Statistics Class 2
Characterizing Data

January 20, 2010

Jobayer Hossain, Ph.D. & Tim Bunnell, Ph.D.

Nemours Bioinformatics Core Facility

Descriptive Statistics

- Summarize or characterize a set of data in a meaningful way
- A set of data is a collection of individual observations.
- Two broad types of data:
  - **Numerical** - integer or real valued data such as height, age, blood pressure, a measured enzyme level, etc.
  - **Categorical** - observations drawn from a finite set of discrete categories or groups. For example sex, eye color, highest degree obtained, etc.

Distributions

- Describe a dataset in terms of the frequency with which specific values or ranges of values are observed, that is, how observations are ***distributed*** over the possible range of values
- **Example 1**: eye colors in a random sample of 26 preschoolers in a daycare center:

Eye Color	Brown	Blue	Hazel	Green
Frequency	13	7	4	2

Distributions continued

Example 2 - The age distribution of a random sample of 26 preschool students (note that each category can be expressed as either a discrete age in years, or as a range of values for age in months):

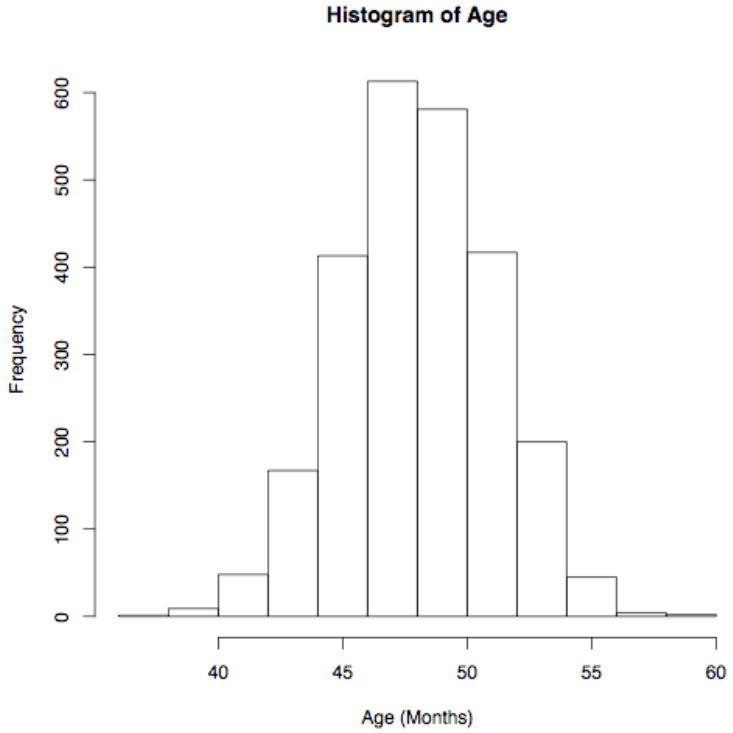
Age	Years	1	2	3	4	5	6
	Months	12-23	24-35	36-47	48-59	60-71	72-84
Frequency		5	3	7	5	4	2

Describing Distributions

- Graphical or qualitative
  - Allows visual inspection of distribution
  - Ex: Histogram or density plot
- Numerical or quantitative
  - Allows numerical comparison of distributions
  - Ex: mean, variance, skewness
- Best to use both so you can see what the numbers are telling you!

Distributions

Age (months)	Frequency
36-37	1
38-39	9
40-41	48
42-43	167
44-45	413
46-47	613
48-49	581
50-51	417
52-53	200
54-55	45
56-57	4
58-59	2

