STA 570 Spring 2011

Lecture 26

Thursday, April 28

Review for Exam

Review for the Final Exam

- What did we cover since the midterm?
- Guide to choosing a statistical method
- Reading/interpreting some SAS output
- ANOVA
- Two samples

Final Exam

- Thurs, 5th May, 13:00-15:00
- Bring
  - Calculator *(**not** a cell phone or similar device)*
  - up to two letter-sized assistance sheets
- Suggested Preparation
  - Old Final Exam
  - Midterm Exam
  - Homework problems
  - Lab problems
  - Lecture Notes
  - Books

Topics Covered

- 1-5 Midterm
- 6 Significance Tests, P-values
- 7 Two-Sample Tests and Confidence Intervals
- 12 Analysis of Variance
- 8 Categorical Data Analysis

(Incomplete) Checklist

- These are some items you should know how to do
  - Midterm material: How to calculate mean, median, mode, percentiles, range, interquartile range, empirical rule, outlier; how to draw bar graphs, histograms, stem-and-leaf plots, and box plots
  - How to adjust the confidence level / alpha when performing multiple tests
  - How to calculate odds ratio / relative risk for a 2x2 table
  - How to decide whether a variable is qualitative/ quantitative, nominal/ordinal/interval
  - How to interpret correlation, and results (P-values) from statistical tests

Which Method to Choose?

- Quantitative Response (Analyzing Means)
  - No explanatory variable
  - Qualitative explanatory variable
- Two Quantitative Responses
  - Two dependent samples
- Qualitative Response (Analyzing Proportions)
  - No explanatory variable
  - Qualitative explanatory variable

Which Method to Choose?

- Quantitative Response (Analyzing Means)
  - No explanatory variable
  - Qualitative explanatory variable
- Two Quantitative Responses
  - Two dependent samples
- Qualitative Response (Analyzing Proportions)
  - No explanatory variable
  - Qualitative explanatory variable

- Quantitative Response (Analyzing Means),
- No explanatory variable (3, 5.2, 6.2, 6.5)
  - Descriptive graphics: Box Plot, Histogram, Stem and Leaf Plot
  - Descriptive numbers: Sample Mean, Median, Quartiles, Variance, Standard Deviation, Interquartile Range, Five-Number-Summary
  - Inference: Confidence interval and test for the mean
    - large sample: z table
    - small sample and normal population: t table

Which Method to Choose?

- Quantitative Response (Analyzing Means)
  - No explanatory variable
  - Qualitative explanatory variable
- Two Quantitative Responses
  - Two dependent samples
- Qualitative Response (Analyzing Proportions)
  - No explanatory variable
  - Qualitative explanatory variable

- Quantitative Response (Analyzing Means)
- Qualitative explanatory variable
- Two Sample Unpaired t-Test
  - Two levels (comparing two groups) (7.1, 7.3)
    - Descriptive: side-by-side box plots, back-to-back histograms and stem and leaf plots; numerical characteristics
    - Inference for large samples: use z scores for confidence interval and test
    - Inference for small samples from normal population distribution:
      - t tests for equal and for unequal variances

- Quantitative Response (Analyzing Means)
- Qualitative explanatory variable
- Analysis of Variance (ANOVA)
  - Several levels (several groups) (12.1, 12.2)
    - Descriptive graphics: scatter plot of response by group, side-by-side box plots
    - Assumptions for Inference: Equal variances and normal population
    - ANOVA F test for the null hypothesis that all means are equal
    - If F test is significant, then post-hoc multiple comparisons to find out which pairs differ significantly. Use Bonferroni method to account for multiple testing.

Which Method to Choose?

- Quantitative Response (Analyzing Means)
  - No explanatory variable
  - Qualitative explanatory variable
  - Quantitative explanatory variable
  - More than one explanatory variable
- Two Quantitative Responses
  - Two dependent samples
- Qualitative Response (Analyzing Proportions)
  - No explanatory variable
  - Qualitative explanatory variable

- Two Quantitative Responses
- Two dependent samples (7.4)
- Paired t-test
- Descriptive graphics: Scatterplot of the two responses by subject (block)
- Inference: t-test for paired samples, confidence interval

Which Method to Choose?

- Quantitative Response (Analyzing Means)
  - No explanatory variable
  - Qualitative explanatory variable
  - Quantitative explanatory variable
  - More than one explanatory variable
- Two Quantitative Responses
  - Two dependent samples
- Qualitative Response (Analyzing Proportions)
  - No explanatory variable
  - Qualitative explanatory variable

- Qualitative Response (Analyzing Proportions)
- No explanatory variable (3.1, 5.3, 6.3, 6.6)
  - Descriptive: Frequency Table, Bar graph, Sample Proportion
  - Inference: Confidence interval and test for the proportion
    - large sample: z table
    - small sample: binomial distribution (appropriate table has to be generated using computer)

Which Method to Choose?

