A Short Introduction to Statistics and Data Analysis - DRAFT DO NOT CITE OUR QUOTE

James T. Durant

September 8, 2013

Outline

1. Introduction
2. Distribution Functions
3. Censored Data Analysis
   - Robust Regression on Order Statistics
   - Maximum Liklihood Estimation
   - Kaplan-Meier
   - Multiple Imputation
4. Confidence Intervals
   - Parametric Confidence Intervals
   - Bootstrapping
   - Chebychev Inequalities

Introduction

Today we will be demonstrating several concepts that we are utilizing in the Exposure Investigation and Data Analysis Team.
We will be using **R** as a platform to demonstrate these concepts. our focus will not be on the mechnics of using **R** or the underlying mathmatics, but to try and illustrate the concepts of what is happening and basic concepts that will help guide thier use.
R is not the sole platform that can perform these analyses, but the concepts are transient to all instances.
Almost all data analysis requires some anaysis and understanding - there are no cookbook techniques

Outline

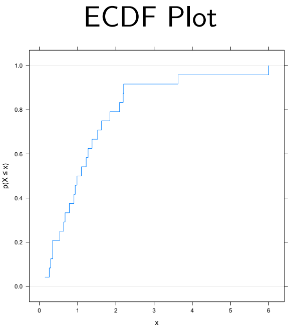
1 Introduction

2 Distribution Functions

3 Censored Data Analysis
- Robust Regression on Order Statistics
- Maximum Liklihood Estimation
- Kaplan-Meier
- Multiple Imputation

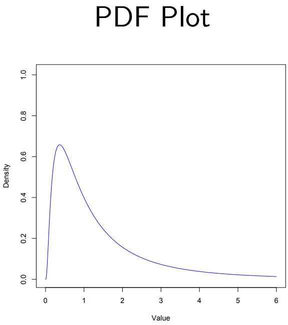
4 Confidence Intervals
- Parametric Confidence Intervals
- Bootstrapping
- Chebychev Inequalities

Emprical Distribution Functions (ECDF)



- Probability that a given value is less than a value
- Based on actual data - with each point having a probability 1/N where N is sample size.
- When plotted looks like stair steps going up from 0 to 1.

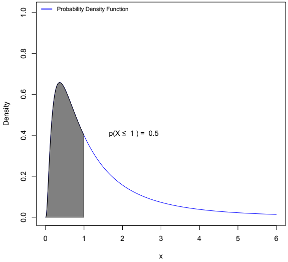
Probability Distribution Functions (PDF)



- Probability density that x is a certain value
- Based on a function - area below the curve must equal to 1
- Higher density means more likely values in that range

Probability Distribution Functions (PDF)

PDF Plot

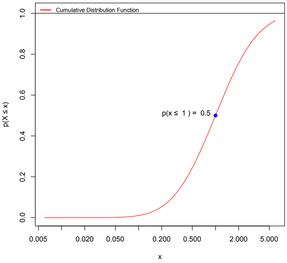


- PDF can be used ot calculate that x is at or below a certain value
- Equal to the area under the PDF curve below a given value
- Here we see probability $X \leq 1$ is 0.5
- With multiple observations, the liklihood is proportional to their densities multiplied together

Cummulative Distribution Functions (CDF)

We can also plot the PDF another way instead of the density on the y axis, we can plot the cumulative probability that $X \leq Value$. This is called the Cumulative Distribution Function (CDF).

CDF Plot

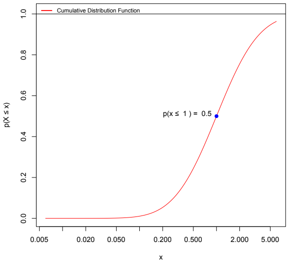


- CDFs can be used to calculate mean
- Area of (Value x CDF(Value)) = mean

ECDF and CDF

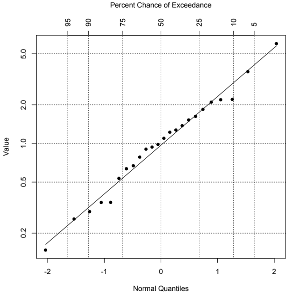
ECDF approximates CDF

ECDF and CDF Plot



- BOTH can be used to calculate mean
- Area of (Value x CDF(Value)) = mean
- Area of (Value x ECDF(Value)) = mean

Probability Plotting



- Compute plotting position (number of standard deviations)
- This is also a probability
- Plot values versus their probability on a scale of standard deviations

Outline

1 Introduction

2 Distribution Functions

3 Censored Data Analysis
- Robust Regression on Order Statistics
- Maximum Liklihood Estimation
- Kaplan-Meier
- Multiple Imputation

4 Confidence Intervals
- Parametric Confidence Intervals
- Bootstrapping
- Chebychev Inequalities

Robust Regression on Order Statistics - [ROS]

Maximum Liklihood Estimation - [MLE]

Kaplan Meier - [KM]

Multiple Imputation - [MI]

Outline

1. Introduction
2. Distribution Functions
3. Censored Data Analysis
   - Robust Regression on Order Statistics
   - Maximum Liklihood Estimation
   - Kaplan-Meier
   - Multiple Imputation
4. Confidence Intervals
   - Parametric Confidence Intervals
   - Bootstrapping
   - Chebychev Inequalities

Choice of Parametric Distribution

The Bootstrap

Limitations of Bootstraps

Chebychev Inequality