Likelihood Ratio and False Positive Risk

John Maindonald, Statistics Research Associates

12 September 2023

Contents

Introduction **1**

1 Basic properties of p-values **2**
1.1 Wear comparison for two shoe materials 3
1.2 What p-values do not, and cannot, provide 5
1.3 Strategies for the use of p-values 5
1.4 Small differences may be of no interest 6

2 Likelihood ratio and false positive risk **7**
2.1 What is the definitive question? 7
2.2 Density curves, under the NULL, and under the alternative . . . 8
2.3 Maximum likelihood ratio versus p-value 9
2.4 False positive risk versus p-value 10

3 Power – how well does a planned experiment discriminate? **11**
3.1 The power of a t or other statistical test 11
3.2 False positive risk, when α is used as cutoff 14

4 How should results be reported? **15**

5 Further reading and references **15**
References 15

Introduction

Under NULL hypothesis assumptions, p-values are uniformly distributed on the unit interval. The common $p \leq 0.05$ strategy for rejecting the NULL hypothesis is justified by a probability for this, under the NULL, that equals 0.05. P-values are often misinterpreted. The notes that follow draw attention to common misunderstandings, and compare and contrast p-values with the insights likelihood based statistics provide.

The p-value probability p relates only to what can be expected under the NULL. For tests that are based on t-statistics, a p-value that equals 0.05 translates to a maximum likelihood ratio that, for degrees of freedom greater than 5, is less than 5.

Decimal numbers that are shown on graphs are given to two significant figures. In the text, they may be given three significant figures.

1 Basic properties of p-values

P-values are commonly used within a Null Hypothesis Significance Testing (NHST) framework. This approach to statistical decision making sets up a choice between a null hypothesis, commonly written H_0, and alternative H_1, with the calculated p-value used to decide whether H_0 should be rejected in favour of H_1. Commonly, H_0 is the hypothesis that a difference of means, or a mean difference, has been drawn from a population with mean $\mu = 0$. In a medical context, a treatment of interest may be compared with a placebo. Then

- Given H_0, the p-value is uniformly distributed on the interval $0 \leq p \leq 1$
  - As a consequence, for any α, $P[p \leq \alpha \mid H_0] = \alpha$
- Under H_1, the p-value is designed to increase as the difference from H_0 increases.

More informative than to report $p \leq 0.05$ is to give a 95% confidence interval for the mean. The NULL hypothesis is rejected at a level of $\alpha = 0.05$ if and only if the interval does not contain 0.

Figure 1 shows the distributions of values for five random samples drawn from the uniform distribution on the interval from 1 to 0. The ordering from 1 to 0 is designed to reflect the decrease in p-value with increasing absolute value of the t or other such statistic.

Under H_0, a fraction α of p-values from independent replications of an experiment will, on average, be less than α. Figure 1, with values less than 0.05 shown in red, is designed to highlight this point for α=0.05 . The values in the first and second samples that are ≤ 0.05 are, to three decimal places.

```
0.047 0.044 0.038 0.02 0.017 0.007
0.029 0.024 0.007
```

Values that are less than α (in the figure, α=0.05) are sampled from a uniform distribution on the interval from α to 0.

The calculated p-value provides more nuanced evidence than comes from merely noting whether it is less than α, typically with $\alpha = 0.05$. A calculated value p is, however, at the upper end of the range of values that under H_0 occur with probability p. It is, under H_0, the expected value for p-values that are in the interval that extends from 0 to $\alpha = 2p$. This suggests that

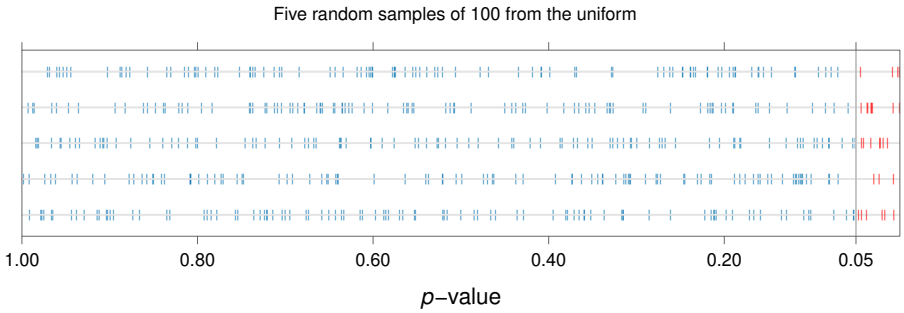

Figure 1: Under the NULL hypothesis, and assuming distributional assumptions are correct, p-values are uniformly distributed on the unit interval. The strip plots each show the distribution of values in a random sample of p-values, under NULL hypothesis assumptions. The ordering from 1 to 0 reflects the decrease in the p-value, for commonly used test statistics, as the absolute value of the test statistic increases. Values less than the commonly used 0.05 threshold are shown in red.

