MSCA 31000 - Introduction to Statistical Concepts

Chapter 1

Q1: A teacher wishes to know whether the males in his/her class have more conservative attitudes than the females. A questionnaire is distributed assessing attitudes and the males and the females are compared. Is this an example of descriptive or inferential statistics?

Answer: This is an example of descriptive statistics. The questionnaire output is used to describe and summarize data at hand and not to generalize results beyond the existing population (male and female students in the class).

Q3: If you are told that you scored in the 80th percentile, from just this information would you know exactly what that means and how it was calculated? Explain.

Answer: We cannot know exactly what 80th percentile means exactly in this question. This is because we have three definitions for percentile and the question does not state what method is being used.

1. Definition 1: lowest score that is greater than 80% of the scores.
2. Definition 2: smallest score that is greater than or equal to 80% of the scores.
3. Definition 3: weighted average of the percentiles computed according to the first two definitions.

1 and 2 will give different results depending on the number of students, scoring system and the ordering.

Q5: Give an example of an independent and a dependent variable.

Answer: For a system of currency conversion from US Dollar to Japanese Yen using an equation: Y = D * c

where,
Y = value in Japanese Yen
D = value in US Dollar
c = conversion factor

Independent variable: US Dollar amount = D
Dependent variable: Japanese Yen amount = Y

Q7: Specify the level of measurement used for the items in Question 6 (bullet items)

Answer:

- Rating of the quality of a movie on a 7-point scale
    - Ordinal scale

    - Potential levels: 7 points from 1 to 7, 0 to 6 and so on.
- Age
    - Interval scale
    - Potential levels: 1, 2, 3, 4 etc.
- Country you were born in
    - Nominal scale
    - Potential levels: India, Japan, USA, UAE
- Favorite Color
    - Nominal scale
    - Potential levels: red, blue, green, yellow
- Time to respond to a question
    - Ratio scale
    - Potential levels: 0.1 sec, 2 sec, 5 sec, 60 sec

Q9: The formula for finding each student's test grade (g) from his or her raw score (s) on a test is as follows: g = 16 + 3s. Is this a linear transformation? If a student got a raw score of 20, what is his test grade?

Answer:
1. It is a linear transformation since the variable s is transformed by multiplying it with a constant (3) and adding a second constant (16).
2. Test grade is **76**.

Q11: Which of the frequency polygons has a large positive skew? Which has a large negative skew?

Answer:
1. Large Positive skew: A
2. Large Negative skew: C

Chapter 2

Q1: Name some ways to graph quantitative variables and some ways to graph qualitative variables.

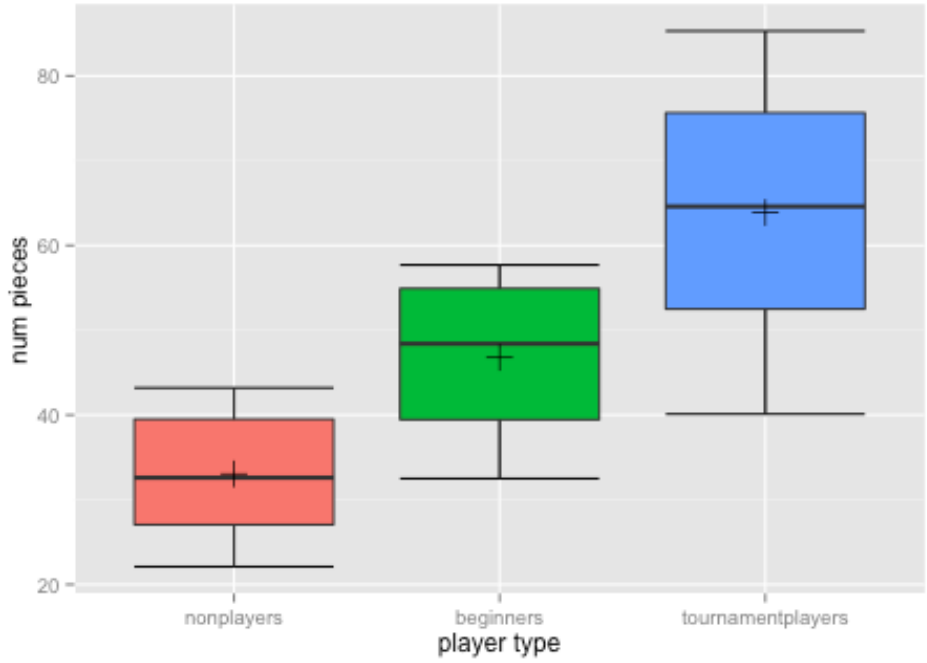
Answer:
Qualitative variable graphing methods

