Stat 504: Statistical Approaches to Interdisciplinary Data Analysis

Introductory session

Overview for Stat 504

- **Goals of the course**:
- Overview/review of key statistical topics,
- Confidence in using statistical software,
- Awareness of some approaches for analyzing data arising from interdisciplinary studies

More overview for Stat 504

- **Course topics**:
- Selected statistical topics:
  - Methods of data collection,
  - Linear model concepts
- Experience using statistical software
- Application of statistical methods for interdisciplinary research in selected articles

Overview for this session

- Discussion of data collection methods
- The importance of visual displays of data

Data Collection

- Surveys: use probability sampling if possible
- Experiments: use randomization of treatments to subjects
- Observational studies: other types of collected data

Surveys

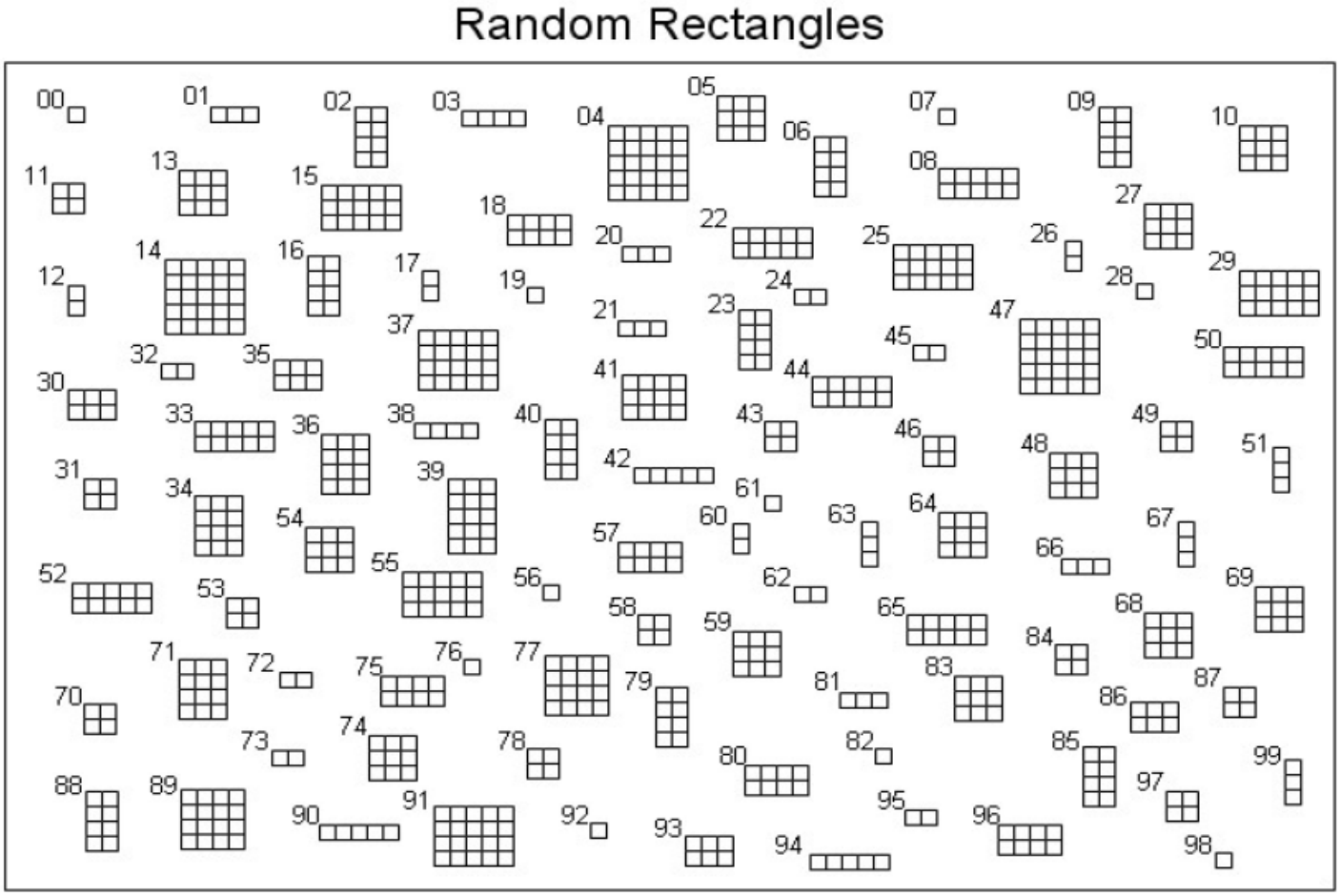
- Is a large sample size enough?
- In 1936, Franklin Delano Roosevelt had been President for one term. The magazine, The Literary Digest, predicted that Alf Landon would beat FDR in that year's election by 57 to 43 percent. The Digest mailed over 10 million questionnaires to names drawn from lists of automobile and telephone owners, and over 2.3 million people responded - a huge sample. But Roosevelt won with 62% of the vote. The size of the Digest's error is staggering.
- How could they have been so far off?

