

Descriptive Statistics

Variables

- A variable is a characteristic which may vary from person to person or from observation to observation.
- Example: Age, annual income, height, sex, marital status, blood group *etc.*
- If a variable takes numerical values, then it is called a quantitative variable. Otherwise (for non-numerical values), it is called a categorical variable.

Example: quantitative – age, income, family size
categorical – sex, marital status

- There are two types of categorical variables. If the categories for a categorical variable can be mean-

1

ingfully ranked, then that categorical variable is called an ordinal variable. Other categorical variables are non-ordinal variables. Examples of ordinal variable: letter grade, alert level in national security system, etc. Examples of non-ordinal variable: color of poinsettia, gender, marital status.

- There are two basic types of quantitative variables. The variables, which take only discrete values, are called discrete variables. Example: family size. The variables, which can take any value on an interval, are called continuous variables. Example: age, height.

2

Sample

- An observational unit is the smallest unit on which some observation is made.
- A sample is a collection of observational units.
- The sample size is the number of observational units in the sample.

Notational Convention

- Denote the variables by uppercase letters and
- the observations by lowercase letters.

Main Goals of Statistics

- Summarize the data: Helps visualize the data.
- Analyze the data and make inference about the population. This is more important.

3

Graphical Summarization

- For categorical variables:
 - Bar chart
 - Pie chart
- For quantitative variables:
 - Dotplot
 - Histogram
 - Stem and leaf plot
 - Boxplot

Numerical Summarization

- Frequency distribution table
- Measures of location or central tendency
 - “Where is the middle of the data?”

4

- Major ones are mean, median and mode.
- Measures of dispersion or spread
 - “How much variation is there in the data?”
 - Major ones are range, variance, standard deviation and inter-quartile range (IQR).

Frequency Distribution Table

- Frequency of a value or category is the number of occurrences of that value or category.
- Relative frequency of a value or category

$$= \frac{\text{Frequency of that value or category}}{\text{Sample size or total frequency}}$$
- Percentage frequency or percentage of a value or category

$$= \text{Relative frequency of that value or category} \times 100\%$$
- The advantage of relative frequency or percentage

5

frequency over frequency is that we can compare the distribution of two data sets if we use relative or percentage frequencies.

- Raw data can be summarized by a frequency distribution table.
- A frequency distribution table is a list of distinct values or categories that appear in the data with the frequency, relative frequency or percentage frequency for each value or category.
- Useful to summarize a large data set, especially for a discrete or categorical variable.

Grouped Frequency Distribution Table

- When there is a large number of distinct values in a data set, the above way to make a frequency distribution table is not very useful. It happens

6

especially for a continuous variable.

- We can make some class intervals and count the frequency for each class interval.
- Each class interval has an upper boundary and a lower boundary.
- Conventionally the upper boundary is excluded from a class interval. So, the class interval 10 - 15 means at least 10 but strictly less than 15.
- A table is formed with the class intervals and their frequencies or relative frequencies or percentages.
- For computing summary statistics from a grouped frequency distribution table, regard as if all the data points in each class interval are at the mid-point of that class interval.

7

Bar Diagram or Bar Chart

- Useful mainly for qualitative or categorical variable.
- For each category a bar of constant width and height proportional to its frequency or relative frequency or percentage in that category or group is drawn.
- Usually bars are equally spaced and not contiguous.

Dotplot

- Useful for small quantitative (especially discrete) datasets.
- Draw a number line covering the range of the data and put a dot above the number line at each observation.
- If there are two or more observations with the same value, stack the dots on top of each other.

8

- It gives an idea about the distribution of the data.

Histogram

- It is drawn based on a grouped frequency distribution table.
- Histogram is made up of a set of contiguous rectangular blocks.
- Class intervals are drawn on the horizontal axis.
- Each block is drawn in such a way that the **area** of the block is proportional to the **percentage** (or frequency or relative frequency) in the corresponding class interval.
- In other words, height of the blocks are proportional to the frequency densities where

frequency density of a class interval

$$= \frac{\text{percentage (or freq. or rel freq.) in that class interval}}{\text{length of the class interval}}$$

- On the vertical axis one has the density scale, i.e., the units on the vertical axis are % (or frequency or relative frequency)/units on the horizontal axis.
- In a histogram, the height of a block represents crowding – percentage (or frequency or relative frequency) per horizontal unit.
- Total area of the blocks is 100% for a percentage type histogram, total frequency for a frequency type histogram, and 1 for relative frequency type histogram.

Stem and Leaf Plot

- If the original data are not in a grouped form, making a histogram may require grouping the data and the original data are not preserved in the histogram. Stem and leaf plot preserves the original data.
- The stem and leaf plot requires either all observations to be integer or all observations to have the same number of decimal places.
- For any observation, the rightmost digit forms the leaf and the other digits form the stem.
- The stems are written in a column in increasing (or decreasing) order. A vertical line is drawn to the right of the stem column.
- The numbers having the same stem are grouped

and the leaves for those numbers are written on the right side of the corresponding stem to the right of the vertical line.

