The logic behind testing hypotheses

We toss a coin 10 times and get 7 tails. Is this sufficient evidence to conclude that the coin is biased?

The logic behind testing hypotheses

We toss a coin 10 times and get 7 tails. Is this sufficient evidence to conclude that the coin is biased?

The **null hypothesis, $\mathbf{H}_0$**, states that "nothing extraordinary is going on". So in this case

$$H_0\text{: } P(T) = \frac{1}{2}$$

The logic behind testing hypotheses

We toss a coin 10 times and get 7 tails. Is this sufficient evidence to conclude that the coin is biased?

The **null hypothesis, $\mathbf{H}_0$**, states that "nothing extraordinary is going on". So in this case

$$\mathrm{H}_0\text{:}\ \ \mathrm{P(T)} = \tfrac{1}{2}$$

The **alternative hypothesis, $\mathbf{H}_A$**, states that there is a different chance process that generates the data. Here we can take

$$\mathrm{H}_A\text{:}\ \ \mathrm{P(T)} \neq \tfrac{1}{2}$$

The logic behind testing hypotheses

We toss a coin 10 times and get 7 tails. Is this sufficient evidence to conclude that the coin is biased?

The **null hypothesis, $\mathbf{H}_0$**, states that "nothing extraordinary is going on". So in this case

$$\mathrm{H}_0\text{: } \mathrm{P(T)} = \frac{1}{2}$$

The **alternative hypothesis, $\mathbf{H}_A$**, states that there is a different chance process that generates the data. Here we can take

$$\mathrm{H}_A\text{: } \mathrm{P(T)} \neq \frac{1}{2}$$

Hypothesis testing proceeds by collecting data and evaluating whether the data are compatible with H_0 or not (in which case one **rejects $\mathbf{H}_0$**).

The logic behind testing hypotheses

A different example: A company develops a new drug to lower blood pressure. It tests it with an experiment involving 1,000 patients.

The logic behind testing hypotheses

A different example: A company develops a new drug to lower blood pressure. It tests it with an experiment involving 1,000 patients.
In this case "nothing extraordinary going on" means that the drug has no effect.

The logic behind testing hypotheses

A different example: A company develops a new drug to lower blood pressure. It tests it with an experiment involving 1,000 patients.

In this case "nothing extraordinary going on" means that the drug has no effect. So

H_0: no change in blood pressure $\qquad$ H_A: blood pressure drops

The logic behind testing hypotheses

A different example: A company develops a new drug to lower blood pressure. It tests it with an experiment involving 1,000 patients.
In this case "nothing extraordinary going on" means that the drug has no effect. So

H_0: no change in blood pressure $\qquad$ H_A: blood pressure drops

Note that in this case the company would like to reject H_0!

The logic behind testing hypotheses

A different example: A company develops a new drug to lower blood pressure. It tests it with an experiment involving 1,000 patients.
In this case "nothing extraordinary going on" means that the drug has no effect. So

H_0: no change in blood pressure $\quad$ H_A: blood pressure drops

Note that in this case the company would like to reject H_0!

So the logic of testing is typically indirect: One assumes that nothing extraordinary is happening and then hopes to reject this assumption H_0.

Setting up a test statistic

A **test statistic** measures how far away the data are from what we would expect if H_0 were true.

Setting up a test statistic

A **test statistic** measures how far away the data are from what we would expect if H_0 were true.

The most common test statistic is the **z-statistic**:

$$z = \frac{\text{observed} - \text{expected}}{\text{SE}}$$

Setting up a test statistic

A **test statistic** measures how far away the data are from what we would expect if H_0 were true.

The most common test statistic is the **z-statistic**:

$$z = \frac{\text{observed} - \text{expected}}{\text{SE}}$$

'Observed' is a statistic that is appropriate for assessing H_0. In the example of the 10 coin tosses, appropriate statistics would be the number of tails or the percent of tails.

Setting up a test statistic

A **test statistic** measures how far away the data are from what we would expect if H_0 were true.

The most common test statistic is the **z-statistic**:

$$z = \frac{\text{observed} - \text{expected}}{\text{SE}}$$

'Observed' is a statistic that is appropriate for assessing H_0. In the example of the 10 coin tosses, appropriate statistics would be the number of tails or the percent of tails.

'Expected' and SE are the expected value and the SE of this statistic, *computed under the assumption that H_0 is true.*

Setting up a test statistic

A **test statistic** measures how far away the data are from what we would expect if H_0 were true.

