Math 3081 (Probability and Statistics)

Lecture #21 of 27 $\sim$ August 10th, 2021

Errors and Misuses of Hypothesis Testing

- Type I and Type II Errors
- Statistical Power
- Use and Misuse of Hypothesis Tests

This material represents §4.1.4 from the course notes, and problems 12-15 from WeBWorK 6.

Type I and Type II Errors, I

When we perform a hypothesis test, there are two possible outcomes (reject H_0 or fail to reject H_0).

- The correctness of the result depends on the actual truth of H_0: if H_0 is false then it is correct to reject it, while if H_0 is true than it is correct not to reject it.
- The other two situations, namely "rejecting a correct null hypothesis" and "failing to reject an incorrect null hypothesis" are refered to as hypothesis testing errors.

Type I and Type II Errors, II

Since these two errors are very different, we give them very different names:

Definition

If we are testing a null hypothesis H_0, we commit a type I error if we reject H_0 when H_0 was actually true. We commit a type II error if we fail to reject H_0 when H_0 was actually false.

We usually summarize these errors with a small table:

H_0 \ Result	Fail to Reject H_0	Reject H_0
H_0 is true	Correct Decision	Type I Error
H_0 is false	Type II Error	Correct Decision

Type I and Type II Errors, III

We would like, in general, to minimize the probabilities of making a type I or type II error.

- The probability of committing a type I error is the significance level α of the test, since by definition this is the probability of rejecting the null hypothesis when it is actually true.
- The probability of committing a type II error is denoted by β. This value is more difficult to calculate, since it will depend on the actual nature in which H_0 is false.
- If we postulate the actual value of the test statistic, we can calculate the probability of committing a type II error.

Type I and Type II Errors, IV

Example: Old curriculum scores are normally distributed with mean 200 and standard deviation 20. New curriculum scores are normally distributed with mean μ and standard deviation 20. The hypotheses are $H_0 : \mu = 200$ and $H_a : \mu > 200$ with a sample of 400 students. H_0 will be rejected if the sample mean $\hat{\mu} > 202$, corresponding to a significance level $\alpha = 2.275\%$.

1. Find the probability of making a type I error.
2. Find the probability of making a type II error if the true mean is actually μ, in terms of μ.
3. Evaluate the type-II error probability for $\mu = 201, 202, 203, 204, 205$.

Type I and Type II Errors, IV

Example: Old curriculum scores are normally distributed with mean 200 and standard deviation 20. New curriculum scores are normally distributed with mean μ and standard deviation 20. The hypotheses are $H_0 : \mu = 200$ and $H_a : \mu > 200$ with a sample of 400 students. H_0 will be rejected if the sample mean $\hat{\mu} > 202$, corresponding to a significance level $\alpha = 2.275\%$.

1. Find the probability of making a type I error.
2. Find the probability of making a type II error if the true mean is actually μ, in terms of μ.
3. Evaluate the type-II error probability for $\mu = 201, 202, 203, 204, 205$.

- Note that the type-I error probability is just the significance level $\alpha = 0.02275$.

Type I and Type II Errors, V

Example: Old curriculum scores are normally distributed with mean 200 and standard deviation 20. New curriculum scores are normally distributed with mean μ and standard deviation 20. The hypotheses are $H_0 : \mu = 200$ and $H_a : \mu > 200$ with a sample of 400 students. H_0 will be rejected if the sample mean $\hat{\mu} > 202$.

2. Find the probability of making a type II error if the true mean is actually μ, in terms of μ.

Type I and Type II Errors, V

Example: Old curriculum scores are normally distributed with mean 200 and standard deviation 20. New curriculum scores are normally distributed with mean μ and standard deviation 20. The hypotheses are $H_0 : \mu = 200$ and $H_a : \mu > 200$ with a sample of 400 students. H_0 will be rejected if the sample mean $\hat{\mu} > 202$.

2. Find the probability of making a type II error if the true mean is actually μ, in terms of μ.

- We want to calculate the probability of failing to reject the null hypothesis when it is false.
- Under the assumption given, the sample mean $\hat{\mu}$ will be normally distributed with mean μ and standard deviation $20/\sqrt{400} = 1$.
- Then, the probability of failing to reject the null hypothesis is $P(N_{\mu,1} \leq 202) = P(N_{0,1} \leq 202 - \mu)$, in terms of μ.

Type I and Type II Errors, VI

Example: Old curriculum scores are normally distributed with mean 200 and standard deviation 20. New curriculum scores are normally distributed with mean μ and standard deviation 20. The hypotheses are $H_0 : \mu = 200$ and $H_a : \mu > 200$ with a sample of 400 students. H_0 will be rejected if the sample mean $\hat{\mu} > 202$.

3. Evaluate the type-II error probability for $\mu = 201, 202, 203, 204, 205$.

Type I and Type II Errors, VI

Example: Old curriculum scores are normally distributed with mean 200 and standard deviation 20. New curriculum scores are normally distributed with mean μ and standard deviation 20. The hypotheses are $H_0 : \mu = 200$ and $H_a : \mu > 200$ with a sample of 400 students. H_0 will be rejected if the sample mean $\hat{\mu} > 202$.

3. Evaluate the type-II error probability for $\mu = 201, 202, 203, 204, 205$.

- For $\mu = 201$ we get $P(N_{201,1} \leq 202) = P(N_{0,1} \leq 1) = 0.8413$.
- For $\mu = 202$ we get $P(N_{202,1} \leq 202) = P(N_{0,1} \leq 0) = 0.5$.
- For $\mu = 203$ we get $P(N_{203,1} \leq 202) = P(N_{0,1} \leq -1) = 0.1587$.
- For $\mu = 204$ we get $P(N_{204,1} \leq 202) = P(N_{0,1} \leq -2) = 0.02275$.
- For $\mu = 205$ we get $P(N_{205,1} \leq 202) = P(N_{0,1} \leq -3) = 0.00135$.

