AP STATISTICS
TOPIC 1: VOCABULARY

PAUL L. BAILEY

1. Chapter 1 - Getting Started

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.
- 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. Electoral polls act at the nominal level: the individual is asked to pick a candidate from a finite list.

Opinion polls often act at the ordinal level: the individual is asked if they strongly agree, agree, disagree, or strongly disagree.

Example 2. 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.

Homework: §1.1 # 1, 2, 3, 4, 6, 9

Date: August 18, 2021.

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.

Random numbers may be supplied a random number table, calculator, or computer.

Example 3. (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.

A *simulation* is a numerical facsimile or representation of a real-world phenomenon.

Other types of sampling:

- *Stratified sampling*: if the individuals can be grouped by strata, 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.

Example 4. Polling firms often stratify by age, sex, and/or ethnicity.

Example 5. If people are lined up in a queue, and a question is asked of every fifth person, this would constitute systematic sampling.

Example 6. There are 1000 prisons in the country; we pick a random subset of five prisons, and measure the incidence of communicable disease. This is cluster sampling.

Example 7. We ask the first ten people we meet on the street if they like goats; this is a convenient sample.

Homework: §1.2 # 3m 6m 8, 10, 14

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

Steps of a Study:

(1) Identify Population (who/what are the individuals?)
(2) Identify Variable (what is to be measured?)
(3) Population or Sample?
(4) Ethics, for example, privacy
(5) Collect Data
(6) Descriptive Statistics (to be discussed later)

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.

Surveys are a means of collecting information to study. These can be observational or experimental. Watch out for hidden bias.

Homework: §1.3 # 1, 2, 3

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
- *Continuous*: the values come from an interval of real numbers

Types of graphs of discrete data.

- Dot Plots: values on horizontal access, evenly spaced dots above each value count frequency
- Bar Graphs: bars have equal width and are uniformly spaced, lengths of bars represent values.
- Circle Graphs (or Pie Chart): percentage * 360° = angle
- Line Graphs: ordered population, (individual, value) pairs plotted and joined by line segments

Changing scale: indicate with a squiggle in the y-axis where there is a jump.

A *Pareto chart* is a bar graph with nominal population, values are frequency, listed in decreasing order.

A *time-series graph* contains data plotted in order of occurrence at regular intervals over a period of time. The data consist of measurements of the same variable for the same subject.

For any graph, provide a title, label the axes, and identify units of measure.

Homework: §2.1 # 1, 4, 5, 7, 11

2.2. **Section 2.2 - Frequency Distributions.**

2.2.1. *Histograms.* We start with a finite list of numeric values. Let a be the minimum values and let b be the maximum value. The *range* of the data is $b - a$.

We decide on number of classes (or *bins*) we wish to consider. Let n be the number of classes.

The *class width* is $\Delta x = \frac{b-a}{n}$.

The *class boundaries* are $x_i = a + i\Delta x$, for $i = 0, 1, \ldots, n$.

The *class midpoint* of the i^{th} class is $\dfrac{x_{i-1} + x_i}{2}$.

The *frequency* of the i^{th} class is the number of values which land within the class boundaries.

A *histogram* is a bar chart with the class across the bottom, in increasing order, and the frequency as the height of the bar.

The *relative frequency* of a class is $\dfrac{f}{n} = \dfrac{\text{Class frequency}}{\text{Total of all frequencies}}$. Typically we may multiply this by 100 to get percentage. We can then construct a *relative frequency histogram.*

The *cumulative frequency* at a given class is the sum of all previous frequencies, including that of the current class.

An *ogive* is a line graph of cumulative frequency.

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.**

A *stem and leaf display* of a set over values has the characteristics of a histogram in the sense it detects frequencies,

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: mode < median < mean

Consider a sequence of histograms where, as the frequencies increase, the bin widths decrease. Eventually the histogram appears to be a continuous curve. The shape of this curve may be call a *distribution*. The horizontal (x) axis represent the possible values of the distribution, and the high indicates the relative frequency.

This curve may have one or more "bumps". Each bump represents a "local maximum". These are the various modes of the data. If there are two bumps, the distribution is called *bimodal*. If there is one bump, the data are essentially "mound shaped", and the distribution is *unimodal*.

If the distribution is unimodal, we may interpret the three averages as follows.

- *Mode:* The x value beneath the highest point on the curve.
- *Median:* The x value which splits the area under the curve in half.
- *Mean:* The x value of the "center of mass" of the area under the curve.

Homework: §3.1 # 3, 6, 16

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 $n = N$

Coefficient of Variation: $\text{COV} = 100 \cdot \dfrac{\sigma}{\mu}$ or $\text{COV} = 100 \cdot \dfrac{s}{\overline{x}}$

z-Value: $z = \dfrac{x-\mu}{\sigma}$ or $z = \dfrac{x-\overline{x}}{s}$

Theorem 1 (Chebechev's Theorem)**.** *Let $k \in \mathbb{N}$. The proportion of data that must lie within k standard deviation of the mean is*

$$1 - \frac{1}{k^2}.$$

At least 75.0% of the data is within 2σ of μ.
At least 88.9% of the data is within 3σ of μ.
At least 93.8% of the data is within 4σ of μ.

Homework: §3.2 # 4, 18

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

3.3.1. *Percentiles.* We often refer to variables as *distributions*; this refers to the fact that the values are distributed through the range.

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

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

3.3.2. *Quartiles.* 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

The *interquartile range* is $Q_3 - Q_1$.

Similarly, we can define *quintiles* and other similar breakdowns of the data.

3.3.3. *Box and Whisker Plots.* The *five-number summary* of a data set consists of the minimum, Q_1, the median, Q_3, and the maximum. We display how the data are distributed around these values with a *box and whisker plot.*

(1) Draw a vertical scale to include the lowest and highest data values.
(2) To the right of the scale draw a box from Q_1 to Q_3.
(3) Include a solid line through the box at the median level.
(4) Draw solid lines, called *whiskers*, from Q_1 to the lowest value and from Q_3 to the highest value.

Department of Mathematics, BASIS Mesa
Email address: paul.bailey@basismesa.org