Inf2 – Foundations of Data Science 2022

Topic: Hypothesis testing and *p*-values

13th January 2023

Recommended reading:

- XKCD comic strip on multiple testing – funny!
- *A hypothesis is a liability* Yanai and Lercher (2020) – thought-provoking and amusing article

1 Principle of hypothesis testing

Hypothesis testing helps us to answer yes/no questions, such as "is chocolate good for you?" or "is a jury selection procedure biased?" There are two aspects to hypothesis testing:

1. Deciding on whether a **hypothesis** or **model** is compatible with data from observational studies and randomised experiments.
2. If the hypothesis is compatible with the data, investigating the mechanisms specific to the data, e.g. the biological effect of chocolate on the body or the process by which a jury panel was selected.

In the course we are going to focus on the statistical aspects (Aspect 1), but it's worth remembering that the question is not answered once we've completed this step – it should prompt further investigation of the question rather than ending the inquiry (Yanai and Lercher, 2020).

Method of hypothesis testing At the core of hypothesis testing are the **null hypothesis** and the **alternative hypothesis**:

The null hypothesis H_0**:** The claim that we initially assume to be true, formalised as a statistical model. e.g. "The jury panel was chosen by random selection from the population in a district."

The alternative hypothesis H_a**:** The claim that is contradictory to H_0, typically not formalised as a statistical model. E.g. "The jury panel was chosen by some other, unspecified, method."

The aim of hypothesis testing is to either reject or not reject the null hypothesis. Note that we do not "accept" the null hypothesis as true, we are just saying that it's not been proved to be false.

Test procedure The procedure to carry out a hypothesis test, which we call the **test procedure**, consists of:

1. Deciding on a **test statistic**, which is a function of the sample data, e.g. the number of Black people in a jury panel.
2. Determining what the distribution of the test statistic would be if it arose from the null hypothesis statistical model.
3. Either:
   (a) Deciding on a **rejection region**, i.e. regions of the distribution of the test statistic under H_0 in which we should reject H_0. Typically, these are the extremities of the distribution. If our test statistic falls into the rejection region, we reject H_0; otherwise, we don't reject it.
   (b) Returning a ***p*-value**, which tells us how compatible the test statistic is with the distribution predicted by chance from H_0.

Application of test procedure to example In the topic on Randomness, sampling and simulation, we looked at the example of Swain versus Alabama (1965), in which the question was "if 8 Black people were chosen for a jury panel of 100 people, but the fraction of Black people in the population was 26%, does this show bias against Black people?" We found the distribution of the test statistic under the null hypothesis by simulating the null hypothesis model of sampling from a Bernoulli distribution with $P(\text{Black}) = 0.26$. In this case probability theory also tells us that the distribution is a binomial distribution with $n = 100$ and $p = 0.26$. We found that there were no replications in which 8 Black members were chosen (Figure 1) – the simulated numbers were always higher.

We did not consider rejection regions or p-values. Since the observed data (8 Black people on the panel; red line in Figure 1) were inconsistent with the range of predictions produced by the null hypothesis, it seemed very clear that we should reject the null hypothesis. But what would we have decided if the number of Black people had sat within the distribution of simulated values, e.g. 15 (magenta line) or 20 (yellow line)?

Rejection regions We might want to specify the rejection region as the bottom 5% of the probability mass, i.e. the region that seems unusually low (Figure 2, left, region to left of orange boundary). If the observed test statistic falls into that region, we might "reject the hypothesis at the 5% level (one-tailed test)". We call this a **one-tailed test** because the rejection region occupies only one tail of the distribution. This is justified, as the alternative hypothesis was implicitly "the number of Black people selected is below the number we would have expected by chance".

