AP STATISTICS EXAM

Review Session 1: Exploring Data

Notes from AP Test Prep Series: AP Statistics by Pearson Education

Graphical Displays

- First step: Answer the W's
  - Who?
  - What?
    - Categorical?
      - Numbers that don't make sense to average
        - Zip codes
        - Jersey numbers
      - Data that can be counted and put in order but not measured
        - Horse-race finishes
        - Team standings
    - Quantitative?

Graphical Displays

- First step: Answer the W's
  - When?
  - Where?
  - How?
  - Why?

Frequency Distribution Table

Table III
Frequency Distribution of Average Reading Scores Tabulated to Nearest Grade Level

Score Interval	Grade Level*	*f*
10.5–11.4	11	2
9.5–10.4	10	6
8.5–9.4	9	14
7.5–8.4	8	22
6.5–7.4	7	19
5.5–6.4	6	12
4.5–5.4	5	10
3.5–4.4	4	3
	N =	88

*Midpoint of class interval.

http://cache.eb.com/eb/image?id=74903&rendTypeId=4

Frequency Distribution Table

- Rules:
  - All classes must be included, even if the frequency of some classes is zero
  - All classes should have the same width. The class intervals should be equal.

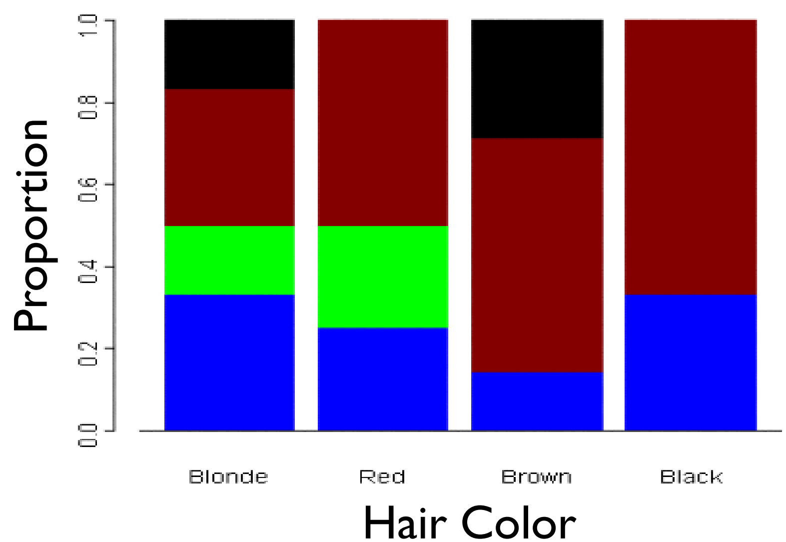
http://cache.eb.com/eb/image?id=74903&rendTypeId=4

Data Analysis Info

- Make a picture, **make a picture**, ***make a picture***!
- One Variable, Categorical Data
  - Do a bar chart or pie chart
    - Bar charts have spaces between each category
    - Order of the category is not important
    - Show either counts or proportions
    - LABEL APPROPRIATELY
    - Describe the chart in the CONTEXT of the data
    - DO NOT describe the shape of a categorical variable

Data Analysis Info

- Make a picture, **make a picture**, ***<u>make a picture</u>***!
- Two-Variable, Categorical Data
  - Do a segmented bar graph
    - Describe in CONTEXT the relationship between the two variables



Data Analysis Info

- Make a picture, **make a picture**, ***<u>make a picture</u>***!
- One Variable, Quantitative Data
  - Histogram, Ogive, Stem-and-Leaf, Dotplot, Boxplot
    - Histograms do not have spaces between the bars, UNLESS there is no data in that interval
    - Describe the shape, center and spread of the distribution in the CONTEXT of the data
    - When working with two data sets, be sure to make comparisons between the two using the same scale

Numerical Descriptions

- Five-Number Summary
  - Minimum, Q1, Median, Q3, Maximum
  - IQR = Q3 – Q1
  - Show a boxplot
    - Stat – Calc – 1 Var Stats
    - 2nd – StatPlot – Modified Boxplot

Numerical Descriptions

- Measures of Center
  - Mean (not resistant to outliers)
  - Median (resistant to outliers)
  - Don't forget weighted means
    - For example, suppose your school reports grades quarterly and you take midterm and final exams. If your grades for each quarter count 20% and the midterm and final exams each count 10%, calculate your final average for the following grades:

