BE/Bi 103: Data Analysis in the Biological Sciences

Justin Bois

Caltech

Fall, 2017

6 Frequentist methods

We have taken a Bayesian approach to data analysis in this class. So far, the main motivation for doing so is that I think the approach is more intuitive. We often think of probability as a measure of plausibility, so a Bayesian approach jibes with our natural mode of thinking. Further, the mathematical and statistical models are explicit, as is all knowledge we have prior to data acquisition. The Bayesian approach, in my opinion, therefore reflects intuition and is therefore more digestible and easier to interpret.

Nonetheless, frequentist methods are in wide use in the biological sciences. They are not more or less valid than Bayesian methods, but, as I said, can be a bit harder to interpret. Importantly, as we will soon see, they can very very useful, and easily implemented, in **nonparametric inference**, which is statistical inference where no model is assumed; conclusions are drawn from the data alone. In fact, most of our use of frequentist statistics will be in the nonparametric context. But first, we will discuss some parametric estimators from frequentist statistics.

6.1 The frequentist interpretation of probability

In the tutorials this week, we will do parameter estimation and hypothesis testing using the frequentist definition of probability. As a reminder, in the frequentist definition of probability, the probability P(A) represents a long-run frequency over a large number of identical repetitions of an experiment. Much like our strategies thus far in the class have been to start by writing Bayes's theorem, for our frequentist studies, we will directly apply this definition of probability again and again, using our computers to "repeat" experiments many time and tally the frequencies of what we see.

The approach we will take is heavily inspired by Allen Downey's wonderful book, *Think Stats* and from Larry Wasserman's All of Statistics. You may also want to watch this great 25-minute talk by Jake VanderPlas, where he discusses the differences between Bayesian and frequentist approaches.

6.2 The plug-in principle

In Bayesian inference, we tried to find the most probable value of a parameter. That is, we tried to find the parameter values at the MAP, or maximum a posteriori probability. We then characterized the posterior distribution to get a credible region for the parameter we were estimating. We will discuss the frequentist analog to the credible region, the **confidence interval** in a moment. For now, let's think about how to get an estimate for a parameter value, given the data.

While what we are about to do is general, for now it is useful to have in your mind a concrete example. Imagine we have a data set that is a set of repeated measurements, such as the repeated measurements of the Dorsal gradient width we studied from the Stathopoulos lab. We have a model in mind: the data are generated from a Gaussian distribution. This means there are two parameters to estimate, the mean μ and the variance σ.

To set up how we will estimate these parameters directly from data, we need to make some definitions first. Let $F(x)$ be the cumulative distribution function (CDF) for the distribution. Remember that the probability density function (PDF), $f(x)$, is related to the CDF by

$$f(x) = \frac{\mathrm{d}F}{\mathrm{d}x}. \tag{6.1}$$

For a Gaussian distribution,

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\,\mathrm{e}^{-(x-\mu)^2/2\sigma^2}, \tag{6.2}$$

which defines our two parameters μ and σ.

A **statistical functional** is a functional of the CDF, $T(F)$. A parameter θ of a probability distribution can be defined from a functional, $\theta = T(F)$. For example, the mean, variance, and median are all statistical functionals.

$$\mu = \int_{-\infty}^{\infty} \mathrm{d}x\, x f(x) = \int_{-\infty}^{\infty} \mathrm{d}F(x)\, x, \tag{6.3}$$

$$\sigma^2 = \int_{-\infty}^{\infty} \mathrm{d}x\, (x-\mu)^2 f(x) = \int_{-\infty}^{\infty} \mathrm{d}F(x)\, (x-\mu)^2, \tag{6.4}$$

$$\text{median} = F^{-1}(1/2). \tag{6.5}$$

Now, say we made a set of n measurements, $\{x_1, x_2, \ldots x_n\}$. You can this of this as a set of Dorsal gradient widths if you want to have an example in your mind. We define the **empirical cumulative distribution function**, $\hat{F}(x)$ from our data as

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \le x), \tag{6.6}$$

with

$$I(x_i \le x) = \begin{cases} 1 & x_i \le x \\ 0 & x_i > x. \end{cases} \tag{6.7}$$

We saw this functional form of the ECDF in our first homework. We can then also define an **empirical distribution function**, $\hat{f}(x)$ as

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \delta(x - x_i), \tag{6.8}$$

where $\delta(x)$ is the Dirac delta function. To get this, we essentially just took the derivative of the ECDF.

