9 Hypothesis Testing

It's easy to get the right answer if you never define what the question is.
—Jerzy Neyman

9.1 Null Hypothesis Significance Testing

My son Andrew is a picky eater and claims that he can taste whether I use his favorite brand of hot sauce or a different one when I make chili. I think he can't tell the difference. He claims he can prove it, once and for all, with an experiment, so I cook up two otherwise-identical batches, one with his preferred sauce and one with the brand he claims to dislike. Out of his sight, I flip a coin to see which one to sample, and then I give him a taste of it. He tries to guess which one it is. We repeat the coin flip and the taste test 30 times, and I record his taste-testing performance.

It seems reasonable to assume that each test X_i is i.i.d. Bernoulli distributed with some unknown probability θ, so the number $T = \sum_{i=1}^{n} X_i$ of times he gets the right answer should be binomially distributed. If Andrew cannot tell the difference between the two hot sauces, then $\theta = \frac{1}{2}$. We call this the *null hypothesis*, denoted H_0. If the null hypothesis H_0 holds, then

$$T \sim \text{Binomial}\left(30, \frac{1}{2}\right).$$

In my particular test Andrew is right 21 times and wrong 9 times. That's better than 50/50, but maybe this just happened by chance. What is the probability he was just lucky, assuming that he really can't tell the difference at all? The probability that he gets exactly 21 right is very low, of course, but so is the probability that he'll get exactly half right. A better question to ask is what is the probability that he got *at least* 21 right by chance? That is, let's compute

$$p = P(T \geq 21 \,|\, H_0).$$

This number is called the *p-value* of the test. We can compute this directly as

$$\begin{aligned} p &= P\left(T \geq 21 \,\middle|\, H_0\right) \\ &= P\left(T \geq 21 \,\middle|\, T \sim \text{Binomial}\left(30, \frac{1}{2}\right)\right) \\ &= \sum_{t=21}^{30} \binom{30}{t} \left(\frac{1}{2}\right)^t \left(1 - \frac{1}{2}\right)^{30-t} \\ &= 2^{-30} \sum_{t=21}^{30} \binom{30}{t} \\ &\approx 0.0213. \end{aligned}$$

The probability of seeing these results, given the null hypothesis, is only 2%, which seems like a reason to reject my null hypothesis that he is randomly guessing. It may be that he really can tell the difference, at least better than just a random guess.

9.1.1 Null Hypothesis Significance Testing

The previous example is typical of much of hypothesis testing in science. A researcher proposes that something has an effect or can be identified in some way. Call this hypothesis H_1. The default assumption H_0 is that H_1 is false. The hypothesis H_0 is called the *null hypothesis*. The steps of *null hypothesis significance testing (NHST)* are as follows.

(i) Design an experiment to test the hypothesis H_1. The experiment should produce data $X_1, \ldots, X_n$ that can be modeled statistically under the assumption that H_0 holds.

(ii) A statistic $T(X_1, \ldots, X_n)$ is a function of the data $X_1, \ldots, X_n$. Identify a statistic T whose statistical distribution is known in the case that H_0 is true and H_1 is false. That is, we want to be able to compute the conditional c.d.f. $P(T \leq t \,|\, H_0)$. Traditional statisticians have developed a large number of different test statistics for a large number of different types of experiments, including a "chi-squared test," a "t-test," an "F-test," and a "z-test," among many others. It is common for researchers to call the whole NHST process by the name of the test statistic being used, but the process is essentially the same, regardless of the choice of test statistic.

(iii) Decide an upper bound α such that if $p < \alpha$, then the null hypothesis should be rejected. The value of α is sometimes called the *significance level* of the test. Note that one should choose α before looking at the data; otherwise, there is a temptation to cheat and choose α in a way that gives the conclusion you want to see.

(iv) Run the experiment and collect the data $\mathbf{x} = (x_1, \ldots, x_n)$. The *p-value* is the probability of observing the result $\mathbf{x}$ *or something more extreme*, given the null hypothesis.

(v) If $p < \alpha$ (as decided in (iii)), the null hypothesis is rejected.[35] Otherwise, the null hypothesis is retained.[36].

If the null hypothesis is retained with $p \geq \alpha$, then H_1 is considered unnecessary to explain the results. But if the null hypothesis is rejected, with $p < \alpha$, then the experiment indicates that something interesting might be going on. If the experiment was well designed, and the statistical model is a good fit, then the data might have provided some support for believing H_1.

Example 9.1.1. In the case of Andrew and his hot sauce, the hypothesis H_1 is that he can identify the difference in sauce by taste. The null hypothesis H_0 is that he cannot. The steps of NHST in this case are as follows.

(i) The experiment is his blind tasting of 30 randomly sampled batches. The null hypothesis H_0, that he can't tell the difference, yields a probabilistic model for each prediction X_i, namely that the X_i are i.i.d. with $X_i \sim \text{Bernoulli}\left(\frac{1}{2}\right)$.

(ii) Assuming the null hypothesis H_0, the statistic $T(X_1, \ldots, X_{30}) = \sum_{i=1}^{30} X_i$ is binomially distributed as $T \sim \text{Binomial}\left(30, \frac{1}{2}\right)$.

(iii) Before running the experiment we should have chosen (iii) a value of α. A common choice is $\alpha = 0.05$.

(iv) Our experiment gave the result $T = 21$ and in this setting the definition of "more extreme than 21" is $T \geq 21$, so the p-value is $p = P(T \geq 21 \,|\, H_0)$, which comes out to 0.0213

(v) Since $p < \alpha = 0.05$, the null hypothesis is rejected, and we conclude that H_0 does not seem to adequately explain the results of the experiment. That is, Andrew's taste test is probably not explained by random chance, so the probability of his guessing the right hot sauce could be greater than $\frac{1}{2}$.

[35] When the null hypothesis is rejected, people often say that the results are *statistically significant*, but this is a misleading expression that seems to suggest that the results are actually meaningful, useful, or important, which is not at all what it means. Thus the expression *statistically significant* should generally be avoided.

[36] When H_0 is retained, people often say that the results are *not statistically significant*, but again, this is a misleading expression that should be avoided

9.1.2 Example: One-sample t-test

I grow grapes in my yard, but I have problems with an infestation by a destructive leafhopper that reduces the productivity of my vines. On average, I have counted 20 of the pests on each leaf. I have been told that spraying a mild solution of soap and water on the leaves at a certain time of year will kill some of the leafhoppers and reduce the severity of the infestation, and that the improvement can be seen after 2 weeks. It seems reasonable to assume that the number of leafhoppers on each leaf before spraying is normally distributed with mean 20, but an unknown variance.[37]

The new hypothesis H_1 that I'd like to test is that two weeks after spraying the soap solution, the number of leafhoppers on each leaf is still normally distributed, but with mean less than 20. The null hypothesis is that even after using the soap solution, the number of leafhoppers remains normally distributed with mean 20.

(i) My experiment will be to spray all the vines with the soap solution, wait two weeks, and then count the number of leafhoppers on 5 randomly chosen leaves to get data $X_1, \ldots, X_5$.

(ii) If the null hypothesis is true, then the sample mean $\widehat{\mu}_5 = \frac{1}{5}\sum_{i=1}^{5} X_i$ should be distributed as $\mathscr{N}(20, \frac{1}{5}\sigma^2)$. So I might want to use want to $\widehat{\mu}_5$ as the test statistic, but since σ^2 is not known, I can't adequately describe the probability distribution of $\widehat{\mu}_5$. I need another test statistic instead. One natural thing to do is to estimate σ^2 from the data and standardize to get the statistic T_n (here $n = 5$):

$$T_n = \frac{\widehat{\mu}_n - \mu}{s/\sqrt{n}}, \tag{9.1}$$

where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \widehat{\mu}_n)^2$ is the unbiased sample variance. If H_0 is true and if s^2 were the true variance σ^2, then T_n would have a standard normal distribution. Unfortunately $s^2 \neq \sigma^2$, but it can be shown that if μ is the true mean, then T_n has a *Student t-distribution* with parameter $\nu = n - 1 = 4$. Thus the null hypothesis (that the true mean μ is 20) implies that

$$T_5 \sim \mathscr{T}(4).$$

(iii) I check the agricultural research literature and see that it is common to choose $\alpha = 0.05$, so I will reject the null hypothesis if $p < 0.05$.

