

Chapter 1: Graphs and Statistics

Variability:

The whole subject of statistics is about studying variability – the fact that things are not the same from person to person, or time to time. In some sense, statistics as a subject can be defined as an organized and scientific approach to studying variability.

Definitions:

Population

The set of units (usually people objects transactions, or events) that we are interested in studying.

Census

A survey of every unit in the population of interest

Parameter

It is a numerical measurement of a variable describing a characteristic of a population.

Sample

A subset of the units in the population.

Statistic

It is a numerical measurement of a variable describing a characteristic of a sample.

Samples are used to learn about populations by calculating statistics which estimate parameters.

There are two types of statistics:

Descriptive

utilizes numerical and graphical methods to look for patterns in a data set, to summarize the information revealed in a data set, and to present that information in a convenient form.

Inferential

- utilizes sample data to make estimates, decisions, predictions, or other generalization about a large set of data.
- Five major components
 1. The population of interest
 2. One or more variables of interest
 3. The sample of population units
 4. The inference about the population
 5. A measure of reliability for the inference.

Reliability

A statement about the degree of uncertainty associated with statistical inference.

Types of Data

1. Archival Data or published source

Data collected for some other purpose other than the current investigator's. Published or stored and then reused to investigate new relationships

2. Designed Experiment

Researcher has control over the units in the study. Example is the controlled clinical trial. Patients are randomly assigned to a treatment groups and the **control** groups.

Control group receives the **placebo**. Only difference in patients is in the treatment of interest. This is when we can conclude causal relationships. Best when possible.

Smoking is the example of not possible. This is why Germany is so horrific.

3. Survey Data

Sampling a population of interest. Political polls, questionnaire analysis, Neilson TV ratings.

4. Observational Studies

When a researcher observes the outcome of interest from a population of interest.

Usually over a period of time. Example smoking studies, birth defects, cancer clusters.

The field of epidemiology deals with this. The big question is what is your control group? How are people matched? Has the problem with confounding. The cause of the differences in groups is not isolated, the variables are said to be confounded with each other. In studying things like heart disease in men, it is difficult to isolate smoking as a cause, because people who smoke a great deal also tend to drink alcohol a great deal. So these two effects are confounded. Thus it is very difficult to assess the effect of smoking alone on heart disease.

Measures of Center

Means

Interested in the center of a distribution. One method is with the means, averages.

We have two quantities: sample mean \bar{x} and population mean μ

If the data is a census, then we calculate μ directly. We use N to denote the sample size of a population. Then

$$\mu = \frac{\sum_{i=1}^N x_i}{N} = \frac{x_1 + x_2 + x_3 + \dots + x_N}{N}$$

If a sample is taken because a census is not possible, then we estimate μ with the sample mean \bar{x} . With a sample, the sample size is denoted with a small n .

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n}$$

The average of first 5 acting pulses yields

$$\mu = \frac{\sum_{i=1}^5 x_i}{5} = \frac{62+110+115+100+90}{5} = 95.4$$

The **median** of a data set is the middle number of a data set when the measurements are put into ascending order.

- if n is odd, the median m is the middle number
- if n is even, the median m is the average of the two middle numbers.

The mode is the number which occurs most frequently.

Dot Plots

- Draw a line from left to right that includes the lowest and highest values in the data
- Place a dot for each observation over the value that it is.

Histograms and Relative Histograms

1. Create Intervals or Classes
 - Make your own class intervals following a natural categorization
 - Make them of equal size
 - Generally want 5-12 intervals
2. Create frequency Table
 - Count the number of observations in each category
 - Calculate the percent for each category (leave as a decimal)
3. Draw a bar graph
 - Y-axis is percent (be sure to label)
 - X-axis is the variable of interest (labeled)
 - Be sure the right or left end point rule is displayed
4. If sample sizes of two groups are not close, relative frequency histograms are better
5. Best for large number of data observations

Stem and Leaf Displays

1. Choose the overriding unit as the stem (can be more than one digit)
 2. Write in stems from low at the top to highest at the bottom
 3. Take the next digit in the observations as the leaf
 - May need to use rounding
 - A leaf can only be one digit
 4. Do not want to have too few categories
 5. Can have split stem plots
 - Split into categories where leaf is digits 0-4 and 5-9
 - Split into categories where leaf is digits 0-1, 2-3, 4-5, 6-7, 8-9
 6. Can have back to back stem plots
 7. Ideal for summarizing data, but retaining the value of each observation.
- Invented by John Tukey

When data is skewed, it is better to use the median as a measure of central tendency than the mean.

