Describing Distributions

With Numbers

April 18, 2012

- Summary Statistics.
- Measures of Center.
- Percentiles.
- Measures of Spread.
- A Summary Statement.
- Choosing Numerical Summaries.

1.0 What Are Summary Statistics?

- Summary statistics are STATISTICS that summarize some aspect of the distribution of a variable.
- The most usual summary statistics describe the CENTER and the SPREAD of a distribution.
- In some cases, we also want to summarize the percentiles of a distribution.

1.0 Summary Statistics

- The purpose of calculating summary statistics are two-fold:
  - Compact reporting of the distribution of a variable.
  - Easy comparison between two distributions.
- Important considerations for calculating summary statistics:
  - Easily interpretable.
  - Robustness or stability.
- We will discuss:
  - how the statistics are defined.
  - how to compute.
  - how to interpret.
  - how to choose.
  - how to "guesstimate"
- Our focus is only quantitative variables.

2.0 Describing Distributions with Numbers

Here are G.P.A.s of 15 students from one section.

3.2	3.7	3.6	3.7	3.3
3.3	3.8	3.2	3.0	3.5
2.5	3.3	3.5	2.3	3.5

The decimal point is at the |

2 | 3
2 | 5
3 | 022333
3 | 5556778

3.0 Measures of Center: The Mean

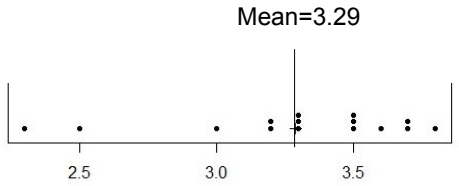
The MEAN $\bar{x}$ (x-bar) of a set of observations is their average. To find the mean of n observations, add the values and divide by n.

$$\bar{x} = \frac{\text{sum of observations}}{\text{n}}$$

3.0 Measures of Center: The Mean

For the G.P.A. data, the mean G.P.A. is:

$$\begin{aligned} \bar{x} &= \frac{3.2 + 3.7 + 3.6 + \cdots + 3.5}{15}, \\ &= \frac{49.4}{15}, \end{aligned}$$



3.1 Measures of Center: The Median

The MEDIAN is the mid-point of the distribution, the number such that half (or more) of the observations are at the median or bigger and half (or more) are at the median or smaller.

3.1 Measures of Center: The Median

For the G.P.A. data, the median G.P.A. is:

1. Order the observations from smallest to largest:

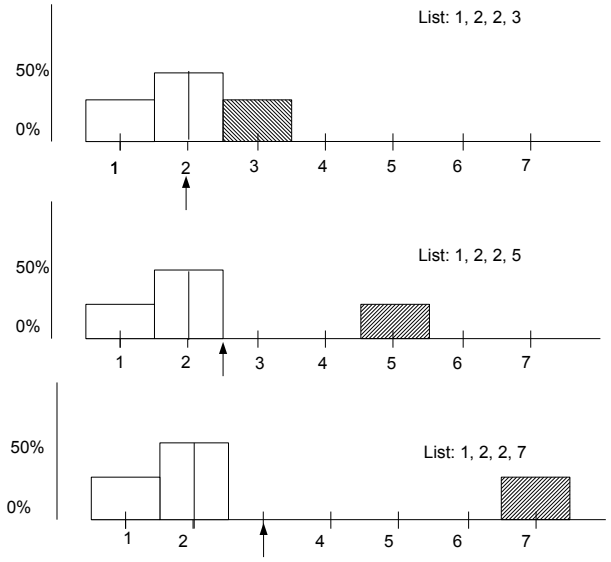
2.3	2.5	3.0	3.2	3.2
3.3	3.3	**3.3**	3.5	3.5
3.5	3.6	3.7	3.7	3.8

2. The median G.P.A. is 3.3 because eight numbers out of 15 are 3.3 or more and eight numbers out of 15 are 3.3 or less.

3.1 Measures of Center: The Median

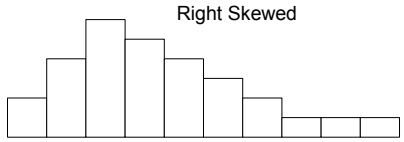
Calculate the median for the following list of numbers:
8, -3, 5, 1, 4, -1.

3.2 Measures of Center and Histograms

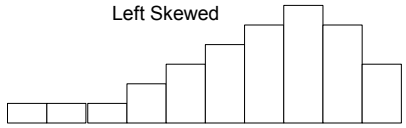


- The average (arrow) moves to the right along with the largest observation.
- The median (straight line at 2) stays where it is.