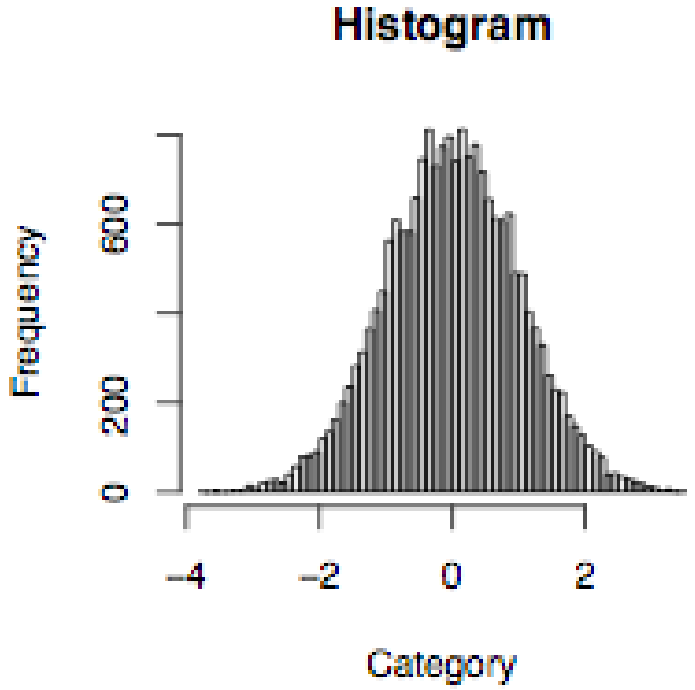


Mean=47.3; S.D.=6.1

A more detailed distribution

Sampled normal distribution

- Mean 0.006
- S.D. 1.003
- Median .003
- Minimum -3.7
- Maximum 3.4
- 1st Quartile -.678
- 3rd Quartile .690



Numerical Properties

Three properties are commonly used to describe distributions:

1. Central Tendency - mean, median, mode, etc.
2. Spread of variability - variance or standard deviation
3. Shape - skewness, kurtosis, etc.

Central Tendency Measures

- **Commonly used methods:** mean, median, and mode etc.
- **Mean**: Summing up all the observation and dividing by number of observations. Mean of 20, 30, 40 is (20+30+40)/3 = 30.
- **Mode**: The value that is observed most frequently. The mode is undefined for sequences in which no observation is repeated. A variable with single mode is unimodal, with two modes is bimodal. E.g. the mode of 2, 3, 3, 4 is 3 and the modes of 2, 3, 3, 4, 5, 5, 6 of are 3 and 5.

Central Tendency Measures (cont.)

Median: The middle value in an ordered sequence of observations. That is, to find the median we need to order the data set and then find the middle value. In case of an even number of observations the average of the two middle most values is the median. For example, to find the median of {9, 3, 6, 7, 5}, we first sort the data giving {3, 5, 6, 7, 9}, then choose the middle value 6. If the number of observations is even, e.g., {9, 3, 6, 7, 5, 2}, then the median is the average of the two middle values from the sorted sequence, in this case, (5 + 6) / 2 = 5.5.

Mean V.S. Median

The median is less sensitive to outliers (extreme scores) than the mean and thus a better measure than the mean for highly skewed distributions, e.g. family income. For example mean of 20, 30, 40, and 990 is (20+30+40+990)/4 =270. The median of these four observations is (30+40)/2 =35. Here 3 observations out of 4 lie between 20-40. So, the mean 270 really fails to give a realistic picture of the major part of the data. It is influenced by extreme value 990.

Mean V.S. Median

- Mean
  - "Best" estimate of expected sample value
  - Easily calculated
  - Well understood
- Median
  - Less sensitive to outliers
  - Less sensitive to distribution shape

Variability Measures

- **Variability (or spread or dispersion)** measures the amount of scatter in a dataset.
- **Commonly used methods**: *range*, *variance*, *standard deviation*, *interquartile range*, *coefficient of variation etc*.
- **Range**: The difference between the largest and the smallest observations. The range of 10, 5, 2, 100 is (100-2)=98. It's a crude measure of variability.

Variability Measures

Variance: The variance of a set of observations is the average of the squares of the deviations of the observations from their mean. Variance of 5, 7, 3? Mean is (5+7+3)/3 = 5 and the variance is

$$\frac{(5-5)^2+(3-5)^2+(7-5)^2}{3-1}=4$$

Standard Deviation (SD) : Square root of the variance. The SD of the above example is 2.

If the distribution is bell shaped (symmetric), then the range is approximately (SD x 6)

Variability Measures

- **Quartiles**: Quartiles are values that divides the sorted dataset in to four equal parts so that each part contains 25% of the sorted data
- The first quartile (Q1) is the value from which 25% observations are smaller and 75% observations are larger. This is the median of the 1st half of the ordered dataset.
- The second quartile (Q2) is the median of the data.
- The third quartile (Q3) is the value from which 75% observations are smaller and 25% observations are larger. This is the median of the 2nd half of the ordered dataset.

Variability Measures

- An example with 15 numbers

  3 6 7 11 13 22 30 40 44 50 52 61 68 80 94

  Q1 Q2 Q3

  The first quartile is Q1=11. The second quartile is Q2=40 (This is also the Median.) The third quartile is Q3=61.

