Sampling to determine proportion of a population having some property

Sampling to determine proportion of a population having some property

So far we have focused on using the mean (and deviation) of a sample to extrapolate some information about the mean of an entire population.

Sampling to determine proportion of a population having some property

So far we have focused on using the mean (and deviation) of a sample to extrapolate some information about the mean of an entire population. These techniques are applicable in many settings, but (in case you didn't notice) they still don't let us how the confidence intervals of opinion polls work, for example.

Sampling to determine proportion of a population having some property

So far we have focused on using the mean (and deviation) of a sample to extrapolate some information about the mean of an entire population. These techniques are applicable in many settings, but (in case you didn't notice) they still don't let us how the confidence intervals of opinion polls work, for example. Fortunately, it doesn't take much reworking of our tools so far to address this setting.

Sampling to determine proportion of a population having some property

So far we have focused on using the mean (and deviation) of a sample to extrapolate some information about the mean of an entire population. These techniques are applicable in many settings, but (in case you didn't notice) they still don't let us how the confidence intervals of opinion polls work, for example. Fortunately, it doesn't take much reworking of our tools so far to address this setting.

So for example, we might be interested in the question: what proportion of the population is left-handed?

So for example, we might be interested in the question: what proportion of the population is left-handed?

- Take sample from class. Not truly random, but probably random enough for a question like this.

So for example, we might be interested in the question: what proportion of the population is left-handed?

- Take sample from class. Not truly random, but probably random enough for a question like this. Let $\hat{p}$ be the proportion of left-handed people.

So for example, we might be interested in the question: what proportion of the population is left-handed?

- Take sample from class. Not truly random, but probably random enough for a question like this. Let $\hat{p}$ be the proportion of left-handed people.

- How well does this approximate the proportion p in the general population?

Left-handedness is a categorical variable, with only two values (YES and NO).

Left-handedness is a categorical variable, with only two values (YES and NO). So it *can't* be normally distributed. What can a histogram look like?

Left-handedness is a categorical variable, with only two values (YES and NO). So it *can't* be normally distributed. What can a histogram look like?

But consider samples of size 10, for example. If we look at what *proportion* of a sample is left-handed, we have 11 possible values: $0, .1, .2, .3, .4, .5, .6, .7, .8, .9, 1$.

Left-handedness is a categorical variable, with only two values (YES and NO). So it *can't* be normally distributed. What can a histogram look like?

But consider samples of size 10, for example. If we look at what *proportion* of a sample is left-handed, we have 11 possible values: $0, .1, .2, .3, .4, .5, .6, .7, .8, .9, 1$.

As the size of the sample increases, the number of possible values for the proportion of left-handed people also increases. We have the following variant of the central limit theorem:

Left-handedness is a categorical variable, with only two values (YES and NO). So it *can't* be normally distributed. What can a histogram look like?

But consider samples of size 10, for example. If we look at what *proportion* of a sample is left-handed, we have 11 possible values: $0, .1, .2, .3, .4, .5, .6, .7, .8, .9, 1$.

As the size of the sample increases, the number of possible values for the proportion of left-handed people also increases. We have the following variant of the central limit theorem:

Theorem 1. *Let X be some random variable of a large population which has values YES and NO. Take SRS of size n from our population, and let $\hat{p}$ be the proportion of the sample which is "YES."*

Theorem 1. *Let X be some random variable of a large population which has values YES and NO. Take SRS of size n from our population, and let $\hat{p}$ be the proportion of the sample which is "YES."*

- *For large n, the sampling distribution of $\hat{p}$ is approximately normal; $N(p, \sigma)$ where*

Theorem 1. *Let X be some random variable of a large population which has values YES and NO. Take SRS of size n from our population, and let $\hat{p}$ be the proportion of the sample which is "YES."*

- *For large n, the sampling distribution of $\hat{p}$ is approximately normal; $N(p, \sigma)$ where*

- *p is the proportion of the entire population which is "YES" and*

Theorem 1. *Let X be some random variable of a large population which has values YES and NO. Take SRS of size n from our population, and let $\hat{p}$ be the proportion of the sample which is "YES."*

