Chapter 3: Producing Data

Brief overview of what we know so far:

Chapter 3: Producing Data

Brief overview of what we know so far:

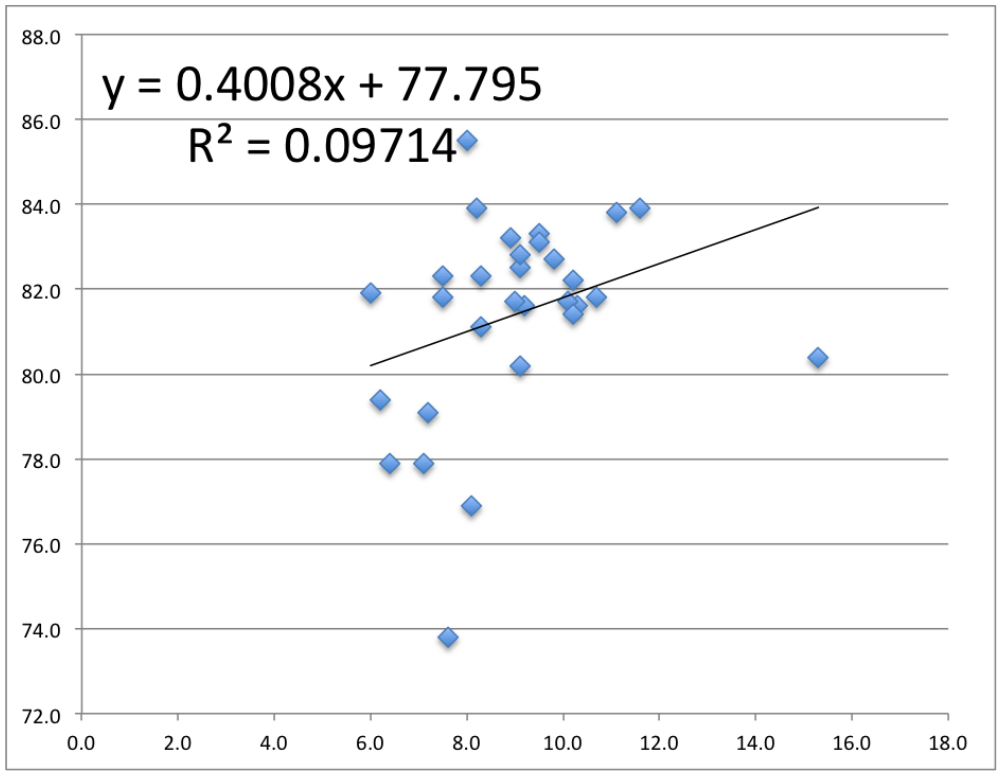
- If we *have* a data set, we can compute summary statistics, study associations between variables, and compute a linear regression in the case of strongly correlated quantitative variables.

	A	B	C	D
1	**Human Development Report 2009**			
2	http://hdr.undp.org/en/reports/global/hdr2009/			
3	**H: Human development index 2007 and its components**			
4	**Life Expectancy: 2007 values**			
5				
6	Country	**Life expectancy at birth (years)**		
7	Afghanistan	43.6		
8	Albania	76.5	MEDIAN	71.6
9	Algeria	72.2	MEAN	68.0
10	Angola	46.5	STD DEV	10.40059
11	Argentina	75.2		
12	Armenia	73.6		
13	Australia	81.4		
14	Austria	79.9		
15	Azerbaijan	70.0		
16	Bahamas	73.2		
17	Bahrain	75.6		
18	Bangladesh	65.7		
19	Barbados	77.0		
20	Belarus	69.0		
21	Belgium	79.5		
22	Belize	76.0		
23	Benin	61.0		
24	Bhutan	65.7		
25	Bolivia	65.4		

Chapter 3: Producing Data

Brief overview of what we know so far:

- If we *have* a data set, we can compute summary statistics, study associations between variables, and compute a linear regression in the case of strongly correlated quantitative variables.



