Sampling error: The fundamental flaw of the severity measure of evidence

May 26, 2023

Contents

1 Introduction

Generally speaking, when we estimate a fixed population parameter based on the observation of a sample (i.e., not the whole population), we know that different samples would have generated different estimates. The magnitude of this difference is called the sampling error.

When an estimator is consistent and unbiased, the sampling error can be reduce to an arbitrarily small difference, by increasing the sample size of a study. This is a desirable property since it implies that the estimates over all possible samples will be closer to the truth and less variable. Consequently, when a test statistic is based on a consistent and unbiased estimator, it is always desirable (although not always possible) to increase the sample size and therefore the power of the test in order to reduce the sampling error.

This does not only ensure that our estimates are closer to the truth on average, but that our rejection of the null hypothesis, when it is false, is not caused by the sampling error of our estimates but by a true discrepancy from the null. In this specific context, a significant and more powerful test (i.e., one that uses a greater sample size) will always provide better evidence against the null hypothesis because one major source of error will be reduced: the sampling error.

Assuming a fixed population parameter to estimate and a test statistic based on a consistent and unbiased estimator of that parameter, I demonstrate without any shadow of a doubt that a popular measure of evidence championed by Deborah Mayo and Aris Spanos (the severity measure) is erroneous because of the sampling error. In fact, I show that the greater the sampling error, the greater the error of that measure.

Why am I so confident? Why 'without any shadow of a doubt'? Because I am presenting mathematical facts: 1-Some statistical tests, like one sided t-tests,

are using statistics that are based on coherent and unbiased estimators such as the sample mean (the test statistic is essentially a centered and standardised sample mean in the one sided t-test). 2- We can increase or decrease the level of sampling error associated with the estimates at will by decreasing or increasing the sample size of the experiment. 3- This is equivalent to decreasing or increasing the power of the associated test, at will, by increasing or decreasing the sample size of the experiment. 4- As we decrease the power of the test (decrease the sample size), the only statistics that can reach the critical region under H1 eventually do so only because of the large sampling error and not because of the underlying truth of the matter: the real (usually unknown difference) between H0 and H1. 5- In that scenario, the test statistics become so deviant that they will inevitably corrupt the severity score because the latter is computed with the estimate that contains the large sampling error. This is all beautifully illustrated in the paper.

This result is problematic for at least two reasons. The first reason has to do with the usefulness of the severity measure. That measure of evidence is incomplete. It fails to capture every dimension of what constitute evidence against the null hypothesis. Even with the full knowledge of the adequacy of a model, it is incomplete because it fails to take sampling error into account.

If someone in my team were to claim that we should reject H0 because the severity score is high for some discrepancy, I would immediately reply: did you take into account the sampling error? Would it be possible to take a different sample of the same population or look at previous studies on that population and see if the results are robust?

The second, and more important reason, why the demonstration is problematic has to do with the actions that need to be taken in order to reduce the sampling error. Under specific conditions, I show that it is in our best interest to work with the largest sample size that we can reasonable obtain in order to reduce that source

of error. If the null hypothesis is false, working with the largest possible sample size will give us the best possible evidence against the null hypothesis by reducing the chance that our test statistic reaches the critical region of a test because of the sampling error. It will also improve the reliability of the severity measure of evidence, should we be inclined to use it.

Here is the catch: Mayo and Spanos(hereafter M & S) are well-known for claiming that more powerful tests do not provide better evidence against the null when the test is significant. I show that this is a mistake and that their own measure of evidence cannot allow them to make such a claim because it is less reliable when the power decreases. They simply cannot embrace the idea of improving their measure of evidence by encouraging the use of greater sample sizes and also claim that this will not improve the evidence against the null. Why bother with improving the measure of evidence then?