(iv) The p-value is the probability that my statistic is equal to or more extreme than the value observed in the experiment. But what is "more extreme?" In this case, I don't think it is very likely that the soap actually increases the number of leafhoppers, so I would consider "more extreme" to be less than the number observed. Thus the p-value will be

$$p = P(T_5 \leq t \,\big|\, H_0) = P(T_5 \leq t \,\big|\, T_5 \sim \mathscr{T}(4)).$$

[37] Normal may not be the best choice here. A Poisson distribution with mean $\lambda = 20$ might be a better choice, since this is about counting. But the count of leafhoppers may not quite meet the conditions for a Poisson distribution to apply, so there's a lot to think about here. But a Poisson with λ this large is approximately normal, so a normal approximation seems like a reasonable first choice.

I perform the experiment by spraying all my vines with the soap solution. After two weeks, I sample 5 leaves and count the number of leafhoppers on each leaf. The results are 9, 13, 24, 16, 21. The mean is only 16.6 and the (unbiased) sample variance is $s_{\mathbf{x}} = 6.025$. Thus, the t-statistic evaluates to

$$T_{\mathbf{x}} = \frac{\widehat{\mu}_{\mathbf{x}} - 20}{s_{\mathbf{x}}/\sqrt{5}} = -1.261.$$

Since $\widehat{\mu}_{\mathbf{x}} < 20$ and $T_{\mathbf{x}}$ is negative, it seems like the soap helped. But maybe the results are just due to chance. The p-value is

$$\begin{aligned} p &= P(T_5 \leq -1.261 \,\big|\, T_5 \sim \mathscr{T}(4)) \\ &= \int_{-\infty}^{-1.261} f_T(x)\,dx \\ &= F_T(-1.261) \\ &= 0.137. \end{aligned}$$

(v) The value of p is much greater than α, so I retain the null hypothesis. This means I conclude that H_0 can adequately explain the data, so it is unnecessary to resort to H_1 to explain the data. That is, there is no reason to assume an effect due to the soap. I accept that the values I observed could just be due to chance. It's too bad that I don't get a reliable way to reduce the number of leafhoppers, but at least I won't have to spend time and money spraying soap on my vines for nothing.

The version of NHST I did for my leaf hoppers is called a *one-sample t-test*, because I used a draw $\mathbf{x} = (x_1, \ldots, x_n)$ of one sample of length $n = 5$, and used the t-statistic $T_{\mathbf{x}} = \frac{\widehat{\mu}_{\mathbf{x}} - \mu}{s_{\mathbf{x}}/\sqrt{n}}$.

Nota Bene 9.1.2. Beware that $p < \alpha$ does not imply that H_1 is true—it only suggests that H_0 alone might not be adequate to explain the data. In the special case of Andrew and the hot sauce, it gives some evidence for Andrew's ability to taste the difference, because the only reasonable alternative to H_0 in this setting is that θ (the probability of a correct guess) is greater than $\frac{1}{2}$. But in general the value of p says little or nothing about the probability that H_1 is true. In the leafhopper example, p was greater than α, so I retained the null hypothesis. But if p had been less than α for the leafhoppers, it would not necessarily mean that the soap had necessarily worked, but only that random chance and the null hypothesis alone were not enough to explain the results of the experiment.

Nota Bene 9.1.3. The p-value is the probability that results as extreme or more extreme could have been observed, given the null hypothesis. Beware that the p-value is

(i) Not the probability that the null hypothesis is correct.

(ii) Not the probability that the alternative hypothesis H_1 is wrong.

(iii) Not the probability that the experiment cannot be replicated.

Many people mistakenly believe one or more of these things about the p-value. And even among those who know, theoretically, that these things are not true, many still use the p-value in their decision making as if these fallacies were true. For more on these fallacies and their consequences, see Section 9.3.

9.1.3 Example: Two-Sample t-Test

There are many other popular tests used for null hypothesis significance testing (NHST) beyond the binomial test (used above for chili and hot sauce) and the one-sample t-test. One of these is the two-sample t-test, where measurements for two different treatments X and Y are taken. For example, instead of spraying all my vines, maybe I spray only half my vines with soap, but leave half unsprayed and then take samples of each. In this case the null hypothesis is that the difference between the two treatments is normally distributed around 0.

The two-sample t-test uses the following test statistic

$$T = \frac{\widehat{\mu}_X - \widehat{\mu}_Y}{\sqrt{s_X^2/n_X + s_Y^2/n_Y}}, \tag{9.2}$$

where n_X and n_Y are the lengths of samples X and Y, respectively, $\widehat{\mu}_X$ and $\widehat{\mu}_Y$ are the sample means of X and Y, respectively, and s_X and s_Y are the unbiased sample variances of X and Y, respectively. In this case one can show that $T \sim \mathscr{T}(\nu)$, where

$$\nu = \frac{\left(\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}\right)^2}{\frac{1}{n_X - 1}\left(\frac{s_X^2}{n_X}\right)^2 + \frac{1}{n_Y - 1}\left(\frac{s_Y^2}{n_Y}\right)^2}.$$

In the special case that $n_X = n_Y = n$, then this reduces to $\nu = 2n - 2$.

The statistic T is considered more extreme than some value t if it is farther from 0 than t, that is, if $|T| > |t|$. Thus, the p-value for a given observation $t = T_{\mathbf{x}}$ is

$$\begin{aligned} p &= P(|T| \geq |t| \,\big|\, H_0) \\ &= P(|T| \geq |t| \,\big|\, T \sim \mathscr{T}(\nu)) \\ &= F_T(-|t|) + 1 - F_T(|t|) = 2F_T(-|t|), \end{aligned}$$

where the last equality follows from the fact that the p.d.f. of $\mathscr{T}(\nu)$ is symmetric around $t = 0$. If $p < \alpha$, the null hypothesis is rejected, and otherwise it is retained.

This is an example of what the marketing community often calls *A/B testing*, where one alternative (A) is compared to another alternative (B).

9.1.4 Example: Pearson Chi-Squared Test

Another common NHST test is the *Pearson chi-squared* test, which is used to test whether a collection of data has been drawn from a given distribution. One example would be to test whether a die was fair by rolling it n times and recording the outcomes $X_1, \ldots, X_n$. The null hypothesis is that X is uniformly distributed on the set $\{1, \ldots, 6\}$. Let N_j denote the number of X_i that are equal to j. Pearson showed that, as $n \to \infty$, the statistic

$$T = T(X_1, \ldots, X_n) = \frac{6}{n} \sum_{j=1}^{6} \left(N_j - \frac{n}{6} \right)^2$$

converges in distribution to a particular gamma distribution $\text{Gamma}(\frac{5}{2}, \frac{1}{2})$ called the *chi-squared distribution with* 5 *degrees of freedom*. A perfect fit would give $N_j = \frac{n}{6}$, so T would be 0. Thus observations of T are more extreme as they become larger. That means that when n is sufficiently large, the p-value for a given draw $\mathbf{x} = (x_1, \ldots, x_n)$ is

$$p = P(T \geq T_{\mathbf{x}} \,|\, H_0) \approx P\left(T \geq T_x \,\middle|\, T \sim \text{Gamma}\left(\frac{5}{2}, \frac{1}{2} \right) \right).$$

If $p < \alpha$, then the null hypothesis, that X is uniformly distributed, is rejected, and we conclude the die is not fair.

