

# Essentials of Statistics and Probability

Dhruv Sharma

May 22, 2007

Department of Statistics, NC State University  
dbsharma@ncsu.edu  
SAMSI Undergrad Workshop

# Overview

## Practical Statistical Thinking

Introduction

Data and Distributions

Variables and Distributions

Graphical and Numeric Description

Density Curves and Normal Distribution

Population and Sample

## Probability: The Language of Statistics

Randomness

Gambling and Probability

Random Variables and Probability Distributions

Sampling Distributions

Law of Large Numbers

Distribution of Sample Mean

Central Limit Theorem

# Death Penalty and Race of Defendant

- ▶ Number of death penalties awarded to defendants in multiple murder cases, in Florida from 1976 to 1987. Ranelet *et al.* (1991).
- ▶ Def (White): Yes 53 No 430 (11% Yes)
- ▶ Def (Black): Yes 15 No 176 (7.9% Yes)
- ▶ What do you make of this?

## Death Penalty and Race of Victim

- ▶ What about the race of the victim?
- ▶ Vic (White) Def (White): Yes 53 No 414 (11.3% Yes)
- ▶ Vic (White) Def (Black): Yes 11 No 37 (22.9% Yes)
- ▶ Vic (Black) Def (White): Yes 0 No 16 (0.0% Yes)
- ▶ Vic (Black) Def (Black): Yes 4 No 139 (2.8% Yes)
- ▶ Now what do you make of this?
- ▶ Lies, Damned Lies and Statistics!

# Variables and Distributions

- ▶ Individuals. Cows on a farm (Daisy, Betty, Polly, Moo, *etc.*)
- ▶ Variable. Any characteristic of an individual. Since we don't know this, lets call it 'x'.
- ▶ Data about individuals. Units in a sample of size 'n', being the number of observations. Variable = weight. Weight of Daisy (Unit 1) =  $x_1$ .

# Variables and Distributions

- ▶ Categorical variable. Places individuals into one of several groups or categories. Examples?
- ▶ Quantitative variable. Takes numerical values for which arithmetic operations such as adding and averaging make sense. Examples?
- ▶ The Distribution of a variable tells us what values it takes and how often it takes these values.

## Saying it with Pictures

- ▶ Pie charts for Categorical data. Can use the counts of individual groups or percentages.
- ▶ Histograms for Quantitative data. Place into groups of ascending order. Can use the counts of individual groups or percentages.
- ▶ The shape of these graphical display methods can tell you about the population.

## Saying it with Numbers: 5 Number Summary

- ▶ 5 Number Summary.
- ▶ Sort the data in ascending order of the values taken by the variable.
- ▶ Median **M** is the midpoint of the distribution. Half the values fall over and half below the median.
- ▶ Quartiles.  $Q_1$  and  $Q_3$  split the distribution at 25% and 75%, like median splits it at 50%.
- ▶ The 5 Number Summary is Min,  $Q_1$ , M,  $Q_3$ , Max.
- ▶ M talks about the center, and its relation with the rest, the spread.

## Saying it with Numbers: Mean and Standard Deviation

- ▶ The **mean** ( $\bar{x}$ ) of  $n$  observations is the average of the values the variable takes.

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

- ▶ The Variance ( $s^2$ ) is a measure of spread and is given by,

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

- ▶ Standard Deviation (S.D) is the square root of the variance.
- ▶ An outlier is a wierdo in the dataset (Kryptonite to mean and S.D!) When you have outliers, the 5 Number Summary should be used.
- ▶ Why Central Tendency and Spread?

# Density Curves and Normal Distribution

- ▶ Sometimes the overall pattern of a large number of observations is so regular that we can describe it by a smooth curve.
- ▶ We have a percentage histogram, a density curve is a line connecting the bars with an area of 100% or 1 under the curve.
- ▶ Commonly data seems to have a Normal distribution. One with a Bell shaped density curve.

# Population and Sample

- ▶ A **sample** is a representative subset of a **population**.
- ▶ Parameter vs. Statistic
- ▶ A parameter is a number that describes a population.  
Unknown quantity we wish to know, say, population mean  $\mu$ .  
Usually a Greek symbol.
- ▶ A statistic is a number that can be computed from only the sample data to estimate an unknown population parameter.  
Say, sample mean  $\bar{x}$ .

