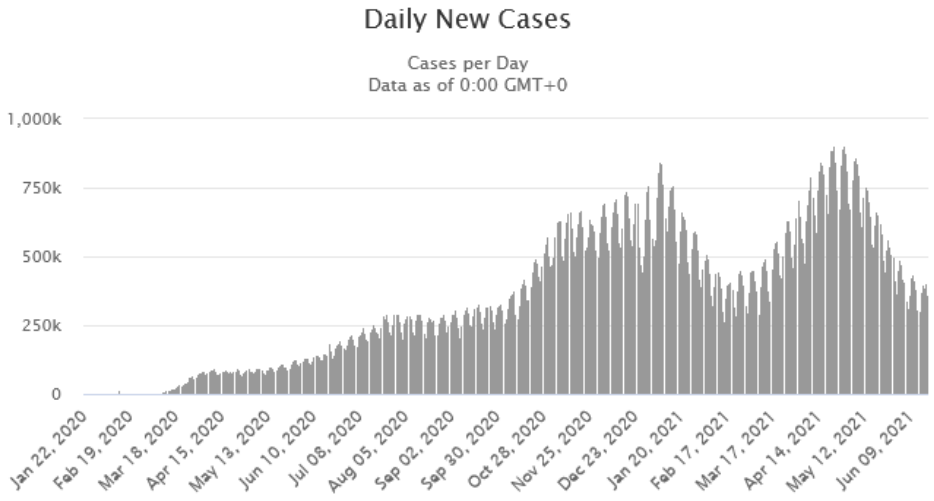
MATH 380

Elementary Probability and Statistics

Dr. Christian K. Hansen



The Three-Step Process in Statistical Analysis.

Step 1: Collecting the Data **Chapter 3**

- *Sampling design*
- *Experimental design*
- *Observational study*

Step 2: Summarizing/Organizing the Data **Chapter 1**

- *Descriptive statistics*
- *Graphical displays*
- *Numerical measures*

Step 3: Drawing Conclusions from the Data **Chapters 4-8**

- *Inferential Statistics*
- *Estimation of unknown parameters*
- *Hypothesis testing*

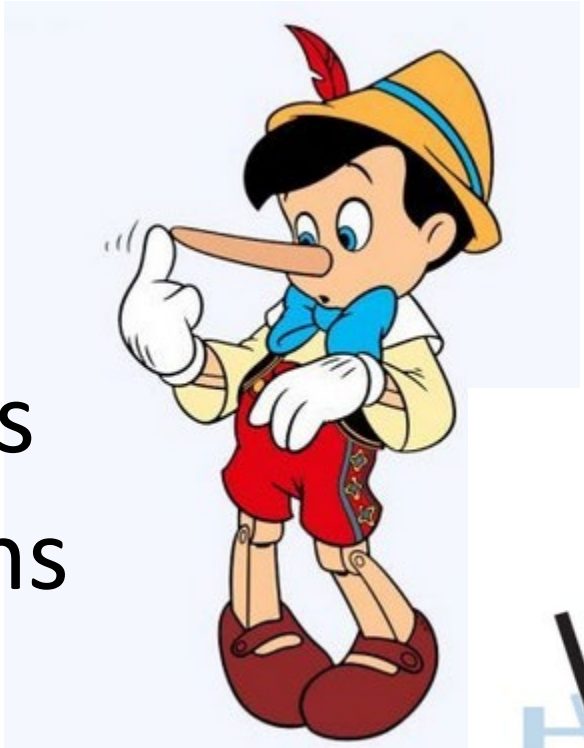
What is Statistics ?

- Statistics is the science of collecting, organizing, summarizing, and drawing conclusions from data.
  (or “Science of Learning from Data”)
- Statistics is an inductive science.
  (Conclusions are expected to be correct in most cases, but are not guaranteed to be correct).

An Old Saying …

There are

- Liars
- Damn Liars
- Statisticians



Looking at Data—Distributions

Introduction

Statistics is the science of learning from data. Data are numerical or qualitative descriptions of the objects that we want to study. In this chapter, we will master the art of examining data.

We begin in **Section 1.1** with some basic ideas about data. We will learn about the different types of data that are collected and how data sets are organized.

Section 1.2 starts our **process** of learning from data by looking at graphs. These visual displays give us a picture of the overall patterns in a set of data. We have excellent software tools that help us make these graphs. However, it takes a little experience and a lot of judgment to study the graphs carefully and to explain what they tell us about our data.

Section 1.3 continues our process of learning from data by computing numerical summaries. These sets of numbers describe key characteristics of the patterns that we saw in our graphical summaries.