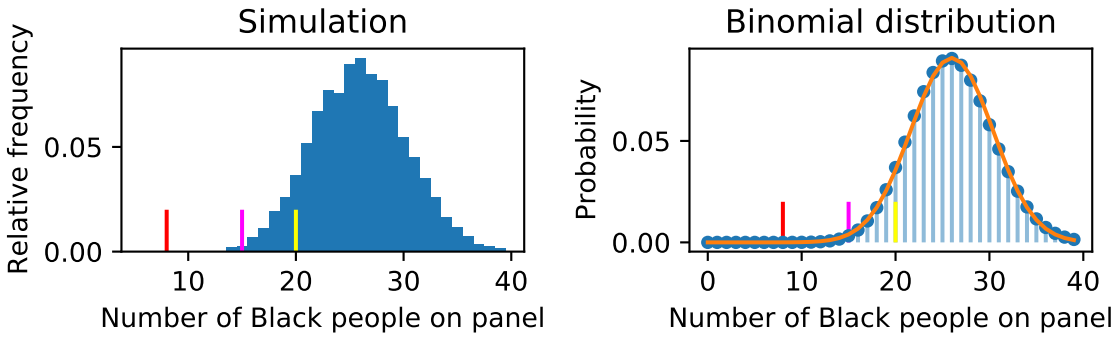

Figure 1: Distributions of number of Black people T_0 (test statistic) on a panel of 100 under the null hypothesis that the jury was randomly selected from a population that is 26% Black and 74% non-Black. Left: distribution arising from 10 000 statistical simulations. The red line indicates the number of Black jurors in Swain versus Alabama (1965), the magenta line indicates $t_0 = 15$ and the yellow line indicates $t_0 = 20$. Right: Binomial distribution (blue dots) for $n = 100$ and $p = 10$. Normal approximation (orange curve) with $\mu = np$ and $\sigma = \sqrt{np(1-p)}$.

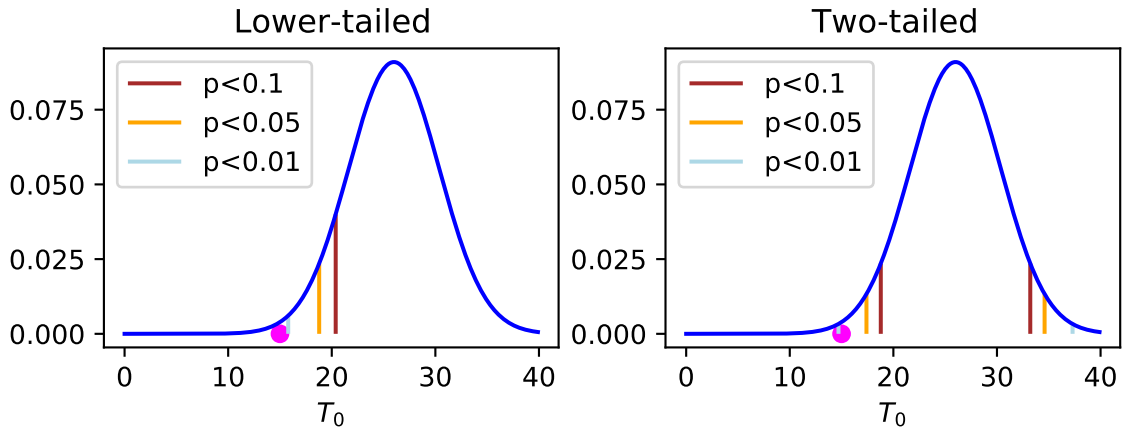

Figure 2: Rejection regions. Lower-tailed (left) and upper-tailed (right) rejection regions are shown for the normal approximation to the distribution of the null hypothesis model in the Swain-Alabama example. The observed statistic $t_0 = 15$ is shown with a magenta dot. It lies in the $p < 0.01$ rejection region for a lower-tailed test and in the $p < 0.05$ rejection region for a two-tailed test.

If we know the distribution of our null hypothesis model, we can look up statistical tables to determine the boundaries of rejection regions. E.g. in this case, the number n is large enough that we can approximate the binomial distribution with a normal distribution with mean $\mu = np$ and variance $\sigma^2 = np(1-p)$. This means that the standardised statistic

$$Z = \frac{T_0 - \mu}{\sigma} \tag{1}$$

is normally distributed. At the edge of the rejection region, this statistic is equal to the z critical value $z_{0.95}$, which has 95% of the probability mass to its right. We can then rearrange Equation 1 to find the edge of the rejection region in terms of the original statistic:

$$T_0 = \mu + \sigma z_{0.95} \tag{2}$$

If a test statistic in a hypothesis test is distributed according to a normal distribution, the hypothesis test is sometimes referred to as a "z-test".

One-talied and two-tailed tests We could have formulated the alternative hypothesis as "the number of Black people selected is different from (i.e. above or below) the number we would have expected by chance". In this case we would perform a **two-tailed test** (Figure 2, right) by setting the rejection regions to be the bottom 2.5% and the top 2.5% of the probability mass of the distribution. We would "reject the hypothesis at the 5% level (two-tailed test)".