Right skewed	median < mean
Symmetric	median = mean
Left skewed	mean < median

Quartiles – are the medians of the two groups separated by the median. [ie take the median of those numbers below Q2 and the median of the numbers above Q2].

The Inner Quartile Range is the middle 50% of the Data. Label $Q_1=25^{\text{th}}$ Percentile and $Q_3=75^{\text{th}}$ Percentile. The $IQR=Q_3-Q_1$

Box-Whisker Plots: An excellent tool to determine if there are any unusual observations – commonly called outliers. Also excellent for comparing two or more groups.

1. Find Q_1 , Q_2 , and Q_3
2. Calculate the Inner-Quartile Range ($IQR=Q_3-Q_1$)
3. Calculate the Inner Fence Posts (the Lower Inner Fence LIF and Upper Inner Fence UIF)
 - LIF= $Q_1-1.5*IQR$
 - UIF= $Q_3+1.5*IQR$
4. Calculate the Outer Fence Posts (the Lower Outer Fence LOF and Upper Outer Fence UOF)
 - LOF= $Q_1-3*IQR$
 - UOF= $Q_3+3*IQR$
5. Set up vertical scale going from lowest observation to highest observation
6. Draw horizontal lines for box at Q_1 , Q_2 , and Q_3 .
7. Draw whiskers to lowest and highest observations within the inner fence posts
8. Denote outliers with 0's for points between inner and outer fence posts
9. Denote extreme outliers with *'s for points beyond the outer fence

Measures of Spread

- **Range**
 - This is simply the highest data value - the lowest data value
- **IQR**
 - Q_3-Q_1 . These are medians of numbers below (or above) Q_2

Standard deviation

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}} = \sqrt{\frac{(62 - 95.4)^2 + (110 - 95.4)^2 + (115 - 95.4)^2 + (100 - 95.4)^2 + (90 - 95.4)^2}{5}}$$

$$= \sqrt{\frac{1115.56 + 213.16 + 384.16 + 21.16 + 29.16}{5}} = \sqrt{\frac{1763.2}{5}} = \sqrt{352.64} = 18.78$$

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

Take five heights – such as

66, 68, 64, 69, 63
66, 62, 60, 70, 72,

Both means are 66 inches. We have very different standard deviations: group1 is 2.28 and group2 is 4.56. There is much higher variation in group2 you can see there is more variation. Even though the means are the same, the standard deviations, are quite different. The higher the standard deviation, the more scattered the data.

Chebyshev's Inequality: Suppose that \bar{x} is the mean of a data set with standard deviation s . Choose a positive quantity, $k > 1$. Then at least $1 - 1/k^2$ of the data lie in the interval $(\bar{x} - k s, \bar{x} + k s)$

So if $k=2$, then we have $1 - 1/4 = 3/4$ so, at least $3/4$ of the data lie in the interval within 2 standard deviations.

Proof of Chebyshev's Inequality for $k=2$.

Define two sets of points –

$A = \{x_i \mid |x_i - \bar{x}| < ks\}$, $B = \{x_i \mid |x_i - \bar{x}| \geq ks\}$. Note the following:

$$\begin{aligned} s^2 &= \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} \geq \frac{\sum_{x \in B} (x_i - \bar{x})^2}{n-1} \geq \frac{\sum_{x \in B} (ks)^2}{n-1} = \frac{k^2 s^2 (\# \text{in } B)}{n-1} \\ &\geq \frac{k^2 s^2 (\# \text{in } B)}{n} = k^2 s^2 \% \text{in } B \end{aligned}$$

Thus

$$s^2 \geq k^2 s^2 \% \text{in } B = k^2 s^2 [1 - \% \text{in } A] \text{ which implies}$$

Divide by s^2 and k^2 to obtain

$$1/k^2 \geq [1 - \% \text{in } A] \text{ which implies}$$

$$\% \text{in } A \geq 1 - 1/k^2$$

So if we check the stock funds data. We find that 95% of the data is within two standard deviations of the mean – higher than the 75% that Chebyshev's implies, but Chebyshev's is a rough estimate.

Kinney text:

p. 4: 1-2

p. 10: 1

p. 17: 1, 3, 9, 11ab

p. 22 1, 3