

STA 291

Summer 2008

Lecture 4

Summary: Numerical Descriptive Techniques

- Measures of Location / Central Tendency
 - Mean, Median, Mode
- Measures of Variation, i.e. “spread” of the data
 - Range, Variance, Standard Deviation, Interquartile Range
- Five Number Summary
 - Minimum, Lower Quartile, Median, Upper Quartile, Maximum

Review: Measuring Central Tendency

- “What is a typical measurement in the sample/population?”
- Mean: Arithmetic average
- Median: Midpoint of the observations when they are arranged in increasing order
- Mode: Most frequent value

Review: Mean vs. Median vs. Mode

- Mean: Interval data with an approximately symmetric distribution
- Median: Interval or ordinal data
- Mode: All types of data

Review: Mean vs. Median vs. Mode

- The mean is sensitive to outliers, median and mode are not
- In general, the median is more appropriate for skewed data than the mean
- In some situations, the median may be too insensitive to changes in the data
- The mode may not be unique

Review: Range

- Range: Difference between the largest and smallest observation
- Very much affected by outliers (one misrecorded observation may lead to an outlier, and affect the range)
- The range does not always reveal different variation about the mean

Review: Box Plots

- Basically a graphical version of the five number summary (unless there are outliers)
- **Box** that contains the central 50% of the distribution (from lower quartile to upper quartile)
- **Line** within the box that marks the median,
- **Whiskers** that extend to the most extreme observations *within 1.5 IQR beyond the quartiles*
- Observations beyond are marked as **outliers**.

Standard Deviation (s) vs. Interquartile Range (IQR)

- Standard Deviation is affected by outliers
- Interquartile Range is not affected by outliers
- *Note: The Range is most affected by outliers*
- For symmetric, bell-shaped distributions, $IQR \sim 4/3 s$ (can be shown mathematically)
- Whenever you use the Median instead of the Mean, you should also use the IQR instead of the Standard Deviation

Variance and Standard Deviation: Interpretation

- Variance and standard deviation cannot be negative
- They equal 0 if and only if all observations are the same
- Standard deviation can easier be interpreted because it is given in the same units as the data
- In general, both are particularly useful when comparing the variation of two or more distributions

There is also: Coefficient of Variation

- Standardized measure of variation
- Idea: A standard deviation of 10 may indicate great variability or small variability, depending on the magnitude of the observations in the data set
- CV = Ratio of standard deviation divided by mean
- Population and sample version
 - Population: $CV = \frac{S}{m}$ Sample: $cv = \frac{s}{\bar{x}}$

Example: Coefficient of Variation

- Which sample has higher relative variability?
(a higher coefficient of variation)
 - Sample A
 - mean = 62
 - standard deviation = 12
 - CV =
 - Sample B
 - mean = 31
 - standard deviation = 7
 - CV =

Use of the Standard Deviation

- Empirical rule: If the histogram of the data is approximately symmetric and bell-shaped, then
 - About **68%** of the data are within **one** standard deviation from the mean
 - About **95%** of the data are within **two** standard deviations from the mean
 - About **99.7%** of the data are within **three** standard deviations from the mean

Example: Highway Gas Mileage for 24 Cars

Stem Leaf	#	Boxplot
5 0	1	
4		
4 00	2	
3 5678	4	+-----+
3 01112	5	*--+-*
2 557889	6	
2 0134	4	+-----+
1 7	1	
1 3	1	

-----+-----+-----+-----+

Multiply Stem Leaf by 10**+1

Notice the
"split stems"

Five number summary of the data:

Maximum =

Upper Quartile =

Median =

Lower Quartile =

Minimum =

Application of the Empirical Rule

- Gas Mileage Data
 - Mean = 29.6
 - Standard Deviation = 8.3
 - 68% of the data (_____ observations) are supposed to be between _____ and _____
 - How many observations are actually within one standard deviation from the mean?
 - 95% of the data (_____ observations) are supposed to be between _____ and _____
 - How many observations are actually within two standard deviations from the mean?

Chebysheff's Theorem

- Let $k > 1$, for any data set the proportion of observations that lie within k standard deviations from the mean is at least

$$1 - \frac{1}{k^2}$$

- This is a conservative estimate of the proportion of observations in an interval centered at the mean.