Type I and Type II Errors, VII

We can see that as the true mean gets further away from the mean predicted by the null hypothesis, the probability of making a type II error drops.

- The idea here is quite intuitive: the bigger the distance between the true mean and the predicted mean, the better our hypothesis test will be better at picking up the difference between them.

If we use the same rejection rule, but instead vary the sample size, the probability of making either type of error will change.

Type I and Type II Errors, VIII

Example: The school wants to gather more data on the new curriculum. Assume as before the scores with the old curriculum are normally distributed with mean 200 and standard deviation 20, and the new curriculum scores also have standard deviation 20. We again test $H_0 : \mu = 200$ against $H_a : \mu > 200$ and reject H_0 if $\hat{\mu} > 202$.

1. Find the probability of a type I error in terms of n.
2. Find the probability of a type I error for $n = 100, 400, 1600$.
3. Find the probability of a type II error in terms of n if the true mean is $\mu = 203$.
4. Find the probability of a type II error for $n = 100, 400, 1600$ if the true mean is $\mu = 203$.

Type I and Type II Errors, IX

Example: Scores with the old curriculum are normally distributed with mean 200 and standard deviation 20, and the new curriculum scores also have standard deviation 20. We again test $H_0 : \mu = 200$ against $H_a : \mu > 200$ and reject H_0 if $\hat{\mu} > 202$.

1. Find the probability of a type I error in terms of n.

Type I and Type II Errors, IX

Example: Scores with the old curriculum are normally distributed with mean 200 and standard deviation 20, and the new curriculum scores also have standard deviation 20. We again test $H_0 : \mu = 200$ against $H_a : \mu > 200$ and reject H_0 if $\hat{\mu} > 202$.

1. Find the probability of a type I error in terms of n.

- To find the probability of a type I error, we assume the null hypothesis is correct, so that $\mu = 200$.
- Then the sample mean $\hat{\mu}$ is normally distributed with mean 200 and standard deviation $\sigma = 20/\sqrt{n}$
- Thus, the probability $P(N_{200,20/\sqrt{n}} > 202) = P(N_(0,1) > \sqrt{n}/10)$.

Type I and Type II Errors, X

Example: Scores with the old curriculum are normally distributed with mean 200 and standard deviation 20, and the new curriculum scores also have standard deviation 20. We again test $H_0 : \mu = 200$ against $H_a : \mu > 200$ and reject H_0 if $\hat{\mu} > 202$.

2. Find the probability of a type I error for $n = 100, 400, 1600$.

Type I and Type II Errors, X

Example: Scores with the old curriculum are normally distributed with mean 200 and standard deviation 20, and the new curriculum scores also have standard deviation 20. We again test $H_0 : \mu = 200$ against $H_a : \mu > 200$ and reject H_0 if $\hat{\mu} > 202$.

2. Find the probability of a type I error for $n = 100, 400, 1600$.

- The probability is $P(N_{200,20/\sqrt{n}} > 202) = P(N_{0,1} > \sqrt{n}/10)$.
- For $n = 100$ this is $P(N_{200,2} > 202) = P(N_{0,1} > 1) = 0.15866$.
- For $n = 400$ this is $P(N_{200,1} > 202) = P(N_{0,1} > 2) = 0.02275$.
- For $n = 1600$ this is $P(N_{200,0.5} > 202) = P(N_{0,1} > 4) = 0.0000316 = 3.16 \cdot 10^{-5}$.

Type I and Type II Errors, XI

Example: Scores with the old curriculum are normally distributed with mean 200 and standard deviation 20, and the new curriculum scores also have standard deviation 20. We again test $H_0 : \mu = 200$ against $H_a : \mu > 200$ and reject H_0 if $\hat{\mu} > 202$.

3. Find the probability of a type II error in terms of n if the true mean is $\mu = 203$.

Type I and Type II Errors, XI

Example: Scores with the old curriculum are normally distributed with mean 200 and standard deviation 20, and the new curriculum scores also have standard deviation 20. We again test $H_0 : \mu = 200$ against $H_a : \mu > 200$ and reject H_0 if $\hat{\mu} > 202$.

3. Find the probability of a type II error in terms of n if the true mean is $\mu = 203$.

- By hypothesis, the sample mean $\hat{\mu}$ will now be normally distributed with mean 203 and standard deviation $\sigma = 20/\sqrt{n} = 1$.
- Thus, the probability of a type II error is $P(N_{203,20/\sqrt{n}} \leq 202) = P(N_{0,1} \leq -\sqrt{n}/20)$.

Type I and Type II Errors, XII

Example: Scores with the old curriculum are normally distributed with mean 200 and standard deviation 20, and the new curriculum scores also have standard deviation 20. We again test $H_0 : \mu = 200$ against $H_a : \mu > 200$ and reject H_0 if $\hat{\mu} > 202$.

4. Find the probability of a type II error for $n = 100, 400, 1600$ if the true mean is $\mu = 203$.

Type I and Type II Errors, XII

Example: Scores with the old curriculum are normally distributed with mean 200 and standard deviation 20, and the new curriculum scores also have standard deviation 20. We again test $H_0 : \mu = 200$ against $H_a : \mu > 200$ and reject H_0 if $\hat{\mu} > 202$.

4. Find the probability of a type II error for $n = 100, 400, 1600$ if the true mean is $\mu = 203$.

- The probability of a type II error is $P(N_{203,20/\sqrt{n}} \leq 202) = P(N_{0,1} \leq -\sqrt{n}/20)$.
- For $n = 100$ this is $P(N_{203,2} \leq 202) = P(N_{0,1} \leq -0.5) = 0.30853$.
- For $n = 400$ this is $P(N_{203,1} \leq 202) = P(N_{0,1} \leq -1) = 0.15866$.
- For $n = 1600$ this is $P(N_{203,0.5} \leq 202) = P(N_{0,1} \leq -2) = 0.02275$.

Type I and Type II Errors, XIII

We saw in an example last time that as the true mean gets further away from the mean predicted by the null hypothesis, the probability of making a type II error drops.

