Understanding common misconceptions about *p*-values

**Q1**: When the sample size in each group of an independent *t*-test is 50 observations (see the Figure above), which statement is correct?

A) The mean of the differences you will observe between the two groups is always 0.

*feedback:* **Although centered on 0, recall that this is a *distribution* of observations, meaning that even if the mean in the population is 0, that does not imply every sample we draw will give a mean of exactly zero.**

B) The mean of the differences you will observe between the two groups is always different from 0.

*feedback:* **While a distribution centered on 0 does not imply that every sample we draw will give a mean of exactly zero, it remains possible that the difference between two groups is exactly zero (e.g., in two groups, all participants answer a 4 on a scale from 1 to 7).**

C) Observing a mean difference of +0.5 or -0.5 is considered surprising, assuming the null-hypothesis is true.

*feedback:* **Correct!**

D) Observing a mean difference of +0.1 or -0.1 is considered surprising, assuming the null-hypothesis is true.

*feedback:* **Remember: Only observations that fall within the red areas of the tails are considered surprising. Because +0.1 and -0.1 do not fall beyond these critical thresholds (but rather reside within the white area of the distribution), they will occur relatively often, and therefore we can not call them ‘surprising’, assuming the null is true.**

**Q2:** In what sense are the null models in the two previous figures (for n = 50 and n = 5000) similar, and in what sense are they different?

A) In both cases, the distributions are centered on zero, and the critical *t*-value is between 1.96 and 2 (for a two-sided test, depending on the sample size). But the larger the sample size, the closer to 0 the mean differences fall that are considered ‘surprising’.

*feedback:* **Correct!**

B) In both cases, a *t*-value of 0 is the most likely outcome, but the critical *t*-value is around 0.4 for n = 50, and around 0.05 for n = 5000.

*feedback:* **Note that in these figures (unlike the very first one you saw), the x-axis refers to the difference score of the means of each group. These values differ importantly from, and should not be interpreted as, *t*-values. Whereas the critical *t*-value remains between 1.96 and 2 (for a two-sided test, depending on the sample size), smaller observations (e.g., difference of 0.05) are considered ‘surprising’ for n = 5000 than for n = 50 (e.g., difference of 0.4).**

C) In both cases, means will vary in exactly the same way around 0, but the Type 1 error rate is much smaller when n = 5000 than when n = 50.

*feedback:* **Careful! There are 2 issues to note: First, we know that with a larger N, the distribution of observations narrows. In other words, while both N = 50 and N = 5000 will vary around a mean of 0, it will not be in exactly the same way: We should expect to observe mean differences closer to 0 in the larger sample. Second, the Type 1 error rate is determined independently of sample size, and equals the alpha-level (i.e. 5% Type 1 errors in the long-run for an alpha of 0.05).**

D) Because the standard error is much larger for n = 50 than for n = 5000, it is much more likely that the null hypothesis is true for n = 50.

*feedback***: The probability to observe a value of 0 does not depend on the sample size.**

If we collected n = 5000, and we would again observe a mean difference of 0.5, it should be clear that this same difference is even more surprising than it was when we collected 50 observations.

**Q3:** Open the app, and make sure it is set to the default settings of a sample size of 50 and an alpha level of 0.05. Look at the distribution of the null model. Set the sample size to 2. Set the sample size to 5000. The app will not allow you to plot data for a ‘group’ size of 1, but with n = 2 you will get a pretty good idea of the range of values you can expect when the true effect is 0, and when you collect single observations (n = 1). Which statement is true for n = 50?

A) When the null hypothesis is true and the standard deviation is 1, if you randomly take 1 observation from each group and calculate the difference score, the differences will fall between -0.4 and 0.4 for 95% of the pairs of observations you will draw.

*feedback:* **This is incorrect - the distribution of mean differences for 1 pair is much wider (as you saw when you set the n to 2 in the app).**

B) When the null hypothesis is true and the standard deviation is 1, with n = 50 per group, 95% of studies where data is collected will observe a mean difference between -0.4 and 0.4.

