Part IV
Hypothesis testing

Do not put your faith in what statistics say until you have carefully considered what they do not say.
William W. Watt

A caricature of one recipe might read: Apply a significance test to each result, believe the result implicitly if the conventional level of significance is reached, believe the null hypothesis otherwise. Such a complete flight from reality and its uncertainties is fortunately rare, but periodically considering its extremism may help us keep our balance.
F. Mosteller & J. W. Tukey, 1977, p 25.

...no one believes an hypothesis except its originator but everyone believes an experiment except the experimenter.
W. I. B. Beveridge, 1950, p 65.

Simple hypotheses[29]

The simplest version of the hypothesis testing problem is as follows: we have two possible models, $p_0(D)$ and $p_1(D)$, and based on the observed data have to make a choice between them. (This is called, appropriately enough, a "simple" hypothesis test; later we'll consider testing between more than just two hypotheses at a time.) The example to keep in mind: we draw samples x_i i.i.d. from a Gaussian distribution. We know the Gaussian's variance is 1 and we know that the mean is either -1 or 1. I.e., we may take

$$p_0(D) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}} e^{-(x_i+1)^2/2}$$

and

$$p_1(D) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}} e^{-(x_i-1)^2/2}.$$

How do we distinguish between these two hypotheses (models of the world)?

Well, given that we've just spent a month or so talking about likelihood-based methods, one obvious approach would be to use maximum likelihood. That is, choose the hypothesis under which the likelihood of the observed data is largest. In other words, we look at the likelihood ratio

$$\frac{p_0(D)}{p_1(D)} = \exp\left(\frac{1}{2}\sum_{i=1}^{N}(x_i-1)^2 - (x_i+1)^2\right) = \exp\left(-2\sum_{i=1}^{N} x_i\right);$$

if this ratio is larger than 1, then $p_0(D) > p_1(D)$ and we decide that the true mean was -1, or otherwise choose 1. This is a straightforward and intuitive thing to do, and we'll see in just a moment that in many cases this approach is in fact optimal.

But first let's look a little more closely at our decision rule here. If we simplify the above likelihood ratio, we see that our decision really comes down to: if $\sum_i x_i > 0$, choose 1, and otherwise choose -1. (Of course in theory $\sum_i x_i$ could equal zero, in which case we could just flip a coin; but this exact-tie case happens with probability zero here, so we'll ignore it for now.) If we recall, $\sum x_i$ was a minimal sufficient statistic for the Gaussian

[29] HMC 8.1.

family with known variance (write the Gaussian as an exponential family to remind yourself of this fact if you've forgotten).

And of course this can be generalized: tests between two hypotheses based on likelihood ratios only depend on the data through sufficient statistics. **Exercise 104**: Prove this, using the product decomposition for sufficiency.

Exercise 105: Let's say we observe N i.i.d. observations x_i from an exponential distribution whose mean is known to be either 1 or 2. Write down the ML test between these two alternatives, in as simple a form as possible. What role does the minimal sufficient statistic of this exponential family play?

Exercise 106: Let's say we observe N i.i.d. observations x_i from a Gaussian distribution whose mean is known and whose variance is known to be either 1 or 2. Write down the ML test between these two alternatives, in as simple a form as possible. What role does the minimal sufficient statistic of this exponential family play?

Decision-theoretic approach

What if we take a more general decision-theoretic point of view? I.e., we have some prior information and a cost function. Now what is the optimal decision rule? Let's write down our expected loss function:

$$E[C(truth, guess(D))] = \sum_{truth=i\in\{-1,1\}} p(i) \sum_D p_i(D) C(i, guess(D)).$$

Here $C(.,.)$ is some cost function (just a two-by-two table of numbers, in this case) and $p(i)$ is the prior probability of model i. Now as usual we want to choose $guess(D)$ in such a way as to minimize the expected cost. It's clear that all we need to do, for each possible data observation D, is to choose $guess(D)$ such that

$$\sum_{truth=i\in\{-1,1\}} p(i)p_i(D)C(i, guess(D))$$

is as small as possible.

