^n e^{-\frac{(x_i - \mu)^2}{2\sigma^2}} = L(x)$$

$$L(x) = \prod_{i=1}^n f_{\theta}(x_i)$$

- where f_{θ} is the density function
- * 3.2) MLE (maximum likelihood estimate): find $\hat{\theta}$ such that $L(x)$ is maximized
e.g.

$$\max_{\hat{\theta}} L(x) \Rightarrow \max_{\hat{\mu}} L(x) = \hat{\mu}^{MLE}$$

- * 4) Simulation and Bootstrap: estimate F as \hat{F} , then simulate values from \hat{F} to get a prior sample $\hat{\Theta}^{(k)} = t(x^{(b)})$
The empirical standard deviation of the $\hat{\Theta}'s$ is the frequentist estimate for $se(\hat{\theta})$
- * 5) Pivotal Statistics: Frequentist use pivotal statistics whenever they are available to conduct stat. tests
e.g. t-test is a pivotal statistic as it does not depend on parameters the distribution might have.

Frequentist Optimality

- Neyman-Pearson lemma optimum hypothesis-testing algorithm:
purpose: choose one of the two possible density functions for observed data x
- * null hypothesis density $f_0(x)$
- * alternative density $f_1(x)$
- let $L(x)$ be the Likelihood Ratio

$$L(X) = \frac{f_1(X)}{f_0(X)}$$

let the testing rule $t_c x$ be:

$$t_c x = \begin{cases} 1(pic f_1(x)), & \ln(L(X)) \geq c \\ 0(pic f_0(x)), & \ln(L(X)) < c \end{cases}$$

- * only rules in the $t_c x$ form can be optimal *problem Steps*
- * 1) define the density functions $f_0(x_i)$ and $f_1(x_i)$ for $f_0(x)$ and $f_1(x)$
e.g.

$$f_0 \sim N(\mu_0, \sigma_0^2)$$

$$f_1 \sim N(\mu_1, \sigma_1^2)$$

$$f_0 \sim N(0, 1)$$

$$f_1 \sim N(0.5, 1)$$

$$f_0(x_i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x_i^2}{2}}$$

$$f_1(x_i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x_i - 0.5^2}{2}}$$

- * 2) calculate the likelihood functions $f_0(X)$ and $f_1(X)$
e.g.

$$f_0(X) = \left[\frac{1}{\sqrt{2\pi}} \right]^n e^{-\frac{1}{2} \sum_{i=1}^n x_i^2}$$

$$f_1(X) = \left[\frac{1}{\sqrt{2\pi}} \right]^n e^{-\frac{1}{2} \sum_{i=1}^n ((x_i - 0.5)^2)}$$

- * 3) calculate the likelihood ratio
e.g.

$$L(X) = \frac{e^{-\frac{1}{2} \sum_{i=1}^n ((x_i - 0.5)^2)}}{e^{-\frac{1}{2} \sum_{i=1}^n x_i^2}}$$

$$L(X) = e^{-\frac{1}{2} [n\bar{x} - \frac{n}{4}]}$$

- * 4) remove all independent variables e.g.

$$L(X) > c \Rightarrow \begin{aligned} e^{-\frac{1}{2} [n\bar{x} - \frac{n}{4}]} &> c_1 \\ -\frac{1}{2} \left[n\bar{x} - \frac{n}{4} \right] &> C_2 \\ n\bar{x} - \frac{n}{4} &> c_3 \\ \bar{x} &> c_4 \\ \bar{x} &> c \end{aligned}$$

- only the mean depends on the sample x
- * 5) the *most powerful* hypothesis test at any type I error rate α is to compare c to a constant.
e.g.

$$\begin{aligned} \alpha &= P(\bar{x} > c | \mu = \mu_0) \\ \alpha &= P((\bar{x} - \mu)\sqrt{n} > (c - \mu)\sqrt{n} | \mu = 0) \\ \alpha &= 1 - P(\bar{x}\sqrt{n} \leq c\sqrt{n} | \mu = 0) \\ \alpha &= 1 - \Phi(c\sqrt{n}) \end{aligned}$$

- Φ is the cumulative density function (CDF) of a normal distribution $N(\mu_0, \sigma_0^2)$
- * 6) calculate c
e.g.

In general:

$$\begin{aligned} \Phi(c\sqrt{n}) &= 1 - \alpha \\ c\sqrt{n} &= \Phi^{-1}(1 - \alpha) \\ c &= 0 + \frac{1}{\sqrt{n}} \Phi^{-1}(1 - \alpha) \quad c = \mu_0 + \frac{1}{\sqrt{n}} \Phi^{-1}(1 - \alpha) \end{aligned}$$

- * 7) calculate β , such that it's minimized
e.g.

$$\begin{aligned} \beta &= P(\bar{x} \leq c | \mu = \mu_1) \\ \beta &= P((\bar{x} - \mu)\sqrt{n} \leq (c - \mu)\sqrt{n} | \mu = 0.5) \\ \beta &= \Phi((c - 0.5)\sqrt{n}) \end{aligned}$$

Notes and Details

- * $1 - \beta$ is the power of the hypothesis test (probability of correctly rejecting $f_0(x)$)

Bayesian Inference

Bayes Rule

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- * Bayes Rule (for one μ) can be written as:

$$g(\mu|x) = c_x L_x(\mu) \Pi(\mu)$$

where:
 μ : an unobserved point in the parameter space Ω
 x : a point in the sample space X
 c_x : normalization constant of the posterior distribution
 $g(\mu|x)$: posterior distribution
 $L_x(\mu)$: likelihood function
 $\Pi(\mu)$: prior distribution

- * Bayes Rule (for two μ_1, μ_2) can be written as:

$$\frac{g(\mu_1|x)}{g(\mu_2|x)} = \frac{g(\mu_1)}{g(\mu_2)} \frac{L_x(\mu_1)}{L_x(\mu_2)}$$

