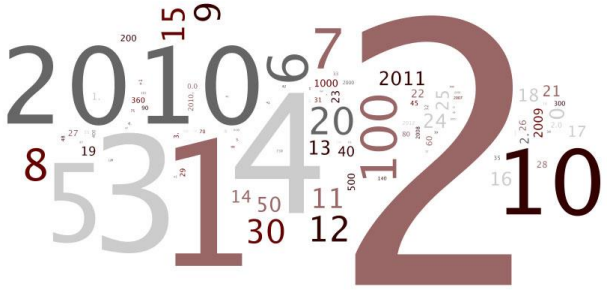
Lecture 1

Introduction

What is ‘Statistics’?

Statistics is the science of collecting, organizing and interpreting data. The goal of statistics is to gain information and understanding from data.

A **statistic** is a number in context.



Statistical inference is making a decision or a conclusion based on the data.

What is ‘data’?

Data are numerical facts with context and we need to understand the context if we are to make sense of the numbers.

How to examine data?

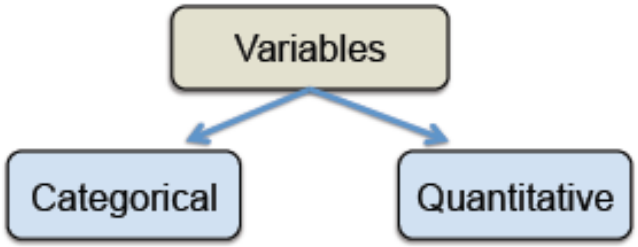
Any set of data contains information about some group of individuals.

Individuals are the objects upon which we collect data. Individuals can be people, animals, and many other things. Sometimes when the objects that we want to study are not people, we often call them **cases**.

The information is organized in variables.

A **variable** is any characteristic of an individual. A variable can take different values for different individuals.

Types of variables:



Categorical variable: a variable that places an individual into one of two or more groups or categories.

Quantitative variable: a variable that takes numerical values for which arithmetic operations such as adding and averaging make sense.

The **distribution** of a variable tells us what values it takes and how often it takes these values.

Ex: Flip a coin once. Let X = # of heads

X: 0 1
P: 1/2 1/2 } → distribution of X

Example: Table below shows part of a data set for students enrolled in an introductory statistics class.

C Q C/Q C C C

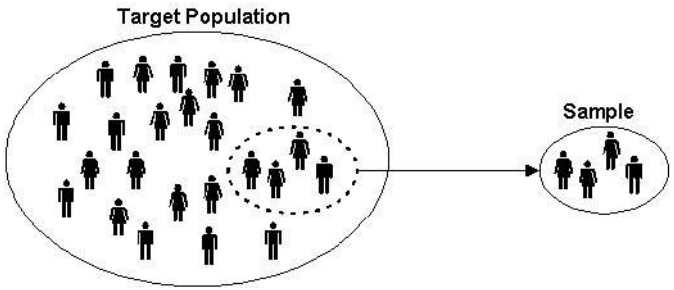
A B C D F

4 3 2 1 0

ID	Total	Grade	Gender	PrevStat	Year
101	899	B	F	Yes	4
102	866	B	M	Yes	3
103	780	C	M	No	3
104	962	A	M	No	1
105	861	B	F	No	4

A **population** is a set of individuals that we are interested in studying.

A **sample** is a subset of the individuals of a population.



Questions to ask when planning a statistical study:

• **Why?** What purpose do the data have? Do we hope to answer some specific questions? Do we want to draw conclusions about individuals other than the ones we actually have data for?

• **Who?** What individuals do the data describe? How many individuals appear in the data?

• **What?** How many variables do the data contain? What are the exact definitions of these variables? What are the units of measurements in which each variable is recorded? Weights for example, might be recorded in pounds, or in kg.

Additional questions you might want to ask yourself: **When? Where? How?**

For the example above:

Who? students

What? 6 variables

Why? keep track of students' grades

How to collect data?

Generally, data can be obtained in four different ways.