*feedback:* **Correct! (because 95% of the data falls between the critical values of
-0.4 and 0.4).**

C) In any study with n = 50 per group, even when the SD is unknown and it is not known if the null hypothesis is true, you should rarely observe a mean difference more extreme than -0.4 or 0.4.

*feedback:* **The frequency that you observe a mean difference more extreme than the critical value (in this case, -0.4 or 0.4) depends not only on sample size, but other important factors, such as alpha-level, mean difference, and which model is true. If the alternative model is true, and the true difference is 0.4, obviously it is much more likely to observe a means difference of approximately 0.4.**

D) As the sample size increases, the expected distribution of means become narrower for the null model, but not for the alternative model.

*feedback:* **As the sample size increases, the expected distribution of means becomes narrower for both the null model and the alternative model. As you can see in the app, increasing N causes the observations of the null- and alternative models to cluster or tighten more closely around their respective distribution means; this is because larger samples yield more precise estimates of population parameters (which visually can be seen as smaller distribution standard errors).**

**Q4:** Open the app once more with the default settings. Set the slider for the alpha level to 0.01 (while keeping the mean difference at 0.5 and the sample size at 50). Compared to the critical value when alpha = 0.05, which statement is true?

A) Compared to an alpha of 0.05, only *less* extreme values are considered surprising when an alpha of 0.01 is used, and only differences larger than 0.53 scale points (or smaller than -0.53) will now be statistically significant.

*feedback:* **Careful. Take another look at the figure in the app. What you should notice is that by shrinking the alpha level from 5% to 1%, the red area in the tail of the null model has become smaller. This is because the critical threshold value has moved further away from the mean of the null distribution (from 0.4 to 0.53). This means that even *more* extreme observations (i.e. *greater* than 0.53 scale points, or *smaller* than - 0.53) are required in order to be considered ‘surprising’. You can think about this as the conditions becoming more stringent. Before, we allowed ourselves a 5% Type 1 error rate; now we’re only allowing this to occur 1% of the time, thus for an observation to be considered significant or surprising, they must reach a *more* (not a *less*) extreme threshold.**

B) Compared to an alpha of 0.05, only *less* extreme values are considered surprising when an alpha of 0.01 is used, and only differences larger than 0.33 scale points (or smaller than -0.33) will now be statistically significant.

*feedback:* **Careful! There are 2 issues here. First, if you’re observing a critical mean difference of 0.33 in the app, then you’ve likely set the alpha level to 0.1 *not* 0.01. Try resetting the alpha value to 0.01 (i.e. 1%) and reanalyzing the figure. What you should notice is that by shrinking the alpha level from 5% to 1%, the red area in the tail of the null model has become smaller. This is because the critical threshold value has moved further away from the mean of the null distribution (from 0.4 to 0.53). This means that even *more* extreme observations (i.e. *greater* than a mean difference of 0.53) are required in order to be considered ‘surprising’. You can think about this as the conditions becoming more stringent. Before, we allowed ourselves a 5% Type 1 error rate; now we’re only allowing this to occur 1% of the time, thus for an observation to be considered significant or surprising, they must reach a *more* (not a *less*) extreme threshold.**

C) Compared to an alpha of 0.05, only *more* extreme values are considered surprising when an alpha of 0.01 is used, and only differences larger than 0.53 scale points (or smaller than -0.53) will be statistically significant.

*feedback:* **Correct!**

D) Compared to an alpha of 0.05, only *more* extreme values are considered surprising when an alpha of 0.01 is used, and only differences larger than 0.33 scale points (or smaller than -0.33) will now be statistically significant.

*feedback:* **Careful! If you’re observing a critical value of a mean difference of 0.33 in the app, you’ve likely set the alpha level to 0.1 *not* 0.01. Try resetting the alpha value to 0.01 and revisit the question.**

**Q5:** Why can’t you conclude that the null hypothesis is true, when you observe a statistically non-significant *p*-value (*p* > alpha)?