The posterior odds ratio is the prior odds ratio times the likelihood ratio

$$L_x(\mu) = \prod_{i=1}^n e^{-\frac{1}{2}(x_i-\mu)^2}$$

Warm-up example

e.g. Find the probability of identical twins. The doctor says that $\frac{1}{3}$ of twin births are identical. A sonogram observed same sex. identical twins are of the same sex, while fraternal have 0.5 probability to be of the same sex.

$$\frac{g(\text{identical}|\text{sameSex})}{g(\text{fraternal}|\text{sameSex})} = \frac{g(\text{identical})}{g(\text{fraternal})} \times \frac{L_{\text{identical}}(\text{sameSex})}{L_{\text{fraternal}}(\text{sameSex})}$$

$$\frac{g(\text{identical}|\text{sameSex})}{g(\text{fraternal}|\text{sameSex})} = \frac{\frac{1}{3}}{1-\frac{1}{3}} \times \frac{1}{2}$$

Flaws in Frequentist Inference

* In Frequentist, if the algorithm changes (even if the data points stay exactly the same), the significance level is different for each algorithm.
 * On Bayesian inference, the algorithm stays the same → the significance level does not change.

A Bayesian/Frequentist Comparison List

Bayesian:	Frequentist:
	* attention is in choosing a prior Π
* attention is in choosing an algorithm $t(x)$	* operates with one parameter (specific question) in many samples
* operates only in one sample with the whole parameter space	* only computes the expected value and the variance (each answer requires an specific algorithm)
* answers all posible questions at once, since the posterior is a as we can come up with many distribution	* is more flexible than Bayes algorithms

Bayesian Reasoning - estimate μ from x if

$\mu \sim N(m, A)$

normal likelihood function (assume a variance of 1):

$$x|\mu \sim N(\mu, 1)$$

the normal posterior:

$$\mu|x \sim N(m + B(x - m), B)$$

where $B = \frac{A=\text{prior variance}}{A+1=\text{total variance}}$, m = prior parameter therefore:

$$\hat{\mu}_{Bayes} = m + B(x - m)$$

Notes and Details

* like in frequentist, the fundamental unit of inference is a family of probability densities.
 * Bayesian inferences assumes the knowledge of a prior density $g(\mu)$, $\mu \in \Omega$

Fisherian Inference and Maximum Likelihood Estimation

* The log-likelihood function is defined as:

$$\ell_x(\theta) = \text{Log}\{f_{\theta}(x)\}$$

$\ell_x(\theta)$: gets the most likely parameters to get the sample x
 $f_{\theta}(x)$: likelihood function (aka. family probability densities)
 θ : vector of parameters

for a fixed x and a variable θ

* The *MLE* is the value of $\theta \in \Omega$ that maximizes $\ell_x(\theta)$

$$MLE : \hat{\theta} = \underset{\theta \in \Omega}{\operatorname{argmax}} \{ \ell_x(\theta) \}$$

* Estimate functions of the true parameter: $\hat{\gamma} = T(\hat{\theta})$

* Good frequentist properties (good bias & variance):

$$\text{bias} = \mu - E(\hat{\mu})$$

μ : true value of the parameter
 $E(\hat{\mu})$: expected value of the estimate

$$\text{variance} = \sum_{i=1}^I (\hat{\mu}^{(i)} - E(\hat{\mu}))^2$$

$\text{variance} = E_F\{(\hat{\mu}^{(i)} - E(\hat{\mu}))^2\}$

* Reasonable Bayesian justification

$P(\theta|x)$: posterior
 c_x : constant
 $\Pi(\theta)$: prior
 $e^{\ell_x(\theta)}$: maximum likelihood estimation

* Fisherian inference assumes a flat prior (aka. unknown prior), so that the MLE $\hat{\theta}^{MLE}$ is a maximizer of $P(\theta|x)$. (The MLE is the highest point of the posterior distribution)
 * As the algorithm does not change, the significance level is not affected by unexpected changes in the algorithm.

e.g. - for a Normal density function

* let $\theta = (\mu, \sigma^2)$

* density function $f_{\theta} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}$ * Since:

$L(x) = \prod_{i=1}^n f_{\theta}(x_i)$ Log-Likelihood function

$$\ell_x(\theta) = \sum_{i=1}^n \text{Log}\{f_{\theta}(x_i)\} = \sum_{i=1}^I \ell_x(\theta)$$

$$\mu^{\hat{MLE}} = \bar{x}$$

$$\sigma^{MLE} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}}$$

* MLE can cause over-fitting identification problems when we fit a lot of parameters in θ (it would become very specific to our sample → may not represent the population)

Fisher Information and the MLE

Log-Likelihood Function

$$\ell_x(\theta) = \text{Log}f_{\theta}(x)$$

Score Function

how higher or lower is the likelihood function value of the sample as θ varies?

$$\dot{\ell}_x(\theta) = \frac{\dot{f}_{\theta}(x)}{f_{\theta}(x)}$$

Expectation of $\dot{\ell}_x(\theta)$

$$E(x) = \int_x x f(x) dx$$

$f(x)$: density function

$$E[\dot{\ell}_x(\theta)] = 0$$

Variance of $\dot{\ell}_x(\theta)$

$$V[x] = \int_x [x - E(x)]^2 f(x) dx$$

$$V[\dot{\ell}_x(\theta)] = \int_x [\dot{\ell}_x(\theta)]^2 f_{\theta}(x) dx$$

Fisher Information I_0

$$I_0 = V[\dot{\ell}_x(\theta)]$$

$$\ddot{\ell}_x(\theta) = \frac{\ddot{f}_{\theta}(x)}{f_{\theta}(x)} - \left(\frac{\dot{f}_{\theta}(x)}{f_{\theta}(x)} \right)^2$$

MLE estimator of θ : $\hat{\theta}^{MLE}$

$$E(\ddot{\ell}_x(\theta)) = -I_0$$

$$\hat{\theta}^{MLE} \sim N\left(\theta, \frac{1}{I_0}\right)$$

e.g. for a normal dist.