So, we have now defined an empirical distribution that is dependent only on the data. We now define a **plug-in estimate** of a parameter θ as

$$\hat{\theta} = T(\hat{F}). \tag{6.9}$$

In other words, to get a plug-in estimate a parameter θ, we need only to compute the functional using the empirical distribution. That is, we simply "plug in" the empirical CDF for the actual CDF.

The plug-in estimate for the median is easy to calculate.

$$\widehat{\text{median}} = \hat{F}^{-1}(1/2), \tag{6.10}$$

or the middle-ranked data point. The plug-in estimate for the mean or variance, seem at face to be a bit more difficult to calculate, but the following general theorem will help. Consider a functional of the form of an expectation value, $r(x)$.

$$\begin{aligned}\int \mathrm{d}\hat{F}(x)\, r(x) = \int \mathrm{d}x\, r(x)\hat{f}(x) = \int \mathrm{d}x\, r(x)\left[\frac{1}{n}\sum_{i=1}^{n}\delta(x - x_i)\right] \\ = \frac{1}{n}\sum_{i=1}^{n}\int \mathrm{d}x\, r(x)\,\delta(x - x_i) = \frac{1}{n}\sum_{i=1}^{n} r(x_i).\end{aligned} \tag{6.11}$$

This means that the plug-in estimate for an expectation value of a distribution is the mean of the observed values themselves. The plug-in estimate of the mean, which has $r(x) = x$, is

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \equiv \bar{x}, \tag{6.12}$$

where we have defined $\bar{x}$ as the traditional sample mean, which we have just shown is the plug-in estimate. This plug-in estimate is implemented in the `np.mean()` function. The plug-in estimate for the variance is

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2. \tag{6.13}$$

This plug-in estimate is implemented in the `np.var()` function.

We can compute plug-in estimates for more complicated parameters as well. For example, for a bivariate distribution, the correlation between the two variables, x and y, is defined with

$$r(x) = \frac{(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}, \tag{6.14}$$

and the plug-in estimate is

$$\hat{\rho} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left(\sum_i (x_i-\bar{x})^2\right)\left(\sum_i (y_i-\bar{y})^2\right)}}. \tag{6.15}$$

6.3 Bias

The **bias** of an estimate is the difference between the expectation value of the estimate and value of the parameter.

$$\text{bias}_F(\hat{\theta}, \theta) = \langle\hat{\theta}\rangle - \theta = \int \mathrm{d}x\, \hat{\theta} f(x) - T(F). \tag{6.16}$$

We often want a small bias because we want to choose estimates that give us back the parameters we expect.

Let's consider a Gaussian distribution. Our plug-in estimate for the mean is

$$\hat{\mu} = \bar{x}. \tag{6.17}$$

In order to compute the the expectation value of $\hat{\mu}$ for a Gaussian distribution, it is useful to know that

$$\langle x\rangle = \int_{-\infty}^{\infty} \mathrm{d}x\, x\, \mathrm{e}^{-(x-\mu)^2/2\sigma^2} = \mu\,. \tag{6.18}$$

Then, we have

$$\langle\hat{\mu}\rangle = \langle\bar{x}\rangle = \frac{1}{n}\left\langle \sum_i x_i \right\rangle = \frac{1}{n}\sum_i \langle x_i\rangle = \langle x\rangle = \mu\,, \tag{6.19}$$

so the bias in the plug-in estimate for the mean is zero. It is said to be **unbiased**.