- In practice, first the stems are formed and then the numbers in the data set are considered one by one and the leaf is written to the right of the appropriate stem.
- The stem and leaf plot has many variations such as considering two rightmost digits for the leaf part, *etc.*

Arithmetic Mean or Average

- The average of a list of numbers is sum of the numbers in the list divided by the number of numbers in the list, i.e., if x_1, x_2, \dots, x_n are the observations, then the arithmetic mean (denoted by \bar{x}) is

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i.$$

- Provides “balance point” of data.
- Not robust, easily influenced by outliers.
- Notation for population mean: μ ,
notation for sample mean: \bar{x} or \bar{X} .

Median

- Median is the middle value. It is a value such that half of the data points are below this value and the other half are above this value.

13

- In a histogram a median is a point on the horizontal axis such that the vertical line passing through the median divides the histogram into two parts of equal area, 50% each.
- Technical definition: A median of a set of numbers is a value such that at least 50% of the numbers are less than or equal to that value and at least 50% are greater than or equal to that value.
- Advantage: Robust, usually not influenced by outliers.
- Disadvantage: difficult to compute for a large data set and unlike arithmetic mean lacks some nice algebraic properties.
- In general median is not unique, but we will use the following method to get a unique median.

14

How to find the median of a list of values?

1. Sort the values in increasing order.
2. Suppose the number of values in the list is n . There are two cases.
 - If n is odd, then the median = $\frac{n+1}{2}$ -th ordered value.
 - If n is even, then the median = $\frac{1}{2}$ [$\frac{n}{2}$ -th ordered value + ($\frac{n}{2} + 1$)-th ordered value].

Mode

- A mode is the most frequent value in a data set.
- Mode is not unique.
- Mode is a good measure of the center only if the distribution is bell shaped.

15

Shape of a Distribution

- The distribution may have different shapes.
- A distribution is **unimodal** if it has only one peak.
- It is called **bimodal** if it has two peaks.
- If the left half of a distribution is the mirror image of the right half, it is **symmetric**.
- If the distribution's right tail is longer than its left tail, it is **positively skewed** or **right skewed**.
- If its left tail is longer than right tail, it is **negatively skewed** or **left skewed**.
- If the distribution is symmetric, mean = median, if right skewed, mean > median and if left skewed, mean < median.

16

Quantiles and Percentiles

- q th quantile separates the bottom q proportion of data from top $(1 - q)$ proportion of data.
- p th percentile separates the bottom $p\%$ from the top $(100 - p)\%$. The p th percentile is $\frac{p}{100}$ th quantile.
- If some number is in k -th percentile, then the **percentile rank** of that number is $k\%$.
- A percentile is a score. A percentile rank is a percent.
- Some important percentiles:
First quartile (Q1): 25th percentile, median or second quartile (Q2): 50th percentile, third quartile (Q3): 75th percentile.

17

How to find q th quantile?

1. Sort the values in increasing order.
2. Suppose there are n values. There are two cases.
 - If nq is not an integer, then round it up to the next higher integer k . Then the q th quantile is the k th ordered value.
 - If nq is an integer, then the q th quantile is $\frac{1}{2}[(nq)$ th ordered value + $(nq + 1)$ th ordered value].

18

Measures of Spread

Range

- Range is the difference between the largest and the smallest values.
- It is easy to compute but very much influenced by the outliers.

Inter-quartile Range (IQR)

- $IQR = Q3 - Q1$
- Measures how far the median of the top half from the median of the bottom half is.
- Not influenced by outliers unlike range or variance.
- Not very commonly used.

19

Variance and Standard Deviation

- Might want to measure dispersion as average distance of data points from the mean:

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

- But this is always 0. One might use mean absolute deviation

$$\frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|$$

but this does not have nice algebraic properties.

- So look at the average squared distance:

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

- The above is the population variance if x_1, \dots, x_n constitute the population.
- The sample variance is defined with a little modi-

20

fication when the x_1, \dots, x_n constitute a sample:

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

- The difference between sample and population variances is the use of $n-1$ or n in the denominator.
- As n grows, the two calculations become closer.
- Variance is 0 if and only if all data points have the same value.
- **Standard deviation** (SD) is square root of variance.
- Advantages: Mathematically useful, frequently used.
- Drawbacks: Easily influenced by outliers, difficult to calculate by hand.

Another measure of spread is the **coefficient of variation** (CV) which is defined as $CV = \frac{SD}{\text{mean}}$.

21

the median and Q1 is greater than the difference between the median and Q3, or the difference between minimum and Q1 is greater than the difference between Q3 and the maximum, then the distribution is left skewed, if less then right skewed, if equal then symmetric.

- We can also compare two or more datasets by drawing parallel box plots with a common single scale.