let $x_i \sim N(\theta, \sigma^2)$

* 1) compute $\ell_x(\theta)$

$$\text{density function } f_{\theta}(x) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$\text{likelihood function } \ell_x(\theta) = -\frac{1}{2} \sum_{i=1}^n \frac{(x_i-\theta)^2}{\sigma^2} - \frac{n}{2} \text{Log}(2\pi\sigma^2)$$

* 2) score function $\dot{\ell}_x(\theta) = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \theta)$

$$\ddot{\ell}_x(\theta) = -\frac{n}{\sigma^2}$$

* 3) compute I_0

as $E(\ddot{\ell}_x(\theta)) = -I_0$, Fisher Information $I_0 = \frac{n}{\sigma^2}$

* 4) compute $\hat{\theta}^{MLE}$

$$E(\dot{\ell}_x(\theta)) = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \theta) = 0, \text{ such that}$$

$$\sum_{i=1}^n x_i = n\theta \Rightarrow \hat{\theta}^{MLE} = \frac{\sum_{i=1}^n x_i}{n} = \bar{x}$$

* 5) compute $se(\hat{\theta}^{MLE})$

for a large n,

$$\hat{\theta}^{MLE} \sim N\left(\theta, \frac{1}{I_0}\right) \Rightarrow \hat{\theta}^{MLE} \sim N\left(\theta, \frac{\sigma^2}{n}\right)$$

$$se(\hat{\theta}^{MLE}) = \frac{1}{I_0} = \frac{\sigma^2}{n}$$

* 6) $se(\hat{\theta}^{MLE}) = \frac{1}{nI_0}$, by Cramer-Rao lower bound.

The MLE has variance at least as small as the best unbiased estimate of θ

Conditional Inference

e.g. An iid sample $x \sim N(\theta, 0)$ has produced estimate $\hat{\theta} = \bar{x}$. however,

$n = 25$ was declined
$$n = \begin{cases} 25, & \text{prob } \frac{1}{2} \\ 100, & \text{prob } \frac{1}{2} \end{cases}$$

* Classical Frequentist rational:

$$sd(\bar{x}) = \sigma_{\bar{x}} = \sqrt{\frac{1}{2} \frac{\sigma^2}{100} + \frac{1}{2} \frac{\sigma^2}{25}} = 0.158$$

* Conditional Inference rational:

$$sd(\bar{x}) = \sqrt{\frac{\sigma^2}{25}} = 0.2$$

* use the likelihood function (based on observation) without the prior

* “just take the sample you have”

1) more relevant inferences (w/what really happened)

2) simpler inferences (no correlation between the result and the sample size selection)

e.g. Observed Fisher Information $I_{(x)}$

$$I_{(x)} = -\ddot{\ell}_x(\hat{\theta}^{MLE})$$

In large samples $I_{(x)} = I_0$. Use $I_{(x)}$ in small samples

$$E[I_{(x)}] = nI_0$$

* 1) compute the log-likelihood

$$f_{\theta}(x) = \frac{1}{\pi} \frac{1}{1 + (x + \theta)^2} \Rightarrow \text{Cauchy density function}$$

$$\ell_x(\theta) = \text{Log}\left(\frac{1}{\pi}\right) + \text{Log}(1) - \text{Log}(1 + (x + \theta)^2)$$

* 2) get its derivative

$$\dot{\ell}_x(\theta) = \frac{2(x - \theta)}{1 + (x + \theta)^2}$$

* 3) get the 2nd derivative

$$\ddot{\ell}_x(\theta) = \frac{-2(1 + (x - \theta)^2) + 4(x - \theta)^2}{(1 + (x - \theta)^2)^2}$$

* 4) get the observed fisher information

$$I_{(x)} = -\ddot{\ell}_x(\hat{\theta}^{MLE})$$

* 5) get the variance of the estimate, even if the distribution does not have a defined variance or expected value

- for 10000 samples of size n with $\theta = 0$, compute $1/I_{(x)}$ and $\hat{\theta}^{MLE}$

- group the 10000 $\hat{\theta}^{MLE}$ values according to quantiles of $1/I_{(x)}$ and calculate the empirical variance for each sample.

* for all samples, the unconditional variance $1/nI_0$ is the same because all the samples are of the same size.

* on the other hand, $I_{(x)}$ will vary from sample to sample ($\hat{\theta}^{MLE}$ is different for each sample). * $I_{(x)}$ is related to the variance.

Permutation and Randomization

* when performing a t-test, it's assumed that the data samples come from a normal distribution.

* small samples may follow a different distribution.

Randomization removes the normality assumption

* Randomization is: taking random groups from the data that are of the same size as the tested groups.

* 1) compute the t-statistic for each randomly sampled pair of groups

* 2) get the t-statistic histogram

Utilizing random generated groups, it's expected the t-values not to be very high \rightarrow construct an empirical distribution of t-values

Parametric Models and Exponential Families

Univariate Families

Name Notation	Density	X	Ω	E Var
Normal $N(\mu, \sigma^2)$	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2}}$	$\mathbb{R}^{(1)}$	$\mu \in \mathbb{R}^{(1)}$ $\sigma^2 \in \mathbb{R}^+$	μ σ^2

* has two parameters, but they are very specific. μ is the location parameter, and σ^2 is the wide/narrow parameter

* model quantities that take positive and/or negative continuous values, if the distribution is symmetric and if there are no too many extreme values

Name Notation	Density	X	Ω	E Var
Poisson $Poi(\lambda)$	$\frac{e^{-\lambda} \lambda^x}{x!}$	\mathbb{N}_0	$\lambda \in \mathbb{R}^+$	λ λ

* if the mean grows/shrinks the variance also grows/shrinks proportionally

* λ must stay positive and is the interval of time of an exponential distribution, which is continuous \rightarrow the expected number of successes can have decimals