Now let's simplify things a little: assume $C(i,i) = 0$, that is, there's no cost associated with choosing the model correctly. Now the optimal decision rule is as follows:

$$guess(D) = \begin{cases} 1 & \text{if } p(-1)p_{-1}(D)C(-1,1) < p(1)p_1(D)C(1,-1) \\ -1 & \text{otherwise.} \end{cases}$$

Thus we see that if the cost of errors is symmetric $C(-1,1) = C(1,-1)$, and each hypothesis as equally probable *a priori*, $p(-1) = p(1)$, then our decision rule is exactly the ML rule described above. So our intuitive ML approach is actually a special case of the decision-theoretic optimal rule. More generally, the optimum rule says that if our prior belief is that -1 is more likely than 1, then it makes sense to "lean" towards -1 in the sense that we will choose -1 even if the likelihood ratio is slightly weighted towards 1.

Exercise 107: Repeat the last two homework problems (the exponential and Gaussian ML hypothesis tests) in this more general decision-theoretic context. What is the decision-theoretically optimal test if the costs of a mistake are $C(1,2) = a$ and $C(2,1) = b$, as a function of the prior distribution? (Assume again that $C(i,i) = 0$.)

Null and alternate hypotheses

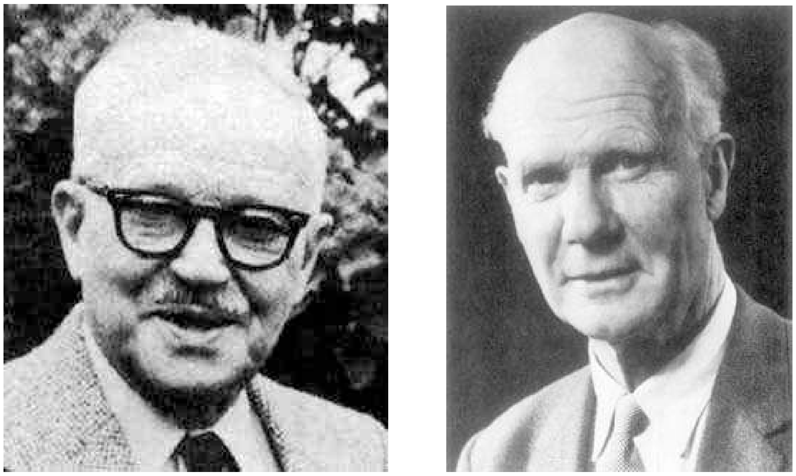

Figure 13: Neyman and Pearson.

In the above discussion we treated both hypotheses equally. In some situations, though, it makes more sense to distinguish between the two hypotheses. For example, if we are testing the fairness of a coin, it might be reasonable to think of the hypothesis that the coin is fair as the “null” hypothesis $p_0(D)$, and the hypothesis that the coin is biased (towards heads, say) as the “alternate” hypothesis, $p_1(D)$. Some specialized terminology has developed in this case:

- the “critical region” A of a test is the region of data space such that if the data D falls in A, we “reject” the null hypothesis; that is, we choose the alternate hypothesis instead.

- the probability $\alpha = \int_{D \in A} p_0(D)$ of incorrectly rejecting the null is called the “size,” or “significance level”; this kind of error is called a “type I” error.

- a “type II” error is when we incorrectly accept the null hypothesis.

- the probability $\int_{D \in A} p_1(D)$ of correctly rejecting the null is called the “power.”

Now, clearly, we want to make the power of our test as large as possible, while making the size as small as possible. These are contrary goals, of course: making A smaller decreases the size but also decreases the power. So one approach is to hold the size fixed at a given level, say $\alpha = 0.05$, and then try to maximize the power over all possible tests with size less than or equal to 0.05.

It turns out this is not hard to do. Conveniently, this optimal test is of exactly the likelihood ratio form we dealt with above.