Chapter 3: Producing Data

Brief overview of what we know so far:

- We can *model* an experimental process using the language of probability. Sample space = outcomes i.e. subjects in a study, patients in a hospital, car parts.

Chapter 3: Producing Data

Brief overview of what we know so far:

- We can *model* an experimental process using the language of probability. Sample space = outcomes i.e. subjects in a study, patients in a hospital, car parts.
- Random variable = quantitative variable, some numerical measurement of each outcome i.e. blood O2 level, mass, temperature, SAT score.

Chapter 3: Producing Data

Brief overview of what we know so far:

- We can *model* an experimental process using the language of probability. Sample space = outcomes i.e. subjects in a study, patients in a hospital, car parts.
- Random variable = quantitative variable, some numerical measurement of each outcome i.e. blood O2 level, mass, temperature, SAT score.
- We can compute 'exact' statistics for random variables, such as mean and standard deviation:

$$E[X] = \mu = \sum_{i=1}^{N} p_i x_i, \qquad \sigma^2 = \sum_{i=1}^{N} p_i \cdot (x_i - \mu)^2$$

3.1: Sources of Data

Where can we get data?

3.1: Sources of Data

Where can we get data?

- Anecdotal data? Not very reliable.

3.1: Sources of Data

Where can we get data?

- Anecdotal data? Not very reliable.
- Available data, e.g. data.gov

3.1: Sources of Data

Where can we get data?

- Anecdotal data? Not very reliable.
- Available data, e.g. data.gov
- Sample surveys - select a subset of a larger population, ask questions, *infer* the population distribution

3.1: Sources of Data

Where can we get data?

- Anecdotal data? Not very reliable.
- Available data, e.g. data.gov
- Sample surveys - select a subset of a larger population, ask questions, *infer* the population distribution
- Census - when the sample is the entire population

3.1: Sources of Data

Where can we get data?

- Anecdotal data? Not very reliable.
- Available data, e.g. data.gov
- Sample surveys - select a subset of a larger population, ask questions, *infer* the population distribution
- Census - when the sample is the entire population
- Observational study: try to observe subjects without attempting to influence response

3.1: Sources of Data

Where can we get data?

- Anecdotal data? Not very reliable.
- Available data, e.g. data.gov
- Sample surveys - select a subset of a larger population, ask questions, *infer* the population distribution
- Census - when the sample is the entire population
- Observational study: try to observe subjects without attempting to influence response
- Experiment: deliberately impose a *treatment* on individuals

Example: Women in STEM Degrees

- Want to study the 'STEM pipeline': are women less likely to earn PhDs in a STEM field than men?

[1]NSF National Survey of College Graduates and Survey of Doctoral Recipients

[2]Miller, Wai 2015 'The bachelors to Ph.D. STEM pipeline no longer leaks more women than men: a 30-year analysis', http://journal.frontiersin.org/article/10.3389/fpsyg.2015.00037/full

Example: Women in STEM Degrees

- Want to study the 'STEM pipeline': are women less likely to earn PhDs in a STEM field than men?
- Observational or experimental? Data source?

[1]NSF National Survey of College Graduates and Survey of Doctoral Recipients

[2]Miller, Wai 2015 'The bachelors to Ph.D. STEM pipeline no longer leaks more women than men: a 30-year analysis', http://journal.frontiersin.org/article/10.3389/fpsyg.2015.00037/full

Example: Women in STEM Degrees

- Want to study the 'STEM pipeline': are women less likely to earn PhDs in a STEM field than men?
- Observational or experimental? Data source?
- *Used existing data from NSCG and SDR*[1]

[1]NSF National Survey of College Graduates and Survey of Doctoral Recipients

[2]Miller, Wai 2015 'The bachelors to Ph.D. STEM pipeline no longer leaks more women than men: a 30-year analysis', http://journal.frontiersin.org/article/10.3389/fpsyg.2015.00037/full

Example: Women in STEM Degrees

- Want to study the 'STEM pipeline': are women less likely to earn PhDs in a STEM field than men?
- Observational or experimental? Data source?
- *Used existing data from NSCG and SDR*[1]
- Sample or census?