- Pie charts
- Vertical Bar charts
- Horizontal Bar charts

Quantitative variable graphing methods

- Stem and Leaf displays
- Histograms
- Frequency Polygons
- Box Plots
- Bar Charts
- Line Graphs
- Scatter Plots

Q3: An experiment compared the ability of three groups of participants to remember briefly-presented chess positions. The data are shown below. The numbers represent the total number of pieces correctly remembered from three chess positions. Create side-by-side box plots for these three groups. What can you say about the differences between these groups from the box plots?
Answer:



Observations:

1. The median for tournament players is higher than for beginners which, in turn, is higher than for non players. This indicates that regular players have better recollection of chess positions.
2. The gap between the adjacents and hinges is large for tournament players as compared to that for non players and beginners. This, combined with the larger H-spread, indicates a wider distribution of values for tournament players. The corresponding distribution of values for non players and beginners is smaller.
3. The mean for non players is slightly greater than the median indicating a positive skew. The corresponding mean for beginners and tournament players is less than the median indicating a negative skew.

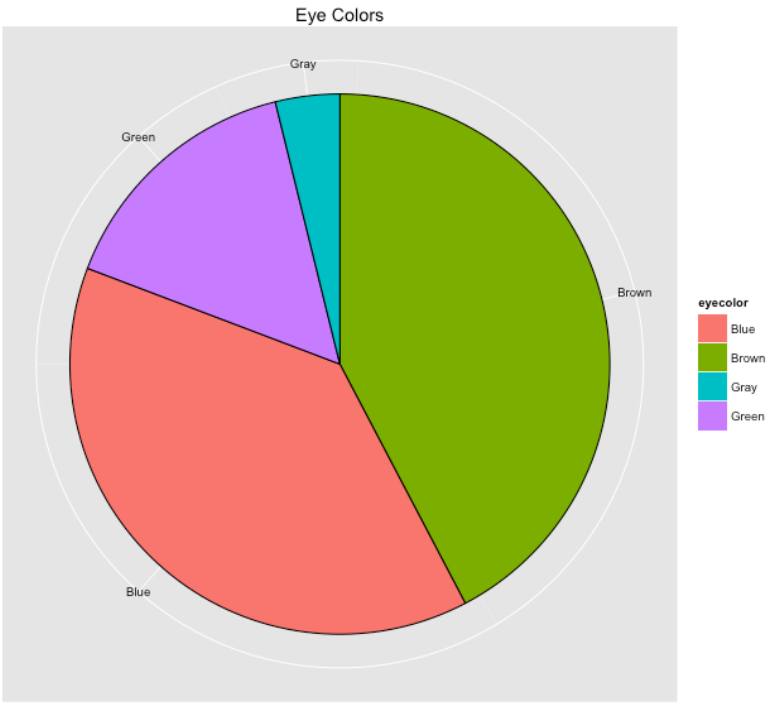
Q5: In a box plot, what percent of the scores are between the lower and upper hinges?

Answer: 50%

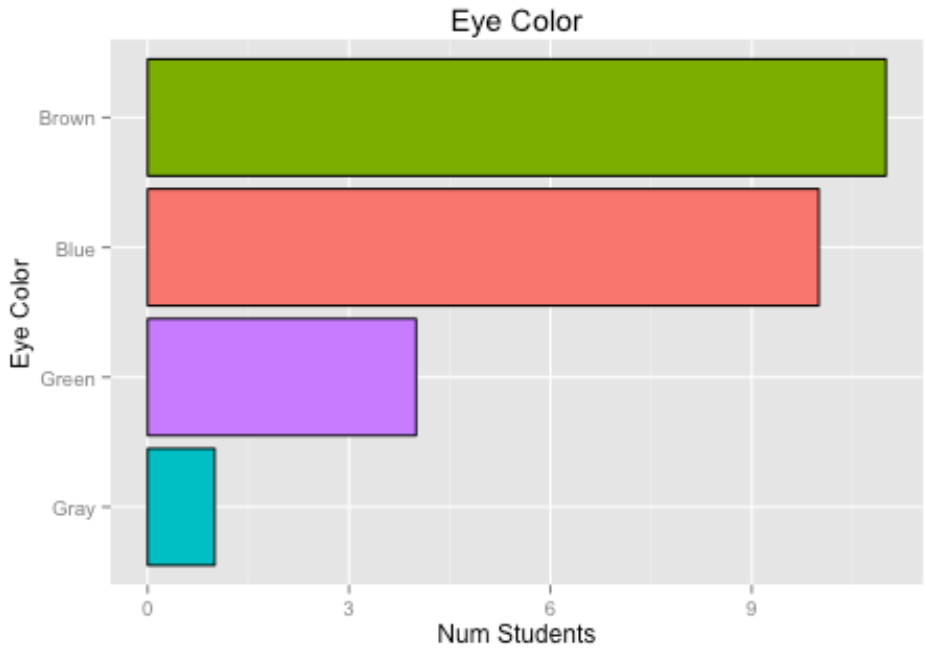
Q7: For the data from the 1977 Stat. and Biom. 200 class for eye color, construct:
a. pie graph
b. horizontal bar graph
c. vertical bar graph
d. a frequency table with the relative frequency of each eye color