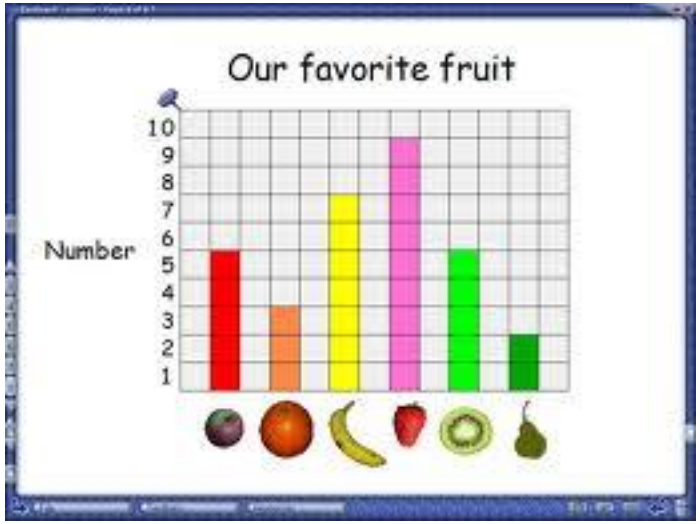
- Published source
- Designed experiment
- Survey
- Observational study

How to explore your data?

Two basic strategies for exploration of data set:

- Begin by examining each variable by itself. Then move on to study the relationships among the variables.

- Each time begin with graphs.



- Then add numerical summaries of specified aspects of the data.

Displaying Distributions with Graphs

Categorical variables

The distribution of a categorical variable lists the categories and gives either the **count** or **percent** of individuals who fall in each category.

Example: How well educated are 30-something young adults?

Education	Count (millions)	Percent
Less than high school	4.6	12.1
High school graduate	11.6	30.5
Some college	7.4	19.5
Associate degree	3.3	8.7
Bachelor's degree	8.6	22.6
Advanced degree	2.5	6.6

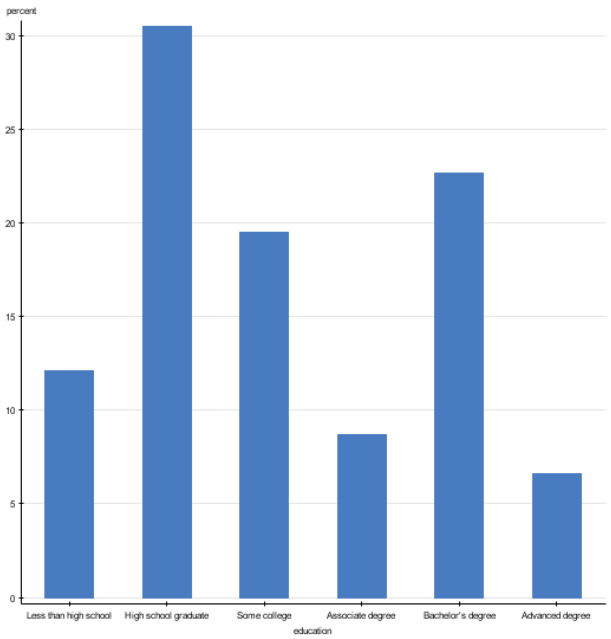
There are two ways to display these data:

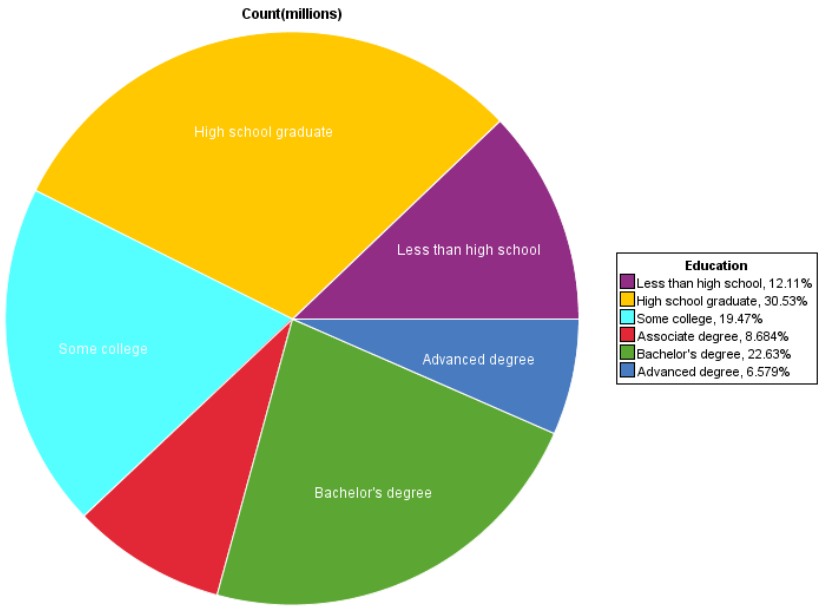
- a bar chart (StatCrunch->Graphics->Bar Plot)