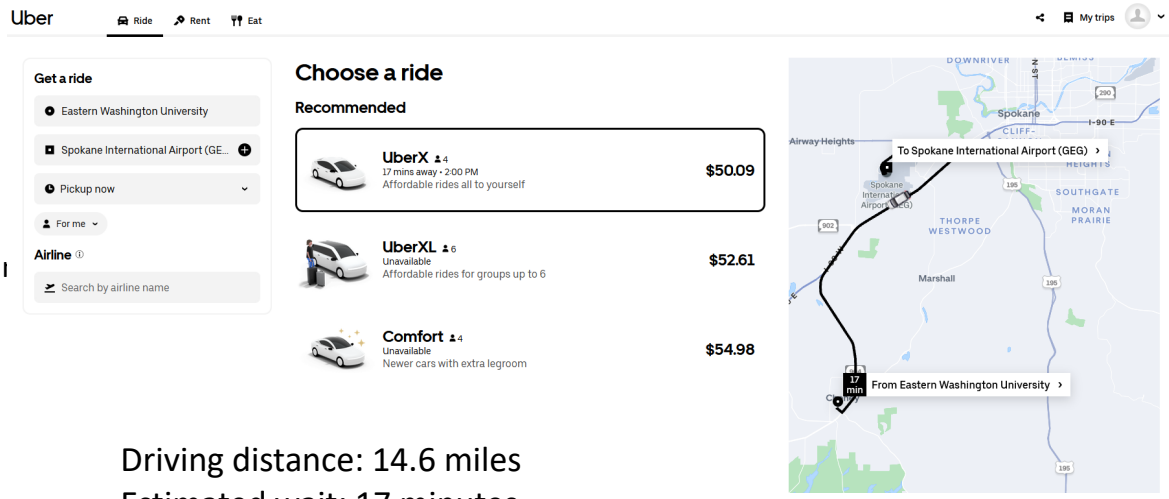
The final section in this chapter helps us make the transition from data summaries to statistical models that are used to draw conclusions and to make predictions. Specifically, we learn about using **density curves** to describe a set of data and are introduced to the **Normal distributions**. These **distributions** can be used to describe many sets of data that we will encounter. They also play a fundamental role in many of the methods of statistical analysis.

Data Science – Some History

- The word “data” is an ambiguous term dating back to the 1600s
- In statistical applications “data” in commonly thought of as a collection of information often represented as observations recorded from one or more variables classified as either quantitative or categorical.
- Prior to computers “data” were mostly recorded and kept on paper. Physical libraries represented the largest collections of “data.”
- Through the computer era (1950 and beyond) “data” became “synonymous” with digital data, i.e. information that can be stored on a computer.
- Though modern times (the Big Data era), “data” is most commonly referring to masses of digital content taking on various formats, including text, pictures, audio and video that is transmitted or streamed between two or more computing devices.
- In the data centered world, virtually everything we earn, own or trade is represented by “data”.
- “Data” is the new universal currency.
- Data management is the management of virtually everything

Case Study: Transportation-as-a-Service

- Uber Ride Explained
  Rider's cost :
  - Base fares
  - Tolls and surcharges
  - Booking fee
  - Surge pricing (for peak and high demai
  - Route-based adjustments
  - Promotions and subscriptions
  - Wait time and cancelation fees
  - Tips



- Driver's earnings:
  - Compensation Based on Actual route driven
  - Service fee deducted (proprietary)
    Estimated 25-30% (not verified)
  - Driver is responsible for vehicle operating cost (~ $0.81/mile)

Driving distance: 14.6 miles
Estimated wait: 17 minutes
Estimated drive time: 19 minutes
Cost per mile: $3.43
Tolls: N/A
Estimated vehicle expense: $24
Estimated driver net earnings: $11+tip

https://www.uber.com/us/en/marketplace/pricing/

Case Study: STA Bus Route Efficiency

- Background
  - The Spokane Transit Authority (STA) operates busses in the urban area of Spokane, WA – a midsize Northwestern US city with a population of approximately 230,000
  - STA operates 53 routes providing 9 million passenger rides annually (2023) covering 1700 stops in the area and with services provided week days between 5:00 AM and 11:00 PM and limited day hours during weekends.
- Cost/Revenue
  - Standard fare is $2.00 per ride (includes transfers).
    Various discounts and subscription options are available to riders who qualify.
    Effective Fall 2022 STA adopted a zero-fare option for Youth.
    Student, faculty and staff of Eastern Washington University are covered 100% by an annual contract between STA and the University
    (students are charged a flat $25 quarterly transportation fee - $38 per semester)
  - Cost recovery was 8.89% in 2023 (subsidized by tax payers)
  - Direct Cost per Revenue Mile: $2.45 (in 2023)
  - Direct Cost per Revenue Hour: $92.35 (in 2023)
- Research Goals
  - Analyze ridership data to study route efficiency and propose improvements
  - Model STA route system using PHM tools

Spokane Transit

* Collaborator:
William Helman
EWU Data Science Student

What is Probability Theory ?

