WHAT IS STATISTICS?

(Part 1)

What is *statistics?*

What is *statistics?*

Statistics is the science of planning studies and experiments; obtaining data; and then organizing, summarizing, presenting, analyzing, and interpreting those data and then drawing conclusions based on them.

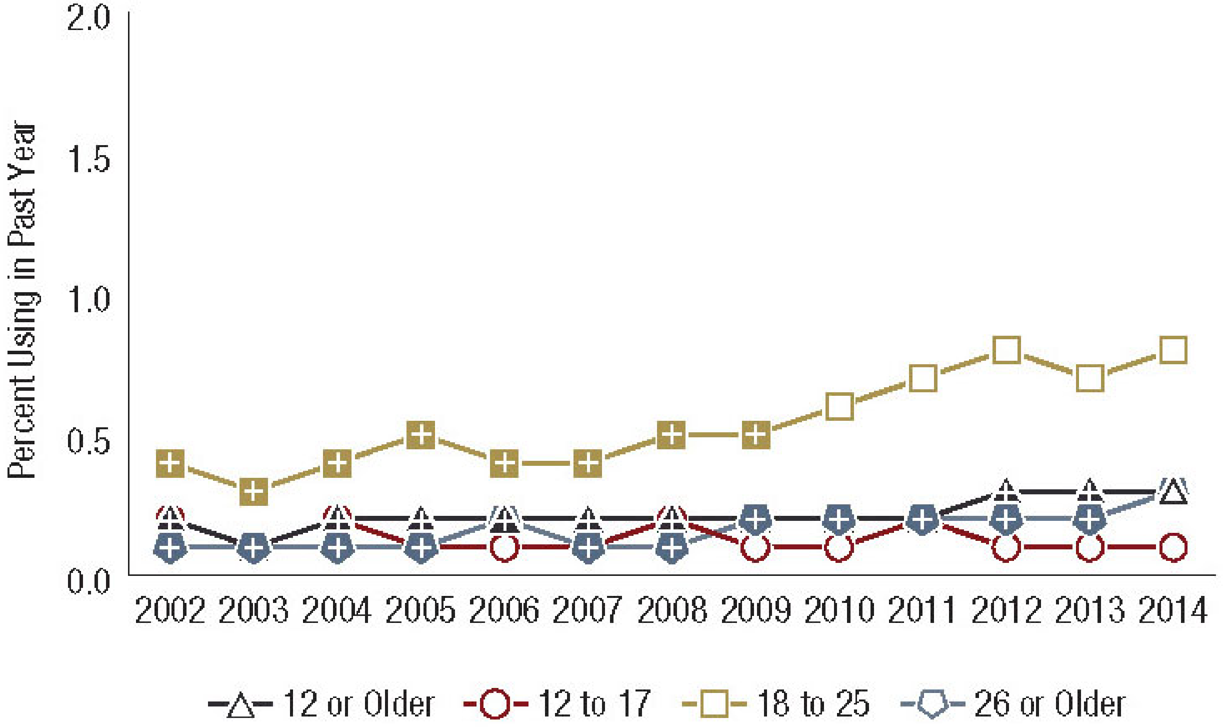
What is *descriptive statistics?*

What is *descriptive statistics?*

Descriptive statistics attempts to summarize and describe relevant characteristics of data.

Simple *descriptive statistics* can often add great clarity to a situation.

Figure 13. Past Year Heroin Use among People Aged 12 or Older, by Age Group: Percentages, 2002-2014



+ Difference between this estimate and the 2014 estimate is statistically significant at the .05 level.

Figure 13 Table. Past Year Heroin Use among People Aged 12 or Older, by Age Group: Percentages, 2002-2014

	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
12 or Older	0.2+	0.1+	0.2+	0.2+	0.2+	0.2+	0.2+	0.2+	0.2+	0.2+	0.3+	0.3+	0.3
12 to 17	0.2	0.1	0.2	0.1	0.1	0.1	0.2	0.1	0.1	0.2	0.1	0.1	0.1
18 to 25	0.4+	0.3+	0.4+	0.5+	0.4+	0.4+	0.5+	0.5+	0.6	0.7	0.8	0.7	0.8
26 or Older	0.1+	0.1+	0.1+	0.1+	0.2	0.1+	0.1+	0.2+	0.2+	0.2+	0.2+	0.2+	0.3

Figure 3.4 Binge Alcohol Use among Adults Aged 18 to 22, by College Enrollment: 2002-2013