[1]NSF National Survey of College Graduates and Survey of Doctoral Recipients

[2]Miller, Wai 2015 'The bachelors to Ph.D. STEM pipeline no longer leaks more women than men: a 30-year analysis', http://journal.frontiersin.org/article/10.3389/fpsyg.2015.00037/full

Example: Women in STEM Degrees

- Want to study the 'STEM pipeline': are women less likely to earn PhDs in a STEM field than men?
- Observational or experimental? Data source?
- *Used existing data from NSCG and SDR*[1]
- Sample or census? *Very large sample (10's of thousands)*

[1]NSF National Survey of College Graduates and Survey of Doctoral Recipients

[2]Miller, Wai 2015 'The bachelors to Ph.D. STEM pipeline no longer leaks more women than men: a 30-year analysis', http://journal.frontiersin.org/article/10.3389/fpsyg.2015.00037/full

Example: Women in STEM Degrees

- Want to study the 'STEM pipeline': are women less likely to earn PhDs in a STEM field than men?
- Observational or experimental? Data source?
- *Used existing data from NSCG and SDR*[1]
- Sample or census? *Very large sample (10's of thousands)*
- Conclusion: while women are still less likely to earn STEM degrees *overall*, those who do earn STEM bachelors are as likely to earn STEM PhD's as men.[2]

[1] NSF National Survey of College Graduates and Survey of Doctoral Recipients

[2] Miller, Wai 2015 'The bachelors to Ph.D. STEM pipeline no longer leaks more women than men: a 30-year analysis', http://journal.frontiersin.org/article/10.3389/fpsyg.2015.00037/full

Example: Class Size

- We want to study the effect of class size on standardized test scores. Observational or experiment?

Example: Class Size

- We want to study the effect of class size on standardized test scores. Observational or experiment?
- Confounding variables?

Example: Class Size

- We want to study the effect of class size on standardized test scores. Observational or experiment?
- Confounding variables?
- Tennessee STAR program: separate kindergarten students, *randomly* place them in smaller or larger classrooms for 4 years, compare standardized test results.

Example: Class Size

- We want to study the effect of class size on standardized test scores. Observational or experiment?
- Confounding variables?
- Tennessee STAR program: separate kindergarten students, *randomly* place them in smaller or larger classrooms for 4 years, compare standardized test results.
- Experimental units/subjects? Treatment? Outcomes/response?

Example: Class Size

- We want to study the effect of class size on standardized test scores. Observational or experiment?
- Confounding variables?
- Tennessee STAR program: separate kindergarten students, *randomly* place them in smaller or larger classrooms for 4 years, compare standardized test results.
- Experimental units/subjects? Treatment? Outcomes/response?
- *The subjects are the students in the study. The treatment is to separate them into different classrooms. The outcome or response variable is their standardized test score.*

Example: Class Size

- We want to study the effect of class size on standardized test scores. Observational or experiment?
- Confounding variables?
- Tennessee STAR program: separate kindergarten students, *randomly* place them in smaller or larger classrooms for 4 years, compare standardized test results.
- Experimental units/subjects? Treatment? Outcomes/response?
- *The subjects are the students in the study. The treatment is to separate them into different classrooms. The outcome or response variable is their standardized test score.*
- A *factor* is an explanatory variable. What is the factor here?

Example: Class Size

- We want to study the effect of class size on standardized test scores. Observational or experiment?
- Confounding variables?
- Tennessee STAR program: separate kindergarten students, *randomly* place them in smaller or larger classrooms for 4 years, compare standardized test results.
- Experimental units/subjects? Treatment? Outcomes/response?
- *The subjects are the students in the study. The treatment is to separate them into different classrooms. The outcome or response variable is their standardized test score.*
- A *factor* is an explanatory variable. What is the factor here?
- *Classroom size!*

Example: Class Size

- We want to study the effect of class size on standardized test scores. Observational or experiment?
- Confounding variables?
- Tennessee STAR program: separate kindergarten students, *randomly* place them in smaller or larger classrooms for 4 years, compare standardized test results.
- Experimental units/subjects? Treatment? Outcomes/response?
- *The subjects are the students in the study. The treatment is to separate them into different classrooms. The outcome or response variable is their standardized test score.*
- A *factor* is an explanatory variable. What is the factor here?
- *Classroom size!*
- Sample or census?