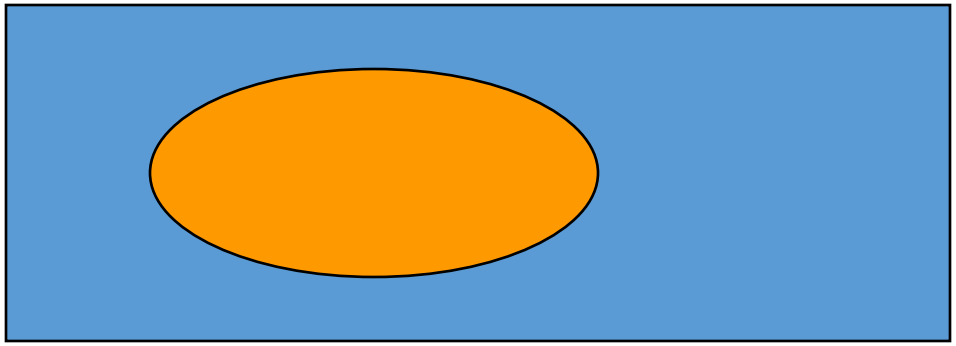
- Probability Theory is the study of randomness.
- The outcome of an experiment is random when the observed outcome is unpredictable, but there is a regularity in the long-term behavior (when the experiment is repeated many times).
- Probability Theory is a deductive science. (Probabilities calculated under correct assumptions are always correct).

Comparison of Probability Theory & Statistics

Probability Theory: Properties of the population are assumed known. We answer questions about the sample based on these properties.

Population: The collection of all subjects of interest to our study

Sample: The collection of subjects actually used in our study



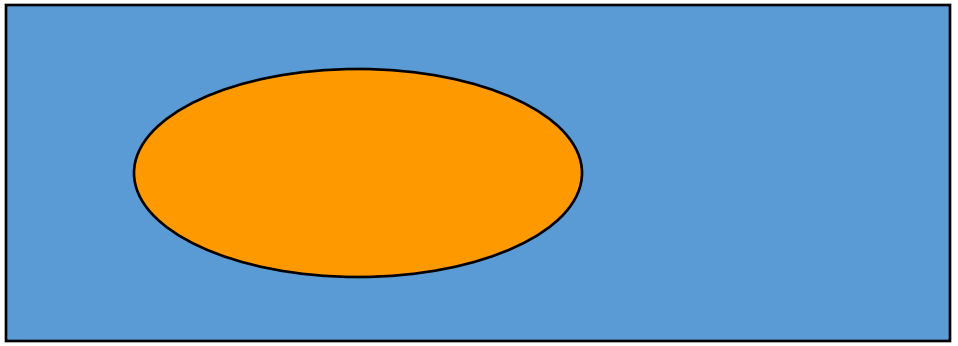
Example: Suppose 80% of all American adults are employed full time.
If we poll a sample of 10 randomly selected American adults, how likely is it that less than half are employed full time?

Comparison of Probability Theory & Statistics

Statistics: We use information in the sample to draw a conclusion about a more general population