- If $p = 0.05$, it is the expected value, under Under H_0, of values that range from 0 to 0.1, and that occur with a frequency of 0.1 (or 10%).
- Note, however! Under H_1, the distribution is no longer uniform, and the range of values for which a calculated value p is the expected value will change.
  - Doubling the calculated p-value, in order to get an equivalent that on average corresponds to the "Reject H_0 when $p < \alpha$ strategy will then, on average, lead to a different rejection rate when the NULL is false!

Rather that making such sense as one can of of the calculated p-value, a better approach is to work with likelihood ratios.

1.1 Wear comparison for two shoe materials

Data from an experiment that compares results from a treatment with a baseline provides a relatively simple setting in which to probe the interpretation that should be placed on a given p-value. Even in this 'simple' setting, the issues that arise for the interpretation of a p-value, and its implication for the credence that should be given to a claimed difference, are non-trivial.

The `MASS::shoes` dataset compares, for each of ten boys, the wear on two different shoe materials. Materials A and B were assigned at random to feet — one to the left foot, and the other to the right. It will be used as a relatively simple setting in which to probe the interpretation that should be placed on a given p-value.

The measurements of wear, and the differences for each boy, were:

```
wear <- with(MASS::shoes, rbind(A,B,d=B-A))
colnames(wear) <- rep("",10)
wear

A 13.2 8.2 10.9 14.3 10.7  6.6 9.5 10.8 8.8 13.3
B 14.0 8.8 11.2 14.2 11.8  6.4 9.8 11.3 9.3 13.6
d  0.8 0.6  0.3 -0.1  1.1 -0.2 0.3  0.5 0.5  0.3
```

Here, the samples are paired The differences will be used for analysis, thus reducing the analysis to that for a single sample t-test. The differences $d_i, i = 1, 2, \ldots, n$ are then used for analysis. The p-value for testing for no difference is obtained by referring the t-statistic for the mean $\bar{d}$ of the d_i to a t-distribution with $n - 1$ degrees of freedom.

The calculation assumes that the differences $d_i, i = 1, 2, \ldots 10$ have been independently drawn from the same normal distribution. The statistic $\sqrt{n}\,\bar{d}/s$, where $\bar{d}$ is the mean of the d_i, and s is the sample standard deviation, can then be treated as drawn from a t-distribution. The p-value for a 2-sided test is then, assuming H0, and as any difference might in principle go in either direction

> the probability of occurrence of values of the t-statistic t that are greater than or equal to $\sqrt{n}\bar{d}/s$ in magnitude

Calculations proceed under the NULL hypothesis assumption that the differences are a random sample from a normal distribution with mean zero:

Mean	SD	n	SEM	t	pval	df
0.41	0.387	10	0.122	3.35	0.00854	9

The p-value can then be interpreted in the following ways:

- We may have decided in advance to set a cutoff α, then lumping together all values less than α
  - With $\alpha = 0.05$, $p = 0.009$ would count as an event, under H_0, with probability 0.05
  - With $\alpha = 0.01$, this would count as an event, under H_0, with probability 0.01
- The calculated p-value presents a more nuanced picture. The probability to which a value of magnitude p then relates is $2p$ rather than p.
  - The calculated p is in the middle of the range from $2p$ to 0. The probability that a p-value will appear in that range is, under the NULL, $2p$.

A 95% (two-sided) confidence interval for the `B-A` wear difference μ is

`shoeStats[['Mean']]` $\pm$ `qt(.97.5,9)*shoeStats[['SEM']]`
i.e., $0.133 < \mu < 0.687$

A 99% confidence interval is $0.012 < \mu < 0.808$

1.2 What p-values do not, and cannot, provide

- A p-value does not give the probability that the NULL is false.
  - It is calculated under the assumption that the NULL hypothesis is true. It does not tell us whether that hypothesis is correct!
- Nor does the p-value give the probability that the results occurred by chance.
  - In order to calculate this, one needs a prior estimate of the frequency with which, over independent repeats of the process that generated the data, true positives can be expected.