- The idea here is quite intuitive: the bigger the distance between the true mean and the predicted mean, the better our hypothesis test will be better at picking up the difference between them.

We also saw in another example that if we increase the sample size (but keep the rejection rule the same) the probabilities of both types of errors will drop.

- The idea again is quite intuitive: the more data we have, the closer our results will be to reality in all situations.

If we fix the significance level α but increase the sample size, the probability of a type II error will change (try guessing how).

Type I and Type II Errors, XIV

Example: The school wants to determine how large a sample size would have been necessary to determine the effectiveness of the new curriculum. Assume the scores with the old curriculum are normally distributed with mean 200 and standard deviation 20, and the new curriculum scores are normally distributed with true mean 203 and standard deviation 20. We test $H_0 : \mu = 200$ against $H_a : \mu > 200$ at the 1% significance level.

1. Find the critical value in terms of n.
2. Find the probability of a type II error in terms of n.
3. Find the probability of a type II error for $n = 100, 400, 900, 1600$.

Type I and Type II Errors, XV

Example: The old curriculum scores are normally distributed with mean 200 and standard deviation 20, and the new curriculum scores are normally distributed with true mean 203 and standard deviation 20. We test $H_0 : \mu = 200$ against $H_a : \mu > 200$ at the 1% significance level.

1. Find the critical value in terms of n.

Type I and Type II Errors, XV

Example: The old curriculum scores are normally distributed with mean 200 and standard deviation 20, and the new curriculum scores are normally distributed with true mean 203 and standard deviation 20. We test $H_0 : \mu = 200$ against $H_a : \mu > 200$ at the 1% significance level.

1. Find the critical value in terms of n.

- What we're looking for is the smallest possible sample mean that would cause us to reject the null hypothesis.
- Under the assumptions of the hypothesis test, $\hat{\mu}$ is normally distributed with mean 200 and standard deviation $\sigma = 20/\sqrt{n}$.
- Since the test is one-tailed, the critical value of $\hat{\mu}$ is the value c such that $P(N_{200,20/\sqrt{n}} > c) = 0.01$.
- Equivalently, this says $P(N_{0,1} > \frac{c-200}{20/\sqrt{n}}) = 0.01$.
- Using a z-table, we can see this occurs when $\frac{c-200}{20/\sqrt{n}} = 2.3263$ and thus $c = 200 + 2.3263 \cdot 20/\sqrt{n}$.

Type I and Type II Errors, XVI

Example: The old curriculum scores are normally distributed with mean 200 and standard deviation 20, and the new curriculum scores are normally distributed with true mean 203 and standard deviation 20. We test $H_0 : \mu = 200$ against $H_a : \mu > 200$ at the 1% significance level.

2. Find the probability of a type II error in terms of n.

Type I and Type II Errors, XVI

Example: The old curriculum scores are normally distributed with mean 200 and standard deviation 20, and the new curriculum scores are normally distributed with true mean 203 and standard deviation 20. We test $H_0 : \mu = 200$ against $H_a : \mu > 200$ at the 1% significance level.

2. Find the probability of a type II error in terms of n.

- In reality, the sample mean $\hat{\mu}$ is normally distributed with mean 203 and standard deviation $\sigma = 20/\sqrt{n}$.
- From the last slide, we reject the null hypothesis if the sample mean is greater than $c = 200 + 2.3263 \cdot 20/\sqrt{n}$.
- Therefore, the probability of a type II error is $P(N_{203,20/\sqrt{n}} \le c) = P(N_{0,1} \le 2.3263 - \dfrac{3\sqrt{n}}{20})$, after doing a bit of simplifying.

Type I and Type II Errors, XX

Example: The old curriculum scores are normally distributed with mean 200 and standard deviation 20, and the new curriculum scores are normally distributed with true mean 203 and standard deviation 20. We test $H_0 : \mu = 200$ against $H_a : \mu > 200$ at the 1% significance level.

3. Find the probability of a type II error for $n = 100, 400, 900, 1600$.

- Now we just plug in the corresponding n to the formula $P(N_{203,20/\sqrt{n}} \leq c) = P(N_{0,1} \leq 2.3263 - \dfrac{3\sqrt{n}}{20})$.
- For $n = 100$ we get $P(N_{0,1} \leq 0.8263) = 0.7957$.
- For $n = 400$ we get $P(N_{0,1} \leq -0.6737) = 0.2503$.
- For $n = 900$ we get $P(N_{0,1} \leq -2.1737) = 0.01486$.
- For $n = 1600$ we get $P(N_{0,1} \leq -3.6737) = 0.0001195$.

Type I and Type II Errors, XXII

We can glean a few general insights from from the examples.

- First, by adjusting the significance level α, we can affect the balance between type I errors and type II errors.
- A smaller α gives a smaller probability of a type I error but a greater probability of a type II error: we are more stringent about rejecting the null hypothesis (so fewer type I errors) but at the same time that means we also incorrectly fail to reject the null hypothesis more (so more type II errors).
- Second, by increasing the sample size, we decrease the probabilities of both error types together (though they do not necessarily drop similar amounts). This is also quite reasonable: the larger the sample, the closer the sample mean should be to the true mean and the less variation around the true mean it will have.

Power, I

With a larger sample size, the test will have a better ability to distinguish smaller deviations away from the null hypothesis.

Definition

If we are testing a null hypothesis H_0, the probability $1 - \beta$ of correctly rejecting the null hypothesis when it is false is called the power of the test.

- The power of the test will depend on the significance level α, the true value of the test parameter, and the size n of the sample.
- For a fixed α and n, we can plot the dependence of the power on the true value of the test parameter to produce what are called power curves.