A *bar chart* displays the distribution of a categorical variable, showing the counts for each category next to each other for easy comparison.

- a pie chart (StatCrunch->Graphics->Pie Chart)

A *pie chart* helps us see what part of the whole each group forms.





Two-way Table

For two categorical variables the raw data are summarized in a two-way table that gives counts of observations for each combination of values of the variables.

Example: A survey of 17,096 students in U.S. four-year colleges collected information on drinking behaviour and alcohol-related problems. The researchers defined "frequent binge drinking" as having five and more drinks in a row three or more times in the past two weeks. Here is the two-way table classifying students by gender and whether or not they are frequent binge drinkers:

	Gender:	
Frequent binge drinker:	Men	Women
Yes	1,630	1,684
No	5,550	8,232

← cell

2 x 2 = 4 cells

Gender is a **column variable**.

Binge drinking is a **row variable**.

Combinations of values for two variables are called **cells**.

We can expand the table by adding the totals for each row, for each column, and the total number of all the observations:

	Gender:		
Frequent binge drinker:	Men	Women	Total
Yes	1,630	1,684	3,314
No	5,550	8,232	13,782
Total	7,180	9,916	17,096

Grand Total

Joint Distribution

For each cell, we can compute a proportion by dividing the cell entry by the total sample size. The collection of these proportions is the **joint distribution** of the two categorical variables.

	Gender:	
Frequent binge drinker:	Men	Women
Yes	0.095	0.099
No	0.325	0.482

Because this is a distribution, the sum of the proportions should be 1. For this example the sum is 1.001. The difference is due to roundoff error.

$9.5\% = 0.095 = \frac{1630}{17096}$ = proportion of all students who are men and binge drinker

Marginal Distribution

A **marginal distribution** is the distribution of a single variable in a two-way table.

	Gender:		
Frequent binge drinker:	Men	Women	Total
Yes	1,630	1,684	3,314
No	5,550	8,232	13,782
Total	7,180	9,916	17,096

For example, the marginal distribution of gender is

	Men	Women
Proportion	0.420	0.580

7180/17096 9916/17096

same

or

	Men	Women
Percent	42%	58%

And here is the marginal distribution of frequent binge drinking:

	Yes	No
Percent	19.4%	80.6%

3314/17096 13782/17096

Relationships among the variables are described by calculating appropriate percents from the counts given.

Example: What percent of the women in our sample are frequent binge drinkers?

	Gender:		
Frequent binge drinker:	Men	Women	Total
Yes	1,630	1,684	3,314
No	5,550	8,232	13,782
Total	7,180	9,916	17,096

$$\frac{1684}{9916} = 17\%$$

Here we *condition* on the value of our 'gender' variable: women.

Conditional Distribution

When we condition on the value of one variable and calculate the distribution of the other variable, we obtain a **conditional distribution**.