To compute the bias of the plug-in estimate for the variance, it is useful to know that

$$\langle x^2\rangle = \int_{-\infty}^{\infty} \mathrm{d}x\, x^2\, \mathrm{e}^{-(x-\mu)^2/2\sigma^2} = \sigma^2 + \mu^2, \tag{6.20}$$

so

$$\sigma^2 = \langle x^2 \rangle - \langle x \rangle^2. \tag{6.21}$$

So, the expectation value of the plug-in estimate is

$$\left\langle \hat{\sigma}^2 \right\rangle = \left\langle \frac{1}{n} \sum_i x_i^2 \right\rangle - \langle \bar{x}^2 \rangle = \frac{1}{n} \sum_i \langle x_i^2 \rangle - \langle \bar{x}^2 \rangle = \mu^2 + \sigma^2 - \langle \bar{x}^2 \rangle . \tag{6.22}$$

We now need to compute $\langle \bar{x}^2 \rangle$, which is a little trickier. We will use the fact that the measurements are independent, so $\langle x_i x_j \rangle = \langle x_i \rangle \langle x_j \rangle$ for $i \neq j$.

$$\begin{aligned}
\langle \bar{x}^2 \rangle &= \left\langle \left(\frac{1}{n} \sum_i x_i \right)^2 \right\rangle = \frac{1}{n^2} \left\langle \left(\sum_i x_i \right)^2 \right\rangle = \frac{1}{n^2} \left\langle \sum_i x_i^2 + 2 \sum_i \sum_{j>i} x_i x_j \right\rangle \\
&= \frac{1}{n^2} \left(\sum_i \langle x_i^2 \rangle + 2 \sum_i \sum_{j>i} \langle x_i x_j \rangle \right) = \frac{1}{n^2} \left(n(\sigma^2 + \mu^2) + 2 \sum_i \sum_{j>i} \langle x_i \rangle \langle x_j \rangle \right) \\
&= \frac{1}{n^2} \left(n(\sigma^2 + \mu^2) + n(n-1) \langle x \rangle^2 \right) = \frac{1}{n^2} \left(n\sigma^2 + n^2 \mu^2 \right) = \frac{\sigma^2}{n} + \mu^2.
\end{aligned} \tag{6.23}$$

Thus, we have

$$\left\langle \hat{\sigma}^2 \right\rangle = \left(1 - \frac{1}{n} \right) \sigma^2. \tag{6.24}$$

Therefore, the bias is

$$\text{bias} = -\frac{\sigma^2}{n} \tag{6.25}$$

An unbiased estimator would instead be

$$\frac{n}{n-1} \hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2. \tag{6.26}$$

Note that in the none of the above analysis depended on $F(x)$ being the CDF of a Gaussian distribution. For any distribution, we define the property of the distribution known as the mean as $\langle x \rangle$ and that known as the variance as $\langle x^2 \rangle - \langle x \rangle^2$.

Comparison to Bayesian treatment. To compare this parameter estimate to a Bayesian treatment, we will consider a Gaussian likelihood in a Jeffreys prior on σ. Recalling Lecture 2, we found that in this case we got $\bar{x}$ as our most probable value of μ, meaning this is the value of μ at the MAP. The most probable value of σ^2 was $\hat{\sigma}^2$. But wait a minute! We just found that was a biased estimator. What gives?

The answer is that we are considering the *maximally probable* values and not the expectation value of the posterior. Recall that the posterior for estimating the parameters of a Gaussian distribution is

$$P(\mu, \sigma \mid \{x_i\}, I) \propto \frac{\mathrm{e}^{-n\hat{\sigma}^2/2\sigma^2}}{\sigma^{n+1}} \exp\left[\frac{n(\mu - \bar{x})^2}{2\sigma^2}\right]. \tag{6.27}$$

After some gnarly integration to compute the normalization constant and the expectation values of μ and σ^2 from this posterior, we get

$$\langle \mu \rangle = \bar{x} \tag{6.28}$$

$$\langle \sigma^2 \rangle = \frac{n}{n-1}\,\hat{s}^2, \tag{6.29}$$

the same as the unbiased frequentist estimators. Note that $\langle \sigma^2 \rangle \neq \langle \sigma \rangle^2$. Remember, in frequentist statistics, we are not computing a posterior distribution describing the parameters. There is no such thing as the "probability of a parameter value" in frequentist probability. A parameter has a value, and that's that. We report a frequentist estimate for the parameter value based on the expectation values of the assumed underlying distribution. We just showed that, at least for a Gaussian, the expectation value of the posterior gives the unbiased frequentist estimate and the MAP gives the plug-in estimate.