A) When calculating *p*-values you always need to take the prior probability into account.

*feedback:* **This is incorrect – *p*-values are calculated without taking prior probabilities that the null-hypothesis or alternative hypothesis are true into account.**

B) You need to acknowledge the probability that you have observed a Type 1 error.

*feedback:* **We have observed a non-significant effect, so it is not relevant that we could also have observed a significant result. Thus, for the question it is relevant that we might have made a Type 2 error, not that we might have made a Type 1 error.**

C) The null hypothesis is never true.

*feedback:* **This is incorrect – although we can have a philosophical discussion about what ‘true’ means, in hypothesis tests the notion is that sometimes the null hypothesis is true (i.e., there is a difference of 0 in the population) and sometimes the alternative hypothesis is true (i.e., there is a difference in the population that is not 0). Whether the null can never be true is actually unknown (or at least I would not know how to scientifically prove it).**

D) You need to acknowledge the probability that you have observed a Type 2 error.

*feedback:* **Correct!**

**Q6:** Why can’t you conclude that the alternative hypothesis is true, when you observe a statistically significant *p*-value (*p* < alpha)?

A) When calculating *p*-values you always need to take the prior probability into account.

*feedback:* **This is incorrect – *p*-values are calculated without taking prior probabilities that the null-hypothesis or alternative hypothesis are true into account.**

B) You need to acknowledge the probability that you have observed a Type 1 error.

*feedback:* **Correct!**

C) The alternative hypothesis is never true.

*feedback:* **This is incorrect – although we can have a philosophical discussion about what ‘true’ means, in hypothesis tests the notion is that sometimes the null-hypothesis is true (i.e., there is a difference of 0 in the population) and sometimes the alternative hypothesis is true (i.e., there is a difference in the population that is not 0).**

D) You need to acknowledge the probability that you have observed a Type 2 error.

*feedback:* **We have observed a significant effect, so it is not relevant that we could also have observed a non-significant result. Thus, for the question it is relevant that we might have made a Type 1 error, not that we might have made a Type 2 error.**

**Q7:** Go to the app: <http://shiny.ieis.tue.nl/d_p_power/>. Set the sample size to 50000, the mean difference to 0.5, and the alpha level to 0.05. Which effects will, when observed, be statistically different from 0?

A) Effects more extreme than -0.01 and 0.01

*feedback:* **Correct!**

B) Effects more extreme than -0.04 and 0.04

*feedback:* **Careful! If you’re observing an effect threshold of a mean difference of -0.04 and 0.04, you’ve likely set your sample size to 5000, not 50000. Try resetting N to 50000 and reanalyze the figure. What you should notice are very narrow null and alternative distributions. Look closely at the null distribution and at the values where the red tails begin. These should fall at -0.01 and 0.01. This is confirmed by the verbal label on the graph which states “Effects larger than a mean difference of 0.01 will be statistically significant.”**

C) Effects more extreme than -0.05 and 0.05

*feedback:* **Make sure you’ve set N to 50000 and reanalyze the figure. What you should notice are very narrow null and alternative distributions. Look closely at the null distribution and at the values where the red tails begin. These should fall at -0.01 and 0.01. This is confirmed by the verbal label on the graph which states “Effects larger than a mean difference of 0.01 will be statistically significant.”**

D) Effects more extreme than -0.12 and 0.12

*feedback:* **Careful! If you’re observing an effect threshold of a mean difference of -0.12 and 0.12, you’ve likely set your sample size to 500, not 50000. Try resetting N to 50000 and reanalyze the figure. What you should notice are very narrow null and alternative distributions. Look closely at the null distribution and at the values where the red tails begin. These should fall at -0.01 and 0.01. This is confirmed by the verbal label on the graph which states “Effects larger than a mean difference of 0.01 will be statistically significant.”**

**Q8:** Let’s assume that the random number generator in R works, and we use rnorm(n = 50, mean = 0, sd = 1) to generate 50 observations, and the mean of these observations is 0.5, which in a one-sample *t*-test yields a *p*-value of 0.03, which is smaller than the alpha level (which we have set to 0.05). What is the probability that we have observed a significant difference (*p* < alpha) just by chance?

