STATS CHAPTER 1-3 SUMMARY

PAUL L. BAILEY

1. Chapter 1

1.1. Section 1.1 - What is Statistics?

Statistics is the study of how to collect, organize, analyze, and interpret numerical information from data.

Individuals are the people or objects included in the study.

A *variable* is a characteristic of the individual to be measured or observed. Each variable takes values from a specific set.

A *quantitative variable* has a value or numerical measurement for which operations such as addition or averaging make sense.

A *qualitative variable* describes an individual by placing the individual into a category or group such as male or female.

In *population data*, the variable is from every individual of interest.

In *sample data*, the variable is from only some of the individuals of interest.

Levels of Measurement dictate appropriate methods of comparing value for a given variable:

- A *nominal level* variable takes values from a structureless set ("in name only").
- An *ordinal level* variable takes values from an ordered set.
- An *interval level* variable admits meaningful substraction.
- A *ratio level* variable admits meaningful division.

Example 1. Temperature measures heat; heat is the amount of kinetic energy of the molecules of a substance at rest. The complete absence of heat has the temperature of absolute zero.

$$0^\circ\mathrm{K} = -273^\circ\mathrm{C} = -460^\circ\mathrm{F}$$

Degrees Celsius is at the interval level, whereas degrees Kelvin is an the ratio level.

Date: February 5, 2009.

1.2. **Section 1.2 - Random Samples.**

A *selection process* is a method of selecting an element from a set.

A selection process is *random* if the probability of selecting any element of the set equals the probability of selecting any other element.

A *sample of n measurements* is a subset of size n.

A *random sample* of n measurements from a population is a subset of the population selected in a manner such that every individual has and equal probability of being selected.

A *simple random sample* of n measurements from a population is a subset of the population selected in a manner such that every sample of size n has an equal chance of being selected. This implies that each element has equal probability of selection.

Example 2. (A random sample which is not a simple random sample.) Let the population be the numbers from one to 100. Let $n = 10$. Select a number between 1 and 100 at random, and along with it the next nine numbers (wrap around, so that 1 follows 100). This sample is random, but not a simple random.

Other types of sampling:

- *Statified sampling*: if the individuals can be grouped by stata, select individuals from each strata
- *Systematic sampling*: if the individuals occur in a random order, select every k^{th} individual
- *Cluster sampling*: if the individuals are grouped in clusters, select all individuals from a random set of clusters
- *Convenience sampling*: select whatever is convenient

1.3. **Section 1.3 - Experimental Design.**

Select the individuals:

- *Census*: use the entire population
- *Sample*: use a subset of the population

Decide on the type of study:

- *Observational*: observations and measurements are conducted in a manner which does not change the variable being measured
- *Experiment*: a treatment is deliberately imposed on the individuals in order to observe a possible change in the variable being measured

A *controlled experiment* consists of

- *experimental group*: receives the treatment
- *control group*: does not receive the treatment

The control group may receive a *placebo* (empty treatment) so they do not know they are the control group. In a *double blind study*, the researchers do not know who is in the control group throughout the course of taking measurements in the experiment.

The control group helps account for the presence of unknown variables that might have an underlying effect on the variable being measured. Such unknown variables are called *lurking* or *confounding* variables.

2. Chapter 2

2.1. **Section 2.1 - Organizing Data - Graphs and Charts.**

Discrete versus continuous data:

- *Discrete*: the values come from a finite set (e.g. the integers); there are only so many possibilities
- *Continuous*: the values come from an interval (e.g. the real numbers), and may range continuously throughout the interval.

One versus two variables: Graphing requires two variables. If we are concerned with a list of n numbers, there is a hidden second variable, which is the numbers 1 through n, giving the sequence in which the numbers are presented.

- Bar Graphs (subclass: Pareto Charts: bar graph sorted by decreasing frequency)
- Pie Charts: percentage * 360° = angle
- Line Graphs: (example Time Plots)

2.2. **Section 2.2 - Histograms.**

2.2.1. *What is it?* Have: list of numbers.
This produces: min=a, max=b.
Decide on number of classes n.
Class width is $\frac{b-a}{n}$ (Excel: class is "bin")
Frequency per class: count and graph.
We call $b-a$ the *range* of the data (population or sample).