9.1.5 Example: A/B Testing of Conversion Rate

A company website has two different possible webpages for a given product. The goal is to get customers to buy the product, and the question is whether one of the two webpages gets more customers to click than the other.

(i) The experiment is to take some customers and randomly assign a of them to page A and the remaining b of them to page B. We'll track the number a_Y of customers that bought from page A and the number b_Y that bought from page B. Also set $a_N = a - a_Y$ to be the number of customers that were shown page A but did not buy, and $B_N = b - b_Y$ the number on page B that did not buy.

(ii) We need a test statistic T whose distribution we can describe under the null hypothesis. The null hypothesis is that it doesn't matter which webpage is visited—customers are equally likely to buy the product from either page. If the two pages are equally likely to convert a customer to buy, then H_0 is the hypothesis that out of the $a+b$ customers, exactly a_Y+b_Y were going to buy, regardless of the website they went to. Let $T = a_Y$ be the number of people who bought from page A. We can describe the probability of seeing $T = a_Y$ in terms of drawing balls from an urn with green (no buy) and red (buy) balls. The urn has $a + b$ balls, of which $a_Y + b_Y$ are red, and $a_N + b_N$ are green. The results of the two webpages correspond to drawing a balls and getting a_Y red and a_N green. The remaining b balls must then have the b_Y that are red and b_N that are green. This is a case of the hypergeometric distribution (see Definition 4.6.1), so the probability of seeing the given result is

$$P(T = a_y) = \frac{\binom{a_Y+b_Y}{a_Y}\binom{a_N+b_N}{a_N}}{\binom{n}{a_Y+a_N}}$$

(iii) For purposes of this example, let's take the significance level to be $\alpha = 0.11$ (an arbitrary choice).

(iv) We now run the experiment with $n = 34$ and find the following result:

	No	Yes
A	8	7
B	4	15

The probability that $T = 7$ is

$$P(T = 7) = \frac{\binom{7+15}{7}\binom{8+4}{8}}{\binom{34}{7+8}} \approx 0.045,$$

but we also need the probability that T is "more extreme." Page A performed worse in this experiment than page B, so more extreme could mean that A is even worse than it is here, that is $T \leq 7$. But more extreme could also include cases where A by chance does better than B. We'll say that such cases are more extreme than the results of our experiment if they are less likely. A quick brute force search of all possible values of T shows that values of $T \geq 13$ are less likely than the value $T = 7$ that we have observed. This means the p-value for this test is

$$p = P(T \leq 7 \text{ or } T \geq 13 \,\big|\, H_0) = F_T(7) + (1 - F_T(12)) \approx 0.075.$$

Here F_T is the c.d.f. for the hypergeometric distribution with parameters $r = a_Y + b_Y = 22$, $g = a_N + b_N = 12$, and $n = a = 15$.

(v) Since $p < \alpha$, we reject the null hypothesis that these two pages have the same conversion rate. That is, we conclude that it is unlikely the data we observed could have occurred under the null hypothesis. Consequently, we want another explanation for the data. Since A performed worse in our test, it seems reasonable to conclude that B probably has a better conversion rate than A.

9.2 Power

9.2.1 Types of Errors

If $p < \alpha$, we reject the null hypothesis, but it is still possible that H_0 holds and the data occurred by chance. In fact, since $p = P(T \geq t \,|\, H_0)$, then the probability is p that the observed data happened by chance, assuming H_0. In this case, by falsely rejecting the null hypothesis, we have made an error, often called a *type I error*.[38] We denote by a the probability of falsely rejecting the null hypothesis, given that the null hypothesis is true.

Alternatively, if $p \geq \alpha$ we conclude that the null hypothesis adequately explains the observed data, and thus we retain the null hypothesis. If the null hypothesis is false in this situation, we have made the error of falsely retaining the null hypothesis, often called a *type II error*. We denote the probability of falsely retaining the null by β. Thus, the probability of correctly rejecting the null hypothesis when H_1 holds is

$$P(\text{reject } H_0 \,|\, H_1) = 1 - \beta$$

Nota Bene 9.2.1. If we let

$$a = P(\text{reject } H_0 \,|\, H_0),$$

then, unfortunately, the probability $1 - \beta$ of correctly rejecting the null hypothesis is not related to a in a clean way. Indeed,

$$\beta = P(\text{reject } H_1 \,|\, H_1).$$

Example 9.2.2. Imagine I have two coins that look identical, but one is fair and the other has probability 66% of coming up heads. I don't know which one is which, so I decide to test one of them by flipping it once. Take as the null hypothesis the assumption that $X \sim \text{Binomial}\left(n, \frac{1}{2}\right)$ and the alternative hypothesis that $X \sim \text{Binomial}\,(n, 0.66)$. If we take $n = 1$ and flip the coin only once, then

$$p = P(X = 1 \,|\, H_0) = \frac{1}{2} = P(X = 0 \,|\, H_0)$$

so for any threshold α less than $\frac{1}{2}$ we never reject the null hypothesis based on a single flip, and the probability a of a false rejection is 0. Similarly, we have

$$P(X = 1 \,|\, H_1) = 0.66 = 1 - P(X = 0 \,|\, H_1),$$

[38] Many statisticians use the names *type I error* and *type II error*, but those names carry little information about what is going on, and thus often confuse students (and everyone else). We prefer to say *falsely reject the null* and *falsely retain the null*.

since we never reject the null, the probabilty β of falsely retaining the null, given H_1, is 1.

Example 9.2.3. The single-coin-flip test was not very useful. If we flip the coin twice more, we have a sample of length 3. Let X be the number of times that heads comes up. Since $X \sim \text{Binomial}(3, \theta)$ for some choice of θ, we can perform a binomial hypothesis test with $n = 3$. Take $X \sim \text{Binomial}\left(3, \frac{1}{2}\right)$ as the null hypothesis, and take the alternative hypothesis to be that $X \sim \text{Binomial}\,(3, 0.66)$. We decide (in advance!) to take $\alpha = 0.2$ so that we reject the null if $p = P(X \geq x \,|\, H_0) < 0.2$. For each value of $x \in \{0, 1, 2, 3\}$ we have

$$p = \frac{1}{8} \sum_{z \geq x} \binom{3}{z} = \begin{cases} 0.125 & \text{if } x = 3 \\ 0.5 & \text{if } x = 2 \\ 0.875 & \text{if } x = 1 \\ 1.0 & \text{if } x = 0 \end{cases}$$

so we only reject the null hypothesis if $x = 3$, and the probability a of false rejection is 0.125. Similarly, we have

$$P(X \geq x \,|\, H_1) = \sum_{z \geq x} \binom{3}{z} (0.66)^z (0.34)^{3-z} = \begin{cases} 0.287 & \text{if } x = 3 \\ 0.732 & \text{if } x = 2 \\ 0.961 & \text{if } x = 1 \\ 1.0 & \text{if } x = 0 \end{cases}$$

since we only reject the null if $x = 3$, the probabilty β of falsely retaining the null, given H_1, is $\beta = P(X < 3 \,|\, H_1) = 0.732$. This better than the single-flip example, but not great.

