More on Describing Distributions

Two more tools for describing continuous distributions

Percentiles

"Below what point in the distribution do 75% of the cases fall?"

Standardized (or "Z") scores

"How many standard deviations from the mean does a particular observation fall?"

Percentiles

Percentile

The value "below which a given percentage of the observations in a distribution falls"

Why would we want to know the point on a distribution below which a given percentage of the cases fall?

Example: "Let's give a tax break to people at or below the 80th percentile of the income distribution"

Example: "Let's give a scholarship to high school students above the 95th percentile of the GPA distribution"

Percentiles

We saw something like this earlier when we talked about cumulative percentages

Values Values	Frequency	Cumulative Frequency	Cumulative Percentage
0	5	5	16.7%
1	10	15	50.0%
2	10	25	83.3%
3	5	30	100.0%

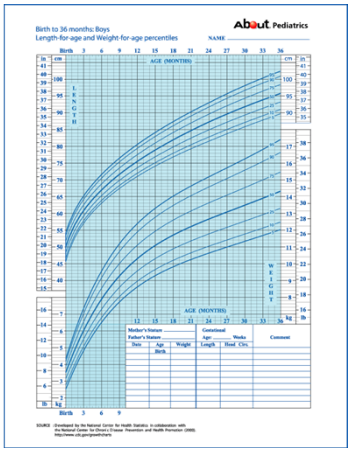
Here it's pretty straightforward to figure out where the Xth percentile is; we simply find the lowest value at which the cumulative percentage equals or exceeds X%

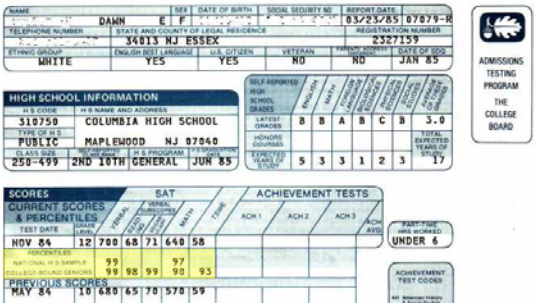
Percentiles

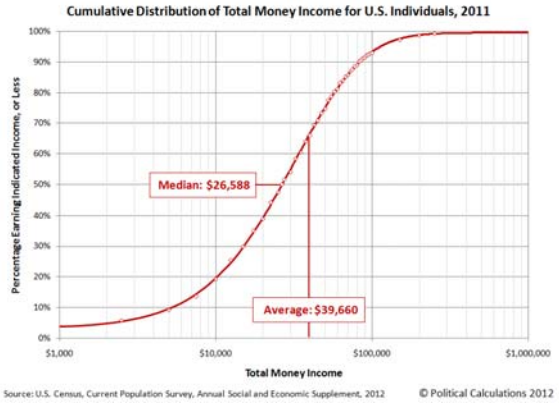
We saw something like this earlier when we talked about cumulative percentages

Values Values	Frequency	Cumulative Frequency	Cumulative Percentage
0	5	5	16.7%
1	10	15	50.0%
2	10	25	83.3%
3	5	30	100.0%

What is the 40th percentile of this distribution? 40% of the cases fall at or below the value 1. What is the 75th percentile? 75% of the vases fall at or below 2







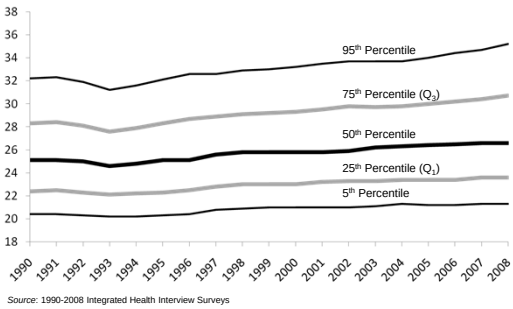
Percentiles

Percentiles are a more general form of the "five number summary" we talked about last time

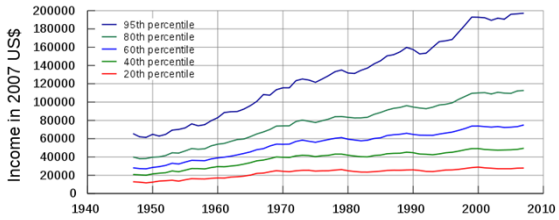
Minimum	➔	0^{th} Percentile
Q_1	➔	25^{th} Percentile
Median	➔	50^{th} Percentile
Q_3	➔	75^{th} Percentile
Maximum	➔	100^{th} Percentile