1st Q	2nd Q	Mid	3rd Q	4th Q	Final E
85	80	82	78	74	71

Numerical Descriptions

- Measures of Spread
  - Standards Deviation (not resistant to outliers)
  - Interquartile Range (resistant to outliers)
  - Range

Comparing Data

- When the data sets have different means and standard deviations,
  - Z-scores

$$z = \frac{x - \mu}{\sigma}$$

Scaling/Shifting Data

- Adding/subtracting to/from each value
  - Adds or subracts the same constant from the mean
  - Measures of spread (standard deviation, range, IQR) remain unchanged
- Multiplying/dividing all the data values
  - Measures of center and spread are affected

Normal Models

- Appropriate for distributions whose shapes are unimodal and roughly symmetric
- $N(\mu, \sigma)$
- 68-95-99.7% (Empirical) Rule

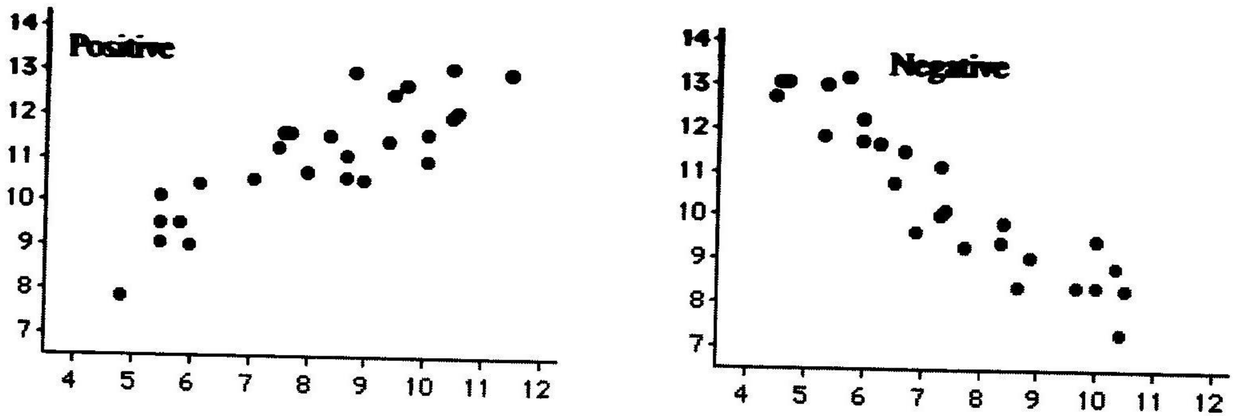
Session 2

EXPLORING RELATIONSHIPS BETWEEN VARIABLES

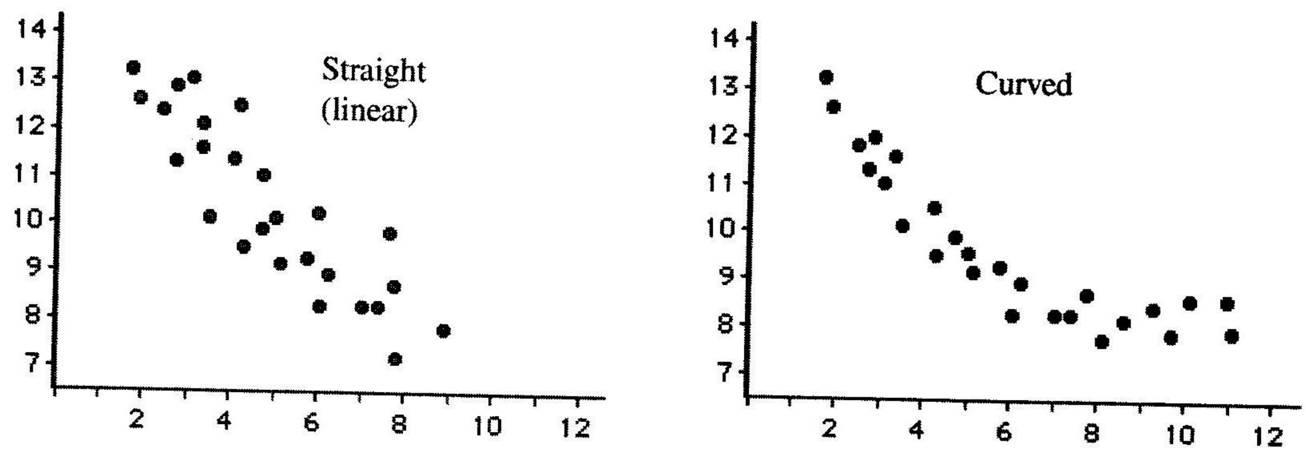
Scatterplots

- Explanatory (predictor) variable goes on the x-axis
- Response variable (the variable you hope to predict or explain) on the y-axis