Resnick (2017) makes the point thus:

> The tricky point is then, that the p-value does not show how rare the results of an experiment are. It's how rare the results would be in the world where the null hypothesis is true. That is, it's how rare the results would be if nothing in your experiment worked, and the difference ... was due to random chance alone. The p-value quantifies this rareness.

What one can say is that

> As the p-value becomes smaller, it becomes less likely that the NULL hypothesis is true.

1.3 Strategies for the use of p-values

A binary choice is not always appropriate.

There are many circumstances where it makes more sense to treat the problem as one of estimation, with the estimate accompanied with a measure of accuracy.

One experiment may not, on its own, be enough

Note comments from Fisher (1935), who introduced the use of p-values, on their proper use:

> No isolated experiment, however significant in itself, can suffice for the experimental demonstration of any natural phenomenon; for the 'one chance in a million' will undoubtedly occur, with no less and no more than its appropriate frequency, however surprised we may be that it should occur to us. In order to assert that a natural phenomenon is experimentally demonstrable we need, not an isolated record, but a reliable method of procedure. In relation to the test of significance, we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result.

In other words, use p-values as a screening device, to identify results that may merit further investigation. This is very different from the way that p-values

have come to be used in most current scientific discourse. A p-value should be treated as a measure of change in the weight of evidence, not a measure of the absolute weight of evidence.

An independent repetition of the experiment provides checks that no statistical analysis can provide. Such checks, which are widedly neglected, are important for reasons that extend beyond checking whether the initial $p \leq \alpha$ was a fluke. For experimental data, they provide a check on biases that may arise from mistakes in procedure.

When p-values are used to choose between a NULL and an alternative, the focus is on how rare the "event" would be if the NULL hypothesis is true. There is no attention to assessing how much more likely it would be if the NULL is false. Likelihood ratios, which will now be discussed, do provide such a comparison. While the detailed discussion will be based around tests and comparisons that work with t-statistics, it will illustrate principles that apply more widely.

1.4 Small differences may be of no interest

Irrespective of the threshold set for finding a difference, both p and the likelihood ratio will detect increasingly small differences from the NULL as the sample size increases. A way around this is to set a cutoff for the minimum difference of interest, and calculate the difference relative to that cutoff.

Effectiveness of soporofic drugs

The use of a cutoff will be illustrated using the dataset `datasets::sleep`. This has the increase in sleeping hours, on the same set of patients, on each of the two drugs. Consider first the result from a regular two-sided test Data, with output from the t-test, are:

```
sleep2 <-with(sleep, Pair(extra[group==2], extra[group==1]))
t <- t.test(sleep2 ~ 1, data = sleep)
```

The t-statistic is 4.06, with $p = 0.0028$. The p-value translates to a maximum likelihood ratio that equals 894.2, which suggests a very clear difference in effectiveness, in favour of drug 2.

It does then seem clear that drug B gives a bigger increase in hours of sleep. How sure can we be that it is large enough to be of substantial consequence?

A test that sets $\mu = 0.8$ hours as the baseline

Suppose, now, that 0.8 hours difference is set as the minimum that is of interest. As we are satisfied that drug B gives a bigger increase, and we wish to check the strength of evidence for an increase that is 0.8 hours of more, a one-sided test is appropriate. Figure 2A compares the densities.

Calculations can be done thus:

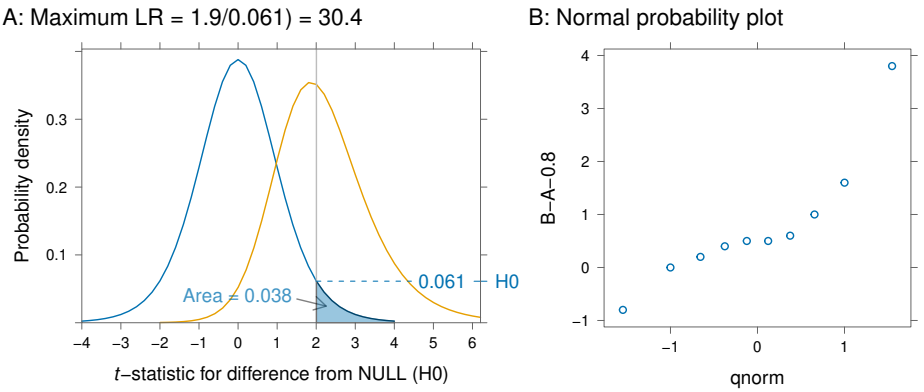

Figure 2: Panel A shows density curves for NULL and for the alternative, for a one-sided test with $t = 2.01$ on 9 degrees of freedom. This is the t-statistic for the data on the effect of soporific drugs when differences are 'B-A-0.8', i.e., interest is in the strength of evidence that differences are at least 0.8 hours. A vertical line is placed at the position that gives the p-value, here equal to 0.038. Panel B shows the normal probability plot for the differences.

```
tinfo <- t.test(sleep2 ~ 1, mu=0.8, alternative = 'greater')
t <- tinfo[['statistic']]; df <- tinfo[['parameter']]
maxlrSleep.8 <-
  with(tinfo, tTOmaxlik(t, df))
```