Justification of using plug-in estimates. Despite the apparent bias in the plug-in estimate for the variance, we will normally just use plug-in estimates going forward. (We will use the hat, e.g. $\hat{\theta}$, to denote an estimate, which can be either a plug-in estimate or not.) Note that the bootstrap procedures we lay out in what follows do not *need* to use plug-in estimates, but we will use them for convenience. Why do this? First, the bias is typically small. We just saw that the biased and unbiased estimators of the variance differ by a factor of $n/(n-1)$, which is negligible for large n. In fact, plug-in estimates tend to have much smaller error than the confidence intervals for the parameter estimate, which we will discuss in a moment. Finally, we saw when connecting to the Bayesian estimates that the expectation value is not necessarily always what we want to describe; sometimes (though certainly not always, perhaps even seldom) the MAP is preferred. In this sense, attempting to minimize bias is somewhat arbitrary.

6.4 Bootstrap confidence intervals

The frequentist analog to a Bayesian credible region is a **confidence interval**. Remember, with the frequentist interpretation of probability, we cannot assign a probability to a parameter value. A parameter has one value, and that's that. We can only describe the long-term frequency of observing results about random variables. So, we can define a 95% confidence interval as follows.

> If an experiment is repeated over and over again, the estimate I compute for a parameter, $\hat{\theta}$, will lie between the bounds of the 95% confidence interval for 95% of the experiments.

While this is a correct definition of a confidence interval, some statisticians prefer another. To quote Larry Wasserman,

> [The above definition] is correct but useless since we rarely repeat the same experiment over and over. A better interpretation is this: On day 1, you collect data and construct a 95 percent confidence interval for a parameter θ_1. On day 2, you collect new data and construct a 95 percent confidence interval for an unrelated parameter θ_2. On day 3, you collect new data and construct a 95 percent confidence interval for an unrelated parameter θ_3. You continue this way constructing confidence intervals for a sequence of unrelated parameters $\theta_1, \theta_2, \ldots$. Then 95 percent of your intervals will trap the true parameter value. There us no need to introduce the idea of repeating the same experiment over and over.

In other words, the confidence interval describes the construction of the confidence interval itself. 95% of the time, it will contain the true (unknown) parameter value. Wasserman's description contains a reference to the *true* parameter value, so if you are going to talk about the true parameter value, his description is useful. However, the first definition of the confidence interval is quite useful if you want to think about how repeated experiments will end up.

We will use the first definition in thinking about how to construct a confidence interval. To construct the confidence interval, then, we will repeat the experiment over and over again, each time computing $\hat{\theta}$. We will then generate an ECDF of our $\hat{\theta}$ values, and report the 2.5th and 97.5th percentile to get our 95% confidence interval. But wait, how will we repeat the experiment so many times?

Remember that the data come from a probability distribution with CDF $F(x)$. Doing an experiment where we make n measurements amounts to drawing n numbers out of $F(x)$[14]. So, we could draw out of $F(x)$ over and over again. The problem

[14] We're being loose with language here. We're drawing out of the distribution that has CDF $F(x)$, but we're saying "draw out of F" for short.

is, we do now know what $F(x)$ is. However, we do have an empirical estimate for $F(x)$, namely $\hat{F}(x)$. So, we could draw n samples out of $\hat{F}(x)$, compute $\hat{\theta}$ from these samples, and repeat. This procedure is called **bootstrapping**.

