Two-sample statistics

Two-sample statistics

In the last class we started to look at two-sample problems.

Two-sample statistics

In the last class we started to look at two-sample problems. Here we have two populations, and wish to understand how some variable differs from one population to another.

Two-sample statistics

In the last class we started to look at two-sample problems. Here we have two populations, and wish to understand how some variable differs from one population to another.

These sort of questions arise in a wide variety of situation when you wish to

Two-sample statistics

In the last class we started to look at two-sample problems. Here we have two populations, and wish to understand how some variable differs from one population to another.

These sort of questions arise in a wide variety of situation when you wish to

- Compare two different treatments. For example, two different manufacturing processes, A and B, or two different medical treatments.

Two-sample statistics

In the last class we started to look at two-sample problems. Here we have two populations, and wish to understand how some variable differs from one population to another.

These sort of questions arise in a wide variety of situation when you wish to

- Compare two different treatments. For example, two different manufacturing processes, A and B, or two different medical treatments. Here your two populations

are those manufactured with process A, and those manufactured with process B.

are those manufactured with process A, and those manufactured with process B.

- Compare some variable of two different populations, such as men and women, or left-handed people and right-handed people.

are those manufactured with process A, and those manufactured with process B.

- Compare some variable of two different populations, such as men and women, or left-handed people and right-handed people.

The methods used are much like for single sample confidence intervals and hypothesis tests.

are those manufactured with process A, and those manufactured with process B.

- Compare some variable of two different populations, such as men and women, or left-handed people and right-handed people.

The methods used are much like for single sample confidence intervals and hypothesis tests. We gave some explicit formulae last time. This time we will first "just work things through" in testing a hypothesis and then give a step-by-step treatment of confidence intervals.

Example of hypothesis testing: exercise 17.38

Example of hypothesis testing: exercise 17.38

This exercise gives some IQ data for some boys and girls from the same midwestern school district and asks if there is a statistically significant difference between the means.

Example of hypothesis testing: exercise 17.38

This exercise gives some IQ data for some boys and girls from the same midwestern school district and asks if there is a statistically significant difference between the means. After keying some numbers into a calculator, we get the following information for our two samples:

Example of hypothesis testing: exercise 17.38

This exercise gives some IQ data for some boys and girls from the same midwestern school district and asks if there is a statistically significant difference between the means. After keying some numbers into a calculator, we get the following information for our two samples:

Population	Mean	Sample Size	Sample mean	Sample s.d
Girls	μ_1	31	$\overline{x}_1 = 105.84$	$s_1 = 14.27$
Boys	μ_2	47	$\overline{x}_2 = 110.96$	$s_2 = 12.12$

1. Our null hypothesis is that boys' IQ scores are the same as girls' IQ scores. That is

$$H_0 : \mu_1 = \mu_2.$$

1. Our null hypothesis is that boys' IQ scores are the same as girls' IQ scores. That is

$$H_0 : \mu_1 = \mu_2.$$

Our alternative hypothesis is that boys have higher IQ scores.

$$H_a : \mu_1 < \mu_2 \text{ or } \mu_1 - \mu_2 < 0.$$

1. Our null hypothesis is that boys' IQ scores are the same as girls' IQ scores. That is

$$H_0 : \mu_1 = \mu_2.$$

Our alternative hypothesis is that boys have higher IQ scores.

$$H_a : \mu_1 < \mu_2 \text{ or } \mu_1 - \mu_2 < 0.$$

We wish to test this using our data.

2. We calculate our two-sample t-statistic

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{105.84 - 110.96}{\sqrt{6.569 + 3.125}} = \frac{-5.12}{3.114} = -1.644.$$

2. We calculate our two-sample t-statistic

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{105.84 - 110.96}{\sqrt{6.569 + 3.125}} = \frac{-5.12}{3.114} = -1.644.$$

3. We calculate our P-value. Since H_a is $\mu_1 - \mu_2 < 0$, we wish to look for $P(t \leq -1.644)$. We use $t(30)$ since 31 is our smaller sample size.