The most common test statistic is the **z-statistic**:

$$z = \frac{\text{observed} - \text{expected}}{\text{SE}}$$

'Observed' is a statistic that is appropriate for assessing H_0. In the example of the 10 coin tosses, appropriate statistics would be the number of tails or the percent of tails.

'Expected' and SE are the expected value and the SE of this statistic, *computed under the assumption that H_0 is true.*

In the example: Using the formulas for the sum of 0/1 labels we get 'expected' $= 10 \times \frac{1}{2} = 5$ and SE $= \sqrt{10}\sqrt{\frac{1}{2} \times \frac{1}{2}} = 1.58$. So

$$z = \frac{7 - 5}{1.58} = 1.27$$

p-values measure the evidence agains H_0

Large values of $|z|$ are evidence against H_0: The larger $|z|$ is, the stronger the evidence.
The strength of the evidence is measured by the
p-value (or: **observed significance level)**:

p-values measure the evidence agains H_0

Large values of $|z|$ are evidence against H_0: The larger $|z|$ is, the stronger the evidence.
The strength of the evidence is measured by the
p-value (or: **observed significance level)**:

The p-value is the probability of getting a value of z as extreme or more extreme than the observed z, assuming H_0 is true.

p-values measure the evidence agains H_0

Large values of $|z|$ are evidence against H_0: The larger $|z|$ is, the stronger the evidence. The strength of the evidence is measured by the
p-value (or: **observed significance level)**:

The p-value is the probability of getting a value of z as extreme or more extreme than the observed z, assuming H_0 is true.

But if H_0 is true, then z follows that standard normal curve, according to the central limit theorem, so the p-value can be computed with normal approximation:

p-values measure the evidence agains H_0

Large values of $|z|$ are evidence against H_0: The larger $|z|$ is, the stronger the evidence. The strength of the evidence is measured by the
p-value (or: **observed significance level)**:

The p-value is the probability of getting a value of z as extreme or more extreme than the observed z, assuming H_0 is true.

But if H_0 is true, then z follows that standard normal curve, according to the central limit theorem, so the p-value can be computed with normal approximation:

The smaller the p-value, the stronger the evidence against H_0. Often the criterion for rejecting H_0 is a p-value smaller than 5%. Then the result is called **statistically significant**.

p-values measure the evidence against H_0

In the example:

p-values measure the evidence against H_0

In the example:

Note that the p-value **does not** give the probability that H_0 is true, as H_0 is either true or not - there are no chances involved. Rather, it gives the probability of seeing a statistic as extreme, or more extreme, that the observed one, assuming H_0 is true.

Distinguishing Coke and Pepsi by taste

It has been said that it is difficult to distinguish Coke and Pepsi by taste alone, without the visual cue of the bottle or can.

Distinguishing Coke and Pepsi by taste

It has been said that it is difficult to distinguish Coke and Pepsi by taste alone, without the visual cue of the bottle or can.

In an experiment that I did in a class at Stanford, 10 cups were filled at random with either Coke or Pepsi. A student volunteer tasted each of the 10 cups and correctly named the contents of seven. Is this sufficient evidence to conclude that the student can tell apart Coke and Pepsi?

Distinguishing Coke and Pepsi by taste

It has been said that it is difficult to distinguish Coke and Pepsi by taste alone, without the visual cue of the bottle or can.

In an experiment that I did in a class at Stanford, 10 cups were filled at random with either Coke or Pepsi. A student volunteer tasted each of the 10 cups and correctly named the contents of seven. Is this sufficient evidence to conclude that the student can tell apart Coke and Pepsi?

"Nothing extraordinary is going on" means that the student does not have any special ability to tell them apart and is just guessing.

Distinguishing Coke and Pepsi by taste

It has been said that it is difficult to distinguish Coke and Pepsi by taste alone, without the visual cue of the bottle or can.

In an experiment that I did in a class at Stanford, 10 cups were filled at random with either Coke or Pepsi. A student volunteer tasted each of the 10 cups and correctly named the contents of seven. Is this sufficient evidence to conclude that the student can tell apart Coke and Pepsi?

"Nothing extraordinary is going on" means that the student does not have any special ability to tell them apart and is just guessing.

To write this down formally we introduce 0/1 labels since we are counting correct answers: $\quad 1 =$ correct answer, $\;0 =$ wrong answer

Distinguishing Coke and Pepsi by taste

It has been said that it is difficult to distinguish Coke and Pepsi by taste alone, without the visual cue of the bottle or can.