2 *p*-values

Principle of *p*-values The principle of p-values is that we set the boundary of the rejection regions to be where the data is, and then report the probability mass in the resulting rejection regions as the ***p*-value**.

Determining *p*-values from statistical simulations Had there been 15 Black people on the panel in Swain versus Alabama (magenta line), a fraction 0.0062 of the 10 000 simulations produced panels with 15 or fewer black members. This would therefore give the p-value $p = 0.0062$, i.e. 0.62%. This certainly calls into question if the observed data is compatible with the null hypotheses.

Suppose that there had been 20 Black people on the jury panel (yellow line in Figure 1). The corresponding rejection region is 20 or fewer Black people on the jury. A fraction of 0.101 of the simulations are in this region, so the p-value is $p = 0.101$. We would tend not to reject the null hypothesis at this size of p-value, but this would not mean that the null hypothesis was true.

Sometimes the p-value is reported relative to a round figure rejection region, e.g. in the case with 15 Black people on the jury, $p < 0.01$, indicating that we could "reject the null hypothesis at the 1% level". However, supplying the actual p-value gives more information than just reporting the rejection region.

Table 1: P-values computed by various methods for various observed values of t_0 in Swain versus Alabama (1965).

	t_0	Simulation	Binomial	Normal
0	8	0	4.73e-06	2.03e-05
1	15	0.0067	0.0061	0.0061
2	20	0.1020	0.1030	0.0857

Determining *p*-values from probability distributions Sometimes it is straightforward to compute the probability distribution implied by the null hypothesis. In the Swain versus Alabama example, it is a binomial distribution with $n = 100$ and $p = 0.26$ (Figure 1, right). As we are looking at a lower-tailed test, the p-value is the cumulative distribution function of the binomial distribution, cut off at t_0, the observed number of Black people on the jury panel:

$$P(T_0 \leq t_0) = B(t_0; n, p) = \sum_{t=0}^{t_0} b(t; n, p) \tag{3}$$

where $b(t; n, p)$ is the probability of t successes in a binomial distribution with n trials and success probability p; $B(t; n, p)$ is the corresponding cumulative distribution function (cdf). Stats packages have functions to compute the cdf for various distributions, and the values for the binomial are shown in Table 1 along with the simulated values.

Also shown is the normal approximation to the binomial, in which we set $\mu = np$ and $\sigma = \sqrt{np(1-p)}$. The p-values are the values of the normal cumulative distribution function at the standardised value

$$z = \frac{t_0 - \mu}{\sigma} \tag{4}$$

Why use rejection regions? The rejection region method works well with printed statistical tables, in which critical values of z and other distributions are available only for particular cut-off values, e.g. 0.01, 0.05. With computer packages it is now possible to define the rejection region relative to the observed data rather than a pre-set cut-off.

Definition of *p*-value We can define the p-value as follows:

> The p-value is the probability, calculated assuming the null hypothesis is true, of obtaining a value of the test statistic at least as contradictory to H_0 as the value calculated from the available sample. (*Modern Mathematical Statistics with Applications*, p. 456)

The whole topic of the interpretation and use of p-values is complex and highly contested. In fact, it took 20 statisticians 2 days and many subsequent days of drafting to produce the American Statistical Association's statement on p-values: The statement by the American Statistical Association (Wasserstein and Lazar, 2016).

What *p*-values are We quote 2 of the 6 points in the statement here. Firstly, what p-values are:

> ***P*-values can indicate how incompatible the data are with a specified statistical model**...
>
> The smaller the p-value, the greater the statistical incompatibility of the data with the null hypothesis, if the underlying assumptions used to calculate the p-value hold. This incompatibility can be interpreted as casting doubt on or providing evidence against the null hypothesis or the underlying assumptions. (*ASA Statement on Statistical Significance and P-values*)

In the Swain versus Alabama example where we imagined there were 15 Black people on the jury, the small p-value ($p = 0.0062$) indicates that the data (here 15 Black people on the panel) are quite incompatible with the null hypothesis statistical model (here that Black and non-Black people were drawn from the population at random). The low p-value casts considerable doubt on the hypothesis. Of course the actual data ($t_0 = 8$) has a vanishingly small p-value (Table 1).

