# Review of last lecture

# Review of last lecture

We have been looking at how data gets accumulated before we ultimately analyze it.

# Review of last lecture

We have been looking at how data gets accumulated before we ultimately analyze it. There are roughly two categories of ways in which data is collected:

# Review of last lecture

We have been looking at how data gets accumulated before we ultimately analyze it. There are roughly two categories of ways in which data is collected: *surveys*, which are usually *sampled* from a larger population;

# Review of last lecture

We have been looking at how data gets accumulated before we ultimately analyze it. There are roughly two categories of ways in which data is collected: *surveys*, which are usually *sampled* from a larger population; and *experimental studies*, where individuals are treated (or not treated, which constitutes a kind of treatment) in various ways and the outcomes of these treatments are measured.

# Review of last lecture

We have been looking at how data gets accumulated before we ultimately analyze it. There are roughly two categories of ways in which data is collected: *surveys*, which are usually *sampled* from a larger population; and *experimental studies*, where individuals are treated (or not treated, which constitutes a kind of treatment) in various ways and the outcomes of these treatments are measured. In all ways of gathering data, one has to be aware of sources of *bias* which can lead to data which does not represent the true situation.

# Review of last lecture

We have been looking at how data gets accumulated before we ultimately analyze it. There are roughly two categories of ways in which data is collected: *surveys*, which are usually *sampled* from a larger population; and *experimental studies*, where individuals are treated (or not treated, which constitutes a kind of treatment) in various ways and the outcomes of these treatments are measured. In all ways of gathering data, one has to be aware of sources of *bias* which can lead to data which does not represent the true situation.

**Example 1.** *Look at the study at* `http://abcnews.go.com` *and say first what kind of study it is, the sample size and other basic data, and what its strengths and weaknesses are.*

# Stratified random sample

# Stratified random sample

In gathering data by survey, the standard method to avoid bias is to take a *simple random sample*,

# Stratified random sample

In gathering data by survey, the standard method to avoid bias is to take a *simple random sample*, which means to assign a number to each individual being studied and then use a random number generator to pick the numbers of those who will be surveyed.

# Stratified random sample

In gathering data by survey, the standard method to avoid bias is to take a *simple random sample*, which means to assign a number to each individual being studied and then use a random number generator to pick the numbers of those who will be surveyed.

An elaboration on this idea, which can help guarantee that various sub-populations are appropriately represented, is the *stratified random sample*.

**Definition 2.** *A stratified random sample is obtained by first dividing the population into groups of individuals, called* strata*. Then a simple random survey is taken of each stratum. The results of these SRS's get combined into a full sample.*

**Definition 2.** *A stratified random sample is obtained by first dividing the population into groups of individuals, called* strata. *Then a simple random survey is taken of each stratum. The results of these SRS's get combined into a full sample.*

**Example 3.** *Use Table B from the book to generated a stratified sample of twenty out of a population with 30 single people, 50 married people, and 20 divorced people. (Assumed that numbers are automatically assigned).*

**Definition 2.** *A stratified random sample is obtained by first dividing the population into groups of individuals, called* strata. *Then a simple random survey is taken of each stratum. The results of these SRS's get combined into a full sample.*

**Example 3.** *Use Table B from the book to generated a stratified sample of twenty out of a population with 30 single people, 50 married people, and 20 divorced people. (Assumed that numbers are automatically assigned). Compare this with what you would get if a simple random survey were applied to the entire population (which we*

*will carry out multiple times with random numbers from random.org, instead of the table).*

*will carry out multiple times with random numbers from random.org, instead of the table).*

In order for the results to be representative, one must be careful that the strata do not get too small.

*will carry out multiple times with random numbers from random.org, instead of the table).*

In order for the results to be representative, one must be careful that the strata do not get too small.

On the other hand, if one starts with an enormous population, one can apply multiple stratified samples.

*will carry out multiple times with random numbers from random.org, instead of the table).*

In order for the results to be representative, one must be careful that the strata do not get too small.

On the other hand, if one starts with an enormous population, one can apply multiple stratified samples. For example, in surveying the US, one might first divide into urban and rural zip-codes from which one samples randomly,

*will carry out multiple times with random numbers from random.org, instead of the table).*

In order for the results to be representative, one must be careful that the strata do not get too small.

On the other hand, if one starts with an enormous population, one can apply multiple stratified samples. For example, in surveying the US, one might first divide into urban and rural zip-codes from which one samples randomly, and then within each zip-code one could choose strata according to race or gender or other socio-economic factors.