In an experiment that I did in a class at Stanford, 10 cups were filled at random with either Coke or Pepsi. A student volunteer tasted each of the 10 cups and correctly named the contents of seven. Is this sufficient evidence to conclude that the student can tell apart Coke and Pepsi?

"Nothing extraordinary is going on" means that the student does not have any special ability to tell them apart and is just guessing.

To write this down formally we introduce 0/1 labels since we are counting correct answers: $1 =$ correct answer, $0 =$ wrong answer

H_0: $P(0) = P(1) = \frac{1}{2}$ H_A: $P(1) > \frac{1}{2}$

Distinguishing Coke and Pepsi by taste

It has been said that it is difficult to distinguish Coke and Pepsi by taste alone, without the visual cue of the bottle or can.

In an experiment that I did in a class at Stanford, 10 cups were filled at random with either Coke or Pepsi. A student volunteer tasted each of the 10 cups and correctly named the contents of seven. Is this sufficient evidence to conclude that the student can tell apart Coke and Pepsi?

"Nothing extraordinary is going on" means that the student does not have any special ability to tell them apart and is just guessing.

To write this down formally we introduce 0/1 labels since we are counting correct answers: $1 =$ correct answer, $0 =$ wrong answer

H_0: $P(0) = P(1) = \frac{1}{2}$ $\qquad$ H_A: $P(1) > \frac{1}{2}$

This is a **one-sided test**: the alternative hypothesis for for P(1) we are interested in is on one side of $\frac{1}{2}$.

Distinguishing Coke and Pepsi by taste

Since we are looking at the sum of ten 0/1 labels, the z-statistic is the same that we had for coin-tossing:

$$z = \frac{\text{observed sum} - \text{expected sum}}{\text{SE of sum}} = \frac{7-5}{1.58} = 1.27$$

Distinguishing Coke and Pepsi by taste

Since we are looking at the sum of ten 0/1 labels, the z-statistic is the same that we had for coin-tossing:

$$z = \frac{\text{observed sum} - \text{expected sum}}{\text{SE of sum}} = \frac{7-5}{1.58} = 1.27$$

But since we do a one-sided test instead of a two-sided test, the p-value is only half as large:

Distinguishing Coke and Pepsi by taste

Since we are looking at the sum of ten 0/1 labels, the z-statistic is the same that we had for coin-tossing:

$$z = \frac{\text{observed sum} - \text{expected sum}}{\text{SE of sum}} = \frac{7-5}{1.58} = 1.27$$

But since we do a one-sided test instead of a two-sided test, the p-value is only half as large:

Since 10.2% is not smaller than 5%, we don't reject H_0: We are not convinced that the student can distinguish Coke and Pepsi.

Distinguishing Coke and Pepsi

A two-sided alternative might also be appropriate:

H_A: $\text{P}(1) \neq \frac{1}{2}$

H_A corresponds to a student who is more likely than not to distinguish Coke and Pepsi, but who may confuse them. Such a student might get one correct answer (say).

Distinguishing Coke and Pepsi

A two-sided alternative might also be appropriate:

H$_A$: P(1) $\neq \frac{1}{2}$

H$_A$ corresponds to a student who is more likely than not to distinguish Coke and Pepsi, but who may confuse them. Such a student might get one correct answer (say).

One has to carefully consider whether the alternative should be one-sided or two-sided, as the p-value gets doubled in the latter case.

Distinguishing Coke and Pepsi

A two-sided alternative might also be appropriate:

H_A: $\text{P}(1) \neq \frac{1}{2}$

H_A corresponds to a student who is more likely than not to distinguish Coke and Pepsi, but who may confuse them. Such a student might get one correct answer (say).

One has to carefully consider whether the alternative should be one-sided or two-sided, as the p-value gets doubled in the latter case.

It is not ok to change the alternative afterwards in order to get the p-value below 5%.

The t-test

The health guideline for lead in drinking water is a concentration of not more than 15 parts per billion (ppb).

The t-test

The health guideline for lead in drinking water is a concentration of not more than 15 parts per billion (ppb).

Five independent samples from a reservoir average 15.6 ppb. Is this sufficient evidence to conclude that the concentration μ in the reservoir is above the standard of 15 ppb?

The t-test

The health guideline for lead in drinking water is a concentration of not more than 15 parts per billion (ppb).

Five independent samples from a reservoir average 15.6 ppb. Is this sufficient evidence to conclude that the concentration μ in the reservoir is above the standard of 15 ppb?