- *For large n, the sampling distribution of $\hat{p}$ is approximately normal; $N(p, \sigma)$ where*

- *p is the proportion of the entire population which is "YES" and*

$$\sigma = \sqrt{\frac{p(1-p)}{n}}.$$

$$\sigma = \sqrt{\frac{p(1-p)}{n}}.$$

Example 2. *Suppose that two-thirds of college students have cheated on an exam.*

$$\sigma = \sqrt{\frac{p(1-p)}{n}}.$$

Example 2. *Suppose that two-thirds of college students have cheated on an exam. What is the probability that in a random sample (taken discretely) of 20 students, 15 or more would have cheated?*

$$\sigma = \sqrt{\frac{p(1-p)}{n}}.$$

Example 2. *Suppose that two-thirds of college students have cheated on an exam. What is the probability that in a random sample (taken discretely) of 20 students, 15 or more would have cheated? What is the probability that 10 or more have cheated?*

Getting back to statistical inference, we would like to do inference aimed at estimating p from $\hat{p}$.

Getting back to statistical inference, we would like to do inference aimed at estimating p from $\hat{p}$. The normalized z-statistic, which is behind the scenes of both confidence intervals and hypothesis testing, would be $z = \frac{\hat{p}-p}{\sigma}$

Getting back to statistical inference, we would like to do inference aimed at estimating p from $\hat{p}$. The normalized z-statistic, which is behind the scenes of both confidence intervals and hypothesis testing, would be $z = \frac{\hat{p}-p}{\sigma}$ where $\sigma = \sqrt{p(1-p)/n}$ as in the theorem.

Getting back to statistical inference, we would like to do inference aimed at estimating p from $\hat{p}$. The normalized z-statistic, which is behind the scenes of both confidence intervals and hypothesis testing, would be $z = \frac{\hat{p}-p}{\sigma}$ where $\sigma = \sqrt{p(1-p)/n}$ as in the theorem. If n is large, then $\hat{p}$ was approximately normal. Thus z will be approximately *standard* normal.

Getting back to statistical inference, we would like to do inference aimed at estimating p from $\hat{p}$. The normalized z-statistic, which is behind the scenes of both confidence intervals and hypothesis testing, would be $z = \frac{\hat{p}-p}{\sigma}$ where $\sigma = \sqrt{p(1-p)/n}$ as in the theorem. If n is large, then $\hat{p}$ was approximately normal. Thus z will be approximately *standard* normal.

In practice, we won't know p.

Getting back to statistical inference, we would like to do inference aimed at estimating p from $\hat{p}$. The normalized z-statistic, which is behind the scenes of both confidence intervals and hypothesis testing, would be $z = \frac{\hat{p}-p}{\sigma}$ where $\sigma = \sqrt{p(1-p)/n}$ as in the theorem. If n is large, then $\hat{p}$ was approximately normal. Thus z will be approximately *standard* normal.

In practice, we won't know p. We use $\hat{p}$ in place of p to get the standard error in place of the standard deviation.

Getting back to statistical inference, we would like to do inference aimed at estimating p from $\hat{p}$. The normalized z-statistic, which is behind the scenes of both confidence intervals and hypothesis testing, would be $z = \frac{\hat{p}-p}{\sigma}$ where $\sigma = \sqrt{p(1-p)/n}$ as in the theorem. If n is large, then $\hat{p}$ was approximately normal. Thus z will be approximately *standard* normal.

In practice, we won't know p. We use $\hat{p}$ in place of p to get the standard error in place of the standard deviation. So we set $s = \sqrt{\hat{p}(1-\hat{p})/n}$, and then $z = \frac{\hat{p}-p}{s}$.

Getting back to statistical inference, we would like to do inference aimed at estimating p from $\hat{p}$. The normalized z-statistic, which is behind the scenes of both confidence intervals and hypothesis testing, would be $z = \frac{\hat{p}-p}{\sigma}$ where $\sigma = \sqrt{p(1-p)/n}$ as in the theorem. If n is large, then $\hat{p}$ was approximately normal. Thus z will be approximately *standard* normal.