* model a quantity that is discrete, it's the number of counts of something

* It's not very flexible as only has one parameter to tweak

Name Notation	Density	X	Ω	E Var
Binomial $Bi(n, \theta)$	$\binom{n}{x} \theta^n (1 - \theta)^{n-x}$	$\{0, \dots, n\}$	$0 \leq \theta \leq 1$	$n\theta$ $n\theta(1 - \theta)$

* model the count of successes as Poisson, but we know the number of trials n

Name Notation	Density	X	Ω	E Var
Gamma $Ga(\nu, \sigma)$	$\frac{x^{\nu-1} e^{-\frac{x}{\sigma}}}{\sigma^{\nu} \Gamma(\nu)}$	\mathbb{R}^+	$\nu > 0$ $\sigma > 0$	$\sigma\nu$ $\sigma^2\nu$

* the Gamma is used to model positive quantities. its common to use the inverse Gamma to model variances.

Name Notation	Density	X	Ω	E Var
Beta $Be(\alpha, \beta)$	$\frac{x^{\alpha-1} (1-x)^{\beta-1}}{B(\alpha, \beta)}$	$0 \leq x \leq 1$	$\alpha > 0$ $\beta > 0$	$\frac{\alpha}{\alpha+\beta}$ var

$$var = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

* as x goes from 0 to 1, it's mostly used to talk about probabilities (aka. probability distribution)

* both the Gamma and Beta have two parameters that convey some degree of flexibility

* Gamma is flexible but not as flexible as Beta

* The Binomial can approximate a Poisson with a large n and small probability.

Multinomial Distribution (a.k.a. multidimensional binomial)

Used when observations take a finite number of possible outcome values L .

* let $\underline{x} = (x_1, \dots, x_L)$ be the vector of counts given the possible outcomes, where x_l is the number of cases/counts having outcome l . e.g. $\underline{x} = (150, 300, 1000, 50)$, where outcome $l = 1$ happened 150 times, and outcome $l = 4$ happened 50 times

* code the outcomes in terms of unit vectors of length L . e.g.

$e_l = (0, \dots, 0, 1, 0, \dots, 0)^T$, where the 1 is in the l^{th} place.

* encode the outcomes as unit vectors with assigned probabilities in π_l , a vector of probabilities.

$$\pi_l = P\{e_l\}, l = 1, 2, 3, \dots, L$$

\underline{x} follows a multinomial distribution f_{π}

$$f_{\pi}(\underline{x}) = \underline{x} \sim Mult_L(n, \pi) = \frac{n!}{x_1! x_2! \dots x_L!} \cdot \prod_{l=1}^L \pi_l^{x_l}$$

where L is the no. of outcomes, n the no. of observations, and π is the prob. vector.

* The multinomial distribution assumes the probabilities are constant.

The parameter space Ω of π is S_L ; a set of probability vectors π such that the components of π are positive quantities for all l 's

$$S_L = \{\pi : \pi_l \geq 0 \forall l \text{ and } \sum_{l=1}^L \pi_l = 1\}$$

The sample space X for \underline{x} is a subset of nS_L with integer components. e.g. for $L = 2$ (a Binomial dist.), $(\pi_1, \pi_2) = (\pi, 1 - \pi)$; $(x_1, x_2) = (x, n - x)$

The mean vector $E(x) = n\pi$

The covariance matrix Σ is given by:

$$\Sigma = n \cdot \left(\begin{bmatrix} \pi_1 & & & \text{zeros} \\ & \pi_2 & & \\ & & \ddots & \\ \text{zeros} & & & \pi_L \end{bmatrix} - \pi \cdot \pi^T \right)$$

The variance of x_l is:

$$V(x_l) = n \cdot \pi_l \cdot (1 - \pi_l)$$

The covariance of x_l is:

$$\text{Cov}(x_l, x_j) = -n\pi_l \cdot \pi_j$$

Multinomial-Poisson relationship

IF {
 S_1, S_2, \dots, S_L are independent Poisson distributions/counts;
 meaning that the counts of each category follow the
 distribution: $S_l \overset{iid}{\sim} \text{Poi}(\mu_l), l = 1, 2, \dots, L$.
 Each Poisson has a different μ_l parameter, which is a vector of
 mean/rate parameters.

}
 THEN {
 the vector of successes is given by:

$$\mathbb{S} | \sum_{l=1}^L S_l \sim \text{Mult}_L \left(\sum_{l=1}^L S_L, \frac{\underline{\mu}}{\sum_{l=1}^L \mu_l} \right)$$

}

IF {
 the number of trials N is distributed Poisson with parameter n

$$N \sim \text{Poi}(n)$$

}

THEN {

$$\text{Mult}_L(N, \underline{\pi}) \sim \text{Poi}(n \cdot \underline{\pi})$$

}

where $\underline{\pi}$ is the probability vector, and $n \cdot \underline{\pi}$ is a vector of
 expected values (means).

For a large n , the approximation

$$\underline{x} \overset{a}{\sim} \text{Poi}(n \cdot \underline{\pi})$$

removes the need to compute multinomial correlations

* the multinomial distribution contains all distributions on
 sample space X composed of \underline{L} discrete categories \rightarrow the
 multinomial dist. can model any distribution

Exponential Families - Poisson Dist.

$$f_{\mu}(x) = \frac{\mu^x e^{-\mu}}{x!}$$

From the ratio of two Poissons $\frac{f_{\mu}(x)}{f_{\mu_o}(x)}$,

$$f_{\mu}(x) = e^{-(\mu - \mu_o)} \cdot \left(\frac{\mu}{\mu_o} \right)^x \cdot f_{\mu_o}(x)$$

given: $\alpha = \log(\frac{\mu}{\mu_o})$, then: $\left(\frac{\mu}{\mu_o} \right)^x = e^{\alpha x}$ and $\mu = e^{\alpha} \mu_o$
 therefore:

$$f_{\mu}(x) = e^{\alpha x} - \Psi(\alpha) \cdot f_{\mu_o}(x)$$

$$\Psi(\alpha) = \mu_o(e^{\alpha-1})$$

Exponential Families - Gamma Dist.