Why *these* values? Other substantive applications may call for different "cut points"

Body Mass Index, 1990-2008



Source: 1990-2008 Integrated Health Interview Surveys



Percentiles

Quantiles

Divisions of the distribution into groups with known (and equal) proportions in each group

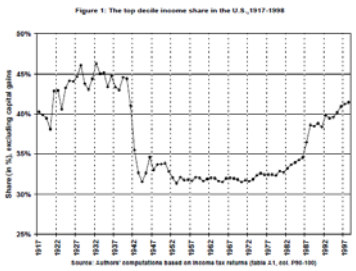
Examples:

Quartiles = divisions of the distribution into four equal sized groups at the 25th percentile, the 50th percentile, and the 75th percentile

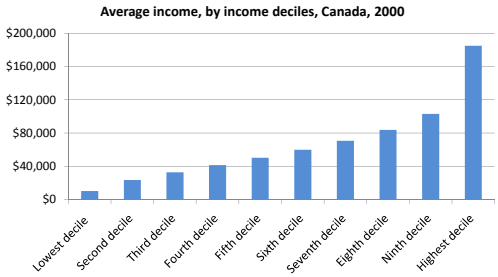
Quintiles = divisions of the distribution into five equal sized groups at the 20th percentile, the 40th percentile, etc.

Deciles = divisions of the distribution into ten equal sized groups at the 10th percentile, the 20th percentile, the 30th percentile, etc.

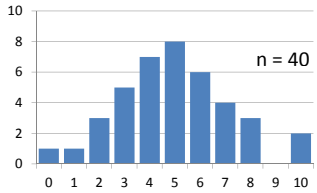
Percentiles



Percentiles



Worksheet



1. What is the 23rd percentile of this distribution?
2. What is the 99th percentile of this distribution?
3. If you wanted to divide this distribution into quintiles—five even sized groups—what values of the distribution would you use to separate the quartiles?

Change in Topic…

Standardized (Z) Scores

Standardized (Z) Score

"A transformation of the observed values of a continuous variable accomplished by subtracting the mean from each value and dividing by the standard deviation"

$$Z_i = \frac{(Y_i - \bar{Y})}{s_Y}$$

Example: Imagine a distribution of Y with a mean of 10 and a standard deviation of 4. How many standard deviations from the mean is a case with Y = 5?

$$Z_i = \frac{(5-10)}{4} = -1.25$$

Standardized (Z) Scores

Z > 0 ➔ Observation has a value of Y above the mean

Z = 0 ➔ Observation has a value of Y equal to the mean

Z < 0 ➔ Observation has a value of Y below the mean

"Standardizing" a distribution by converting all of the observed values of Y to Z-scores results in a "standardized distribution" with a mean of 0 and a standard deviation of 1

Standardized (Z) Scores

Because all standardized distributions have a mean of 0 and a standard deviation of 1, it becomes possible to compare values across distributions that have different means and standard deviations

Example: Who is more unusually tall: A man who is 72 inches tall or a woman who is 67 inches tall?

$\bar{Y}_{Men}$ = 69.1" and s_{Men} = 2.8

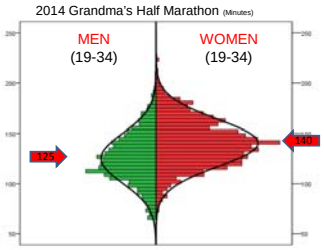
$\bar{Y}_{Women}$ = 63.8" and s_{women} = 2.6

$$Z_{72"\text{Man}} = \frac{(72.0-69.1)}{2.8} = 1.04 \qquad Z_{67"\text{Woman}} = \frac{(67.0-63.8)}{2.6} = 1.23$$

Worksheet

Who is more unusual? A man finishing the race in 109 minutes or a woman finishing in 124 minutes?

