

# STATISTICS 7

Professor Jessica Utts

<http://www.ics.uci.edu/~jutts/7>

## STATISTICS 7, FALL 2009

**Lecture:** MWF, 1:00 – 1:50, 100 Engineering Lecture Hall  
**Instructor:** Professor Jessica Utts, 2212 Bren Hall, 929-824-0649, [jutts@uci.edu](mailto:jutts@uci.edu)  
**Office Hours:** Tues 2-4 or by appointment; *Until 10/16:* Fri 10:30-noon; *After 10/16:* Mon 10:30-noon (to correspond with homework due dates).

**TAs and Discussions:** Discussions are Mon in 134 Steinhaus Hall at 3, 4, 5 and 6pm. You must attend the discussion for which you are enrolled, not one at another time. See “Grading” for additional details about format of the discussions and attendance.

<b>TA:</b>	<b>Yan He:</b> Discussions 1 and 2 (3pm, 4pm)	<b>Thao Duong:</b> Discussions 3 and 4 (5pm, 6pm)
Email	<a href="mailto:yhe@uci.edu">yhe@uci.edu</a>	<a href="mailto:tduong@uci.edu">tduong@uci.edu</a>
Location	2013 Donald Bren Hall	2013 Donald Bren Hall
Office Hrs	Mon, 5-6pm and Thurs, 2:00 - 4:00pm	Mon, 3-5pm and Fri, 2:00-3:00pm

### Office hours by day:

<b>Mon, 3-5 (Duong), 5-6 (He)</b> After 10/16 10:30am to noon (Utts)	<b>Tues, 2-4</b> (Utts)	<b>Thurs, 2-4</b> (He)	<b>Fri, 2-3 (Duong)</b> Until 10/16 10:30am to noon (Utts)
---	----------------------------	---------------------------	---

**Textbook:** *Mind On Statistics, 3<sup>rd</sup> edition* by Jessica Utts and Robert Heckard

**Computing:** We will use free software called “R Commander.” You will be instructed in its use.

**Web Site:** <http://www.ics.uci.edu/~jutts/7> includes basic information, daily schedule, practice exams, homework assignments and solutions, links to related sites, etc. Grades posted at eee.

### GRADING, EXAMS, HOMEWORK, CLICKERS and DISCUSSION SECTIONS

(See next page for more details):

**Exams:** There will be 2 midterms, each worth **25%**. Dates are **Mon Oct 19** and **Wed Nov 18**.  
**Final Exam:** **Wed, Dec 9, 1:30-3:30 pm**, comprehensive but more focus on material covered after the second midterm exam. The final exam counts for **35%** of your grade.  
**Homework:** Homework is worth **5.5%** of your grade. See the next page for details.  
**Clickers:** Questions are worth **4.5%** of your grade, with **0.5%** for **each of the 9 weeks** they will be given. Optionally, you may substitute online quizzes. See the next page for details.  
**Discussions:** There will be “hands-on” activities in 7 of the 10 Monday discussion sections. You are required to participate in 5 out of the 7 and will receive 1 point for each, up to a maximum of 5 points, or **5%** of your grade. See next page for details.

**Grade Disputes** will not be considered until the end of the quarter. Write a cover note including the *total number of points* in question and *your rationale* and hand it in by the time of the final exam. If the total points in question will *change* your letter grade (including +/-), your rationale will be considered.

### Classroom Etiquette:

1. I make every effort to start and end on time. Please respect the 50 minutes of class time; do not pack up early - it disturbs those around you. I have another class at 2pm, so will definitely end on time!
2. If you have to leave early, please sit near the doors.
3. If you are bored and want to work on something else, feel free to do so, *but* be very discreet. For example, do *not* read the newspaper.
4. Please turn off cell phone sounds and other electronic equipment sounds.

