Interface of Biology and Mathematics
Descriptive Statistics

David B. Damiano

Department of Mathematics and Computer Science
College of the Holy Cross
Worcester, MA
ddamiano@holycross.edu

July 2, 2008

Contents

Introduction

Descriptive Statistics

- Thinking about physical/real-world problems in a consistent manner is made easier by the use of numbers
- This generates large–often VERY LARGE–collections of numbers
- We want to know
  - General properties of these collections
  - How particular numbers relate to the entire collection
- And, based on these properties and relations, we want to be able to make statements about the underlying reality

DATA–THE STARTING POINT

Examples

- The populations of towns in Worcester County
- The number of men/women in Worcester by age
- The results of repeating an experiment many times
- The number of heads resulting from 100 tosses of coin
- The number of times a number of heads results from 100 people tossing a coin 100 times
- The number of people in a random sample of 500 people who believe global warming is a result of human activity
- The number of pea plants with yellow seeds that result from second generation hybrids of (pure) green seeded plants with (pure) yellow seeded plants.

TYPES OF DATA

Definitions

- A VARIABLE is a characteristic of the subjects/objects of a study
- A variable can be QUALITATIVE or QUANTITATIVE
  - Handedness–right-handed, left-handed, ambidextrous–is qualitative
  - Population size is quantitative
- A quantitative variable can be DISCRETE or CONTINUOUS
  - Number of offspring is discrete
  - Mass of an object is continuous
- A DATA SET is a collection of values for one or more variables associated with the objects of study

A First Example

Populations of towns and cities in Worcester County

City/Town	Population (2000)
Ashburnam	5,546
Athol	11,299
Auburn	15,901
Barre	5,113
⋮	⋮
Winchendon	9,611
Worcester	175, 898

FAMILIAR TERMS I

- N is the number of cities/towns

$$N = 60$$

- TOTAL Population: add the populations of each city/town

$$5,546 + 11,299 + 15,901 + \cdots + 9,611 + 175,898 = 750,963$$

- AVERAGE or MEAN Population: Divide the total population by the number of cities/towns

$$\frac{750,963}{60} = 12,516.05 \approx 12,516$$

- MODE The data value that occurs most frequently in the data set

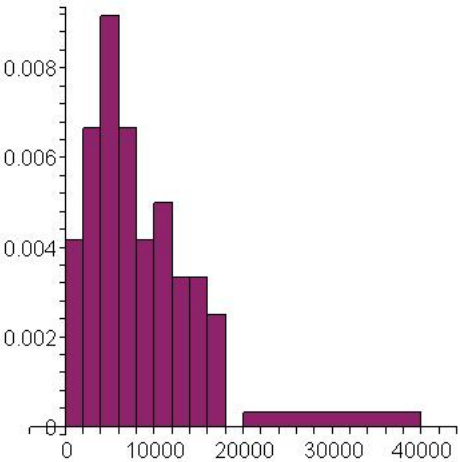
  No two populations are the same so . . .
  anyone is the mode

Familiar Terms II

- MEDIAN is the value separating the upper and lower halves of the data when the data is put in order
  - When N is odd, pick the middle value
  - When N is even, pick the average of the closest values to the middle
$$\frac{7,380+7,481}{2} = 7,430.5$$
- nth PERCENTILE is the value for the data so that n% of the data is at or below this value.
  - The 25th percentile is the first quartile $= 4,148$
  - The 50th percentile is the median $= 7,430.5$
  - The 75th percentile is the third quartile $= 13,182$

The Histogram for Worcester County City/Town Populations

(Histogram represents 97% of the data.)



Histogram Terminology and Properties I

- A Histogram is a graph consisting of rectangular Blocks resting on a horizontal number line.
- The Scale on the number line is the scale for the data
- The number line is divided into Class Intervals
  - Class intervals are ranges on the number line
    for example, between 2,000 and 4,000
  - Class intervals touch but do not overlap
  - There are enough class intervals so (almost) each data value falls into some class interval

Histogram Terminology and Properties II

- For each class interval there is a Block resting on the interval
  - The Area of each block is the Percentage of Data that falls into the corresponding class interval
  - The Height of each block is

$$\frac{\text{Area of the block}}{\text{Width of the class interval}} = \frac{\text{\% of data in class interval}}{\text{Width of the class interval}}$$

  - The scale on the vertical axis is a Density Scale with units

    Per Cent per Horizontal Unit

- The sum of the areas of the blocks in a histogram is (usually) 100%
- If all the class intervals have the same width, the density scale is equivalent to a Frequency Scale

CONSTRUCTING A HISTOGRAM FROM RAW DATA

- Choose class intervals
  - Enough to get a sense of the distribution of data
  - Consider using more intervals in ranges where the data is denser
  - If data takes only integer values, split intervals at an integer plus a $\frac{1}{2}$ (see below)
- BIN the data (assign data to class intervals)
  - Use a consistent rule where intervals touch
  - For example, values common to adjacent class intervals assigned to the right-interval
- Determine percentage of data in each class interval
- Determine the height of each block by the rule

$$\frac{\text{Area of the block}}{\text{Width of the class interval}} = \frac{\text{\% of data in class interval}}{\text{Width of the class interval}}$$

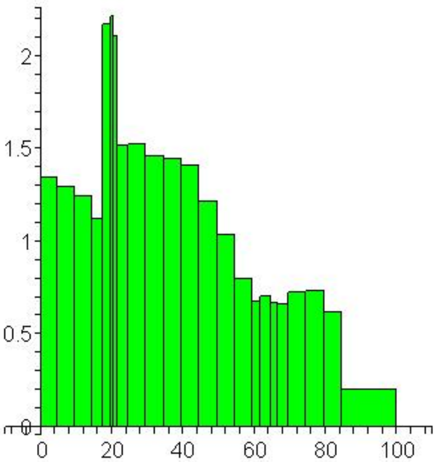
- Sketch the blocks

WOMEN IN WORCESTER BY AGE (2000)