2. We calculate our two-sample t-statistic

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{105.84 - 110.96}{\sqrt{6.569 + 3.125}} = \frac{-5.12}{3.114} = -1.644.$$

3. We calculate our P-value. Since H_a is $\mu_1 - \mu_2 < 0$, we wish to look for $P(t \leq -1.644)$. We use $t(30)$ since 31 is our smaller sample size. (see p. 452 for a more accurate way to determine degrees of freedom).

   From the calculator, $P(t \leq -1.644) = .0553$.

4. We draw our conclusion: If we assume H_0 is *true*, then the probability of seeing samples like the ones we have is $.0553$.

4. We draw our conclusion: If we assume H_0 is *true*, then the probability of seeing samples like the ones we have is $.0553$. This is moderately low, so our assumption that H_0 was true is probably wrong. So, this is moderate evidence that boys score higher on IQ tests than girls.

4. We draw our conclusion: If we assume H_0 is *true*, then the probability of seeing samples like the ones we have is $.0553$. This is moderately low, so our assumption that H_0 was true is probably wrong. So, this is moderate evidence that boys score higher on IQ tests than girls. Which is in turn evidence that small differences in tests such as IQ tests do not accurately reflect much of anything.

4. We draw our conclusion: If we assume H_0 is *true*, then the probability of seeing samples like the ones we have is $.0553$. This is moderately low, so our assumption that H_0 was true is probably wrong. So, this is moderate evidence that boys score higher on IQ tests than girls. Which is in turn evidence that small differences in tests such as IQ tests do not accurately reflect much of anything.

5. We can ask the calculator to *do* the test for us. This is under STAT, TESTS, 4:2-SampTTest. We get $df = 56.93$, $t = -1.64$, $P = .053$.

We *still* need to do step 4 (conclusion) above. And we need to do it carefully, because we've possibly lost track of what all our numbers mean.

Recap of confidence intervals for difference of means of two populations

Recap of confidence intervals for difference of means of two populations

In order to use t-statistics to study two-sample problems, we need the following conditions to be satisfied, just as for single-sample problems.

Recap of confidence intervals for difference of means of two populations

In order to use t-statistics to study two-sample problems, we need the following conditions to be satisfied, just as for single-sample problems.

- An independent SRS from each populations. For example if we were trying to sample men and women, it won't work to take a random sample of men, and then to take their wives or girlfriends as the other sample.

Recap of confidence intervals for difference of means of two populations

In order to use t-statistics to study two-sample problems, we need the following conditions to be satisfied, just as for single-sample problems.

- An independent SRS from each populations. For example if we were trying to sample men and women, it won't work to take a random sample of men, and then to take their wives or girlfriends as the other sample.
- Both populations need to be normally distributed, or...

Recap of confidence intervals for difference of means of two populations

In order to use t-statistics to study two-sample problems, we need the following conditions to be satisfied, just as for single-sample problems.

- An independent SRS from each populations. For example if we were trying to sample men and women, it won't work to take a random sample of men, and then to take their wives or girlfriends as the other sample.
- Both populations need to be normally distributed, or...

- If distributions aren't close to normal but no outliers and no strong skewedness, need sample sizes over 15.

- If distributions aren't close to normal but no outliers and no strong skewedness, need sample sizes over 15.
- Generally sample sizes greater than 40 are OK even with strongly skewed distributions or outliers.

- If distributions aren't close to normal but no outliers and no strong skewedness, need sample sizes over 15.
- Generally sample sizes greater than 40 are OK even with strongly skewed distributions or outliers.

If these conditions hold then we can follow these steps to compare these different populations, finding a confidence interval for the difference of means $\mu_1 - \mu_2$.

- If distributions aren't close to normal but no outliers and no strong skewedness, need sample sizes over 15.
- Generally sample sizes greater than 40 are OK even with strongly skewed distributions or outliers.

If these conditions hold then we can follow these steps to compare these different populations, finding a confidence interval for the difference of means $\mu_1 - \mu_2$.

- Compute the standard error of the two samples, $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$.

- If distributions aren't close to normal but no outliers and no strong skewedness, need sample sizes over 15.
- Generally sample sizes greater than 40 are OK even with strongly skewed distributions or outliers.

If these conditions hold then we can follow these steps to compare these different populations, finding a confidence interval for the difference of means $\mu_1 - \mu_2$.