2.2.2. *Distribution.* : how do the frequencies vary across the range?

- Rectangular
- Mound shaped
- Skewed Left
- Skewed Right
- Bimodal

2.3. **Section 2.3 - Stem and Leaf.**

3. Chapter 3

3.1. **Section 3.1 - Averages and Variations.** Three types of averages:

- *Mode*: the most frequent value
- *Median*: the middle value (if even number, mean of middle two)
- *Mean*: sum divided by count

Notation:

- $\sum x$ - the sum of the values in the population or sample
- N = population size
- n = sample size
- $\mu = \frac{\sum x}{N}$ (population mean)
- $\overline{x} = \frac{\sum x}{n}$ (sample mean)

Consider: which average works with which data level?

- *Nominal*: mode
- *Ordinal*: mode; to lesser extent, median
- *Interval*: all
- *Ratio*: all

Variations of Mean:

- *Trimmed Mean*: eliminate top and bottom 5 percent, then mean
- *Weighted Mean*: $\frac{\sum wx}{\sum w}$ where each value x is weighted by w

Averages of Skewed Data

- Mound: mean = mode = median
- Skewed Left: mean < median < mode
- Skewed Right: mod < median < mean

3.2. **3.2 - Measures of Variation.**

Range: the difference between the max and the min.
Population Difference: $x - \overline{x}$
Population Sum of Squares: $\sum(x-\mu)^2$
Population Variance: $\sigma^2 = \dfrac{\sum(x-\mu)^2}{N}$
Population Standard Deviation: $s = \sqrt{\dfrac{\sum(x-\mu)^2}{N}}$
Sample Difference: $x - \overline{x}$
Sample Sum of Squares: $\sum(x-\overline{x})^2$
Sample Variance: $s^2 = \dfrac{\sum(x-\overline{x})^2}{n-1} = \dfrac{\sum x^2 - (\sum x)^2/n}{n-1}$
Sample Standard Deviation: $s = \sqrt{\dfrac{\sum(x-\overline{x})^2}{n-1}} = \sqrt{\dfrac{\sum(x-\overline{x})^2}{n-1}}$
Relate these? $\sigma = s\sqrt{\dfrac{n-1}{n}}$ for $n = N$

3.3. **3.3 - Percentiles and Box-and-Whisker Plots.**

Given a distribution, a value x in its range is in the p^{th} *percentile* if p percent of the data are less than or equal to x, and p is the smallest such integer.

To compute the percentile, sort the data $a = x_1 \leq x_2 \leq \cdots \leq x_n = b$. Here, a is the min and b is the max. Now let x be between a and b: $a \leq x \leq b$. Let k be the largest integer such that $x_k \leq x$. If $\rho(x)$ is the percentile, then

$$\rho(x) = \left\lfloor \frac{100k}{n} \right\rfloor.$$

Given a distribution, a value x in its range is in the q^{th} *quartile* if $25q$ percent of the data are less than or equal to x, and q is the smallest such integer.

Define the number $Q1$, $Q2$, and $Q3$ by

- $Q2$ is the median
- $Q1$ is the median of the data between the minimum and the median
- $Q3$ is the median of the data between the median and the maximum

4. Suggested Exercises

Old §1.1 # 7 equals New §1.1 # 9
Old §1.2 # 15 equals New §1.2 # 15
Old §1.3 # 1 equals New §1.3 # 3
Old §1.3 # 2 equals New §1.3 # 4
Old §2.1 # 2 equals New §2.2 # 4
Old §2.2 # 1 equals New §2.1 # 7
Old §2.3 # 12 equals New §2.3 # 10
Old §3.1 # 3 equals New §3.1 # 9
Old §3.2 # 2 and 3 equals New §3.2 # 13
Old §3.3 # 2 equals New §3.2 # 16
Old §3.4 # 11 equals New §3.3 # 7
Old §3.4 # 14 equals New §3.3 # 10

Department of Mathematics and CSci, Southern Arkansas University
E-mail address: plbailey@saumag.edu