In practice, we won't know p. We use $\hat{p}$ in place of p to get the standard error in place of the standard deviation. So we set $s = \sqrt{\hat{p}(1-\hat{p})/n}$, and then $z = \frac{\hat{p}-p}{s}$. To get a confidence interval with certainty $C\%$, we choose z^* a

critical value for C, and then with confidence $C\%$ we know p is between $\hat{p} - z^* \times s$ and $\hat{p} + z^* \times s$.

critical value for C, and then with confidence $C\%$ we know p is between $\hat{p} - z^* \times s$ and $\hat{p} + z^* \times s$.

Example 3. *Use an in-class survey to estimate the percentage of left-handers with 90 and 95 percent confidence.*

critical value for C, and then with confidence $C\%$ we know p is between $\hat{p} - z^* \times s$ and $\hat{p} + z^* \times s$.

Example 3. *Use an in-class survey to estimate the percentage of left-handers with 90 and 95 percent confidence.*

To do inference, we need to know we are reasonably close to a normal distribution. Here are some conditions:

To do inference, we need to know we are reasonably close to a normal distribution. Here are some conditions:

- Our sample is a SRS.

To do inference, we need to know we are reasonably close to a normal distribution. Here are some conditions:

- Our sample is a SRS.
- The population is at least 10 times the sample size.

To do inference, we need to know we are reasonably close to a normal distribution. Here are some conditions:

- Our sample is a SRS.
- The population is at least 10 times the sample size.
- The sample size is "large enough." (At least 15 successes and 15 failures.)

To do inference, we need to know we are reasonably close to a normal distribution. Here are some conditions:

- Our sample is a SRS.
- The population is at least 10 times the sample size.
- The sample size is "large enough." (At least 15 successes and 15 failures.)

Unfortunately, even for relatively large n, this can be not so close to Normal. Fix (recommended to always use): the "Plus four" confidence interval.

To do inference, we need to know we are reasonably close to a normal distribution. Here are some conditions:

- Our sample is a SRS.
- The population is at least 10 times the sample size.
- The sample size is "large enough." (At least 15 successes and 15 failures.)

Unfortunately, even for relatively large n, this can be not so close to Normal. Fix (recommended to always use): the "Plus four" confidence interval. Let $\overline{p} = \frac{\text{successes}+2}{n+4}$.

To do inference, we need to know we are reasonably close to a normal distribution. Here are some conditions:

- Our sample is a SRS.
- The population is at least 10 times the sample size.
- The sample size is "large enough." (At least 15 successes and 15 failures.)

Unfortunately, even for relatively large n, this can be not so close to Normal. Fix (recommended to always use): the "Plus four" confidence interval. Let $\overline{p} = \frac{\text{successes}+2}{n+4}$.

Then the C% confidence interval is between $\overline{p} - z^* \sqrt{\frac{\overline{p}(1-\overline{p})}{n+4}}$ and $\overline{p} + z^* \sqrt{\frac{\overline{p}(1-\overline{p})}{n+4}}$.

Then the C% confidence interval is between $\overline{p} - z^* \sqrt{\frac{\overline{p}(1-\overline{p})}{n+4}}$ and $\overline{p} + z^* \sqrt{\frac{\overline{p}(1-\overline{p})}{n+4}}$.

Example 4. *Redo our estimate for left-handers using the "plus four" confidence interval.*

Then the C% confidence interval is between $\overline{p} - z^* \sqrt{\frac{\overline{p}(1-\overline{p})}{n+4}}$ and $\overline{p} + z^* \sqrt{\frac{\overline{p}(1-\overline{p})}{n+4}}$.

Example 4. *Redo our estimate for left-handers using the "plus four" confidence interval.*

Example 5. *Establish some confidence intervals (both the usual and plus four) for polls found at:*
`http://www.usatoday.com/news/polls/tables/live/2`

Example 6. *Find some polls on the web which publish their sample size and margin of error, and determine with what certainty the number being measured is within that margin or error.*

Hypothesis testing for population proportions

Hypothesis testing for population proportions

Example: Opinions about healthcare

Hypothesis testing for population proportions

Example: Opinions about healthcare

A large poll conducted in 2001 showed that 39% of Americans favored universal health care. Recently a poll of 400 people found that 50% favored universal health care.