Chebysheff's Theorem

Value of k	Interval	Proportion of observations
2	$\mu - 2s < x < \mu + 2s$	$\geq .75$
2.5	$\mu - 2.5s < x < \mu + 2.5s$	$\geq .84$
3	$\mu - 3s < x < \mu + 3s$	$\geq .89$
4	$\mu - 4s < x < \mu + 4s$	$\geq .93$

Analyzing Linear Relationships Between Two Quantitative Variables

- Is there an association between the two variables?
- Positive or negative?
- How strong is the association?
- Notation
 - Response variable: Y
 - Explanatory variable: X

Sample Measures of Linear Relationship

- Sample Covariance:

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{1}{n-1} \left(\sum x_i y_i - \frac{1}{n} \sum x_i \sum y_i \right)$$

- Sample Correlation Coefficient:

$$r = \frac{s_{xy}}{s_x s_y}$$

- Population measures: Divide by N instead of n-1

Properties of the Correlation I

- The value of r does not depend on the units (e.g., changing from inches to centimeters), whereas the covariance does
- r is standardized
- r is always between -1 and 1 , whereas the covariance can take *any* number
- r measures the ***strength and direction of the linear association*** between X and Y
- $r > 0$ positive linear association
- $r < 0$ negative linear association

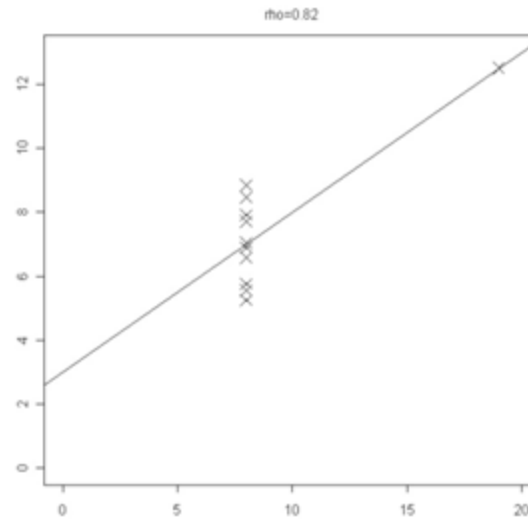
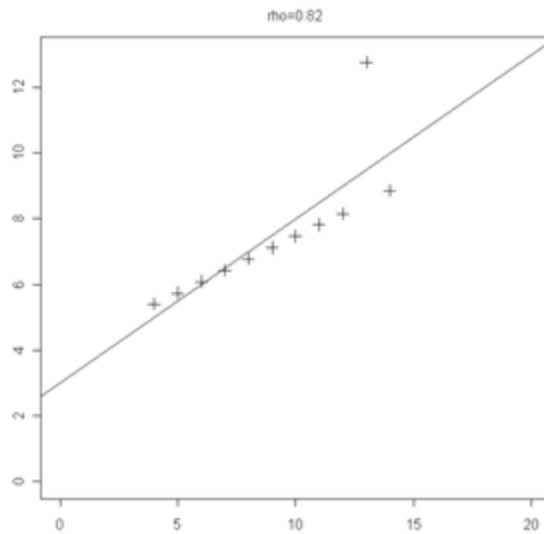
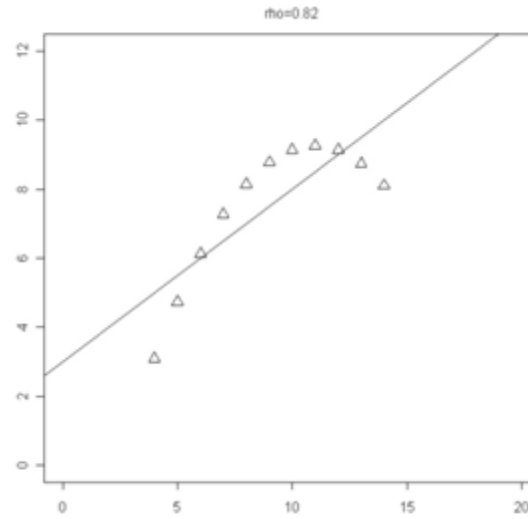
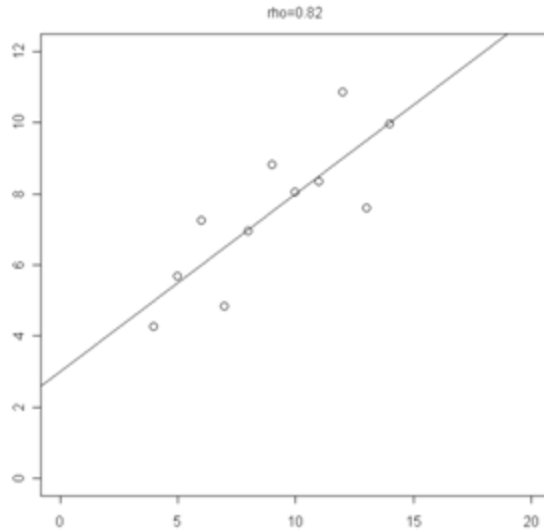
Properties of the Correlation II