2 More Power is better when tests are based on consistent and unbiased estimators

In this section, I will discuss a very specific kind of experiment in order to make my point: the one-tailed t-test. The test statistics used in such tests are based on a consistent and unbiased estimator of the mean of the population: the sample mean of independent and normally distributed observations. The test statistics, which will be taken as the evidence against H0 (should the test be significant) is essentially an unbiased and consistent estimator of the mean that is re-centered and standardised. This is the link between 'estimating' and 'making a statistical inference'. The one-tailed t-test combines both theory testing and estimation theory, and this makes sense. T-tests would be of little interest if we did not have a good estimator of the mean in order to make our inferences.

I am not claiming that M&S theory on statistical evidence requires that a test statistics be an unbiased and coherent estimator of the parameter on which we want to make an inference. I am saying that it is the case in this very well-known and basic test. This is just a mathematical fact. From an estimation perspective, it is obvious that it is better to have the largest sample possible in order to reduce the sampling error and to track the truth: the estimator is coherent and unbiased. I am assuming here of course that we want to track the truth in science. I hope this is not controversial.

What seems to be controversial however are the following arguments that are at the heart of this paper: 1- When we perform a t-test, we also want our test statistics to track the truth by reducing that very same sampling error as much as possible. There would be an epistemic/logical incoherence if it were better to increase the sample size to improve the estimation of a population parameter and yet not desirable to increase the sample size (and therefore the power of a test) in order to make a statistical inference about the same parameter. Consequently, the best possible t-test will be the most powerful test possible. When such as test is significant, it will provide the best possible evidence against H0. 2- In this scenario, claiming that a less powerful test can provide better evidence against the null than a more powerful one, is tantamount to saying that we can obtain better better evidence against H0 simply by increasing the sampling error of our estimates. This is ridiculous at best.

In order to avoid any unnecessary complications and useless debates, please remember the following claim: the independence and normality of all the observations of every experiment discussed in this paper is a given. It is known by everyone. There is no doubt possible about the i.i.d. and normality of the observation. The model used in every experiment mentioned in this paper is never misspecified. I make sure of it since I control the simulations (the reader can also have control

since the code is in the Appendix, the box is not black it is open for everyone to play with). If the severity score fails in this generous situation, it will fail harder in more complex and realistic scenarios where this knowledge is unavailable.

In this special context, the sample size does not need to be above 20 or some other arbitrary number in order to satisfy the hypothesis of normality. Do not worry about it. It is your lucky day. The normality is a God given knowledge here. The sample size could be 9 and the test would be valid. The sample size will however determine the sampling error of the estimator/test-statistics. The smaller the sample size, the greater the sampling error.

Moreover, please keep in mind that when I mention the concept of power like in the statement "More power is better because it reduces the sampling error of a consistent and unbiased estimator" I mean the probability of rejecting the null hypothesis when it is false. I am not concerned at all with so-called pre-data power calculations. It is an important aspect of research, but not one that is relevant for this particular paper. When the sample size of the experiments increases, its power (the real and most likely unknown power) increases and the sampling error decreases. That is all that matters in this paper.

2.1 The demonstration

Here is the scenario: we have a population with a true and unknown fixed mean of 1.2. We aim to test if the mean of the population is strictly greater than 1 with a one-tailed t-test. We know that any given sample will be comprised of independent observations that are generated by the same normal distribution. In reality we would be lucky to have this knowledge. If the severity measure fails in this simple and advantageous scenario, it will not do better when we have less than absolute knowledge about the i.i.d. nature of the observations.

We do not aim to detect a difference of 0.2 (the true difference between the mean under H0 and the mean under H1). We do not perform a pre-data power calculation to determine the appropriate sample size that would give us the best chance possible of detecting a difference of 0.2. As far as we are concerned, the true difference could be as large as 500 or it could be 0. We simply do not know. We want to know if H0 is false.

Now here is the question, which sample size will give us the best evidence against H0, should we obtain a significant test? 9, 64, 169, 324, 1089, 2304, 4624, 8649? The answer lies in the distribution of the estimates that we will obtain given that the test is significant. Those distributions are presented in Figure 1. The first boxplot on the left illustrates the distribution of all the estimates of the means that triggered a significant result over 500,000 samples of size 9. All the other boxplots represent the same thing except that the sample size gradually increases. The percentage displayed in the graph is the percentage of the 500,000 tests that actually triggered a significant result. As the sample size increases and therefore the power, that percentage increases.