A) 3%

*feedback:* **The probability that this result is due to chance does not equal the p-value. Because we know the null hypothesis is true (the random number generator works) there is no other explanation for a low *p*-value than random variation or chance.**

B) 5%

*feedback:* **The probability that this result is due to chance does not equal the Type 1 error rate. There is a 5% chance to observe a Type 1 error if the null hypothesis is true, but we have already observed a Type 1 error. Because we know the null hypothesis is true (the random number generator works) there is no other explanation for a low *p*-value than random variation or chance.**

C) 95%

*feedback:* **The probability that this result is due to chance does not equal 1- the Type 1 error rate. There is a 5% chance to observe a Type 1 error if the null hypothesis is true, but we have already observed a Type 1 error. Because we know the null hypothesis is true (the random number generator works) there is no other explanation for a low *p*-value than random variation or chance.**

D) 100%

*feedback:* **Correct!**

**Q9:** Which statement is true?

A) The probability that a replication study will yield a significant result is 1-*p*.

*feedback:* **Remember: The *p*-value in one study cannot inform us about the *p*-value we will observe in future studies, such as a replication study. On the other hand, the level of statistical power informs us about how frequently we should yield a significant result (e.g., 80% power means we should observe significant results 80% of the time, if a true effect exists).**

B) The probability that a replication study will yield a significant result is 1-*p* multiplied by the probability that the null-hypothesis is true.

*feedback:* **Remember: The *p*-value in one study cannot inform us about the *p*-value we will observe in future studies, such as a replication study. On the other hand, the level of statistical power informs us about how frequently we should yield a significant result (e.g., 80% power means we should observe significant results 80% of the time, if a true effect exists).**

C) The probability that a replication study will yield a significant result is equal to the statistical power of the replication study (if there is a true effect), or the alpha level (if there is no true effect).

*feedback:* **Correct!**

D) The probability that a replication study will yield a significant result is equal to the statistical power of the replication study + the alpha level.

*feedback:* **Careful! This answer contains components which are correct, but when taken together are wrong. Here’s why: *If a true effect exists*, then the level of statistical power informs us about how frequently we should yield a significant result (e.g., 80% power means we should observe significant results 80% of the time). On the other hand, *if the effect is null (or non-existent)*, then significant results will be observed only 5% of the time in the long run (i.e. the Type 1 error rate given an alpha of 0.05). Therefore, *either* the statistical power *or* the alpha level equals the probability of replication, depending on if there *is* or *isn’t* a true effect.**

**Q10:** Does a non-significant *p*-value (i.e., *p* = 0.65) mean that the null-hypothesis is true?

A) No - the result could be a Type 2 error, or a false negative.

*feedback:* **Correct! A non-significant result can be a true negative (when the null-hypothesis is true) or a false negative (or a Type 2 error, when the alternative hypothesis is true).**

B) Yes, because it is a true negative.

*feedback:* **Incorrect. You can only observe a true negative if the null-hypothesis is true. And a p-value can never be used to draw conclusions about the probability of a hypothesis.**

C) Yes, if the *p*-value is larger than the alpha level the null-hypothesis is true.

*feedback:* **Incorrect. It is possible to observe p-values larger than the alpha level when the alternative hypothesis is true (this is known as a false negative). And a p-value can never be used to draw conclusions about the probability of a hypothesis.**

D) No, because you need at least two non-significant p-values to conclude the null-hypothesis is true.

*feedback:* **Incorrect. If the statistical power is low, it is possible to observe two non-significant result when the alternative hypothesis is true. And a *p*-value can never be used to draw conclusions about the probability of a hypothesis.**

**Q11:** What is a correct way to present a non-significant *p*-value (e.g., *p* = 0.34 assuming an alpha level of 0.05 is used in an independent *t*-test)?