Theorem 6 (Neyman-Pearson lemma)**.** *The likelihood ratio test*

$$A = \{D : \frac{p_1(D)}{p_0(D)} \geq T_\alpha\},$$

with the threshold T_α chosen so that the size of the test is equal to α, is the most powerful test of size α.

Proof. Let A_1 be the critical region of any other test of size α. We need to prove that A is at least as powerful as A_1. We have

$$\begin{aligned}
\int_{D \in A} p_1(D) - \int_{D \in A_1} p_1(D) &= \int_{D \in A \cap A_1^c} p_1(D) - \int_{D \in A_1 \cap A^c} p_1(D) \\
&\geq T \int_{D \in A \cap A_1^c} p_0(D) - T \int_{D \in A_1 \cap A^c} p_0(D) \\
&= T \int_{D \in A} p_0(D) - T \int_{D \in A_1} p_0(D) \\
&= T\alpha - T\alpha = 0.
\end{aligned}$$

□

To return to our Gaussian example above, we have that the most powerful test of size α is to choose 1 whenever $\sum_{i=1}^N x_i > T_\alpha$, where T_α is chosen such that

$$\alpha = \int_{T_\alpha}^{\infty} \frac{1}{\sqrt{2N\pi}} e^{-(u+1)^2/2N} du.$$

So, to sum up, all three of the approaches we've looked at — ML, decision-theoretic, and Neyman-Pearson — all say basically the same thing: for simple hypothesis testing, the optimal thing to do is to use a test based on the likelihood ratio.

Exercise 108: What is the most powerful test between two exponential distributions with mean 1 and 2, at some fixed size α? What is the power of this test as a function of the number of samples N?

Exercise 109: What is the most powerful test between two uniform distributions, $U([0,1])$ and $U([0,2])$, at some fixed size α? What is the power of this test as a function of the number of samples N?

A side note: it is possible to use many of the same tricks we developed earlier to describe the asymptotic power as N becomes large. We won't go into the details, but if we look at the log-likelihood ratio

$$\sum_{i=1}^{N} \log \frac{p_0(x_i)}{p_1(x_i)},$$

it's clear that we may apply the LLN, CLT, etc. to elucidate the asymptotic behavior here; again we find that the Kullback-Leibler divergence plays a key role in determining the asymptotic behavior of the error (the main difference here is that the hypothesis testing error goes to zero *exponentially* in N, while we saw in the last section that the estimation error, at least in the mean-square setting, goes to zero like $1/N$). We leave the details to the interested reader.

Compound alternate hypotheses[30]

More generally, we're interested in *compound* hypothesis tests: were the data generated by a model $\theta \in H_0$ or $\theta \in H_A$, where H_0 and H_A are two disjoint *sets* of models, the null and alternate sets respectively.

We'll start with the simplest case: the null hypothesis is simple (that is, H_0 consists of just one model, $p_0(D)$), but the alternate is compound. In this case it's reasonable to ask if there is a test with a given size which maximizes the power over every single alternate hypothesis $p_1(D) \in H_A$. (Remember, the size of a test only depends on p_0, so the size will be the same for all p_1 here.) Such a test is called a "uniformly most powerful test," or UMP test for short.

From the Neyman-Pearson theory, we already know that a UMP test has to be based on likelihood ratio tests. This makes it clear that UMP tests often simply don't exist: **Exercise 110**: Prove that no UMP test exists when we are drawing data from a Gaussian of variance 1, if the null hypothesis is $\mu = 0$ and the alternate hypothesis is $\mu \neq 0$. (Hint: break the alternate hypothesis into two sets, $\mu > 0$ and $\mu < 0$, and look at the likelihood ratio tests for each of these individual alternate hypotheses. If these tests are not the same, then argue that no UMP test can exist.)