- $r = 1$ when all sample points fall exactly on a line with positive slope (*perfect positive association*)
- $r = -1$ when all sample points fall exactly on a line with negative slope (*perfect negative association*)
- The larger the absolute value of r , the stronger is the degree of linear association

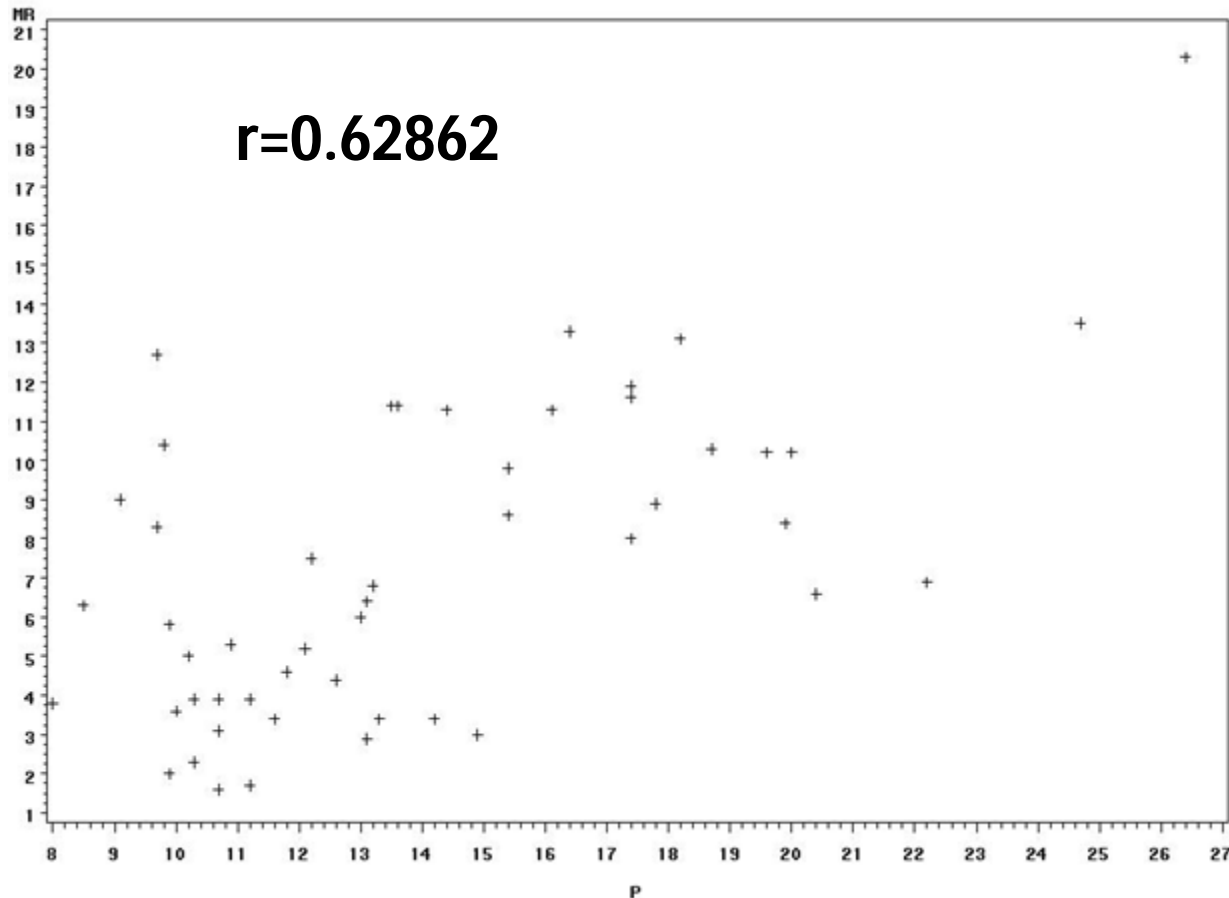
Properties of the Correlation III

- If r is close to 0, this does not necessarily mean that the variables are not associated
- It only means that they are not *linearly associated*
- The correlation treats X and Y symmetrically
- That is, it does not matter which variable is explanatory (X) and which one is response (Y), the correlation remains the same