## GRADES: EXAMS, HOMEWORK, CLICKERS, ONLINE QUIZZES AND DISCUSSIONS

### Grading summary:

Midterm 1:	25.0%
Midterm 2:	25.0%
Final exam:	35.0%
Homework:	5.5%
Clicker/Quizzes:	4.5%
Discussion sections:	5.0%
TOTAL:	100%

### Homework (5.5%):

One or two problems will be assigned with each lecture. There are due as follows:

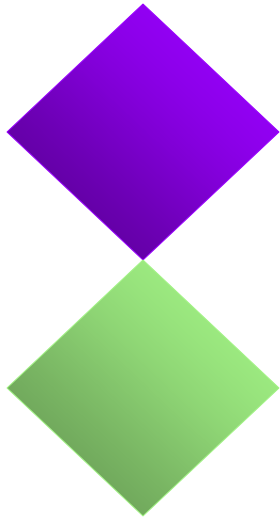
- Until midterm 1, they are due the *Friday* after they are assigned. Thus, problems assigned on Friday, Monday and Wednesday, are due that Friday. Due dates are **October 2, 9 and 16**.
- After midterm 1, they are due the *Monday* after they are assigned. Thus, problems assigned on Mon, Wed, Fri of one week are due the Monday of the next week. Due dates are **Oct 26, Nov 2, 9, 16, 30**. The only exception is the last week of class, when problems are due on Friday, **Dec. 4<sup>th</sup>**.
- Notice that there are 9 assignment due dates. Each assignment will be graded on the basis of 10 points. The lowest of the 9 scores will be dropped. The remaining 8 scores will be averaged, and will be worth **5.5%** of your total grade.
- An “*assignment week*” consists of the 2 or 3 lectures making up one homework assignment.

### Clickers and Online quizzes (4.5%):

- For each “assignment week” (defined above) there will be 5 clicker quiz questions given intermittently during the lectures. For each question you answer you will receive 1 point, and for each *correct* answer you will receive an additional point, for a total possible maximum of 10 points.
- If you prefer, for each “assignment week” you can substitute an online quiz for the clicker points. The quiz will consist of 5 multiple choice questions, randomly selected for you from a larger set. You will receive 2 points for each correct answer and 0 points for each incorrect answer, for a total possible maximum of 10 points. Quizzes will be released the evening of the last lecture of the “assignment week” (Wed or Fri) and must be completed by 1pm on the day of the next lecture (Fri or Mon), the same day the homework is due for that “assignment week.” Quizzes will be timed and you will have a maximum of 20 minutes to complete them. You have only one attempt.
- If you complete both the clicker questions and the online quiz, the higher of your two scores for that assignment week will be counted.
- Your score for each of the 9 assignment weeks counts as 0.5% of your total grade, so the combined total for clicker and/or online quizzes is **4.5%** of your grade.

### Discussion Sections (5%):

- There are a total of 10 discussion dates, but discussions on the day of the first midterm (Monday, October 19<sup>th</sup>) will be cancelled.
- Of the remaining 9 discussion dates, on Monday, October 5 you will be introduced to the R Commander software, and on Monday, Nov. 16 discussions will be used to review for Midterm 2.
- In the remaining 7 discussion sections we will do hands-on projects that give you experience in various areas of statistical practice. You are required to participate in **5 out of the 7**. You will receive **1 point for each** of these, for a total possible maximum of **5 points**, or **5%** of your grade.
- If you participate in more than 5 of the 7 hands-on discussion sections, I will consider your extra participation if your grade is borderline between two possible grades. Also, exam problems may ask about what was done in discussion sections.



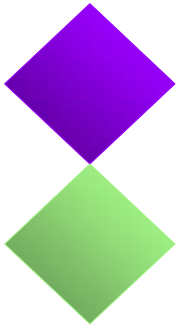
**Chapter 1**

---

# **Statistics Success Stories and Cautionary Tales**

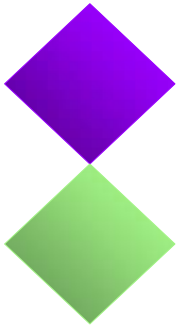
---

# 1.1 What is Statistics?



*Statistics* is a collection of procedures and principles for gathering data and analyzing information in order to help people make decisions when faced with uncertainty.