Is there a simple way to guarantee that a UMP test exists? Well, by the same logic we used above, it's enough to establish that the LR test is independent of $\theta \in H_A$. Here's an example: look at our old friend the exponential family in canonical form. Let's write down the loglikelihood ratio given i.i.d. observations:

$$\begin{aligned}
\log \frac{p_0(D)}{p_{\theta \in H_A}(D)} &= \log \frac{\exp[\theta_0 \sum_i k(x_i) + \sum_i s(x_i) + Ng(\theta_0)]}{\exp[\theta \sum_i k(x_i) + \sum_i s(x_i) + Ng(\theta)]} \\
&= [\theta_0 - \theta] \sum_i k(x_i) + N[g(\theta_0) - g(\theta)].
\end{aligned}$$

It's not too hard to see that a test of the form $T(D) = \sum k(x_i) > c$, for some constant c, leads to a UMP test of the hypothesis that $\theta_0 > \theta$ here. **Exercise 111**: Complete this argument.

How do we choose the test when no UMP test exists? Well, this puts us back in the situation we've encountered before, when e.g. no uniformly

[30] HMC 8.2.

optimal decision rule or UMVUE was available: we have to look at other optimality criteria, e.g. minimax or Bayesian criteria. For example, a Bayesian would choose a hypothesis according to its posterior probability ratio, namely

$$\frac{P(H_0|D)}{P(H_A|D)} = \frac{p(H_0)p_0(D)}{\int_{\theta \in H_A} p(\theta)p_\theta(D)}.$$

More generally, we can always choose some test statistic, $T(D)$, compute its sampling distribution under the null hypothesis $p_0(T)$, and then ask if the observed $T(D)$ is "significantly different" than what we would have expected under the null hypothesis. For example, if we see that $T(D)$ falls outside of the interval defined by the 1st and 99th quantile of $p_0(T)$, then it is reasonable to suspect that D was not, in fact drawn from $p_0(D)$, but rather from some other distribution, and we would reject the null hypothesis here. (This test would not be guaranteed to be optimal in any sense — and importantly, depending on your choice of your test statistic T, your test might have much more power against some alternatives than others — but nonetheless this idea often leads to useful tests in the real world.)

This leads naturally to the concept of a "power curve": namely, for a given test (at some given size), we plot the power $\int_A p_\theta(x)dx$ as a function of $\theta \in H_A$. This function plays a similar role to the risk function played in our decision theory section: basically, we want to make the power as large as possible over the "relevant" part of the parameter space (where "relevant" here depends in some sense on which θ are allowed, or most probable, etc.), just as we wanted to make the risk function as small as possible over the relevant part of the parameter space.

We can also make use of the large sample estimation theory we learned not so long ago: recall that if we have a consistent estimator $\hat{\theta}$ for θ and know that

$$\sqrt{N}\left(\hat{\theta}_N - \theta\right) \rightarrow_D \mathcal{N}(0, V(\theta)),$$

and a consistent estimator of $\sqrt{V(\theta)}$, $\hat{\sigma}$, then as we discussed earlier we may build tests for θ_0 versus $\theta > \theta_0$ based on $\sqrt{N}(\hat{\theta} - \theta)/\hat{\sigma}$ with an asymptotically correct significance level.

Compound null and alternate hypotheses

In this last section we talk about the general case: both null and alternate hypotheses are compound. To find a test of size α here, we would look for some critical region A such that

$$\int_A p_\theta = \alpha \ \forall \ \theta \in H_0.$$

Sometimes this isn't even possible, and we have to relax our standards and look for tests such that

$$\int_A p_\theta \leq \alpha \ \forall \ \theta \in H_0.$$

Sometimes it is possible to find a test whose size is the same for all $\theta \in H_0$, though: for example, if we can find a test statistic T whose distribution is the same for all $\theta \in H_0$, then clearly a test constructed on this statistic will have size independent of $\theta \in H_0$ as well.