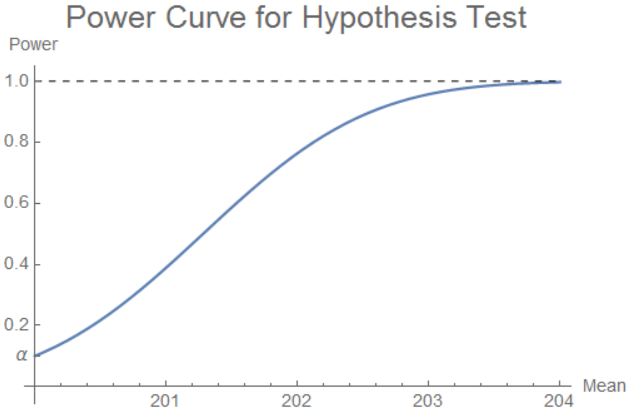
Power, II

To plot a power curve, we need only perform a calculation like the one in the last example.

- Specifically, first we calculate the critical value, and then we calculate the probability of correctly rejecting the null hypothesis based on the true value of the test statistic.
- For the test we analyzed above, of testing $H_0 : \mu = 200$ against $H_a : \mu > 200$ with significance level $\alpha = 0.10$ and a sample size $n = 400$, we want to reject the null hypothesis if $\hat{\mu} > 201.282$, and so the power of the test if the true mean is x is $P(N_{x,1} \geq 201.282 = P(N_{0,1} \geq 201.282 - x)$, whose graph is plotted on the next slide.

Power, III

Here is a fairly typical power curve, from the last example:



Notice that the power increases as the true mean (the x-variable) increases away from the null-hypothesis mean of 200.

Power, IV

Here are a few brief observations:

- As is suggested by the plot, the limit of the power as the true mean approaches the null hypothesis mean is equal to α. (This follows by noting the moderately confusing fact that the type II error coincides with the complement of the type I error in the limit.)
- Furthermore, the power increases monotonically as the true parameter value moves away from the null hypothesis mean, and approaches 1 as the true parameter value becomes large.

Power, V

Although it may seem that we would always want the power of the test to be as large as possible, there are certain non-obvious drawbacks to this desire.

- Specifically, if the power is very large even for small deviations away from the null hypothesis parameter, then the test will frequently yield statistically significant results even when the sample parameter is not very far away from the null hypothesis parameter.

Power, VI

In some – perhaps most – situations a high power may seem good, but sometimes it is not.

- For example, suppose we want to test whether the new curriculum actually improves scores above the original mean $\mu = 200$.
- If the power is sufficiently high, the hypothesis test will indicate a statistically significant result whenever the the sample mean $\hat{\mu} > 200.001$.
- Now, it certainly is useful to know that the true mean is statistically significantly different from 200, but in most situations we would not view this difference as "practically useful".

Power, VII

This issue is usually framed as statistical significance versus practical significance.

- With large samples, we may obtain a statistically significant difference from the hypothesized mean (perhaps even with an exceedingly small *p*-value), yet the actual difference is negligibly small and not actually important in practice.
- This highlights one issue with relying solely on *p*-values on a measure of evidence quality: it is possible to set up tests (e.g., by using a very large sample) that yield extremely small *p* values even if the actual result is practically meaningless.
- Another, more philosophical, point here is that the null hypothesis is rarely (if ever) exactly true: thus, if we take a sufficiently large sample size, we can identify as statistically significant whatever tiny deviation actually exists, even if this deviation is not practically relevant.

Some Comments About Significance, I

With these observations in mind, we can see that the precise choice of the significance level α is entirely arbitrary (which has been illustrated by the somewhat eclectic selection of values in the examples we have given so far).

- The only particular considerations we have are whether the choice of α yields acceptably low probabilities of making a type I or type II error.
- In some situations, we would want to be extremely sure, when we reject the null hypothesis, that it was truly outlandishly unlikely to have observed the given data by chance: this corresponds to requiring α to be very small.

More Comments About Significance, II

- For example, if the result of the hypothesis test is regarding whether the numbers in a company's accounting ledgers are real or manufactured to cover up embezzling, we would want to be very sure that any seeming discrepancies were not merely random chance.
- However, in other situations (e.g., in the sciences) where the statistical test is merely one component of broader analysis of a topic, we should view the result of a hypothesis test as more of a suggestion for what to investigate next.
- If the p-value is very small, then it suggests that the alternative hypothesis may be correct, and further study is warranted.
- If the p-value is large, then it suggests that the null hypothesis is correct, and that additional study is not likely to yield different results.

These Comments Have A Bit Of An Edge, III

For various historical reasons, the significance level $\alpha = 0.05$ is very commonly used.

- The value strikes a balance between requiring strong evidence (only a 5% chance that the result could have arisen by chance) but not so strong as to tend to ignore good evidence suggesting the null hypothesis is false (which becomes likely with smaller values of α).
- Indeed, many authors, both in the past and the present, often call a result with $p < 0.05$ "statistically significant" (with no qualifier) and a result with $p < 0.01$ "very statistically significant" (and if $p < 0.001$, one also sometimes sees "extremely statistically significant").
- Such statements entirely ignore the actual nuances of what p-values measure, and should be assiduously avoided!

Now It Sounds A Bit Like A Rant, IV

The history of $\alpha = 0.05$ is often summarized as follows (adapted from a paper from the American Statistical Association's statement on p-values):

Q: Why do so many colleges and grad schools teach $p = 0.05$?

A: Because that's still what the scientific community and journal editors use.

Q: Why do so many people still use $p = 0.05$?

A: Because that's what they were taught in college or grad school.

Rant Impending, V

As I would hope you understand by now, *p*-values come on a sliding scale, and there are no obvious lines to draw.

- A hypothesis test with $p = 0.051$ provides almost the same level of evidence against the null hypothesis as a hypothesis test with $p = 0.049$, and there is simply no practical distinction that should be made between the two.
- Nonetheless, the prevalence of the view that results are not worth reporting unless they have $p < 0.05$ has led to various undesirable, and very real, negative consequences.
- One such problem is the lack of reporting of experiments that had negative or "statistically insignificant" results (which is also partly a cultural issue in research, more generally), which leads to a bias in the resulting literature.

Very Much Seeming Like A Rant, VI

There are various other related factors that can also contribute to an overall bias in reported results of hypothesis tests.