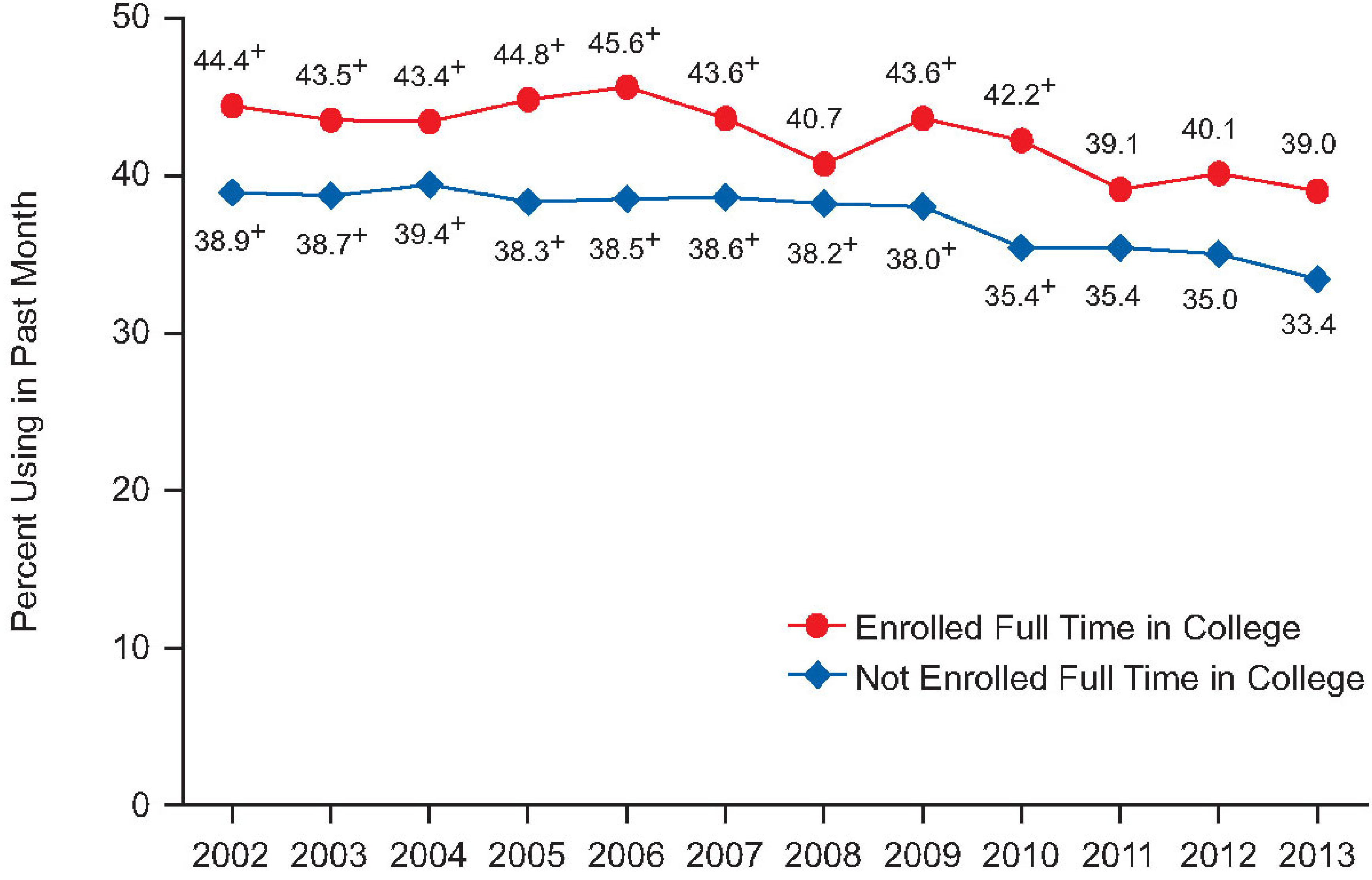
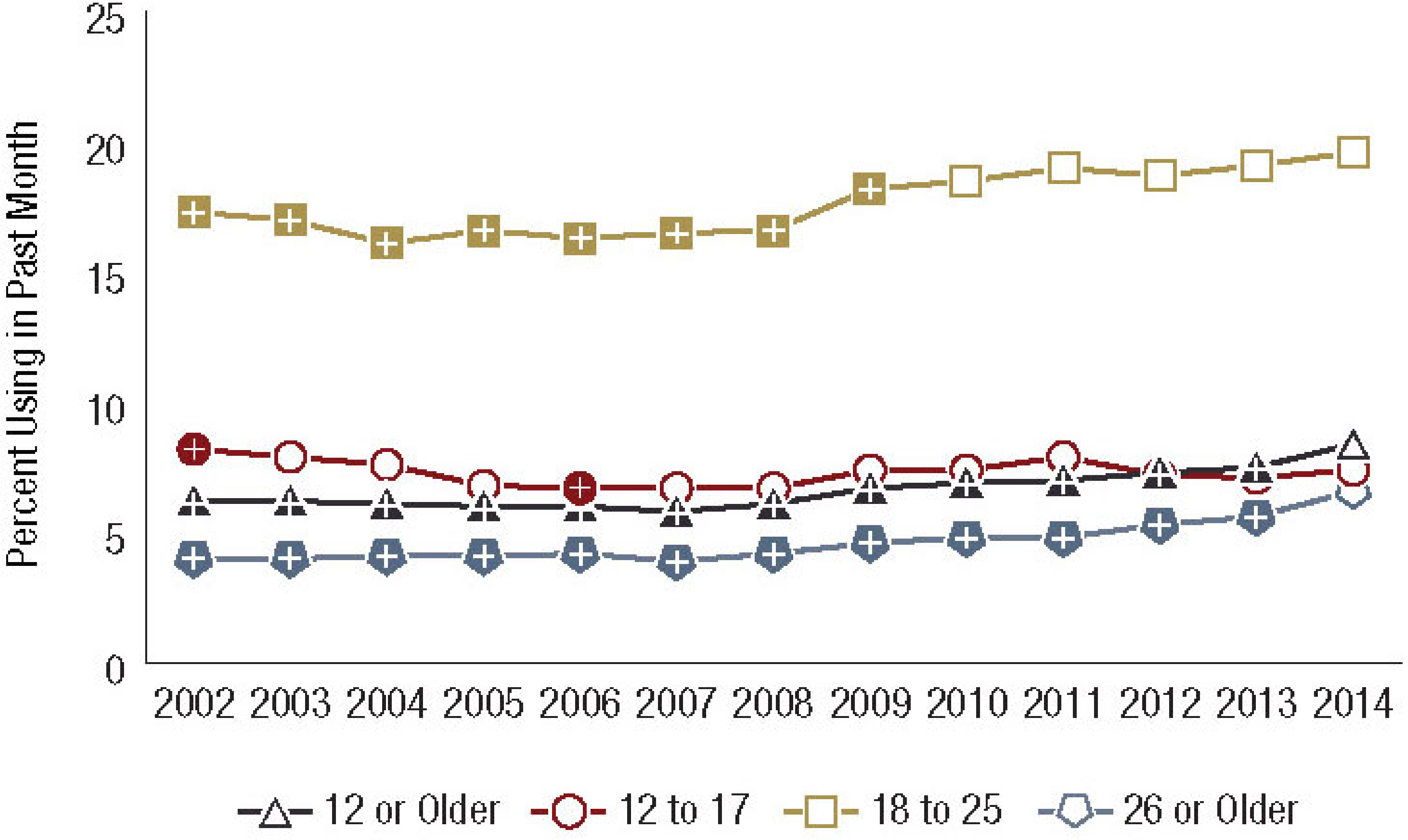
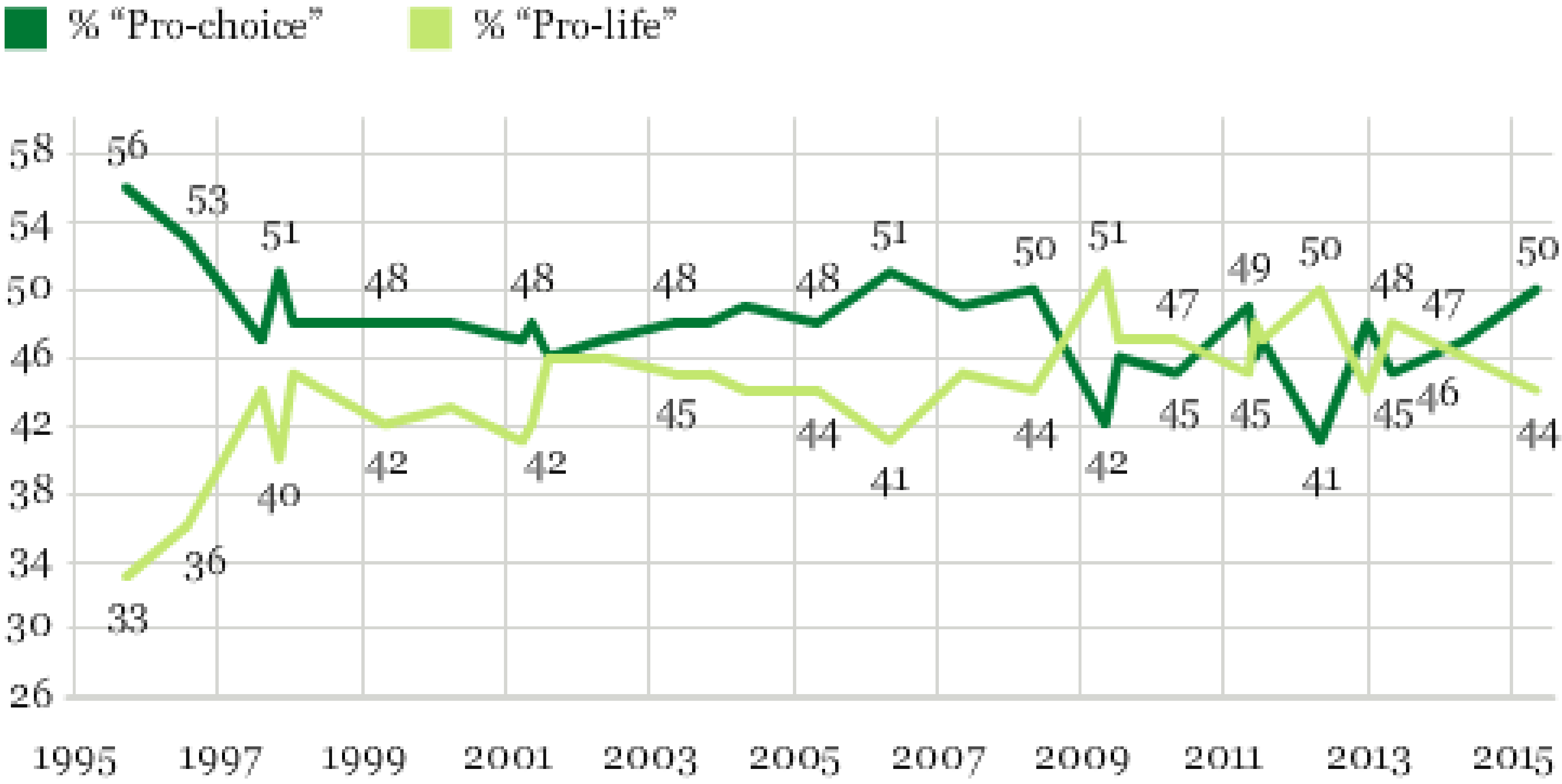