Example 9.2.4. In the previous example, we can do much better if we flip the coin a lot more, say 100 times, and perform a binomial hypothesis test with $n = 100$. Take $X \sim \text{Binomial}\left(100, \frac{1}{2}\right)$ as the null hypothesis, and take the alternative hypothesis to be that $X \sim \text{Binomial}\,(100, 0.66)$. We decide to set $\alpha = 0.01$ so that we reject the null if $p = P(X \geq x \,|\, H_0) < 0.01$. For each value of $x \in \{0, 1, \ldots, 100\}$ we have

$$p_{\mathbf{x}} = 2^{-100} \sum_{z \geq x} \binom{100}{z},$$

and a straightforward numerical computation using the the inverse of the binomial c.d.f. shows that this is less than 0.01 whenever $x \geq 63$. So we only reject the null hypothesis if $x \geq 63$, and the probability a, of a false rejection given H_0 is $p_{63} = 0.006$. Similarly, we have

$$P(X < x \,|\, H_1) = \sum_{z<x} \binom{100}{z} (0.66)^z (0.34)^{3-z},$$

and a numerical computation shows that the probability $P(X < 63 \,|\, H_1)$ of falsely retaining the null given H_1 is $\beta = 0.229$. Thus, the probability of each type of error is much less with this test than with the two previous tests ($n = 0$ and $n = 3$).

9.2.2 Power

The previous examples show that some statistical tests are more useful, and less likely to make errors, than others. If there are just two possible hypotheses, we define the *power* of a statistical test to be

$$\pi = P(\text{reject } H_0 \,|\, H_1) = 1 - \beta.$$

In case of just two hypotheses, the power is also often called the *true positive rate*.

Example 9.2.5. Consider the case of the leafhoppers and soap on my grapes for the single-sample t-test with $n = 5$ and $\alpha = 0.05$, as in Section 9.1.2. Suppose the only two hypotheses being compared are $H_0 : \mu = 20$ and $H_1 : \mu = 15$. We reject the null if

$$p = P(T < t \,|\, H_0) = P(T < t \,|\, T \sim \mathscr{T}(4)) = F_T(t)$$

is less than $\alpha = 0.05$. This implies that the null is rejected if $t < F_T^{-1}(0.05) = -2.1318$. Considering t now as a random variable itself, the probability of a false rejection is $a = P(\text{reject } H_0 \,|\, H_0) = P(t < -2.1318 \,|\, H_0) = 0.05 = \alpha$.

The probabilility of falsely retaining the null, given the alternative is

$$\begin{aligned}
\beta &= P(\text{retain } H_0 \,|\, H_1) \\
&= P(t \geq -2.1318 \,|\, H_1) \\
&= P\left(\frac{\widehat{\mu}_5 - 20}{s_5/\sqrt{5}} \geq -2.1318 \,\middle|\, \mu = 15 \right) \\
&= P\left(\frac{\widehat{\mu}_5 - 15 - 5}{s_5/\sqrt{5}} \geq -2.1318 \,\middle|\, \mu = 15 \right) \\
&= P\left(\frac{\widehat{\mu}_5 - 15}{s_5/\sqrt{5}} - \frac{5}{s_5/\sqrt{5}} \geq -2.1318 \,\middle|\, \mu = 15 \right) \\
&= P\left(Z \geq \frac{5}{s_5/\sqrt{5}} - 2.1318 \,\middle|\, Z \sim \mathscr{T}(4) \right), \qquad (9.3)
\end{aligned}$$

where the last step follows from the fact that if $X \sim \mathcal{N}(15, \sigma^2)$ then $Z = \frac{\widehat{\mu}_5 - 15}{s_5/\sqrt{5}} \sim \mathcal{T}(4)$. Of course in (9.3) the quantity s_5 is itself a random variable, but given a draw $\mathbf{x}$, we can estimate it as $s_{\mathbf{x}}$. In the case of the grapes example, we found $s_{\mathbf{x}} = 6.025$, so we can estimate β as

$$\begin{aligned}\beta &\approx P\left(Z \geq \frac{\sqrt{5}}{s_{\mathbf{x}}} - 2.1318 \,\middle|\, T \sim \mathcal{T}(4)\right) \\ &= 1 - F_T(-0.2762) \\ &= 0.6020,\end{aligned}$$

where F_T is the c.d.f. of $\mathcal{T}(4)$. Thus, the probablity of falsely retaining the null, given the alternative, is pretty high, and the power $\pi = 1 - \beta \approx 0.3980$ of this test is poor.

Unfortunately, when the alternative hypothesis is not one specific value for θ, then the power computation becomes more tricky. For example, in the one-sample t-test, the alternative hypothesis can be of the form $\mu = \theta > \mu_0$. In this case, we compute the power $\beta(\theta) = P(\text{reject } H_0 \,|\, \mu = \theta)$ for all values of $\theta > \mu_0$. A straightforward computation shows that $1 - a = \sup_{\theta > \mu_0} \beta(\theta)$.

But values of μ near μ_0 are not usually useful or interesting to identify. What matters is when μ is farther away from μ_0. In the case of the leafhoppers and the grapes, if spraying soap only reduces the average number of leafhoppers from 20 to 19, then it doesn't matter how good the test is nor how small the p-value is—it's not worth bothering to spray the soap. In such a case, we might say that the alternative hypothesis should be limited to a smaller range, say $\mu \in [0, 15]$. In this case compute

$$\beta_{\max} = \sup_{\theta \in [0,15]} \beta(\theta)$$

and take the power of the test to be $\pi = 1 - \beta_{\max}$.

Example 9.2.6. Consider the example of my son and the hot sauce at the beginning of Section 9.1. The test we did there only computed the probability that he could get 21 or more right out of 30, given my null hypothesis of $\theta = 0.5$. I'd like to know the power of the test we did, but for that we need to decide the threshold α, let's say $\alpha = 0.05$. Also, it involves an infinite number of values of θ, so we let H_θ be that he can guess the sauce correctly with probability θ. To compute the power, first compute which values of x would cause us to retain H_0; that is, find all x such that

$$P(X \geq x \,|\, H_0) = 2^{-30} \sum_{k=x}^{30} \binom{30}{k} \geq 0.05.$$

A numerical search shows that this holds for all $x \leq 20$. Now we can compute $\beta(\theta) = P(\text{retain } H_0 \,|\, H_\theta)$ for any choice of $\theta > \frac{1}{2}$ as

$$\begin{aligned}
\beta(\theta) &= P(\text{retain } H_0 \,|\, H_\theta) \\
&= P(X \leq 20 \,|\, H_\theta) \\
&= P(X \leq 20 \,|\, X \sim \text{Binomial}(30, \theta)) \\
&= \sum_{k=0}^{20} \binom{30}{k} \theta^k (1-\theta)^{30-k} \\
&\approx \frac{1}{\sqrt{60\pi\theta(1-\theta)}} \int_{-\infty}^{20.5} e^{\frac{(t-30\theta)^2}{60\theta(1-\theta)}}\, dt,
\end{aligned}$$

where the last step follows from the normal approximation of the binomial distribution (see Volume 2, Sect. 6.3.2). The worst case, that is, the largest probability of error, occurs when $\beta(\theta)$ is maximized. One can show that $\beta(\theta)$ is monotonically decreasing in θ, so the probability of error increases as $\theta \downarrow \frac{1}{2}$.

But if the true value of θ is 0.5001, that doesn't really seem good enough for my son to claim that he can tell the difference between the two sauces. Assume that he and I agree that the cutoff should be $\theta = 0.65$. That is, we agree that a mean θ of anything less than 0.65 we take to be the same as random guessing, and a mean of 0.65 or greater we interpret as "he can tell the difference." So we are testing H_0 versus H_θ for $\theta \geq 0.65$. The fact that $\beta(\theta)$ is monotonically decreasing in θ means that the highest probability of error occurs at 0.65 and is $\beta(0.65) \approx 0.6424$. Thus, the power of this test is $\pi = 0.3576$.

