Lecture Notes

Ch. 2 Data Description and Simple Inference

Christopher Kinson

Department of Statistics
University of Illinois at Urbana-Champaign

Exercise: CU Jail Data

```
data cujail;
 length employment_status $25. race $12. sex $6.
   citizenship $8. offense_level $24.
   high_level_crime $12. city $24. state $8.;
 infile 'filelocation/ccso_subset_2.csv' dlm=','
   dsd missover firstobs=2;
 input employment_status $ jacket_number race $ sex $
   citizenship $ age_at_arrest days_in_jail
   offense_level $ high_level_crime $ city $ state $;
run;
```

Exercise: CU Jail Data

1. How many people who were arrested lived outside of Champaign-Urbana?
2. When focusing on age at arrest, how many observations are missing?
3. Has anyone over the age of 70 and in Urbana been arrested?
4. Create a scatter plot of days in jail vs age at arrest for nonmissing ages only. What do you see?
5. Now create a scatter plot of days in jail vs age at arrest for black and white people arrested ages 35 and older. Describe the plot.

Any Questions?

Let's move on to Ch. 2 (of the textbook)
"Data Description and Simple Inference"

Outline of Chapter 2

1. Provide descriptions of interesting features within the data
   - Data visualizations (plots)
     - **proc sgplot**
     - **proc sgscatter**
     - **proc univariate**
   - Numerical summaries (tables)
     - **proc univariate**
     - **proc corr**

2. Discuss various forms of statistical inference
   - Hypothesis testing and confidence intervals
     - **proc univariate**
     - **proc ttest**
     - **proc npar1way**
     - **proc corr**

Descriptions & Visualizations

- Plots are your friend but can be an enemy if you are not careful.
- Your hope is to create visually appealing plots that you can easily interpret
- Do pay attention to axes, legends, titles, and other information in plots
- Some typical plots are: scatter plots, bar plots, box plots, line graphs, histograms, QQ plots, and density curve plots
- In SAS, **proc sgplot** produces several types of plots
- **proc sgscatter** produces multiple plots in a single panel, e.g., a scatter plot matrix
- **proc univariate**

Descriptions & Visualizations (cont.)

Box plot features include:

- minimum
- lower quartile
- mean - diamond inside the box
- median - line inside the box
- upper quartile
- maximum
- spread - the box itself
- possible outliers - dots beyond the box plot
- symmetry - when looking above and below the median

Descriptions & Visualizations (cont.)

For scatter plots, be sure to use descriptors when commenting on the visual features and relationships among the variables.

Here are some typical descriptors (adjectives):

