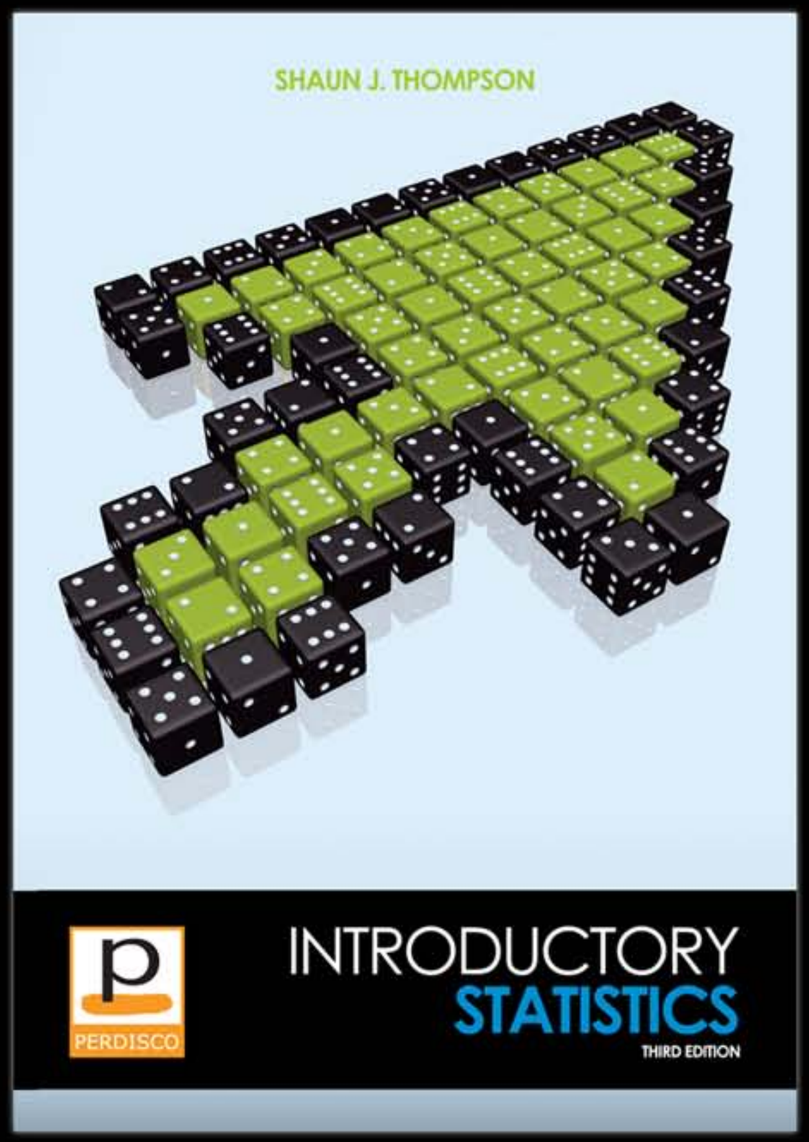


These lecture slides are designed to accompany:

Introductory Statistics, Third Edition

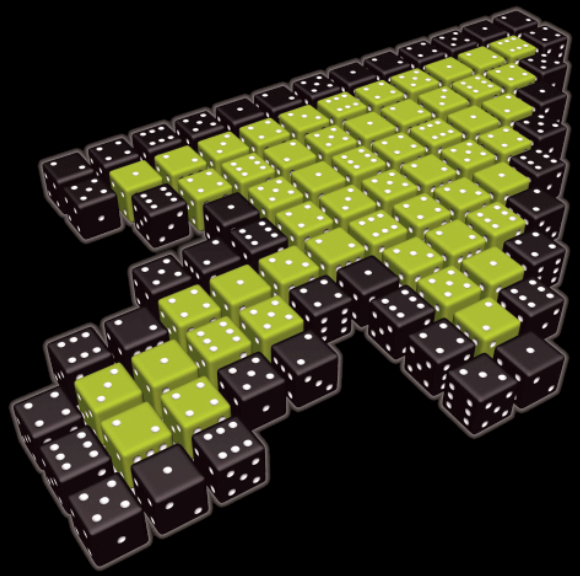
Other features include:

- Chapter summary videos
- MP3 audio podcasts
- VirtualTutor e-learning
- AntiCheat and AutoGrade homework
- Detailed instructor resources

To find out more, visit: perdisco.com/stats

Chapter 1

Introduction to statistics



Statistical concepts

- Why study statistics?
- Suppose you're in a situation where you want to answer a question or explain something
  - e.g. Do men earn more than women?