$$f_{\underline{\alpha}}(x) = \frac{x^{\nu-1} \cdot e^{-\frac{x}{\sigma}}}{\sigma^{\nu} \Gamma(\nu)}$$

$\underline{\alpha} = (\alpha_1, \alpha_2) = (-\frac{1}{\sigma}, \nu) \in A \subseteq \{\alpha_1 < 0; \alpha_2 > 0\}$
 $\underline{y} = (y_1, y_2) = (\gamma, \log(x))$
 $\Psi(\underline{\alpha}) = \alpha_2 \log(-\alpha_1) + \log(\Gamma(\alpha_2))$

IF {
 $\underline{x} = (x_1, \dots, x_n)$ is iid from $f_{\mu}(x) = e^{\alpha^T \underline{y} - \Psi(\underline{\alpha})} \cdot f_{\mu_o}(x)$ and
 $\underline{y}_i = t(x_i)$
 }
 THEN {

$$f_{\underline{\alpha}}(\underline{x}) = e^{n(\underline{\alpha}^T \bar{\underline{y}} - \Psi(\underline{\alpha}))} \cdot f_o(\underline{x})$$

with: $\bar{\underline{y}} = \sum_{i=1}^n \frac{\underline{y}_i}{n}$
 }

$\Psi(\alpha)$ can be computed numerically by doing:

$$\Psi(\alpha) = \log \int_{\text{sample space } X} e^{\alpha y} f_o(x) dx$$

where $f_o(x)$ is the pdf in question

James-Stein Estimator vs Bayes vs MLE

Estimate μ from x if $\mu \sim N(m, A)$

$$\hat{\underline{\mu}}^{Bayes} = \underline{M} + B(\underline{x} - \underline{M})$$

where m = prior parameter, $B = \frac{A}{A+1} = \frac{\text{prior variance}}{\text{total variance}}$,
 $\underline{M} = [m, m, \dots, m]$

$$\hat{\underline{\mu}}^{MLE} = \underline{x}$$

$$\hat{\underline{\mu}}^{JS} = \hat{\underline{M}} + \hat{B}(\underline{x} - \hat{\underline{M}})$$

where $\hat{\underline{M}} = \bar{x}$, $\hat{\underline{M}} = [\hat{M}, \hat{M}, \dots, \hat{M}]$, $\hat{B} = \frac{1-N-3}{\sum_{i=1}^N (x_i - \bar{x})^2}$
 (for N points)

Expected Squared Error

$$E\{\|\hat{\underline{\mu}}^{Bayes} - \underline{\mu}\|^2\} = NB$$

$$E\{\|\hat{\underline{\mu}}^{MLE} - \underline{\mu}\|^2\} = N$$

$$E\{\|\hat{\underline{\mu}}^{JS} - \underline{\mu}\|^2\} = NB + 3(1 - B)$$

$\hat{\underline{\mu}}^{JS}$ has a bigger ESE than $\hat{\underline{\mu}}^{Bayes}$ as \underline{M} and B are estimated,
 but still better than $\hat{\underline{\mu}}^{MLE}$ if $N \leq 4$ observations.

James-Stein Theorem

IF $x_i | \mu_i \sim N(\mu_i, 1)$ for $i = 1, 2, \dots, N$ with $N \geq 4$; THEN

$$E\{\|\hat{\underline{\mu}}^{JS} - \underline{\mu}\|^2\} < E\{\|\hat{\underline{\mu}}^{MLE} - \underline{\mu}\|^2\}$$

for all choices of $\underline{\mu} \in \mathbb{R}^N$ (not Bayesian reasoning any more)

* JS gets observations from Normal distributions which have
 different means for each observation (estimate different means
 for each observation). * JS is a shrinks the effects of individual
 observations towards the common mean * extreme shrinkage is
 to say each observation is the average of all observations * void
 shrinkage is to say each observation is its own average (as in
 MLE) * JS is in between.

Ridge Regression vs Linear Regression

Linear Regression

based on MLE, it assumes a n-dimensional vector
 $\underline{y} = (y_1, \dots, y_n)^T$ from a linear model $\underline{y} = x\beta + \underline{\epsilon}$
 where:

β a unknown p-dimensional parameter vector
 $\underline{\epsilon}$ uncertain values (aka. independent random variables or
 independent draws from a dist.)
 x known data points
 \underline{y} outcomes taken from the linear model

Thus: $\epsilon \sim (0, \sigma^2 I_n)$ where:

mean = 0

the variance I_n is the identity matrix of size n

$$\hat{\beta} = \underset{\beta}{\text{argmin}} \{\|\epsilon\|^2\} = \underset{\beta}{\text{argmin}} \{\|\underline{y} - x\beta\|^2\}$$

differentiating:

$$\hat{\beta}^{OLS} = S^{-1} \cdot x^T y$$

where: $S = x^T x$

standard error:

$$\hat{\beta}^{OLS} \sim (\beta, \sigma^2 \cdot S^{-1})$$

Ridge Regression

$$\hat{\beta} = \underset{\beta}{\text{argmin}} \{\|\underline{y} - x \cdot \beta\|^2 + \lambda \cdot \|\beta\|^2\}$$

where: $\|\beta\|^2 = \beta_1^2 + \beta_2^2 + \dots + \beta_p^2$

if β coefficients are small, the better the results \rightarrow as the
 variance decreases by introducing some bias. λ is how much
 the sum of squares is penalized.

differentiating:

$$\hat{\beta}^{Ridge} = (S + \lambda \cdot I_n)^{-1} \cdot x^T y$$

standard error:

$$\hat{\beta}^{Ridge} \sim ((S + \lambda \cdot I_n)^{-1} \cdot S \cdot \beta, \sigma^2 \cdot (S + \lambda \cdot I_n)^{-1} \cdot S \cdot (S + \lambda \cdot I_n)^{-1})$$

Ridge is a regularized regression, meaning that the variables need to be rescaled as the coefficients have to be on the same scale.

