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:

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

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

For a mean: For the difference between two means:

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

$$se(\bar{x_1}, \bar{x_2}) = \sqrt{\frac{s_1{}^2}{n_1} + \frac{s_2{}^2}{n_2}}$$

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

$$\hat{\beta}_0 = \hat{Y} - \hat{\beta}_1\hat{X}$$

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

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

$$se(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

* predicted values = fitted curve given x:

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

* residuals $\hat{\epsilon}$:

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 + \beta_1 X_i$$

* 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 $w(x_k)$

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

    $for\ k = 1, ..., n$
    d is the distance between x_i
    and the kth 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 $G(x_k)$

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

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

LOESS: $\hat{y_k} = \sum_k w(x_k)G(x_k)(a + bx_k + cx_k{}^2)^2$
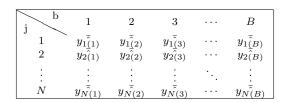
* 5) A series of new smoothed values is the result. The procedure can be repeated to get a more precise curve fitting.

Bootstrapping

* bootstrap principle:
$\sigma_{(\text{sampling w/replacemnt})} = \sigma_{(\text{across samples})}$
* bootstrap iterations: B
* original sample: $(x_i, y_i)_{i=1}^{N}$
* bootstrap samples: $(x_{j(b)}, y_{j(b)})_{j \epsilon I}$ for $b = 1, ..., B$,
$I = \{1, ..., N\}$, and j is the index that is randomly sampled from I
* for each b, compute $y_{\hat{j(b)}}$ using LOWESS or any other model

j \ b	1	2	3	$\cdots$	B
1	$y_{1\hat{(1)}}$	$y_{1\hat{(2)}}$	$y_{1\hat{(3)}}$	$\cdots$	$y_{1\hat{(B)}}$
2	$y_{2\hat{(1)}}$	$y_{2\hat{(2)}}$	$y_{2\hat{(3)}}$	$\cdots$	$y_{2\hat{(B)}}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
N	$y_{N\hat{(1)}}$	$y_{N\hat{(2)}}$	$y_{N\hat{(3)}}$	$\cdots$	$y_{N\hat{(B)}}$

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

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

* sort $i\bar{\hat{(b)}}$ 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 $H_a : \mu\{=, > \text{ or } <\}\mu_0$
* $t - statistict$ standarices 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$(p - value < \alpha)\{$ reject H_o and accept H_a $\}$
else $\{$ cant reject H_o $\}$
* α is the predetermined value of significance (usually 0.05)
* if$(t$ is of the 'wrong' sign$)p - value = 1 - p - value_{chart}$

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

$$t = \frac{\bar{x_1} - \bar{x_2}}{se(\bar{x_1}, \bar{x_2})}$$

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
* $x = (x_1, ..., x_n)$ is the data vector (aka. *the sample's values*)
* $X = (X_1, ..., X_n)$ 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,

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

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

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

* **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. $expected - truevalues$) $\qquad var = E_F\{(\hat{\Theta} - \mu)^2\}$

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.

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

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

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

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

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

* 3.2) MLE (maximum likelihood estimate): find $\hat{\theta}$ such that $L(x)$ is maximized
e.g.

$$\overset{\max}{\theta}\, L(x) \Rightarrow \overset{\max}{\mu}L(x) = \hat{\mu}^{MLE}$$

* 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 *prblem 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.

$$f_0 \sim N(\mu_0, \sigma^2_0) \qquad\qquad f_1 \sim N(\mu_1, \sigma^2_1)$$
$$f_0 \sim N(0, 1) \qquad\qquad f_1 \sim N(0.5, 1)$$
$$f_0(x_i) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x_i^2}{2}} \qquad f_1(x_i) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x_i-0.5^2}{2}}$$

* 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.*

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

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 $N(\mu_0, \sigma^2_0)$
* 6) calculate c
e.g. In general:

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

* 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:

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

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
$g(\mu)$: 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

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

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 fraternals have 0.5 probability to be of the same sex.

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

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

Flaws in Frequentist Inference

* In Frequentist, if the algorith 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 posible questions at once, since the posterior is a distribution

Frequentist:
* attention is in choosing a prior $\prod$
* 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 $g(\mu), \mu \epsilon \Omega$

Fisherian Inference and Maximum Likelihood Estimation

* The log-likelihood function is defined as:

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

for a fixed x and a variable θ

$\ell_x(\theta)$: 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 $\theta \epsilon \Omega$ that maximizes $\ell_x(\theta)$

$$MLE : \hat{\theta} = \overset{argmax}{\theta \epsilon \Omega} \{\ell_x(\theta)\}$$

* Estimate functions of the true parameter: $\hat{\gamma} = T(\hat{\theta})$
* Good frequentist properties (good bias & variance):

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

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

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

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

Likelihood and Maximum Likelihood

Fisher Information and the MLE

Conditional Inference

Permutation and Randomization

Notes and Details

Parametric Models and Exponential Families

Univariate Families

Osamu Katagiri - A01212611, linkedin.com/osamu-katagiri/