Figure 3. Past Month Marijuana Use among People Aged 12 or Older, by Age Group: Percentages, 2002-2014



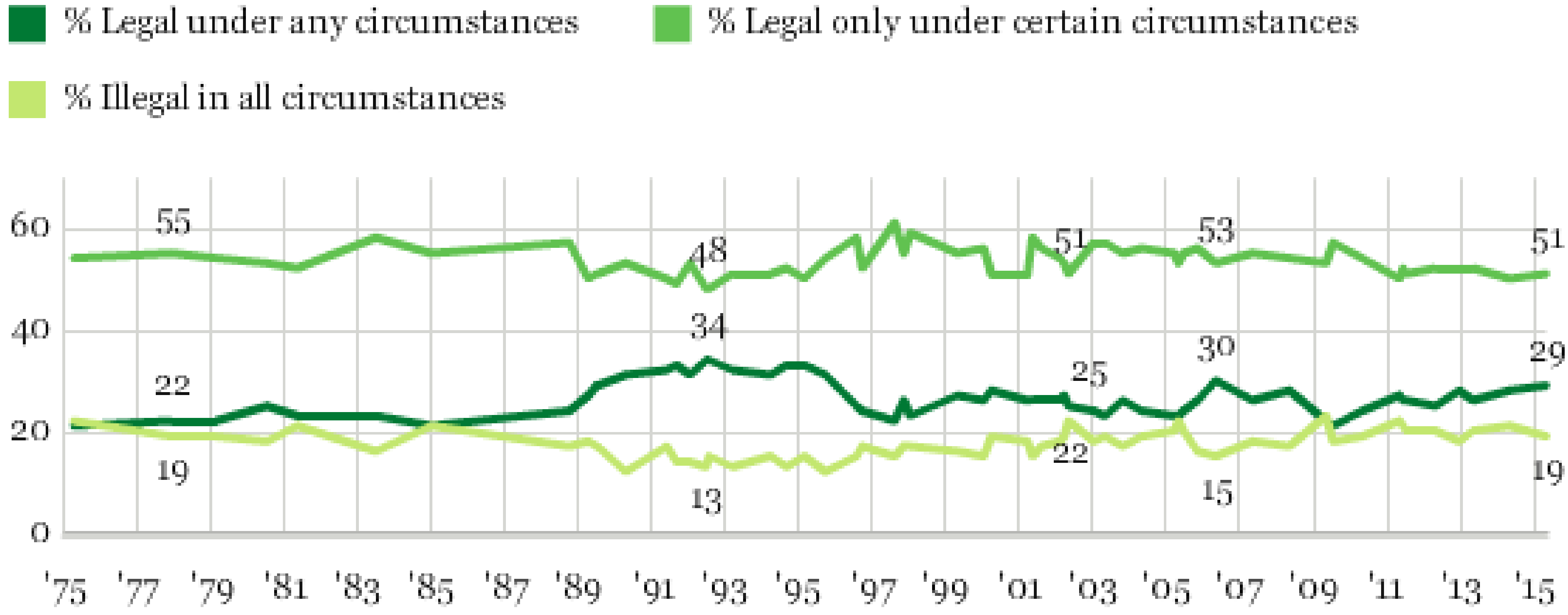
U.S. Adults' Self-Identified Position on the Abortion Issue



Question wording: With respect to the abortion issue, would you consider yourself to be pro-choice or pro-life?

GALLUP

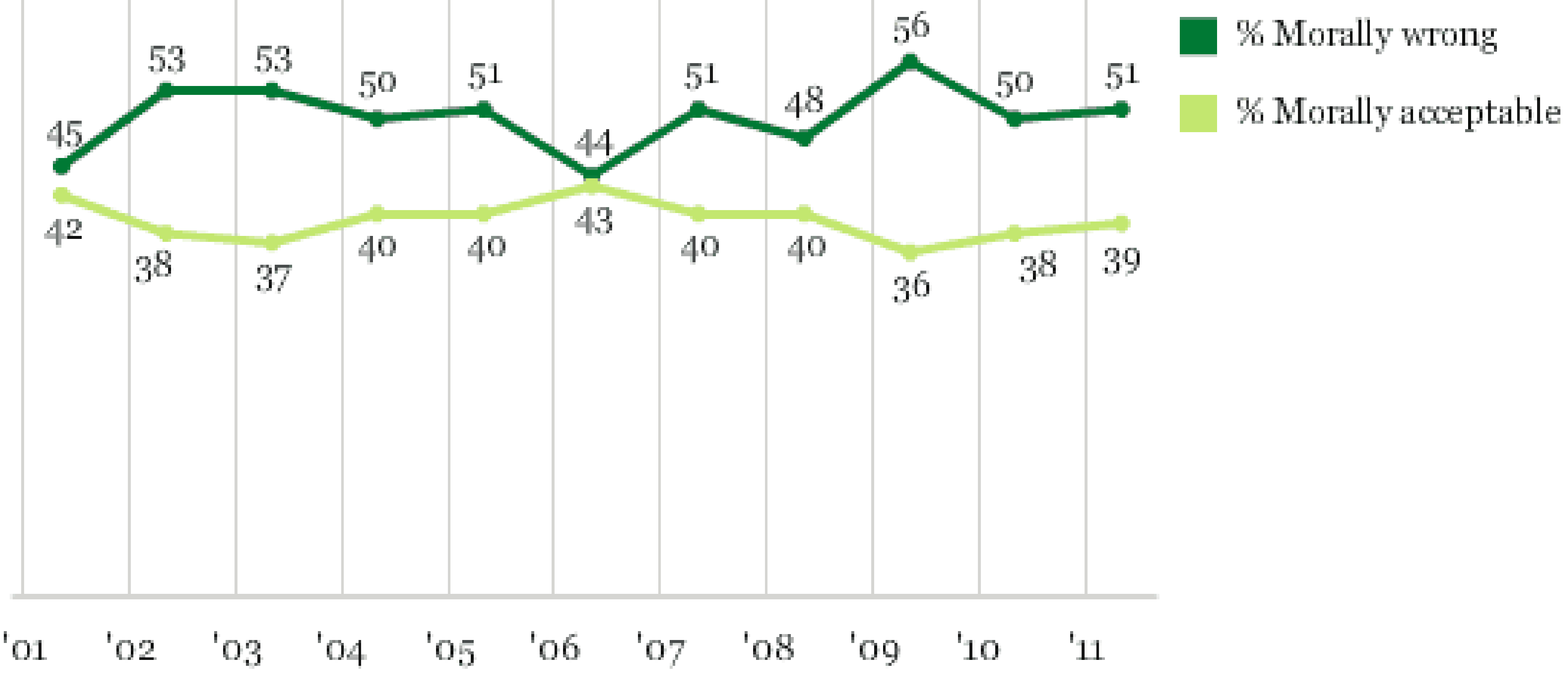
Do you think abortions should be legal under any circumstances, legal only under certain circumstances, or illegal in all circumstances?