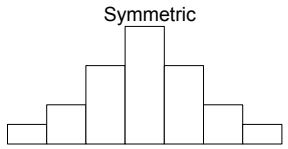
3.2 Measures of Center and Histograms



Average is bigger than median



Average is smaller than the median



Average is equal to the median

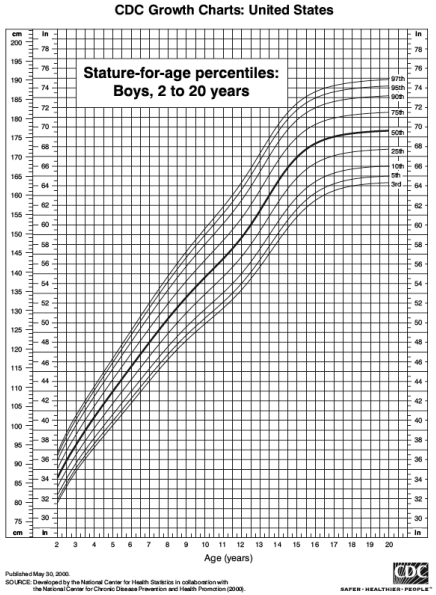
4.0 Percentiles

Definition

The cth percentile of a distribution is defined so that c% (or more) of the observations are at or bigger than it and 100-c% (or more) of the observations are at or smaller than it.

- The median is the 50th percentile of a distribution.
- The 25th percentile of a distribution is called the lower quartile $Q1$.
- The 75th percentile of a distribution is called the upper quartile $Q3$.

4.0 Percentiles



4.1 Calculating Percentiles

- Back to the G.P.A.

2.3	2.5	3.0	3.2	3.2
3.3	3.3	**3.3**	3.5	3.5
3.5	3.6	3.7	3.7	3.8

- Median = 3.3 (shown in box).
- $Q1$ = median of the lower 8 observations = 3.2. lower half including the median
- $Q3$ = median of upper 8 observations = 3.5, 3.6 or 3.55. upper half including the median

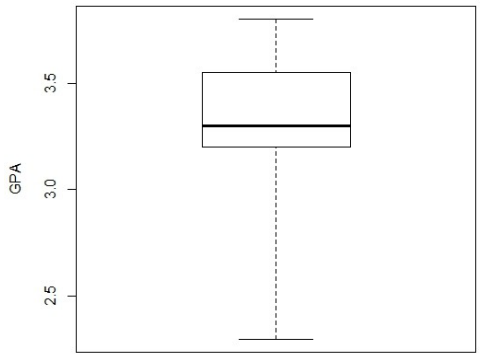
4.2 The Five Number Summary

Definition

The FIVE NUMBER SUMMARY of a distribution consists of the smallest observation, the first quartile, the median, the third quartile, and the largest observation, written in order from smallest to largest.

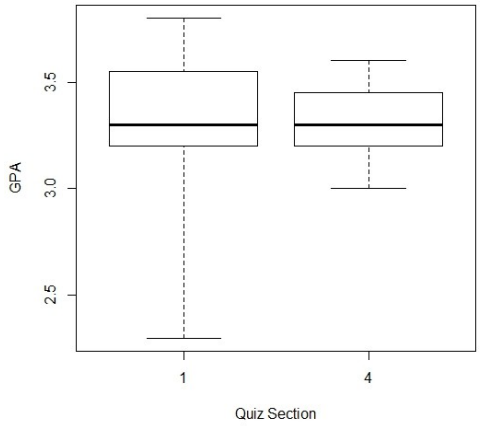
- These five numbers offer a complete summary of a distribution.
- It is typically represented as a box-and-whisker plot.

4.3 The Box-and-Whisker Plot



- A central box spans the quartiles.
- A line in the box marks the median.
- Lines extend from the box to the smallest and largest observation.

4.3 The Box-and-Whisker Plot: Comparing Distributions



- First compare the medians.
- The quartiles show the spread of the middle half of the data.

5.0 Measures of Spread

- Maximum - Minimum. range of the distribution
- Inter-quartile range (I.Q.R.) = $Q3$ - $Q1$.
  - The I.Q.R. is the range of the middle 50% of a distribution.
  - Some people call a data point an outlier if it is more than 1.5 times I.Q.R below $Q1$ or above $Q3$. 1.5 I.Q.R. rule
- Standard Deviation (S.D.)

Definition

The STANDARD DEVIATION measures the average distance (or deviation) of the observations from their arithmetic mean.