The t-statistic is 2.01, with $p = 0.038$. The maximum ratio of the likelihoods, given in Figure 2A as 3.5, is much smaller than the value of $\frac{1-p}{p} = 25.4$.

The normal probability plot shows a clear departure from normality. At best, the p-values give ballpark indications.

There are other ways to calculate a likelihood ratio. In principle, one might calculate the average for all values where $\bar{d}$ is greater than the cutoff. This, however, requires an assumed distribution for $\bar{d}$ under the alternative. It can never exceed the maximum value, calculated as in Figure 2A

2 Likelihood ratio and false positive risk

2.1 What is the definitive question?

Comments in Berkson (1942) highlight the point that p-values relate only to what can be expected under the NULL

> If an event has occurred, the definitive question is not, 'Is this an event which would be rare if the null hypothesis is true?' but 'Is there an alternative hypothesis under which the event would be relatively

frequent?'

By contrast, likelihood ratio statistics do address what Berkson identifies as "the definitive question".

2.2 Density curves, under the NULL, and under the alternative

Subsection 1.1 gave the following statistical summary information, for the ten observations in the shoe wear dataset:

```
  Mean     SD      n    SEM      t   pval     df
  0.41   0.39     10   0.12    3.3 0.0085      9
```

Here, in order to obtain a graph where the features of interest show up more clearly, we will take the first seven observations only from the shoe wear dataset. This is done for purposes of illustration only – the analysis that properly reflects the data is the analysis that is based on all 10 observations.

```
d 0.8 0.6 0.3 -0.1 1.1 -0.2 0.3
 Mean    SD     n   SEM     t  pval    df
  0.4  0.47     7  0.18   2.3 0.065     6
```

Figure 3 compares the density curves, under H0 and under an alternative H1 for which the estimated mean of the t-distribution is $t = \sqrt{n}\bar{d}/s$. Under the alternative, the t-statistic becomes the non-centrality parameter. Because this is subject to sampling error, the distribution is positively skewed and the mode, which gives the maximum likelihood, is to the left of the mean.

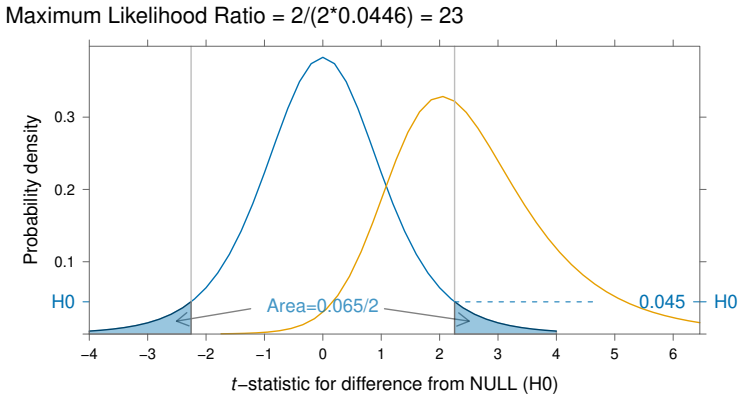

Figure 3: Density curves for NULL and for the alternative, for a two-sided test with $t = 2.26$ on 6 degrees of freedom. Vertical lines are placed at the positions that give the p-value, here equal to 0.065. Panel B shows the normal probability plot for the 'B-A' differences in the dataset.

The function `tOlr::tOmaxlik()` can be used to calculate the maximum likelihood under the alternative, at the same time calculating the maximum likelihood ratio. For Figure 3, degrees of freedom are 6, and the t-statistic is 2.256.