# 1.2 Seven Statistical Stories With Morals



- **Case Study 1.1: Who Are Those Speedy Drivers?**
- **Case Study 1.2: Safety in the Skies**
- **Case Study 1.3: Did Anyone Ask Whom You've Been Dating?**
- **Case Study 1.4: Who Are Those Angry Women?**
- **Case Study 1.5: Does Prayer Lower Blood Pressure?**
- **Case Study 1.6: Does Aspirin Reduce Heart Attack Rates?**
- **Case Study 1.7: Does the Internet Increase Loneliness and Depression?**

# Case Study 1.1 *Who Are Those Speedy Drivers?*



**Question:** What's the fastest you have ever driven a car? \_\_\_\_\_ mph.

**Data:** 87 male and 102 female students from large statistics class at University.

**Males:** 110 109 90 140 105 150 120 110 110 90 115 95 145 140 110 105  
85 95 100 115 124 95 100 125 140 85 120 115 105 125 102 85 120 110  
120 115 94 125 80 85 140 120 92 130 125 110 90 110 110 95 95 110 105  
80 100 110 130 105 105 120 90 100 105 100 120 100 100 80 100 120 105  
60 125 120 100 115 95 110 101 80 112 120 110 115 125 55 90

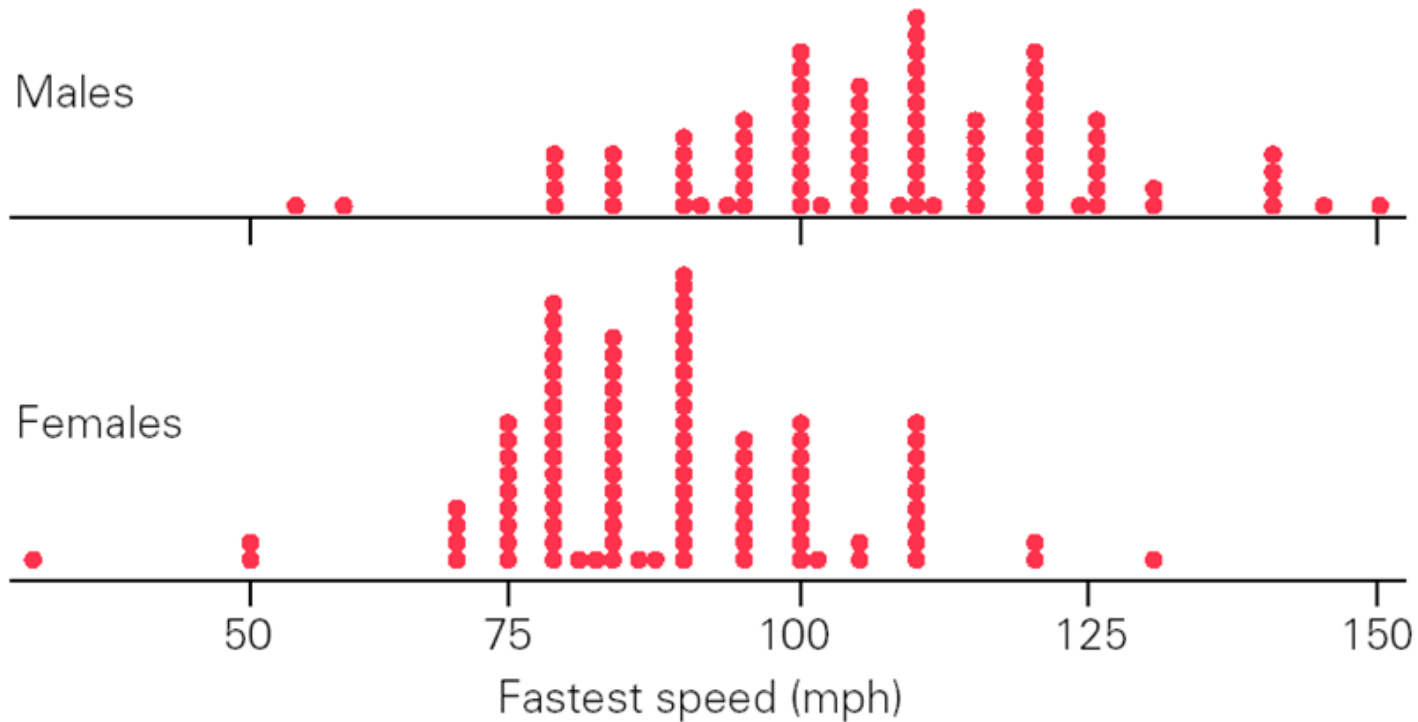
**Females:** 80 75 83 80 100 100 90 75 95 85 90 85 90 90 120 85 100 120 75  
85 80 70 85 110 85 75 105 95 75 70 90 70 82 85 100 90 75 90 110 80 80  
110 110 95 75 130 95 110 110 80 90 105 90 110 75 100 90 110 85 90 80  
80 85 50 80 100 80 80 80 95 100 90 100 95 80 80 50 88 90 90 85 70 90 30  
85 85 87 85 90 85 75 90 102 80 100 95 110 80 95 90 80 90