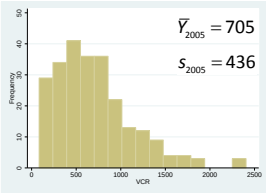
$\bar{Y}_{Men} = 125$
$s_{Men} = 15$
$\bar{Y}_{Women} = 140$
$s_{women} = 11$



Standardized (Z) Scores

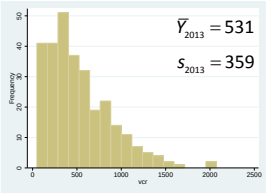
Here are violent crime rates (violent crimes per 100,000 residents) in 2005 and 2013 for cities in America with at least 100,000 people (from the Uniform Crime Reports)

2005



Minneapolis	1454
St Paul	876

2013



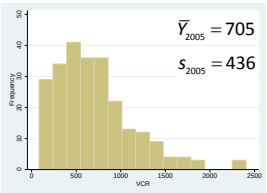
Minneapolis	1019
St Paul	747

How happy should Minneapolis and St Paul be about how much their violent crime rates decline between 2005 and 2013?

Standardized (Z) Scores

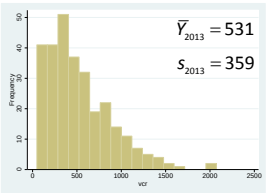
Here are violent crime rates (violent crimes per 100,000 residents) in 2005 and 2013 for cities in America with at least 100,000 people (from the Uniform Crime Reports)

2005



Minneapolis	Z=1.72
St Paul	Z=0.39

2013



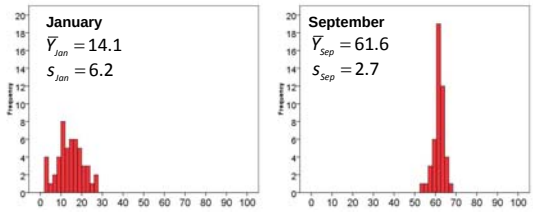
Minneapolis	Z=1.36
St Paul	Z=0.60

Both saw declines in violent crime rates, but Minneapolis's rate declined faster than the national average while St Paul's declined slower than the national average

Worksheet

Mean Monthly Temperature in St Paul, MN, 1957-2006

What is more unusual? A January that averages 25 degrees or a September that averages 65 degrees?

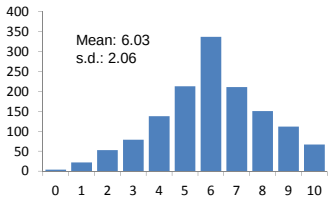


Vocabulary Test

Example: **BEAST** 1. afraid 2. words 3. large 4. animal 5. separate

A. **SPACE**	1. school 2. noon 3. captain 4. room 5. board
B. **BROADEN**	1. efface 2. make level 3. elapse 4. embroider 5. widen
C. **EMANATE**	1. populate 2. free 3. prominent 4. rival 5. come
D. **EDIBLE**	1. auspicious 2. eligible 3. fit to eat 4. sagacious 5. able to speak
E. **ANIMOSITY**	1. hatred 2. animation 3. disobedience 4. diversity 5. friendship
F. **PACT**	1. puissance 2. remonstrance 3. agreement 4. skillet 5. pressure
G. **CLOISTERED**	1. miniature 2. bunched 3. arched 4. malady 5. secluded
H. **CAPRICE**	1. value 2. a star 3. grimace 4. whim 5. inducement
I. **ACCUSTOM**	1. disappoint 2. customary 3. encounter 4. get used to 5. business
J. **ALLUSION**	1. reference 2. dream 3. eulogy 4. illusion 5. aria

Worksheet



	Cum. %
0	0%
1	2%
2	6%
3	11%
4	21%
5	37%
6	61%
7	76%
8	87%
9	95%
10	100%

- What is <u>your</u> Z-score?
- In what percentile is <u>your</u> score?

BREAK

Sampling

Minnesota Senate - McFadden vs. Franken

Candidates		Minnesota Snapshot
 Al Franken (D)* Bio | Campaign Site	 Mike McFadden (R) Bio | Campaign Site	RCP Average: Franken +10.0 RCP Ranking: Likely Dem 2014 Key Races: Governor | MN-1 | MN-2 | MN-7 | MN-8 ----------PAST KEY RACES---------- 2012: President | Senate | MN-6 | MN-8 2010: Governor | MN-1 | MN-6 | MN-7 | MN-8 2008: President | Senate | MN-3 | MN-6 2006: Senate | Governor | MN-6 2004: President