- Quantitative Response (Analyzing Means)
  - No explanatory variable
  - Qualitative explanatory variable
  - Quantitative explanatory variable
  - More than one explanatory variable
- Two Quantitative Responses
  - Two dependent samples
- Qualitative Response (Analyzing Proportions)
  - No explanatory variable
  - Qualitative explanatory variable

- Qualitative Response (Analyzing Proportions)
- Qualitative explanatory variable (8)
- Chi-Squared Test for Contingency Tables
  - Descriptive: Contingency Table
  - Two levels (comparing two groups)
    - Descriptive numbers: difference of proportions, odds ratio
    - Inference for large samples: Test and confidence interval using z-scores (7.2), or Chi-squared test of independence for a 2x2 table, using chi-squared distribution with df=1 (8.2)
    - Inference for small samples: Fisher's exact test (7.3)
  - Several levels (several groups) (8.2)
    - Inference: Chi-squared test of independence, using chi-squared distribution with df=(r-1)(c-1)

Large Sample or Small Sample

- In general, focus on how to do the calculations by hand in the large sample case (e.g., using z-scores)
- In practice, small sample calculations are done using the computer
- What sample sizes are considered large?

Which Method to Choose?

1. In samples of 150 college graduates and 150 of high school graduates, the mean hourly wage in 1994 was \$15.71 for a college graduate and \$9.92 for a high school graduate.
2. The incarceration rate in 1994 in the nation's prisons was 646 per 100,000 male residents and 45 per 100,000 female residents
3. A study compares the mean level of contributions to political campaigns in Pennsylvania by registered Democrats, registered Republicans, and unaffiliated voters.

P-value Question

- *What does the following mean and/or how can it be interpreted?*
- **“the P-value is 0.03”**

Multiple Choice Question

The P-value for testing the null hypothesis mu=100 (two-sided) is P=.001. This indicates

a) There is strong evidence that mu = 100
b) There is strong evidence that mu does not equal 100
c) There is strong evidence that mu > 100
d) There is strong evidence that mu < 100
e) If mu were equal to 100, it would be unusual to obtain data such as those observed

P-value and Confidence Interval

Assume that the P-value is P=0.043 for a test of the null hypothesis mu=2, with two-sided alternative.

What conclusion can we make about a 95% confidence interval for mu?

ANOVA Example

- Plastic containers for motor oil are blow molded in a machine that has three molding stations. The production engineer is concerned that the weights of the molded containers at different stations are not uniform and a substantial fraction of the containers are not close to the design weight of 51.5 grams. A random sample of eight containers is taken from each station.
- BSS=2.194, WSS=4.577
- Calculate the F test statistic.
- If the overall F test is significant, we need to perform 6x5/2=15 pairwise comparisons. How should we choose the individual alpha* to guarantee a multiple comparison (simultaneous) error rate of alpha=0.05?

Statistics: Making Sense of Data

Methods for Collecting, Describing, Analyzing, and Drawing Conclusions from Data

These methods are used for…

Design
- Planning research studies
- How best to obtain the required data

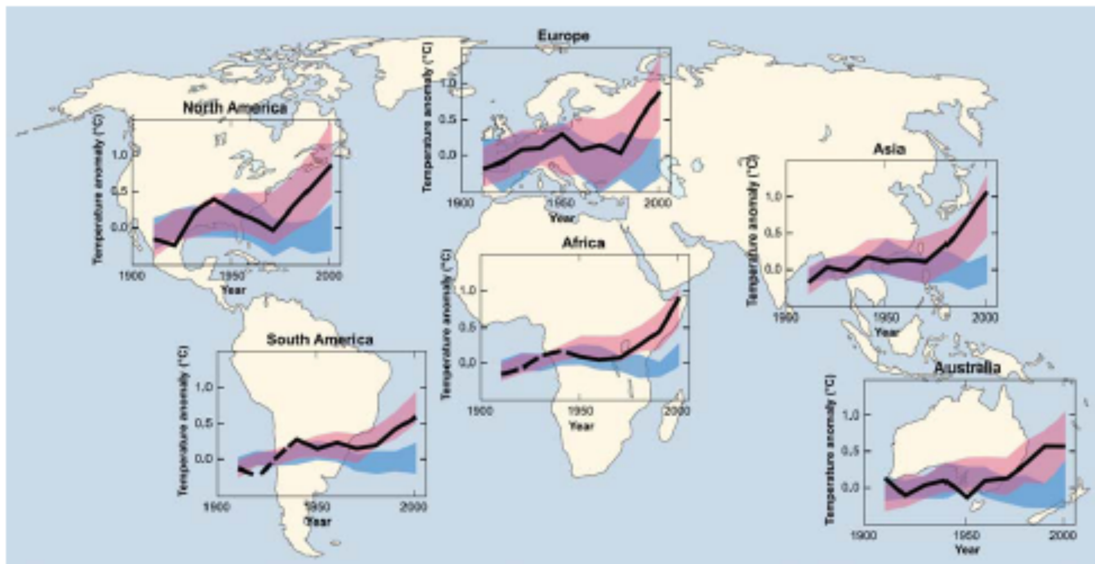
Description
- Summarizing data
- Exploring patterns in the data
- Extract/condense information
- Graphical pictures of the data