Example: Correlation = 0.82



Scatter Diagram of Murder Rate (Y) and Poverty Rate (X) for the 50 States



[Correlation and Scatterplot Applet](#)

[Correlation by Eye Applet](#)

[Simple Regression Analysis Tool](#)

Model Assumptions and Violations

- **Factors Influencing the Correlation**
- The sample correlation depends on the range of X -value sampled
- When a sample has a much narrower range of variation in X than the population, the sample correlation tends to underestimate the population correlation
- The sample (X,Y) values should be a random sample of the population
- It should be representative of the X population values as well as the Y values

Correlation: Example

- For a sample of 100 people, the correlation coefficient between $X =$ years of education and $Y =$ annual income (in dollars) equals 0.50
 - a) Suppose instead Y refers to annual income in thousands of dollars. State the correlation.
 - b) Suppose that Y is treated as the explanatory variable and X is treated as the response variable. Will the correlation coefficient change in value?

Example Data Sets

- One Variable Statistical Calculator
- Modify the data sets and see how mean and median, as well as standard deviation and interquartile range change
- Look at the histograms and stem-and-leaf plots – does the empirical rule apply?
- Make yourself familiar with the standard deviation
- Interpreting the standard deviation requires some experience

Covariance: Example

X = Hours per week studying and doing homework for STA 291
Y = STA 291 First Exam Score

x_i	x-Mean	x-Deviation	y_i	y-Mean	y-Deviation	Cross Product
1			45			
5			80			
12			100			
					Sum	
					n-1	
					Cov	

Method of Least Squares

- There are many possible ways to choose a straight line through the data
- Goal: Make the vertical distances between the observations and the straight line as small as possible
- Vertical distances: residuals
- The sum of the residuals should be zero
- There are many possible ways to choose a straight line through the data such that the sum of the residuals is zero

Method of Least Squares

- Better Goal: Minimize the sum of the squared residuals

$$\sum (y_i - \hat{y}_i)^2$$

- The squared residuals are the squared vertical distances between the straight line and the data
- Correlation by Eye Applet Minimize the MSE
- This method is called the ***method of least squares*** (Gauss)

Method of Least Squares

- This leads to the ***prediction equation*** or ***least squares equation***

$$\hat{y} = b_0 + b_1 \cdot x$$

- With the following coefficients

slope $b_1 = \frac{s_{xy}}{s_x^2} = \frac{\sum_i (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$