OLS is a special case of Ridge where $\lambda = 0$

Logistic Regression

In OLS y can take values in \mathbb{R} , or $y \in \mathbb{R}$. However, to predict proportions then $y = p$ should $y_i = p_i \in \{0, 1\} \forall i$

for each observation the odds ratio is: $\lambda_i = \log\left(\frac{p_i}{1-p_i}\right)$

for the 1-dimension case, $\lambda_i = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \cdot x_i + \epsilon_i$ using MLE, estimate β_0 , β_1 and therefore λ_i
 $\hat{\lambda}(x) = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_i$

$$\hat{\lambda}_i = \log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right)$$

$$\hat{p}_i = (1 + e^{-(\hat{\beta}_0 + \hat{\beta}_1 \cdot x_i)})^{-1}$$

This transformation does not work well when $x = 0$ or $x = 1$ as the OLS loss function $\beta^{min} \|\lambda - x\beta\|^2$ increases with the right prediction and decreases with a wrong prediction.

The Deviance Function has the opposite behaviour...

$$D = p_i \cdot \log\left(\frac{p_i}{\hat{p}_i}\right) + (1 - p_i) \cdot \log\left(\frac{1 - p_i}{1 - \hat{p}_i}\right)$$

multiply assuming independent sampling to get the loss function:

$$D(\hat{p}_i | p_i) = 2n_i \left[p_i \cdot \log\left(\frac{p_i}{\hat{p}_i}\right) + (1 - p_i) \cdot \log\left(\frac{1 - p_i}{1 - \hat{p}_i}\right) \right]$$

Then minimize the loss function to estimate $(\hat{\beta}_0, \hat{\beta}_1)$

Generalized Linear Models - GLMs

GLMs extend linear regression to Binomial, Poisson, Gamma or any exponential distribution.

* GLMs transform an estimation problem in to a regression problem where the regression parameters are to be estimated.

GLM - exponential family

start with 1-parameter exponential family:

$$f_{\lambda}(y) = e^{\lambda y - \gamma(\lambda)} \cdot f_o(y)$$

where: the observed data $\underline{y} = (y_1, y_2, \dots, y_N)^T$ is assumed to come from $y_i \stackrel{ind}{\sim} f_{\lambda_i}(\cdot)$ for $i = 1, \dots, N$

write $\underline{\lambda}$ as a regression equation to avoid N estimations (one for each λ_i)

$$\underline{\lambda} = \underline{x} \cdot \underline{\alpha}$$

where:

α is a coefficients vector to assess the importance of each x
 x the covariance matrix from the data

the likelihood of \underline{y} for an exponential family is:

$$f_{\underline{\lambda}}(\underline{y}) = e^{\underline{\lambda} \cdot \underline{y} - \gamma(\underline{\lambda})} \cdot f_o(\underline{y})$$

let $\underline{\lambda} = \underline{x} \cdot \underline{\alpha}$, $\underline{z} = \underline{x}^T y$, $\Psi(\alpha) = \sum_{i=1}^N \gamma(x_i^T \cdot \alpha)$ such that

$$f_{\underline{\alpha}}(\underline{y}) = e^{\underline{\alpha}^T \underline{z} - \Psi(\alpha)} \cdot f_o(\underline{y})$$

GLM - Binomial Distribution

$$\lambda = \log\left(\frac{\pi}{1 + \pi}\right)$$

$$\gamma(\lambda) = n \log(1 + e^{\lambda})$$

GLM - Poisson Distribution

$$\lambda = \log(\mu)$$

$$\gamma(\lambda) = e^{\lambda}$$

GLM - Parameter Estimation

$(\mu_{\lambda}, \sigma_{\lambda}^2)$ denotes the expectation and variance of a univariate density $f_{\lambda}(y)$ in terms of the exponential family properties

$$y \sim (\mu_{\lambda}, \sigma_{\lambda}^2)$$

a N-dimensional vector y from $f_{\underline{\alpha}}(y)$ has mean and covariance matrix:

$$y \sim (\underline{\mu}(\underline{\alpha}), \Sigma(\underline{\alpha}))$$

where:

$$\underline{\mu}(\underline{\alpha}) = [\mu_{\lambda_1}, \mu_{\lambda_2}, \dots, \mu_{\lambda_N}]$$

$$\Sigma(\underline{\alpha}) = \begin{bmatrix} \sigma_{\lambda_1}^2 & & & \text{zeros} \\ & \sigma_{\lambda_2}^2 & & \\ & & \ddots & \\ \text{zeros} & & & \sigma_{\lambda_N}^2 \end{bmatrix}$$

* MLE estimate of α is to satisfy: $x^T [y - \mu(\alpha)] = 0$

where:

y is the data

$\mu(\alpha)$ is the means vector (adjust α to better describe y)

For the Normal distribution, $\hat{\alpha}^{MLE} = (x^T x)^{-1} \cdot x^T y$

For other family distributions $\hat{\alpha}$ is to be solved numerically.

$$* \hat{\alpha}^{MLE} \stackrel{a}{\sim} (\alpha, (x^T \cdot \Sigma(\alpha) \cdot x)^{-1})$$

where $(x^T \cdot \Sigma(\alpha) \cdot x)^{-1}$ is the variance of $\hat{\alpha}^{MLE}$

The DEVIANCE FUNCTION D() can be used to get MLE estimates of α . The D() between two densities f_1 and f_2 is:

$$D(f_1, f_2) = 2 \cdot \int_{sampleSpace} f_1(y) \cdot \log\left(\frac{f_1(y)}{f_2(y)}\right) dy$$

Deviance for:

Normal (known σ^2)

$$\left(\frac{\mu_1 - \mu_2}{\sigma}\right)^2$$

Binomial

$$2n \left[\pi_1 \cdot \log\left(\frac{\pi_1}{\pi_2}\right) + (1 - \pi_1) \cdot \log\left(\frac{1 - \pi_1}{1 - \pi_2}\right) \right]$$

Poisson

$$2 \cdot \mu_1 \left[\left(\frac{\mu_2}{\mu_1} - 1\right) \cdot \log\left(\frac{\mu_2}{\mu_1}\right) \right]$$

Hoeffding's Lemma \rightarrow the MLE $\hat{\alpha}$ is the choice of α that minimizes the total deviance. (as OLS minimizes the sum of squares)

Regression Trees

(aka. fancy averaging) - a technique to estimate regression surfaces using adapting partitioning.