GALLUP

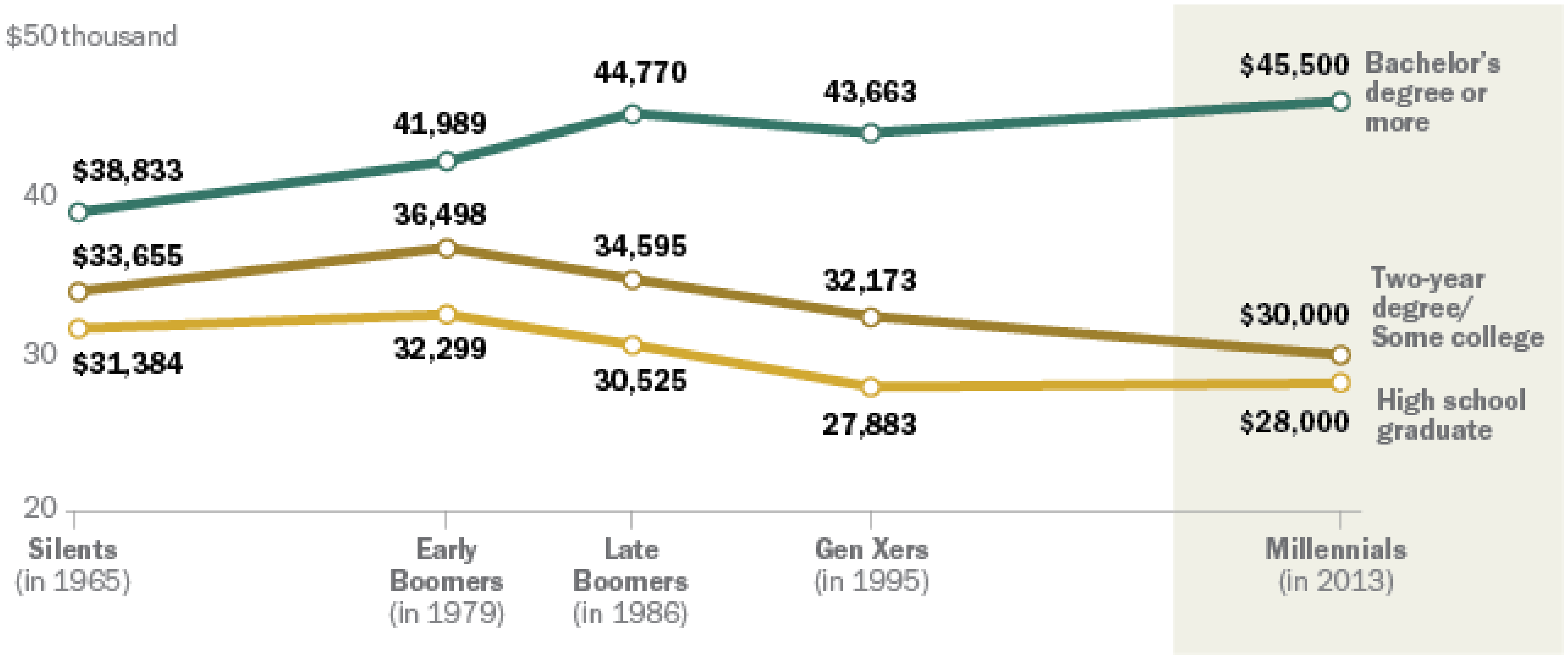
Americans' Views on the Morality of Abortion

Regardless of whether or not you think it should be legal, for each one, please tell me whether you personally believe that in general it is morally acceptable or morally wrong. How about -- abortion?



GALLUP

Median annual earnings among full-time workers ages 25 to 32, in 2012 dollars



Source: Pew Research Center tabulations of the 2013, 1995, 1986, 1979 and 1965 March Current Population Survey (CPS) IPUMS

PEW RESEARCH CENTER

What is *inferential statistics?*

What is *inferential statistics?*

Inferential statistics attempts to make inferences or generalizations about a population and also determine if results statistically significant.

What are *data?* (*data* is the plural of the singular *datum*)

What are *data?* (*data* is the plural of the singular *datum*)

Data are collections of observations such as measurements, genders, or survey responses.

What is a *population?*

What is a *population?*

A *population* is the complete collection of all measurements or data that are being considered.

What is a *sample?*

What is a *sample?*

A *sample* is a proper subcollection of members selected from a population.

What is a *census?*

What is a *census?*

A *census* is a collection of data from every member of a population.

What is a *random sample?*

What is a *random sample?*

A *random sample* is a sample in which each element of the population has an equally likely chance of being selected.

What is a *random sample?*

A *random sample* is a sample in which each element of the population has an equally likely chance of being selected.

We take random samples in an attempt to prevent *bias.*

What is *bias?*

What is *bias?*

Bias-

1. A particular tendency or inclination, especially one that prevents unprejudiced consideration of a question.

2. Statistics: A systematic as opposed to a random distortion of a measure as a result of a sampling procedure.

What is a *simple random sample of size n?*

(1.4)

What is a *simple random sample of size n?*

In a *simple random sample of size n*, every possible sample of size n has the same chance of being selected. If we pick a state at random and select it’s two senators, then this is a random sample where each senator has a 1 in 50 chance of being selected, but it is not a simple random sample of size 2 since not every possible sample of size 2 can be selected.