	Gender:		
Frequent binge drinker:	Men	Women	Total
Yes	1,630	1,684	3,314
No	5,550	8,232	13,782
Total	7,180	9,916	17,096

For example, the conditional distribution of binge drinking of women is

	Yes	No
Percent	17.0%	83.0%

↳ 100% − 17%

Similarly, the conditional distribution of binge drinking for men is

	Yes	No
Percent	22.7%	77.3%

↳ 100% − 22.7%

$\frac{1630}{7180}$

Conclusion:

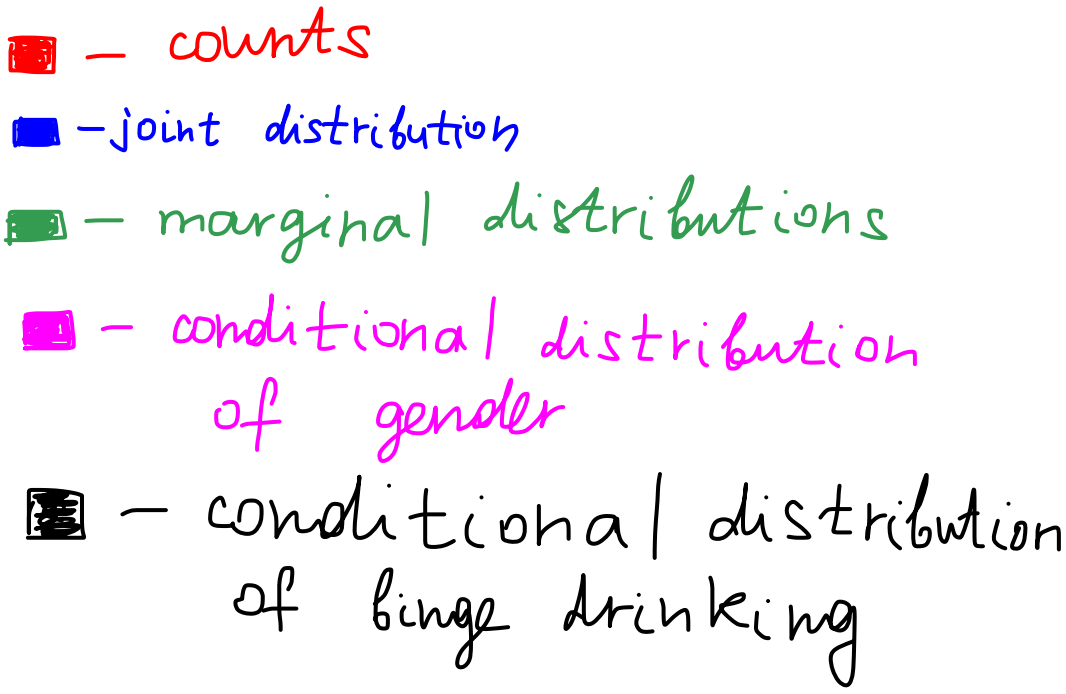
men are more likely to be binge drinkers than women.

Contingency Table: StatCrunch->Stat->Tables->Contingency

Contingency table results:
Rows: Frequent binge drinker
Columns: Gender

Cell format
Count
(Row percent)
(Column percent)
(Total percent)

	Men	Women	Total
Yes	1630 (49.19%) (22.7%) (9.534%)	1684 (50.81%) (16.98%) (9.85%)	3314 (100.00%) (19.38%) (19.38%)
No	5550 (40.27%) (77.3%) (32.46%)	8232 (59.73%) (83.02%) (48.15%)	13782 (100.00%) (80.62%) (80.62%)
Total	7180 (42%) (100.00%) (42%)	9916 (58%) (100.00%) (58%)	17096 (100.00%) (100.00%) (100.00%)



What is independence?

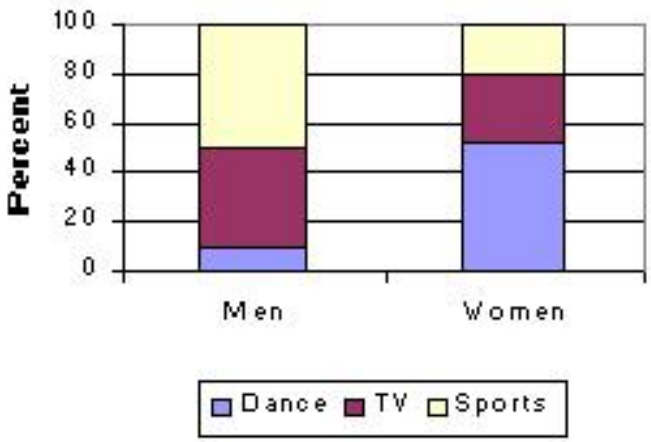
If the conditional distribution of one variable does not depend on the other variable (hence, each conditional distribution resembles the marginal distribution), these two variables are said to be independent.