# Cautions about sampling

# Cautions about sampling

Some common problems with surveys are that, short of stalking people or offering large amounts of money for participation, you cannot guarantee that the people you may have chosen perfectly random to participate will in fact participate. This is known as *nonresponse*.

# Cautions about sampling

Some common problems with surveys are that, short of stalking people or offering large amounts of money for participation, you cannot guarantee that the people you may have chosen perfectly random to participate will in fact participate. This is known as *nonresponse*. This problem, or the method in which one stratifies a survey, or any other number of factors, can lead to counting individuals from one group less than others, known as *undercoverage*.

# Cautions about sampling

Some common problems with surveys are that, short of stalking people or offering large amounts of money for participation, you cannot guarantee that the people you may have chosen perfectly random to participate will in fact participate. This is known as *nonresponse*. This problem, or the method in which one stratifies a survey, or any other number of factors, can lead to counting individuals from one group less than others, known as *undercoverage*.

Basic question design can change results wildly (“pray while smoking” vs. “smoke while praying”).

Basic question design can change results wildly ("pray while smoking" vs. "smoke while praying").

But even if one is meticulous about representing all groups appropriately, phrasing things carefully, getting a high response rate, you still don't know how good your data is until you know something about extrapolating from a small data set to a large one.

Basic question design can change results wildly ("pray while smoking" vs. "smoke while praying").

But even if one is meticulous about representing all groups appropriately, phrasing things carefully, getting a high response rate, you still don't know how good your data is until you know something about extrapolating from a small data set to a large one. We will come back to this question in some detail after we study the mathematics of probability.

# Experimental design

# Experimental design

Experiments require a lot more energy than surveys,

# Experimental design

Experiments require a lot more energy than surveys, but that energy can be worth it in that one can *isolate variables*.

# Experimental design

Experiments require a lot more energy than surveys, but that energy can be worth it in that one can *isolate variables*. For example, in seeing whether wine consumption effects heart disease, it could be that looking at data by country we are missing some other driving factors which mirror wine consumption (lifestyle, diet, exercise...).

# Experimental design

Experiments require a lot more energy than surveys, but that energy can be worth it in that one can *isolate variables*. For example, in seeing whether wine consumption effects heart disease, it could be that looking at data by country we are missing some other driving factors which mirror wine consumption (lifestyle, diet, exercise...). In an experimental study, we could gather groups of people who behave identically in all of these other ways,

# Experimental design

Experiments require a lot more energy than surveys, but that energy can be worth it in that one can *isolate variables*. For example, in seeing whether wine consumption effects heart disease, it could be that looking at data by country we are missing some other driving factors which mirror wine consumption (lifestyle, diet, exercise...). In an experimental study, we could gather groups of people who behave identically in all of these other ways, split them into groups who do and do not drink a glass of wine three times a week,

# Experimental design

Experiments require a lot more energy than surveys, but that energy can be worth it in that one can *isolate variables*. For example, in seeing whether wine consumption effects heart disease, it could be that looking at data by country we are missing some other driving factors which mirror wine consumption (lifestyle, diet, exercise...). In an experimental study, we could gather groups of people who behave identically in all of these other ways, split them into groups who do and do not drink a glass of wine three times a week, and then

observe their cardiac health over a long period of time.

observe their cardiac health over a long period of time. Such a study would be much more accurate, and much more expensive than a survey!

observe their cardiac health over a long period of time. Such a study would be much more accurate, and much more expensive than a survey! Keep this in mind any time you see data.

Just measuring something can have an effect, so to get an appropriate analysis one must have at least two groups on which one experiments.

Just measuring something can have an effect, so to get an appropriate analysis one must have at least two groups on which one experiments. One group may be a *control group*, which is treated "normally".

Just measuring something can have an effect, so to get an appropriate analysis one must have at least two groups on which one experiments. One group may be a *control group*, which is treated "normally". In medical (and some other similar areas), the way one gives no treatment without a patient knowing is by giving a *placebo* or "dummy treatment."

Just measuring something can have an effect, so to get an appropriate analysis one must have at least two groups on which one experiments. One group may be a *control group*, which is treated "normally". In medical (and some other similar areas), the way one gives no treatment without a patient knowing is by giving a *placebo* or "dummy treatment." To minimize the effects of hidden variables, one *randomizes* (that is, picks at random members of) the groups getting treated.