Outlier

- An outlier is a value in a data set that is much different from almost all other values in the data set.
- The outliers are important for many reasons. It may indicate a problem with experimental method or a mistake in recording the data or about a special

23

Box Plot

- Minimum, Q1, median, Q3 and maximum of a data set are referred to as the **five-number summary**.
- A box plot is a graphical representation of the five-number summary.
- A vertical box plot is formed by drawing three horizontal line segments at the three quartiles and then by drawing two vertical line segments to join the horizontal line segments. Then two vertical line segment, one from the first quartile to the smallest value and another from the third quartile to the largest value, are drawn.
- The box plot gives a rough idea about the symmetry of the distribution. If the difference between

22

circumstance.

- The above definition of outlier is subjective. To make it objective, we define lower and upper fences of a data set.
- Lower fence of the data = $Q1 - 1.5 \times IQR$.
Upper fence of the data = $Q3 + 1.5 \times IQR$.
- A value in the data set is an outlier if it is greater than the upper fence or less than the lower fence of the data.

Modified Box Plot

- A modified box plot is a box plot where the outliers are drawn as separate points.
- The outliers are marked by asterisks.
- The line segment from Q1 to minimum in the reg-

24

ular box plot is replaced by a line segment from Q1 to the minimum non-outlying value.

- The line segment from Q3 to maximum in the regular box plot is replaced by a line segment from Q3 to the maximum non-outlying value.

Empirical Rule

- About 68% of the data fall within 1 SD from the mean
- About 95% of the data fall within 2 SD from the mean
- Almost all (about 99.7%) of the data fall within 3 SD from the mean

Units

- All measures of location have the units same as the units of the variable.

25

- The range, mean absolute deviation, standard deviation and IQR have the units same as the units of the variable. The units of the variance are the squared units of the variable.

Effects of Shift and Scaling

- If a constant is added to each of the observations in the sample,
 - the same constant is added to the arithmetic mean, median, mode and quantiles,
 - all measures of spread remain unchanged,
- If all the observations in a sample are multiplied by a constant,
 - the arithmetic mean, median and mode are multiplied by the same constant; if the multiplier

26

constant is positive, then the quantiles are also multiplied by the same constant,

- all measures of spread discussed earlier except variance are multiplied by the absolute value of the multiplier constant. Variance is multiplied by the square of the multiplier constant.

- Any linear transformation of data is a combination of the above two procedures and we can find the effects of any linear transformation on the measures of location and spread. Suppose $y_i = ax_i + b$ for some constants a and b . Then

1. $\bar{y} = a\bar{x} + b$.
2. Median of $y = a \cdot$ median of $x + b$.
3. Mode of $y = a \cdot$ mode of $x + b$.
4. q th quantile of $y = a \cdot q$ th quantile of $x + b$ if

27

$a > 0$.

5. q th quantile of $y = a \cdot (1 - q)$ th quantile of $x + b$ if $a < 0$.
6. $Range(y) = |a| \cdot Range(x)$.
7. $IQR(y) = |a| \cdot IQR(x)$.
8. Mean absolute deviation of $y = |a| \cdot$ mean absolute deviation of x .
9. $Var(y) = a^2 Var(x)$.
10. $SD(y) = |a| \cdot SD(x)$.

28

Population and Sample

- A Population is the collection of individuals or objects about which we want to make some conclusion.
- A sample is a part of a population.
- We cannot make the observation on all individuals in the population. That's why a sample is chosen and the observations are made on the individuals in the sample.
- We make conclusions about the population based on the observations made on the sample. This is called statistical inference.
- Example: If we want to find the proportion of people in the US who are cigarette smokers, it is not easy to ask every person in the US whether he or

29

she is a smoker. But we can take a sample from the US population and ask the question to each person in the sample. Then we can make conclusion about the proportion of smoker in the US population based on what we observe in the sample.

- To make this kind of statistical inference the sample has to be a **good representative** of the population.
- There are two considerations:
 - Sample size: If the sample size is too small it may not have enough observational units to represent all different types of unit in the population. Also the sampling error is more with smaller sample size.
 - Bias: Systematic error in sampling. This is more

30

important than the sample size. We will discuss these in more details in chapter 8.

- Remember that the conclusions that are made about the population from the observed sample is valid only for the population from which the sample is selected.

Parameter and Statistic

- A parameter is a numerical measure of the population. Example: Proportion of people having cancer in the US.
- A statistic is a numerical measure of the sample. Example: Proportion of people having lung cancer in a random sample taken from the US population.
- Notice that the word “statistics” has two meanings
 - the subject statistics and as the plural of statistic.

31

- Some important parameters and statistics and their notations:

Measure	Statistic (Sample Value)	Parameter (Population Value)
Proportion	\hat{p}	p
Mean	\bar{y} or \bar{x}	μ
Standard Deviation	s	σ

32