Example: Class Size

- We want to study the effect of class size on standardized test scores. Observational or experiment?
- Confounding variables?
- Tennessee STAR program: separate kindergarten students, *randomly* place them in smaller or larger classrooms for 4 years, compare standardized test results.
- Experimental units/subjects? Treatment? Outcomes/response?
- *The subjects are the students in the study. The treatment is to separate them into different classrooms. The outcome or response variable is their standardized test score.*
- A *factor* is an explanatory variable. What is the factor here?
- *Classroom size!*
- Sample or census? *Sample*

Example: Anesthetic Safety

To investigate the safety of anesthetics used in surgery, records from 850,000 operations in 34 hospitals were collected. The death rates from 4 common anesthetics are shown below:

Anesthetic	A	B	C	D
Death Rate	1.7%	1.7%	3.4%	1.7%

- Observational or experimental?

Example: Anesthetic Safety

To investigate the safety of anesthetics used in surgery, records from 850,000 operations in 34 hospitals were collected. The death rates from 4 common anesthetics are shown below:

Anesthetic	A	B	C	D
Death Rate	1.7%	1.7%	3.4%	1.7%

- Observational or experimental?
- *Observational*
- Explanatory variable? Response variable?

Example: Anesthetic Safety

To investigate the safety of anesthetics used in surgery, records from 850,000 operations in 34 hospitals were collected. The death rates from 4 common anesthetics are shown below:

Anesthetic	A	B	C	D
Death Rate	1.7%	1.7%	3.4%	1.7%

- Observational or experimental?
- *Observational*
- Explanatory variable? Response variable?
- *The anesthetic used is the explanatory variable and the death rate is the response variable*

Example: Anesthetic Safety

To investigate the safety of anesthetics used in surgery, records from 850,000 operations in 34 hospitals were collected. The death rates from 4 common anesthetics are shown below:

Anesthetic	A	B	C	D
Death Rate	1.7%	1.7%	3.4%	1.7%

- Observational or experimental?
- *Observational*
- Explanatory variable? Response variable?
- *The anesthetic used is the explanatory variable and the death rate is the response variable*
- Conclusion?

Example: Anesthetic Safety

To investigate the safety of anesthetics used in surgery, records from 850,000 operations in 34 hospitals were collected. The death rates from 4 common anesthetics are shown below:

Anesthetic	A	B	C	D
Death Rate	1.7%	1.7%	3.4%	1.7%

- Observational or experimental?
- *Observational*
- Explanatory variable? Response variable?
- *The anesthetic used is the explanatory variable and the death rate is the response variable*
- Conclusion?
- *We can't conclude much here. There are probably lurking/confounding variables. Anesthetic C might be used in more serious cases that have a higher death rate anyway.*

Example: Heart Attacks

- Want to study the effect of aspirin and beta carotene on the rate of heart attack

Example: Heart Attacks

- Want to study the effect of aspirin and beta carotene on the rate of heart attack
- Physician's health study (1980-1995): 21,996 male physicians randomly divided into 4 groups, 2 placebo groups, 1 aspirin, 1 beta carotene. Outcome: 239 heart attacks in placebo, 139 in the aspirin group.