Polling Data						
Poll	Date	Sample	MoE	Franken (D)	McFadden (R)	Spread
Final Results	--	--	--	53.2	42.9	Franken +10.3
RCP Average	10/16 - 10/30	--	--	50.0	40.0	Franken +10.0
KSTP/SurveyUSA*	10/27 - 10/30	596 LV	4.1	51	40	Franken +11
Star Tribune/Mason-Dixon*	10/20 - 10/22	800 LV	3.5	48	39	Franken +9
CBS News/NYT/YouGov	10/16 - 10/23	2430 LV	3.0	51	41	Franken +10
All Minnesota Senate - McFadden vs. Franken Polling Data						

http://www.realclearpolitics.com/epolls/2014/senate/mn/minnesota_senate_mcfadden_vs_franken-3902.html

Sampling

Population

All of the individuals (or "units") about whom we wish to make conclusions

Sample

A subset of this population; it is this subset that we actually observe (or survey or interview or whatever)

Sampling

The process of selecting individuals from the population for inclusion in the sample

Sampling

We actually observe…	*We want to draw conclusions about…* a **Sample** of cases	*We want to draw conclusions about…* a **Population** of Cases
a **Sample** of cases	Descriptive Statistics	Inferential Statistics
a **Population** of Cases	(Why would we do this?)	Descriptive Statistics

Sampling

Population Parameter

An attribute of the entire population. For example, the mean or variance of some variable in the full population

Sample Statistic (or Estimate)

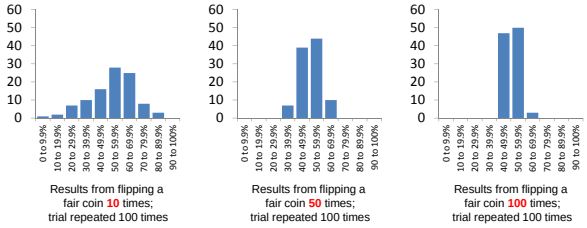
An attribute of a sample, sometimes used to make inferences about the corresponding population parameter

Sampling Error

The difference between a population parameter and the sample statistic being used to estimate it

Sampling Examples

Question: If you flip a fair coin, what percentage of the time will it come up "heads?"



Results from flipping a fair coin **10** times; trial repeated 100 times

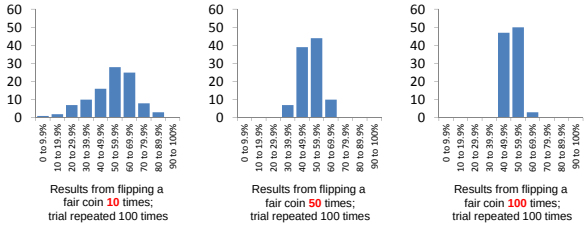
Results from flipping a fair coin **50** times; trial repeated 100 times

Results from flipping a fair coin **100** times; trial repeated 100 times

Population Parameter: Percentage of times "heads" comes up among all fair coins
Sample Statistic: Percentage of times "heads" comes up out of 10 flips of a fair coin

Sampling Examples

Question: If you flip a fair coin, what percentage of the time will it come up "heads?"



Results from flipping a fair coin **10** times; trial repeated 100 times

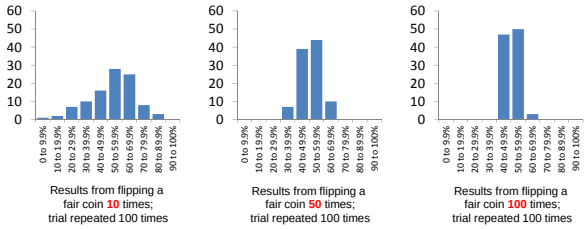
Results from flipping a fair coin **50** times; trial repeated 100 times

Results from flipping a fair coin **100** times; trial repeated 100 times

This is a dumb example because we know the population parameter … it's 50%
(It's also too convenient, because all fair coins are the same)

Sampling Examples

Question: If you flip a fair coin, what percentage of the time will it come up "heads?"



Results from flipping a fair coin **10** times; trial repeated 100 times

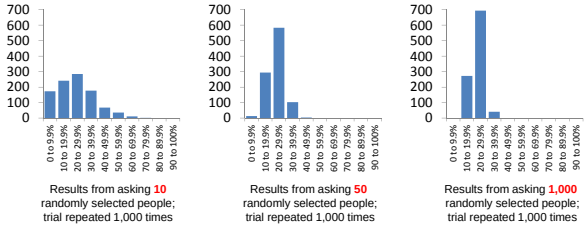
Results from flipping a fair coin **50** times; trial repeated 100 times

Results from flipping a fair coin **100** times; trial repeated 100 times

Note that (a) there is sampling error and (b) the size of the sampling error goes down as the sample size (i.e., number of flips) goes up

Sampling Examples

Question: What percentage of Americans were high school graduates in 1940?