Age Range	Number	Age Range	Number
Under 5	5,424	45 to 49 years	5,452
5 to 9 years	5,809	50 to 54 years	4,651
10 to 14 years	5,582	55 to 59 years	3,575
15 to 17 years	3,014	60 and 61 years	1,210
18 and 19 years	3,904	62 to 64 years	1,894
20 years	1,987	65 and 66 years	1,193
21 years	1,892	67 to 69 years	1,782
22 to 24 years	4,082	70 to 74 years	3,262
25 to 29 years	6,840	75 to 79 years	3,285
30 to 34 years	6,535	80 to 84 years	2,756
35 to 39 years	6,496	85 years and over	2,797
40 to 44 years	6,312	Total	89,734

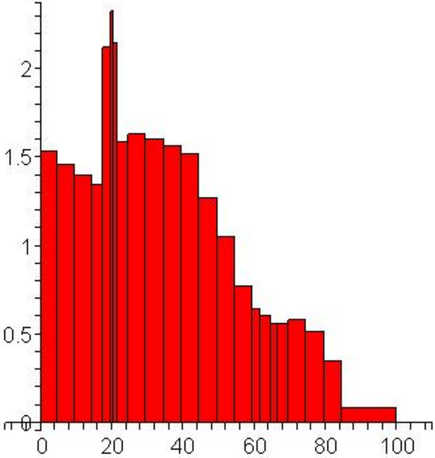
THE HISTOGRAM FOR WOMEN IN WORCESTER BY AGE (2000)

(Histogram represents 100% of the data.)



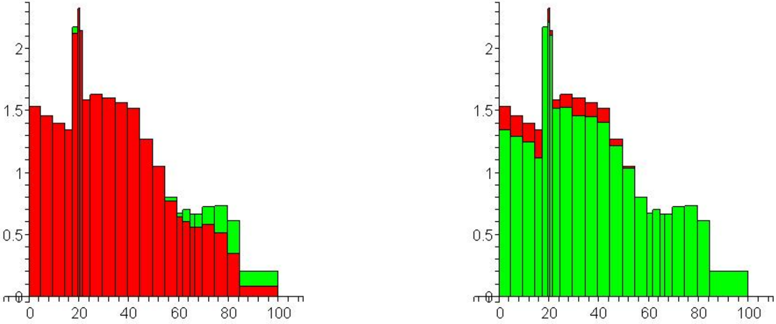
THE HISTOGRAM FOR MEN IN WORCESTER BY AGE (2000)

(Histogram represents 100% of the data.)



A Comparison

(Histograms represents 100% of the data.)



Spread of Data

We want a measure of the "spread" of data

- The mean serves as the "center" of the data
- The distance of a data point from the mean should be an ingredient
- Intuitive (wrong) version
  - For each data point, find distance from the mean
  - Average these distances
- What could be wrong with the above?

Standard Deviation Spread of Data

Spread of data, correct version

- For each data point compute the distance from the mean of the data set by subtraction
- Square these distances (to get positive numbers) (Square)
- Average the squared distances (Mean)
- Take the square root of the average of the squared distances (Root)
- Read in reverse order we have Root-Mean-Square
- The Root-Mean-Square of a data set is its Standard Deviation

The Standard Deviation of Worcester County Populations

Including Worcester

- Population Mean ≈ 12570.
- Sum of squares of population difference

$$(5546-12570)^2+(11299-12570)^2+\cdots+(175898-12570)^2 = 31,252,134,145$$

- Mean Square

$$\frac{31,252,134,145}{60} \approx 520,868,902$$

- Standard Deviation (Root-Mean-Square)

$$SD = \sqrt{520,868,902} \approx 22,823$$

The Standard Deviation of Worcester County Populations

Excluding Worcester

- Population Mean $\approx 9,802$.
- Sum of squares of population difference

$$(5546-12570)^2+(11299-12570)^2+\cdots+(9611-12570)^2+ \\ = 4,124,034,337$$

- Mean Square

$$\frac{4,124,034,337}{59} \approx 69,898,887$$

- Standard Deviation (Root-Mean-Square)

$$SD = \sqrt{69,898,887} \approx 8,361$$

The Normal Curve I

Is there a standard form for data?

- NO!... but
- In certain important situations, up to a change of units, YES!
- It goes by many names, referring to the shape of the histogram
  - A “bell-shaped” curve
  - The “Gaussian” curve
  - The “normal” curve

THE NORMAL CURVE II

When does the normal curve apply?

- Measurement error–very important for us
- Random processes, like tossing a coin–extremely important
- Variability in sampling
- Boot strapping

Properties of the Normal Curve

The Normal Curve

- is always positive (above the horizontal or x-axis)
- is symmetric about the vertical line $x = 0$
- has turning or inflection points 1 standard unit from 0
- has area equal to 100% in standard units (units of SDs)
- has approximately 64% of its area within 1 standard unit of 0
- has approximately 95% of its area within 2 standard units of 0
- has approximately 99% of its area within 3 standard unit of 0

FORMULA(S) FOR THE NORMAL CURVE

Expressed in terms of e, the base for the natural logarithm

- $e \approx 2.718$
- Centered at 0 with standard deviation 1

$$\frac{100\%}{\sqrt{2\pi}} e^{-x^2/2}$$

- Centered at the mean μ with standard deviation σ

$$\frac{100\%}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{\sigma^2}}$$