

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 $w(x_k)$

$$w(x_k) = \left(1 - \left| \frac{x_i - x_k}{d} \right|^3 \right)^3$$

fork = 1, ..., n
d is the distance between x_i and the k^{th} neighbouring point

- * 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 $G(x_k)$

$$G(x_k) = \begin{cases} \left(1 - \left(\frac{|y_i - \hat{y}_i|}{6 \text{median}(|y_i - \hat{y}_i|)} \right)^2 \right)^2, & \left| \frac{|y_i - \hat{y}_i|}{6 \text{median}(|y_i - \hat{y}_i|)} \right| < 1 \\ 0, & \left| \frac{|y_i - \hat{y}_i|}{6 \text{median}(|y_i - \hat{y}_i|)} \right| \geq 1 \end{cases}$$

if $(p - \text{value} < \alpha)$ { reject H_o and accept H_a }
 else { can't reject H_o }

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

j \ b	1	2	3	...	B
	$y_{1(1)}$	$y_{1(2)}$	$y_{1(3)}$...	$y_{1(B)}$
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 $(t \text{ is of the 'wrong' sign}) p - \text{value} = 1 - p - \text{value}_{chart}$

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

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

- * $\hat{\theta}$ is sample specific, is a realization of $\hat{\Theta} = t(x)$. Typically, μ is the expected value of producing an estimate using $t(x)$ when x comes from F

$$E_F(\hat{\Theta}) = \mu$$

- * **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)

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

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.

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

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

- * 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:

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

- * 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.

$$f_0 \sim N(\mu_0, \sigma_0^2) \quad f_1 \sim N(\mu_1, \sigma_1^2)$$

$$f_0 \sim N(0, 1) \quad f_1 \sim N(0.5, 1)$$

$$f_0(x_i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x_i^2}{2}} \quad f_1(x_i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x_i - 0.5^2}{2}}$$

- * 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.

$$\alpha = P(\bar{x} > c | \mu = \mu_0)$$

$$\alpha = P((\bar{x} - \mu)\sqrt{n} > (c - \mu)\sqrt{n} | \mu = 0)$$

$$\alpha = 1 - P(\bar{x}\sqrt{n} \leq c\sqrt{n} | \mu = 0)$$

$$\alpha = 1 - \Phi(c\sqrt{n})$$

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

$$\Phi(c\sqrt{n}) = 1 - \alpha$$

$$c\sqrt{n} = \Phi^{-1}(1 - \alpha)$$

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

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

$$\beta = P(\bar{x} \leq c | \mu = \mu_1)$$

$$\beta = P((\bar{x} - \mu)\sqrt{n} \leq (c - \mu)\sqrt{n} | \mu = 0.5)$$

$$\beta = \Phi((c - 0.5)\sqrt{n})$$

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:

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

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

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.

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

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

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 \rightarrow the significance level does not change.

A Bayesian/Frequentist Comparison List

Bayesian:

* attention is in choosing an algorithm $t(x)$

* operates only in one sample with the whole parameter space

* answers all possible questions at once, since the posterior is a distribution

Frequentist:

* attention is in choosing a prior Π
* operates with one parameter (specific question) in many samples

* only computes the expected value and the variance (each answer requires an specific algorithm)

* is more flexible than Bayes as we can come up with many algorithms

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:

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

for a fixed x and a variable θ

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

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

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

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

* 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

* Reasonable Bayesian justification

$$P(\theta|x) = c_x \Pi(\theta) e^{\ell_x(\theta)}$$

* 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

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

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

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

* MLE can cause over-fitting identification problems when we fit a lot of parameters in θ (it would become very specific to our sample \rightarrow 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

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

$$\ddot{\ell}_x(\theta) = \frac{\ddot{f}_\theta(x)}{f_\theta(x)} - \left(\frac{\dot{f}_\theta(x)}{f_\theta(x)} \right)^2 \quad E(\ddot{\ell}_x(\theta)) = -I_0$$

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

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

e.g. for a normal dist.

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

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

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

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 $\hat{\theta}^{MLE}$

$E(\dot{\ell}_x(\theta)) = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \theta) = 0$, 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 not 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.