***Which gender has driven faster? How to summarize data?***

# Case Study 1.1 *Who Are Those Speedy Drivers?*



**Dotplot:** each dot represents the response of an individual student.



# Case Study 1.1 *Who Are Those Speedy Drivers?*



**Five-number summary:** the lowest value, the cutoff points for  $\frac{1}{4}$ ,  $\frac{1}{2}$ , and  $\frac{3}{4}$  of the data, and the highest value.

	Males (87 Students)		Females (102 Students)	
<b>Median</b>	110		89	
<b>Quartiles</b>	95	120	80	95
<b>Extremes</b>	55	150	30	130

Note:  $\frac{3}{4}$  of men have driven 95 mph or more, only  $\frac{1}{4}$  of women have done so.

**Moral:** *Simple summaries of data can tell an interesting story and are easier to digest than long lists.*

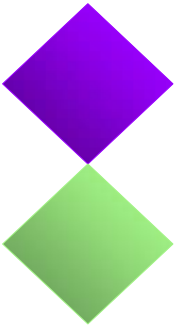
# **READ ON YOUR OWN:**

**Case Study 1.2** *Safety in the Skies?*

**Case Study 1.3** *Did Anyone Ask  
Whom You've Been Dating?*

**Case Study 1.4** *Who Are Those Angry Women?*

**Case Study 1.7** *Does the Internet Increase  
Loneliness and Depression?*



# Case Study 1.5 *Does Prayer*

## *Lower Blood Pressure?*

**“Attending religious services lowers blood pressure more than tuning into religious TV or radio, a new study says”**

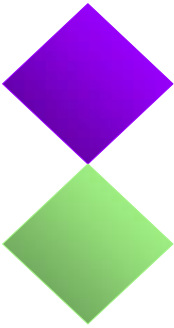
***USA Today* headline read:**

**“Prayer can lower blood pressure.” (Davis, 1998)**

Based on *observational study*, followed 2391 people 6 years.

“People who attended a religious service once a week and prayed or studied the Bible once a day were 40% less likely to have high blood pressure than those who don’t go to church every week and prayed and studied the Bible less.”

Researchers *did observe a relationship*, but it’s a **mistake to conclude** prayer actually *causes* lower blood pressure.