Example: Heart Attacks

- Want to study the effect of aspirin and beta carotene on the rate of heart attack
- Physician's health study (1980-1995): 21,996 male physicians randomly divided into 4 groups, 2 placebo groups, 1 aspirin, 1 beta carotene. Outcome: 239 heart attacks in placebo, 139 in the aspirin group.
- Placebo effect: individuals in the placebo group might *think* they have received treatment

Example: Heart Attacks

- Want to study the effect of aspirin and beta carotene on the rate of heart attack
- Physician's health study (1980-1995): 21,996 male physicians randomly divided into 4 groups, 2 placebo groups, 1 aspirin, 1 beta carotene. Outcome: 239 heart attacks in placebo, 139 in the aspirin group.
- Placebo effect: individuals in the placebo group might *think* they have received treatment
- Control group: the placebo groups

Example: Heart Attacks

- Want to study the effect of aspirin and beta carotene on the rate of heart attack
- Physician's health study (1980-1995): 21,996 male physicians randomly divided into 4 groups, 2 placebo groups, 1 aspirin, 1 beta carotene. Outcome: 239 heart attacks in placebo, 139 in the aspirin group.
- Placebo effect: individuals in the placebo group might *think* they have received treatment
- Control group: the placebo groups
- Treatment group: the groups who received the treatment

Example: Heart Attacks

- Want to study the effect of aspirin and beta carotene on the rate of heart attack
- Physician's health study (1980-1995): 21,996 male physicians randomly divided into 4 groups, 2 placebo groups, 1 aspirin, 1 beta carotene. Outcome: 239 heart attacks in placebo, 139 in the aspirin group.
- Placebo effect: individuals in the placebo group might *think* they have received treatment
- Control group: the placebo groups
- Treatment group: the groups who received the treatment
- Bias: the placebo effect might cause bias.

Example: Heart Attacks

- Want to study the effect of aspirin and beta carotene on the rate of heart attack
- Physician's health study (1980-1995): 21,996 male physicians randomly divided into 4 groups, 2 placebo groups, 1 aspirin, 1 beta carotene. Outcome: 239 heart attacks in placebo, 139 in the aspirin group.
- Placebo effect: individuals in the placebo group might *think* they have received treatment
- Control group: the placebo groups
- Treatment group: the groups who received the treatment
- Bias: the placebo effect might cause bias.
- Randomization: use a random number generator, etc to separate groups to eliminate other biases.

Examples: Polls

- Polls are a good way to predict the outcome of an election, etc.

Examples: Polls

- Polls are a good way to predict the outcome of an election, etc.
- Hard to eliminate bias! Ex: polling companies only call house phones, so the outcome is biased towards individuals with home phones. In recent elections, this eliminates many young voters from the polls.

Examples: Polls

- Polls are a good way to predict the outcome of an election, etc.
- Hard to eliminate bias! Ex: polling companies only call house phones, so the outcome is biased towards individuals with home phones. In recent elections, this eliminates many young voters from the polls.
- Ex: 1936 presidential election. 10 million questionnaires were mailed to magazine subscribers, car owners, and addresses in telephone books

Examples: Polls

- Polls are a good way to predict the outcome of an election, etc.
- Hard to eliminate bias! Ex: polling companies only call house phones, so the outcome is biased towards individuals with home phones. In recent elections, this eliminates many young voters from the polls.
- Ex: 1936 presidential election. 10 million questionnaires were mailed to magazine subscribers, car owners, and addresses in telephone books
- *Selection and Non-response Bias*: this is during the depression, so many people did not have money for magazines, cars, phones.

Experiment Designs

- Double-blind: neither the experimenters nor the subjects know which treatment they are administering/being administered.

Experiment Designs

- Double-blind: neither the experimenters nor the subjects know which treatment they are administering/being administered.
- Matched pair design: divide population into exactly two groups where pairs are 'matched' to eliminate other variables. E.g. pair people of same sex, age, income, etc. Then apply treatment to one group and compare the outcome.

Experiment Designs

- Double-blind: neither the experimenters nor the subjects know which treatment they are administering/being administered.
- Matched pair design: divide population into exactly two groups where pairs are 'matched' to eliminate other variables. E.g. pair people of same sex, age, income, etc. Then apply treatment to one group and compare the outcome.
- Block design: group subjects, then randomize within the groups.