For our example:

Gender and drinking are not independent, i.e. there might be an association between these variables

Segmented Bar Charts

Example: What activities are preferred by women/men?



Each bar is divided into "segments", such that the length of each segment indicates proportion or percentage of observations in a second variable.

Conclusion:

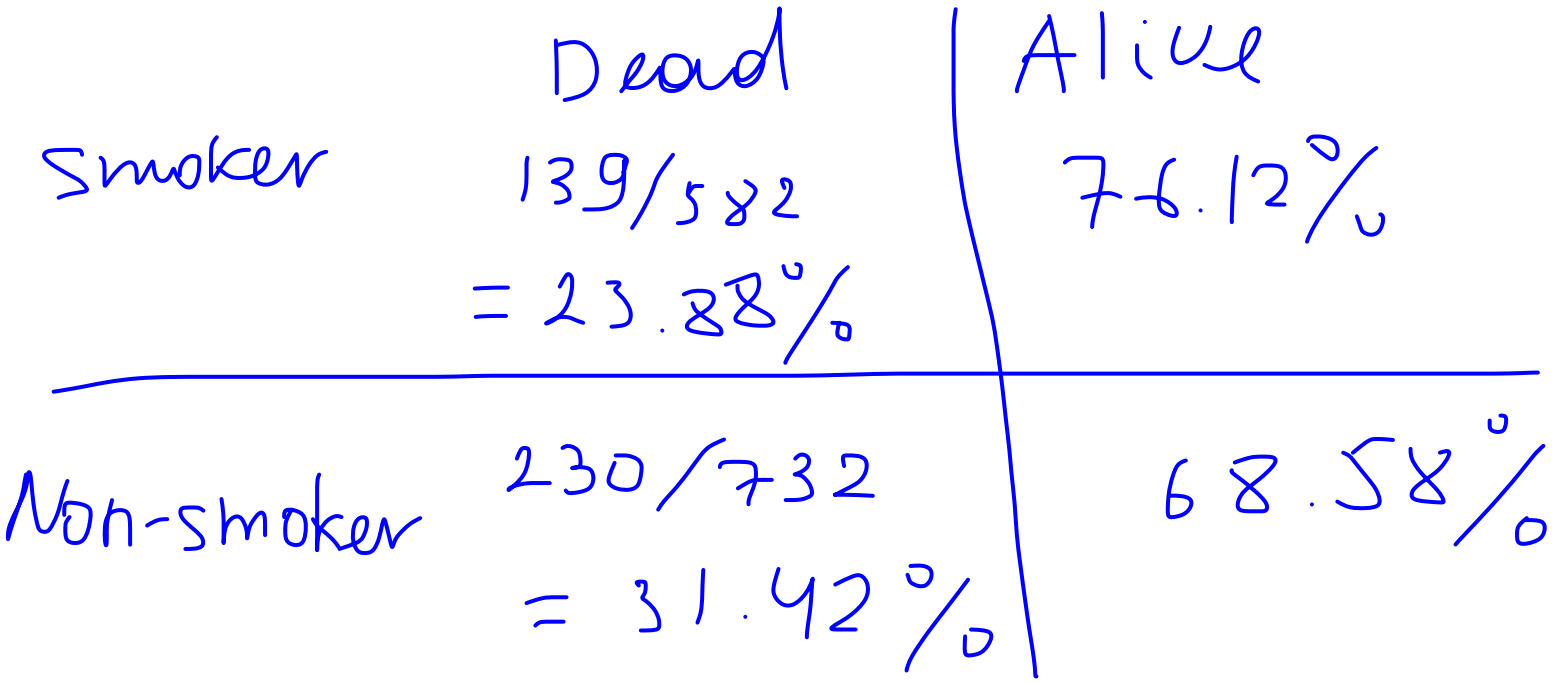
Women will more likely go dancing than men