## Case Study 1.5 *Does Prayer*

### *Lower Blood Pressure?*

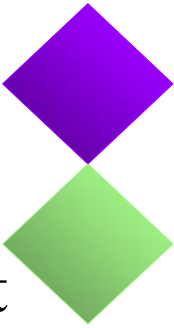
In observational studies, **groups can differ** by important ways that may contribute to the observed relationship.

People who attended church regularly may have ...

- been less likely to smoke or drink alcohol;
- had a better social network;
- been somewhat healthier and able to go to church.

These other factors are possible *confounding variables*.

**Moral:** *Cause-and-effect conclusions cannot generally be made based on an observational study.*



# Case Study 1.6 *Does Aspirin Reduce Heart Attack Rates?*

## Physician's Health Study (1988)

5-year *randomized experiment* ...

- 22,071 male physicians of age 40 - 84;
- *randomly assigned* to one of two *treatment* groups;
- Group 1 = aspirin every other day;  
Group 2 = *placebo*;
- Physicians *blinded* as to which group they were in.



# Case Study 1.6 *Does Aspirin Reduce Heart Attack Rates?*

TABLE 1.1 ■ The Effect of Aspirin on Heart Attacks

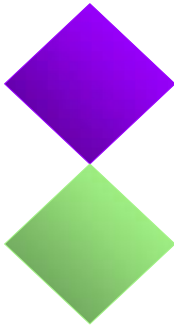
Treatment	Heart Attacks	Doctors in Group	Attacks Per 1000 Doctors
Aspirin	104	11,037	9.42
Placebo	189	11,034	17.13

Aspirin group: 9.42 heart attacks per 1000 participants

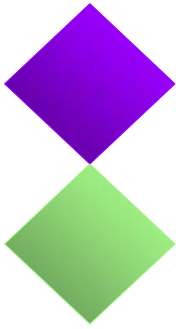
Placebo group: 17.13 heart attacks per 1000 participants

**Randomization** => other important factors (age, diet, etc) should have been similar for both groups. Only important difference should be whether they took aspirin or placebo.

**Moral:** *Unlike with observational studies, cause-and-effect conclusions can generally be made on the basis of randomized experiments.*



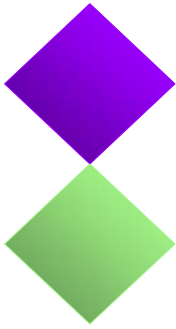
# 1.3 The Common Elements in the Seven Stories



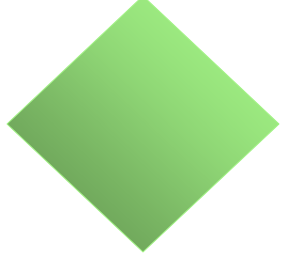
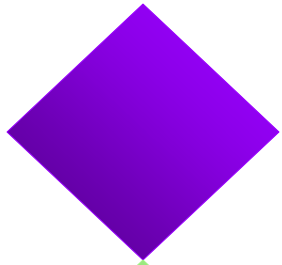
In every story, *data are used to make a judgment about a situation.*

This is what *statistics* is all about.

# The Discovery of Knowledge



- *Asking the right question(s).*
- *Collecting useful data, which includes deciding how much is needed.*
- *Summarizing and analyzing data, with the goal of answering the questions.*
- *Making decisions and generalizations based on the observed data.*
- *Turning the data and subsequent decisions into new knowledge.*



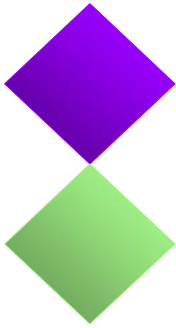
**Chapter 2**

---

# **Turning Data Into Information**

---

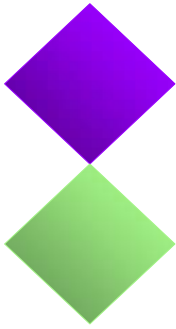
# 2.1 Raw Data



- **Raw data** are numbers and category labels that have been collected but have not yet been processed in any way.
- When measurements are taken from a subset of a population, they represent **sample data**.
- When all individuals in a population are measured, the measurements represent **population data**.
- **Descriptive statistics:** summary numbers for either population or a sample.