Here is what Figure 1 shows: As the power of the test decreases, we see that the observed means lift off from the ground truth (1.2) and that the variance of the observed means increases. This is called the inflation of the effect size. It is uniquely caused by the increasing sampling error that is introduced as the sample size and the power of the test decreases. You can recreate this figure with the code bellow. The code contains a seed, but the results on the graph are so robust that you can change it without any significant differences.

Conversely, as the power increases, the observed means are less variable and closer to the ground truth because the estimator is unbiased and consistent and because more of them fall withing the critical region of the test as the impact of the sampling error because less and less significant. In fact, the boxplot on the

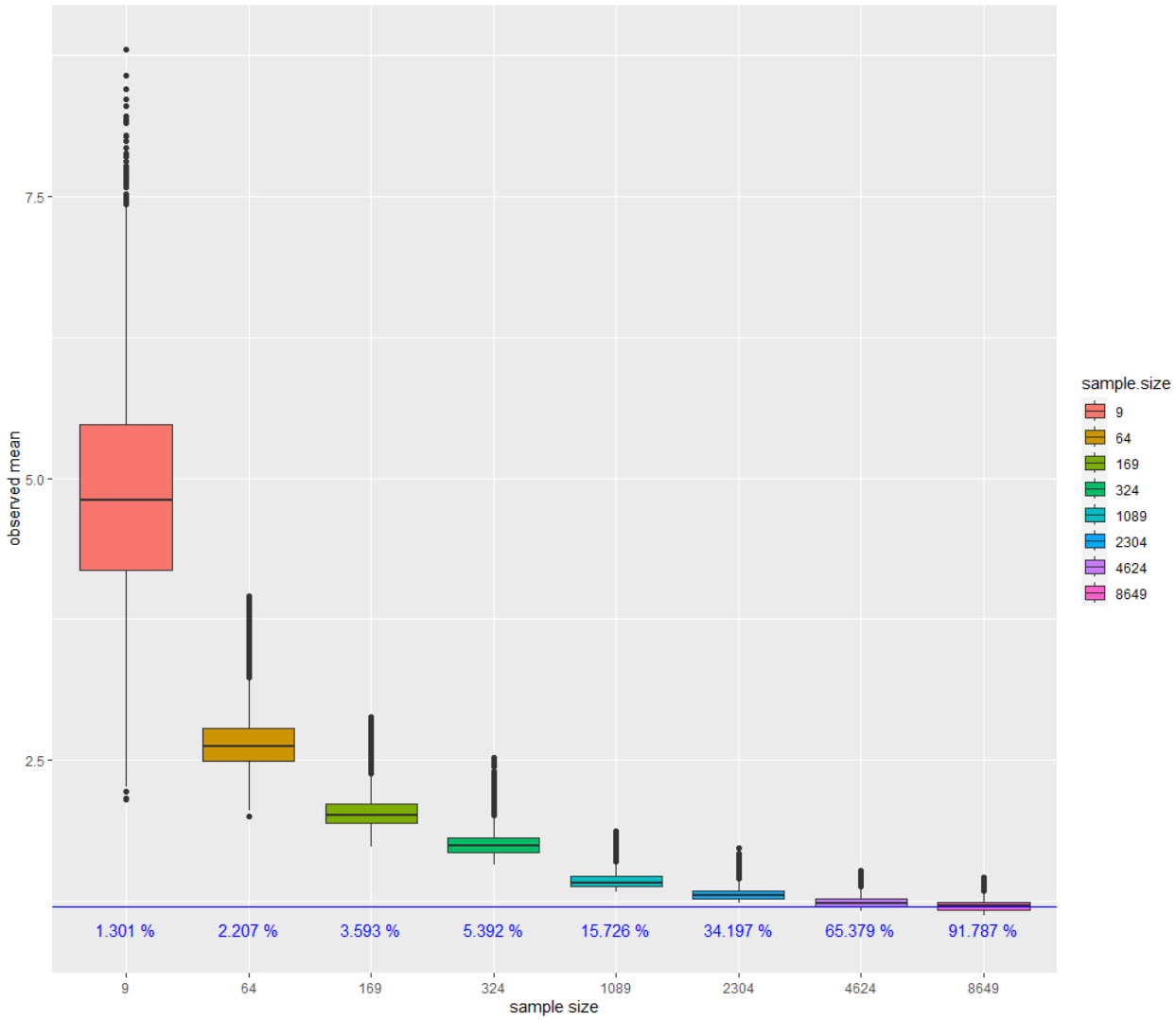

Figure 1: Each boxplot describes the observed distribution of every sample means that triggered a significant (p-value bellow 0.01) result under H1 over 500,000 samples of a given size specified on the x-axis. The blue horizontal line is at 1.2 which is the true unknown and fixed mean of the population that we estimate. The percentage represents the percentage of the 500,000 samples that generated a significant result under H1.

right side is right on the money. Estimates with that sample size provide the best evidence against H0 out of all the other estimates obtained with smaller sample sizes. Results are robust: we obtain a significant result almost 92 % of the time and each estimates are closer to the truth of 1.2. What else could we want from a statistical point of view? Clearly, the more power (the greater the sample size) the better evidence against the null. The answer to the question is 8649.

But let us see what the severity score tells us. I call it a score, because it refers to a function that outputs a real number between 0 and 1 and that number is meant to tell us if we have good evidence for the existence of a particular range of discrepancy from the null given an observed test statistics. In the case at hand, we could ask if a significant test statistic warrants the existence of difference strictly greater than 0 or any other real number greater than 0. The score is computed by evaluating the cumulative distribution of a Student's law evaluated at a point of interest, with the observed statistic.