Answer:

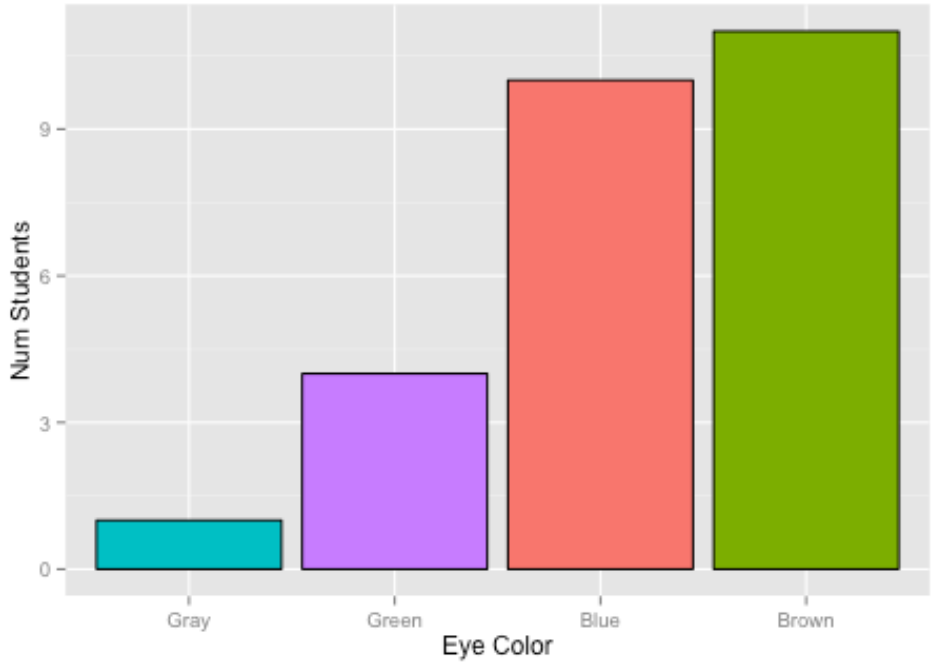
Pie graph



Horizontal bar graph



Vertical bar graph



Frequency table

	eyecolor	numstudents	relativefrequency
1	Brown	11	0.42
2	Blue	10	0.38
3	Green	4	0.15
4	Gray	1	0.04

Q9: Which of the box plots below has a large positive skew? Which has a large negative skew?

Answer: **Plot B** has a large positive skew. **Plot C** has a large negative skew.

Chapter 3

Q1: Make up a dataset of 12 numbers with a positive skew. Use a statistical program to compute the skew. Is the mean larger than the median as it usually is for distributions with a positive skew? What is the value for skew?

Answer:

data <- data.frame(num = c(1,1,2,2,2,2,3,3,4,5,6,7))

- mean: **3.167**
- median: **2.5**

Yes, value of mean is greater than median, as expected for a positive skewed distribution.

Value of skew is **0.7613364**

Q3: Make up three data sets with 5 numbers each that have:

a. the same mean but different standard deviations.
b. the same mean but different medians.
c. the same median but different means.

Answer:

a. Same mean, different std dev

	Dataset A	Dataset B	Dataset C
	1	-1	-100
	2	2	5
	3	3	5
	4	5	5
	5	6	100
mean	3	3	3
std dev	1.870829	2.738613	70.76369

b. same mean, different median

	Dataset A	Dataset B	Dataset C
	1	-1	-100
	2	2	5
	3	2	5
	4	6	5
	5	6	100
mean	3	3	3
median	3	2	5

c. same median, different mean

	Dataset A	Dataset B	Dataset C
	1	1	1
	2	3	2
	3	3	3
	4	3	20
	5	25	24
mean	3	7	10
median	3	3	3

Q5: A sample of 30 distance scores measured in yards has a mean of 7, a variance of 16, and a standard deviation of 4.

(a) You want to convert all your distances from yards to feet, so you multiply each score in the sample by 3. What are the new mean, variance, and standard deviation?