Calculations that give the maximum likelihood under the alternative, the maximum likelihood ratio, and other statistical information, then proceed thus:

```
stats7 <- list(t=2.256, df=6)  # t is rounded to 2dp
maxlr7 <- with(stats7, tOlr::tOmaxlik(t, df))
```

The values returned, to three significant figures are:

```
maxlik   tmax   lik0
 2.033  2.033 0.0446
```

Whereas the t-statistic was 2.256, the maximum likelihood estimate for the difference from the NULL, on the scale of the t-statistic, was 2.033

Likelihood ratios offer useful insights on what p-values may mean in practice. In the absence of contextual information that gives an indication of the size of the difference that is of practical importance, the ratio of the maximum likelihood when the NULL is false to the likelihood when the NULL is true gives a sense of the meaning that can be placed on a p-value. If information is available on the prior probability, or if a guess can be made, it can be immediately translated into a false positive risk statistic.

Likelihood ratio statistics directly address the question whether, under an alternative hypothesis, the observed data would be relatively more likely. They are, for this reason, in principle preferable to p-values. They are important, both for the light that they shed on p-values, and as alternatives to p-values.

2.3 Maximum likelihood ratio versus p-value

As noted earlier, the maximum likelihood ratio is calculated by dividing the maximum likelihood for the alternative by the likelihood for the NULL.

Figure 4 gives the maximum likelihood ratio equivalents of p-values, for a range of sample sizes, for p-values that equal 0.05, 0.01, and 0.001, and for a range of degrees of freedom. The comparison is always between a point NULL (here μ=0) and the alternative $\mu > 0$. For 6 or more degrees of freedom $p = 0.05$ translates to a ratio that is less than 5.0, while it is less than 4.5 for 10 or more degrees of freedom, and less than 4 for 13 or more degrees of freedom.

The ratio is higher for low degrees of freedom because of the way that the shape of the distribution changes. Other uncertainties enter. Departures from assumptions are of greatest consequence in those contexts where distributional asssumptions will detect only the most extreme departures from assumptions — i.e., when degrees of freedom are small. Experience with comparable historical data can be especially useful in those circumstances.

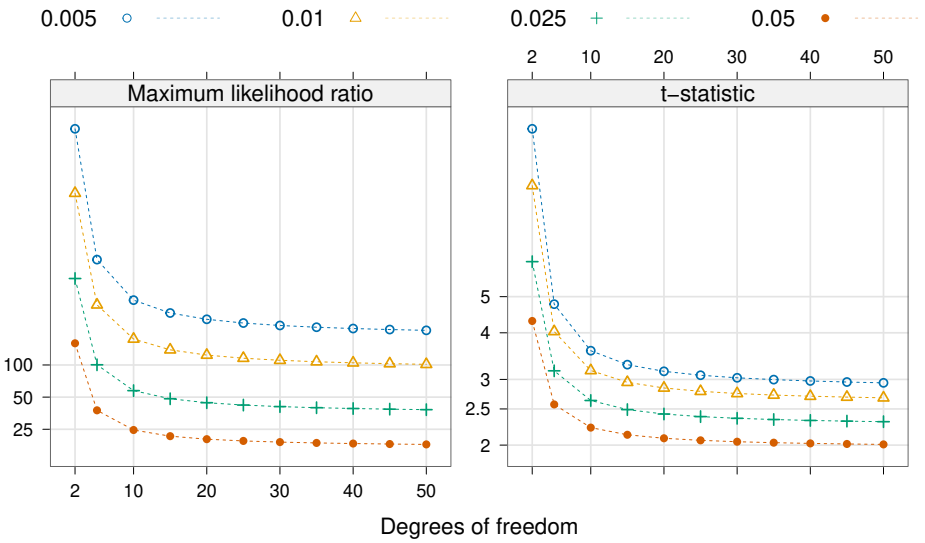

Figure 4: Ratio of the maximum likelihood under the alternative to the likelihood under the NULL, for three different choices of p-value, for a range of sample sizes, and for a range of degrees of freedom.

An observed $p = 0.05$ can be taken as representative of p-values that range from $\alpha = 0.1$ to 0, with odds against that are 9:1. This is commonly seen as providing strong evidence in favour of the alternative. The case for rejecting the NULL looks much less convincing when this is translated into a maximum likelihood ratio of the order or 4 or 5 in favour of the alternative.

Rather than focusing on the maximum likelihood ratio, one can compare the point NULL that we have been assuming with a point alternative, and a likelihood ratio that will usually be smaller.

2.4 False positive risk versus p-value

The false positive risk is the probability, under one or other decision strategy, that what is identified as a positive will be a false positive? False positive risk calculations require an assessment of the prior probability `prior` $= \pi$ of the alternative H1, with `1-prior` as the prior probability of H0. In the absence of such an assessment, all that can be said is that the NULL hypothesis becomes less likely as the p-value becomes smaller.