Inference
- Make predictions based on the data
- "Infer" from sample to population
- Generalize

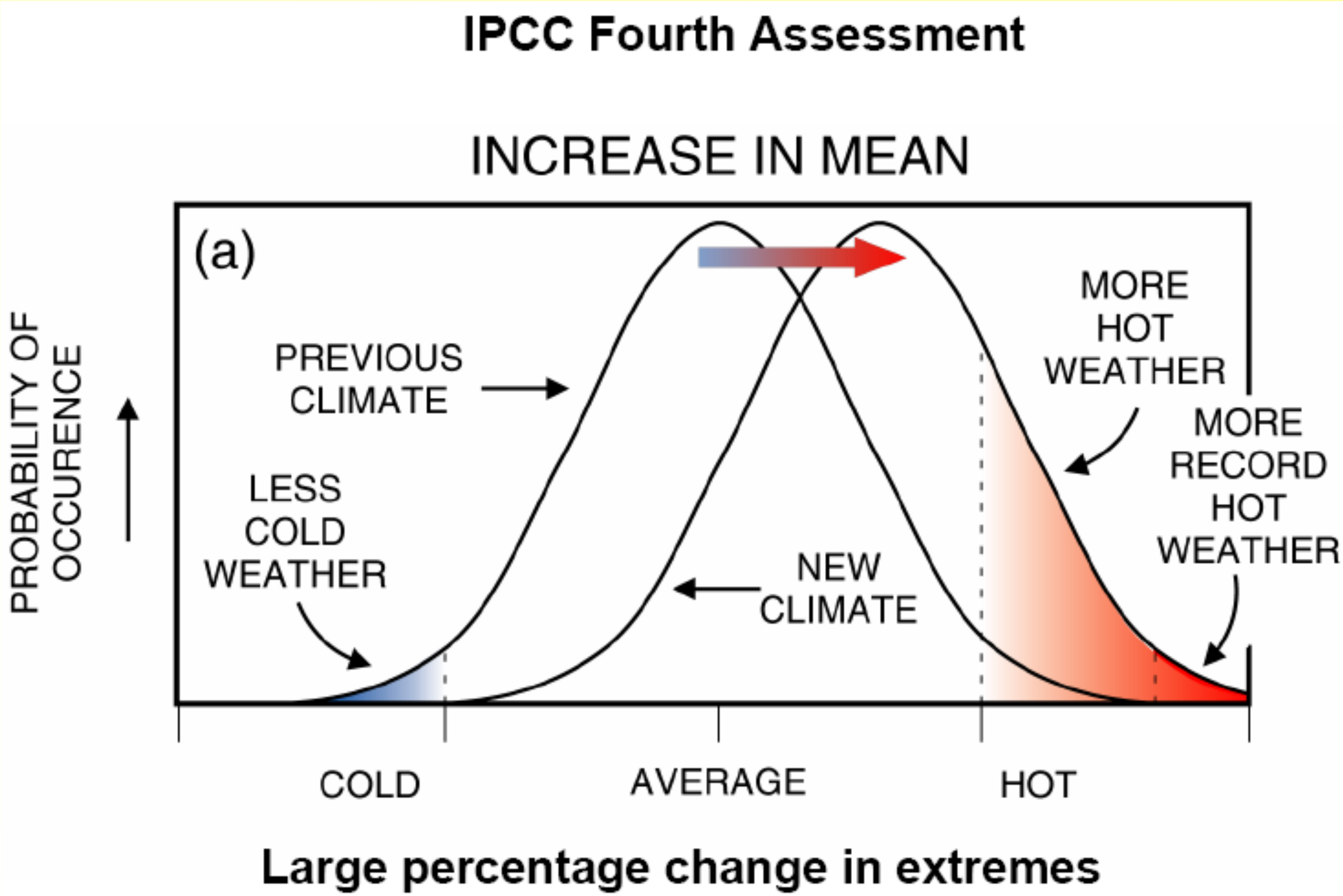
- Hypothesis: *Climate change is due to natural factors*

IPCC (2007):

Most of the observed increase in global averaged temperature … is ***very likely*** *due to … increase in GHG concentrations.*



- Data provides evidence against hypothesis



IPCC Fourth Assessment
INCREASE IN MEAN
(a)
PROBABILITY OF OCCURENCE
PREVIOUS CLIMATE
LESS COLD WEATHER
MORE HOT WEATHER
MORE RECORD HOT WEATHER
NEW CLIMATE
COLD
AVERAGE
HOT
Large percentage change in extremes

Last Slide: Ten Common Mistakes and Famous Fallacies Related to Statistics

- Inference based on non-random data
- The true meaning of “statistically significant”
- Margin of error
- Correlation vs. causation
- Confounding factors
- Multiple testing without adjustment
- Regression Fallacy
- Prosecutor’s Fallacy
- Texas Sharpshooter’s Fallacy
- Simpson’s Paradox