What is a *voluntary response sample?*

(1.2)

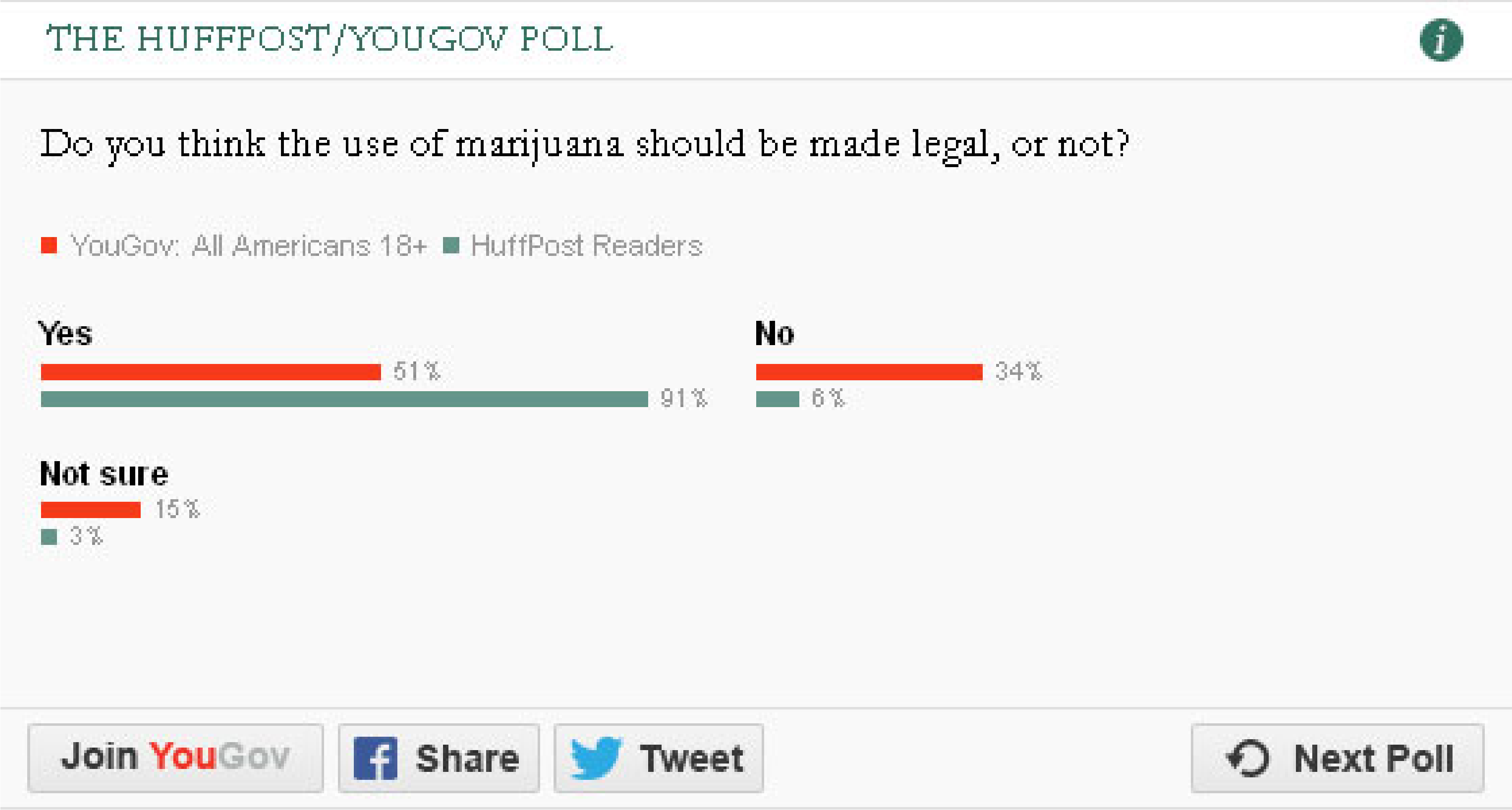
What is a *voluntary response sample?*

A *voluntary response sample* is one in which the respondents themselves choose or decide whether to be included. In other words, the sample members are self-selected volunteers.

What is a *voluntary response sample?*

A *voluntary response sample* is one in which the respondents themselves choose or decide whether to be included. In other words, the sample members are self-selected volunteers.

Voluntary response samples are not necessarily valid since they are very susceptible to bias.



THE HUFFPOST/YOUGOV POLL

Do you think the use of marijuana should be made legal, or not?

🟥 YouGov: All Americans 18+ 🟩 HuffPost Readers

Yes

51%

91%

No

34%

6%

Not sure

15%

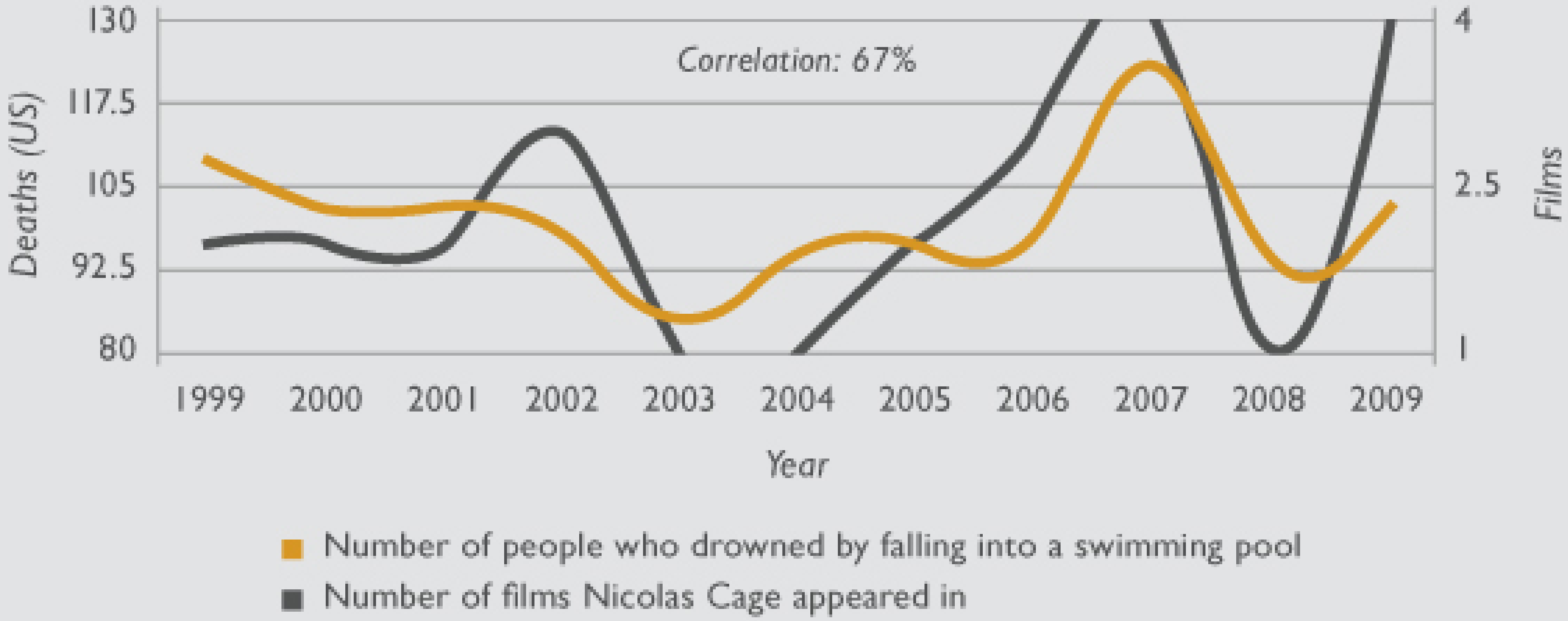
3%

A *correlation* between two variables means that there is a mathematical relationship between the two variables. Usually, either one of them increases as the other increases, or one of them decreases as the other increases.