A) The null-hypothesis was confirmed, *p* > 0.05

*feedback:* **Incorrect. What’s important to recall here is that *p*-values are a statement about the probability of the data, *not* a statement about the probability of a theory or hypothesis. Thus, a correct interpretation of a non-significant *p*-value is that ‘there was no statistically significant difference’ or ‘there was no difference large enough to yield a *p* < .05.’**

B) There was no difference between the two conditions, *p* > 0.05

*feedback:* **Incorrect. Careful! While it might sound like semantics, stating ‘no difference’ versus ‘no statistically significant difference’ is a small but important distinction. When one simply states that there is ‘no difference’ between conditions, it erroneously implies that it is 100% probable that the null-hypothesis is true. Because *p*-values are statements about the probability of data, and *not* a statement about the probability of a theory or hypothesis, conclusions drawn should be specific to the findings of the data, rather than broad claims about a theory.**

C) The observed difference was not statistically different from 0.

*feedback:* **Correct!**

D) The null hypothesis is true.

*feedback:* **Incorrect: It is possible to observe p-values larger than the alpha level when the alternative hypothesis is true (this is known as a false negative). And a p-value can never be used to draw conclusions about the probability of a hypothesis.**

**Q12:** Does observing a significant *p*-value (*p* < .05) mean that the null-hypothesis is false?

A) No, because *p* < .05 only means that the alternative hypothesis is true, not that the null hypothesis is wrong.

*feedback:* **This is incorrect - when p < .05, you can only conclude that there is a statistically significant difference in your sample. Because *p*-values are never a statement about the probability of a hypothesis or theory, you cannot claim that the alternative hypothesis is true, nor that the null hypothesis is false/wrong.**

B) No, because *p*-values are never a statement about the probability of a hypothesis or theory.

*feedback:* **Correct!**

C) Yes, because an exceptionally rare event has occurred.

*feedback:* **Incorrect: It is possible that the observed p < .05 is a Type 1 error, and thus, that the null-hypothesis is true.**

D) Yes, because the difference is statistically significant.

*feedback:* **Incorrect: It is possible that the observed p < .05 is a Type 1 error, and thus, that the null-hypothesis is true.**

**Q13:** Is a statistically significant effect always a practically important effect?

A) No, because in extremely large samples, extremely small effects can be statistically significant, and small effects are never practically important.

*feedback:* **Incorrect. While extremely large samples can lead to small effects being statistically significant, it would be incorrect to state that small effects are never practically important. A small effect can very well have important impact such as the theoretical example provided above whereby 12 cents saved per person per year (i.e. small effect) could amount to over 2 million if combined across individuals (i.e. large impact). Practical importance, as such, is always a matter of context and should be assessed in terms of a cost-benefit analysis.**

B) No, because the alpha level could in theory be set to 0.20, and in that case a significant effect is not practically important.

*feedback:* **Incorrect: The alpha level only determines the Type 1 error rate, not the importance of the effect.**

C) No, because how important an effect is depends on a cost-benefit analysis, not on how surprising the data is under the null hypothesis.

*feedback:* **Correct!**

D) All of the above are true.

*feedback:* **Incorrect - only one of the answers is correct!**

**Q14:** What is the correct definition of a *p*-value?

A) A *p*-value is the probability of that the null hypothesis is true, given data that is as extreme or more extreme than the data you have observed.

*feedback:* **Incorrect – the *p*-value is not the probability of a theory or hypothesis, but the probability of the data.**

B) A *p*-value is the probability of that the alternative hypothesis is true, given data that is as extreme or more extreme than the data you have observed.

*feedback:* **Incorrect – the *p*-value is not the probability of a theory or hypothesis, but the probability of the data.**

C) A *p*-value is the probability of observing data that is as extreme or more extreme than the data you have observed, assuming the alternative hypothesis is true.

*feedback:* **Incorrect – the *p*-value is calculated assuming the null hypothesis is true.**

D) A *p*-value is the probability of observing data that is as extreme or more extreme than the data you have observed, assuming the null hypothesis is true.

*feedback:* **Correct!**