Experiment Designs

- Double-blind: neither the experimenters nor the subjects know which treatment they are administering/being administered.
- Matched pair design: divide population into exactly two groups where pairs are 'matched' to eliminate other variables. E.g. pair people of same sex, age, income, etc. Then apply treatment to one group and compare the outcome.
- Block design: group subjects, then randomize within the groups.
- Example: 3.18 in textbook, cancer treatments.

Sampling Design and Basic Inference (3.3/3.4)

We want to *infer* some facts about a *population*. To do so, we need to extract a *sample*.

Sampling Design and Basic Inference (3.3/3.4)

We want to *infer* some facts about a *population*. To do so, we need to extract a *sample*.

– Example: Of the 227,762 housing units (houses or apts) in Tucson, how many own a gun?

Sampling Design and Basic Inference (3.3/3.4)

We want to *infer* some facts about a *population*. To do so, we need to extract a *sample*.

– Example: Of the 227,762 housing units (houses or apts) in Tucson, how many own a gun?

– We assume there is a 'true' number, i.e. if we took a *census*, we would find (say) 32,005 homes ($\approx 14.05\%$) with guns.

Sampling Design and Basic Inference (3.3/3.4)

We want to *infer* some facts about a *population*. To do so, we need to extract a *sample*.

– Example: Of the 227,762 housing units (houses or apts) in Tucson, how many own a gun?

– We assume there is a 'true' number, i.e. if we took a *census*, we would find (say) 32,005 homes ($\approx 14.05\%$) with guns.

– To *infer* this number, we would probably design a *sample* - contact say 1000 homes, ask if they have a gun.

Sampling Design and Basic Inference (3.3/3.4)

We want to *infer* some facts about a *population*. To do so, we need to extract a *sample*.

– Example: Of the 227,762 housing units (houses or apts) in Tucson, how many own a gun?

– We assume there is a 'true' number, i.e. if we took a *census*, we would find (say) 32,005 homes ($\approx 14.05\%$) with guns.

– To *infer* this number, we would probably design a *sample* - contact say 1000 homes, ask if they have a gun.

– How do we design such an experiment so that our sample is accurate?

Sampling Design and Basic Inference (3.3/3.4)

We want to *infer* some facts about a *population*. To do so, we need to extract a *sample*.

– Example: Of the 227,762 housing units (houses or apts) in Tucson, how many own a gun?

– We assume there is a 'true' number, i.e. if we took a *census*, we would find (say) 32,005 homes ($\approx 14.05\%$) with guns.

– To *infer* this number, we would probably design a *sample* - contact say 1000 homes, ask if they have a gun.

– How do we design such an experiment so that our sample is accurate?

– Bias: sampling bias, response bias, question wording

Sampling Design and Basic Inference (3.3/3.4)

We want to *infer* some facts about a *population*. To do so, we need to extract a *sample*.

– Example: Of the 227,762 housing units (houses or apts) in Tucson, how many own a gun?

– We assume there is a 'true' number, i.e. if we took a *census*, we would find (say) 32,005 homes ($\approx 14.05\%$) with guns.

– To *infer* this number, we would probably design a *sample* - contact say 1000 homes, ask if they have a gun.

– How do we design such an experiment so that our sample is accurate?

– Bias: sampling bias, response bias, question wording

– What if we took a *different* sample of 1000 people? Would we end up with the same sample statistic? Maybe we found 10% the first time, 7% the second time...

Population and Sample

– The *population* is the *entire* set of individuals that we want to know about. A *sample* is a subset of this group.
– A *parameter* is a **fixed, constant number** associated to a population.
– A *statistic* is computed from a **sample**.

Example

Twenty fourth-year English students are randomly selected to be on a committee to evaluate changes in the statistics requirement for the major. They are asked to vote 'yes' or 'no' to the question 'should stats be required?' Suppose there are 92 total fourth year English majors.

– What is the population and parameter?