To get the terminology down, a **bootstrap sample**, $\mathbf{x}^*$, is a set of n x values drawn from $\hat{F}(x)$. A **bootstrap replicate** is the estimate $\hat{\theta}^*$ obtained from the bootstrap sample $\mathbf{x}^*$. To generate a bootstrap sample, consider an array of measured values $\mathbf{x}$. We draw n values out of this array, *with replacement*. This is equivalent to sampling out of $\hat{F}(x)$.

So, the recipe for generating a bootstrap confidence interval is as follows.

1) Generate B independent bootstrap samples. Each one is generated by drawing n values out of the data array with replacement.
2) Compute $\hat{\theta}$ for each bootstrap sample to get the bootstrap replicates.
3) The $100(1-\alpha)$ percent confidence interval consists of the percentiles $100\alpha/2$ and $100(1-\alpha/2)$ of the bootstrap replicates.

This procedure works for any estimate $\hat{\theta}$, be it the mean, median, variance, skewness, kurtosis, or any other esoteric thing you can think of. Note that we use the empirical distribution, so there is never any assumption of an underlying "true" distribution. Thus, we are doing *nonparametric inference* on what we would expect for parameters coming out of unknown distributions; we only know the data. We will not discuss Bayesian nonparameterics, but they are generally not nearly as straightforward.[15] In this way, frequentist procedures are often useful in the nonparametric context.

There are plenty of subtleties and improvements to this procedure, but this is most of the story. We will discuss bootstrap confidence intervals for regression parameters in the tutorials, but we have already covered the main idea.

6.5 Hypothesis tests

The frequentist analog to model comparison is hypothesis testing. But we should be careful, it is an analog, but *most definitety not the same thing.* It is important to note that frequentist hypothesis testing is different from Bayesian model comparison in that in frequentist hypothesis tests, we will only consider how probable it is to get the observed data under a specific hypothesis, often called the **null hypothesis**. It is

[15] But Bayesian nonparametrics is a fascinating and useful field. The basic idea is that you have infinite dimensional priors over models and proceed with Bayesian inference from there. A new book on the subject, *Fundamentals of Nonparametric Baysian Inference*, by Ghosal and van der Vaart, is a good, complete reference.

just a name for the hypothesis you are testing. We will not assess other hypotheses nor compare them. Remember that the probability of a hypothesis being true is not something that makes any sense to a frequentist.

A frequentist hypothesis test consists of these steps.

1) Clearly state the null hypothesis.
2) Define a **test statistic**, a scalar value that you can compute from data. Compute it directly from your measured data.
3) *Simulate* data acquisition for the scenario where the null hypothesis is true. Do this many times, computing and storing the value of the test statistic each time.
4) The fraction of simulations for which the test statistic is at least as extreme as the test statistic computed from the measured data is called the **p-value**, which is what you report.

We need to be clear on our definition here. The p-value is the probability of observing a test statistic being at least as extreme as what was measured if the null hypothesis is true. It is exactly that, and nothing else. It is not the probability that the null hypothesis is true.

Importantly, **a hypothesis test is defined by the null hypothesis, the test statistic, and what it means to be at least as extreme.** That uniquely defines the hypothesis test you are doing. All of the named hypothesis tests, like the Student-t test, the Mann-Whitney U-test, Welch's test, etc., describe a specific hypothesis with a specific test statistic, with a specific definition of what it means to be at least as extreme (e.g., one-tailed or two-tailed). I can never remember what these are, nor do I encourage you to; you can always look them up. Rather, you should just clearly write out what your test is in terms of the hypothesis, test statistic, and definition of extreme.

Now, the real trick to doing a hypothesis test is step 3, in which you simulate the data acquisition assuming the null hypothesis was true. I will demonstrate two hypothesis tests and how we can simulate them. For both examples, we will consider the commonly encountered problem of performing the same measurements under two different conditions, control and test. You might have in mind the example of Dorsal gradient widths for wild type Dorsal versus those of the Dorsal-Venus construct.