9.3 Pitfalls of Null Hypothesis Significance Testing

Without modern statistics, we find it unlikely that people would take seriously a claim about the general population of women, based on two survey questions asked to 100 volunteers on the internet and 24 college students. But with the p-value, a result can be declared significant and deemed worth publishing in a leading journal.
—Andrew Gelman and Eric Loken

Despite its heavy use in many scientific disciplines, null hypothesis significance testing (NHST) is fraught with pitfalls, and its overuse and abuse have misled researchers into many false conclusions. There is significant reason to believe many published research findings are false, largely because of misuse of NHST. In this section we discuss a few of the many pitfalls in using NHST. A much more complete discussion can be found in [WL16, MGG+19, GL13, GKV04], and [Ioa05].

9.3.1 p is the Wrong Measure

The first problem with NHST is that the p-value is poor justification for accepting or rejecting a statistical hypothesis—it just doesn't tell us what we want to know.

Using p-values as a way to judge whether to reject the null hypothesis amounts to committing the prosecutor's fallacy (see Volume 2, Sect. 5.3.2). What we actually want is $P(H_0 \,|\, \mathbf{x})$ or $P(H_0 \,|\, T \geq t)$, but what the p-value gives us is $P(T \geq t \,|\, H_0)$. By Bayes rule, we have

$$P(H_0 \,|\, T \geq t) = \frac{P(T \geq t \,|\, H_0)P(H_0)}{P(T \geq t)} = \frac{p \cdot P(H_0)}{P(T \geq t)}, \tag{9.4}$$

so even when p is very small, there are two ways the desired probability $P(H_0 \,|\, T \geq t)$ could still be large.

First, if there is a strong reason to believe the null hypothesis before the current experiment, the prior $P(H_0)$ should be large, which makes the resulting probability (9.4) large. Second, the denominator $P(T \geq t)$ could be very small. If there are only two possible hypotheses, H_0 and the alternative H_1, then the law of total probability gives $P(T \geq t) = P(T \geq t \,|\, H_0)P(H_0) + P(T \geq t \,|\, H_1)P(H_1)$. So, if there is reason to believe that $P(H_1)$ is small or if $P(T \geq t \,|\, H_1)$ is small, then the ratio (9.4) will be close to 1.

Finally, in the common case that there is more than one plausible alternative to the null hypothesis (say $H_1, \ldots, H_k$), then the prior information about all the alternatives and all the conditional probabilities $P(T \geq t \,|\, H_i)$ must be taken into account for each alternative.

Therefore, a small value of p by itself provides little justification for rejecting or retaining H_0. It can only be meaningful when considered in conjunction with the prior for both H_0 and all other possible hypotheses $H_1, H_2, \ldots, H$, and also the conditional probabilities $P(T \geq t \,|\, H_i)$. Unfortunately, many practitioners of NHST fail to consider the priors or these other probabilities.

9.3.2 The Null Hypothesis is Usually Obviously False

In the two-sample t-test, the null hypothesis is that the difference $X - Y$ between two random variables X and Y is normally distributed with mean 0, and the alternative hypothesis is that the mean is anything other than 0. But it is extremely unlikely that the difference between real-life random variables ever has mean precisely 0. As the famous statistician John Tukey said, "All we know about the world teaches us that the effects of A and B are always different—in some decimal place—for any A and B. Thus asking 'are the effects different?' is foolish."

In the one-sample t-test, the null hypothesis is that the random variable of interest is normally distributed with mean equal to some given μ, but again, in real life it is unlikely that the mean will be exactly μ, or that the random variable will be distributed precisely according to $\mathcal{N}(\mu, \sigma^2)$. In the case of leafhoppers on grape leaves (Section 9.1.2) the number of leaf hoppers is always a nonnegative integer, but with the normal distribution the probability of a negative number is not zero, and the probability of an integer is zero, so the normal distribution can't possibly be exactly the correct distribution. A better approximation might be a Poisson distribution, but even then it is not a perfect fit.

Moreover, the null hypothesis requires that the there be no systematic error. That means that

(i) The statistical model must be perfectly correct, including noise models.

(ii) There can be no measurement error.

(iii) There can be no problems with reliability or validity.

(iv) There can be no bias in the samples.

(v) There can be no failure to randomize treatment assignment.

(vi) There can be no missing data.

(vii) There can be no nonresponses on surveys.

(viii) There can be no noncompliant participants.

(ix) There can be no confounding variables.

(x) There can be no double-blinding errors.

If any of these is imperfect, the null hypothesis is already known to be false, but this in no way supports the adoption of the alternative hypothesis H_1. Instead there are many alternative hypotheses to choose from, including models for the measurement error, models for reliability and validity, models of sampling bias, and so forth. Therefore, if p is small, it does not indicate that the hypothesis H_1 is true, but only that there is a problem with using the null hypothesis H_0 to explain the data.

Nota Bene 9.3.1. A small value of p only indicates that there may be a problem with one of the many the assumptions (the model, the experiment design, etc), but it does not identify which one.

But large values of p only indicate that this one test did not identify a problem. Maybe this happened because there is truly no difference, but maybe it happened because the bias, the errors, and the noise mostly canceled each other out, or maybe because the specific test is insensitive to the problems. In any case, a given p-value alone tells very little about the truth of the null hypothesis or of the alternative hypothesis.

Beware also that similar problems hold for other statistical measures—this is not exclusively limited to tests using p-values.

9.3.3 Multiple Comparison

Performing NHST repeatedly needs careful additional analysis. When I did the soap experiment and t-test on the grape leaves, there was an $\alpha = 5\%$ chance of falsely rejecting the null hypothesis. If there had been two different species of pests on my grape vines and I counted the number of each of those pests on my five leaves after applying soap and performed a t-test on each one, then the probability of falsely rejecting the null hypothesis in each test would remain at 5%, but the probability that at least one of those two tests gives a false rejection of the null would be higher. This is called the *multiple comparison problem*.