- Compute the standard error of the two samples, $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$.

- Use the $t(k)$ distribution where k is one less than the smaller of the two sample sizes.

- Use the $t(k)$ distribution where k is one less than the smaller of the two sample sizes.

- Find t^* as in the one-sample case so that C% of the area is between $-t^*$ and t^*.

- Use the $t(k)$ distribution where k is one less than the smaller of the two sample sizes.

- Find t^* as in the one-sample case so that C% of the area is between $-t^*$ and t^*. We can look this up in Table C.

- Use the $t(k)$ distribution where k is one less than the smaller of the two sample sizes.

- Find t^* as in the one-sample case so that C% of the area is between $-t^*$ and t^*. We can look this up in Table C.

- With C% confidence,

- Use the $t(k)$ distribution where k is one less than the smaller of the two sample sizes.

- Find t^* as in the one-sample case so that C% of the area is between $-t^*$ and t^*. We can look this up in Table C.

- With C% confidence, we can say the true difference of means is between $(\overline{x_1} - \overline{x_2}) - t^* SE$ and $(\overline{x_1} - \overline{x_2}) + t^* SE$.

Example 1. *Mean body temperatures:*

Example 1. *Mean body temperatures: In one study, 65 men and 65 women have their temperature taken (in similar conditions).*

Example 1. *Mean body temperatures: In one study, 65 men and 65 women have their temperature taken (in similar conditions). The male mean is 98.105 with a standard deviation of 0.699.*

Example 1. *Mean body temperatures: In one study, 65 men and 65 women have their temperature taken (in similar conditions). The male mean is 98.105 with a standard deviation of 0.699. The female mean is 98.394, with a standard deviation of 0.743.*

Example 1. *Mean body temperatures: In one study, 65 men and 65 women have their temperature taken (in similar conditions). The male mean is 98.105 with a standard deviation of 0.699. The female mean is 98.394, with a standard deviation of 0.743. Give a 95% confidence interval for the difference between these means and test the hypothesis that women have higher temperatures than men at the 0.05 level.*

Example 1. *Mean body temperatures: In one study, 65 men and 65 women have their temperature taken (in similar conditions). The male mean is 98.105 with a standard deviation of 0.699. The female mean is 98.394, with a standard deviation of 0.743. Give a 95% confidence interval for the difference between these means and test the hypothesis that women have higher temperatures than men at the 0.05 level. What if the data were drawn from samples of only 20 men and 22 women?*

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. 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.

- But how can we know how well this approximates proportion p from 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.

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:

Theorem 2. *Let X be some random variable of a large population which has values YES and NO.*

Theorem 2. *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 3. *Suppose that two-thirds of college students have cheated on an exam.*

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

Example 3. *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 3. *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?*

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

Example 3. *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 $\hat{p}$. 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.

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 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 $\hat{p} + z^* \times s$.

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

Example 4. *Use an in-class survey to estimate the percentage of left-handers with 90 and 95 percent confidence.*

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

Example 4. *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.)

Unfortunately, even for relatively large n, this can be not so close to Normal.

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

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 $\overline{p} = \frac{\text{successes}+2}{n+4}$.

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 $\overline{p} + z^* \sqrt{\frac{\overline{p}(1-\overline{p})}{n+4}}$.

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 $\overline{p} + z^* \sqrt{\frac{\overline{p}(1-\overline{p})}{n+4}}$.

Example 5. *Redo our estimate for left-handers using the "plus four" confidence interval.*

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 $\overline{p} + z^* \sqrt{\frac{\overline{p}(1-\overline{p})}{n+4}}$.

Example 5. *Redo our estimate for left-handers using the "plus four" confidence interval.*

Example 6. *Establish some confidence intervals (both the usual and plus four) for polls found at:*

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 $\overline{p} + z^* \sqrt{\frac{\overline{p}(1-\overline{p})}{n+4}}$.

Example 5. *Redo our estimate for left-handers using the "plus four" confidence interval.*

Example 6. *Establish some confidence intervals (both the usual and plus four) for polls found at:* `http://www.usatoday.com/news/polls/tables/live/2`