Test and control come from the same distribution. Here, the null hypothesis is that the distribution F of the control measurements is the same as that G of the test, or $F = G$. To simulate this, we can do a **permutation test**. Say we have n measurements from control and m measurements from test. We then concatenate arrays of the control and test measurements to get a single array with $n + m$ entries. We then randomly scramble the order of the entries (this is implemented in `np.random.permuation()`). We take the first n to be labeled "control" and the last m to be labeled "test." In this way, we are simulating the null hypothesis: whether or not a sample is test or control makes no difference.

For this case, we might define our test statistic to be difference of means, or difference of medians. These can be computed from the two data sets and are a scalar value.

Test and control have the same mean. The null hypothesis here is exactly as I have stated, and nothing more. To simulate this, we shift the data sets so that they have the same mean. In other words, if the control data are $\mathbf{x}$ and the test data are $\mathbf{y}$, then we define the mean of all measurements to be

$$\bar{z} = \frac{n\bar{x} + m\bar{y}}{n+m}. \tag{6.30}$$

Then, we define

$$x_{\text{shift},i} = x_i - \bar{x} + \bar{z}, \tag{6.31}$$

$$y_{\text{shift},i} = y_i - \bar{y} + \bar{z}. \tag{6.32}$$

$$\tag{6.33}$$

Now, the data sets $\mathbf{x}_{\text{shift}}$ and $\mathbf{y}_{\text{shift}}$ have the same mean, but everything else about them is the same as $\mathbf{x}$ and $\mathbf{y}$, respectively.

To simulate the null hypothesis, then, we draw bootstrap samples from $\mathbf{x}_{\text{shift}}$ and $\mathbf{y}_{\text{shift}}$ and compute the test statistic from the bootstrap samples, over and over again.

In both of these cases, no assumptions were made about the underlying distributions. Only the empirical distributions were used; these are nonparametric hypothesis tests.

6.5.1 Interpretation of the p-value

If the p-value is small, the effect is said to be **statistically significant**. But what is small? I strongly discourage a bright line p-value used to deem a result statistically significant or not. You computed the p-value, it has a specific meaning; you should report it. I do not see a need to convert a computed value, the p-value, into a Boolean, True/False on whether or not we attach the word "significant" to the result.

The question the p-value addresses is rarely the question we want to ask. For example, say we are doing a test of the null hypothesis that two sets of measurements have the same mean. In most cases, which of the following questions are we interested in asking:

1) How different are the means of the two samples?
2) Would we say there is a statistically significant difference of the means of the two samples? Or, more precisely, what is the probability of observing a difference in means of the two samples at least as large as the the observed difference in means, if the two samples in fact have the same mean?

The second question is convoluted and often of little scientific interest. I would say that the first question is much more relevant. To put it in perspective, say we made trillions of measurements of two different samples and their mean differs by one part per million. This difference, though tiny, would still give a low p-value, and therefore often be deemed “statistically significant.” But, ultimately, it is the size of the difference, or the *effect size* we care about.

6.5.2 What is with all those names?

You have no doubt heard of many named frequentist hypothesis tests, like the Student-t test, Welch’s t-test, the Mann-Whitney U-test, and countless others. What is with all of those names? It helps to think more generally about how frequentist hypothesis testing is usually done.

To do a frequentist hypothesis test, people unfortunately do not do what I laid out above, but typically follow the following prescription (borrowing heavily from the treatment in Gregory’s excellent book).

1) Choose a null hypothesis. This is the hypothesis you want to test the truth of.
2) Choose a suitable test statistic that can be computed from measurements *and* has a predictable distribution. For the example of two sets of repeated measurements, we can choose as our statistic

$$\begin{aligned}
T &= \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{S_D\sqrt{n_1^{-1} + n_2^{-1}}},\\
\text{where } S_D^2 &= \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2},\\
\text{with } S_1^2 &= \frac{1}{n_1 - 1}\sum_{i\in D_1}(x_i - \bar{x}_1)^2, \qquad (6.34)
\end{aligned}$$