- **Inter-quartile Range**: Difference between Q3 and Q1. Inter-quartile range of the previous example is 61- 40=21. The middle half of the ordered data lie between 40 and 61.

Variability Measures

- **Deciles**: If data are ordered and divided into 10 parts, then cut points are called Deciles
- **Percentiles**: If data are ordered and divided into 100 parts, then cut points are called Percentiles. 25th percentile is the Q1, 50th percentile is the Median (Q2) and the 75th percentile of the data is Q3.

  For example, Consider the following ordered set of data: 3, 5, 7, 8, 9, 11, 13, 15.

  25^{th} percentile = 5 + (7-5)x.25= 5.5

Variability Measures

- **Coefficient of Variation (CV)**: The standard deviation of data divided by it's mean. It is usually expressed in percent.

$$\text{Coefficient of Variation} = \frac{\text{Standard deviatin}}{\text{Mean}} \times 100$$

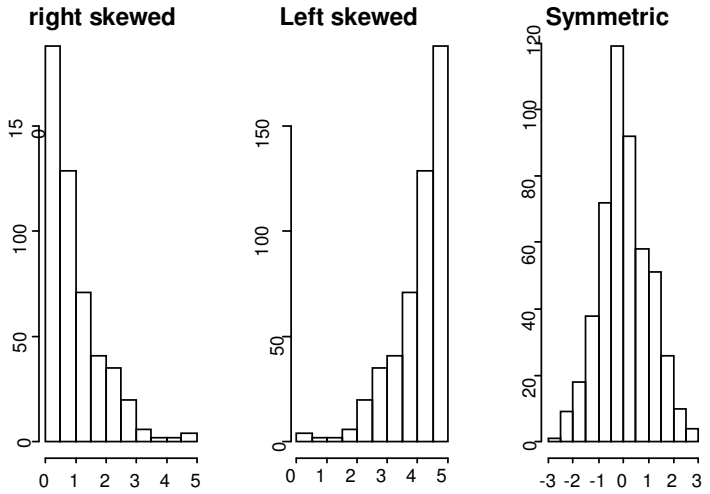
E. g. Mean and standard deviation of 5, 7, and 3 are 5 and 2 respectively. The CV of this data is (2/5)x 100= 40%

Distribution Shape Measures

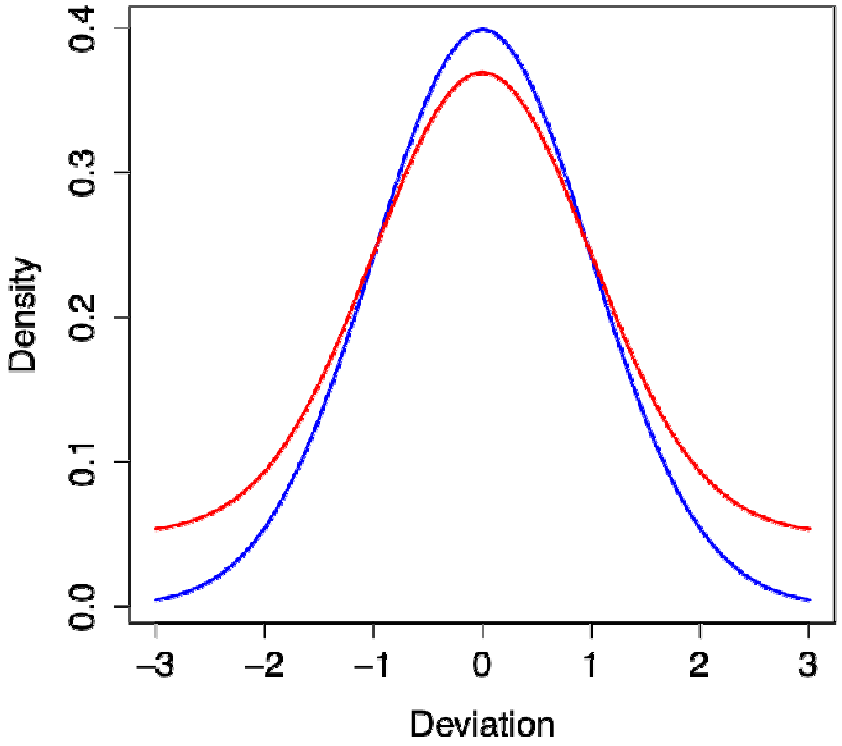
- Shape of data is measured by
  - Skewness
  - Kurtosis

Skewness

- Measures of asymmetry of data
  - Positive or right skewed: Longer right tail
  - Negative or left skewed: Longer left tail
  - Symmetric: Bell shaped
  - Graphically:



Kurtosis



Kurtosis relates to the relative flatness or peakedness of a distribution. A standard normal distribution (blue line: $\mu = 0$; $\sigma = 1$) has kurtosis = 0. A distribution like that illustrated with the red curve has kurtosis > 0 with a lower peak relative to its tails.