At a given step k , of the partitioning algorithm, the mean of group k is

$$m_k = \sum_{i \in \text{group } k} \frac{y_i}{N_k}$$

and the sum of squares of group k is

$$S_k^2 = \sum_{i \in \text{group } k} (y_i - m_k)^2$$

The total S_k^2 is given by:

$$S_k^2 = S_{k_{left}}^2 + S_{k_{right}}^2 + \frac{N_{k_{left}}^2 N_{k_{right}}^2}{N_k} (m_{k_{left}} - m_{k_{right}})^2$$

the purpose is to maximize the 3rd term (aka. the information gain) in order to get a smaller S_k^2

Impurity

The impurity of a node measures the deviation from the predicted behaviour of the node. \rightarrow it's the probability that the prediction is incorrect usually measured by the Gini impurity or the mse (sum of squares error)

The Gini impurity is as follows:

$$\sum_{\text{group } i} p(i) \cdot (1 - p(i))$$

where $p(i)$ is the probability of success of the node. e.g.

$\sum_{\text{group } i} p(i) \cdot (1 - p(i)) = \frac{5}{95+5} \left(\frac{95}{95+5}\right) + \frac{10}{150+10} \left(\frac{150}{150+10}\right)$, where 5 points were wrong about outcome A, 95 right about outcome A, 10 wrong about outcome B and 150 right about outcome B.

Life Tables & Hazard Rates

What is the probability of dying/failing at time i ?

Discrete time

$$S_i = \sum_{j \geq i} f_j = P_r(x \geq i)$$

is the probability of surviving past time $i - 1$, where
 S_i is the marginal survival rate ($S_{0,i-1} = S_{i,j}$)
 f_j probability of dying at time j
 x length of a lifetime

$$h_i = \frac{f_i}{S_i} = P_r(x = i | x \geq i)$$

is the probability of dying exactly at time i given survival past time $i - 1$

$$S_{i,j} = \prod_{k=i}^j (1 - h_k) = P(x > j | x \geq i)$$

is the conditional surviving rate past age j given survived time $i - 1$

parameter estimates:

$$\hat{h}_i = \frac{y_i}{n_i} \text{ death rate at time } i$$

$$S_{t_0,j} = \prod_{k=t_0}^j (1 - \hat{h}_k)$$

It can be used to learn the probability of a t_0 old to survive past time t_n

Continuous time

$$S(t) = \int_t^\infty f(x) dx = P(T \geq t)$$

is the survival function, where $f(x)$ is density function (pdf).
 $S(t) = 1 - F(t)$ as it's the reverse cdf (cumulative distribution function)

$$h(t) = \frac{f(t)}{S(t)}$$

is the hazard rate density function

$$S(t) = e^{-\int_0^t h(x) dx}$$

is the continuous survival function in terms of the hazard function

Censored Data

Censored data includes information in which the number of subjects can change due to a death but also due to a loss of tracking.

If n is the no. of subjects at risk, y the no. of deaths, l the no. of lost subjects, h the hazard rate $\frac{y}{n}$, then: The Kaplan-Meier Estimate

$$\hat{S}_j = \prod_{k \leq j} \left(\frac{n - k}{n - k + 1} \right)^{d_{(k)}}$$

can be used as the survival function for censored data.
 $d_{(k)} = l_k$ and $k = monthno$.

$$sd(\hat{S}_j) = \hat{S}_j \left[\sum_{k \leq j} \frac{y_k}{n_k(n_k - y_k)} \right]^{\frac{1}{2}}$$

is the standard error of Kaplan-Meier Estimate

log-rank test

When working with "regular data", t -test can be implemented to compare survival times of two treatments.

However, when dealing with "censored data", log -rank test is to be used.

* Say the Null Hyp as the hazard rate for time i is the same for treatment A and in treatment B $Ho(t) : h_{A,i} = h_{B,i}$

* Ask "If n_d patients are randomly drawn from n , what is the probability distribution for the number of drawn patients to be in treatment A?"

n_d the total no. of deaths, n_s the total no. survivals, n_A the total no. of subjects in treatment A, and n_B the total no. of subjects in treatment B.

If $Ho(t)$ is TRUE, then:

$$E(y) = \frac{n_A \cdot n_d}{n}$$

is the mean of the hyper-geometric distribution

$$V(y) = \frac{n_A n_B n_d n_s}{n^2(n-1)}$$

the variance of $E(y)$

If $Ho(t)$ is FALSE, then: it's more probable to draw a subject from treatment B, which has a higher hazard rate.

* For all point in time, calculate y_i , E_i , V_i and compute the log-rank statistic z

$$z = \frac{\sum_{i=1}^N (y_i - E_i)}{\left(\sum_{i=1}^N V_i \right)^{\frac{1}{2}}}$$

the numerator is close to zero if Ho is TRUE

* Finally, compare z to normal critical values.