- You can collect and analyze data
- Then, you use the data to draw conclusions

- That's statistics!

Populations and samples

- **Population** is the collection of all things we are interested in
  - e.g. salaries across the country for both men & women
- **Sample** is the subset that we collect and analyze
  - e.g. survey of men and women, asking about salaries

- We study the sample so that we can know more about the population

Parameters and statistics

- These are what we want to know ***about*** populations and samples
- **Parameters** measure population characteristics
  - e.g. national average salary for a female
- **Statistics** measure sample characteristics
  - e.g. average salary for a female in the survey
- We calculate a statistic by analyzing a sample
- Then estimate a parameter based on statistic

Inference and description

- **Descriptive** statistics:
  - collecting, presenting, measuring sample data
  - typically very precise and certain
    - e.g. average female salary in survey is exactly $57,000

- **Inferential** statistics:
  - drawing conclusions based on descriptive findings
  - typically imprecise and uncertain
    - e.g. national average female salary is probably around $57,000

Why not go straight to the population?

- If we want to know about the population, why bother with a sample?
- Why not study the whole population directly?
- Good question. Answer: Sometimes we do!
  - e.g. national election or census
- But this is costly, and sometimes impossible
- You can get away with using sample data, provided you know what to do with it!

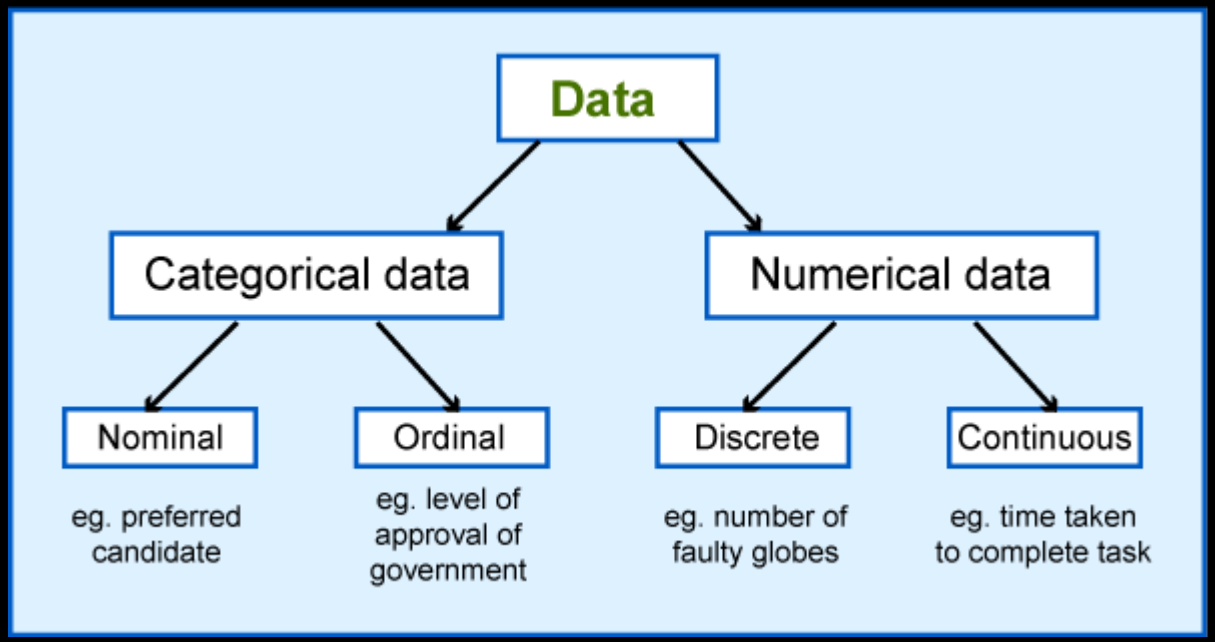
Data

- Basic definition: data are pieces of information

- A **variable** is an observable characteristic
  - e.g. annual salary
- **Data** are the observed values of a variable
  - e.g. survey 50 people, get 50 annual salaries, have 50 data values

Data types

- Two main types: categorical and numerical
- Each type can be broken down further



Categorical data

- Takes values from a number of qualitative options
- These options are known as **categories**.