When analyzing a scatterplot, discuss…Direction

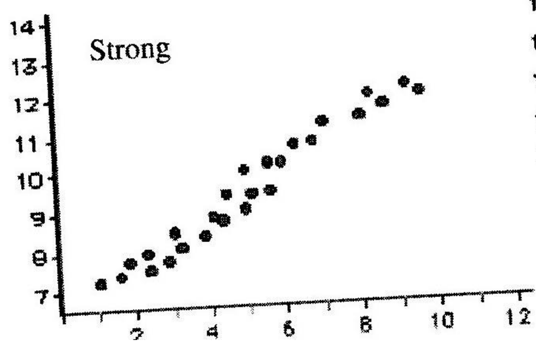


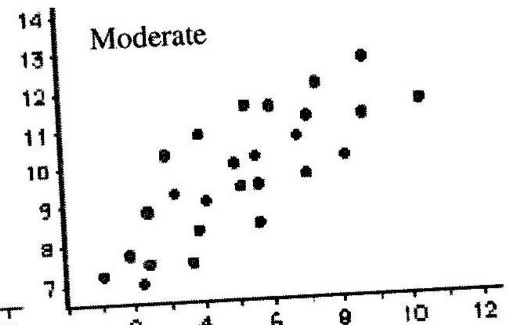
Form

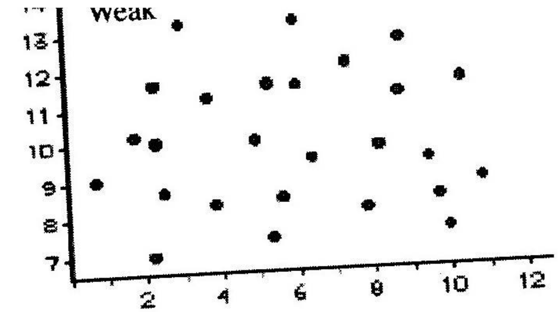


37

When analyzing a scatterplot, discuss…Strength







Note:

- Association does NOT imply causation
- Causation can only by assessed through a randomized, controlled experiment

Correlation Coefficient (r)

$$r = \frac{1}{n-1} \sum \left(\frac{x - \bar{x}}{s_x}\right)\left(\frac{y - \bar{y}}{s_y}\right)$$

Facts about the correlation coef.

- No units
- Quantitative variables
- Sign indicates direction of association
- Between -1 and 1
- Linear only; NO CURVES
- Not resistant to outliers
- Not affected by changes in scale or center
- NO CAUSATION

Session 3

GATHERING DATA

Understanding Randomness

- A random event is one whose outcome we can't predict
- BUT…long-run predictability is helpful
  - Example: We can't predict whether a flipped coin will land on heads or tails, but we CAN predict that in the long run, the percentage of each will be about 50%.

Performing a Simulation

- See handout

1. Identify the trial to be repeated
2. State how you'll model the random occurrence of an outcome
3. Explain how you will simulate the trial
4. Define the response variable
5. Run several trials
6. Summarize the result across the trials
7. Describe what your simulation shows
8. Draw conclusions about the real world

Random Number Table Note

- Mark the table so that your method can be followed by the reader
- Indicate the response variable (yes or no) for each trial

Terminology of Sampling

- Population
  - The entire group of individuals
- Sample
  - A smaller group of individuals selected from the population
- Sampling frame
  - A list of individuals from the population of interest from which the sample is drawn
    - For example, population = high school students, but our sample comes from private schools, then our sampling frame does not represent the population

Terminology of Sampling

- Census
  - A sample that consists of the entire population
- Sampling Variability
  - The natural tendency of randomly-drawn samples to differ