Hypothesis testing for population proportions

Example: Opinions about healthcare

A large poll conducted in 2001 showed that 39% of Americans favored universal health care. Recently a poll of 400 people found that 50% favored universal health care. We'd like to determine, based on this, if more people now favor universal health care.

1. Let p be the (true) proportion of people i favor universal health care.

1. Let p be the (true) proportion of people i favor universal health care. (We'll assume the 2001 poll is accurate so that the proportion of people then favoring universal health care is $p_0 = 39\%$).

1. Let p be the (true) proportion of people i favor universal health care. (We'll assume the 2001 poll is accurate so that the proportion of people then favoring universal health care is $p_0 = 39\%$).

   So H_0 is

1. Let p be the (true) proportion of people i favor universal health care. (We'll assume the 2001 poll is accurate so that the proportion of people then favoring universal health care is $p_0 = 39\%$).

   So H_0 is "$p = p_0$",

1. Let p be the (true) proportion of people i favor universal health care. (We'll assume the 2001 poll is accurate so that the proportion of people then favoring universal health care is $p_0 = 39\%$).

   So H_0 is "$p = p_0$", and H_a is

1. Let p be the (true) proportion of people i favor universal health care. (We'll assume the 2001 poll is accurate so that the proportion of people then favoring universal health care is $p_0 = 39\%$).

   So H_0 is "$p = p_0$", and H_a is "$p > p_0$".

1. Let p be the (true) proportion of people i favor universal health care. (We'll assume the 2001 poll is accurate so that the proportion of people then favoring universal health care is $p_0 = 39\%$).

   So H_0 is "$p = p_0$", and H_a is "$p > p_0$". In English: Our "null hypothesis" is that 39% of people now support universal health care.

1. Let p be the (true) proportion of people i favor universal health care. (We'll assume the 2001 poll is accurate so that the proportion of people then favoring universal health care is $p_0 = 39\%$).

   So H_0 is "$p = p_0$", and H_a is "$p > p_0$". In English: Our "null hypothesis" is that 39% of people now support universal health care. Our "test" is to determine if it is likely that *more* than 39% of people now support universal health care.

2. We pretend that H_0 is true (that is that $p = p_0$), and we calculate our z-statistic based on that:

2. We pretend that H_0 is true (that is that $p = p_0$), and we calculate our z-statistic based on that:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} =$$

2. We pretend that H_0 is true (that is that $p = p_0$), and we calculate our z-statistic based on that:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{.5 - .39}{\sqrt{\frac{.39(1-.39)}{400}}} = \frac{.11}{.0243875} = 4.51.$$

2. We pretend that H_0 is true (that is that $p = p_0$), and we calculate our z-statistic based on that:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{.5 - .39}{\sqrt{\frac{.39(1-.39)}{400}}} = \frac{.11}{.0243875} = 4.51.$$

3. We calculate our P-value.

2. We pretend that H_0 is true (that is that $p = p_0$), and we calculate our z-statistic based on that:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{.5 - .39}{\sqrt{\frac{.39(1-.39)}{400}}} = \frac{.11}{.0243875} = 4.51.$$

3. We calculate our P-value.

$$P(Z \geq 4.51) = .00000324.$$

2. We pretend that H_0 is true (that is that $p = p_0$), and we calculate our z-statistic based on that:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{.5 - .39}{\sqrt{\frac{.39(1-.39)}{400}}} = \frac{.11}{.0243875} = 4.51.$$

3. We calculate our P-value.

$$P(Z \geq 4.51) = .00000324.$$

This P-value is:

2. We pretend that H_0 is true (that is that $p = p_0$), and we calculate our z-statistic based on that:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{.5 - .39}{\sqrt{\frac{.39(1-.39)}{400}}} = \frac{.11}{.0243875} = 4.51.$$

3. We calculate our P-value.

$$P(Z \geq 4.51) = .00000324.$$

This P-value is: the probability of picking a sample of 400 so that at least 200 of them favor universal health

care if in the population only 39% favor it.

care if in the population only 39% favor it.