Number of People Who Drowned by Falling Into a Swimming Pool

- correlates with -

Number of Films Nicolas Cage Appeared In

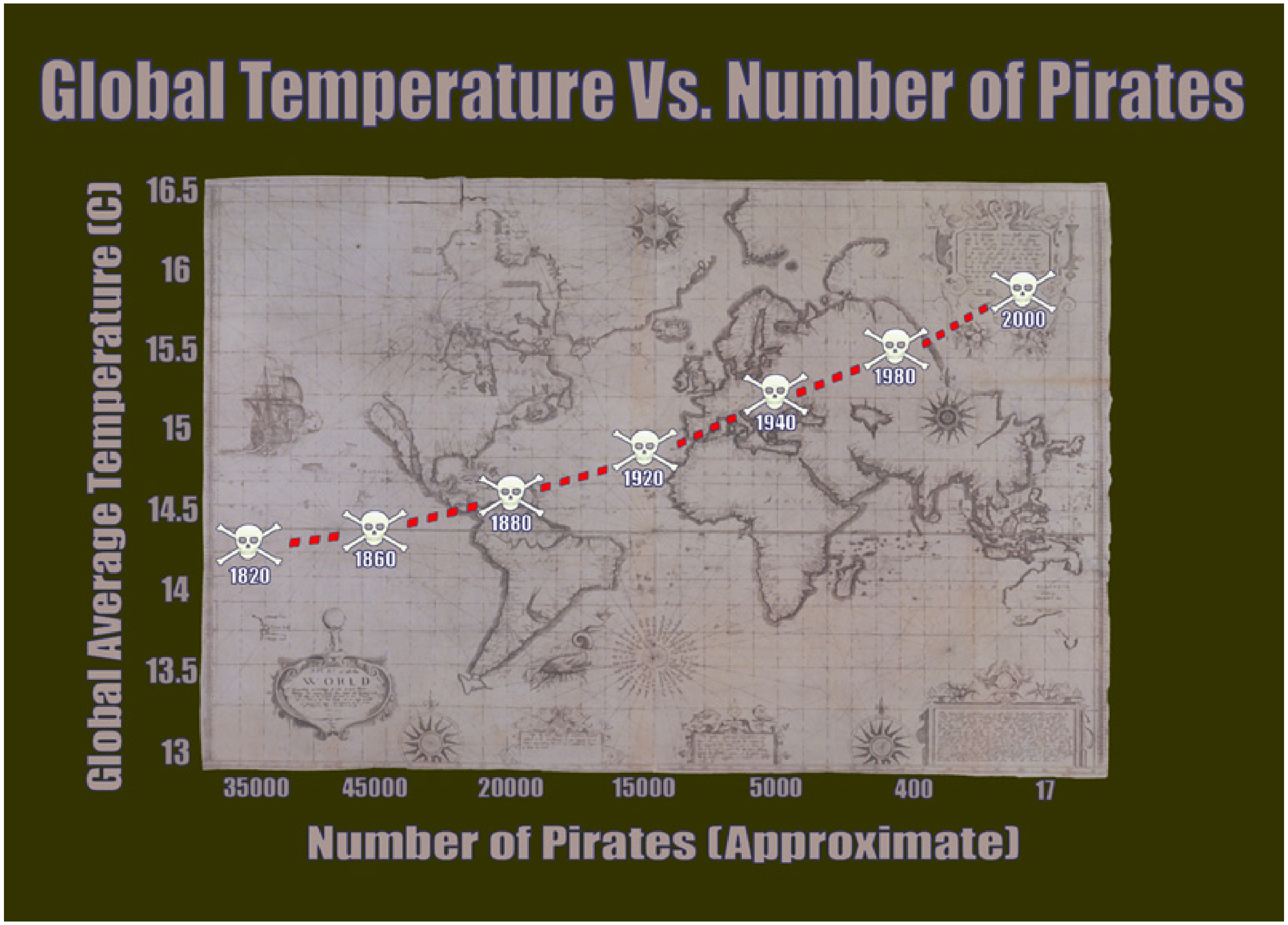


Sources: CDC & IMDB, tylervigen.com

Forbes / Leadership

MAR 23, 2012 @ 11:30 AM 102,768 VIEWS

True Fact: The Lack of Pirates Is Causing Global Warming



Does *correlation imply causality?*

Does *correlation imply causality?*

Never conclude that correlation proves causality. Correlation is a measure of the strength of the relationship between two variables. However, just because two variables are related does not mean that one causes the other. ***Correlation does not imply causality.*** However, given that caveat, there may be situations where causality is a reasonable or educated guess.