(b) You then decide that you only want to look at the distance past a certain point. Thus, after multiplying the original scores by 3, you decide to subtract 4 feet from each of the scores. Now what are the new mean, variance, and standard deviation?

Answer:

If a variable X has a mean of μ, a standard deviation of σ, and a variance of σ2, then a new variable Y created using the linear transformation

$Y = bX + A$

will have:

- mean of $b\mu+A$
- a standard deviation of $b\sigma$
- a variance of $b^2\sigma^2$

(a) b = 3, A = 0
- mean: 21
- variance: 144
- std dev: 12

(b) b = 3, A = -4
- mean: 17
- variance: 144
- std dev: 12

Q7: For the test scores in question #6, which measures of variability (range, standard deviation, variance) would be changed if the 22.1 data point had been erroneously recorded as 21.2?

Answer: Standard deviation and variance will change. Range remains unchanged.

Q9: For the numbers 1, 3, 4, 6, and 12:
a. Find the value (v) for which $\Sigma(X-v)^2$ is minimized.
b. Find the value (v) for which $\Sigma|x-v|$ is minimized.

Answer:

a. Smallest squared deviation is minimized by MEAN. Hence, v = 5.2.
b. Smallest absolute deviation is minimized by MEDIAN. Hence, v = 4.

Q11: An experiment compared the ability of three groups of participants to remember briefly-presented chess positions. The data are shown below. The numbers represent the number of pieces correctly remembered from three chess positions. Compare the performance of each group. Consider spread as well as central tendency.

Answer:

Central tendency

	non players	beginners	tournament players
mean	33.04	46.79	63.89
median	32.6	48.4	64.6
trimean	32.9375	47.7925	64.3375

	non players	beginners	tournament players
trimmed mean (20%)	33.1375	47.2125	64.1875
mode	43.2	na	na
geometric mean	32.12256	45.96684	62.07393

Spread

	non players	beginners	tournament players
range	22.1 - 43.2	32.5 - 57.7	40.1 - 85.3
inter quartile range	12.45	15.475	23.15
variance	64.53378	81.55211	244.0299
std dev	8.033292	9.030621	15.62146

Q13: True/False: The best way to describe a skewed distribution is to report the mean.

Answer: False.

Q15: Compare the mean, median, trimean in terms of their sensitivity to extreme scores

Answer: The mean is highly sensitive to extreme scores. The median is not sensitive to extreme scores. The trimean is also not sensitive to extreme scores if the scores do not shift the H-spread.

Q17: A set of numbers is transformed by taking the log base 10 of each number. The mean of the transformed data is 1.65. What is the geometric mean of the untransformed data?

Answer: 44.66836.

Q19: The histogram is in balance on the fulcrum. What are the mean, median, and mode of the distribution (approximate where necessary)?

Answer:

- mean: 4.5
- median: 4
- mode: 1

Chapter 4

Q1: Describe the relationship between variables A and C. Think of things these variables could represent in real life.

Answer: There is a negative association between the variables. The points are not clustered together. The relationship between the variables is potentially non-linear, since the points do not fall on a straight line.

In real life, these variables could potentially represent:

1. Price and demand for a product
2. Marginal Utility curve for a product
3. Half life of radioactive elements

Q3: Make up a data set with 10 numbers that has a negative correlation.

Answer:

X	Y
10	100
20	90
30	80
40	70
50	60
60	50
70	40
80	30
90	20
100	10

Q5: Would you expect the correlation between High School GPA and College GPA to be higher when taken from your entire high school class or when taken from only the top 20 students? Why?

Answer: I would expect the correlation to be higher when taken from the top 20 High School students. This is because top students in High School with a high GPA have a higher probability of also scoring a high GPA in College.

Q7: For this same class, the relationship between the amount of time spent studying and the amount of time spent socializing per week was also examined. It was determined that the more

hours they spent studying, the fewer hours they spent socializing. Is this a positive or negative association?

Answer: Negative association.

Q9: Students took two parts of a test, each worth 50 points. Part A has a variance of 25, and Part B has a variance of 36. The correlation between the test scores is 0.8.

(a) If the teacher adds the grades of the two parts together to form a final test grade, what would the variance of the final test grades be?
(b) What would the variance of Part A - Part B be?

Answer:

(a) 109
(b) 13

Q11: True/False: It is possible for variables to have r=0 but still have a strong association.

Answer: False. r = 0 implies no linear relationship, which means that variables cannot have an association.

Q13: True/False: After polling a certain group of people, researchers found a 0.5 correlation between the number of car accidents per year and the driver's age. This means that older people get in more accidents.

Answer: True.

Q15: True/False: To examine bivariate data graphically, the best choice is two side by side histograms.

Answer: False. The best choice would be a scatter plot.