Recall our model for measurements:

$$\text{measurement} = \mu + \text{measurement error}$$

The t-test

The health guideline for lead in drinking water is a concentration of not more than 15 parts per billion (ppb).

Five independent samples from a reservoir average 15.6 ppb. Is this sufficient evidence to conclude that the concentration μ in the reservoir is above the standard of 15 ppb?

Recall our model for measurements:

$$\text{measurement} = \mu + \text{measurement error}$$

So it may be that the concentration μ is below 15 ppb, but measurement error results in an average of 15.6 ppb.

The t-test

The health guideline for lead in drinking water is a concentration of not more than 15 parts per billion (ppb).

Five independent samples from a reservoir average 15.6 ppb. Is this sufficient evidence to conclude that the concentration μ in the reservoir is above the standard of 15 ppb?

Recall our model for measurements:

$$\text{measurement} = \mu + \text{measurement error}$$

So it may be that the concentration μ is below 15 ppb, but measurement error results in an average of 15.6 ppb.

H_0: $\mu = 15$ ppb $\qquad$ H_A: $\mu > 15$ ppb

The t-test

The health guideline for lead in drinking water is a concentration of not more than 15 parts per billion (ppb).

Five independent samples from a reservoir average 15.6 ppb. Is this sufficient evidence to conclude that the concentration μ in the reservoir is above the standard of 15 ppb?

Recall our model for measurements:

$$\text{measurement} = \mu + \text{measurement error}$$

So it may be that the concentration μ is below 15 ppb, but measurement error results in an average of 15.6 ppb.

H_0: $\mu = 15$ ppb $\qquad$ H_A: $\mu > 15$ ppb

We can try a z-test for the average of the measurements:

$$z = \frac{\text{observed average} - \text{expected average}}{\text{SE of average}} = \frac{15.6 \text{ ppb} - 15 \text{ ppb}}{\text{SE of average}}$$

since the measurement error has expected value zero.

The t-test

SE of average $= \frac{\sigma}{\sqrt{n}}$, but the standard deviation σ of the measurement error is unknown.

The t-test

SE of average $= \frac{\sigma}{\sqrt{n}}$, but the standard deviation σ of the measurement error is unknown.

We can estimate σ by s, the sample standard deviation of the measurements. However:

The t-test

SE of average $= \frac{\sigma}{\sqrt{n}}$, but the standard deviation σ of the measurement error is unknown.

We can estimate σ by s, the sample standard deviation of the measurements. However:

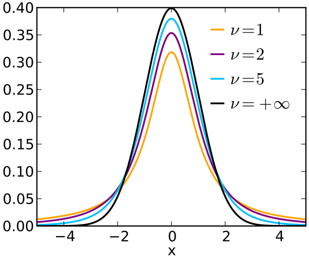
If we estimate σ and n is small $(n \leq 20)$, then the normal curve is not a good enough approximation to the distribution of the z-statistic. Rather, an appropriate approximation is **Student's t-distribution with** $n-1$ **degrees of freedom**:

The t-test

SE of average $= \frac{\sigma}{\sqrt{n}}$, but the standard deviation σ of the measurement error is unknown.

We can estimate σ by s, the sample standard deviation of the measurements. However:

If we estimate σ and n is small ($n \leq 20$), then the normal curve is not a good enough approximation to the distribution of the z-statistic. Rather, an appropriate approximation is **Student's t-distribution with $n-1$ degrees of freedom**:



The t-test

The fatter tails account for the additional uncertainty introduced by estimating σ by $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$.

The t-test

The fatter tails account for the additional uncertainty introduced by estimating σ by $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$.

Using the **t-test** in place of the z-test is only necessary for small samples: $n \leq 20$ (say).

The t-test

The fatter tails account for the additional uncertainty introduced by estimating σ by $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$.

Using the **t-test** in place of the z-test is only necessary for small samples: $n \leq 20$ (say).

In that case it is also better to replace the confidence interval $\bar{x} \pm z$ SE by

$$\bar{x} \pm t_{n-1}\mathsf{SE}$$

More on testing

- *Statistically significant does not mean that the effect size is important*:

More on testing

- *Statistically significant does not mean that the effect size is important*:

  Suppose the sample average shows a lead concentration that is only slightly above the health standard of 15 ppb: say the sample average is 15.05 ppb.

More on testing

- *Statistically significant does not mean that the effect size is important*:

  Suppose the sample average shows a lead concentration that is only slightly above the health standard of 15 ppb: say the sample average is 15.05 ppb.

  That may not be of practical concern, even though the test may be highly signficant: Statistical significance convinces us that there is an effect, but it doesn't say how big the effect is.