## 2.2 Types of Data

- Raw data from **categorical variables** consist of *group or category names* that don't necessarily have a logical ordering. *Examples:* eye color, country of residence.
- Categorical variables for which the categories have a logical ordering are called **ordinal variables**. *Examples:* highest educational degree earned, tee shirt size (S, M, L, XL).
- Raw data from **quantitative variables** consist of *numerical values* taken on each individual. *Examples:* height, number of siblings.



# Asking the Right Questions

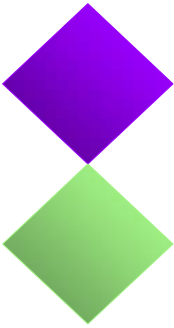
## *One Categorical Variable*

**Question 1a:** How many and what percentage of individuals fall into each category?

*Example:* What percentage of college students favor the legalization of marijuana, and what percentage of college students oppose legalization of marijuana?

**Question 1b:** Are individuals equally divided across categories, or do the percentages across categories follow some other interesting pattern?

*Example:* When individuals are asked to “randomly” choose a number from 1 to 10, are all numbers equally likely to be chosen?



# Asking the Right Questions

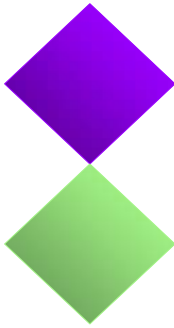
## *Two Categorical Variables*

**Question 2a:** Is there a relationship between the two variables, so that the category into which individuals fall for one variable seems to depend on which category they are in for the other variable?

*Example:* In Case Study 1.6, we asked if the risk of having a heart attack was different for the physicians who took aspirin than for those who took a placebo.

**Question 2b:** Do some combinations of categories stand out because they provide information that is not found by examining the categories separately?

*Example:* The relationship between smoking and lung cancer was detected, in part, because someone noticed that the *combination* of being a nonsmoker and having lung cancer is unusual.



# Explanatory and Response Variables



Many questions are about the **relationship** between *two variables*.

It is useful to identify one variable as the **explanatory variable** and the other variable as the **response variable**.

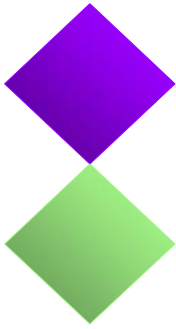
In general, the *value of the explanatory variable* for an individual is thought to **partially explain** the *value of the response variable* for that individual.

Example:

Explanatory variable: aspirin or placebo

Response variable: heart attack or not

## 2.3 Summarizing One or Two Categorical Variables



### Numerical Summaries

- Count how many fall into each category.
- Calculate the percent in each category.
- If two variables, create a “two-way table.”  
Have the categories of the *explanatory* variable define the *rows*.  
*Compute row percentages* and compare.

## Example 2.2 *Lighting the Way to Nearsightedness*

Survey of  $n = 479$  children.

Those who slept with nightlight or in fully lit room before age 2 had higher incidence of nearsightedness (myopia) later in childhood.

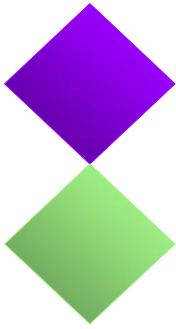
TABLE 2.3 ■ Nighttime Lighting in Infancy and Eyesight

Slept with:	No Myopia	Myopia	High Myopia	Total
Darkness	155 (90%)	15 (9%)	2 (1%)	172
Nightlight	153 (66%)	72 (31%)	7 (3%)	232
Full Light	34 (45%)	36 (48%)	5 (7%)	75
Total	342 (71%)	123 (26%)	14 (3%)	479