- When analyzing collected data, it is important to examine outliers (points far away from the norm), since they may be the results of errors in data collection or otherwise unrepresentative of the desired sample.
- The presence of outliers often has a large effect on the results of a hypothesis test, especially one that relies on an estimate of a standard deviation or variance, and in some situations it is entirely reasonable to discard outliers.

Yes, It's Definitely A Rant, VII

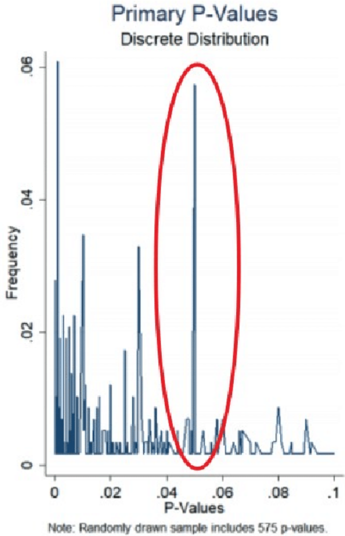
However, the process of handling outliers can rise to the level of scientific misconduct if it is done after the fact.

- The phenomenon called *p*-hacking involves massaging the underlying data used for a statistical test (e.g., by removing additional outliers, or putting back outliers that were previously removed) so that it yields a *p*-value less than 0.05 rather than greater than 0.05.
- Various analyses of published *p*-values have uncovered a large proportion of *p*-values (much larger than would be expected from typical hypothesis tests) that are just below $\alpha = 0.05$.

Ranting With Figures, VIII

Here is a sample of *p*-values published in medical research papers:

Citation: Havenaar, Matthias. "Is medical research facing a replication crisis?" (2018)



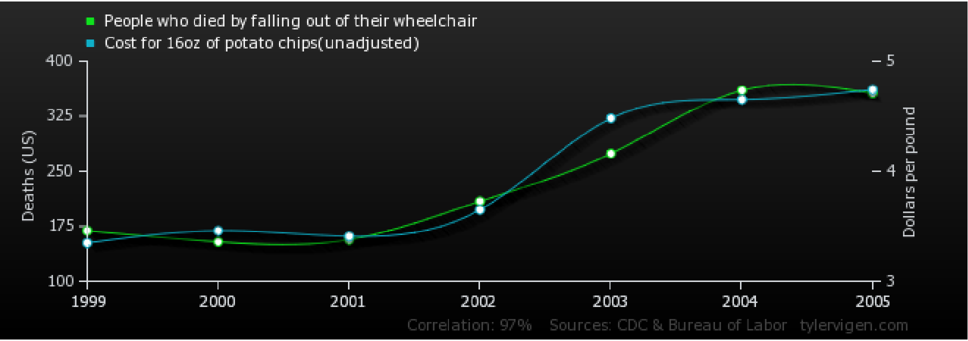
More Ranting, IX

Another related issue is that of performing multiple comparisons on the same set of data.

- This procedure is sometimes (more uncharitably) referred to as data dredging: sifting through data to find signals in the underlying noise.
- The difficulty with performing multiple comparisons is that there is a probability α that any given hypothesis test will yield a statistically significant result even though the null hypothesis is true, and these probabilities add up if we perform more tests.
- For example, if we perform 40 hypothesis tests where the null hypothesis is actually true at the $\alpha = 0.05$ significance level, we will have a probability $1 - 0.95^{40} \approx 87\%$ of getting at least one statistically significant result (i.e., making a type I error), even though there is no actual result to find.

Ranting With More Pictures, X

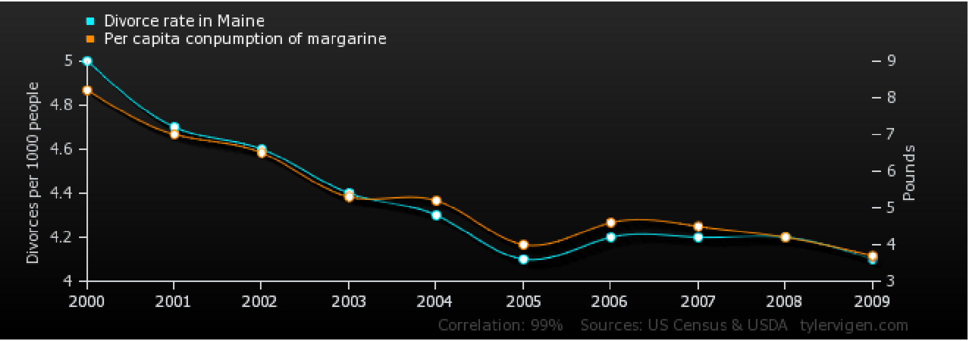
We can illustrate this phenomenon (rather amusingly) with some examples from Spurious Correlations:



Citation: Tyler Vigen (“Spurious Correlations”, tylervigen.com/spurious-correlations)

Ranting With More Pictures, XI

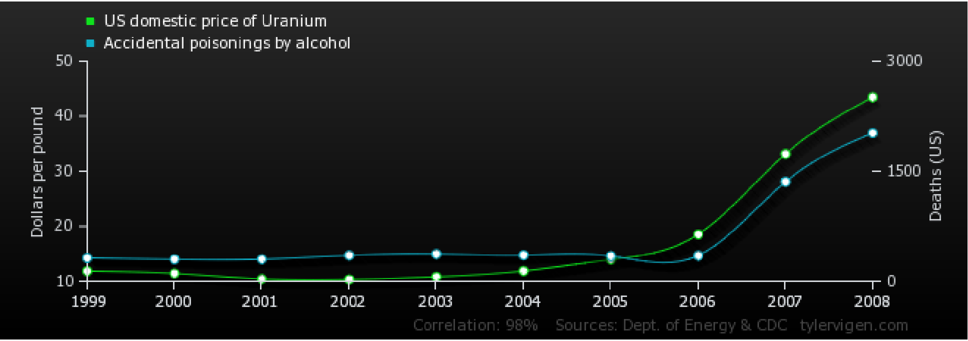
We can illustrate this phenomenon (rather amusingly) with some examples from Spurious Correlations:



Citation: Tyler Vigen (“Spurious Correlations”, tylervigen.com/spurious-correlations)

Ranting With More Pictures, XII

We can illustrate this phenomenon (rather amusingly) with some examples from Spurious Correlations:



Citation: Tyler Vigen ("Spurious Correlations", tylervigen.com/spurious-correlations)

Yet More Ranting, XIII

When actually performing a large number of hypothesis tests, one should correct for the fact that multiple tests were performed on the same data.