More on testing

- *Statistically significant does not mean that the effect size is important*:

  Suppose the sample average shows a lead concentration that is only slightly above the health standard of 15 ppb: say the sample average is 15.05 ppb.

  That may not be of practical concern, even though the test may be highly signficant: Statistical significance convinces us that there is an effect, but it doesn't say how big the effect is.

  Reason: A large sample size n makes SE $= \frac{\sigma}{\sqrt{n}}$ small, so even a small exceedance over the limit by (say) 0.05 ppb may give a statistically significant result.

More on testing

- *Statistically significant does not mean that the effect size is important*:

  Suppose the sample average shows a lead concentration that is only slightly above the health standard of 15 ppb: say the sample average is 15.05 ppb.

  That may not be of practical concern, even though the test may be highly signficant: Statistical significance convinces us that there is an effect, but it doesn't say how big the effect is.

  Reason: A large sample size n makes SE $= \frac{\sigma}{\sqrt{n}}$ small, so even a small exceedance over the limit by (say) 0.05 ppb may give a statistically significant result.

  Therefore it is helpful to complement a test with a confidence interval: In the above case a 95% confidence interval for μ might be [15.02 ppb, 15.08 ppb].

More on testing

- There is a general connection between confidence intervals and tests:

More on testing

- There is a general connection between confidence intervals and tests:

  A 95% confidence interval contains all values for the null hypothesis that will not be rejected by a two-sided test at a 5% significance level.
  (A 5% **significance level** means that the threshold for the p-value is 5%).

More on testing

- There is a general connection between confidence intervals and tests:

  A 95% confidence interval contains all values for the null hypothesis that will not be rejected by a two-sided test at a 5% significance level.
  (A 5% **significance level** means that the threshold for the p-value is 5%).
- There are two ways that a test can result in a wrong decision:

More on testing

- There is a general connection between confidence intervals and tests:

  A 95% confidence interval contains all values for the null hypothesis that will not be rejected by a two-sided test at a 5% significance level.
  (A 5% **significance level** means that the threshold for the p-value is 5%).

- There are two ways that a test can result in a wrong decision:

  H_0 is true, but was erroneously rejected $\rightarrow$ Type I error ('false positive')

More on testing

- There is a general connection between confidence intervals and tests:

  A 95% confidence interval contains all values for the null hypothesis that will not be rejected by a two-sided test at a 5% significance level.
  (A 5% **significance level** means that the threshold for the p-value is 5%).
- There are two ways that a test can result in a wrong decision:

  H_0 is true, but was erroneously rejected $\rightarrow$ Type I error ('false positive')

  H_0 is false, but we fail to reject it $\rightarrow$ Type II error

More on testing

- There is a general connection between confidence intervals and tests:

  A 95% confidence interval contains all values for the null hypothesis that will not be rejected by a two-sided test at a 5% significance level.
  (A 5% **significance level** means that the threshold for the p-value is 5%).
- There are two ways that a test can result in a wrong decision:

  H_0 is true, but was erroneously rejected $\rightarrow$ Type I error ('false positive')

  H_0 is false, but we fail to reject it $\rightarrow$ Type II error

  Rejecting H_0 if the p-value is smaller than 5% means P(type I error) $\leq 5\%$

The two-sample z-test

Last month, the President's approval rating in a sample of 1,000 likely voters was 55%. This month, a poll of 1,500 likely voters resulted in a rating of 58%. Is this sufficient evidence to conclude that the rating has changed?

The two-sample z-test

Last month, the President's approval rating in a sample of 1,000 likely voters was 55%. This month, a poll of 1,500 likely voters resulted in a rating of 58%. Is this sufficient evidence to conclude that the rating has changed?

We want to assess whether

$p_1 =$proportion of all likely voters approving last month

is equal to

$p_2 =$proportion of all likely voters approving this month

The two-sample z-test

Last month, the President's approval rating in a sample of 1,000 likely voters was 55%. This month, a poll of 1,500 likely voters resulted in a rating of 58%. Is this sufficient evidence to conclude that the rating has changed?

We want to assess whether

$p_1 =$proportion of all likely voters approving last month

is equal to

$p_2 =$proportion of all likely voters approving this month

"nothing unusual is going on" means $p_1 = p_2$.

The two-sample z-test

Last month, the President's approval rating in a sample of 1,000 likely voters was 55%. This month, a poll of 1,500 likely voters resulted in a rating of 58%. Is this sufficient evidence to conclude that the rating has changed?

