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:

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

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:

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

For the difference between two means:

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

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

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

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

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

any regression is a conditional mean $\hat{Y_i} = E(Y_i|X_i)$
* Y : response variable
* X : covariate/predictor/feature
* predicted values = fitted curve given x:

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

Hypothesis Testing

Notes

Frequentist Inference

Frequentism in Practice

Frequentist Optimality

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

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