Five Number Summary

- **Five Number Summary**: The five number summary of a distribution consists of the smallest (Minimum) observation, the first quartile (Q1), the median(Q2), the third quartile, and the largest (Maximum) observation written in order from smallest to largest.

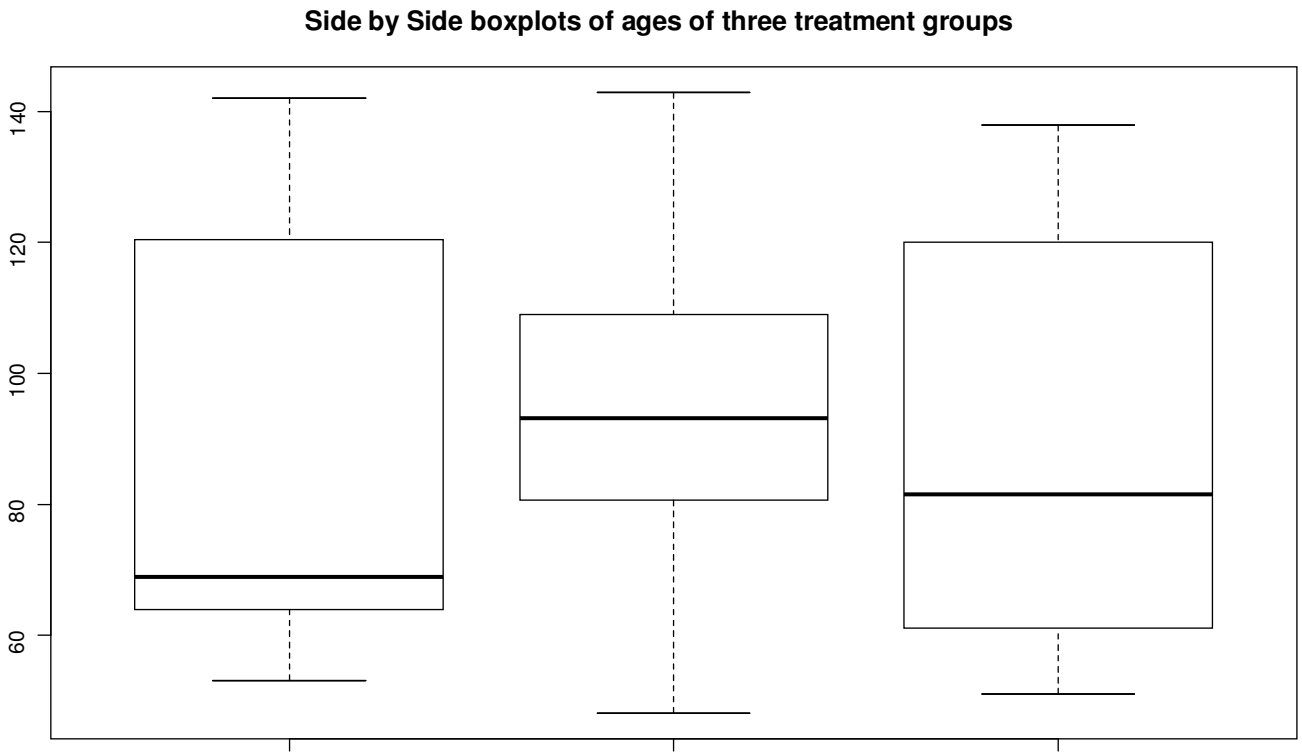
Choosing a Summary

- The five number summary is usually better than the mean and standard deviation for describing a skewed distribution or a distribution with extreme outliers. The mean and standard deviation are reasonable for symmetric distributions that are free of outliers.
- In real life we can't always expect symmetry of the data. It's a common practice to include number of observations (n), mean, median, standard deviation, and range as common for data summarization purpose. We can include other summary statistics like Q1, Q3, Coefficient of variation if it is considered to be important for describing data.