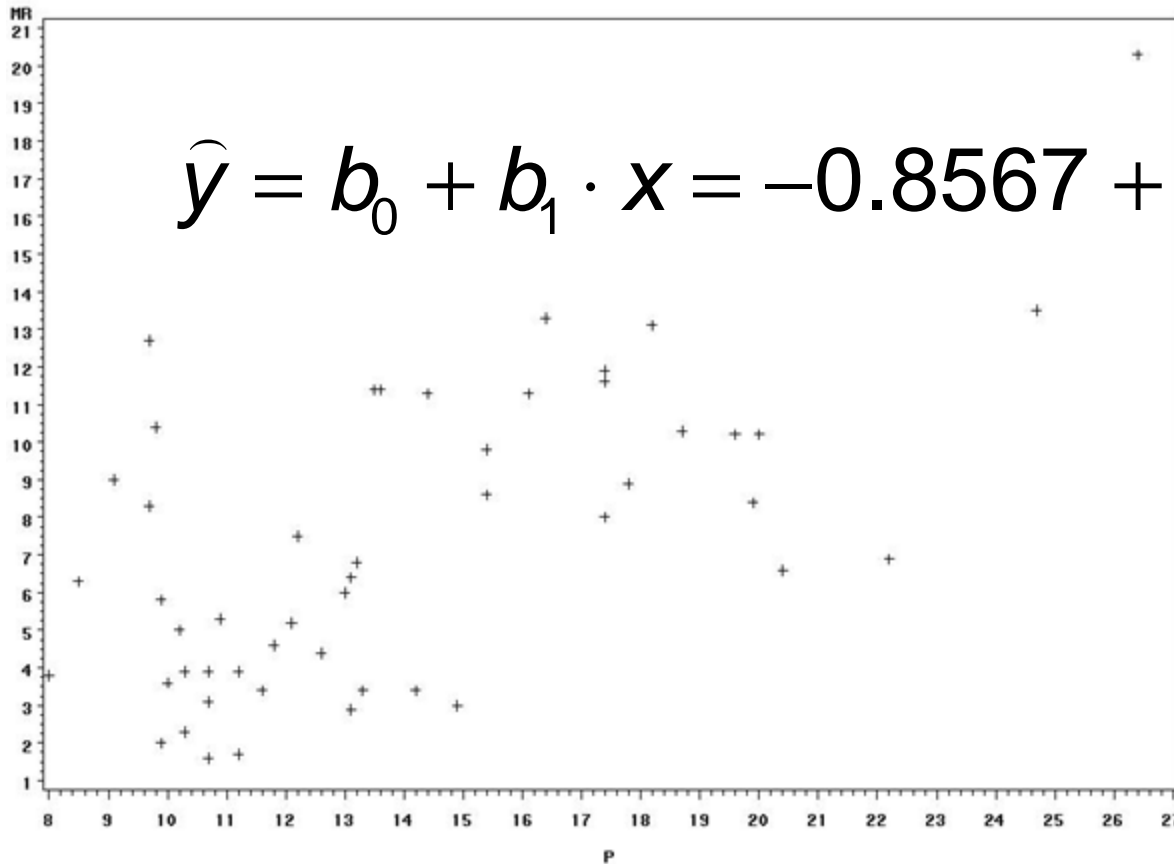
intercept $b_0 = \bar{y} - b_1 \cdot \bar{x}$

- In practice, the calculations for slope and intercept are done using the computer, not by hand

Method of Least Squares

- The least squares estimates b_0 and b_1 in the prediction equation are the values that make the sum of squared residuals minimal
- The equation is called prediction equation because it can be used to predict values of the response variable when knowledge about the explanatory variable is available

Scatter Diagram of Murder Rate (y) and Poverty Rate (x) for the 50 States



$r=0.62862$

- Any other linear equation will lead to a larger sum of squared residuals
- The observed data points fall closer to this line than to any other line

Interpretation of Slope and Intercept

- Slope: rise/run
 - Change in y (rise) for a one-unit increase in x (run)
- Intercept
 - Intersection of the straight line with the (vertical) y -axis
 - (Hypothetical) predicted value of y for $x=0$
 - Often, the intercept has little practical meaning because the data does not have observations with $x=0$

Example

- Someone claims that the prediction equation $\hat{y} = 0.5 + 7.0x$ approximates the relationship between y =college GPA and x =high school GPA (both on a four-point scale)
 - a) Is this realistic? Why or why not?
 - b) Suppose that the prediction equation is actually $\hat{y} = 0.5 + 0.7x$. Interpret the slope.
 - c) Using the prediction equation in b), find the predicted GPA for a student having a high school GPA of (i) 3.0, (ii) 4.0
 - d) Suppose the prediction equation is $\hat{y} = x$. Identify the y -intercept and slope, and interpret their values.

Example (Data from the 50 States)

- y = violent crime rate
- x = poverty rate
- The relation can be approximated by the straight line $y = 210 + 25x$
- Interpretation of y -intercept and slope?
- y = violent crime rate
- x = percentage of high school graduates
- Approximated by $y = 1756 - 16x$
- Interpretation of y -intercept and slope?

Linear Function, Slope

- When $b_1 = 0$, the value of x has no influence on the value of y
- $b_1 > 0$: positive relationship between the variables
- $b_1 < 0$: negative relationship

Slope (b_1) vs. Correlation (r)

- The slope b_1 of the prediction equation tells us the direction of the association between the two variables
 - Positive b_1 : Slope upward
 - Negative b_1 : Slope downward
- The slope does not tell us the strength of the association
 - It depends on the units and can be made arbitrarily small or large by choice of units
 - It does **not** treat X and Y symmetrically
- A measure of the strength of the linear association is the correlation coefficient r