Surveys

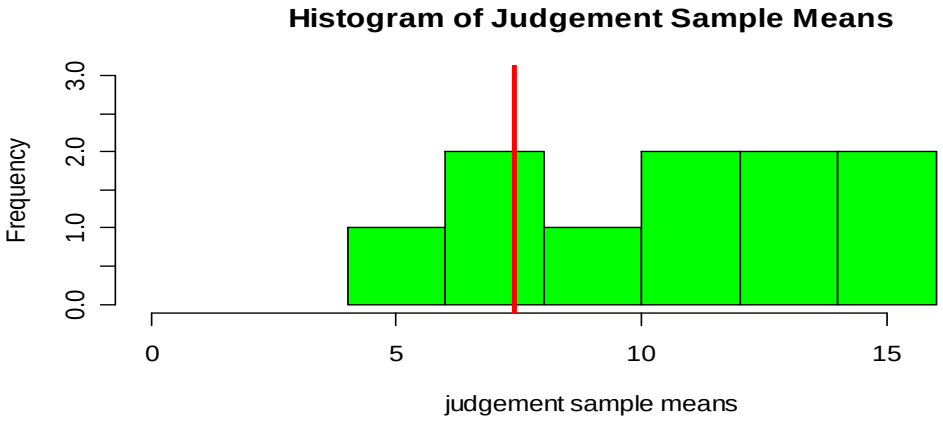
- The key to conducting a scientific survey is to use probability sampling when possible
- Even data from large samples cannot substitute for taking a probability sample. The Literary Digest survey had 2.3 million respondents but was badly wrong. On the other hand, scientific surveys commonly make accurate estimates for the entire country using only 1000-1500 respondents

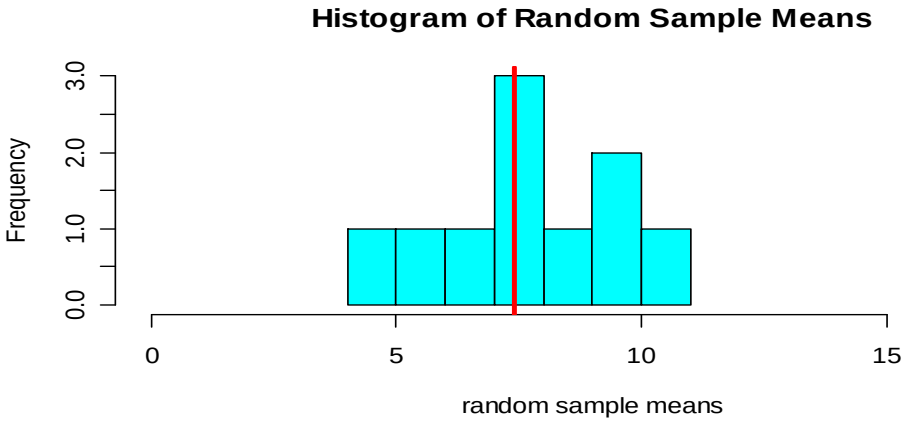
A rectangle sampling activity

Source: Key Curriculum, *Activity Based Statistics*



Rectangle sampling results





Do all surveys require probability sampling?

Experiments

- Random assignment of treatments to subjects is the key
- There are many examples of studies that did not use randomization that gave unreliable results

The Portacaval Shunt

In patients with cirrhosis of the liver, this operation was thought to be helpful

Source: Freedman et al, *Statistics*, 1991

Design	Marked enthusiasm	Moderate enthusiasm	None
No controls	24	7	1
Control, no randomization	10	3	2
Randomized controlled	0	1	3

Can all research studies use randomization?

- Does cigarette smoking cause lung cancer in humans?