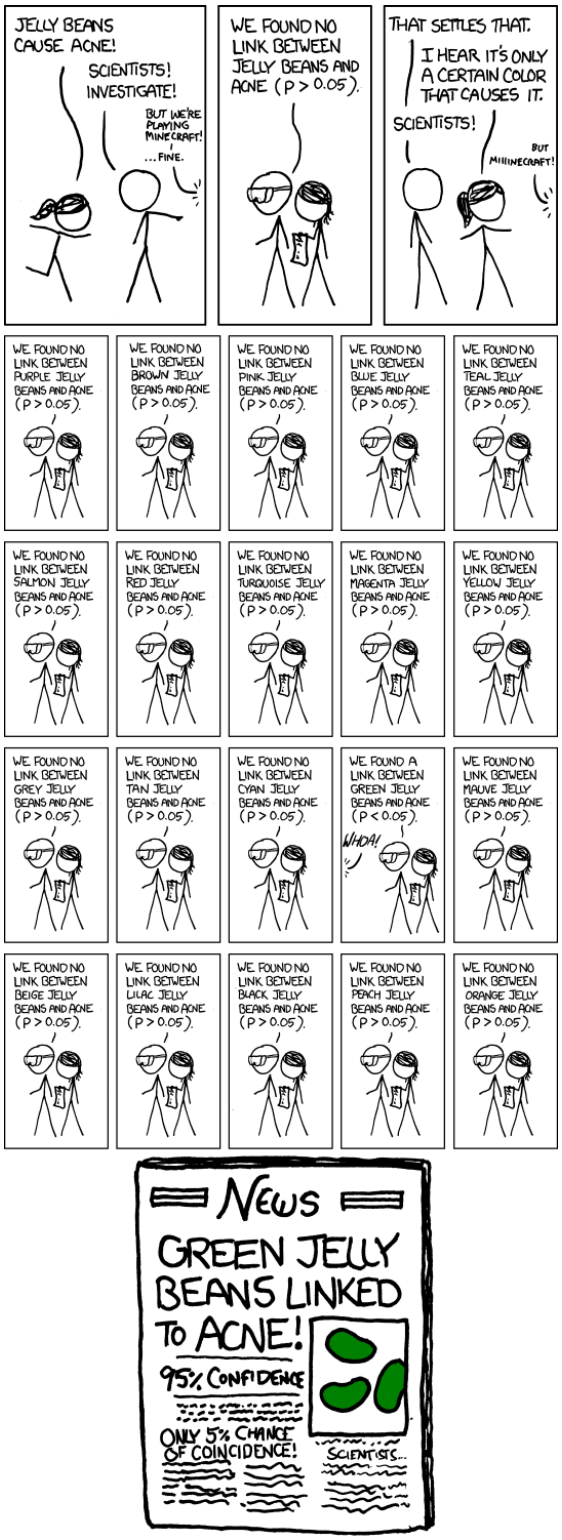

Figure 9.1: The multiple comparison problem. `https://xkcd.com/882/`

Specifically, let R_i denote the event of false rejection in the ith test, so that $R = R_1 \cup R_2$ is the event of at least one false rejection. Assuming that the events R_1 and R_2 are independent, we have

$$P(R) = P(R_1 \cup R_2) = P(R_1) + P(R_2) - P(R_1 \cap R_2) = 2\alpha - \alpha^2 > \alpha.$$

More generally, if m independent null hypothesis significance tests are performed, each with a significance level of α, then one can show

$$P(R) = 1 - (1 - \alpha)^m. \tag{9.5}$$

As m grows large, the probability of at least one false null rejection approaches 1. For an illustration of this idea, see Figure 9.1.

One simple way to address the multiple comparison problem is to replace α by $\frac{\alpha}{m}$. This is called the *Bonferroni method*. Specifically, we reject the null hypothesis for the ith test only if the p-value for that test is less than α/m.

Theorem 9.3.2. *Using the Bonferroni method, the probability of falsely rejecting the null hypothesis at least once in m tests is no more than α.*

Proof. As above, denote by R_i the event that the ith test falsely rejects the null hypothesis and by $R = \bigcup_{i=1}^m R_i$ the event that at least one test falsely rejects the null hypothesis. Note that (9.5) cannot be used because the events R_i are not necessarily independent. But we have

$$P(R) = P\left(\bigcup_{i=1}^m R_i\right) \leq \sum_{i=1}^m P(R_i) = \sum_{i=1}^m \frac{\alpha}{m} = \alpha. \quad \square$$

There are several other, more sophisticated, ways to address the multiple comparison problem, but we do not treat them here. The important thing to remember is that multiple comparisons must be thought about as a whole, and significance tests must be adjusted to compensate for multiple comparisons.

9.3.4 Forking Paths

In some cases an evil scientist intentionally performs multiple tests on a data set, finds one situation where the null hypothesis is rejected without correcting for the multiple comparisons, and then reports only that one case, as if it were an isolated test. This is called *p-hacking*, and it is, of course, unethical. But a similar effect can easily occur even when good scientists are trying hard to be careful and honest about their work.

A significant risk for a professional scientist performing many experiments and using NHST regularly is that their entire scientific career essentially amounts to one giant case of multiple testing. If they perform 1,000 different experiments in a lifetime, using a significance level of $\alpha = 0.05$, they will have roughly 50 false null rejections. Those experiments for which the null is retained are usually considered uninteresting and are not published, but null rejections are usually considered supportive of an alternative hypothesis and published. This means that roughly 50 of that scientist's published papers are expected to be false. This is not to imply any ill intent on the part of the scientist—only that by pure chance, even when they use standard, accepted scientific practices, a large number of their published results will be false.

Even more subtle is the situation that Gelman and Loken call *the garden of forking paths* [GL13] and is also sometimes called *researcher degrees of freedom*. This occurs when a scientist performs only a single analysis, apparently avoiding the multiple comparison problem, but they use the data to decide which analysis to perform. The problem is that there are potentially many analyses that could be performed, and when the choice of which to perform is determined by the data, then it is essentially similar to performing all the analyses and discarding all but the chosen one. At multiple points in the analysis, a decision must be made about what to compare, or how to perform the analysis—a forking path. When each path forks multiple times, the total number of potential analyses can grow very large. Using the data to make each decision essentially amounts to performing all the analyses—a multiple comparison—even though the researcher is not examining the corresponding p-values or doing any conscious fishing for significant results.

9.3.5 Thresholding is Problematic

The previous discussion shows that the threshold of $\alpha = 0.05$ is probably too high. With that threshold a large number of the published papers of even good, responsible researchers would be expected to be false. Consequently, some people have proposed tightening the standards and using a threshold of $\alpha = 0.005$. And in particle physics, the standard is even tighter, with the threshold for "evidence" of a particle being $\alpha = 0.003$, and the threshold for claiming "discovery" of a particle is $\alpha = 0.0000003$.

But having any threshold at which to retain or reject the null is problematic, regardless of how small the threshold may be. Here are some of the problems.

(i) The p-value is continuous, but a threshold embraces values just barely below the threshold while rejecting those just barely above. This leads to a false dichotimization (or in some cases trichotomization) of evidence. One result is labeled "significant" while another result, almost identical, is labeled "not signficant."

(ii) The threshold choice is arbitrary and has no ontological basis. Ronald Fisher, the earliest proponent of NHST, advocated for using 5% because it was about two standard deviations from the mean in a normal distribution and therefore was "convenient." But convenience is a poor reason for deciding to accept or reject a hypothesis.

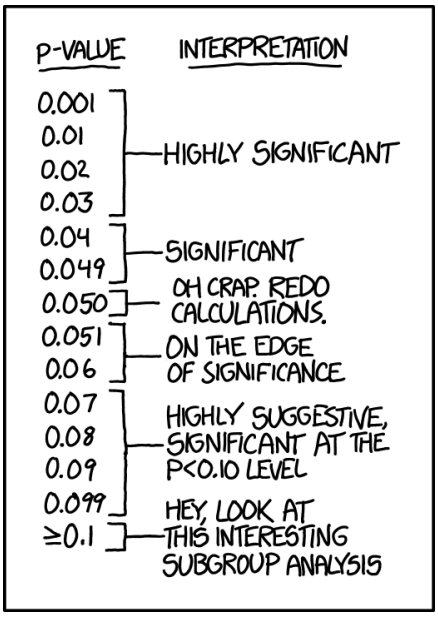

Figure 9.2: Thresholding p-values. Source: `https://xkcd.com/1478/`. Rollover text: *If all else fails use "significant at the* $p > 0.05$ *level" and hope no one notices.*

(iii) Thresholding p-values rewards running many noisy or badly designed studies with low power, instead of a few low-noise, well-designed studies with high power, because noisy, low-power studies can be done more cheaply, and the noise often allows p to go below the threshold.

The well-intentioned recommendation of some scientists to use smaller significance levels, like $\alpha = 0.005$, could actually reduce the reliability of published studies, because well-designed studies might not quite reach the threshold, while poorly designed studies with lots of variance in the results would still occasionally meet the standard. Since the poorly designed studies are usually cheaper and easier to conduct, we expect many more of those to be conducted than well-designed studies. Thus, the small percentage of the many poorly designed studies that happen to meet the threshold would still account for a large percentage of total published work, maybe even larger than they do now.

9.3.6 Fallacious Reasoning

Here is a summary of some common types of fallacious reasoning of p-value users, including even many statisticians:

(i) Incorrectly take a low value of p as support for an alternative hypothesis.

(ii) Mistakenly dichotomize reality. This includes drawing conclusions from sharp threshold as if p were discontinuous, and also handling continuous treatments and effects as if they were discrete and binary.