Correlation Coefficient

- The correlation coefficient r can also be interpreted as a standardized version of the slope b_1 of the prediction equation
- It can be calculated using b_1 and the standard deviations of x and y :

$$r = \frac{s_x}{s_y} \cdot b_1$$

where b_1 is the slope of the (estimated) prediction equation

$$\hat{y} = b_0 + b_1 \cdot x$$

Correlation and Slope

- The value of r does not depend on the units – it is a standardized regression coefficient
- r is always between -1 and 1
- Recall that b_0 and b_1 could take *any* value
- r measures the ***strength of the linear association*** between X and Y
- r has the same sign as the slope b_1
- r is symmetric in x and y

Scatterplot

- How to decide whether a linear function may be used?
- *Always plot the data first*
- Recall: A **scatterplot** is a plot of the values (x,y) of the two variables
- Each subject is represented by a point in the plot

Effect of Outliers

- Outliers have a substantial effect on the (estimated) prediction equation
- In the murder rate vs. poverty rate example, DC is an outlier
- Prediction equation with DC:
$$\hat{y} = -10.13 + 1.32x$$
- Prediction equation without DC:
$$\hat{y} = -0.86 + 0.58x$$

Effect of Outliers

- Removing the outlier causes a large change in the results
- Observations whose removal causes substantial changes in the prediction equation, are called ***influential***
- It may be better not to use one single prediction equation for the 50 states and DC
- In reporting the results, it has to be noted that the outlier DC has been removed
- Correlation and Regression Applet

Prediction

- The prediction equation $\hat{y} = b_0 + b_1 x$ is used for predictions about the response variable y for different values of the explanatory variable x
- For example, based on the poverty rate, the predicted murder rate for Arizona is

$$b_0 + b_1 x = -0.8567 + 0.5842 \times 20 = 10.83$$

Dependent Variable	Predicted Value	Residual	
10.2	10.8281	-0.6281	(Arizona)
6.6	11.0618	-4.4618	(Kentucky)

Residuals

- The difference between observed and predicted values of the response variable ($y - \hat{y}$) is called a ***residual***
- The residual is negative when the observed value is smaller than the predicted value
- The smaller the absolute value of the residual, the better is the prediction
- The sum of all residuals is zero

Correlation: Example

- Three data pairs (x,y) where x = number of books read for pleasure in the last year, y = daily average number of hours spent watching television
- $(0,5)$ $(5,3)$ $(10,1)$
- a) Construct a scatter plot. State the prediction equation (in this case, it is possible to do this without using the least squares formula) and interpret
- b) Report the correlation coefficient between x and y and interpret

Regression Toward the Mean

- Sir Francis Galton (1880s): correlation between x =father's height and y =son's height is about 0.5
- Interpretation: If a father has height one standard deviation below average, then the predicted height of the son is 0.5 standard deviations below average
- More Interpretation: If a father has height two standard deviations above average, then the predicted height of the son is $0.5 \times 2 = 1$ standard deviation above average
- Tall parents tend to have tall children, but not **so** tall
- This is called "regression toward the mean"
———statistical term "regression"

Exam I, Monday, June 23

- **When: 6:45 – 8:30 PM**
- **Where: CB 343 (usual classroom)**
- **The Exam will cover Lectures 1-4, that is, Chapters 1-5 of the textbook**

Exam I, Monday, June 23

- Please bring a calculator.
- Any technology that can receive/transmit information wirelessly is not permitted during the exam
- Examples include:
 - cell phones, PDAs with wireless/Bluetooth/IR capabilities, laptops/tablet PCs

In the unlikely instance that a breach of academic integrity is suspected, it will be dealt with in strict accordance with the University of Kentucky policy on academic integrity.

The University of Kentucky regards cheating as a very serious offense for which the minimum penalty is failure in the course (see the Student Code, Section 6.4.0).

Exam I, Monday, June 23

- Here is a list of items for which you should know how to calculate them
 - Median, mode
 - Percentiles
 - Range, Interquartile Range
 - Empirical rule
 - Outlier
- You should also know how to draw histograms, stem-and-leaf plots, and box plots, and know how to interpret slope and correlation coefficient in a linear regression
- The formulas for calculating the variance and standard deviation will be provided

Exam I, Monday, June 23

Concepts, definitions

- - population (parameter describe popu.)
 - sample,
 - size n
 - SRS, stratified random sampling
 - (statistic describe sample)

Exam I, Monday, June 23

- Population parameter: often unknown, but is a fixed number.
- Sample statistic: known after collecting the sample and do the calculation, but it is random due to the random sampling.
- Statistic = ? Parameter