- strong
- weak
- negative
- positive
- linear
- nonlinear
- no relationship

Let's look at plots used in the real world

Visualization from OkCupid

2009

OkCupid QuickMatch Scores

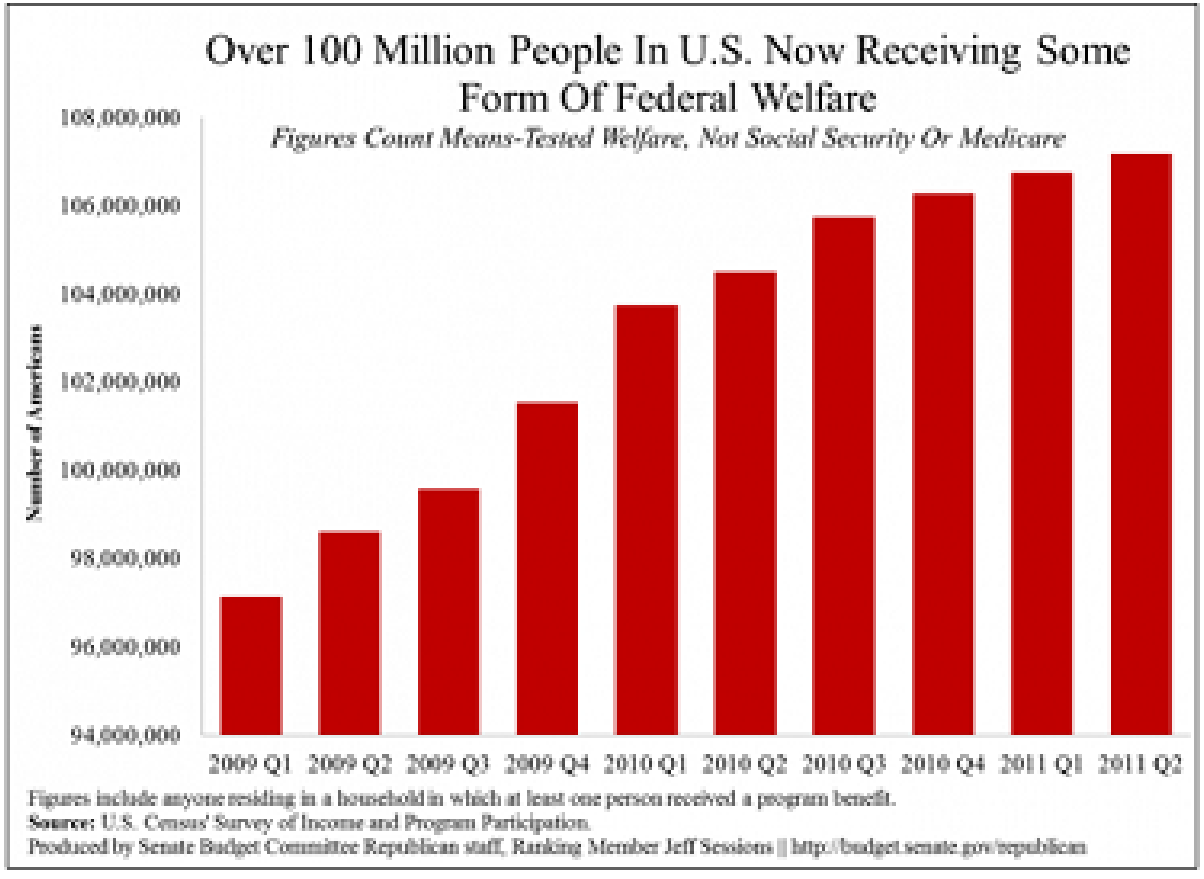
men rating women	ASIAN women	BLACK women	LATINA women	WHITE women
ASIAN men rating...	11%	-16%	-1%	7%
BLACK men rating...	3%	-3%	3%	-3%
LATINO men rating...	7%	-22%	6%	9%
WHITE men rating...	6%	-18%	2%	10%

women rating men	ASIAN men	BLACK men	LATINO men	WHITE men
ASIAN women rating...	10%	-14%	-12%	16%
BLACK women rating...	-11%	16%	-4%	0%
LATINA women rating...	-16%	-4%	11%	10%
WHITE women rating...	-12%	-6%	1%	17%

Source: OkCupid Blog, Race and Attraction, 2009–2014

Visualization from USA Today

A new chart set to be released later today by the Republican side of the Senate Budget Committee details a startling statistic: "Over 100 Million People in U.S. Now Receiving Some Form Of Federal Welfare."



Descriptions & Numerical Summaries

- Using basic statistics and other numeric information are foundational to data analysis
- Some measures we'll want to discuss are: mean, median, mode, variance, standard deviation, extreme observations, and correlation
- Central Tendency
- Spread
- Outliers
- Association

Descriptions & Numerical Summaries

Central Tendency

- The (sample) mean (or average) is the sum of all parts divided by the number of parts
- The sample median is the midpoint of a set of values
- Median is more robust than the mean
- Mean shifts as more extreme values occupy a dataset
- We know that the normal distribution is bell shaped
  - it is Unimodal (has 1 mode)
  - Mean $=$ Median
  - the distribution is symmetric

Descriptions & Numerical Summaries (cont.)

Spread

- Variability or spread of distribution also relevant for normal distribution
  - Standard deviation, $\sigma = 1$ for normal distribution
  - Variance, $\sigma^2 = 1$

Outliers

- Be suspicious of extreme observations
- Look more closely at these values to ensure they are not erroneous
- Extreme values have very low probability for the normal distribution
- $1.5 * IQR$ rule is for detecting outliers (but may not always be reliable)
  - IQR = interquartile range = $Q_3 - Q_1$

Descriptions & Numerical Summaries (cont.)

Association

- Correlation generally measures linear association of pairs of variables
  - Pearson correlation to be specific
  - Use if populations are normal or if linear association is assumed
  - range $(-1, +1)$
  - Closer to $+1$ means stronger positive linear association
  - Closer to -1 means stronger negative linear association
  - Near 0 means no linear association
- Spearman rank correlation (monotonic, nonlinear association)
  - Use if populations strongly deviate from normality or if linearity is not assumed
  - range $(-1, +1)$

Let's introduce a data set!