(iii) Confuse statistical significance and practical importance (actual significance). This often happens because they ignore the magnitude of the effects. Effects that are tiny and of no importance are mistaken as being meaningful just because they are statistically significant.

(iv) Incorrectly take a low value of p as evidence of causality.

(v) Ignore the true cost functions of an incorrect decision. If wrongly rejecting the null could lead to severe health risks, for example, we should be very cautious to avoid rejecting the null falsely. If wrongly rejecting it has little or no cost, we can safely reject the null with much less evidence.

(vi) Ignore prior knowledge of the truth or falsity of the null hypothesis.

9.3.7 Partial Solutions

Some basic principles of good science can partially help avoid the abuse of NHST. Here are a few.

(i) Always decide on your hypothesis before examining the data. This helps avoid p-hacking, multiple testing, and the garden of forking paths.

(ii) Account for the many other possible explantations of small or large p. The null hypothesis should not be considered alone, but rather in competition with all the other possible hypotheses.

(iii) Consider prior knowledge, both mathematical (Bayes) and common sense, that could lead to accept/reject independent of p.

(iv) Remember that p is continuous—not discrete.

(v) Always report all data, not just the results, and not just the data for which you have statistically significant results.

Gossett, the discoverer of the t-distribution, used in the t-test, was unhappy with people (especially with Fisher, the first big promoter of NHST) who used p-values as a null-hypothesis test procedure. Gossett maintained that such decisions must take into account the "advantage to be gained by following the result of the experiment, compared with the increased cost of the new method, if any, and the cost of each experiment" [Zil08].

Exercises

Note to the student: Each section of this chapter has several corresponding exercises, all collected here at the end of the chapter. The exercises between the first and second line are for Section 1, the exercises between the second and third lines are for Section 2, and so forth.

You should **work every exercise** (your instructor may choose to let you skip some of the advanced exercises marked with *). We have carefully selected them, and each is important for your ability to understand subsequent material. Many of the examples and results proved in the exercises are used again later in the text. Exercises marked with ⚠ are especially important and are likely to be used later in this book and beyond. Those marked with † are harder than average, but should still be done.

Although they are gathered together at the end of the chapter, we strongly recommend you do the exercises for each section as soon as you have completed the section, rather than saving them until you have finished the entire chapter.

9.1. A sample of 5 leaves for my grape vines is a little skimpy, so I decide to check 3 more leaves. On these I find 18, 14, and 15 leafhoppers, respectively, so now I have 8 data points. Compute the new test statistic T_8 (see (9.1)) and the corresponding p-value. Does this give me enough reason to reject the null hypothesis at the $\alpha = 0.05$ level? Hint: Many numerical systems have built in methods for computing the c.d.f. of the t-distribution.

9.2. Instead of spraying all my grapes with soap, I decide to spray only half of them—after all maybe the number of leafhoppers would change over two weeks regardless of spraying. Two weeks after spraying, I take a random sample of 5 treated leaves and find 21, 18, 9, 22, and 11 leafhoppers. I also take a random sample of 5 untreated leaves and find 19, 28, 21, 18, and 25 leafhoppers. Apply the two-sample t-test to this situation.

(i) What is the null hypothesis?

(ii) What is the value of the T statistic (9.2)?

(iii) What should be meant by "more extreme" in the definition of the p-value for this test?

(iv) What is the p-value for this test?

(v) Does NHST suggest retaining or rejecting the null hypothesis at the $\alpha = 0.05$ level?

9.3. A six-sided die is rolled 50 times and the numbers 1 through 6 are observed 8, 10, 15, 6, 4, and 7 times, respectively. We would like to know what this says about whether the die is fair. Use NHST and the Pearson chi-squared test to investigate this.

(i) What is the null hypothesis in this case?

(ii) What is a reasonable value of α to choose for rejection of the null hypothesis?

(iii) What is $T_{\mathbf{x}}$?

(iv) What is the p-value for this test?

(v) What is your conclusion?

9.4. The manufacturer of a certain product wants has a call-in help center for customers who have trouble using the product, but staffing the call center is expensive, so they set up some extra webapges with additional instructions and FAQs in the hope of reducing the number of calls that come in. Assume that the number of calls N coming in before the new webpages was Poisson distributed $N \sim \text{Poisson}(70)$ with an average rate of 70 per day. The first day after the change, the number of calls was 50. We assume that the number of calls is still Poisson distributed, but we hope that the rate λ is less than 70. Use NHST to determine whether to retain or reject the null hypothesis that $\lambda = 70$.

(i) What is a reasonable threshold α for which to reject the null hypothesis?

(ii) For computing the p-value, we need to consider those values of N that are more extreme than the observed value. What range of values of N is more extreme than the observed value of $N = 50$?

(iii) Compute the value of p. What is your conclusion?

9.5. As in the two-coin examples (Examples 9.2.2–9.2.4), assume we have two identical-looking coins, one fair, and one with with probability $\theta > \frac{1}{2}$ of heads. Assume that one coin is chosen, it is flipped n times, the sum X is recorded, and NHST with the binomial test is used to determine whether it is the fair coin or not. Let H_0 be that the coin is fair and reject the null if $p < \alpha$. Let $\alpha = 0.05$, let $n = 50$, and let $\theta = 0.75$.

(i) Find the smallest value of x such that H_0 will be rejected if x is observed.

(ii) Find the probability $a = P(\text{reject } H_0 \,|\, H_0)$.

(iii) Find the probability $\beta = P(\text{retain } H_0 \,|\, H_\theta)$.

(iv) Find the power of this test.

9.6. Assume the same setup as in the previous problem, with $\alpha = 0.05$ and $\theta = 0.75$, but n is not fixed yet. We want to choose n in order to make the test powerful. Find the smallest value of n that will make the previous test have power at least 0.8.

9.7. Given the same setting as Exercise 9.4 with $\alpha = 0.05$

(i) Find the largest value $n_{\max}$ of N that would still result in the retention of the null hypothesis.

(ii) Let $\lambda = 55$ and find the probability $\beta(\lambda) = P(\text{retain } H_0 \,|\, H_\lambda)$ of falsely retaining the null.

(iii) What is the power of this test for $\lambda = 55$?

9.8. Let $X_1, \ldots, X_n$ be a sample of length n from $\mathcal{N}(\mu, \sigma^2)$, where σ^2 is known, but μ is unknown. Let the null hypothesis be that $\mu \leq 0$ and the alternative hypothesis be that $\mu > 0$. Consider a test that rejects H_0 if $\widehat{\mu}_n > c$ for some fixed c.

(i) For arbitrary $\theta > 0$, express $\beta(\theta)$ in terms of c, σ, θ and the c.d.f. Φ of the standard normal.

(ii) Show that $\beta(\theta)$ is a decreasing function of θ, so the greatest probability of error and the lowest power occurs as $\theta \to \theta_0^+$.

(iii) Show that $\beta(\theta)$ is continuous everywhere, so that $\sup_{\theta>0} \beta(\theta) = \beta(0)$.

(iv) For any given value of $\alpha \in (0,1)$, find the value of c that will make $\sup_{\theta>0} \beta(\theta) = 1 - \alpha$.

9.9. For a sample $X_1, \ldots, X_n$ of length $n > 3$ from $\mathscr{N}(\mu, \sigma^2)$ with unknown μ and σ^2, assume that we use a single-sample t-test, with the statistic $T = \frac{\widehat{\mu}_n - \mu}{s_n/\sqrt{n}}$. Take as the null hypothesis $H_0 : \mu = \theta_0$ and as the alternative hypothesis $H_\theta : \mu = \theta$ for an arbitrary $\theta < \theta_0$. Thus, H_0 will be rejected if t is small enough that $p = P(T < t \,\big|\, H_0) < \alpha$.