**Note:** Study *does not prove* sleeping with light actually *caused* myopia in more children.



# Visual Summaries for Categorical Variables



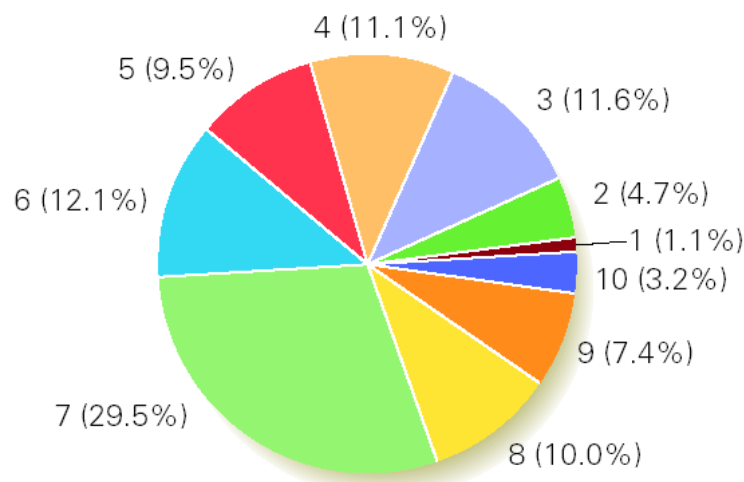
- **Pie Charts:** useful for summarizing a single categorical variable if not too many categories.
- **Bar Graphs:** useful for summarizing one or two categorical variables and particularly useful for making comparisons when there are two categorical variables.

# Example 2.3 *Humans Are Not Good Randomizers*

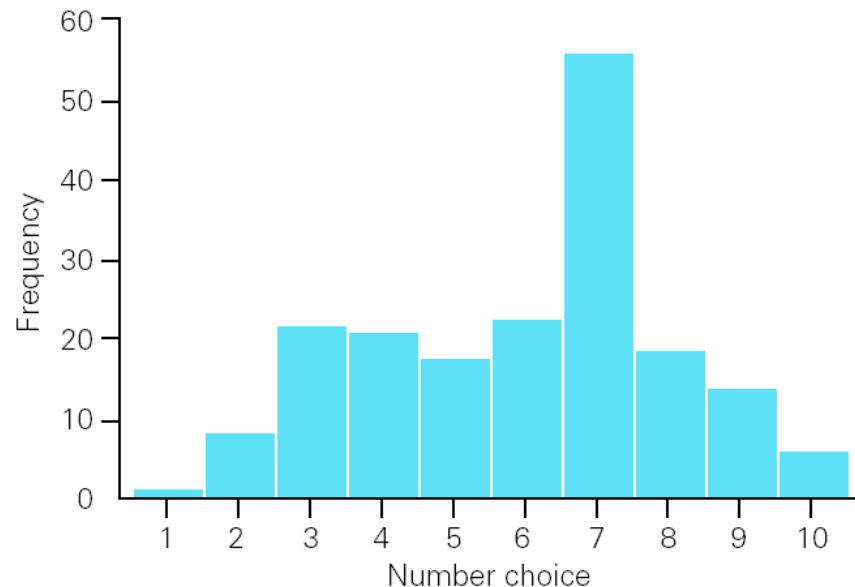


Survey of  $n = 190$  college students.

“Randomly pick a number between 1 and 10.”