Within each boxplot let us take the median observed mean (or the one that is the closest to the median) and let us see what happens when we ask the severity measure if we have good evidence for a difference that is strictly greater than the truth of 0.2 (see Appendix to reproduce the results). Let us put the following claim to the test:

> **Severity Principle (full)**. Data x_0 (produced by process G) provides good evidence for hypothesis H (just) to the extent that test T severely passes H with x_0. (Mayo and Spanos 2011, p.162).

For each respective sample sizes on Figure 1, starting with 9 and ending with 8649, we obtain the following severity scores: 0.9939735, 0.9844467, 0.9900222, 0.9742364, 0.9229936, 0.8304057, 0.6731266, 0.540871. It would appear that would have very good evidence for a difference from the null that is strictly greater

than 0.2 with a low powered test. Yet, this is incorrect: the real difference is not strictly greater than 0.2. The problem is that the severity score is computed with the sampling error. The greater that error, the more erroneous its results.

There is no way around it: this is the demonstration promised in the introduction. Figure 1 shows, without any shadow of a doubt, that the sampling error of lower powered tests inflate the observed statistics when the test is significant and will also inflate the severity score by the very nature of its mathematical form: the cumulative distribution of a student law evaluated at a given point of interest with the test statistics! It is the fact that the score is calculated with the observed statistics that makes all the difference because the observed statistics contains sampling error.

If that score has to have any chance of being correct, we need to keep the sampling error to a bare minimum and, with the kind of scenario discussed here, this can only be done by increasing the power of the test. Increasing the power will not only produce better evidence against H0 by producing estimates that are less variable and less biased when the test is significant, it will also reduce the pervasive effect of the sampling error on the severity measure of evidence. The problem is that M & S refuse to acknowledge that more powerful test provide better evidence against the null and the reasons for this are misguided. How can they claim that more powerful tests do not provide better evidence against the null when their measure of evidence fails to be adequate as the power decreases (see Figure 1)?