Terminology of Sampling

- Parameter
  - A number that characterizes some aspect of the population
- Statistic
  - Value calculated for sample data

Name	Parameter	Statistic
Mean	μ	$\bar{x}$
Standard Deviation	σ	s
Correlation	ρ	r
Regression Coefficient (slope)	β	b
Proportion	p	$\hat{p}$

Sampling Designs

- Sample size
  - The number of individuals selected from our sampling frame
- Probability Sample
  - Chosen using a random mechanism in such a way that each individual has the same chance of being selected
- Random Sample
  - Chosen using a random mechanism in such a way that the probability of each sample being selected can be computed (with or without replacement)

Sampling Designs

- Simple Random Sample (SRS)
  - A random sample chosen without replacement so that in an SRS of size n, an individual could be selected only once for that sample

Sampling Designs

- Stratified Random Sample
  - The population is divided into strata (homogeneous groups) before simple random sampling is applied
    - Example: A tv station wants information from its viewers about events they are likely to watch during the Olympics. The stations suspects that there will be a difference between responses from men and women. They stratify by gender to help reduce variation.

Sampling Designs

- Cluster Sample
  - The population exists in readily-defined heterogeneous clusters (groups). The sample is an SRS of the clusters.
    - A large school wants to sample 9th grade students about summer reading requirements. Students are assigned to homerooms alphabetically. A random sample of 9th grade homerooms is selected, with all students in each selected classroom participating.

Sampling Designs

- Systematic Sampling
  - A sample is selected according to a predetermined scheme. (Note: This never produces a simple random sample.)
    - When there is reason to believe that order of the list is not associated with the responses sought, this method gives a representative sample.
      - List seniors alphabetically. Choose every 10th student, starting with a randomly selected number.

Sampling Designs

- Multistage Sampling
  - May combine several methods of sampling
  - Produces a final sample in stages, each sample taken from the one before
  - Does NOT produce an SRS

Sampling Designs

- Convenience
  - Sampling individuals who are conveniently available.
  - Does not produce an SRS
  - Not likely to represent the population
  - Likely to cause bias

Sources of Bias

- Undercoverage
- Response bias
- Nonresponse bias
- Voluntary response bias (self-selected surveys)

Observational Study vs. Randomized Comparative Experiment

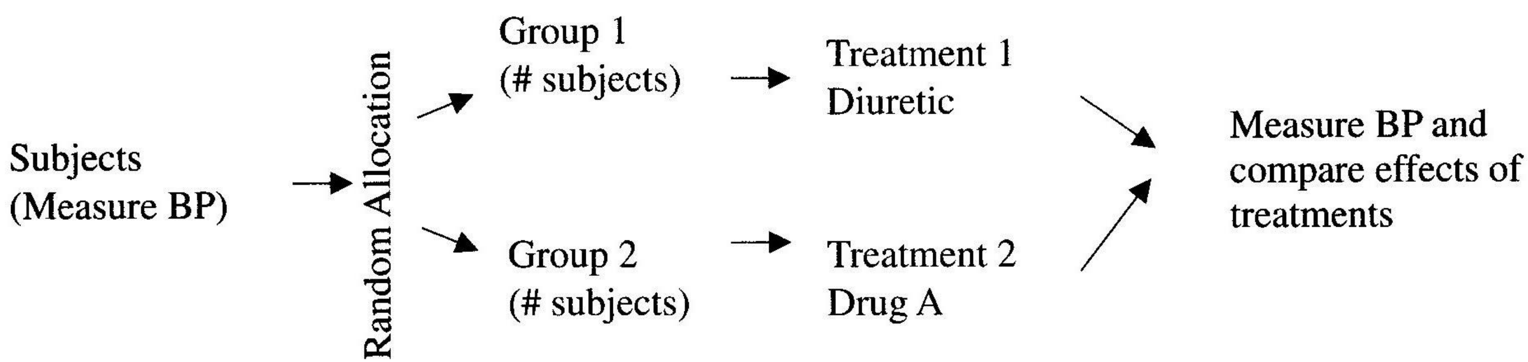
- Observational study
  - Researchers observe individuals, record variables, but NO TREATMENT IS IMPOSED
  - You CAN NOT prove cause-and-effect from an observational study

