Chapter 1
Exploring Statistics

Introduction

1.1 Sources of Data

1.2 Describing Data

1.3 Probability and Statistical Inference

Introduction

- Data and Statistics
- Statistical Studies

Data and Statistics

Data are numerical or qualitative descriptions of the objects that we want to study.

Statistics has traditionally focused on a collection of methods that translate data into answers to our questions.

A **statistical study** starts with a general topic of interest and ends with answers to specific research questions. In between, we will collect or find data and subject it to statistical analysis.

The Six (Four?) Steps in a Statistical Study

1. **Topic of interest.** Develop a general idea of the area to be studied.

2. **Research questions.** Translate general ideas into research questions.

3. **Collect or find data.** The research questions help us decide on the type of data to be collected.

4. **Data.** Carefully collected data is vital to ensure reliable results.

5. **Analysis.** Use statistical methods to process data and generate answers to the research questions.

6. **Conclusions.** Results are presented to maximize comprehension.

I call this the PCAI process (1,2=P; 3,4=C; 5=A; 6=I) : Pose the question; Collect data; Analyze the data; Interpret the results…

1.1 Sources of Data

- Sampling from a Population
- Comparative Studies

Sampling from a Population

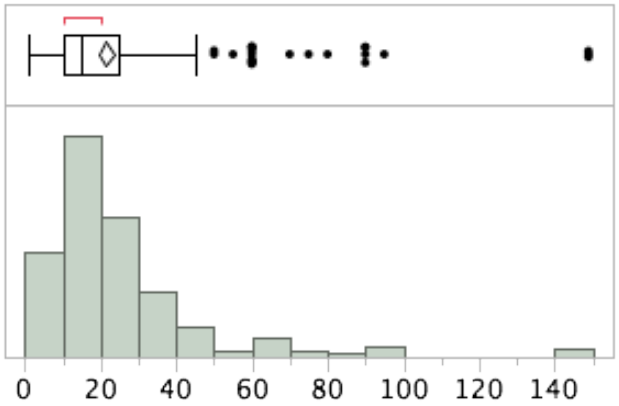
To answer the question “How much time does a resident of New Hanover or Brunswick county spend getting to work each day? (i.e, what is the typical commute time of workers in NH and Brunswick Counties?)” we **sample** from the **population** of workers in these counties and ask them – this is done by the US Census Bureau in a regular way using the American Community Survey (http://www.census.gov/newsroom/releases/archives/american_community_survey_acs/cb13-215.html?intcmp=sldr2)

Good sampling methods allow us to make good inferences about the population.

This is a fundamental concept in statistics – the next slide shows the analysis of the sample…

Distributions

JWMNP



Quantiles

100.0%	maximum	149
99.5%		149
97.5%		82.75
90.0%		40
75.0%	quartile	25
50.0%	median	15
25.0%	quartile	10
10.0%		5
2.5%		2
0.5%		1
0.0%	minimum	1

Summary Statistics

Mean	21.469626
Std Dev	20.108174
Std Err Mean	0.9719653
Upper 95% Mean	23.380058
Lower 95% Mean	19.559194
N	428

Some of the analyses that you see here that we'll learn about in STT 215 are:
Boxplots, outliers, histograms, quantiles, moments, standard errors, confidence intervals, etc.
Not only how to compute them (mostly using technology), but also how to interpret the output when we allow JMP to do all the hard work!

Comparative Studies

Another common setting for a statistical study involves the **comparison** of two groups.

Selection of the groups should be done using **randomization.**

Random selection of individuals from a population allows us to answer questions about the population from which the sample was drawn. A random sample is more likely to be representative of the population than a so-called “convenience” sample or a “voluntary” sample.

Similarly, in a comparative study, the *random assignment* of subjects to different treatments allows us to make a valid comparison of treatments.

Here’s an interesting example from the MIT Poverty Action Lab concerning cash transfers to the poor…

Here's the abstract of the paper – I've edited out some of the econ stuff to concentrate on the PCAI aspects…

Household Response to Income Changes: Evidence from an Unconditional Cash Transfer Program in Kenya∗ by Johannes Haushofer†, Jeremy Shapiro‡ November 15, 2013

Abstract

This paper studies the response of poor rural households in rural Kenya to large temporary income changes. Using a randomized controlled trial, households were randomly assigned to receive unconditional cash transfers of at least USD 404 from the NGO GiveDirectly. … We randomized at both the village and household levels; further, within the treatment group, we randomized recipient gender (wife vs. husband), transfer timing (lump-sum transfer vs. monthly installments over 9 months), and transfer magnitude (USD 404 vs. USD 1,520). … Intriguingly, recipient gender does not affect the household response to the program. … Transfer recipients experience large increases in psychological well-being, and several types of transfers lead to reductions in levels of the stress hormone cortisol. Together, these results suggest that unconditional cash transfers have significant impacts on consumption and psychological well-being.