- Data is **nominal** if no natural order to categories
  - e.g. preferred political candidate from a list of three
- Data is **ordinal** if there *is* a natural order
  - e.g. qualitative level of approval of a given candidate (e.g. strongly approve, approve, disapprove …)

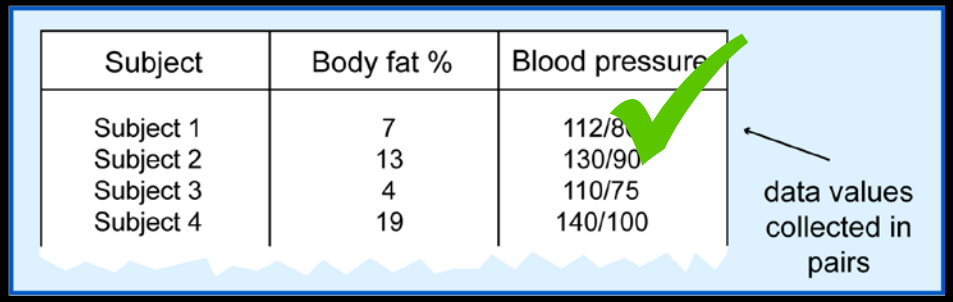
Numerical data

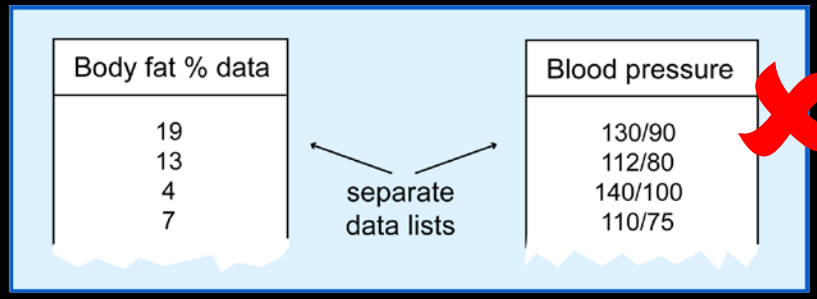
- Takes values that are quantitative, i.e. numbers
  - e.g. recording the number of faulty products in a batch

- Variable is **discrete** if its values are countable
  - e.g. number of faulty products in a batch
- Variable **continuous** if values are on a spectrum
  - e.g. time taken to complete a task

Data from more than one variable

- Can look at more than one variable at a time
- Usually looking for a **relationship**
  - e.g. are body fat % and blood pressure related?
- Must collect data in pairs to study this





- Note: Relationships are not always cause & effect!

Data sources

- Either primary or secondary
- **Primary** source: the person using and analyzing the data actually collected it
  - e.g. holding a survey and analyzing the results
- **Secondary** source: the person using and analyzing the data did not collect it themselves
  - e.g. accessing data from a government archive
- Secondary is easier and cheaper
- But it might not be available

Collecting data

- Data from a primary source must be collected
- The method of collection becomes important
- Sample must faithfully represent population

- Two main methods of data collection: **observational studies** and **experiments**
- Both involve observing responses from **subjects**
- They differ in how we treat the subjects

Observational studies

- Data are observed and recorded based on responses from subjects
- No effort is made to affect those responses
- That is, they are just observed
  - e.g. questionnaire survey
  - e.g. measure and record blood pressure in 100 people

Experiments

- Used when you are studying a causal relationship
- That is, looking to see if an **explanatory variable** has an effect on a **response variable**

- Split subjects up into different groups
- Apply different levels of the explanatory variable
- See if there is any difference in response variable

Example of experiment

- Studying effect of **caffeine** on **blood pressure**

- Split 100 people up into two groups of 50
- Give one group caffeine, the other a placebo
- Measure blood pressure in everyone, and see if there is any difference between the groups

Comparing data collection methods

- Which method is best?
- It depends …

- Observational studies are used to study:
  - one variable
  - multiple unrelated variables
  - non-causal relationships
- Experiments are used to study:
  - causal relationships

Comparing data collection methods (cont'd)

- Experiments **can** prove cause and effect
- Observational studies can't
- Why? Because observation has no control over the variables

- But: experiments can have more ethical issues than observational studies
- Have you the right to **treat** the subjects that way?

Sample design

- Observational studies involve samples
- These samples must be carefully selected!