Observational Study vs. Randomized Comparative Experiment

- Experiment
  - Treatment is imposed
  - Can determine cause-and-effect relationship
    - Explanatory variable
    - Response variable

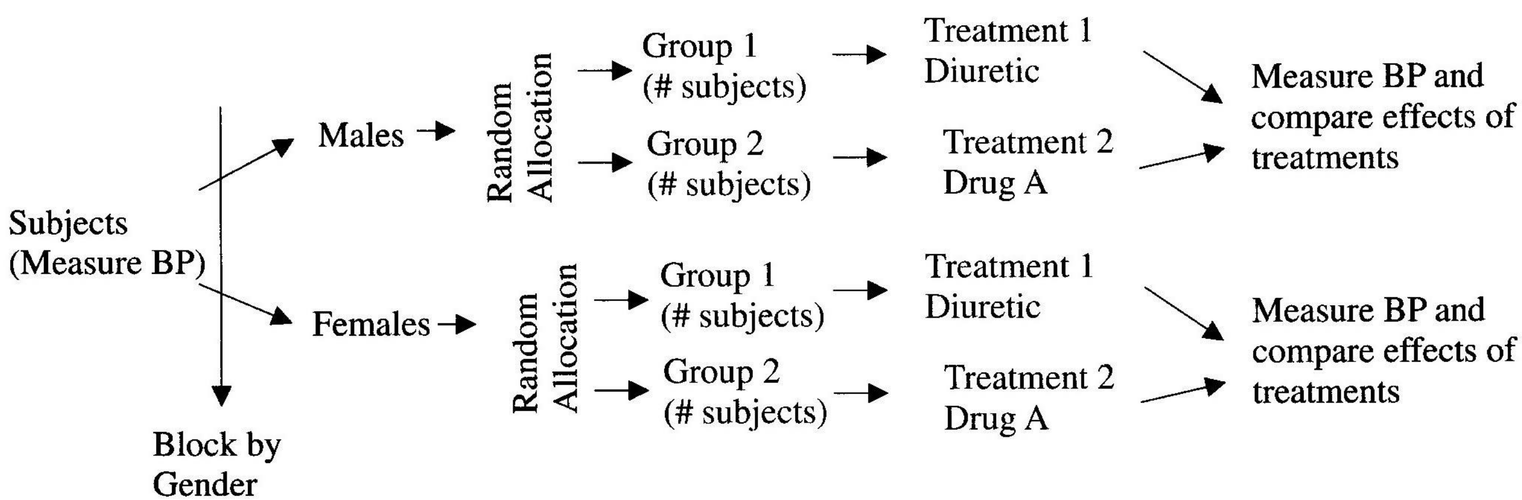
Experimental Design

- Completely randomized Experiment



Experimental Design

Block Design



Experimental Design

- Matched-Pairs Design
  - A form of block design
    - One Subject
      - One subject receives both treatments
      - Note: Randomize the order of the treatment
      - Example: One person works a puzzle while listening to classical music, then works a similar puzzle while listening to rock music. Randomize which music is played first to rule out improvement from experience.

Experimental Design

- Matched-Pairs Design
  - A form of block design
    - Two subjects
      - Two subjects with common characteristics are paired
      - One subject receives one treatment, the other receives the other treatment
      - Example: Marathon runners are matched by weight, build, and running times. One wears a new running shoe, the other wears the old shoe. Difference is then compared.

Principles of Experimental Design

1. Control
    - Reduce variability by controlling sources of variation
2. Randomize
    - Randomization to treatment groups reduces bias cause by lurking variables
3. Replicate
    - Include many subjects
    - Others should be able to reproduce the experiment
4. Block

Other Considerations

- Blinding
  - Single-Blind
    - Subjects don't know which treatment group they have been assigned to, OR
    - Evaluators don't know how subjects have been allocated to treatment groups
  - Double-Blind
    - Neither the subjects nor the evaluators know how the subjects have been allocated to treatment groups

Other Considerations

- Confounding
  - This occurs when we can’t separate the effects of a treatment (explanatory variable) from the effects of other influences (confounding variables)

Other Considerations

- Statistical Significance
  - When an observed difference is too large for us to believe that it is likely to have occurred by chance

Placebo Effect

- The tendency in humans to show a response whenever they think a treatment is in effect.
- Use a control group to contradict this tendency.