Symbols

μ (mu)	<i>population mean</i>
σ (sigma)	<i>population standard deviation</i>
σ^2 (sigma-square)	<i>population variance</i>
x or x_i (x-i)	<i>observation</i>
\bar{x} (x-bar)	<i>sample mean</i>
s	<i>sample standard deviation</i>
s^2	<i>sample variance</i>
Σ	<i>summation symbol</i>

Mean, Median, Mode

- Example: Salary distribution at a college

Salary Range (in \$ 1,000)	Frequency	Cumulative Frequency	Relative Frequency	Relative Cumulative Frequency
20-29	4			
30-39	3			
40-49	2			
50-59	1			
> 60	1			

Mean vs. Median vs. Mode

- Mean: Interval data with an approximately symmetric distribution
- Median: Interval or ordinal data
- Mode: All types of data

	Mode	Median	Mean
Nominal (unordered categories)	Yes	No	No
Ordinal ()	Yes	Yes	No
Interval ()	Yes	Yes	Yes

Measures of Variation

- Range: maximum - minimum
- Variance (sample / population):

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

$$\mathbf{s}^2 = \frac{\sum (x_i - \mathbf{m})^2}{N}$$

- Standard Deviation (sample / population):

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

$$\mathbf{s} = \sqrt{\frac{\sum (x_i - \mathbf{m})^2}{N}}$$

- Interquartile Range: Q3-Q1

Variance Step By Step

1. Calculate the mean
2. For each observation, calculate the deviation
3. For each observation, calculate the squared deviation
4. Add up all the squared deviations
5. Divide the result by $(n-1)$

(To get the standard deviation,
take the square root of the result)

Sample Variance: Example

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

- Calculate the variance and standard deviation for the following data set:
91, 83, 92, 93, 91

(Note: A table such as on the next slide will not be provided in the exam)

Variance and Standard Deviation: Example

Observation	Mean	Deviation	Squared Deviation
Sum of the Squared Deviations			
$n-1$			
Variance = Sum of the Squared Deviations / $(n-1)$			
Standard Deviation = Square Root of the Variance			

Use of the Standard Deviation

- Empirical rule: If the histogram of the data is approximately symmetric and bell-shaped, then
 - About **68%** of the data are within **one** standard deviation from the mean
 - About **95%** of the data are within **two** standard deviations from the mean
 - About **99.7%** of the data are within **three** standard deviations from the mean

Use of the Standard Deviation

- In 2006, scores of 242 students in a STA 291 exam had a mean of 76.1 and a standard deviation of 13.8
- Mean and median were close to each other, suggesting a roughly symmetric distribution
- How many students would you expect to have scored between...

Example

- 20 Exam scores:
78, 92, 85, 48, 76, 88, 62, 56, 60, 97,
33, 68, 95, 56, 68, 87, 100, 96, 49, 87
- Create a stem-and-leaf plot.
- Find the median.

Example

- Number of goals in 2007 FIFA Women's World Cup games until semi-finals:

11, 4, 1, 0, 2, 7, 4, 2, 2, 2, 1, 3, 5, 3, 4, 2, 9, 4, 5,
5, 2, 4, 2, 1, 3, 1, 3, 5, 3, 4

- Calculate the five-number summary.
- Determine whether there are any outliers.
- Sketch a box plot.

Method of Least Squares

- The prediction equation describes the straight line that minimizes the sum of the squared residuals

$$\sum (y_i - \hat{y}_i)^2$$

- The squared residuals are the squared vertical distances between the straight line and the data points

Method of Least Squares

- This leads to the ***prediction equation*** or ***least squares equation***

$$\hat{y} = b_0 + b_1 \cdot x$$

- With the following coefficients

slope

$$b_1 = \frac{s_{xy}}{s_x^2} = \frac{\sum_i (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

intercept

$$b_0 = \bar{y} - b_1 \cdot \bar{x}$$

Correlation Coefficient and Slope

- The correlation coefficient r is a standardized version of the slope b_1 of the prediction equation

$$r = \frac{s_x}{s_y} \cdot b_1$$

where b_1 is the slope of the (estimated) prediction equation

$$\hat{y} = b_0 + b_1 \cdot x$$

- Advanced Interpretation:
 - Slope: If x increases by **one unit**, then y is expected to increase by **b_1 units**
 - Correlation coefficient: If x increases by **one standard deviation**, then y is expected to increase by **r standard deviations** (does not depend on units!)