- The sample should represent the population
- If it doesn't, the sample is **biased**

Bias

- **Non-response bias**: subjects fail to respond
  - e.g. some people refuse to answer a survey
- **Undercoverage**: the population sampled from (i.e. the **frame**) doesn't cover whole population
  - e.g. missing out on cell phone users in phone survey
- **Non-random sampling**: sample not selected in random fashion
  - e.g. using your friends as a survey sample

Making a sample random

- Non-random sampling: very common form of bias
- Need good **sampling plan** to ensure it is random

- Ideally: Every sample should stand an equal chance of being selected from the population

- Such samples are **simple random samples**

Simple random sample

- Say you want to select 200 people from 5000

1. Assign a number to each person, 1 to 5000
2. Randomly select 200 numbers, 1 to 5000
3. Choose the people assigned to these numbers

- This is very simple, but time-consuming

Systematic sample

- Used to save time on simple random sampling – only select **one** sample item at random

1. Assign a number to each person, 1 to 5000
2. Split list into 200 smaller lists (25 in each)
3. Select **one** random number between 1 and 25
4. Select the person assigned to this number, and every 25th person thereafter

Stratified sample

- Used if you know the population contains groups (**strata**) that will respond differently

- Say your 5000 people are 70% blue-collar and 30% white-collar
- Force sample of 140 blue-collar & 60 white-collar

- Ensures sample represents demographics properly

Cluster sample

- Used if you know the population contains a group (**cluster**) that acts as a mini-population
  - e.g. one suburb in a town of many suburbs
- You can use the cluster as your sample
- Or you can sample from the cluster

- Careful! Does the group **really** represent the whole population?
  - e.g. does suburb represent whole town, or is it different?

Experimental design

- Experiments study the effect of one variable on another variable
- **Explanatory variable:** the variable causing the effect
- **Response variable:** the variable being effected

Applying treatments

- The explanatory variable is also called a **factor**
- The values of this variable are called **levels**
- Split subjects into different groups
- Apply a different level of the factor to each group
- Record and compare responses between groups
  - e.g. studying the effect of **room light** on **mood**
    - two levels: dim light and ample light, so two groups
    - apply dim light to one group, ample light to other
    - observe and compare moods between groups

Randomizing the groups

- Important to **randomize** the groups
- That is, assign subjects to the groups randomly

- Why? Because it minimizes sample variation
- Example: Studying effect of room light on mood
  - moods will differ, regardless of room light
  - randomly assigning people to two groups increases the chance that this will ‘average out’
  - so any effect of light on mood will be more pronounced

Blocking the groups

- Can be appropriate to **block** the groups

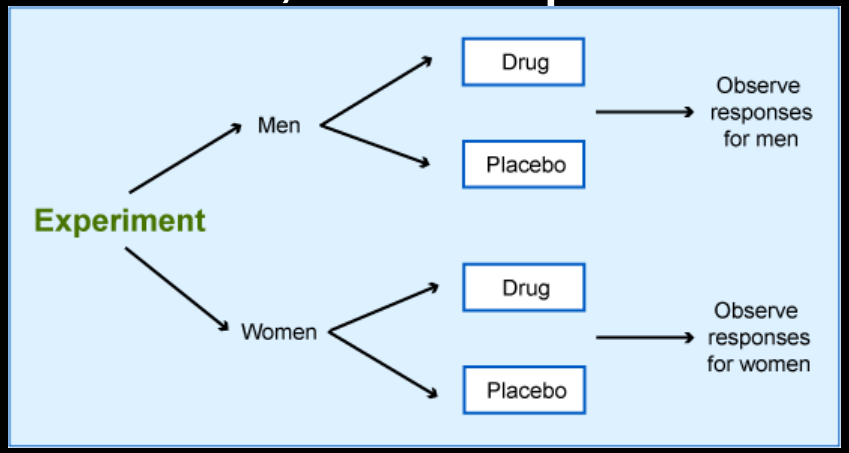
- Can subjects be broken down into demographics (**blocks**) that respond differently?
- If so, break subjects into different blocks first
- Then each block becomes its own experiment

- This makes experiments more powerful
- Removes effect of factors that don't interest us

Example of blocking

- Imagine studying the effect of a drug on iron levels
- But men and women have different iron levels

- So split subjects up into men and women first
- Then run an experiment for men, and separate one for women



Placebos and control groups

- When testing drugs, one group gets a placebo
- This group is called the **control group**
- Control group is **any** group that is left ‘untreated’

- Important to use control groups
- Allows you to say that the explanatory variable **is** the cause of any change
- Otherwise, some other factor might be causing it

Blinding the experiment

- If subjects know the treatment they are receiving, this might impact results
- That is why placebos are used!
- Subjects should be **blinded** to their treatment

- But people administering treatments or observing results might also impact them!
- So anyone involved in any of this must be blinded
- This is known as **double blinding** the experiment

Chapter 1 complete

We discussed:

- Statistical concepts
- Data
- Collecting data
- Sample design
- Experimental design