Video Game Sales Data

The data set is a subset of the enriched video games sales data from Kaggle. It contains the sales and attributes, including critic's review score, for video games on Sony's Playstation 2 and Microsoft's XBox 360. The video games included in this data set were randomly selected but are specifically rated as either E (everyone) or T (teen) and of the Action and Sports video game genres. North American (NA_sales) and Global sales are measured in millions.

```
data vgsales;
 infile 'filelocation/vgsalessubs.csv' firstobs=2
   dsd  ;
 input Genre $ Rating $ Platform $ Name $
   Year_of_Release Publisher $ NA_Sales
   Global_Sales Critic_Score;
run;
```

Video Game Sales Data

What questions might a data analyst have about this dataset?

What might be their goals for analysis?

Example: Video Game Sales Data

Let's do produce some visualizations and numerical summaries

Statistical Inference

Statistical Inference: Hypothesis Testing

Why?

- Often we want to set up experiments based on our beliefs
- These experiments rely on measurements that can be observed repeatedly
- Using these measurements, we can draw some conclusions that support or oppose our beliefs using probability

Statistical Inference: Hypothesis Testing (cont.)

What are we testing?

- population location parameter **proc univariate**
  - mean or median
  - one-sample **proc ttest**
  - two-sample **proc ttest** or **npar1way**
- goodness of fit test **proc univariate**
- equality of variance test **proc ttest** for two-sample case
- population correlation **proc corr**

Statistical Inference: Hypothesis Testing (cont.)

Goodness of Fit Tests

- We may be interested in testing if one population follows a specified distribution
- 3 tests are used regardless of the specified distribution
  - Kolmogorov-Smirnov
  - Anderson-Darling
  - Cramer-von Mises
- These 3 tests are based on the empirical distribution function (EDF)
- Think of the EDF as the estimated version of the cumulative distribution function

Statistical Inference: Hypothesis Testing (cont.)

Goodness of Fit Tests specifying the Normal Distribution

- If we specify the distribution as normal, the tests used are
  - Shapiro-Wilk test
  - and the three EDF-based tests mentioned previously
- For all of these goodness of fit tests, the
  - Null hypothesis: the population follows a specified distribution
  - Alternative hypothesis: the population does not follow that distribution
  - Very small p-value: there is significant evidence against the null
    - *the specified distribution may not be appropriate*

Statistical Inference: Hypothesis Testing (cont.)

When the population distribution is normal

- the parameter of interest is the mean μ
- t test is appropriate
- two-sided tests results **proc univariate** or **ttest**
- one-sided tests results **proc ttest**
- To verify whether normality is appropriate, we can look at the goodness of fit by:
  - plotting the histogram and probability plots
  - interpreting the goodness of fit tests

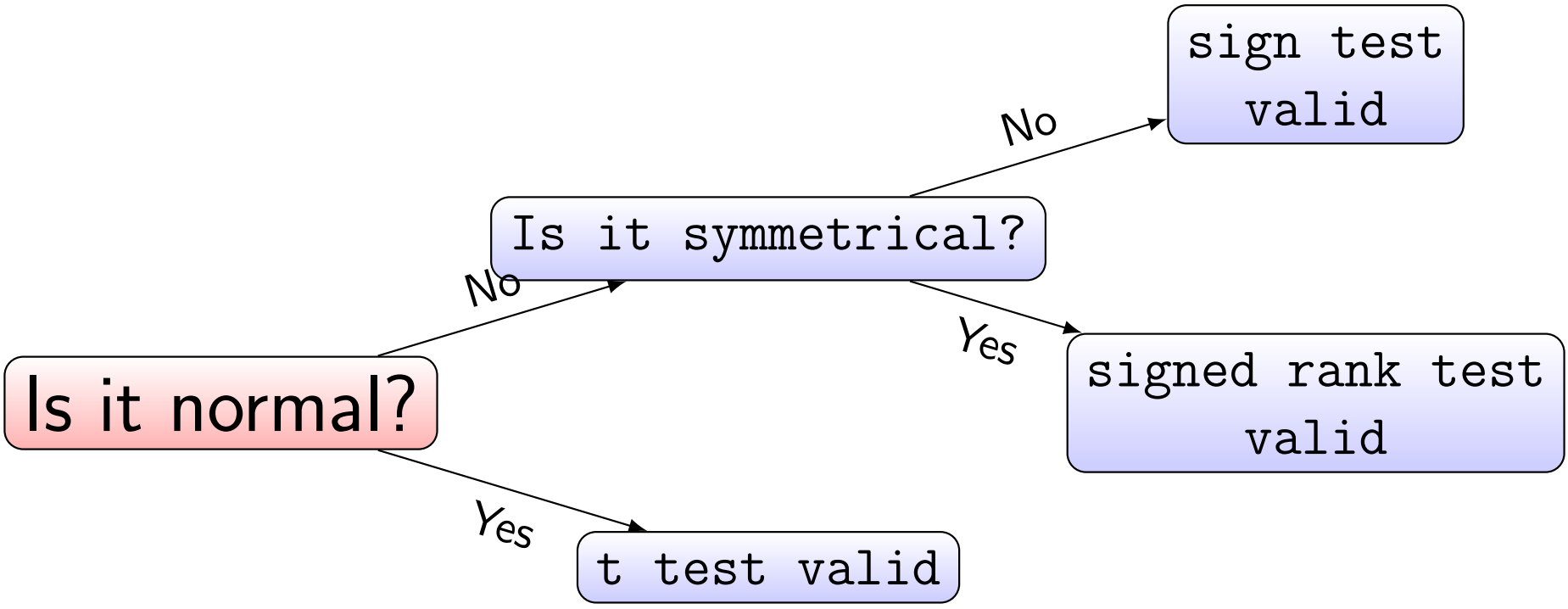
Statistical Inference: Hypothesis Testing (cont.)