They claim that there is a mistaken view "wherein an α level rejection is taken as more evidence against the null, the higher the power of the test" (Mayo and Spanos 2006, p.344) In a recent article, Spanos even claims that a less powerful test, of the kind discussed here, provides better evidence against the null because the observed statistics are larger and the confidence intervals are further away from the parameter under the null hypothesis (Spanos 2022, p.16), but we now know

that this is a mistake. Estimations will inflate, and so will the confidence intervals, as the sample size decreases when a test is significant.

Of course, very powerful test will be able to detect very small discrepancies but this has nothing to do with the quality of the evidence against the null. If I claim to have no food in my refrigerator, it does not make it better evidence against that claim if someone finds two apples instead of one. It is a mistake to equate the magnitude of a different with the quality of the evidence for that difference. It seems that M & S make that mistake.

Some things have to change. First, we need to acknowledge that the severity measure of evidence is incomplete because sampling error needs to be taken into account when measuring the quality of the evidence against the null: more sampling error means less evidence against the null. Second, M & S's view according to which more powerful tests do not provide better evidence against the null when tests are significant is false. I have given a compelling counterexample with Figure 1.

3 Conclusion

In conclusion, I would like to address eight common objections to the theses I have been presenting in this paper. Not all of them are brilliant, but they sure seem to have convinced brilliant people. Please pay close attention to Objection 6 and my reply. I believe it is the main point of contention with M&S. They strongly believe that my claim to the effect that more powerful significant tests provide better evidence against the null is a fallacy.

Objection 1: M&S never claimed that parameter estimation is part of their approach and that parameter estimates should be unbiased when conditioned on statistically significant results.

Reply to Objection 1: They do not need to claim such a thing. Their approach applies to t-tests. T-tests are based on unbiased estimators. Those estimators will generated estimates with large sampling error when the sample size is low. Those estimates will trigger significant results every now and then (at lest with probability alpha (the degree of a test). When they do, they do not provide good evidence against the null because they are artifact of the the sampling error. They also corrupt the severity score because it is computed with the estimates that contain, in their magnitude, the sampling error. These facts are at the foundation of this paper.

Objection 2: M&S do not claim that their severity measure if a measure of evidence.

Reply to objection 2: When someone claims that they have a function that can determine if an observation is good evidence for a given hypothesis, I call that a 'measure of evidence'. The quote I give in the paper to that effect speaks for itself.

Objection 3: Figure 1 seems to show that the severity score is behaving the way it should. When the power is low, we obtain greater score than when the power is high. I fail to see the problem.

Reply to Objection 3: Yes, it behaves the way it is meant to. I argued that this way is misguided. The score for the claim "there is a discrepancy strictly greater than 0.2" was very high with a sample size of 9 and low for a sample size of 8649. The only difference between the two scenarios is that the sampling error is greater with a sample size of 9. Conclusion: we need to increase the sample size of our tests (i.e. increase their power) if we want to reduce the sampling error, obtain a more reliable severity score, and better evidence against the null hypothesis.

Objection 4: Looking at Figure 1, I understand that someone would have been lucky to obtain a significant result with only 9 observations. But it's OK to be lucky. One can have lucky evidence against H0.

Reply to Objection 4: The problem is not so much that the observation of a sig-

nificant result is lucky (although that is a problem when you cannot replicate your results). The problem is the inflation of the estimates. They corrupt the severity score by making it totally unreliable. There is no luck in a measure telling you that you have good evidence that there is a greater discrepancy from the null than the true one.

Objection 5: What you, the author of this paper, have shown here is that if we make the statistical tests unreliable by decreasing the power (pre-data), then the severity measure is false (post-data). This is not a problem for the (post-data) severity measure of evidence. It is good if we assume that the test procedure is reliable to begin with (pre-data).