Visualizing Data

- Long ago, computation was difficult and researchers spent much time graphing their data in various ways before attempting to analyze the data
- Now, tiny handheld computers (cellphones next?) can conduct statistical analyses far more sophisticated than was possible in those times long ago

The Good and Bad of the Computer Revolution

- The Good: Many more options are available for analyzing data (and also understanding data graphically)
- The Bad: Many researchers have become too reliant on the power of the computer to deliver computational results, often without an adequate understanding of those results and potential limitations of the methods that they use

Roles of Visual Displays of Data

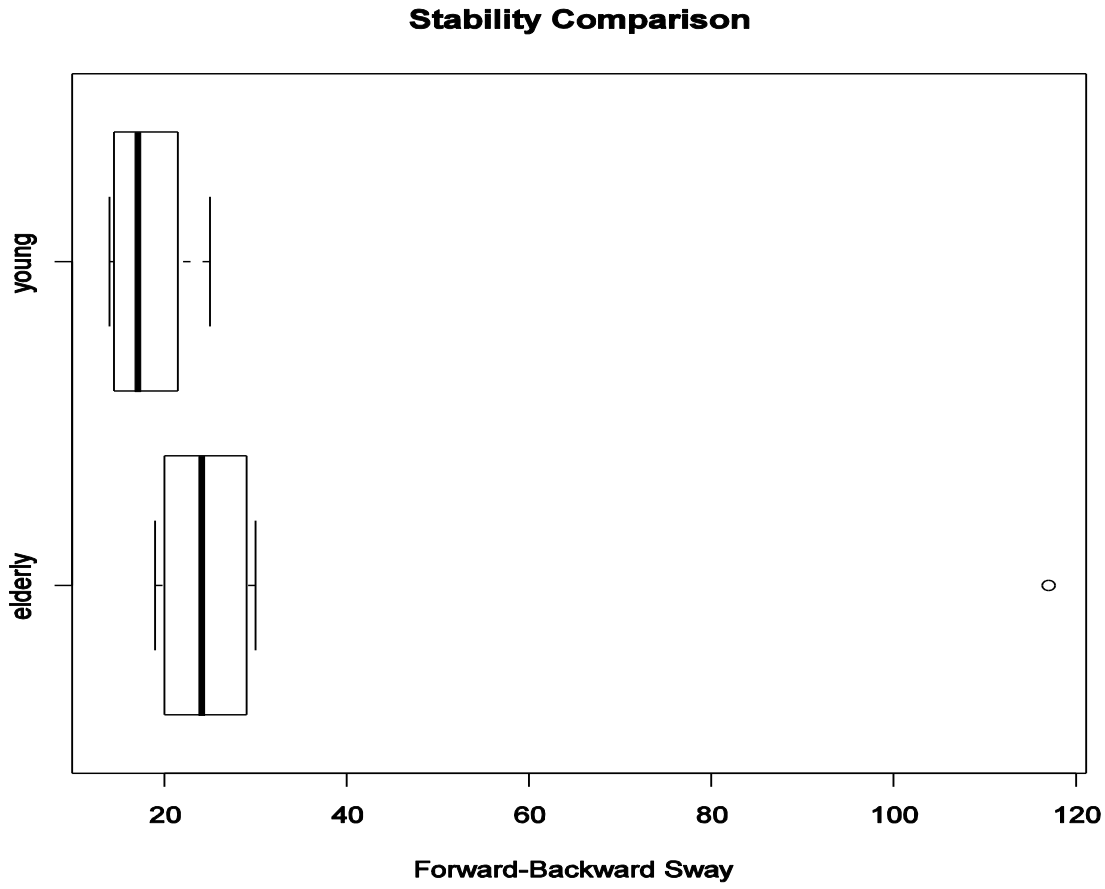
- Quality control: Graphs offer a quick way to look for problems
- Understanding Analyses: If you are reporting results from an analysis of data, you should have a graph to help explain the results
- Assessing Assumptions: Statistical analyses have assumptions, which are usually best examined with graphs

Stability example

adapted from data at http://lib.stat.cmu.edu/DASL/Datafiles/Balance.html

- Forward/backward sway was measured in 8 ‘young’ and 9 ‘elderly’ subjects on a force platform
- A two-sample t test yields t = 1.39 on 15 df with P = .18 (two-tailed). On the basis of this test it would seem that there is no evidence of difference in average forward/backward sway between the two groups