Results from asking **10** randomly selected people; trial repeated 1,000 times

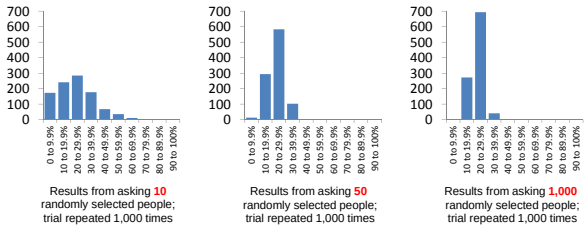
Results from asking **50** randomly selected people; trial repeated 1,000 times

Results from asking **1,000** randomly selected people; trial repeated 1,000 times

This is a *also* a dumb example because we know the population parameter … it's 23%
(I know this because we have access to the full 1940 U.S. Census data)

Sampling Examples

Question: What percentage of Americans were high school graduates in 1940?



Results from asking **10** randomly selected people; trial repeated 1,000 times

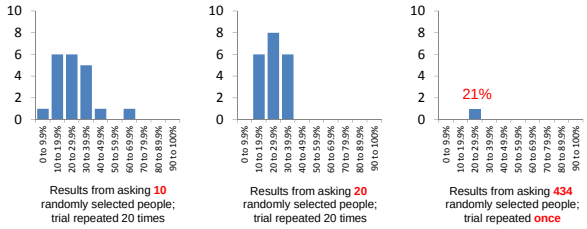
Results from asking **50** randomly selected people; trial repeated 1,000 times

Results from asking **1,000** randomly selected people; trial repeated 1,000 times

Note again that (a) there is sampling error and (b) the size of the sampling error goes down as the sample size (i.e., number of people sampled) goes up

Sampling Examples

Question: What percentage of Americans say it is true that humans evolved from an earlier species?



Results from asking **10** randomly selected people; trial repeated 20 times

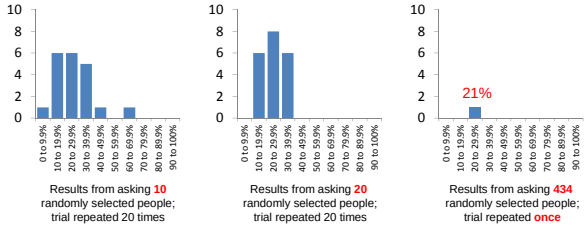
Results from asking **20** randomly selected people; trial repeated 20 times

Results from asking **434** randomly selected people; trial repeated **once**

This is a *not* a dumb example … we do not know the population value.
However, if the sampling procedures are sound, the value is certainly close to 21%

Sampling Examples

Question: What percentage of Americans say it is true that humans evolved from an earlier species?



Results from asking **10** randomly selected people; trial repeated 20 times

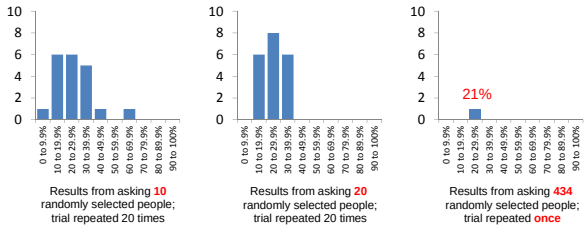
Results from asking **20** randomly selected people; trial repeated 20 times

Results from asking **434** randomly selected people; trial repeated **once**

Again, in the left two figures sampling error goes down as sample size goes up

Sampling Examples

Question: What percentage of Americans say it is true that humans evolved from an earlier species?



Results from asking **10** randomly selected people; trial repeated 20 times

Results from asking **20** randomly selected people; trial repeated 20 times

Results from asking **434** randomly selected people; trial repeated **once**

How confident should we be in the 21% result, obtained from just one sample of size n=434?

Sampling Error

How much sampling error should we expect?