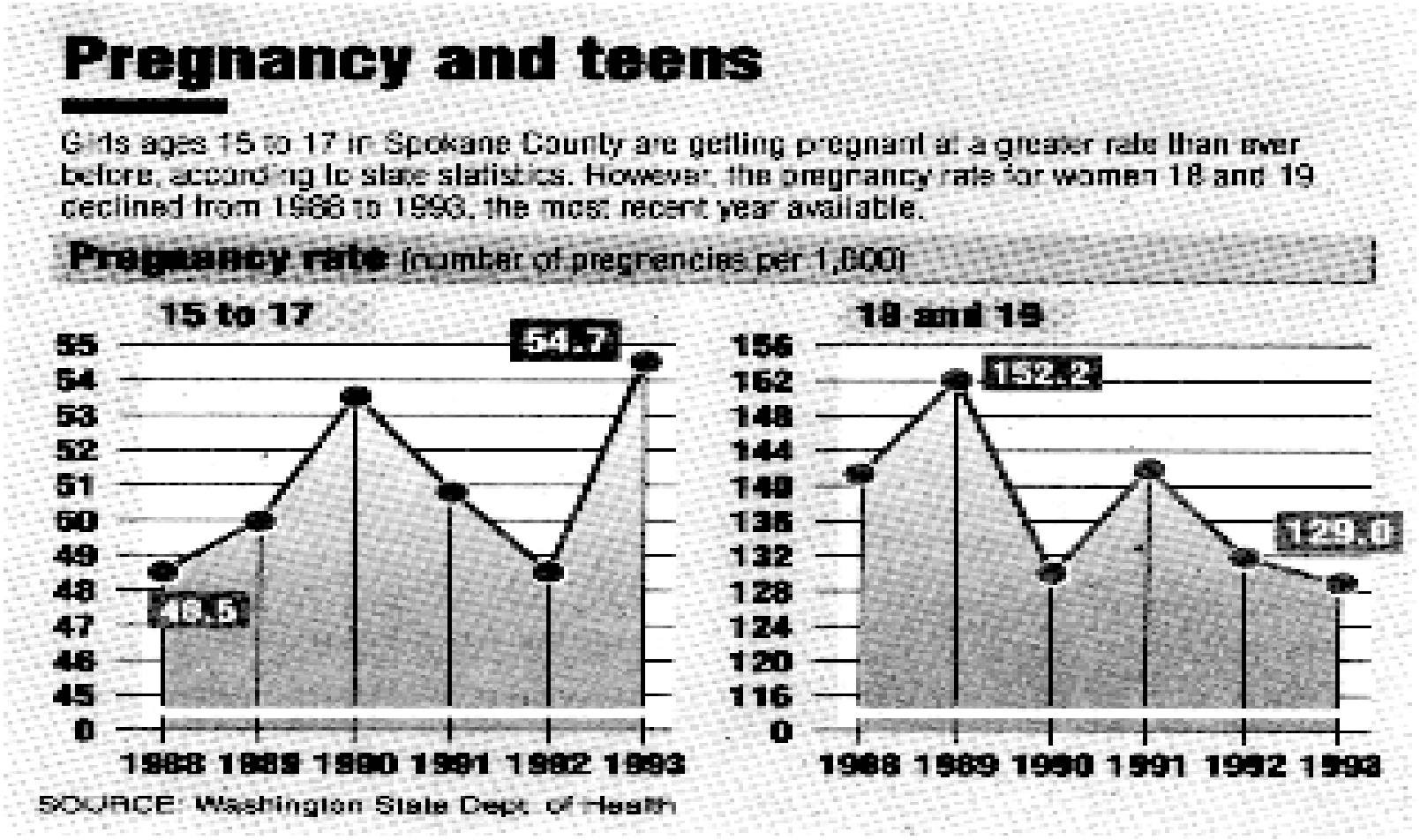
Population: The collection of all subjects of interest to our study

Sample: The collection of subjects actually used in our study



Example: If we ask 20 randomly selected American adults if they are employed full time and 16 of them are employed full time, what can we conclude about the entire population of American adults?

Pregnancy and Teens



Pregnancy and Teens

Pregnancy rates increase among younger local teens

By Jeanette White
Staff writer

Pregnancies are on the rise among Spokane County girls from 15 to 17 years old, but fewer older teenagers are getting pregnant.

Of girls who do become pregnant, far fewer are having abortions than in the past.

"More kids are choosing to keep their children," said Melinda Harmon, the state's teen pregnancy prevention coordinator.