Graphical Presentation

- Boxplot :
  - A boxplot is a graph of the five number summary. The central box spans the inter quartile range.
  - A line within the box marks the median.
  - Lines extending above and below the box mark the smallest and the largest observations (i.e. the range).
  - Outlying samples may be additionally plotted outside the range.

Side by Side Boxplots



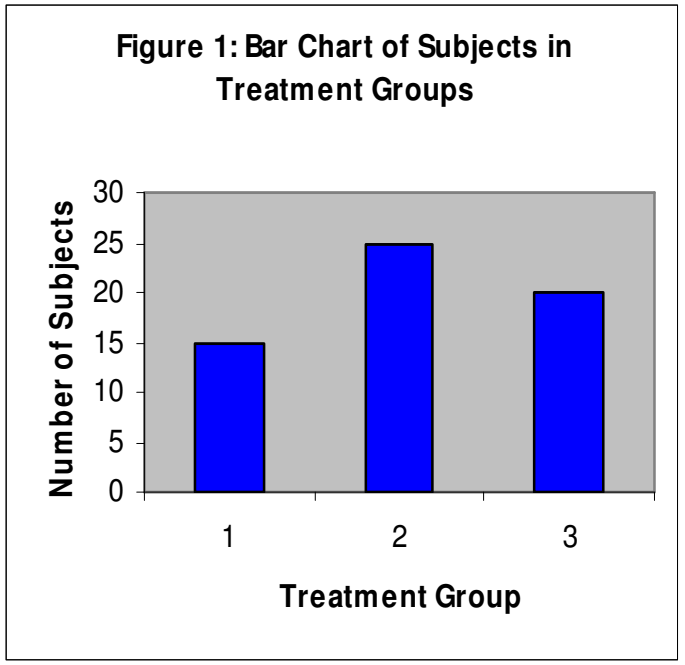
Additional Graphical Presentations

- Many other common graphical presentations of data
- Bar graph - histogram with only a few categories
- Pie chart - good to describe distribution of categories within a closed domain.
- Line graph - good for illustrating functional relations among means (topic for a later lecture)

Data Description: Categorical Variable

The distribution of 60 patients in three treatment group

Treatment Group	Frequency	Proportion	Percent (%)
1	15	(15/60)=0.25	25.0
2	25	(25/60)=.417	41.7
3	20	(20/60)=0.333	33.3
Total	60	1.00	100

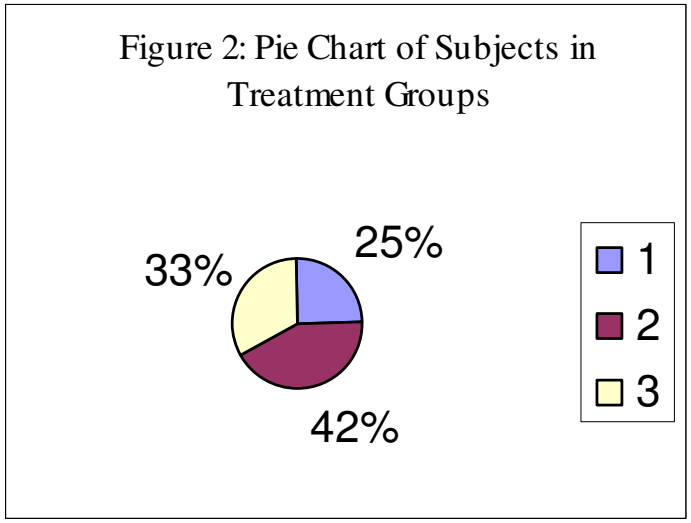


Bar Diagram: Lists the categories and presents the percent or count of individuals who fall in each category.

Data Description: Categorical Variable

The distribution of 60 patients in three treatment group

Treatment Group	Frequency	Proportion	Percent (%)
1	15	(15/60)=0.25	25.0
2	25	(25/60)= 0.417	41.7
3	20	(20/60)=0.333	33.3
Total	60	1.00	100



Pie Chart: Lists the categories and presents the percent or count of individuals who fall in each category.

SPSS Demo

- Import data default in to SPSS
- Calculate summary statistics of numerical and categorical variables
- Graphs: Bar chart, Pie Chart, Boxplots and Histograms