When the population distribution is not normal

- the parameter of interest is the median M
- If data is symmetrical (about the median)
  - signed rank test is appropriate **proc univariate**
- If data is not symmetrical (about the median)
  - sign test is appropriate **proc univariate**
- If data is clearly skewed, then it is asymmetrical.
- If data has mean, median, and a single mode all at the same value, then it is symmetrical.
- Data could be symmetrical so long as what's happening above and below the median looks similar.

Considering One Distribution

For one-sample hypothesis testing,



SAS Programming for Testing One Distribution

Here's an example of using **proc univariate**.

```
proc univariate data=dataName mu0=aNumber
   normaltest cibasic alpha=aDecimal;
 var variable1 variable2;
 ods select TestsForLocation TestsForNormality
   BasicIntervals;
run;
```

Here's an example of using **proc ttest**.

```
proc ttest data=dataName h0=aNumber;
  var variable1 variable2;
  ods select ConfLimits TTests;
run;
```

Statistical Inference: Hypothesis Testing (cont.)

When both population distributions are normal

- the parameter of interest is the difference in means
- the order of the difference is that the first category listed is population 1 and the second listed is population 2.
  - You can also tell by looking at the Diff(1-2) value.
  - $H_0 : \mu_1 - \mu_2 = 0 \quad \text{vs} \quad H_A : \mu_1 - \mu_2 \neq 0$
- two-sample t test is appropriate **proc ttest**
  - Pooled results are valid if variances *are equal* (assumed or tested)
  - Satterthwaite results are valid if variances are not equal
- The equality of variances (homogeneity) test indicates which situation is appropriate
  - H_0: both population variances are equal
  - H_A: they are not equal