What *p*-values are not Secondly, what they are not:

> ***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.**
>
> Researchers often wish to turn a p-value into a statement about the truth of a null hypothesis, or about the probability that random chance produced the observed data. The p-value is neither. It is a statement about data in relation to a specified hypothetical explanation, and is not a statement about the explanation itself. (*ASA Statement on Statistical Significance and P-values*)

"Statistical significance" A widespread practice in scientific literature is to take p-values of less than $p = 0.05$ as indicating **statistical significance**, i.e. that the null hypothesis should be rejected. Values of less than 0.05 indicate weak evidence against the null hypothesis. Sometimes higher thresholds are used, e.g. $p = 0.01$ and $p = 0.001$. In scientific papers and the output from stats packages you will sometimes see these values indicated with asterisks:

- * means significant at least at the $p < 0.05$ level
- ** means significant at least at the $p < 0.01$ level
- *** means significant at least at the $p < 0.001$ level

There is no "correct" answer about what the right level of significance is. The $p < 0.05$ value was suggested in a paper by the statistician Ronald Fisher[1], who invented the hypothesis test,

[1]Fisher studied under Pearson, and developed a huge body of modern statistics. He also edited the *Annals of Eugenics* and had controversial views on race.

	Caucasian	Black/AA	Hispanic	Asian/PI	Other	Total
Population %	54	18	12	15	1	100
Observed panel numbers	780	117	114	384	58	1453
Expected panel numbers	784.62	261.54	174.36	217.95	14.53	1453.00
$\frac{(\text{Observed}-\text{Expected})^2}{\text{Expected}}$	0.03	79.88	20.90	126.51	130.05	357.36

Table 2: Alameda County jury panel data. The top row shows the estimated proportions of 5 ethnic groups (Caucasian, Black/African American, Hispanic, Asian/Pacific Islander and Other) in Alameda County. The second row (Observed panel numbers) shows the total number in each group on 11 jury panels from 2009–2010. There was a total of 1453 on the 11 jury panels (final column). The third row (Expected panel numbers) shows the numbers we would expect from each group if the panels had been selected randomly from the population. The final row $\frac{(\text{Observed}-\text{Expected})^2}{\text{Expected}}$ shows the disparity between the observed and expected using this formula. The total disparity is in the final column.

but it simply seemed "convenient" to him for his purposes. As in the discussion on confidence intervals (How big should a confidence interval be?), the value we choose to use may depend on the application. For example, we would demand a very low p-value when testing the null hypothesis that a new drug has no effect on the death rate of patients. We might accept a slightly higher p-value for the hypothesis that it has no positive effect on symptoms. In less mission-critical scientific applications, a higher p-value will be acceptable.

3 Testing for goodness of fit to a model

Multiple categories In the example so far, there have been just two categories: Black and non-Black. In 2010 the North California branch of the American Civil Liberties Union (ACLU) investigated the numbers of Caucasian, Black/African American, Hispanic, Asian/Pacific Islander and Other people on jury panels in Alameda County. The found the data shown in the first two rows of Table 2.

We want to test the following null and alternative hypotheses, which are essentially the same as for the case with two categories:

The null hypothesis H_0**:** The jury panels were chosen by random selection from the population in a district.

The alternative hypothesis H_a**:** The jury panels were chosen by some other, unspecified, method.

With two categories, it's easy to see that the number of Black people could be a test statistic. But in this case, there are 4 numbers that describe the outcome of any simulation (we can always compute the number in the 5th category if we know the total number and the numbers in 4 categories). We can't have 4 test statistics, so we need to create a statistic that indicates the disparity between the observed and expected outcomes.

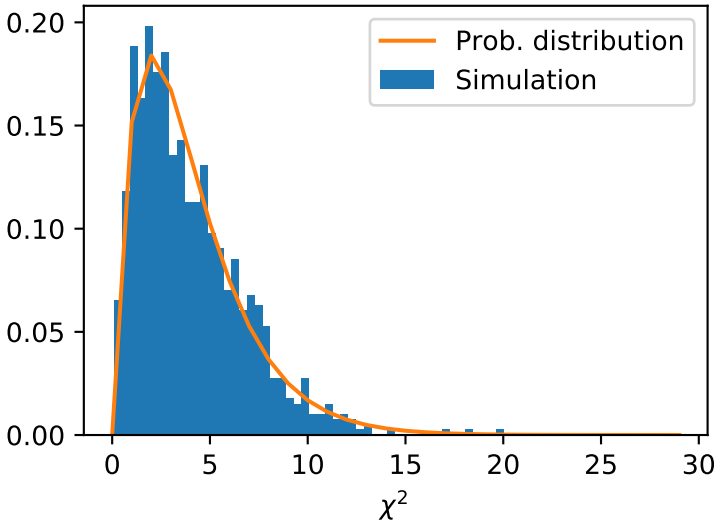

Figure 3: Distribution of χ^2 for jury panel selection in Alameda County. Simulations shown in blue and theoretical χ^2 distribution with 4 degrees of freedom shown in orange.