Population and Sample

– The *population* is the *entire* set of individuals that we want to know about. A *sample* is a subset of this group.
– A *parameter* is a **fixed, constant number** associated to a population.
– A *statistic* is computed from a **sample**.

Example

Twenty fourth-year English students are randomly selected to be on a committee to evaluate changes in the statistics requirement for the major. They are asked to vote 'yes' or 'no' to the question 'should stats be required?' Suppose there are 92 total fourth year English majors.

– What is the population and parameter?
– What is the sample and statistic?

Population and Sample

– The *population* is the *entire* set of individuals that we want to know about. A *sample* is a subset of this group.
– A *parameter* is a **fixed, constant number** associated to a population.
– A *statistic* is computed from a **sample**.

Example

Twenty fourth-year English students are randomly selected to be on a committee to evaluate changes in the statistics requirement for the major. They are asked to vote 'yes' or 'no' to the question 'should stats be required?' Suppose there are 92 total fourth year English majors.

– What is the population and parameter?
– What is the sample and statistic?
– Why might we select fourth year students instead of another group?

Simple Random Samples and Stratified Samples

– A *simple random sample* chooses individuals from the entire population uniformly at random. We use a random number table or computer to generate the randomness.
– A *stratified* random sample divides the population into groups of similar individuals, then performs a SRS on each *strata*.
– A *multistage* sample divides the population into *clusters* (not necessarily similar!), then performs SRS to choose a **sample** of clusters, then performs another SRS to sub-sample within each cluster. Example: select 10 states at random, then in each selected state select 10 counties at random, then in each county select 100 people at random.

Sample Design Example

Example

Identify the following as an SRS, stratified, or multistage sample.

– There are 10 sections of 263. A random sample of 3 sections is chosen, then 8 students from each section are chosen at random.

Sample Design Example

Example

Identify the following as an SRS, stratified, or multistage sample.

- There are 10 sections of 263. A random sample of 3 sections is chosen, then 8 students from each section are chosen at random.
- 5 students from the UA student body are selected to win a prize using a random number generator

Sample Design Example

Example

Identify the following as an SRS, stratified, or multistage sample.

- There are 10 sections of 263. A random sample of 3 sections is chosen, then 8 students from each section are chosen at random.
- 5 students from the UA student body are selected to win a prize using a random number generator
- An online poll asks users of Hulu.com what their favorite TV show is.

Sample Design Example

Example

Identify the following as an SRS, stratified, or multistage sample.

- There are 10 sections of 263. A random sample of 3 sections is chosen, then 8 students from each section are chosen at random.
- 5 students from the UA student body are selected to win a prize using a random number generator
- An online poll asks users of Hulu.com what their favorite TV show is.
- Students are UA are selected to take a stats competency exam. We first divide students into male and female, then choose students at random from these groups.

Polls

Things to watch out for:

- How are individuals selected to take the poll, i.e. what is the sampling design? (Ex: polls reported on cable news)
- What is the response/nonresponse rate (i.e. how many people responded out of those selected to be given the poll)?
- How are the questions worded? (Some phrases/wording might elicit a stronger response from one group or another)

Using a random number table or generator

– Table B in the book contains a list of randomly generated numbers.
– Example: suppose we have a list of 100 students. Number them 1-100. Read off 10 numbers from table B:

$$03802, 77320, 07886, 87065, 42090, 55494, 16698, 16297, 22897, 98163$$

These are 5 digit numbers, so we only use the first 2 digits to determine our sample. We select students:

Using a random number table or generator

– Table B in the book contains a list of randomly generated numbers.
– Example: suppose we have a list of 100 students. Number them 1-100. Read off 10 numbers from table B:

$$03802, 77320, 07886, 87065, 42090, 55494, 16698, 16297, 22897, 98163$$

These are 5 digit numbers, so we only use the first 2 digits to determine our sample. We select students:

$$3, 77, 7, 87, 42, 55, 16, 22, 98$$

(We don't select 16 twice!)