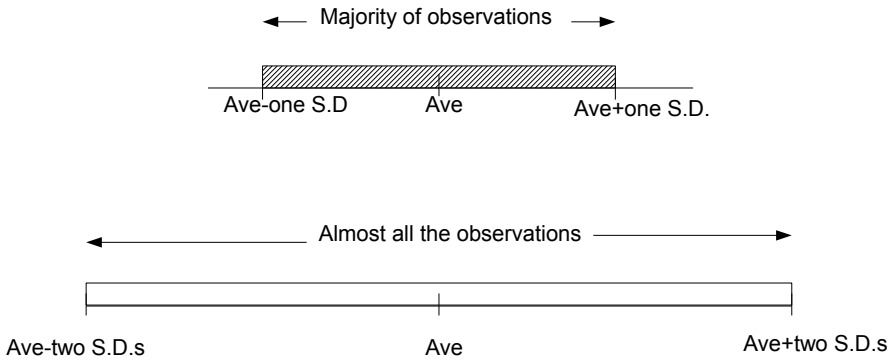
5.1 Calculating Standard Deviations

Find the S.D. for this list of numbers: 2, -6, 12, 4, 3.

- Step 1: Find the average for the list of numbers. The answer is 3.
- Step 2: Find the deviation of each value from this average: -1, -9, 9, 1, 0.
- Step 3: The S.D. tells the "average size" of a deviation.
  - Step 3.1: Square each deviation: 1, 81, 81, 1, 0. SQUARE
  - Step 3.2: Calculate the average of this list but dividing by $(n-1)$ instead of n: The answer is 41. MEAN
  - Step 3.3: Take the square-root of 41. The answer is 6.4. ROOT
- The standard deviation is 6.4. has the same units as the list of numbers

5.2 Interpreting Standard Deviations

The standard deviation (S.D.) says how far numbers on a list are from their average. Most entries of the list will be somewhere around one S.D. from the average. Very few will be more than two or three S.D.s away.



5.2 Guesstimating Standard Deviations

Each of the following lists has an average of 50. For which one is the standard deviation the biggest? smallest?

1. 0, 20, 40, 50, 60, 80, 100.
2. 0, 48, 49, 50, 51, 52, 100.
3. 0, 1, 2, 50, 98, 99, 100.

5.2 Guesstimating Standard Deviations

Below are sketches of histograms for three lists of numbers. Match the sketch with the description that fits.

(I) ave $\approx$ 3.5, S.D. $\approx$ 1

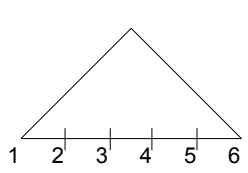
(II) ave $\approx$ 3.5, S.D. $\approx$ 0.5

(III) ave $\approx$ 3.5, S.D. $\approx$ 2

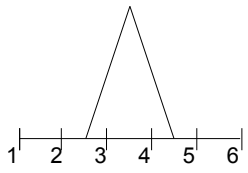
(IV) ave $\approx$ 2.5, S.D. $\approx$ 1

(V) ave $\approx$ 2.5, S.D. $\approx$ 0.5

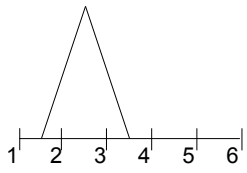
(VI) ave $\approx$ 4.5, S.D. $\approx$ 0.5



(a)



(b)



(c)

5.2 Guesstimating Standard Deviations

Household size in the U.S. has a mean of 2.5 people approximately. Which of these numbers would be a good guess for the standard deviation? 0.014, 0.14, 1.4 and 14?

5.3 A Short-cut for Calculating S.D.s

- There is a very useful short-cut for calculating the S.D. of a list with only two different numbers, a big one and a small one. (Each number can be repeated many times).
- In this case, the S.D. can be calculated using:

$$\left(\begin{matrix}\text{big}\\ \text{number}\end{matrix} - \begin{matrix}\text{small}\\ \text{number}\end{matrix}\right) \times \sqrt{\begin{matrix}\text{fraction with}\\ \text{big number}\end{matrix} \times \begin{matrix}\text{fraction with}\\ \text{small number}\end{matrix}}.$$

- Find the S.D. of the list of numbers: 1, -2, -2.
- Find the S.D. of the list of numbers: -1, -1, -1, 1.
- Can you use the short cut to calculate the standard deviation of the list: 1, 2, 3, 4?

6.0 A Summary Statement

The distribution of math S.A.T. scores for a subset of STAT 220 students is describe shape here . The test scores range from minimum to maximum, and tend to be around average, give or take standard deviation, or so.

7.0 Choosing Numerical Summaries

Advice

Use means and standard deviations for distributions that are roughly symmetric and with no outliers. Use the five number summary otherwise.