We want to assess whether

$p_1 =$proportion of all likely voters approving last month

is equal to

$p_2 =$proportion of all likely voters approving this month

"nothing unusual is going on" means $p_1 = p_2$. It's common to look at the difference $p_2 - p_1$ instead:

$$\mathsf{H}_0 : \; p_2 - p_1 = 0 \qquad\qquad \mathsf{H}_1 : \; p_2 - p_1 \neq 0$$

The two-sample z-test

Last month, the President's approval rating in a sample of 1,000 likely voters was 55%. This month, a poll of 1,500 likely voters resulted in a rating of 58%. Is this sufficient evidence to conclude that the rating has changed?

We want to assess whether

$p_1 =$proportion of all likely voters approving last month

is equal to

$p_2 =$proportion of all likely voters approving this month

"nothing unusual is going on" means $p_1 = p_2$. It's common to look at the difference $p_2 - p_1$ instead:

$$\mathrm{H}_0 : \ p_2 - p_1 = 0 \qquad\qquad \mathrm{H}_1 : \ p_2 - p_1 \neq 0$$

p_1 is estimated by $\hat{p}_1 = 55\%$, p_2 by $\hat{p}_2 = 58\%$.

The two-sample z-test

Last month, the President's approval rating in a sample of 1,000 likely voters was 55%. This month, a poll of 1,500 likely voters resulted in a rating of 58%. Is this sufficient evidence to conclude that the rating has changed?

We want to assess whether

$p_1 =$proportion of all likely voters approving last month

is equal to

$p_2 =$proportion of all likely voters approving this month

"nothing unusual is going on" means $p_1 = p_2$. It's common to look at the difference $p_2 - p_1$ instead:

H$_0$: $p_2 - p_1 = 0$ H$_1$: $p_2 - p_1 \neq 0$

p_1 is estimated by $\hat{p}_1 = 55\%$, p_2 by $\hat{p}_2 = 58\%$. The central limit theorem applies to the difference $\hat{p}_2 - \hat{p}_1$ just as it does to $\hat{p}_1$ and $\hat{p}_2$. So we can use a z-test:

The two-sample z-test

We can use a z-test for the difference $\hat{p}_2 - \hat{p}_1$:

The two-sample z-test

We can use a z-test for the difference $\hat{p}_2 - \hat{p}_1$:

$$z = \frac{\text{observed difference} - \text{expected difference}}{\text{SE of difference}} = \frac{(\hat{p}_2 - \hat{p}_1) - (p_2 - p_1)}{\text{SE of difference}}$$

The two-sample z-test

We can use a z-test for the difference $\hat{p}_2 - \hat{p}_1$:

$$z = \frac{\text{observed difference} - \text{expected difference}}{\text{SE of difference}} = \frac{(\hat{p}_2 - \hat{p}_1) - (p_2 - p_1)}{\text{SE of difference}}$$

An important fact is that if $\hat{p}_1$ and $\hat{p}_2$ are independent, then

$$\text{SE}(\hat{p}_2 - \hat{p}_1) = \sqrt{(\text{SE}(\hat{p}_1))^2 + (\text{SE}(\hat{p}_2))^2}.$$

The two-sample z-test

We can use a z-test for the difference $\hat{p}_2 - \hat{p}_1$:

$$z = \frac{\text{observed difference} - \text{expected difference}}{\text{SE of difference}} = \frac{(\hat{p}_2 - \hat{p}_1) - (p_2 - p_1)}{\text{SE of difference}}$$

An important fact is that if $\hat{p}_1$ and $\hat{p}_2$ are independent, then

$$\text{SE}(\hat{p}_2 - \hat{p}_1) = \sqrt{(\text{SE}(\hat{p}_1))^2 + (\text{SE}(\hat{p}_2))^2}.$$ So

$$z = \frac{(\hat{p}_2 - \hat{p}_1) - 0}{\sqrt{\sqrt{\frac{p_1(1-p_1)}{1000}}^2 + \sqrt{\frac{p_2(1-p_2)}{1500}}^2}} =$$

The two-sample z-test

We can use a z-test for the difference $\hat{p}_2 - \hat{p}_1$:

$$z = \frac{\text{observed difference} - \text{expected difference}}{\text{SE of difference}} = \frac{(\hat{p}_2 - \hat{p}_1) - (p_2 - p_1)}{\text{SE of difference}}$$

An important fact is that if $\hat{p}_1$ and $\hat{p}_2$ are independent, then

$$\text{SE}(\hat{p}_2 - \hat{p}_1) = \sqrt{(\text{SE}(\hat{p}_1))^2 + (\text{SE}(\hat{p}_2))^2}. \qquad \text{So}$$

$$z = \frac{(\hat{p}_2 - \hat{p}_1) - 0}{\sqrt{\sqrt{\frac{p_1(1-p_1)}{1000}}^2 + \sqrt{\frac{p_2(1-p_2)}{1500}}^2}} = \frac{0.03}{0.0202} = 1.48$$

The two-sample z-test

The two-sample z-test is applicable in the same way to the difference of two sample means in order to test for equality of two population means.