# The Story so Far

- ▶ Statistics is a mathematical science pertaining to the collection, analysis, interpretation or explanation, and presentation of data.
- ▶ Beware, Statistics could lie!
- ▶ What are they not telling you?

# The Truth about where Statisticians come from!

- ▶ People like to gamble, take chances against great odds. Probability was born out of gambling.
- ▶ Chevalier De Mere (17th century), gambled on dice. 1 six in 4 throws = 50.77% but 2 sixes in 24 throws of 2 = 49.14%.
- ▶ Contacted Fermat and Pascal. Led to the birth of Probability Theory.
- ▶ Chance behaviour is unpredictable in the short run but has a regular and predictable pattern in the long run.

# Basic Terminology

- ▶ Random Phenomenon. If individual outcomes are uncertain but there is a regular distribution of outcomes in a large number of repetitions.
- ▶ The probability of any outcome of a random phenomenon is the proportion of times the outcome would occur in a very long series of repetitions.
- ▶ Toss a coin. Count the number of heads. After a while the proportion of heads will tend to be half.

# Probability Models

- ▶ The **sample space  $S$**  of a random phenomenon is the set of all possible outcomes.
- ▶ An **event** is any outcome or a set of outcomes of a random phenomenon. An event is a subset of the sample space.
- ▶ A **probability model** is a mathematical description of a random phenomenon consisting of two parts: a sample space and a way of assigning probabilities to events.

# Probability Rules

- ▶ Rule 1: The probability  $P(A)$  of any event  $A$  satisfies  $0 \leq P(A) \leq 1$ .
- ▶ Rule 2: If  $S$  is the sample space in a probability model, then  $P(S) = 1$ .
- ▶ Rule 3: For any event  $A$ ,  $P(\text{not } A) = 1 - P(A)$ .
- ▶ Two events  $A$  and  $B$  are **disjoint** if they have no outcomes in common and so can never occur simultaneously. If  $A$  and  $B$  are disjoint,  $P(A \text{ or } B) = P(A) + P(B)$ . (Addition rule for disjoint events).

# Random Variable

- ▶ A **random variable** is a variable whose value is a numerical outcome of a random phenomenon. Usually  $X$ ,  $Y$  or  $Z$  (capital letters).
- ▶ The **probability distribution** of a random variable  $X$  tells us what values  $X$  can take and how to assign probabilities to those values.

# Probability Density Function

- ▶ Back to density curves, "Math Imitates Life".
- ▶ The density curve could be seen as a math function with area 1 under the curve.
- ▶ If  $X$  is a random variable taking values  $x$ , then its p.d.f can be seen as  $f(x)$ , following the probability rules.

# The Normally Distributed Random Variable

- ▶ Most prevalent distribution where  $X$  takes values on the Real line.
- ▶ Its p.d.f is given by,

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

- ▶  $\mu$  describes central tendency and  $\sigma$  describes spread.
- ▶ Notation,  $X \sim N(\mu, \sigma^2)$ .
- ▶  $X \sim N(0, 1)$  is known as a Standard Normal Distribution.

# Law of Large Numbers

- ▶ Draw observations at random from any population with finite mean  $\mu$ . As the number of observations drawn increases, the mean  $\bar{x}$  of the observed values gets closer and closer to the mean  $\mu$  of the population.

## Distribution of Sample Mean

- ▶ The **sampling distribution** of a statistic is the distribution of values taken by the statistic in all possible samples of the same size from the same population.
- ▶ Suppose that  $\bar{x}$  is the mean of an SRS of size  $n$  drawn from a large population with mean  $\mu$  and S.D  $\sigma$ . Then the mean of the sampling distribution of  $\bar{x}$  is  $\mu$  and its S.D is  $\frac{\sigma}{\sqrt{n}}$ .

# Central Limit Theorem

- ▶ If a population has the  $N(\mu, \sigma^2)$  distribution, then the sample mean  $\bar{x}$  of  $n$  independent observations has the  $N(\mu, \frac{\sigma^2}{n})$  distribution.
- ▶ Draw an SRS of size  $n$  from any population with mean  $\mu$  and finite S.D  $\sigma$ . When  $n$  is large, the sampling distribution of the sample mean  $\bar{x}$  is approximately normal:  $\bar{x}$  is approximately  $N(\mu, \frac{\sigma^2}{n})$ .

Thank you.