Simpson's Paradox

Example: Smoking and 20yr survival rate for 1314 English women:

	Dead	Alive	Total
Smoker	139	443	582
Non-Smoker	230	502	732

Let us involve some basic math:



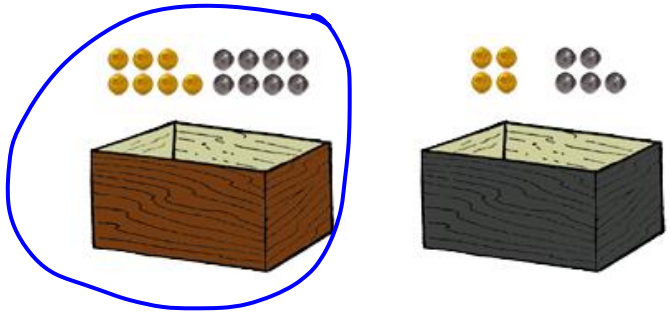
What would you conclude? Is it better to be a smoker?

Unlikely...

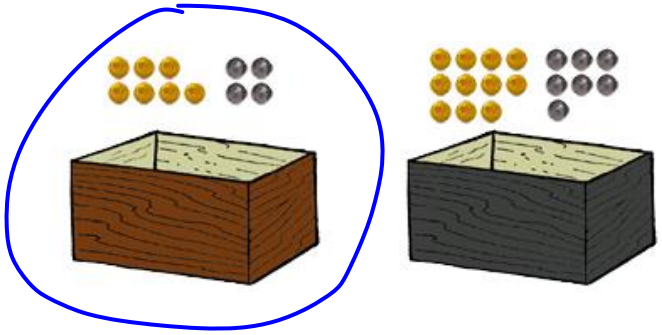
What happened here has a name: **Simpson's Paradox**.

To see how it works let's conduct the following experiment:

We have two tables. There are two boxes on each table: a brown one and a gray one.



Scenario 1



Scenario 2

You are given one chance to extract a gold ball from the boxes. Which box will you choose for each table?

Table 1:

brown box: 7 g and 8 s

gray box: 4 g and 5 s

$P(\text{gold from brown}) = \frac{7}{15} > \frac{4}{9} = P(\text{gold from gray})$

Table 2:

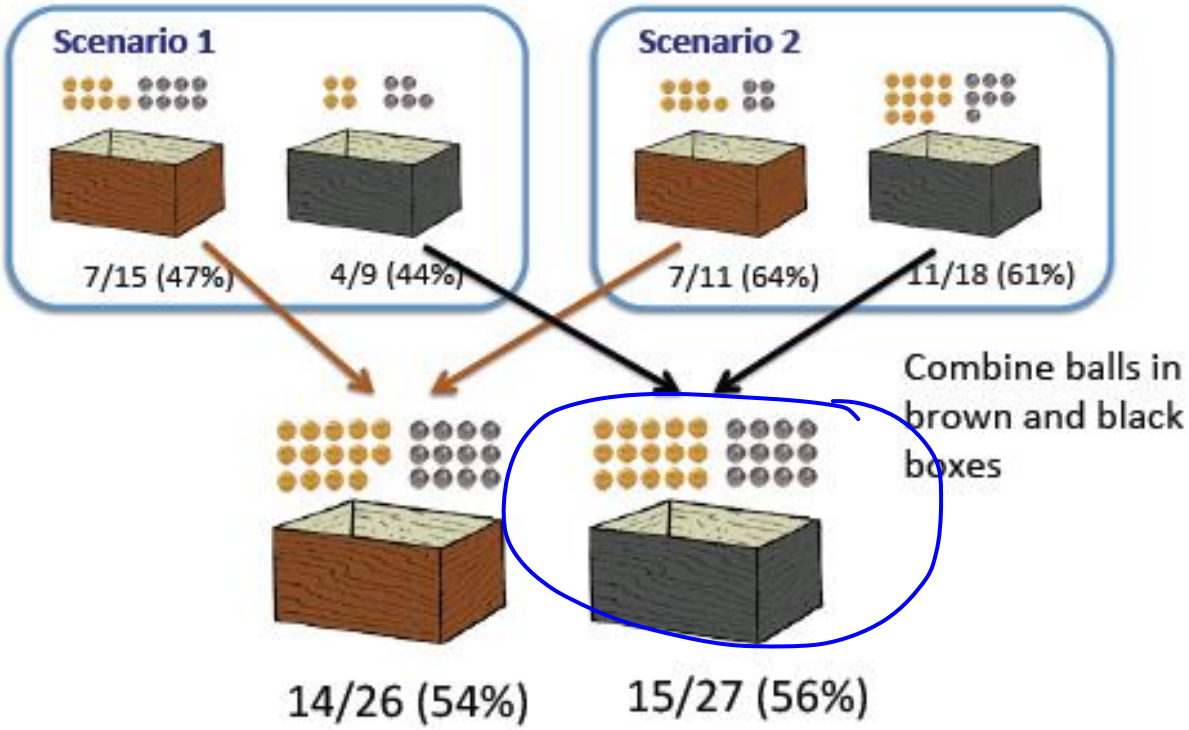
brown box: 7 g and 4 s

gray box: 11 g and 7 s

$P(\text{gold from brown}) = \frac{7}{18} > \frac{4}{11} = P(\text{gold from gray})$

In both cases we would choose the brown box.

Now we put all the balls from the brown boxes into one big brown box, and all the balls from the gray boxes into one big gray box:



Suddenly, we get a reversed conclusion. Now we would choose the gray box!

Definition: An association that holds for all of several groups can reverse direction when the data are combined to form a single group. This reversal is called **Simpson's paradox**.

History: Simpson's paradox is named after Edward Simpson, who first described this paradox in the 1951 paper "The Interpretation of Interaction in Contingency Tables."

Pearson and Yule each observed a similar paradox half a century earlier than Simpson, so Simpson's paradox is sometimes also referred to as the Simpson-Yule effect.

Anytime that data is aggregated, watch out for this paradox to show up.

Now let's look at the real life example.

Example: We want to test two drugs.

We give each drug to a group of people and then count the number of successes (improvements) and failures (no change) for each group.

	Success	Failure	Total
Drug 1	100	100	200
Drug 2	110	80	190

S F

Drug 1 50% 50%

Drug 2 57.9% 42.1%

Conclusion: Drug 2 seems slightly better

What we find out next is that the data were aggregated over gender.

Same example, but split by gender:

Male:

	Success	Failure	Total
Drug 1	60	20	80
Drug 2	100	50	150

	Success	Failure
Drug 1	60/80 = 75%	25%
Drug 2	100/150 = 66.7%	33.3%

Female:

	Success	Failure	Total
Drug 1	40	80	120
Drug 2	10	30	40

	Success	Failure
Drug 1	40/120 = 33.3%	66.7%
Drug 2	10/40 = 25%	75%

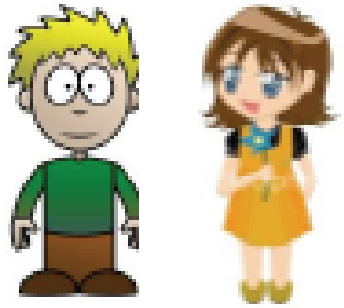
Now drug 1 is better

The conclusion of the study has been reversed.

Lurking Variable

- In the drug example: Gender is a **lurking variable.**

  – A variable that has an important effect and yet is not included amongst the variables under consideration.



- Solutions for lurking variables – eliminate them, or make them part of the study.

• A nice example for a lurking variable is the strong positive association between:

- the number of firefighters at a fire and
- the amount of damage.

• The lurking variable is the size of the blaze, which "causes" both damage and the large number of fire fighters.

- Back to the first example: Better to be a smoker?

- Paradox goes away if the data are broken down by age. Smokers are more likely to die in all but one age group.

- **Take home message**: Be careful when you average across different levels of variables!