The two-sample z-test

The two-sample z-test is applicable in the same way to the difference of two sample means in order to test for equality of two population means.

If the two samples are independent, then again

$$\mathsf{SE}(\bar{x}_2 - \bar{x}_1) \;=\; \sqrt{(\mathsf{SE}(\bar{x}_1))^2 + (\mathsf{SE}(\bar{x}_2))^2}$$

The two-sample z-test

The two-sample z-test is applicable in the same way to the difference of two sample means in order to test for equality of two population means.

If the two samples are independent, then again

$$\mathsf{SE}(\bar{x}_2 - \bar{x}_1) \;=\; \sqrt{(\mathsf{SE}(\bar{x}_1))^2 + (\mathsf{SE}(\bar{x}_2))^2}$$

and $\mathsf{SE}(\bar{x}_1) = \frac{\sigma_1}{\sqrt{n_1}}$ is estimated by $\frac{s_1}{\sqrt{n_1}}$.

The two-sample z-test

The two-sample z-test is applicable in the same way to the difference of two sample means in order to test for equality of two population means.

If the two samples are independent, then again

$$\mathsf{SE}(\bar{x}_2 - \bar{x}_1) \;=\; \sqrt{(\mathsf{SE}(\bar{x}_1))^2 + (\mathsf{SE}(\bar{x}_2))^2}$$

and $\mathsf{SE}(\bar{x}_1) = \frac{\sigma_1}{\sqrt{n_1}}$ is estimated by $\frac{s_1}{\sqrt{n_1}}$.

All of the above two-sample tests require that the two samples are independent. They are also applicable in special situations where the samples are dependent, e.g. to compare the treatment effect when subjects are randomized into treatment and control groups.

The paired-difference test

Do husbands tend to be older than their wives?

The paired-difference test

Do husbands tend to be older than their wives?

The ages of five couples:

Husband's age	Wife's age	age difference
43	41	2
71	70	1
32	31	1
68	66	2
27	26	1

The paired-difference test

Do husbands tend to be older than their wives?

The ages of five couples:

Husband's age	Wife's age	age difference
43	41	2
71	70	1
32	31	1
68	66	2
27	26	1

The two-sample t-test is not applicable since the two samples are not independent.

The paired-difference test

Do husbands tend to be older than their wives?

The ages of five couples:

Husband's age	Wife's age	age difference
43	41	2
71	70	1
32	31	1
68	66	2
27	26	1

The two-sample t-test is not applicable since the two samples are not independent. Even if they were independent, the small differences in ages would not be significant since the standard deviations are large for husbands and also for the wives.

The paired-difference test

Since we have paired data, we can simply analyze the differences obtained from each pair with a regular t-test, which in this context of **matched pairs** is called **paired t-test**:

The paired-difference test

Since we have paired data, we can simply analyze the differences obtained from each pair with a regular t-test, which in this context of **matched pairs** is called **paired t-test**:

H_0: population difference has mean zero

The paired-difference test

Since we have paired data, we can simply analyze the differences obtained from each pair with a regular t-test, which in this context of **matched pairs** is called **paired t-test**:

H_0: population difference has mean zero

$t = \frac{\bar{d}-0}{\mathsf{SE}(\bar{d})}$, where d_i is the age difference of the ith couple.

The paired-difference test

Since we have paired data, we can simply analyze the differences obtained from each pair with a regular t-test, which in this context of **matched pairs** is called **paired t-test**:

H$_0$: population difference has mean zero

$t = \frac{\bar{d}-0}{\mathsf{SE}(\bar{d})}$, where d_i is the age difference of the ith couple.

$\mathsf{SE}(\bar{d}) = \frac{\sigma_d}{\sqrt{n}}$. Estimate σ_d by $s_d = 0.55$.