Just measuring something can have an effect, so to get an appropriate analysis one must have at least two groups on which one experiments. One group may be a *control group*, which is treated "normally". In medical (and some other similar areas), the way one gives no treatment without a patient knowing is by giving a *placebo* or "dummy treatment." To minimize the effects of hidden variables, one *randomizes* (that is, picks at random members of) the groups getting treated. Finally, when possible one should keep everyone involved unaware of which individuals are being treated by which

treatment; such a protocol is called *double-blind*.

treatment; such a protocol is called *double-blind*.

**Example 4.** *Is the data gathering at* `http://www.shelbyst` *an experiment or a survey?*

treatment; such a protocol is called *double-blind*.

**Example 4.** *Is the data gathering at* `http://www.shelbyst` *an experiment or a survey?*

**Example 5.** *Discuss the experiment described at the following websites in relation to the terminology developed above.*

`http://www.iht.com/articles/2005/04/13/healthsci`

`http://www.cancer.gov/newscenter/pressreleases/H`

**Example 6.** *Design an experiment to measure the effects on safety of having car headlights on during normal daytime driving.*

**Example 6.** *Design an experiment to measure the effects on safety of having car headlights on during normal daytime driving. Discuss the problem(s) with doing the same for measuring the effect of taking a driver's education course.*

**Example 6.** *Design an experiment to measure the effects on safety of having car headlights on during normal daytime driving. Discuss the problem(s) with doing the same for measuring the effect of taking a driver's education course.*

# Big question: when are studies or experiments "statistically significant"?

# Big question: when are studies or experiments "statistically significant"?

What if our drug study just happened to randomly select the people who were going to overcome their illness anyways?

# Big question: when are studies or experiments "statistically significant"?

What if our drug study just happened to randomly select the people who were going to overcome their illness anyways? (what are the chances of such a selection?)

# Big question: when are studies or experiments "statistically significant"?

What if our drug study just happened to randomly select the people who were going to overcome their illness anyways? (what are the chances of such a selection?) Is it enough to have a 50% better result for one group over another? 10%? 2%?

# Big question: when are studies or experiments "statistically significant"?

What if our drug study just happened to randomly select the people who were going to overcome their illness anyways? (what are the chances of such a selection?) Is it enough to have a 50% better result for one group over another? 10%? 2%?

As with our discussion of sample size for surveys, we will develop the mathematics of probability first and then come back to the question of statistical significance.

# Basic probability

# Basic probability

The theory of probability let's us answer questions such as "what are the chances that at random these ten people will get well (in our study) but those ten people will not?"

# Basic probability

The theory of probability let's us answer questions such as "what are the chances that at random these ten people will get well (in our study) but those ten people will not?" In general probability theory helps in predict what will happen for a random process *over many trials*.

# Basic probability

The theory of probability let's us answer questions such as "what are the chances that at random these ten people will get well (in our study) but those ten people will not?" In general probability theory helps in predict what will happen for a random process *over many trials*.

"Randomness" is a more difficult concept than you would think, but there are some clear examples.

# Basic probability

The theory of probability let's us answer questions such as "what are the chances that at random these ten people will get well (in our study) but those ten people will not?" In general probability theory helps in predict what will happen for a random process *over many trials*.

"Randomness" is a more difficult concept than you would think, but there are some clear examples. Observationally, the outcome of a coin toss is random since we can't predict heads or tails.

# Basic probability

The theory of probability let's us answer questions such as "what are the chances that at random these ten people will get well (in our study) but those ten people will not?" In general probability theory helps in predict what will happen for a random process *over many trials*.

"Randomness" is a more difficult concept than you would think, but there are some clear examples. Observationally, the outcome of a coin toss is random since we can't predict heads or tails.

Simultaneously, however, there is a long-term *pattern* to the outcome.

Simultaneously, however, there is a long-term *pattern* to the outcome. If enough tosses are made, approximately half will be heads, and half tails.

Simultaneously, however, there is a long-term *pattern* to the outcome. If enough tosses are made, approximately half will be heads, and half tails.

This does *not* mean that every "heads" will be followed by a "tail." Just that if you toss long enough, you expect to see half heads and half tails.

Famous question: If you toss a coin 5 times in a row and get heads every time, what is the probability of getting tails the next time?

Famous question: If you toss a coin 5 times in a row and get heads every time, what is the probability of getting tails the next time?

Same kind of question: If you have 4 girls already, are you more likely to have a boy as your next child?