Jackknife

Used to assign a standard error to a $\hat{\theta}$

* given a iid sample $\underline{x} = (x_1, \dots, x_n)$ from an unknown

probability distribution F on some sample space X

* some algorithm $S()$ can be applied to compute $\hat{\theta} = S(x)$

* let $x_{-i} = (x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_n)$ (remove x_i)

* so, $\hat{\theta}_{-i} = S(x_{-i})$

$$\hat{se}_{Jack} = \left[\frac{n-1}{n} \cdot \sum_{i=1}^n (\hat{\theta}_{-i} - \hat{\theta}(\cdot))^2 \right]^{\frac{1}{2}}$$

where: $\hat{\theta}(\cdot) = \sum_{i=1}^n \frac{\hat{\theta}_{-i}}{n}$

i) get each one of the $\hat{\theta}_{-i}$

ii) numerically compute $\hat{\theta}(\cdot)$

iii) do $\hat{\theta}_{-i} - \hat{\theta}(\cdot)$

iv) apply JK equation

* works for any distribution * it depends on local derivatives - non-smooth statistics can result in erratic behaviour

Non-parametric Bootstrap

* replaces an unknown distribution F with an estimate $\hat{F} = \underline{x}$

* each bootstrap sample provides a bootstrap replication of

$\hat{\theta}^{*,b} = S(\underline{x}^{*,b})$ for $b = 1, 2, \dots, B$

$$\hat{se}_{boot} = \left[\frac{\sum_{b=1}^B (\hat{\theta}^{*,b} - \hat{\theta}^{*,\cdot})^2}{n-1} \right] = se(\hat{F})$$

where $\hat{\theta}^{*,\cdot} = \frac{\sum_{b=1}^B \hat{\theta}^{*,b}}{B}$

* "moves" the data much more than Jackknife

* does not depend on local derivatives

* can estimate any other measure of variability

* more computational intensive than Jackknife

Re-sampling Plans

Say $\underline{P} = (P_1, \dots, P_n)^T$ is a vector of weights with

$\sum_{i=1}^n P_i = 1$ and $P_i \geq 0 \forall i$

$\hat{\theta}^* = S(\underline{P})$ denotes $\hat{\theta}$ under a particular resampling vector \underline{P} .

(\underline{x} is fixed)

For the sample mean,

$$S(\underline{x}) = \sum_{i=1}^n \frac{x_i}{n}$$

$$\hat{\theta}^* = S(\underline{P}) = \sum_{i=1}^n P_i \cdot x_i$$

For the sample variance,

$$S(\underline{x}) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

$$\hat{\theta}^* = \frac{n}{n-1} \left[\sum_{i=1}^n P_i x_i^2 - \left(\sum_{i=1}^n P_i x_i \right)^2 \right]$$

The original estimate is $\hat{\theta} = S(\underline{P}_o)$, where:

$$\underline{P}_o = \frac{1}{n} \cdot [1, 1, 1, \dots, 1] \text{ (of } n \text{ components)}$$

which is the initial/original vector of weights.

The i^{th} iteration of the Jackknife estimate is:

$$\hat{\theta}_{-i} = S(P_{-i})$$

, where $P_{-1} = \frac{1}{n-1} [1, 1, 1, \dots, 1]$ where the 0 is located in the i^{th} element being removed.

A Bootstrap sample has weight vector

$$\underline{P}_b = \frac{(k_1, k_2, \dots, k_n)^T}{n}$$

with k_b being the no. of times each observation is chosen from the finite bootstrap sample in iteration b

$$k_b \sim Mult_L(n, \underline{P_o})$$

where L is the no. of categories/outcomes, and n is the no. of samples/draws.

$$f(k_b) = \frac{n!}{k_1!k_2!\dots k_n!} \cdot \prod_{l=1}^{L=n} P_{k_l}^{k_l} = \frac{n!}{k_1!k_2!\dots k_n!} \cdot \frac{1}{n^n}$$

is the multinomial pdf of k_b

Infinitesimal Jackknife

let's use the "convex combination" of vectors $\underline{P_o}$ and $\underline{P_{-i}}$

$$\begin{aligned} \underline{P_i} &= \underline{P_o} + \varepsilon(\underline{P_{-i}} - \underline{P_o}) \\ D_i &= \lim_{\varepsilon \rightarrow 0} \frac{S(P_i(\varepsilon)) - S(\underline{P_o})}{\varepsilon \|\underline{P_i} - \underline{P_o}\|} \\ \hat{se}_{IJ} &= \left(\frac{D_i^2}{n^2} \right)^{frac{1}{2}} \end{aligned}$$

computationally, shall be done with a small ε

Moving Blocks Bootstrap

Used in time series data, where x occurs in a meaningful order.

Let B_m be the set of continuous blocks of length m . *e.g.*

$$B_3 = \{(x_1, x_2, x_3), (x_2, x_3, x_4), \dots, (x_{n-2}, x_{n-1}, x_n)\}$$

Choose a m big enough such that the correlation between the $x's$ is negligible

Then use regular 1-sample bootstrap to randomly draw n/m block numbers

$$\hat{se}_{bootB} = \left[\sum_{b=1}^B \frac{\hat{\theta}^{*,b} - \hat{\theta}^{*,\cdot}}{B-1} \right]^{\frac{1}{2}}$$

$$\text{where } \hat{\theta}^{*,\cdot} = \sum_{b=1}^B \frac{\hat{\theta}^{*,b}}{B}$$

Multisample Bootstrap

Used when subjects in data are classified in 2 or more groups; and want to compute an estimator (e.g. median) for each group and then compare between them.

Do not sample from the n_{total} observations, instead re-sample with replacement n_1 samples of group 1 and n_2 samples of group 2 separately.

Then perform t-student for the means of the estimator. (as each bootstrap sample provides a bootstrap replication of $S()$)

Parametric Bootstrap

In regular bootstrap sample \underline{x} is used as \hat{F} ; where every outcome/category has the same probability ($1/n$), by keeping the y-axis at a fixed width histogram-wise.

Parametric Bootstrap is about keeping categories in the x-axis at a fixed width histogram-wise.

* it uses $f_{\hat{\theta}}(x)$ as \hat{F} , therefore the vector of weights P is given by $f_{\hat{\theta}}(x)$