**FIGURE 2.1** Pie chart of numbers picked



**FIGURE 2.2** Bar graph of numbers picked

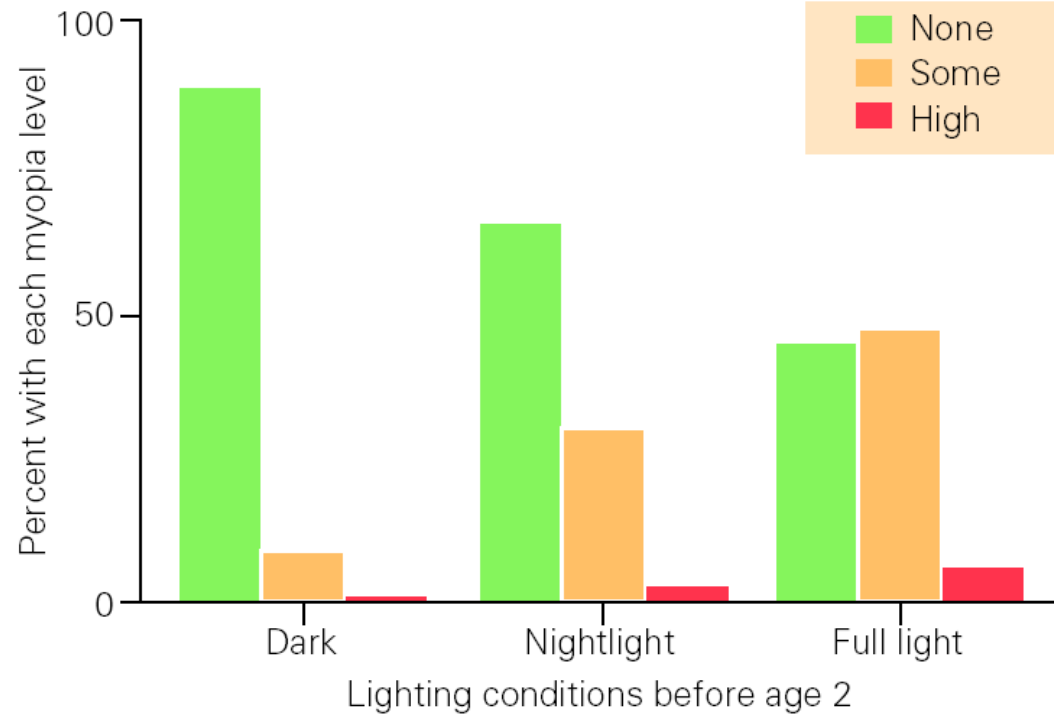
**Results:** Most chose 7, very few chose 1 or 10.

# Example 2.4 *Revisiting Nightlights and Nearsightedness*

Survey of  $n = 479$  children.

**Response:**  
Degree of Myopia

**Explanatory:**  
Amount of Sleep-time Lighting



**FIGURE 2.3** Bar chart for myopia and nighttime lighting in infancy

# Summary of Key Points

- Recognize an observational study (no cause/effect) vs a randomized experiment.

Exercise 1.10: “*Breast-Fed Youth Found to Do Better in School.*” (Newspaper headline)

Exercise 1.23: *Volunteers randomly assigned to nicotine patch or placebo, 46% vs 20% quit.*

- Categorical data/ quantitative data  
*For the two examples above?*
- Explanatory variable/ response variable  
*For the two examples above?*

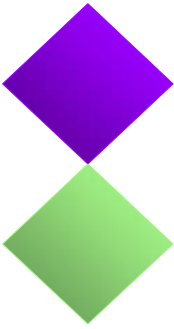


## **Summarizing one categorical variable:**

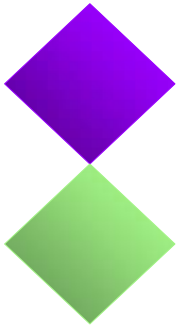
- Pie Charts
- Bar Graphs

## **Summarizing two categorical variables:**

- Two-way table (explanatory variable as rows)
- Row percents (or column percents), use for comparisons
- Bar graphs



# Homework; Sections to Read



- Read Chapter 1
- Read Chapter 2, sections 2.1 to 2.3 (up to page 24)
- Exercises, **due Friday, October 2:**
  - 1.10 (pg. 9-10)
  - 2.20 (pg. 59)