1.2 Describing Data

- Describing Data for One Variable
- Describing Data for Relationships

Numerical Summaries of Data

Common numerical summaries involve measures of center and measures of spread.

The most common measure of center is the **mean,** sometimes called the *average.* It is found by dividing the sum of the data values by the number of data values.

A measure of spread often used in conjunction with the mean is the **standard deviation.**

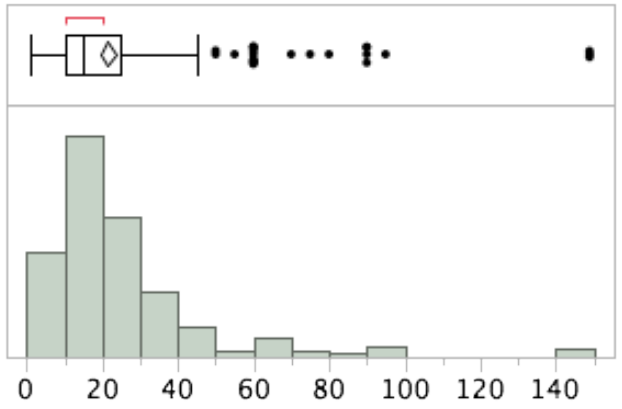
These numerical summaries and others will be discussed in Chapter 3.

Graphical Summaries of Data

Graphical presentations of data are used to visualize the data for the benefit of better understanding the data. We will see that analysis of center, shape, and spread of graphs is key to understanding the data and drawing conclusions to research questions.

Distributions

JWMNP



Quantiles

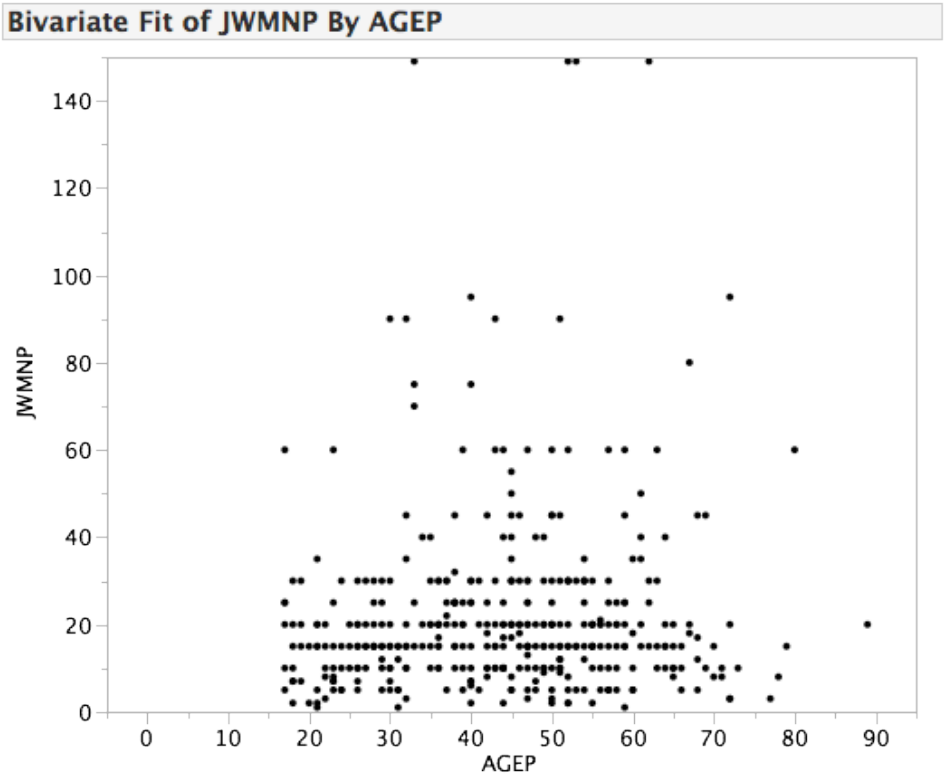
100.0%	maximum	149
99.5%		149
97.5%		82.75
90.0%		40
75.0%	quartile	25
50.0%	median	15
25.0%	quartile	10
10.0%		5
2.5%		2
0.5%		1
0.0%	minimum	1