For any value of the maximum likelihood ratio `lr`, the false positive risk can then be calculated as `(1-prior)/(1-prior+prior*lr)`.

Figure 5 gives the false positive risk equivalents of p-values, for a range of sample sizes, for p-values that equal 0.05, 0.01, and 0.001, for a range of degrees of freedom, and for priors $\pi = 0.1$ and $\pi = 0.5$ for the probibility of H1.

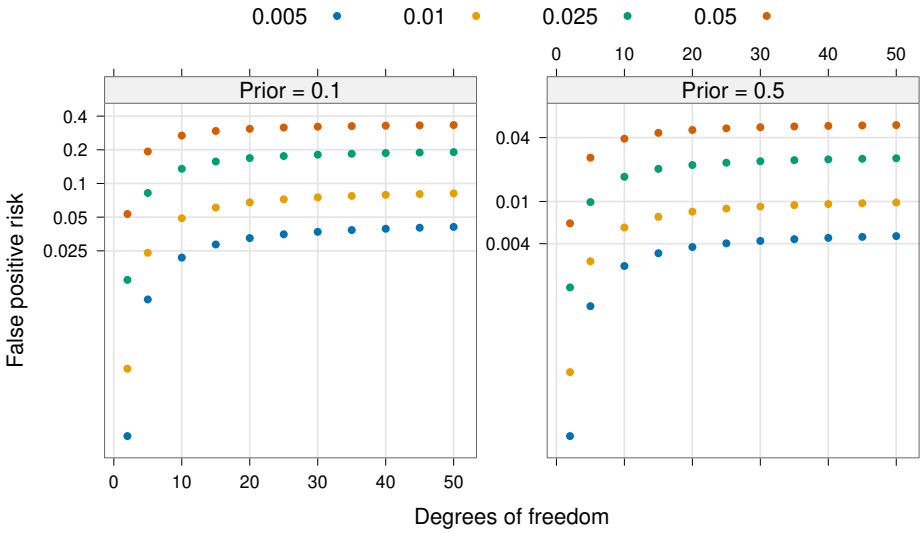

Figure 5: False positive risk, for three different choices of p-value, for a range of sample sizes, and for a range of degrees of freedom.

3 Power – how well does a planned experiment discriminate?

The discussion will assume that we are testing $\mu = 0$ against $\mu > 0$ (one-sided test), or $\mu \neq 0$ (two-sided test). (As noted earlier, it is often more appropriate to use as the baseline a value of μ that is non-zero. Working with a non-zero baseline is simplest for a one-sided test.)

For purposes of designing an experiment, researchers should want confidence that the experiment is capable of detecting differences in the mean, or (for an experiment that generates one-sample data) the mean difference, that are more than trivial in magnitude.

3.1 The power of a t or other statistical test

The power is the probability that, if H1 is true, the calculated p-value will be smaller than a chosen threshold α. Experiments that have low power can waste effort, to little purpose.

For designing an experiment, setting a power is usually done relative to a baseline difference of 0. There is, however, no reason why power should not be set relative to a baseline that is greater than 0. Once experimental results are in, what is more relevant than the power is the minimum mean difference or (for a two-sample test) difference in means that one would like to be able to detect.

Figure 6 is designed to illustrate the notion of power graphically. The densities

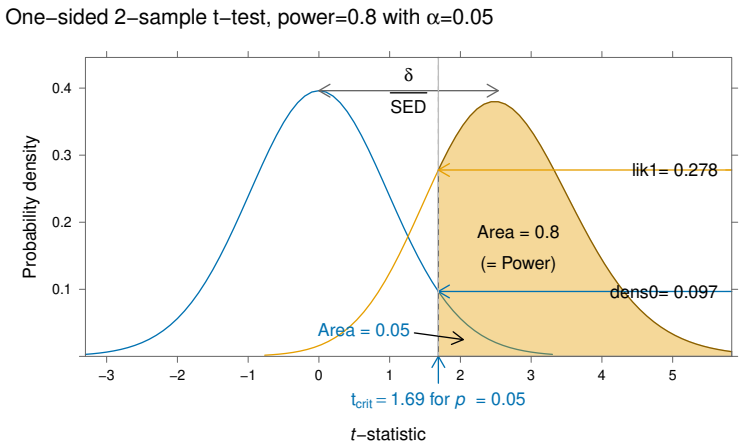

Figure 6: This illustrates graphically, for a one-sided t-test, the t-statistic for the difference in means required to achieve a given power. For this graph, the t-statistic is calculated with 18 degrees of freedom. The two density curves are separated by the amount that gives $power = 0.8$ for $\alpha = 0.05$.