Reply to Objection 5: Sampling error is a property of the data and in order for the severity measure to be reliable, we need to reduce the sampling error to a minimum by increasing the sample size and the power to a maximum (within operational constraints). This will also generate better evidence against the null when it is false (assuming a test based on an unbiased and consistent estimator). The problem is that M & S's views, expressed in the quotes given above, are incompatible with this solution. They simply refuse to acknowledge that a more powerful test will provide better evidence against the null. It would not make sense for them to systematically encourage the usage of greater sample sizes. They would have to claim that greater sample sizes are better for the reliability of the severity measure of evidence but that the evidence against the null will not be better. Why bother then?

Objection 6: The paper argues for the following claims: If there is a high probability that test T will reject H0 when $\mu = \mu'$ for μ' , a value in H1 — that is, if the power of test T against $\mu = \mu'$ is high—then observing a sample mean M is good evidence that μ is as large as μ' (where $\mu' = \mu 0 + k$ standard errors (SE).) The flaw in this paper's claim can be seen with an entirely informal example.

Suppose one is testing if a treatment yields 0 benefit versus various magnitudes of positive benefit, measured in terms of the percentage of patients cured. Suppose the test is practically guaranteed to reject H0. In fact, H1, the drug cures practically everyone. The test has high power to detect H1. Bt merely finding a statistically significant M does not warrant H1.

Reply to Objection 6: I do not make such a foolish claim. The paper never argues for "the following claims:" It will not be so easy to dismiss the result shown in this paper. I do not believe that the result of a powerful test will provide good evidence for a grandiose interpretation of H1 such as "the drug practically cures everyone" and I certainly do not want to encourage this belief.

However, I believe that the result of such a test will provide great evidence against H0 which says that the drug has no effect. Please do not interpret my work in such a way that I would endorse the claim "the drug will cure everyone". I do not understand why anyone would do that anyway in good faith. Nowhere in this paper have I claimed that the result of a statistical test can provide evidence for a wide range of different discrepancies from the null. I'm making a point about the evidence for rejecting H0, not for accepting any kind of interpretation of H1.

The reason for this should be clear by now. I advocate for the highest possible sample size in order to obtain the best evidence against the null when the test is significant. When the test statistic is based on a consistent and unbiased estimator, it will track the truth more efficiently, as shown in Figure 1. If the truth points towards the existence of a small discrepancy from the null, then so be it. That is the discovery.

I am certainly not going to be the advocate for outlandish discrepancies that are incompatible with my estimate and I am certainly not going to reach any conclusion if I suspect that my estimate is corrupted with a large sampling error. Objection 6 is basically trying to turn the tables on my own arguments. The fact is that

severity score will warrant the existence of ridiculously large discrepancies when the estimate contains a large sampling error. This is what I have shown in this paper. Again, this is a mathematical fact. The only way around it is to reduce the sampling error by increasing the sample size and the power of the test.

Objection 7: You, the author of this paper, claim that more powerful tests provide better evidence against the null but consider this scenario: H0: the urn contains 0 yellow balls, 95 red balls and 5 blue balls. H1: The urn contains 1 yellow ball, 94 red balls and 5 blue balls. The rejection rule is: if you observe a blue or yellow ball after one draw, reject H0. The probability of rejecting H0 when true is 5% and the probability of rejecting H0 when H1 true is 5%. The power is very low. However, imagine that you observe a yellow ball, you would have conclusive evidence against H0 even if the power is small. You can change the scenario just a little bit, increase the power slightly, and it will have no impact on the quality of the evidence when you observe a yellow ball.

Reply to Objection 7: This example relies on the fact that the parameter space is not the same under H0 and H1. There is no "yellow" category under H0. Therefore, the observation of a yellow ball is not even in the critical region. If this kind of thought experiment has any usefulness here is for me to make the following caveat: my conclusions apply only for tests that are such that their parameter space is the same under H0 and H1 and that are based on consistent and unbiased estimators.

Objection 8: The references are skimpy.

Reply to Objection 8: That's all I need to make my point.