Summary Statistics

Mean	21.469626
Std Dev	20.108174
Std Err Mean	0.9719653
Upper 95% Mean	23.380058
Lower 95% Mean	19.559194
N	428

Describing Data for Relationships

The relationship between two variables is often investigated through statistical analysis. For example, is there an association between the commute time of a resident of NH/Brunswick County and their age?? We can begin to answer this question through the use of another graphical summary: **the scatterplot.**



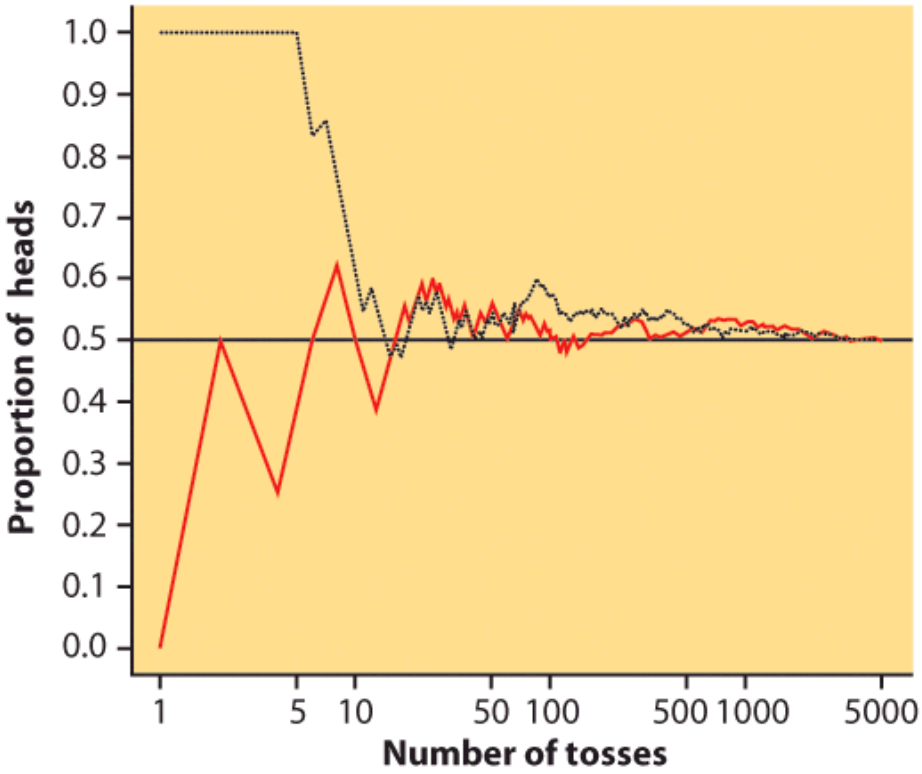
The pattern of the points may lead to an answer to the question. A numerical summary, the **correlation,** will be used to support our conclusion and may allow the use of a **least-squares line.**

1.3 Probability and Statistical Inference

- Probability Models
- Probability and the Practice of Statistics

Probability Models

The use of dice, spinners, and coins are prevalent in statistics courses as a way to introduce a **model** for many situations. Tossing coins is easy to do and understand. Two outcomes, heads and tails, are easy to tabulate, and the proportion of heads (or tails) is easy to calculate. A key characteristic of this experiment is that the outcome of one toss does not influence the outcome of subsequent trials – the so-called **independence of the trials.**



Probability and the Practice of Statistics

When we perform a statistical study, the data we collect and analyze tell us something about the sample we used in our study. For most statistical studies, we are interested in more. For example, what can our data tell us about the population from which we drew the sample?

The tools of **statistical inference** allow us to answer this question and more. The two primary tools discussed in this course are **confidence intervals** and **significance testing.**

Confidence intervals rely on an estimate of some quantity from a sample. Using this estimate, we give a measure of our uncertainty in our estimate called the **margin of error.**

We use the estimate and margin of error to form an interval that may or may not contain the true value.

Assignment for next time:

Go to the Stats Portal and sign up…talk with me if you have particular problems (financial aid?)…

Read Chapter 1 in the book and work your way through the examples – don't worry too much about the details and the terminology because we'll be spending the whole semester on the main topics discussed in this chapter …

We'll start Chapter Two next time on Design of Experiments and Sampling…