- Various methods exist for this, such as the Bonferroni correction, which states that the desired significance level α should be divided by the total number of tests performed.
- The idea is simply that we want a total probability of approximately α (among all the tests) of obtaining at least one type I error among all the tests.
- Thus, if we perform 5 different tests on the same data using the typical $\alpha = 0.05$, we should actually test at the level $\alpha = 0.01$ in order to have an overall total probability of approximately 0.05 of obtaining at least one type I error.

Continued Ranting, XIV

Multiple hypothesis testing on the same data is not necessarily a problem if we report the results of all of the tests and give the actual p-values for each test, since then it is straightforward to correct for multiple tests.

- However, a much more serious issue occurs when we only report the statistically significant results without noting (or correcting for) the fact that other hypothesis tests were also performed and not reported: it is then entirely possible that most of the reported results are false.
- The extent to which false research findings are an actual problem in scientific research is disputed, and varies substantially by field, but is obviously a fundamental concern!
- See Ioannadis, "Why Most Published Research Findings are False", PLoS Medicine (2005) for an argument that this is a serious and widespread problem.

Ranting With Hypothetical Examples, XV

To illustrate the problem here, consider the hypothetical "Journal of Advanced Augury", which publishes papers related to augury, the classical Roman practice of interpreting the will of the gods by studying flights of birds.

- The editorial board of JAA, trained in the basic practice of statistics, will only accept papers demonstrating a result with a p-value in their "statistically significant" range $p < 0.05$.
- So, imagine augurs all over the world, dutifully observing birds and writing down their corresponding predictions, then compiling all of their results that come out with $p < 0.05$ for the journal.

Ranting With Hypothetical Examples, XVI

Under the quite reasonable assumption that none of these results is in any sense real[1], only about 1 in 20 of them will pass the cutoff for publication.

- Nevertheless, this will still represent plenty of results, all of which seem impeccably valid: after all, they all passed the journal's threshold for statistical significance!
- The problem is that we do not see the 95% of experiments performed that did not pass the $p < 0.05$ threshold.
- This is sometimes called the file drawer problem: that experiments not meeting the statistical threshold for publication are not published (they instead go into a file drawer), which results in a bias in the published results.

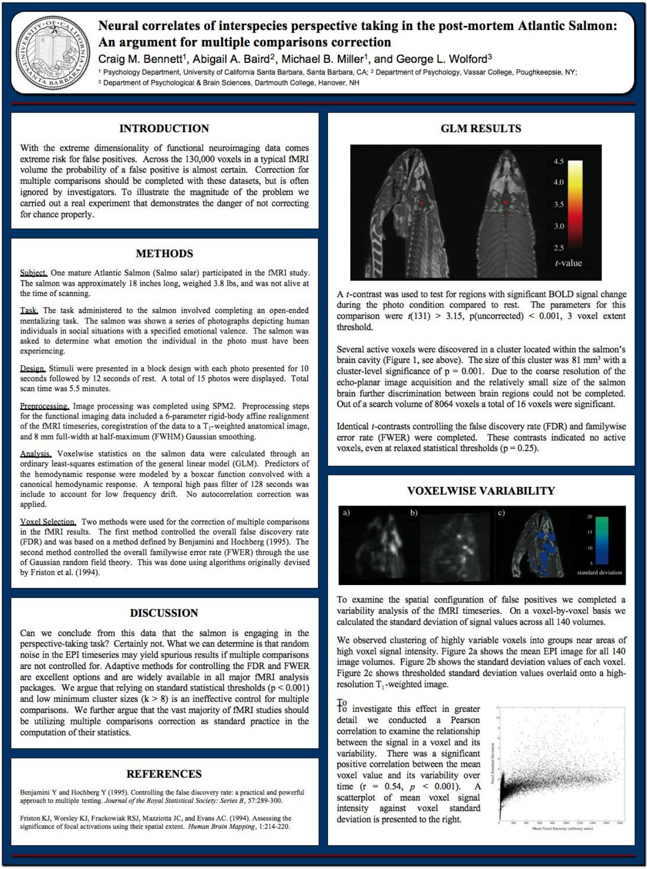
[1] As everyone knows, haruspicy is at least twice as effective as augury in predicting the future.

Ranting With Dead Fish, XVII

I would like to mention another real paper[2] that caused a bit of a splash (so to speak) in psychology.

- The researchers applied functional MRI (fMRI) to a dead salmon and used standard statistical analysis on the 130,000 voxels in the resulting data to identify several clusters in the salmon's brain cavity that correlated well with an emotional identification task.
- The point of the paper was to illustrate the dangers inherent in performing large numbers of tests on the same data set without correcting properly for the likely appearance of a large number of false positives.

[2] Bennet CM, Baird AA, Miller MB, Wolford GL, "Neural Correlates of Interspecies Perspective Taking in the Post-Mortem Atlantic Salmon: An Argument For Proper Multiple Comparisons Correction", *Journal of Serendipitous and Unexpected Results*, 2011.



Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction

Craig M. Bennett[1], Abigail A. Baird[2], Michael B. Miller[1], and George L. Wolford[3]

[1] Psychology Department, University of California Santa Barbara, Santa Barbara, CA; [2] Department of Psychology, Vassar College, Poughkeepsie, NY; [3] Department of Psychological & Brain Sciences, Dartmouth College, Hanover, NH

INTRODUCTION

With the extreme dimensionality of functional neuroimaging data comes extreme risk for false positives. Across the 130,000 voxels in a typical fMRI volume the probability of a false positive is almost certain. Correction for multiple comparisons should be completed with these datasets, but is often ignored by investigators. To illustrate the magnitude of the problem we carried out a real experiment that demonstrates the danger of not correcting for chance properly.