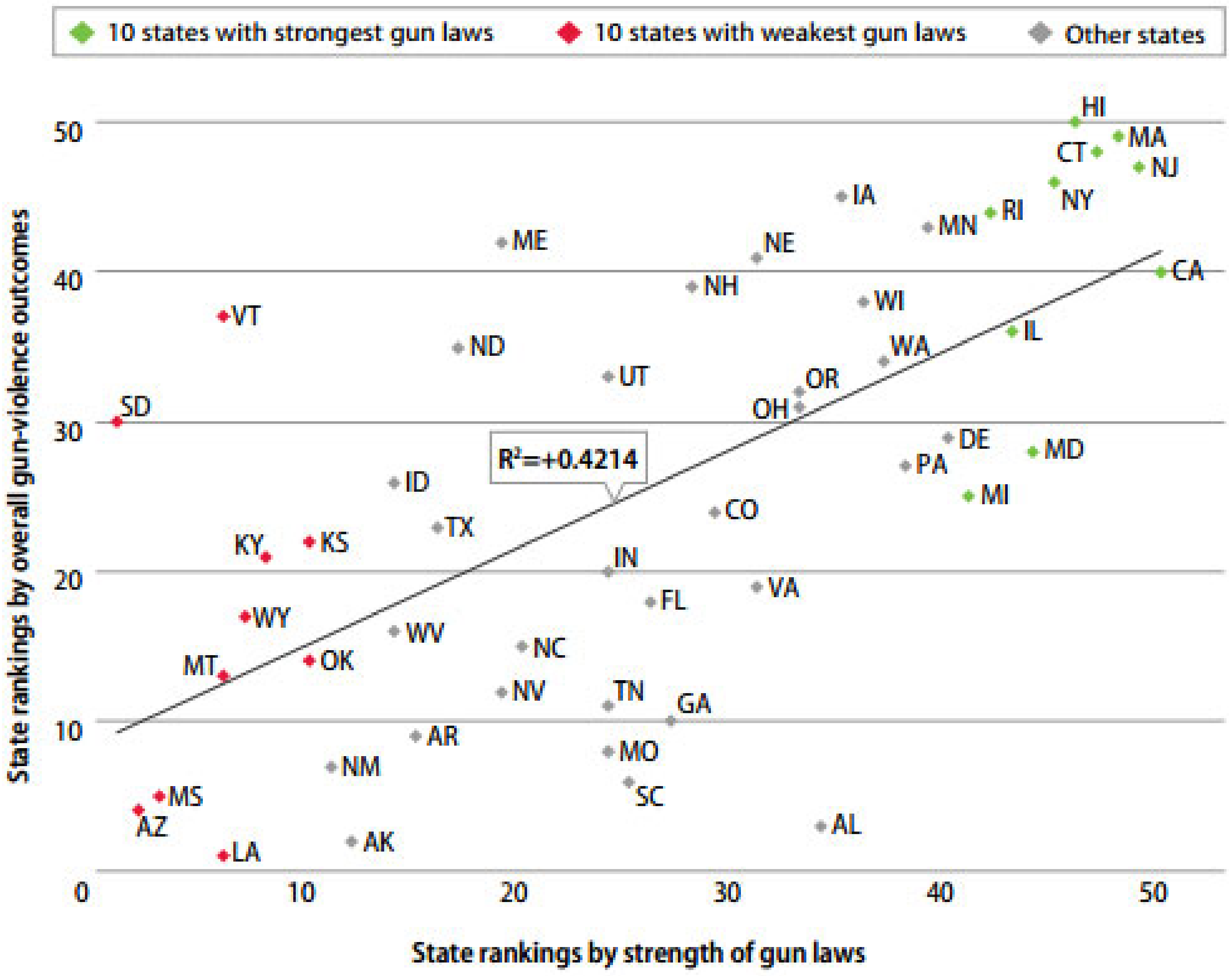
Does *correlation imply causality?*

Never conclude that correlation proves causality. Correlation is a measure of the strength of the relationship between two variables. However, just because two variables are related does not mean that one causes the other. ***Correlation does not imply causality***. However, given that caveat, there may be situations where causality is a reasonable or educated guess.

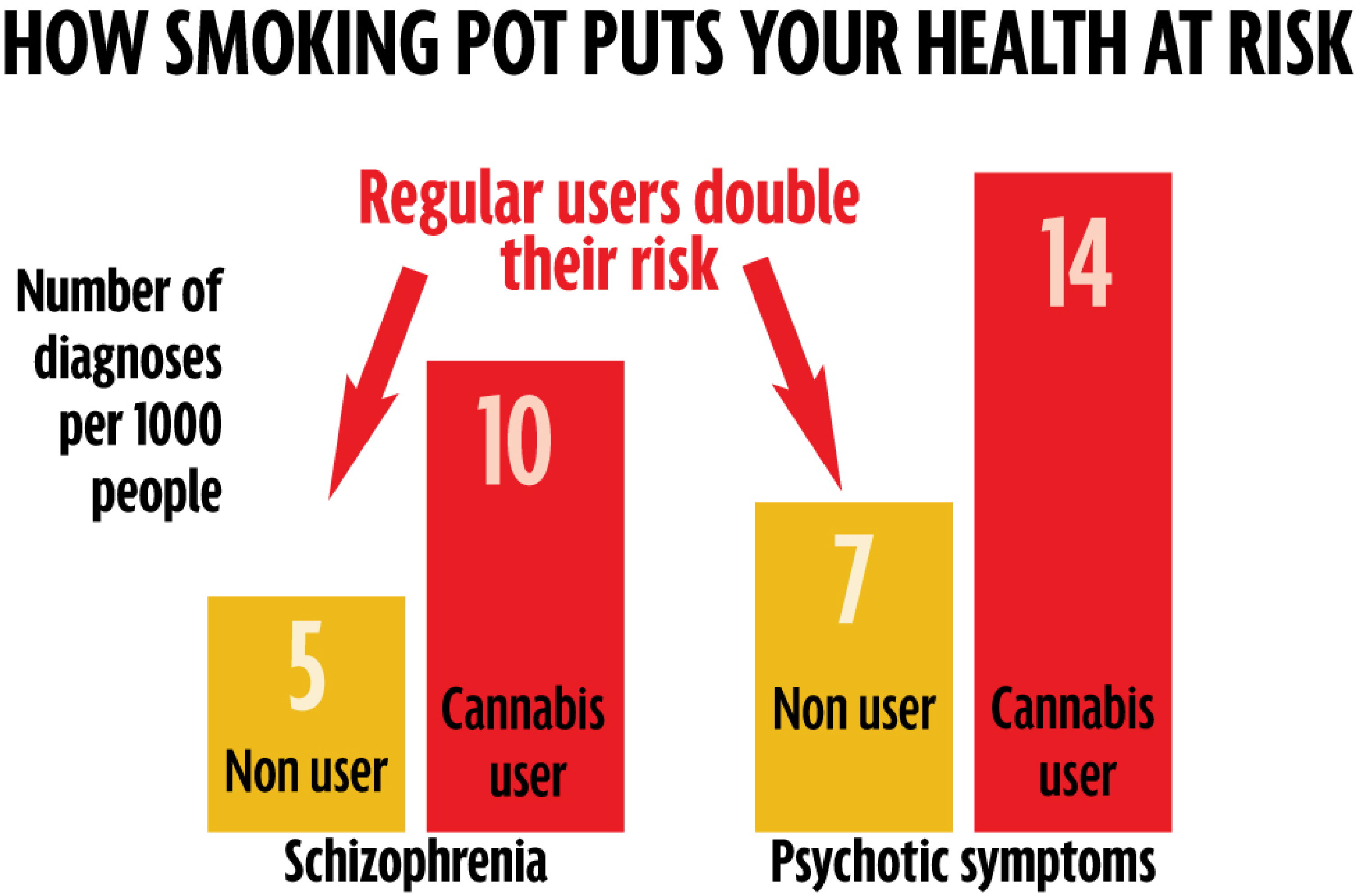
NOTE: When two variables have a correlation with one another, you will often hear people say that the variables are "linked." However, that does not mean that one variable is the cause of the other.

Do gun laws cause a reduction in gun violence?

Correlation between state gun laws and gun-violence outcomes



Does marijuana cause psychosis/schizophrenia?



In doing any statistical study, the following factors should be considered:

- *Are the results statistically significant?*

In doing any statistical study, the following factors should be considered:

- *Are the results statistically significant?*

Example: For now, think of statistical significance as meaning simply the occurrence of something that is unlikely to have been due to chance. In other words, we measure the statistical significance of an event in terms of whether it was a probable or an improbable outcome.

In doing any statistical study, the following factors should be considered:

- *Are the results statistically significant?*

Example: For instance, if you flip a “fair” coin 100 times and get heads each time, that is highly improbable. Thus, you would deem the result statistically significant, and you would question your assumption that it was a “fair” coin.

In doing any statistical study, the following factors should be considered:

- *Are the results statistically significant?*

Example: **A standard for significance that is often used in statistics is the .05 or 5% level of significance. In other words, if the likelihood of some event occurring by chance is .05 or less, then <u>we'll conclude that the result is statistically significant</u>.**

In doing any statistical study, the following factors should be considered:

- *Are the results statistically significant?*
- *Do the results have practical significance?*

In doing any statistical study, the following factors should be considered:

- *Are the results statistically significant?*
- *Do the results have practical significance?*

Example: When the number of records or subjects is large, even small changes have a tendency to be statistically significant. However, even if a change in the average math SAT score from, say, 550 to 551 turns out to be statistically significant, would you really consider that result to have practical significance?