and S_2^2 similarly defined. The T statistic is the difference of the difference of the observed means and the difference of the true means, weighted by the spread in the data. This is a reasonable statistic for determining something about means from data. This is the appropriate statistic when σ_1 and σ_2 are both unknown but assumed to be equal. (When they are assumed to be unequal, you need to adjust the statistic you use. This test is called Welch's t-test.) It can be derived that this statistic has the Student-t distribution,

$$P(t) = \frac{1}{\sqrt{\pi \nu}} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \left(\frac{t^2}{\nu}\right)\right)^{-\frac{\nu+1}{2}}, \tag{6.35}$$

$$\text{where } \nu = n_1 + n_2 - 2. \tag{6.36}$$

3) Evaluate the statistic from measured data. In the case of the Student-t test, we compute T.
4) Plot $P(t)$. The area under the curve where $t > T$ is the p-value, the probability that we would observe our data under the null hypothesis. Reject the null hypothesis if this is small.

As you can see from the above prescription, item 2 can be tricky. Coming up with test statistics *that also have a distribution that we can write down* is difficult. When such a test statistic is found, the test usually gets a name. The main reason for doing things this way is that most hypothesis tests were developed before computers, so we couldn't just bootstrap our way through hypothesis tests. (The bootstrap was invented by Brad Efron in 1979.) Conversely, in the approach we have taken, sometimes referred to as "hacker stats," we can invent any test statistic we want, and we can test is by numerically "repeating" the experiment, in accordance with the frequentist interpretation of probability.

So, I would encourage you not to get caught up in names. If someone reports a p-value with a name, simply look up the things you need to define the p-values (the hypothesis being tested, the test statistic, and what it means to be as extreme), and that will give you an understanding of what is going on with the test.

That said, many of the tests with names have analytical forms and can be rapidly computed. Most are included in the `scipy.stats` module. I have chosen to present a method of hypothesis testing that is intuitive with the frequentist interpretation of probability front and center. It also allows you to design your own tests that fit a null hypothesis that you are interested in that might not be "off-the-shelf."

6.5.3 Warnings about hypothesis tests

There are many.

1) An effect being statistically significant does not mean the effect is significant in practice or even important. It only means exactly what it is defined to mean:

an effect is unlikely to have happened by chance under the null hypothesis. Far more important is the **effect size**.

2) The p-value is **not** the probability that the null hypothesis is true. It is the probability of observing the test statistic being at least as extreme as what was measured if the null hypothesis is true. I.e., if H_0 is the null hypothesis,

$$\text{p-value} = P(\text{test stat at least as extreme as observed} \mid H_0). \tag{6.37}$$

It is not the probability that the null hypothesis is true given that the test statistic was at least as extreme as the data.

$$\text{p-value} \neq P(H_0 \mid \text{test stat at least as extreme as observed}). \tag{6.38}$$

We often actually want the probability that the null hypothesis is true, and the p-value is often erroneously interpreted as this to great peril.

3) Null hypothesis significance testing does not say anything about alternative hypotheses. Rejection of the null hypothesis does not mean acceptance of any other hypotheses.

4) P-values are not very reproducible, as we will see in the tutorials when we do "dance of the p-values."

5) Rejecting a null hypothesis is also kind of odd, considering you computed

$$P(\text{test stat at least as extreme as observed} \mid H_0). \tag{6.39}$$

This does not really describe the probability that the hypothesis is true. This, along with point 4, means that the p-value better be *really* low for you to reject the null hypothesis.

6) Throughout the literature, you will see null hypothesis testing when the null hypothesis is not relevant at all. People compute p-values because that's what they are supposed to do. The Dorsal gradient might be an example: of course the gradients will be different; we have made a big perturbation. We slapped a giant glowing barrel onto the Dorsal protein. Again, it gets to the point that **effect size** is waaaaay more important than a null hypothesis significance test.

Given all these problems with p-values, I generally advocate for their abandonment. I am not the only one. They seldom answer the question scientists are asking and lead to great confusion.