shown are for a two-sample comparison (equal variances) with $n = 19$ in each sample. Calculations proceed by first calculating the separation between means required, with $\alpha = 0.05$, to give a power that equals 0.8, and from this the non-centrality parameter, thus:

```
n <- 19; df <- 2*(n-1); sd <- 1.5; sed <- sd*sqrt(2/n)
## Calculate difference delta between means that gives power=0.8
delta <- power.t.test(n=19, sd=sd, sig.level=0.05,
                      power=0.8, type="two.sample",
                      alternative = "one.sided")[['delta']]
## Calculate the non-centrality parameter ncp
ncp <- delta/sed # sed is Standard Error of Difference
```

The comparison is between densities of t-statistics, both with degrees of freedom 36, the first with noncentrality parameter `ncp` $= 0$, and the second with noncentrality parameter `ncp` $=$ `delta/sed` $= 2.535$.

The same graph will result irrespective of the standard deviation. It is at the same time the graph that will be obtained for a single sample t-test with $n = 37$, now with `delta` equal to the mean difference rather than the difference in means. The two density curves are in each case separated, on the scale of the t-statistic, by the amount required for the test to have a *power* that equals 0.8 for $\alpha = 0.05$.

Here are the calculations:

Once experimental results are obtained and a p-value has been calculated, the alternative of interest is the minimum difference δ in means (or, in the one-sample

case, mean difference) that was set before the experiment as of interest to the researcher.

As an example of a power calculation, suppose that we want to have an 80% probability of detecting, at the $\alpha = 0.05$ level, a difference δ of 1.4 or more. Assume, for purposes of an example, that the experiment will give us data for a two-sample two=sided test. Assume further that the standard deviation of treatment measurements is thought to be around 1. As this is just a guesstimate, we build in a modest margin of error, and take the standard deviation to be 1.5 for purposes of calculating the sample size. We then do the calculation:

```
power.t.test(type='two.sample', alternative='two.sided', power=0.8,
             sig.level=0.05, sd=1.5, delta=1.4)[['n']]
[1] 19.03024
```

With the results in, the relevant alternative to H0, for purposes of calculating a likelihood ratio, has $\delta = 1.4$. Suppose, then, that the experimental results yield a standard deviation of 1.2, assuming that the standard deviation is the same for both treatments.

Figure 7 (left panel) plots maximum likelihood ratios, and likelihood ratios, for the choices $\delta = 1.0$ and $\delta = 1.4$, against p-values. Results are for a two-sample two-sided test with $n = 19$ in each sample. Results are presented for $\delta = 1.0$ as well as for $\delta = 1.4$, in order to show how the likelihood ratio changes when δ changes.

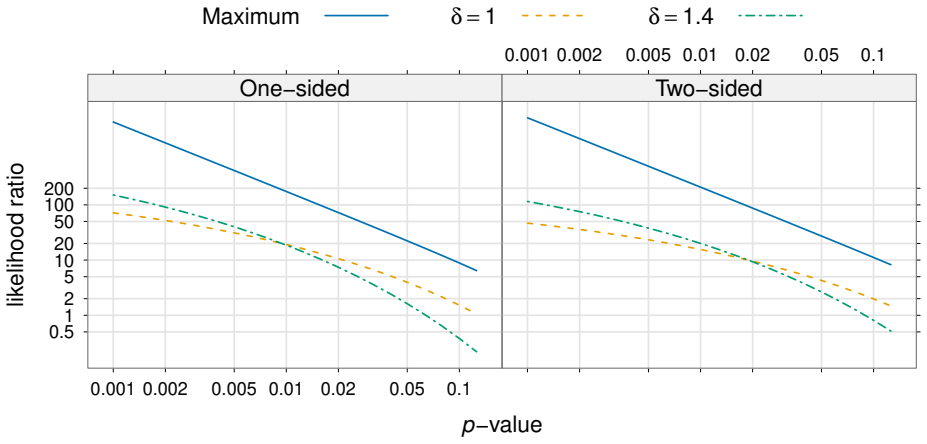

Figure 7: Ratio of likelihood under the alternative to the likelihood under the NULL, as a function of the calculated p-value, with $n = 9$ in each sample in a two-sample test, and with $\delta = 0.6s$ set as the minimum difference of interest. The graph may, alternatively, be interpreted as for $n = 19$ in a one-sample test, now with $\delta = 1.225s$. The left panel is for one-sided tests, while the right panel is for two-sided tests.

