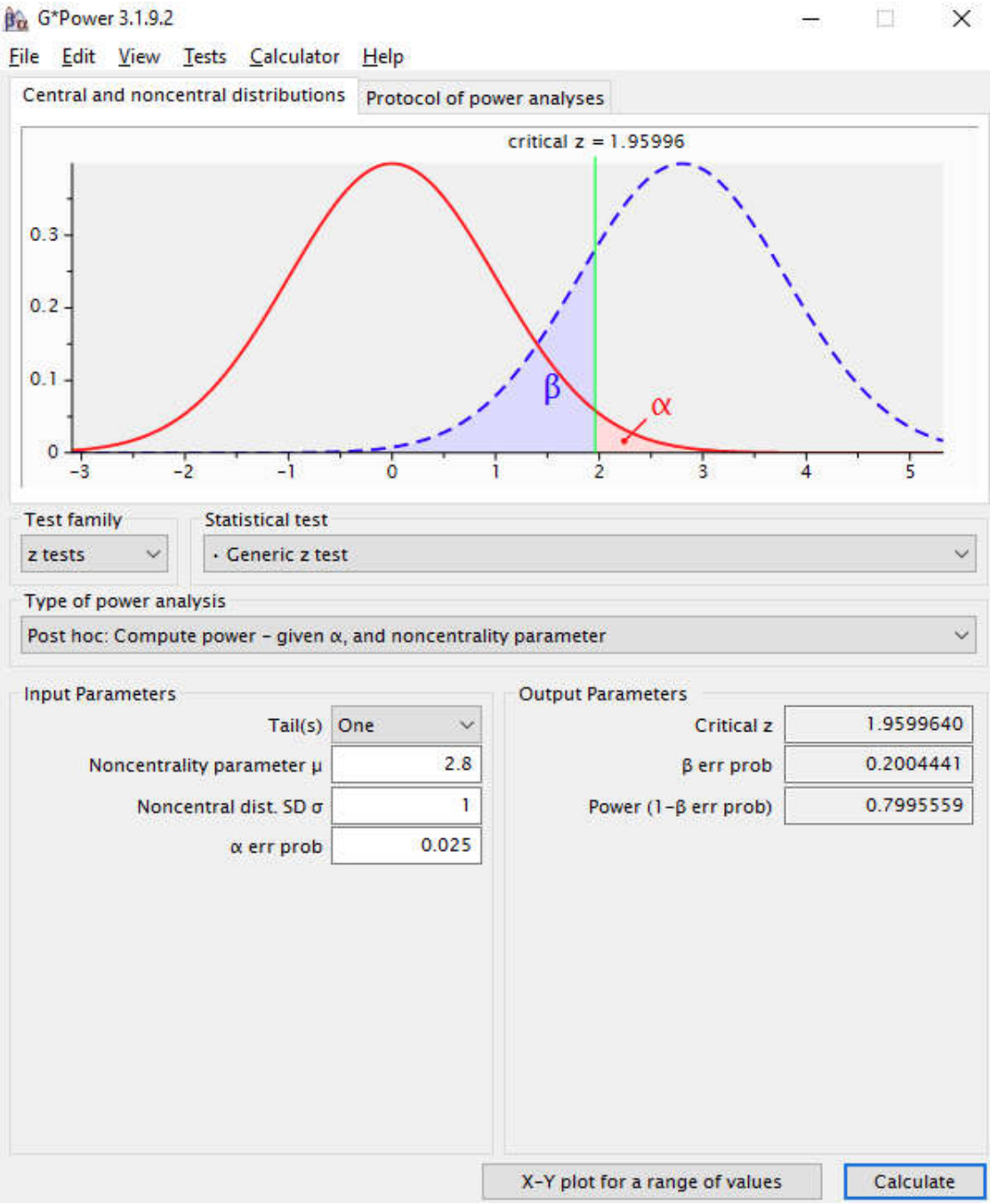
Replicability-Index

Improving the replicability of empirical research

An Attempt at Explaining Null-Hypothesis Testing and Statistical Power with 1 Figure and 1,500 Words

February 26, 2017 Uncategorized

Is a Figure worth 1,500 words?



Gpower. http://www.gpower.hhu.de/en.html

Significance Testing

1. The red curve shows the sampling distribution if there is no effect. Most results will give a signal/noise ratio close to 0 because there is no effect (0/1 = 0)

2. Sometimes sampling error can produce large signals, but these events are rare

3. To be sure that we have a real signal, we can chose a high criterion to decide that there was an effect (reject H0). Normally, we use a 2:1 ratio (z > 2) to do so, but we could use a higher or lower criterion value. This value is shown by the green vertical line in the Figure

4. z-score greater than 2 leaves only 2.5% of the red distribution. This means we would expect only 2.5% of outcomes with z-scores greater than 2 if there is no effect. If we would use the same criterion for negative effects, we would get another 2.5% in the lower tail of the red distribution. Combined we would have 5% of cases where we have a false positive, that is, we decide that there is an effect when there was no effect. This is why we say, p < .05 to call a result significant. The probabilty (p) of a false positive result is no greater than 5% if we keep on repeating studies and using z > 2 as the criterion to claim an effect. If there is never an effect in any of the studies we are doing, we end up with 5% false positive results. A false positive is also called a type-I error. We are making the mistake to infer from our study that an effect is present when there is no effect.