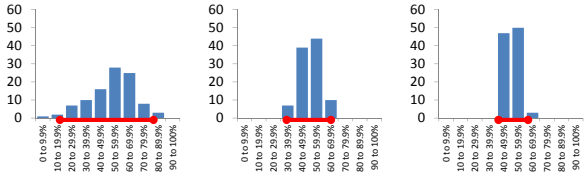
When we are talking about percentages, a good guess is given by the *conservative margin of error*

If the sample is representative of the population from which it was drawn, the sample percentage and the population percentage will differ by less than the conservative margin of error at least 95% of the time

$$\text{conservative margin of error} = \pm\frac{1}{\sqrt{n}} \times 100\%$$

Sampling Error

Question: If you flip a fair coin, what percentage of the time will it come up "heads?"



Results from flipping a fair coin **10** times; trial repeated 100 times

$\pm\frac{1}{\sqrt{10}} \times 100\% = \pm 32\%$

Results from flipping a fair coin **50** times; trial repeated 100 times

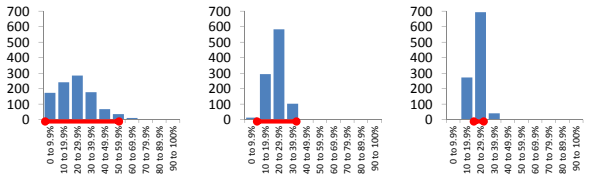
$\pm\frac{1}{\sqrt{50}} \times 100\% = \pm 14\%$

Results from flipping a fair coin **100** times; trial repeated 100 times

$\pm\frac{1}{\sqrt{100}} \times 100\% = \pm 10\%$

Sampling Error

Question: What percentage of Americans were high school graduates in 1940?



Results from asking **10** randomly selected people; trial repeated 1,000 times

$$\pm\frac{1}{\sqrt{10}}\times 100\% = \pm 32\%$$

Results from asking **50** randomly selected people; trial repeated 1,000 times

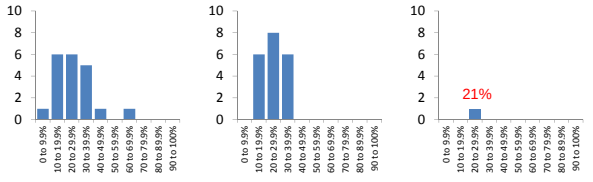
$$\pm\frac{1}{\sqrt{50}}\times 100\% = \pm 14\%$$

Results from asking **1,000** randomly selected people; trial repeated 1,000 times

$$\pm\frac{1}{\sqrt{1000}}\times 100\% = \pm 3\%$$

Sampling Error

Question: What percentage of Americans say it is true that humans evolved from an earlier species?



Results from asking **10** randomly selected people; trial repeated 20 times

$$\pm\frac{1}{\sqrt{10}}\times 100\% = \pm 32\%$$

Results from asking **20** randomly selected people; trial repeated 20 times

$$\pm\frac{1}{\sqrt{50}}\times 100\% = \pm 14\%$$

Results from asking **434** randomly selected people; trial repeated **once**

$$\pm\frac{1}{\sqrt{100}}\times 100\% = \pm 10\%$$

Where do you think the error bars should go?

Worksheet

I randomly sampled 120 U of M undergraduate students (by selecting students at random from the registrar's list of currently enrolled students). All 120 students that I sampled agreed to honestly and accurately answer all of my questions.

40% of those students admitted that at some point in their lives they have stolen something worth $50 or more

1. Construct a confidence interval for the proportion of all U of M undergraduate students who have ever stolen something worth $50 or more
2. How large would our sample have to be for our confidence interval to be ±1% ?

Change in Topic…

Methods for Sampling

We need samples to be representative of the populations from which they were drawn. What could go wrong?

Selection Bias

The procedure used to sample cases from the population is systematically flawed or biased

Non-Response Bias

The cases selected to be in the sample cannot be included or refuse to be included in the sample

Measurement Error

Our measures of sample members' attributes are flawed

Methods for Sampling

We rarely know the "true" values of population parameters in advance, and we can rarely observe entire populations

We infer population parameters based on sample estimates

This works (on average) if the sample is large and representative of the population from which it was drawn

How do we avoid systematic selection biases? How do we ensure that samples are representative of populations?

(Learn about avoiding non-response bias and measurement error in Sociology 3801)

Methods for Sampling

Why might the following samples be systematically biased? That is, what may lead to the samples not accurately reflecting the populations from which they are drawn?