The power, calculated relative to a specific choice of α, is an important consideration when an experiment is designed. The aim is, for a simple randomized trial of the type considered here, to ensure an acceptably high probability that a treatment effect δ that is large enough to be of scientific interest, will be detectable given a threshold α for the resultant p-value. Once experimental results are available, the focus should shift to assessing the strength of the evidence that the treatment effect is large enough to be of scientific interest, i.e., that it is of magnitude δ or more.

Any treatment effect, however small, contributes to shifting the balance of probability between the NULL and the alternative. By contrast, the maximum likelihood ratio depends only on the estimated treatment effect. What is really of interest, as has just been noted, is the strength of evidence that the treatment effect is of magnitude δ or more.

3.2 False positive risk, when α is used as cutoff

The use of a cutoff α, as a basis for a decision-making strategy, is a less nuanced use of the evidence than when there is attention to the specific p-value or, equivqlently, to the t-statistic. Assume that experiments are designed to have a power P_w to accept H1 when $p <= \alpha$. Then the false positive risk is:

$$\frac{\alpha(1-\pi)}{\alpha(1-\pi)+\pi P_w}$$

In the case where $\pi = 0.5$, and P_w is 0.8 or more, this is always less than 1.25 α. Note again that what are modeled here are the properties of a strategy for choosing between H0 and H1. Thus, with $\alpha = 0.5$, it makes no distinction between, for example, $p = 0.05$ and $p = 0.01$ or less.

What choices of cutoff α, and of power, make sense?

The conventional choice has been $\alpha = 0.05$, with 0.8 for the power. In recent years, in the debate over reproducibility in science, a strong case has been made for a choice of $\alpha = 0.01$ or $\alpha = 0.005$ for the cutoff. Such a more stringent cutoff makes sense for purposes of deciding on the required sample size. It does not deal with the larger problem of binary decision making on the basis of a single experiment.

A higher power alters the tradeoff between the type I error α, and the type II error $\beta = 1$ - P_w, where P_w is the power. In moving from $P_w = 0.8$ to $P_w = 0.9$ while holding the sample size constant, one is increasing the separation between the distribution for the NULL and the distribution for the alternative H1.

4 How should results be reported?

P-values have come to have a central role in the reporting of scientific results. It is commonly assumed that an individual *p*-value that equals 0.05 provides 19 to 1 evidence against the NULL hypothesis, and in favour of the alternative. Two points are

- The probability to which it most directly relates is the probability of obtaining such a result, given the NULL
- For this purpose, the relevant probability is 0.1, not 0.05

The maximum likelihood ratio for the alternative against the NULL depends on the degrees of freedom. It is less than 4.5 for degrees of freedom greater than 5.

Results should come with evidence of relevant checks on distributional assumptions. Where degrees of freedom are small (e.g., 4 or less), and there is no evidence from comparable data from earlier studies on which to rely, checks are in general unlikely to be effective. The uncertainty that this generates should be acknowledged.

Meaningful data are a richer source of information than can be satisfactorily summarized in a single statistic. Consider the use of multiple forms of statistical summary, each offering its own perspective, and supported by relevant graphs.

5 Further reading and references

See especially Colquhoun (2017), Wasserstein, Schirm, and Lazar (2019), and other papers in the American Statistician supplement in which Wasserstein's editorial appeared. Code used for the calculations is based on David Colquhoun's code that is available from https://ndownloader.figshare.com/files/9795781.

References

Berkson, Joseph. 1942. "Tests of Significance Considered as Evidence." *Journal of the American Statistical Association* 37 (219): 325–35.

Colquhoun, David. 2017. "The Reproducibility of Research and the Misinterpretation of p-Values." *Royal Society Open Science* 4 (12): 171085. https://royalsocietypublishing.org/doi/suppl/10.1098/rsos.171085.

Fisher, Ronald A. 1935. *The Design of Experiments*. Oliver and Boyd.

Resnick, Brian. 2017. "What a Nerdy Debate about p Values Shows about Science – and How to Fix It." *Vox* 31. https://www.vox.com/science-and-health/2017/7/31/16021654/p-values-statistical-significance-redefine-0005.

Wasserstein, Ronald L., Allen L. Schirm, and Nicole A. Lazar. 2019. "Moving to a World Beyond 'p«<0.05'." *The American Statistician* 73 (sup1): 1–19. https://doi.org/10.1080/00031305.2019.1583913.