Suppose we call the population proportions of each of k groups p_i and the observed numbers in each group n_i. The total number sitting on jury panels is $n = \sum_i n_i$. We can compute the numbers we would expect to be on jury panels as np_i (third row of table). One measure of disparity would be the sum of the squared differences:

$$\sum_{i=1}^{k} (n_i - np_i)^2$$

This looks at the *absolute* squared differences between the expected and observed values for each category. If we expected $np_1 = 100$ in one category and observed $N_1 = 95$, this expected-observed pair would contribute 25 to the sum. A difference of $np_2 = 10$ (expected) and $N_2 = 5$ (observed) would also contribute 25 to the sum. However, in *relative* terms, the difference between the first expected-observed pair is 5%, whereas in the second pair it is 50%.

This motivates us to look at the scale the disparity measure by dividing by the expected number in each category, to create a statistic that we call **chi-squared**, written using the Greek symbol χ^2:

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} \tag{5}$$

The components of the χ^2 statistic are seen in the final row of Table 2, as is the value of $\chi^2 = 357.36$ for the observed values (in the "Total" column).

Statistical simulation We can now run a statistical simulation to generate the expected distribution of χ^2. For each repetition we simulate the numbers in each category by drawing

[2]Sometimes you may see the letter X used instead of χ.

from a multinomial distribution with parameters n and p_i. We then compute and store χ^2 for that simulation, which gives us the simulated distribution shown in blue in Figure 3. We can immediately see that the observed value of $\chi^2 = 357.36$ is off the scale of the graph, indicating that it has a much bigger value than is compatible with the null hypothesis, so we reject the null hypothesis.

Chi-squared distribution It turns out that, providing every expected value np_i is greater than 5, the χ^2 statistic is distributed approximately according to a χ^2 probability distribution with $k-1$ degrees of freedom, shown in orange in Figure 3. The fit between the probability distribution the simulated distribution is clear.

Goodness-of-fit Large values of χ^2 statistic (i.e. the upper tails of the distribution) indicate a poor **goodness-of-fit** between the model and the data. χ^2 tests therefore tend to be **upper-tailed**. However, the statistic had a very low χ^2, we might be suspicious that the data had been fiddled with.

The χ^2 statistic can be used to assess the goodness-of-fit of many types of model and data, not just this proportion example. If we find a χ^2 with a p-value greater than desired cut-off, this suggests that we should not reject the model.

Testing for independence with two-way contingency tables We may have multiple populations (e.g. males and females) and multiple categories (e.g. depressed or not depressed). We can arrange these in a **two-way contingency table** (Table 3).

We want to test the null hypothesis that being depressed is independent of if you are male of female. In other words $P(X = x, Y = y) = P(X = x)P(Y = y)$. Using a notation similar to Table 3 (right), we can write this probability as $p_{ij} = p_{i\bullet}p_{\bullet j}$, where $p_{i\bullet}$ is the marginal probability of an item being in category i and $p_{\bullet j}$ is the marginal probability of an item being in category j. Our best estimates of the marginal probabilities are

$$p_{i\bullet} = \frac{n_{i\bullet}}{n_{\bullet\bullet}} \text{ and } p_{\bullet j} = \frac{n_{\bullet j}}{n_{\bullet\bullet}} \tag{6}$$

Therefore the best estimates of the number of in each cell are

$$\hat{e}_{ij} = n_{\bullet\bullet}p_{ij} = \frac{n_{i\bullet}n_{\bullet j}}{n_{\bullet\bullet}} \tag{7}$$

The χ^2 statistic is computed as

$$\chi^2 = \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}} = \sum_i \sum_j \frac{(n_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}} \tag{8}$$

In this case it is 4.433.

