# Algorithms, Evidence, and Data Science Cookbook

## Part I: Classic Statistical Inference

* **Population:** the entire group
* **Sample:** a subset of the population
* **Mean:** $\mu$ 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:

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

Variance of a sample:

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

* **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

$$\hat{\beta_0} = \hat{Y} - \hat{\beta_1}\hat{X}$$

$$\hat{\beta_1} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$se(\hat{\beta_0}) = \hat{\sigma}^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right] \quad se(\hat{\beta_1}) = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

* predicted values = fitted curve given $x$:

$$\hat{Y}(x) = \hat{\beta_0} + \hat{\beta_1}x$$

* residuals $\hat{\epsilon}$:

$$\hat{\epsilon_i} = Y_i - \hat{Y_i} = Y_i - \hat{\beta_0} + \beta_1 X_i$$

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

* 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 $\hat{y_{j(b)}}$ using LOWESS or any other model

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

* for each $j$ row, the standard deviation $\sigma_j^{boot}$ is

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

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

### Hypothesis Testing

### T-test, one-sample

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

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

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 $H_o$ }
* $\alpha$ 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 $\mu_0$ to perform a one-sample t-test

### unpaired two-sample t-test

the two populations are independent
* $H_o : \mu_1 = \mu_2$
* $H_a : \mu_1\{=, > or <\}\mu_2$
* $t - statistict$

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

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** $\alpha$**:** probability of rejecting a true $H_o$
* **Type II error** $\beta$**:** probability of failing to reject a false $H_o$

### Notes

* the OLS confidence intervals work asymptotically → 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 $\theta$

usually,

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

where $t(x)$ is the algorithm

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

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

$\mu$ is the expected value of producing an estimate using $t(x)$ when $x$ 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 → 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 → modelling noise or over-fitting)

$$bias = \mu - \theta$$

(aka. $expected - true values$)

$$var = E_F\{(\hat{\Theta} - \mu)^2\}$$

### Frequentist principles

* usually defines parameters with infinite sequence of trials → hypothetical data sets $X^{(1)}, X^{(2)}, \ldots$ generate infinite samples $\hat{\Theta}^{(1)}, \hat{\Theta}^{(2)}, \ldots$ * 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.*

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

* 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

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

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 *e.g.*

### Notes and Details

## Bayesian Inference

### Two Examples

### Uninformative Prior Distributions

### Flaws in Frequentist Inference

### A Bayesian/Frequentist Comparison List

### Notes and Details

## Fisherian Inference and Maximum Likelihood Estimation

### Likelihood and Maximum Likelihood

### Fisher Information and the MLE

### Conditional Inference

### Permutation and Randomization

### Notes and Details

## Parametric Models and Exponential Families

### Univariate Families