ISTA 116 Lab: Week 8

Midterm Review

Last Revised October 10, 2011

1 Data types

Univariate, Bivariate, Multivariate

1.1 Categorical data

Examples:

- Sex (Levels: Male, Female)
- Eye Color (Levels: Blue, Brown, Green, Other)
- Home state (Levels: Alaska, Arkansas, Arizona, . . .)

1.2 Numeric data

Discrete: consecutive values with nothing in between *possible* (usually, whole numbers).

Examples:

- The number of correct answers each student gets on a 5 question math exam.
- The number of times it takes a person to pass a driving test.

Continuous: there's always another value possible between any two, no matter how close.

Examples:

- Sound pressure levels (decibels) measured at rock concerts.
- Time between 1st and 2nd place at horse races.

2 Plots

Example data set: heights of 4 year olds:

38 24 40 36 36 41 38
24 40 41 45 37 36 36
39 40 36 43 33 39 30

1. How do you draw a strip chart?
2. Stem-and-Leaf plot:

2	4 4
3	0 3 6 6 6 6 6 7 8 8 9 9
4	0 0 0 1 1 3 5

3. Histogram: what would the histogram look like for this data? What if we divide into smaller bins?

3 Central Tendency & Variability

3.1 Mean

Calculated by summing all the data points and dividing by the number of data points:

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

What are some problems with the mean?

3.2 Median

Calculated by finding the center data point if there are an odd number of data points, or the point midway between the 2 center data points if even.:

$$Q_2 = \begin{cases} x_{\frac{n+1}{2}} & \texttt{if } n \texttt{ is odd} \\ Mean(x_{\frac{n}{2}}, x_{\frac{n}{2}+1}) & \texttt{if } n \texttt{ is even} \end{cases}$$

3.3 Variance

- Deviation scores for each point: $\text{Deviation}_i = x_i - \bar{x}$
- Variance is the average squared deviation:

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

- Standard Deviation is the square root of the variance: $\sqrt{s^2}$
- What is the standard deviation of the 4-year-old heights?
- What are some problems with standard deviation?

3.4 *z*-scores

z-scores show how far data points are from the mean, independent of the units:

$$z_i = \frac{(x_i - \bar{x})}{s} \tag{3}$$

3.5 Inter-Quartile Range (IQR) & H-spread

- What are Quartiles?
- IQR is the distance between Q_1 and Q_3: $IQR = Q_3 - Q_1$
- Hinges are calculated by splitting the data into 2 subsets and finding the medians of these subsets
  - The lower-hinge H_1 is the median of the data at or below Q_2.

  - The upper-hinge H_3 is the median of the data at or above Q_2.
- The H-spread is similar to the IQR: H-spread $= H_3 - H_1$.
- What is the H-spread for the 4-year-old heights?

3.6 Five number summary & Box-and-whisker plots

- The five-number summary contains $(Q_0, H_1, Q_2, H_3, Q_4)$
- The box-and-whisker plot presents this information visually.
- What does the box-and-whisker plot look like for the 4-year-old data?
- How can we tell if a distribution is skewed or symmetric?

4 Contingency Tables

Consider the following table dealing with votes on the debt-ceiling deal:

		Vote		
		Yea	Nay	No Vote
Party	GOP	171	59	10
	Dem	68	77	48

1. What type of information does this table show?
2. What are the marginal frequencies for Party and Vote?
3. How are the joint and marginal proportions calculated?
4. Calculate the distribution of votes conditioning on Party.

5 Correlation

Consider the dataset from Web Quiz 5:

ID	Hrs of Sleep	Errors
1	3	5
2	9	1
3	2	11
4	6	6
5	5	7

1. Calculate Pearson's correlation coefficient using the following formula:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\sum_{i=1}^{n}(x_i - \bar{x})^2)(\sum_{i=1}^{n}(y_i - \bar{y})^2)}} \tag{4}$$

   - Break this down into pieces: find $\bar{x}$ and $\bar{y}$ (the means of each column).
   - Make new columns for $(x_i - \bar{x})$ and $(y_i - \bar{y})$.
   - Now find the value of the numerator by multiplying the numbers in each row and summing the results.
   - Make new columns for $(x_i - \bar{x})^2$ and $(y_i - \bar{y})^2$.
   - Sum the squared columns, multiply the results, and take the square root to get the denominator.
   - Now divide the numerator by the denominator to get r.

2. Spearman's ρ uses a similar procedure, but starts with ranks.

$$\rho = \frac{\sum_{i=1}^{n}(\text{Rank}(x_i) - \frac{n+1}{2})(\text{Rank}(y_i) - \frac{n+1}{2})}{\sqrt{\sum_{i=1}^{n}(\text{Rank}(x_i) - \frac{n+1}{2})^2 \sum_{i=1}^{n}(\text{Rank}(y_i) - \frac{n+1}{2})^2}} \tag{5}$$

   - Begin by making new columns that have the ranks:

ID	Rank(Sleep)	Rank(Errors)
1	2	2
2	5	1
3	1	5
4	4	3
5	3	4

   - Find the median of the ranks (3 for both).

- Make new columns that have the Ranks − the median: $(\text{Rank}(x_i) - \frac{n+1}{2})$ and $(\text{Rank}(y_i) - \frac{n+1}{2})$.
- Multiply each row and sum the results to get the numerator.
- Make new columns that square the $(\text{Rank}(x_i) - \frac{n+1}{2})$ and $(\text{Rank}(y_i) - \frac{n+1}{2})$ terms.
- Sum each of those, multiply the results, and take the square root to get the denominator.
- Divide the numerator by the denominator to get ρ.