Introductory Statistics

Ivo Ugrina

University of Split

September 9, 2019

Uncertainty

Data Collection

Descriptive Statistics

Inferential Statistics

Probability Models

Populations and Samples

Descriptive Statistics

Describing Data Sets

Describing Data Sets

- Frequency Tables
- Relative Frequency Tables
- Graphs
- Histograms

Summarizing Data Sets

Describing Data Sets

- Sample Mean

$$\overline{x} = \frac{1}{n}\sum x_i$$

- Sample Median
- Sample Mode
- Sample Variance

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum (x_i - \overline{x})^2$$

- Sample Percentiles/Quantiles/Quartiles
- Box Plots

Random Variables

Random Variables

- Sample Space
- Concept
- Types of Random Variables
- Probability Density Function - PDF
- Cummulative Distribution Function - CDF
- Bernoulli RV
- Binomial RV
- Uniform RV
- Normal RV

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Exponential RV
- Student's t-distribution (RV)

Elements of Probability

Elements of Probability

- Axioms of Probability
- Sample Spaces Having Equally Likely Outcomes
- Conditional Probability
- Expectation
- Variance
- Law(s) of Large Numbers

(Sampling) Statistics

(Sampling) Statistics

- The Sample Mean
- Law of Large Numbers
- Convergence in Distribution
- Central Limit Theorem(s) - CLT

(Point) Parameter Estimation

(Point) Parameter Estimation

- Introduction
- Unbiased, Biased
- An example in R - Mean/Median
- Let's play a game - (Throwing) DARTS!
- Estimator "qualities"?
- Maximum likelihood estimators
- Bayesian approach (concept)
- *Bootstrap*

Interval (Paramter) Estimation

Interval (Parameter) Estimation

- Introduction
- Normal - Mean, Variance known

$$Z = \sqrt{n}\frac{\overline{X} - \mu_0}{\sigma} \sim N(0,1)$$

- Normal - Mean, Variance unknown

$$Z = \sqrt{n}\frac{\overline{X} - \mu_0}{\hat{\sigma}} \sim t(n-1)$$

Interval (Parameter) Estimation

- Normal - Variance, Mean unknown

$$(n-1)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1}$$

- Normal - Variance, Mean known ?
- Computational statistics (why?)

Hypothesis Testing

Hypothesis Testing

- Introduction
- Decision theory, Risk analysis
- Null hypothesis (*simple, composite*)
- Test!
- "*the objective of a statistical test of H0 is not to explicitly determine whether or not H0 is true but rather to determine if its validity is consistent with the resultant data*
- Errors (*type I error, type II error*)
- (implicit) Parametrisation
- Significance

Hypothesis Testing

- z-test

$$Z = \sqrt{n}\frac{\overline{X} - \mu_0}{\sigma}$$

- One-sided? Two-sided?
- t-test

$$T = \sqrt{n}\frac{\overline{X} - \mu_0}{\hat{\sigma}}$$

- Hands-on example in R
- p-values
- *the probability that, when the null hypothesis is true, the statistical summary (such as the sample mean difference between two groups) would be equal to, or more extreme than, the actual observed results*

Hypothesis Testing

- Uniformity of p-values? (hands-on example in R)
- Analytical solutions and Computational statistics. Importance of computers in statistics.
- Non-parametric tests
- Ranking?
- Permutation/Randomization tests
- A best test for XYZ? Simulation studies?

Regression

Basic structure

$$Y = f(X) + \epsilon$$

ϵ is random *error term* with zero mean, independent of X

Simple Linear Regression

Model

$$Y = \beta_0 + \beta_1 X + \epsilon$$

$$Y \approx \beta_0 + \beta_1 X$$

- ϵ is random *error term* with zero mean, independent of X
- β_0 and β_1 are unknown constants representing/called *intercept* and *slope*. Often called *coefficients* or *parameters*.

Estimating coefficients

- First, we need to have data,

$$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$$

$$X = (x_1, \cdots, x_n), \quad Y = (y_1, \cdots, y_n)$$

- With the data the goal is to obtain coefficient estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for which the linear model

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

fits the data *as close as possible*

Selected Topics (a glimpse)

Selected Topics

- Feature selection (both-ways)
- PCA
- Design of Experiments
- Example with Batch Correction and Normalization