4 Appendix

######## t t e s t u n d e r H1##

```
library("tidyverse")
library("ggplot2")
library("data.table")
set.seed(1829345)
simulations<-list()
significant_results<-list()
sim.size<-500000
estimate<-NA
stderr<-NA
pvalue<-NA
sample.size<-((seq(3, 100, by=5))**2)[c(1, 2, 3, 4, 7, 10, 14, 19)]
for(n in sample.size){
  for (i in 1:sim.size){
    obs<-rnorm(n, 1.2, sd=5)
    test<-t.test(obs, mu=1, alternative='greater')
    estimate[i]<-test$estimate
    stderr[i]<-test$stderr
    pvalue[i]<-test$p.value
      }
  simulations[[which(sample.size==n)]]<-list(estimate, stderr, pvalue)
  estim<-(simulations[[which(sample.size==n)]][[3]]<=0.01) *
  simulations[[which(sample.size==n)]][[1]]
  estim<-replace(estim,estim==0,NA)
  err<-(simulations[[which(sample.size==n)]][[3]]<=0.01) *
  simulations[[which(sample.size==n)]][[2]]
  err<-replace(err,err==0,NA)
```

```
  significant_results[[which(sample.size==n)]]<-list(estim, err)
  print(n)
}
df <-as.data.frame(t(sapply(significant_results,"[[",1)))
df$sample.size <- as.factor(sample.size)
percent<-apply(apply(df[, -(sim.size+1)], 1, is.na), 2,
function(x) 100-((sum(x)*100)/(sim.size)))
percent<-as.data.frame(t(rbind(percent,sample.size)))
df_long <- melt(as.data.table(df), id = ('sample.size'))
ggplot() +
  geom_boxplot(data=na.omit(df_long), aes(x = sample.size,
  y = value, fill= sample.size))+
  geom_hline(yintercept=1.2, color = "blue") +
  geom_text(aes(x=as.factor(percent$sample.size), y=2,
  label=paste(round(percent$percent, 3), "%")), nudge_y=-1, color='blue')+
  labs(x="sample size",y="observed mean")
estim_var<-apply(df[, -(sim.size+1)], 1, var, na.rm=T)
estim_var
df_err <-as.data.frame(t(sapply(significant_results,"[[",2)))
fun<-function(x){
  x<-x[!is.na(x)]
  x[which.min(abs(x-median(x)))]
}
mid_estim<-apply(df[, -(sim.size+1)], 1, fun)

n=1.2
```

```
pt((mid_estim[1]-n)/df_err[1, which(df[1,]==mid_estim[1])],
8, lower.tail = TRUE)
pt((mid_estim[2]-n)/df_err[2, which(df[2,]==mid_estim[2])],
63, lower.tail = TRUE)
pt((mid_estim[3]-n)/df_err[3, which(df[3,]==mid_estim[3])],
168, lower.tail = TRUE)
pt((mid_estim[4]-n)/df_err[4, which(df[4,]==mid_estim[4])],
323, lower.tail = TRUE)
pt((mid_estim[5]-n)/df_err[5, which(df[5,]==mid_estim[5])],
1088, lower.tail = TRUE)
pt((mid_estim[6]-n)/df_err[6, which(df[6,]==mid_estim[6])],
2303, lower.tail = TRUE)
pt((mid_estim[7]-n)/df_err[7, which(df[7,]==mid_estim[7])],
4623, lower.tail = TRUE)
pt((mid_estim[8]-n)/df_err[8, which(df[8,]==mid_estim[8])],
8648, lower.tail = TRUE)
```

References

Mayo, D. G. and A. Spanos (2006). Severe testing as a basic concept in a Neyman–Pearson philosophy of induction. *The British Journal for the Philosophy of Science 57*(2), 323–357.

Mayo, D. G. and A. Spanos (2011). Error statistics. *Philosophy of statistics 7*, 152–198.

Spanos, A. (2022). Severity and trustworthy evidence: Foundational problems vs. misuses of frequentist testing. *Philosophy of Science*, 1–31.