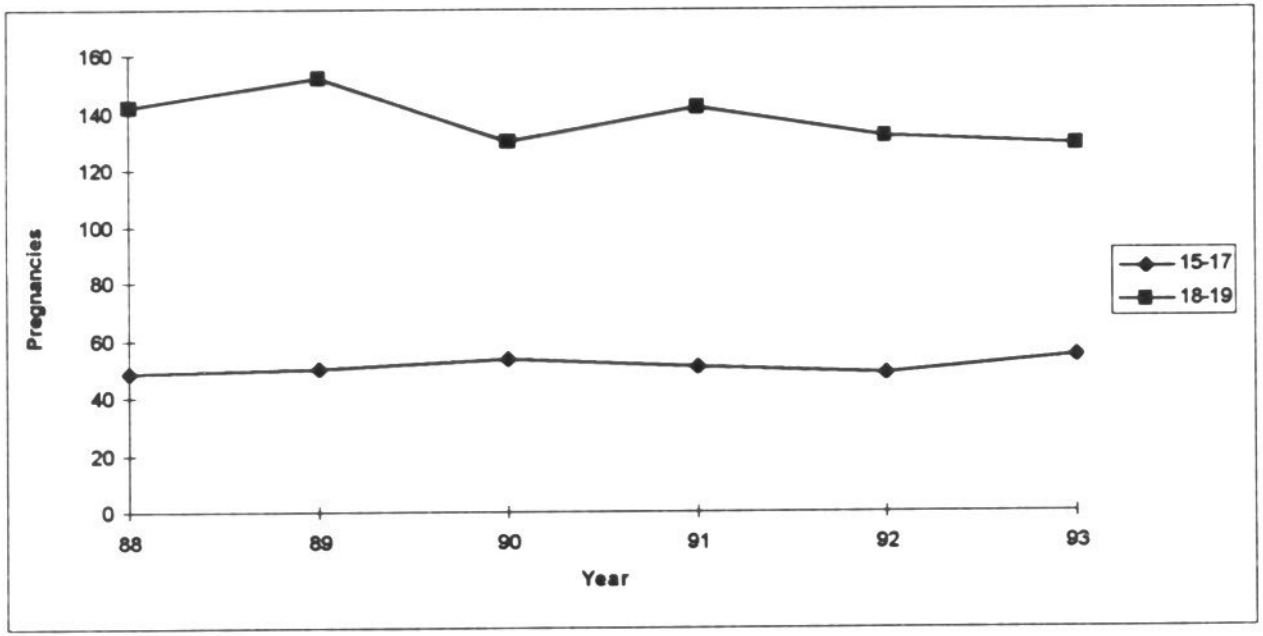
One in 21 Spokane County teens 15 to 17 years old became pregnant in 1988. That rose to one in 18 in 1993, according to the state Health Department's latest statistics.

Among 18- and 19-year-old girls, one in seven got pregnant in 1988, compared with one in eight in 1993.

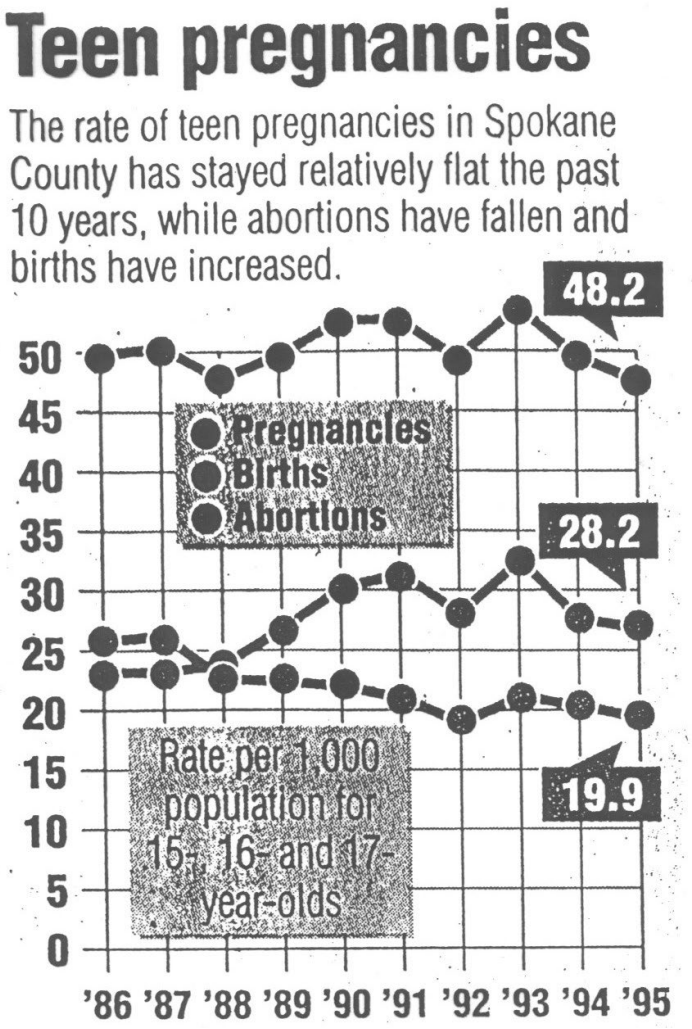
In a striking change, nearly half of pregnant Spokane County teens 15 and older had abortions in 1988, compared with just more than a third

Continued: **Pregnancy/A12**

Pregnancy and Teens (cont.)



Pregnancy and Teens (cont.)



SOURCE: Spokane Regional Health District

Staff graphic