We assume that the numbers of depressed and non-depressed, and males and females are fixed. In general, if there are I rows and J columns in the table there are $(I-1)(J-1)$ degrees

	Female	Male	Total
Depressed	30	12	42
Not depressed	2048	1663	3711
Total	2078	1675	3753

	Population 1	Population 2	Total
Category 1	n_{11}	n_{12}	$n_{1\bullet}$
Category 2	n_{21}	n_{22}	$n_{2\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	$n_{\bullet\bullet}$

Table 3: Left: Contingency table of the number of depressed and not depressed people in a population of females and males; data based on a prospective study Bornioli et al. (2020). Right: General symbolic version of the two-way contingency table. There are I rows and J columns. The number of items falling into a cell in the ith row and jth column is denoted n_{ij}. The total in the ith row is denoted $n_{i\bullet} = \sum_{j=1}^{J} n_{ij}$, the total in the jth column is $n_{\bullet j} = \sum_{j=1}^{I} n_{ij}$. The grand total is $n_{\bullet\bullet} = \sum_i n_{i\bullet} = \sum_j n_{\bullet j}$.

	Female	Male
Depressed	23.25	18.75
Not depressed	2054.75	1656.25

	Population 1	Population 2
Category 1	$\hat{e}_{11}$	$\hat{e}_{12}$
Category 2	$\hat{e}_{21}$	$\hat{e}_{22}$

Table 4: Expected numbers in contingency table, in example (left) and in symbols from Equation 7 (right).

of freedom. In this case there is therefore only 1 degree of freedom; specifying n_{11} (or any other cell) allows us to compute the values of all the other cells. We therefore look up the cumulative distribution function of χ^2 with 1 degree of freedom, to find that $p < 0.035$, so this is significantly different from independence at the 5% level.

4 Issues in hypothesis testing

Type I and Type II errors Regardless of whether we use a one-tailed test or a two-tailed test, *if* the null hypothesis were true, there is 5% chance that an observed test statistic in the rejection region might really have arisen by chance. By rejecting H_0, we would have made a **Type I error**: rejecting the null hypothesis when it is true. To reduce the risk of making a Type I error, we could make the rejection region smaller, e.g. the bottom 1%. However, we would also have increased the chance of making a **Type II error**: not rejecting the null hypothesis when it is false.

There is no right answer about what size of rejection region to use – it depends on what the consequences of Type I versus Type II errors are.

Decisions based on confidence intervals

> Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold. (ASA Statement, point 5). (Wasserstein and Lazar, 2016)

$p \leq 0.05$ does not mean that the false; it is one point in a spectrum. However, it is often seen as the "holy grail" of scientific research.

Data snooping and *p*-value hacking It is very tempting to try out many experiments in order to get a *p*-value of less than 0.05. However, the more experiments are run, the more chance there is of Type I errors – i.e. rejecting the null hypothesis when it is true.

"Data snooping" or "*p*-value hacking" is the practice of rerunning experiments or selecting subsets of datasets until a statistically significant result is achieved. It is harder to publish negative results than positive results in academic journals, so there is an incentive to data snoop. Some statistically significant results in the literature will be Type I errors, which makes it important to replicate experimental results.

The ASA statement says:

> **Proper inference requires full reporting and transparency.**
> *P*-values and related analyses should not be reported selectively. Conducting multiple analyses of the data and reporting only those with certain *p*-values (typically those passing a significance threshold) renders the reported *p*-values essentially uninterpretable. Cherry-picking promising findings, also known by such terms as data dredging, significance chasing, significance questing, selective inference, and "p-hacking," leads to a spurious excess of statistically significant results in the published literature and should be vigorously avoided... (*ASA Statement on Statistical Significance and P-values*)

Multiple testing Suppose we undertake multiple tests on the same dataset is problematic, and find that one of the tests is significant. As we increase the number of tests, the probability of a Type I error increases (XCKD comic in reading). If we undertake 20 tests, there's a 0.95^{20} chance of not having a Type I error, and therefore a $1 - 0.95^{20} = 0.64$ chance of a type I error. There are ways to compute more stringent cut-offs in these cases, for example the Bonferroni correction.

References

Bornioli, A., Lewis-Smith, H., Slater, A., and Bray, I. (2020). Body dissatisfaction predicts the onset of depression among adolescent females and males: a prospective study. *J Epidemiol Community Health*.

Wasserstein, R. L. and Lazar, N. A. (2016). The ASA statement on *p*-values: Context, process, and purpose. *The American Statistician*, 70(2):129–133.

Yanai, I. and Lercher, M. (2020). A hypothesis is a liability. *Genome Biol*, 21:231.