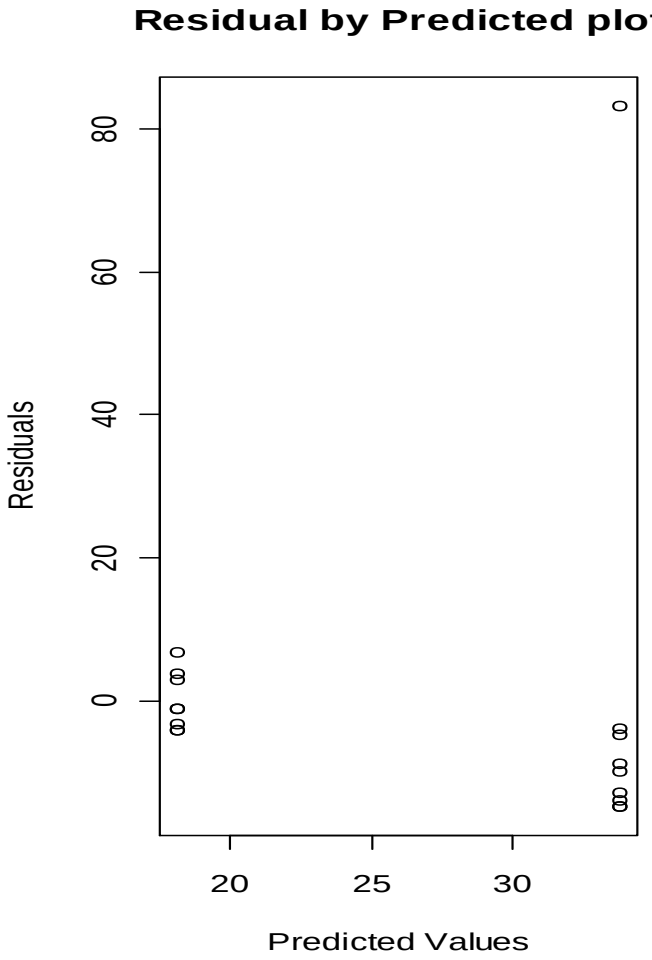
A plot of the data – how does it help us understand the result?

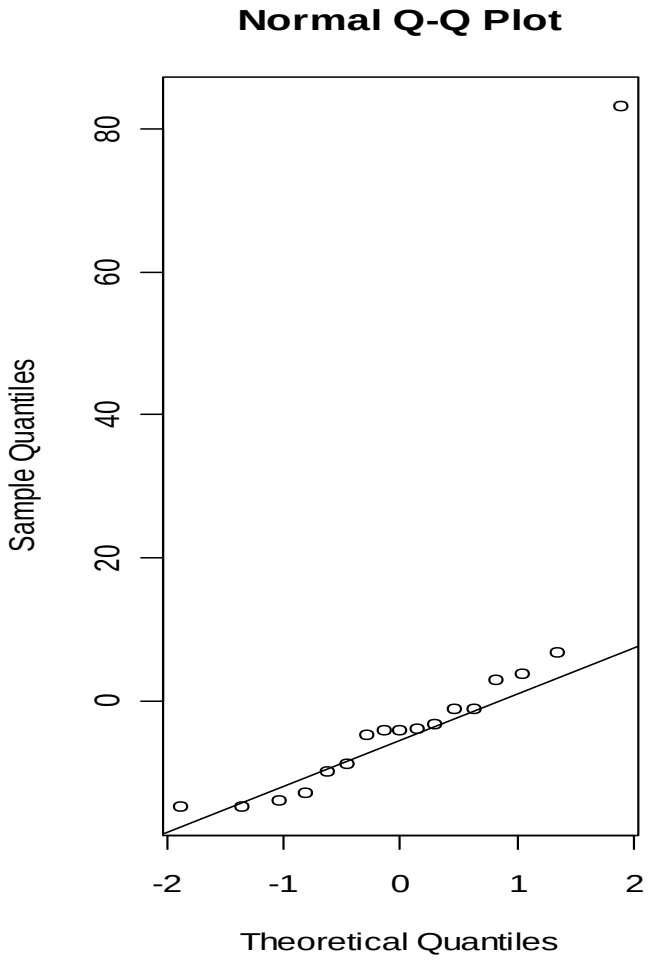


Plots for Assessing Assumptions

- For the stability data the model can be written as:
- $Y_{ij} = \mu_i + \varepsilon_{ij}$
- Where the errors ε_{ij} are assumed to be independent and normally distributed with common mean 0 and variance σ^2
- We can obtain estimates of the errors from the fitted model and make plots to assess the normality and common variance assumptions

Assessing Assumptions for the Stability t test





Conclusion after viewing plots

- There actually appears to be a difference between groups that is obscured by a single outlier
- The moral of the story is that you should wait to make final conclusions until after viewing graphical summaries of the data