In doing any statistical study, the following factors should be considered:

- *Are the results statistically significant?*
- *Do the results have practical significance?*

Example: On the other hand, when the number of records or subjects is small, even “large” changes may be statistically insignificant. Nonetheless, while a change in enrollment of 20 students at a small, private school may not be statistically significant, the difference in revenue that results may be practically significant.

What should we remember about *reported versus measured results?*

(1.2)

What should we remember about *reported versus measured results?*

Beware of ***reported*** *versus* ***measured results***. An example of “reported” data would be when you ask someone their weight instead of measuring it.

(1.2)

What should we remember about *small samples?*

What should we remember about *small samples?*

The smaller the size of the sample, the less likely it is to be representative of the population.

What should we remember about *loaded questions?*

(1.2)

What should we remember about *loaded questions?*

Loaded questions can influence the way a person responds.

97% yes: Should the President have the line item veto to eliminate waste?

57% yes: Should the President have the line item veto?

(1.2)

What should we remember about *question order?*

What should we remember about *question order?*

Be aware that even ***question order*** *can affect responses.* The response to the first question might condition your response to the second question.

1. “Should individuals in this country be free to do whatever they want as long as they are not hurting anybody?”

2. “Should people be allowed to carry guns openly on college campuses?”

(1.2)

What should we remember about *nonresponses?*

(1.2)

What should we remember about *nonresponses?*

*Be aware of how many **nonresponses** there are in the data.* For example, some people refused to respond to recent census questionnaires.

What should we remember about *missing data?*

What should we remember about *missing data?*

*Be aware of the impact of **missing data.*** For example, the recent census results may be affected by an inability to collect data on some homeless people, immigrants or low income minorities.

What should we remember about *precise numbers?*

What should we remember about *precise numbers?*

Be wary of ***precise numbers.*** Another problem that can occur in statistics with percents is displaying too many digits. For example, uncertainty often exists when it comes to real world measurements, and in statistics we are often dealing with "best estimates" of various quantities. Consequently, when there is some uncertainty in our data, but we display our results to several decimal places, we may be suggesting a higher level of accuracy than is actually present. Often we may round results to two or three decimal places.

What should we remember about *percentages?*

(1.2)

What should we remember about *percentages?*

Beware of incorrect uses of ***percentages***. Percent means "per one hundred." Keep that in mind. Also, people sometimes use percents in ways that might be misleading.

What should we remember about *percentages?*

*Beware of incorrect uses of **percentages**.* Percent means "per one hundred." Keep that in mind. Also, people sometimes use percents in ways that might be misleading.

In 2013, 7.5 percent of Americans, 12 or older, who were surveyed had used marijuana in the past month. By 2014, this figure had increased by 12 percent. What was the percentage of Americans, 12 or older, who had used marijuana in the past month in 2014?

(1.2)

What should we remember about *percentages?*

*Beware of incorrect uses of **percentages**.* Percent means "per one hundred." Keep that in mind. Also, people sometimes use percents in ways that might be misleading.

In 2013, 7.5 percent of Americans, 12 or older, who were surveyed had used marijuana in the past month. By 2014, this figure had increased by 12 percent. What was the percentage of Americans, 12 or older, who had used marijuana in the past month in 2014?

8.4%

What should we remember about *percentages?*

*Beware of incorrect uses of **percentages**.* Percent means "per one hundred." Keep that in mind. Also, people sometimes use percents in ways that might be misleading.

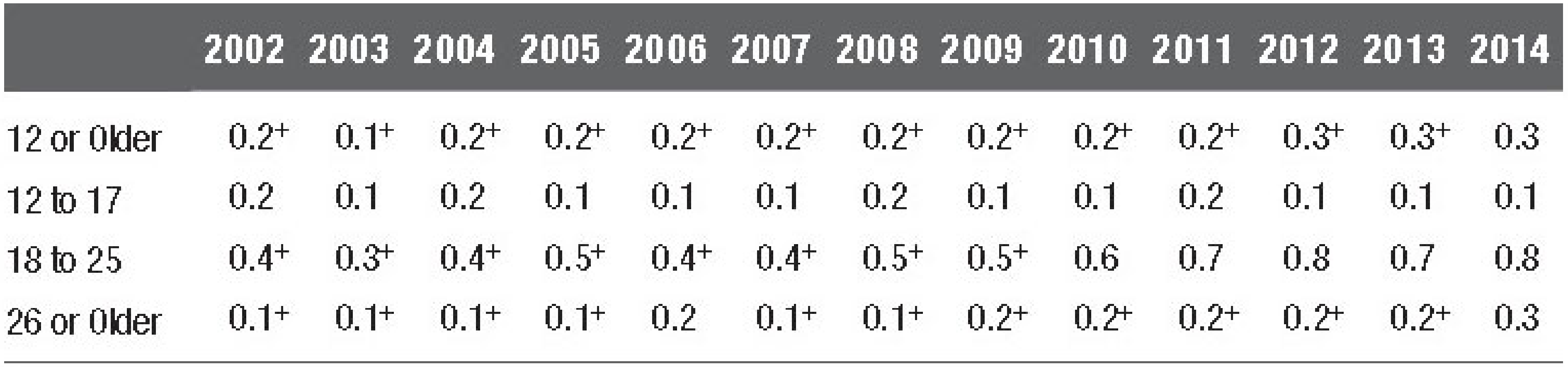
Also, people sometimes report percentages when they want to make the changes look more dramatic, and they will report numbers when they want to make changes appear less dramatic.

Example: Which headline below sounds more dramatic?

- **From 2002 to 2014, heroin use increased by 50%!**

- **From 2002 to 2014, heroin use increased from 0.2% to 0.3%.**

Figure 13 Table. Past Year Heroin Use among People Aged 12 or Older, by Age Group: Percentages, 2002-2014



	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
12 or Older	0.2$^+$	0.1$^+$	0.2$^+$	0.2$^+$	0.2$^+$	0.2$^+$	0.2$^+$	0.2$^+$	0.2$^+$	0.2$^+$	0.3$^+$	0.3$^+$	0.3
12 to 17	0.2	0.1	0.2	0.1	0.1	0.1	0.2	0.1	0.1	0.2	0.1	0.1	0.1
18 to 25	0.4$^+$	0.3$^+$	0.4$^+$	0.5$^+$	0.4$^+$	0.4$^+$	0.5$^+$	0.5$^+$	0.6	0.7	0.8	0.7	0.8
26 or Older	0.1$^+$	0.1$^+$	0.1$^+$	0.1$^+$	0.2	0.1$^+$	0.1$^+$	0.2$^+$	0.2$^+$	0.2$^+$	0.2$^+$	0.2$^+$	0.3

What is an *outlier?*

What is an *outlier?*

Outlier: An *outlier* is an unusual element of data, generally an observation that is numerically distant from the rest of the data.