METHODS

Subject. One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning.

Task. The task administered to the salmon involved completing an open-ended mentalizing task. The salmon was shown a series of photographs depicting human individuals in social situations with a specified emotional valence. The salmon was asked to determine what emotion the individual in the photo must have been experiencing.

Design. Stimuli were presented in a block design with each photo presented for 10 seconds followed by 12 seconds of rest. A total of 15 photos were displayed. Total scan time was 5.5 minutes.

Preprocessing. Image processing was completed using SPM2. Preprocessing steps for the functional imaging data included a 6-parameter rigid-body affine realignment of the fMRI timeseries, coregistration of the data to a T_1-weighted anatomical image, and 8 mm full-width at half-maximum (FWHM) Gaussian smoothing.

Analysis. Voxelwise statistics on the salmon data were calculated through an ordinary least-squares estimation of the general linear model (GLM). Predictors of the hemodynamic response were modeled by a boxcar function convolved with a canonical hemodynamic response. A temporal high pass filter of 128 seconds was include to account for low frequency drift. No autocorrelation correction was applied.

Voxel Selection. Two methods were used for the correction of multiple comparisons in the fMRI results. The first method controlled the overall false discovery rate (FDR) and was based on a method defined by Benjamini and Hochberg (1995). The second method controlled the overall familywise error rate (FWER) through the use of Gaussian random field theory. This was done using algorithms originally devised by Friston et al. (1994).

DISCUSSION

Can we conclude from this data that the salmon is engaging in the perspective-taking task? Certainly not. What we can determine is that random noise in the EPI timeseries may yield spurious results if multiple comparisons are not controlled for. Adaptive methods for controlling the FDR and FWER are excellent options and are widely available in all major fMRI analysis packages. We argue that relying on standard statistical thresholds ($p < 0.001$) and low minimum cluster sizes ($k > 8$) is an ineffective control for multiple comparisons. We further argue that the vast majority of fMRI studies should be utilizing multiple comparisons correction as standard practice in the computation of their statistics.

REFERENCES

Benjamini Y and Hochberg Y (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57:289-300.

Friston KJ, Worsley KJ, Frackowiak RSJ, Mazziotta JC, and Evans AC. (1994). Assessing the significance of focal activations using their spatial extent. *Human Brain Mapping*, 1:214-220.

GLM RESULTS

A t-contrast was used to test for regions with significant BOLD signal change during the photo condition compared to rest. The parameters for this comparison were $t(131) > 3.15$, p(uncorrected) < 0.001, 3 voxel extent threshold.

Several active voxels were discovered in a cluster located within the salmon's brain cavity (Figure 1, see above). The size of this cluster was 81 mm^3 with a cluster-level significance of p = 0.001. Due to the coarse resolution of the echo-planar image acquisition and the relatively small size of the salmon brain further discrimination between brain regions could not be completed. Out of a search volume of 8064 voxels a total of 16 voxels were significant.

Identical t-contrasts controlling the false discovery rate (FDR) and familywise error rate (FWER) were completed. These contrasts indicated no active voxels, even at relaxed statistical thresholds (p = 0.25).

VOXELWISE VARIABILITY

To examine the spatial configuration of false positives we completed a variability analysis of the fMRI timeseries. On a voxel-by-voxel basis we calculated the standard deviation of signal values across all 140 volumes.

We observed clustering of highly variable voxels into groups near areas of high voxel signal intensity. Figure 2a shows the mean EPI image for all 140 image volumes. Figure 2b shows the standard deviation values of each voxel. Figure 2c shows thresholded standard deviation values overlaid onto a high-resolution T_1-weighted image.

To
To investigate this effect in greater detail we conducted a Pearson correlation to examine the relationship between the signal in a voxel and its variability. There was a significant positive correlation between the mean voxel value and its variability over time ($r = 0.54$, $p < 0.001$). A scatterplot of mean voxel signal intensity against voxel standard deviation is presented to the right.

Ranting With Dead Fish, XIX

Just because it's so delightful, here are some quotes from the methods section of the paper:

- "One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon measured approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning. It is not known if the salmon was male or female, but given the post-mortem state of the subject this was not thought to be a critical variable."
- "Foam padding was placed within the head coil as a method of limiting salmon movement during the scan, but proved to be largely unnecessary as subject motion was exceptionally low."

Ranting With Dead Fish, XX

More from the methods section of the paper:

- "A mirror directly above the head coil allowed the salmon to observe experiment stimuli. The task administered to the salmon involved completing an open-ended mentalizing task. The salmon was shown a series of photographs depicting human individuals in social situations with a specified emotional valence, either socially inclusive or socially exclusive. The salmon was asked to determine which emotion the individual in the photo must have been experiencing."
- "The photo stimuli were presented in a block design, with each block consisting of four photos presented individually for 2.5 seconds each (10 seconds total) followed by 12 seconds of rest. A total of 12 blocks of photo presentation were completed with 48 photos presented during the run."

Ranting With Dead Fish, XXI

The authors spend most of the paper discussing how to avoid making the same statistical errors that lead one to believe that a dead salmon is capable of identifying emotional states (and why some authors don't seem to want to do that):

- "Any time that multiple tests are completed without proper correction it has the potential to impact the conclusions drawn from the results."
- "The control of false positives is not a matter of difficulty, as all major analysis packages for fMRI include some form of multiple comparisons correction. Rather it seems to be the case that investigators do not want to jeopardize their results through a reduction in statistical power. While we must guard against the elimination of legitimate results through Type II error, the alternative of continuing forward with uncorrected statistics cannot be an option."