Using a random number table or generator

– Table B in the book contains a list of randomly generated numbers.
– Example: suppose we have a list of 100 students. Number them 1-100. Read off 10 numbers from table B:

$$03802, 77320, 07886, 87065, 42090, 55494, 16698, 16297, 22897, 98163$$

These are 5 digit numbers, so we only use the first 2 digits to determine our sample. We select students:

$$3, 77, 7, 87, 42, 55, 16, 22, 98$$

(We don't select 16 twice!)
– On TI84: MATH:PRB:randInt(min,max,num):STO>L1

Using a random number table or generator

– Table B in the book contains a list of randomly generated numbers.
– Example: suppose we have a list of 100 students. Number them 1-100. Read off 10 numbers from table B:

$$03802, 77320, 07886, 87065, 42090, 55494, 16698, 16297, 22897, 98163$$

These are 5 digit numbers, so we only use the first 2 digits to determine our sample. We select students:

$$3, 77, 7, 87, 42, 55, 16, 22, 98$$

(We don't select 16 twice!)
– On TI84: MATH:PRB:randInt(min,max,num):STO>L1
– In excel: RAND() generates a single random number from 0 to 1. To find a random *real number* between a and b, use

$$\text{RAND}() * (b - a) + a$$

To find a random *integer* between a and b, simply wrap the above in INT().

Sampling Distribution

A central goal of statistics: use samples to determine behavior of a population. You must *repeat* an experiment or sample to justify inference

Sampling Distribution

A central goal of statistics: use samples to determine behavior of a population. You must *repeat* an experiment or sample to justify inference

– Do we choose a *single* large sample or *many* small samples? It depends. Large samples might lead to more experimental error, small samples might exhibit bias or systemic error.

Sampling Distribution

A central goal of statistics: use samples to determine behavior of a population. You must *repeat* an experiment or sample to justify inference

– Do we choose a *single* large sample or *many* small samples? It depends. Large samples might lead to more experimental error, small samples might exhibit bias or systemic error.

– Bias vs variability: bias is systemic error, variability is natural because of randomness (similar to accuracy vs. precision)

Sampling Distribution

A central goal of statistics: use samples to determine behavior of a population. You must *repeat* an experiment or sample to justify inference

– Do we choose a *single* large sample or *many* small samples? It depends. Large samples might lead to more experimental error, small samples might exhibit bias or systemic error.

– Bias vs variability: bias is systemic error, variability is natural because of randomness (similar to accuracy vs. precision)

– Ideal situation: choose *many* large samples.

Sampling Distribution

A central goal of statistics: use samples to determine behavior of a population. You must *repeat* an experiment or sample to justify inference

– Do we choose a *single* large sample or *many* small samples? It depends. Large samples might lead to more experimental error, small samples might exhibit bias or systemic error.

– Bias vs variability: bias is systemic error, variability is natural because of randomness (similar to accuracy vs. precision)

– Ideal situation: choose *many* large samples.

– The *sampling distribution* of a statistic: choose many samples. For each sample, compute the statistic. Make a histogram of the resulting numbers.

Sampling Distribution: Example

Sampling Distribution: Example

– Example: selecting groups of 100 students at random, we compute their average high school GPA. We do this 10 times and arrive at the following list:

$$3.25, 2.99, 3.41, 3.22, 3.05, 3.63, 3.55, 3.68, 3.15, 3.24$$

We can make a histogram with this list! The actual average (population mean) is 3.4, does our sampling procedure appear to give an *unbiased estimator* of the actual population mean?

Sampling Distribution: Example

– Example: selecting groups of 100 students at random, we compute their average high school GPA. We do this 10 times and arrive at the following list:

$$3.25, 2.99, 3.41, 3.22, 3.05, 3.63, 3.55, 3.68, 3.15, 3.24$$

We can make a histogram with this list! The actual average (population mean) is 3.4, does our sampling procedure appear to give an *unbiased estimator* of the actual population mean? – How might we reduce bias or variability?