The paired-difference test

Since we have paired data, we can simply analyze the differences obtained from each pair with a regular t-test, which in this context of **matched pairs** is called **paired t-test**:

H$_0$: population difference has mean zero

$t \; = \; \frac{\bar{d}-0}{\mathsf{SE}(\bar{d})}$, where d_i is the age difference of the ith couple.

$\mathsf{SE}(\bar{d}) = \frac{\sigma_d}{\sqrt{n}}$. Estimate σ_d by $s_d = 0.55$. Then $t = \frac{1.4-0}{0.55/\sqrt{5}} \; = \; 5.69$

The paired-difference test

Since we have paired data, we can simply analyze the differences obtained from each pair with a regular t-test, which in this context of **matched pairs** is called **paired t-test**:

H_0: population difference has mean zero

$t \;=\; \frac{\bar{d}-0}{\mathsf{SE}(\bar{d})}$, where d_i is the age difference of the ith couple.

$\mathsf{SE}(\bar{d}) = \frac{\sigma_d}{\sqrt{n}}$. Estimate σ_d by $s_d = 0.55$. Then $t = \frac{1.4-0}{0.55/\sqrt{5}} \;=\; 5.69$

The independence assumption is in the sampling of the couples.

The sign test

What if didn't know the age difference d_i but only if the husband was older or not?

The sign test

What if didn't know the age difference d_i but only if the husband was older or not?

We can test

H_0: half the husbands in the population are older than their wives

The sign test

What if didn't know the age difference d_i but only if the husband was older or not?

We can test

H_0: half the husbands in the population are older than their wives

using 0/1 labels and a z-test, just as we tested whether a coin is fair:

The sign test

What if didn't know the age difference d_i but only if the husband was older or not?

We can test

  H_0: half the husbands in the population are older than their wives

using 0/1 labels and a z-test, just as we tested whether a coin is fair:

$$z = \frac{\text{sum of 1s} - \frac{n}{2}}{\text{SE of sum}} = \frac{5 - \frac{5}{2}}{\sqrt{5}\frac{1}{2}} = 2.24 \quad \text{since } \sigma = \frac{1}{2} \text{ on } H_0.$$

The sign test

What if didn't know the age difference d_i but only if the husband was older or not?

We can test

H$_0$: half the husbands in the population are older than their wives

using 0/1 labels and a z-test, just as we tested whether a coin is fair:

$$z = \frac{\text{sum of 1s} - \frac{n}{2}}{\text{SE of sum}} = \frac{5 - \frac{5}{2}}{\sqrt{5}\frac{1}{2}} = 2.24 \quad \text{since } \sigma = \frac{1}{2} \text{ on H}_0.$$

The p-value of this **sign-test** is less significant than that of the paired t-test. This is because the latter uses more information, namely the size of the differences.

The sign test

What if didn't know the age difference d_i but only if the husband was older or not?

We can test

H$_0$: half the husbands in the population are older than their wives

using 0/1 labels and a z-test, just as we tested whether a coin is fair:

$$z = \frac{\text{sum of 1s} - \frac{n}{2}}{\text{SE of sum}} = \frac{5 - \frac{5}{2}}{\sqrt{5}\frac{1}{2}} = 2.24 \quad \text{since } \sigma = \frac{1}{2} \text{ on } \mathrm{H}_0.$$

The p-value of this **sign-test** is less significant than that of the paired t-test. This is because the latter uses more information, namely the size of the differences. On the other hand, the sign test has the virtue of easy interpretation due to the analogy to coin tossing.

Mini quiz

1. True or false:
   a. The p-value depends on the data.
   b. If the p-value is smaller than 5%, then there is less than a 5% chance that the null hypothesis is true.
   c. If the null hypothesis is true, then there is less than a 5% chance to get a p-value that is smaller than 5%.
   d. If a data scientist does many tests, then even if the all the null hypotheses are true, a certain proportion will be rejected in error.

2. For each of the following situations, indicate which test is appropriate to address the respective question: z-test, t-test, two-sample z-test, sign test, or paired-difference test.

a. You want to test whether plain M&Ms really contain 24% blue M&Ms as claimed on the manufacturer's web site. You sample 500 plain M&Ms at random and count the fraction of blue M&Ms.

b. A high school principal wants to find out whether the average SAT score of this year's graduating class is higher than last year's. She samples 13 students from this year's graduating class at random and wants to compare their average SAT score to the average SAT score from last year's graduating class.

c. To investigate whether there are difference in scholastic abilities between first-borns and second-born siblings, 600 families that have at least two children were randomly selected. The scholastic abilities of the first-born and the second-born siblings were assessed with a test and are to be compared.