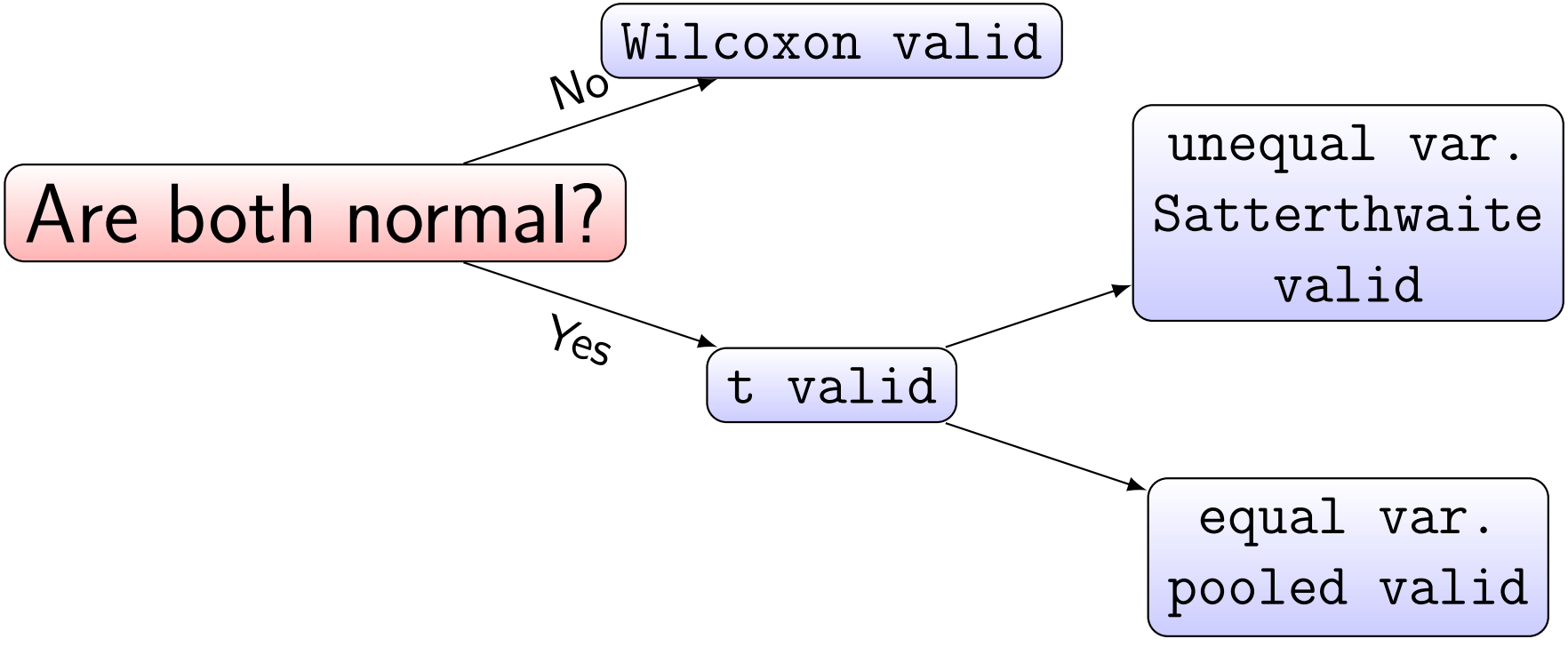
Statistical Inference: Hypothesis Testing (cont.)

When one of the two population distributions is not normal

- the ranks are computed using the Wilcoxon rank-sum test
  - Looking at the Wilcoxon Scores table and the Wilcoxon Two-Sample Test statistic
  - H_0: the two populations come from the same distribution
  - two-sided H_A: one of the populations tends to have larger values
- Wilcoxon rank-sum test is appropriate **proc npar1way ... wilcoxon**

Considering Two Distributions

For two-sample hypothesis testing,



SAS Programming for Testing Two Distributions

Testing for the mean differences between the two classes, when both are normal.

```
proc ttest data=dataName;
 class classVariable;
 var variable1;
 ods select TTests Equality ConfLimits;
run;
```

Testing for the mean differences between the two classes, when one of the two is not normal.

```
proc npar1way data=dataName wilcoxon;
 class classVariable;
 var variable1;
 ods select WilcoxonScores WilcoxonTest;
run;
```

Statistical Inference: Hypothesis Testing (cont.)

Testing for Population Correlation

- the populations are normally distributed or linearly associated
- the default correlation measure is based on the Pearson calculation
- If the populations deviate severely from normality or if the relationship is nonlinear, try using Spearman rank correlation
- Null hypothesis: a pair of variables is not correlated
- Alternative hypothesis: that pair is correlated
- Very small p-value: *the pair of variables have some correlation*

Data Transformations

- Often, certain variables may be far from normal (often skewed)
- Some examples include salaries and light intensity
- The log transformation is typically used
- In SAS, creating a new variable is done through an assignment statement as in: `variable=expression`
- A log transformation would look like: `newVariable = log(oldVariable)`

Exercise: Video Game Data

1. Make a scatter plot of numeric variables of interest to you. Describe what you see.
2. Make a histogram of those same variables from (part 1). Is either distribution normal?
3. What do the 4 tests for normality determine for each variable?
4. Test for the population correlations of those variables (from part 1) by platform separately. What conclusions can we draw?
5. Create a log transform of the NA sales variable. Use `lnNAsales = log(NA_sales)`. Was there clear skewness before or after the log transformation?
6. Perform the appropriate test for the mean difference in NA_sales for the two ratings.