(i) For an arbitrary α, compute the maximal value $t_{\max}$ of t for which the null hypothesis will be rejected. Express you answer in terms of α and the inverse c.d.f. of the appropriate t-distribution.

(ii) Generalize the computation in Example 9.2.5 to give an estimate for $\beta(\theta) = P(\text{retain } H_0 \,\big|\, H_\theta)$ in terms $t_{\max}$, the c.d.f. F_T of $T \sim \mathscr{T}(n-1)$, and $s_{\mathbf{x}}$.

(iii) Show that $\beta(\theta)$ is an increasing function of θ on the interval $\theta \in (-\infty, \theta_0)$, so the greatest possibility of error and the lowest possible power occurs as $\theta \to \theta_0^-$.

(iv) Show that $\lim_{\theta \to \theta_0^-} \beta(\theta) = 1 - \alpha$ and $\inf_{\theta \in (-\infty, \theta_0)} \pi(\theta) = \alpha$.

9.10. Let R_i denote the event of a false rejection of the null hypothesis in test i at significance level α, and let $R = \bigcup_{i=1}^m R_i$ be the probabilty of at least one false rejection. Assume that the R_i are independent.

(i) Prove (9.5). Hint: This can be done in at least two ways, the first using set complements, intersections, and independence, and the second by using inclusion-exclusion (see Volume 2, Theorem 1.6.16) to show

$$P(R) = \sum_{i=1}^{m} (-1)^{i+1} \binom{m}{i} \alpha^i.$$

(ii) Assume that $\alpha = 0.05$. Compute the minimum number m of repeated tests necessary to make the probability $P(R)$ of false rejection greater than 90%.

9.11. Suppose you have a treatment that you suspect may alter performance on a certain task. You compare the means of your control and experimental groups (say, 20 subjects in each sample). Furthermore, suppose you use a two-sample t-test and your result is significant (t = 2.7, degrees of freedom = 18, p = 0.01). Which of the statements below logically follow from the premises above?

(i) You have disproved the null hypothesis (the hypothesis that there is no difference between the control and the experiment groups).

(ii) You have found the probability that the null hypothesis is true.

(iii) You have proved the experimental (alternative) hypothesis (that there is a difference between the population means).

(iv) You can deduce the probability that the experimental hypothesis is true.

(v) You know the probability that you are making the wrong decision if you decide to reject the null hypothesis.

(vi) You have a reliable experimental finding in the sense that if, hypothetically, the experiment were repeated a great number of times, you would obtain a significant result on 99% of occasions.

9.12. Consider a dataset consisting of answers from a survey of women shortly before the 2012 presidential election asking about their political attitudes and voting intentions, religiosity, age, ethnicity, parenthood status, relationship status (single not dating, single but dating, living with a partner, engaged, or married), and which day of their menstrual cycle the respondant was in at the time of answering the questionnaire. Each woman was surveyed only once.

The authors' stated goal in the study was to test how fertility influenced women's politics, religiosity, and voting. The authors claimed that their study showed ($p < 0.05$) that ovulation had "drastically different effects on single versus married women. Ovulation led single women to become more liberal, less religious, and more likely to vote for Barack Obama. In contrast, ovulation led married women to become more conservative, more religious, and more likely to vote for Mitt Romney." Here, by "married" they meant any of "living with a partner," "engaged," or "married."

In addition to the hypothesis of an interaction between ovulation, marital status, and political preference, list at least ten other hypotheses that could have been tested with this dataset and that would have been consistent with the stated goals. These are some of the many possible forking paths.

9.13. Consider a study from 2015 that claimed to show that chocolate helps with weight loss. The study took 18 different measurements—weight, cholesterol, sodium, blood protein levels, sleep quality, well-being, and more—from 15 people. Those 15 people were divided into three groups—a control group, a diet-without-chocolate group, and a diet-with-chocolate group. The diet-without-chocolate group followed a standard low-carb diet for 21 days. The diet-with-chocolate group followed the same low-carb diet but also took 1.5 oz (42 grams) of chocolate. The control group was instructed to make no changes to their regular diet. The 18 measurements were taken for all fifteen people both before and after the 21 day period. The diet-with-chocolate group had a (mean) loss of 3.2% of their body weight (denote this by $\widehat{\mu}_C = 0.032$, and the diet-without chocolate group had a (mean) loss of 3.1% of their body weight ($\widehat{\mu}_0 = 0.031$), while the control group stayed relatively constant.

(i) Name three potential sources of systematic error that might have affected the results of this study.

(ii) Assuming that there was no systematic error, and that changes due to chocolate in each of the 18 different measurements are independent random variables, taking $\alpha = 0.05$ on each test, what is the probability of finding at least one statistically significant difference among the 18? (assuming the null hypothesis in each case)

(iii) Even if they had only run one test—the test of the hypothesis that choclolate helps you lose more weight, what the NHST tested was $P(\text{data (or more extreme)} \,|\, H_0)$, but what we want to know is $P(H_0|\text{data})$.

To analyze this more carefully, let W_C be the random variable of percentage weight loss in choclate eaters and W_0 be the random variable of percentage weight loss in choclate abstainers. Assume that $W_0 \sim \mathcal{N}(\mu_0, \sigma_0^2)$ and $W_C \sim \mathcal{N}(\mu_C, \sigma_C^2)$. Let $D = \widehat{\mu}_C - \widehat{\mu}_0$ be the difference in sample means.

For simplicity assume there are only two possible values for μ_C, with H_0 being the hypthesis that $\mu_C = \mu_0$, and H_1 being the hypothesis that $\mu_C = \mu_0 + 0.003$ (a difference of 10%, as claimed in the abstract of the paper).

Since the study did not report $\widehat{\sigma}_0^2$ nor $\widehat{\sigma}_C^2$, assume that the true variances are the same $\sigma_0^2 = \sigma_C^2 = \sigma^2$. As a rough estimate, assume that body weight for the average study participant fluctuates about 3% in the course of a month (due, among other things, to menses and variation in eating and drinking habits), so let's take $\sigma^2 = 0.03^2$.

Given hypothesis H_1, assuming that there are $15/3 = 5$ people in each group, and given all the previous assumptions, what is the distribution of the random variable $D = \widehat{\mu}_C - \widehat{\mu}_0$? Hint: The number of people in each group affects the answer.

(iv) Given hypothesis H_0 and all the previous assumptions, what is the distribution of the random variable $D = \widehat{\mu}_C - \widehat{\mu}_0$?

(v) Given what science and medicine and most people's own life experience have already shown about candy and body weight, it seems likely that $P(H_1) \ll P(H_0)$, but to be very generous to the study's claims, let's assume that the prior probabilities are equal: $P(H_1) = P(H_0) = 0.5$. Use Bayes' rule (the continuous case, with p.d.f.s) and all the different estimates we have made so far to give a rough (lower) estimate of

$$P(H_0 \,|\, D = 0.001) = \frac{f_D(0.001 \,|\, H_0)P(H_0)}{f_D(0.001 \,|\, H_0)P(H_0) + f_D(0.001 \,|\, H_1)P(H_1)}.$$

Notes

Exercise 9.11 is from a quiz given to, and failed by, many experienced practitioners and teachers of NHST, see [GKV04]. Exercise 9.8 is based on Example 10.2 of [Was04]. Exercise 9.12 is based on an actual study described in [GL13]. Exercise 9.13 is based on the study [BKHD15], which was an actual published study, but soon after publication the authors revealed that the entire study, while not falsifying any data, was intentionally designed to demonstrate the consequences of bad practices in NHST and the irresponsible ways that many journalists report on bad research (see [Boh15]). The hoax completely worked, in that journalists worldwide wrote, and millions of people believed, that this study showed chocolate accelerates weight loss.