4. Conclusion: If only 39% of people now support universal health care, then the probability of getting 50% in favor of universal health care by taking a random sample of 400 people would be only $.00000324$.

care if in the population only 39% favor it.

4. Conclusion: If only 39% of people now support universal health care, then the probability of getting 50% in favor of universal health care by taking a random sample of 400 people would be only $.00000324$.

   This is strong evidence that H_0 is false, and that more than 39% of people now support universal health care.

care if in the population only 39% favor it.

4. Conclusion: If only 39% of people now support universal health care, then the probability of getting 50% in favor of universal health care by taking a random sample of 400 people would be only $.00000324$.

   This is strong evidence that H_0 is false, and that more than 39% of people now support universal health care. So we reject the null hypothesis at almost any sensible significance level.

Notes on applying z-test for proportions, and more examples

Notes on applying z-test for proportions, and more examples

In order for these confidence intervals and hypothesis tests to be valid...

Notes on applying z-test for proportions, and more examples

In order for these confidence intervals and hypothesis tests to be valid...

1. We need our sample to be large enough so that both np_0 and $n(1-p_0)$ are larger than 10.

Notes on applying z-test for proportions, and more examples

In order for these confidence intervals and hypothesis tests to be valid...

1. We need our sample to be large enough so that both np_0 and $n(1-p_0)$ are larger than 10.

2. What else do we need to know about our sample?

Notes on applying z-test for proportions, and more examples

In order for these confidence intervals and hypothesis tests to be valid...

1. We need our sample to be large enough so that both np_0 and $n(1-p_0)$ are larger than 10.

2. What else do we need to know about our sample? It needs to be an SRS. This is hard, and is usually the biggest reason to be skeptical about polls.

Notes on applying z-test for proportions, and more examples

In order for these confidence intervals and hypothesis tests to be valid...

1. We need our sample to be large enough so that both np_0 and $n(1-p_0)$ are larger than 10.

2. What else do we need to know about our sample? It needs to be an SRS. This is hard, and is usually the biggest reason to be skeptical about polls.

Example 7. *Suppose we take a SRS of 90 people under the age of 22, and discover 11 of them have at some point had an STI. Use this to estimate the proportion of the total population which have had an STI?*

Example 7. *Suppose we take a SRS of 90 people under the age of 22, and discover 11 of them have at some point had an STI. Use this to estimate the proportion of the total population which have had an STI? (Note: half of the top ten most frequently reported diseases in the U.S., accounting for over four-fifths of all cases, are STI's).*

Example 7. *Suppose we take a SRS of 90 people under the age of 22, and discover 11 of them have at some point had an STI. Use this to estimate the proportion of the total population which have had an STI? (Note: half of the top ten most frequently reported diseases in the U.S., accounting for over four-fifths of all cases, are STI's).*

Example 8. *In a certain population, we expect 46% of the population to get a cold in a 3 month period.*

Example 7. *Suppose we take a SRS of 90 people under the age of 22, and discover 11 of them have at some point had an STI. Use this to estimate the proportion of the total population which have had an STI? (Note: half of the top ten most frequently reported diseases in the U.S., accounting for over four-fifths of all cases, are STI's).*

Example 8. *In a certain population, we expect 46% of the population to get a cold in a 3 month period. We give 264 volunteers 1000 mgs of vitamin C per day for 3 months. At the end of the period 119 people (45%) have gotten colds. So the proportion of our sample that have*

gotten colds is $\hat{p} = .45$.

gotten colds is $\hat{p} = .45$.

Let p *be the proportion of people being treated with vitamin C who will get colds. We would like to know if there is evidence that* p *is smaller than* $.46$.

gotten colds is $\hat{p} = .45$.

Let p *be the proportion of people being treated with vitamin C who will get colds. We would like to know if there is evidence that* p *is smaller than* $.46$.