Here's an example: Gaussian data of unknown variance; we want to test the null $\mu = 0$ versus the alternate $\mu > 0$. Thus H_0 is the set of all Gaussians with mean 0, and H_A is the set of all Gaussians with mean greater than zero. Before we used the sample mean as our test statistic, but this won't work here because its distribution clearly depends on the variance (and therefore the size of any test based on the sample mean will depend on the underlying unknown true variance). But we know that the sample mean and sample variance are sufficient for this family; why don't we look at the standardized sample mean,

$$T(D) = \frac{\bar{x}}{\bar{\sigma}},$$

with

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

and

$$\bar{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2.$$

It's not hard to show that the distribution of $\bar{x}/\bar{\sigma}$ is independent of the variance under the null $\mu = 0$ (**Exercise 112**: prove this); this statistic (or rather, the normalized statistic $\sqrt{N-1}\bar{x}/\bar{\sigma}$) is called a "t-statistic," and its distribution was originally derived by a statistician working for the Guinness brewery who disguised his identity when publishing his work (to avoid getting in trouble for revealing trade secrets) under the pseudonym "Student." Thus the t-statistic is also called "Student's t," and the commonsense process of dividing by the sample standard deviation is known as "studentizing." **Exercise 113**: Derive the distribution of Student's t.

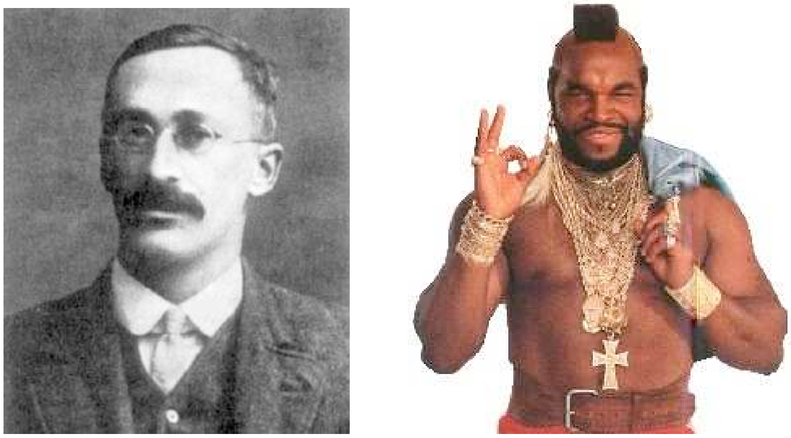

Figure 14: Gosset (aka "Student").

Another example: Gaussian again, but with unknown mean and we'd like to test $\sigma^2 = 1$ versus $\sigma^2 > 1$. Clearly the sample variance is independent of the mean, so we can use this as our test statistic; the null distribution of $N\bar{\sigma}^2$ is a chi-square with $N-1$ degrees of freedom, and thus we can use a test of the form $\bar{\sigma}^2 > c$, for some c chosen such that the size of the test is α. **Exercise 114**: Prove that $N\bar{\sigma}^2$ is a chi-square with $N-1$ degrees of freedom. (Hint: start by proving that the sample variance $\bar{\sigma}^2$ and the sample mean $\bar{x}$ are independent if x_i are i.i.d. Gaussian.)

Outside of the Gaussian family it is a little harder (though not impossible) to find test statistics which are independent of $\theta \in H_0$; nonetheless, again, we may use our large-sample theory to take advantage of this nice property of the Gaussian distribution.

More generally we can always turn back to our Bayesian methods: a Bayesian would choose a hypothesis according to its posterior probability ratio, namely

$$\frac{P(H_0|D)}{P(H_A|D)} = \frac{\int_{\theta \in H_0} p(\theta) p_\theta(D)}{\int_{\theta \in H_A} p(\theta) p_\theta(D)}$$

in general. This is often simplified into the maximum likelihood ratio test, based on

$$\lambda = \frac{\max_{\theta \in H_0} p_\theta(D)}{\max_{\theta \in H_A} p_\theta(D)} :$$

we reject if λ is sufficiently small. (Again, we often resort to large sample theory to determine exactly how small is "sufficiently small.") This maximal likelihood ratio test is sometimes easier to compute numerically than the integral-based Bayesian test, and the two tests turn out to behave similarly asymptotically (this may be shown, again, using expansions of the loglikelihood similar to those we used in establishing the asymptotic behavior of the MLE).