Ranting With Another Hypothetical Example, XXII

Let's do an almost-real example: consider trying to identify genes that cause breast cancer.

- Suppose about 100 of the 20,000 protein-coding genes in the genome will, if suitably mutated, increase the risk of developing breast cancer.
- Now, we collect some data and do statistics. Suppose that our test will pick up an actually-related gene 100% of the time, but if we pick one of the others, then we have a 5% probability of getting $p < 0.05$ and flagging it.
- We then publish every one of the results where $p < 0.05$. What is the probability that the identified gene *actually has anything to do with breast cancer*?

Ranting With Another Hypothetical Example, XXII

Let's do an almost-real example: consider trying to identify genes that cause breast cancer.

- Suppose about 100 of the 20,000 protein-coding genes in the genome will, if suitably mutated, increase the risk of developing breast cancer.
- Now, we collect some data and do statistics. Suppose that our test will pick up an actually-related gene 100% of the time, but if we pick one of the others, then we have a 5% probability of getting $p < 0.05$ and flagging it.
- We then publish every one of the results where $p < 0.05$. What is the probability that the identified gene *actually has anything to do with breast cancer*?

In fact, we worked out this kind of calculation in the second week of class: you can use Bayes' formula to see that only about 9.1% of the published results are correct (100 real vs 999.5 false positives).

Ranting With Another Hypothetical Example, XXIII

If you remember back to the “testing for a disease” Bayes’ formula example, you’ll remember that the probability of having the disease given a positive test was actually fairly low, in the situation where the disease is rare.

- The same sort of effect is observable here: if very few of the studied results are actually real, then most of the results that have $p < 0.05$ will be false positives.
- The solution to this problem, as in the disease testing example, is to test again.
- For scientific studies, this means performing a replication study: a study designed to test whether a particular result is reproducible by a different team working independently.

Nearly Done Ranting, XXIV

When spurious results are reported as significant, followup studies will (at least in theory) eventually show that the original results were erroneous – this is the virtue of the scientific method.

- But this phenomenon of having subsequent studies widely being unable to replicate the results of the originals has led to a replication crisis in various fields, since it suggests that most of these original results were actually false.
- Although one can reasonably adopt the viewpoint that, *eventually*, incorrect results will be identified and extirpated, having many false results believed to true creates a substantial waste of resources (in having to perform unnecessary replication studies and, more broadly, building additional research on a faulty foundation).

Ranting About How To Fix Things, XXV

Various fixes for these issues have been proposed, such as decreasing the *p*-value threshold for publication (the most common value suggested is $p < 0.01$). But this change, if adopted, would not solve the problem on its own.

- One reason why is that in most things under reasonable scientific examination, it is rarely true that there is absolutely zero effect, especially since we don't often test things for no reason.
- For example, consider a drug trial. Since biological processes are so tightly interconnected, it is deeply unlikely that taking a drug will have absolutely zero effect on an underlying condition.
- So suppose the actual effect is extremely small, say, that on average, 1 patient in 1,000 sees a net positive effect.

Ranting About Another Hypothetical Example, XXVI

Perhaps everyone is scored on a quality-of-life measure, and the measure with the drug is 1/100th of a point better on average. Even with such a small effect, strictly speaking the null hypothesis is now false: the average effect is not zero.

- So now suppose we run a gigantic trial with 1,000,000 participants to decide whether the drug is effective.
- It is quite likely we will be able to detect the improvement, even though it is very small.
- If for example the quality-of-life scores are normally distributed with mean 8.00000 and standard deviation 1.00000, and the observed sample mean is 8.01000, then the p-value for the resulting one-sided z-test is $P(N_{8,0.001} \geq 8.01) = P(N_{0,1} \geq 10) = 7.62 \cdot 10^{-24}$ (tiny!).
- Nonetheless, the drug's actual effect is practically nonexistent: 1/100th of a point on a 10-point scale on average.

The Penultimate Slide of The Rant, XXVII

I have mentioned these issues because it is very important to be sanguine about the limitations of hypothesis testing, and how easy it is to misuse or misinterpret the results of hypothesis tests.

- Ultimately, there can be no “magic fix” for these issues: statistical testing is fundamentally an approximation, and there is always a positive probability of getting an incorrect result.
- Although these issues are not inherently issues of statistics, I feel it does you (the students) a major disservice if you only learn how to apply the basic tests, without being told about all of the easy-to-make mistakes that go along with using statistics in the real world.

The Actual End of The Rant, XXVIII

When designing an experiment and a hypothesis test, the best we can do is to identify an appropriate significance level α and an appropriate sample size n.

- We select α to balance the possibility of making a type I error against the possibility of making a type II error.
- We select n to balance the possibility of making any type of error with the difficulty and expense of obtaining the necessary data, and with the likelihood that there probably is some practically irrelevant deviation from the null hypothesis.
- Then we must conduct followup analyses and replication studies to make sure any observed results are truly real and practically significant.

Guidelines For *p*-Values

In 2016, the American Statistical Association released guidelines for interpretation and usage of *p*-values:

1. *p*-values can indicate how incompatible the data are with a specified statistical model.
2. *p*-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold.
4. Proper inference requires full reporting and transparency.
5. A *p*-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis.

Guidelines For Statistical Inference

I will also quote the conclusion of the ASA's guidelines:

- Good statistical practice, as an essential component of good scientific practice, emphasizes principles of good study design and conduct, a variety of numerical and graphical summaries of data, understanding of the phenomenon under study, interpretation of results in context, complete reporting and proper logical and quantitative understanding of what data summaries mean.
- No single index should substitute for scientific reasoning.

From Wasserstein and Lazar, "The ASA Statement on *p*-Values: Context, Process, and Purpose" (2016)
`https://amstat.tandfonline.com/doi/full/10.1080/00031305.2016.1154108`

Summary

We discussed more facets of type I and type II errors.

We discussed the power of a statistical test and some of its properties.

We discussed some other issues about using (and misusing) hypothesis tests.

Next lecture: The t distribution, t-statistic confidence intervals