Statistical Power

5. Now that you understand significance testing (LOL), we can introduce the concept of statistical power. Effects can be large or small. For example, gender differences in height are large, gender differences in the number of sexual partners are small. Also studies can have a lot of sampling error or very little sampling error. A study of 10 men and 10 women may accidentally include 2 women who are on the basketball team. A study of 1000 men and women is likely to be more representative of the population. Based on the effect size in the population and sample size, the true signal (effect size in the population) to noise (sampling error) ratio can differ. The higher the signal to noise ratio is, the further away the sampling distribution of the real data (the blue curve) will be. In the figure below the population effect size and sampling error produced a z-score of 2.8, but actual samples will never produce this value. Sampling error will again produce different z-scores above or below the expected value of 2.8. Most samples will produce values close to 2.8, but some samples will produce more extreme deviations. Samples that overestimate the expected value of 2.8 are not a problem because these values are all greater than the criterion for statistical significance. So, in all of these samples we will make the right decision to infer that an effect is present when an effect is present. A so called true positive result. Even if sampling error leads to a small underestimation of the expected value of 2.8, the values can still be above the criterion for statistical significance and we get a true positive result.

6. When sampling error leads to more extreme underestimation of the expected value of 2.8, samples may produce results with a z-score less than 2. Now the result is no longer statistically significant. These cases are called false negatives or type-II errors. We fail to infer that an effect is present, when there actually is an effect (think about a faulty pregnancy test that fails to detect that a woman is pregnant). It does not matter whether we actually infer that there is no effect or remain indecisive about the presence of an effect. We did a study where an effect exists and we failed to provide sufficient evidence for it.

7. The Figure shows the probability of making a type-II error as the area of the blue curve on the left side of the green line. In this example, 20% of the blue curve is on the left side of the green line. This means 20% of all samples with an expected value of 2.8 will produce false negative results.

8. We can also focus on the area of the blue curve on the right side of the green line. If 20% of the area is on the left side, 80% of the area must be on the right side. This means, we have an 80% probability to obtain a true positive result; that is, a statistically significant result where the observed z-score is greater than the criterion z-score of 2. This probability is called statistical power. A study with high power has a high probability to discover real effects by producing z-scores greater than the criterion value. A study with low power has a high probability to produce a false negative result by producing z-scores below the criterion value.

9. Power depends on the criterion value and the expected value. We could reduce the type-II error and increase power in the Figure by moving the green line to the left. As we reduce the criterion to claim an effect, we reduce the area of the blue curve on the left side of the line. We are now less likely to encounter false negative results when an effect is present. However, there is a catch. By moving the green line to the left, we are increasing the area of the red curve on the right side of the red curve. This means, we are increasing the probability of a false positive result. To avoid this problem we can keep the green line where it is and move the expected value of the blue line to the right. By shifting the blue curve to the right, a smaller area of the blue curve will be on the left side of green line.

10. In order to move the blue curve to the right we need to increase the effect size or reduce sampling error. In experiments it may be possible to use more powerful manipulations to increase effect sizes. However, often increasing effect sizes is not an option. How would you increase the effect size of sex on sexual partners? Therefore, your best option is to reduce sampling error. As sampling error decreases, the blue curve moves further to the right and statistical power increases.

Practical Relevance: The Hunger Games of Science: With high power the odds are always in your favor

10. Learning about statistical power is important because the outcome of your studies does not just depend on your expertise. It also depends on factors that are not under your control. Sampling error can sometimes help you to get significance by giving you z-scores higher than the expected value, but these z-scores will not replicate because sampling error can also be your enemy and lower your z-scores. In this way, each study that you do is a bit like playing the lottery or a box of chocolates. You never know how much sampling error you will get. The good news is that you are in charge of the number of winning tickets in the lottery. A study with 20% power, has only 20% winning tickets. The other 80% say, “please play again.” A study with 80% power has 80% winning tickets. You have a high chance to get a significant result and you or others will be able to redo the study and again have a high chance to replicate your original result. It can be embarrassing when somebody conducts a replication study of your significant result and ends up with a failure to replicate your finding. You can avoid this outcome by conducting studies with high statistical power.

11. Of course, there is a price to pay. Reducing sampling error often requires more time and participants. Unfortunately, the costs increase exponentially. It is easier to increase statistical power from 20% to 50% than to increase it from 50% to 80%. It is even more costly to increase it from 80% to 90%. This is what economists call diminishing marginal utility. Initially you get a lot of bang for your buck, but eventually the costs for any real gains are too high. For this reason, Cohen (1988) recommended that researchers should aim for 80% power in their studies. This means that 80% of

your initial attempts to demonstrate an effect will succeed when your hard work in planning and conducting a study produced a real effect. For 20% of the study you may either give up or try again to see whether your fist study produced a true negative result (there is no effect) or a false negative result (you did everything correctly, but sampling error handed you a losing ticket. Failure is part of life, but you have some control over the amount of failures that you encounter.

12. The End. You are now ready to learn how you can conduct power analysis for actual studies to take control your fate. Be a winner, not a loser.