Population	Sample
Newborn babies in Minneapolis	1,000 babies recently born in Minneapolis hospitals
Homeless people in St Paul	500 homeless people who are on the streets or in shelters
Adults in Minnesota	784 adults in Minnesota who could be contacted by telephone

Methods for Sampling

N = The number of individuals in the population

Sampling Frame = A list of all of the *N* individuals in the population

n = The number of individuals in the sample

n/*N* = The *probability of selection* … that is, the average probability that each of the *N* individuals in the population will be sampled

Methods for Sampling

In order to draw a representative sample from a population … that is, in order to avoid selection bias … two things must be true:

1. We must have a sampling frame
2. We must know every individual's probability of selection

Otherwise, we cannot determine whether the sample accurately reflects the population (because we cannot know whether segments of the population have been systematically excluded)

Methods for Sampling

Simple Random Sampling (SRS)

Individuals are selected from the population based strictly on chance, and the probability of selection is the same (*n*/*N*) for every individual in the population

A researcher simply selects *n* cases at random from the sampling frame

For practical reasons, SRS is not always very efficient

Methods for Sampling

Stratified Random Sampling

First, all individuals in the population are divided into strata on the basis of some relevant characteristic. Second, chance determines which individuals are sampled from each of these strata

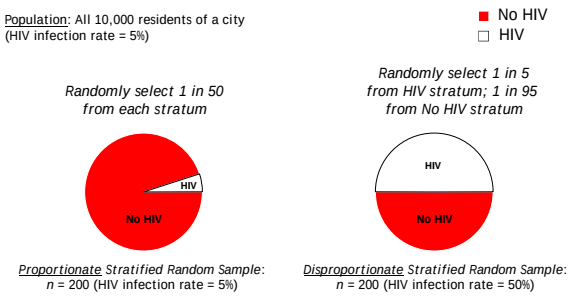
Proportionate Stratified Random Sampling

Each individual still has an *n*/*N* chance of being sampled

Disproportionate Stratified Random Sampling

n/*N* differs across strata for reasons of efficiency, but is the same within each strata

Methods for Sampling



Methods for Sampling

Cluster sampling

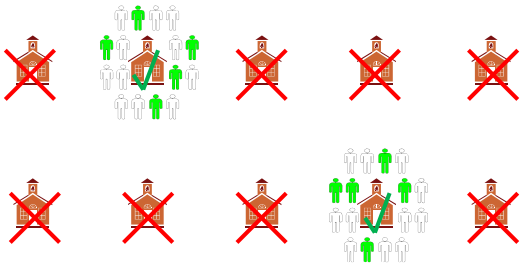
Cases are randomly selected from within naturally occurring "clusters" that are themselves randomly selected

As with stratified sampling, the goal is efficiency

Examples

- Students are clustered within schools
- Employees are clustered within companies
- Residents are clustered within neighborhoods

Methods for Sampling



= student selected at random from within randomly selected school

Worksheet

Imagine you want to sample 100 members of the Elk's Club to interview them about something. Which of the sampling procedures described above might be most effective and efficient? Why?

Imagine you want to compare current Minnesotans who were born in the United States to current Minnesotans who were born in Honduras. Which of the sampling procedures described above might be most effective and efficient? Why?

Methods for Sampling

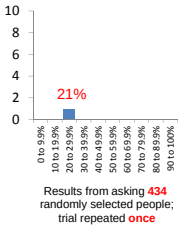
Simple random sampling, stratified random sampling, cluster sampling, and other methods are just procedures for drawing representative samples

We know that we can expect sampling error even when we use well-designed sampling procedures that successfully avoid systematic bias

However, if we avoid systematic bias by using procedures that generate representative samples, we can fairly precisely quantify the amount of sampling error

Sampling Examples

Question: What percentage of Americans say it is true that humans evolved from an earlier species?



How confident should we be in the 21% result, obtained from just one sample of size n=434?

If the sample is really representative of the population, then we are 95% certain that the population value is $21\% \pm (1/\sqrt{434}) \times 100\% = 21\% \pm 4.8\%$

Want More?

A Good Overview:

http://ccnmtl.columbia.edu/projects/qmss/samples_and_sampling/types_of_sampling.html

Another:

http://cnx.org/content/m16014/latest/