Blaise Pascal, 1623-1662; Probability: 1654

**Notation 7. [Notation by example]** • *We let $X$ be the result of a coin toss. It can be $H$ or $T$. $\{H, T\}$ is called the* sample space*.*

**Notation 7. [Notation by example]**

- *We let* $X$ *be the result of a coin toss. It can be* $H$ *or* $T$*.* $\{H, T\}$ *is called the* sample space*.*

- *We write* $P(X = H) = .5$ *to mean the* probability *of the result of the coin toss being heads is* $.5$*.*

**Notation 7. [Notation by example]**

- *We let $X$ be the result of a coin toss. It can be $H$ or $T$. $\{H, T\}$ is called the* sample space*.*

- *We write $P(X = H) = .5$ to mean the* probability *of the result of the coin toss being heads is $.5$.*

- *$P(X = H \text{ or } X = T) = 1$ since those are the only possibilities.*

**Example 8.** *Let $X$ be the result of two successive coin tosses. What is the sample space?*

**Example 8.** *Let $X$ be the result of two successive coin tosses. What is the sample space? What are the probabilities of each member of the sample space?*

Notice that if we add the probability for each event in our sample space we get 1.

**Example 8.** *Let $X$ be the result of two successive coin tosses. What is the sample space? What are the probabilities of each member of the sample space?*

Notice that if we add the probability for each event in our sample space we get 1.

**Example 9.** *Basically same example, different sample space: answer the same questions when $X$ is the number of heads in two successive coin tosses.*

# Probability formalities

# Probability formalities

We have yet to define probability (and will not do so in the way that mathematicians ultimately do).

# Probability formalities

We have yet to define probability (and will not do so in the way that mathematicians ultimately do). But we can list properties which it must obey.

# Probability formalities

We have yet to define probability (and will not do so in the way that mathematicians ultimately do). But we can list properties which it must obey.

- For any event $A$, $0 \leq P(A) \leq 1$. Probability $1$ means $A$ is certain to happen, probability $0$ means $A$ is certain not to happen.

# Probability formalities

We have yet to define probability (and will not do so in the way that mathematicians ultimately do). But we can list properties which it must obey.

- For any event $A$, $0 \leq P(A) \leq 1$. Probability $1$ means $A$ is certain to happen, probability $0$ means $A$ is certain not to happen.

- If there are $D$ outcomes in the sample space which are a priori equally likely, then the chance of achieving one

of $N$ of these outcomes is $\frac{N}{D}$.

of $N$ of these outcomes is $\frac{N}{D}$.

- $P(A \text{ doesn't happen}) = 1 - P(A)$.

of $N$ of these outcomes is $\frac{N}{D}$.

- $P(A \text{ doesn't happen}) = 1 - P(A)$.
- If event $A$ and event $B$ have no outcomes in common,

$$P(A \text{ or } B) = P(A) + P(B).$$

of $N$ of these outcomes is $\frac{N}{D}$.

- $P(A \text{ doesn't happen}) = 1 - P(A)$.
- If event $A$ and event $B$ have no outcomes in common,

$$P(A \text{ or } B) = P(A) + P(B).$$

- If the outcome of event $X$ is unrelated to the outcome of event $Y$ (they are *independent*) then

$$P(X = A \text{ and } Y = B) = P(X = A) \times P(Y = B)$$

**Example 10.** *You roll two dice; each die has an equal probability* $(1/6)$ *of showing any number out of* $\{1, 2, 3, 4, 5, 6\}$*. What is the probability of getting a 12? An 11? A 7?*

**Example 10.** *You roll two dice; each die has an equal probability* $(1/6)$ *of showing any number out of* $\{1, 2, 3, 4, 5, 6\}$*. What is the probability of getting a 12? An 11? A 7?*

**Example 11.** *Toss two coins (a nickel and a penny). If we are told that at least one came up heads, what is the probability of the other coming up heads?*

**Example 12.** *Monty Hall easy version. There are 3 doors; you don't know what is behind any of them. You are told that there is a car behind one door, and a goat behind the other two. You get whatever is behind the door you choose as a prize. So you pick a door at random. What is your chance of getting a car?*

**Example 13. [Monty Hall problem]** *Monty Hall wants to confuse you a little bit. You get to pick a door.* Then *(whichever door you pick, whether there is a goat behind it or a car) Monty opens a different door that has a goat behind it. (He can always do this, even if you chose a goat door, because there are 2 goats).*

*Now Monty gives you the opportunity to switch doors to the other unopened door. Should you switch?*