6. Hypothesis testing

6.1. **Introduction.** We now take a look at *hypothesis tests,* which are a rather different kind of statistical procedure. We will again observe the values of a random sample $X_1, \ldots, X_n$, drawn from a distribution that depends on an unknown parameter θ, but rather than try to guess the exact value of θ, we would now like to reach a decision between a *null hypothesis* H_0 and an *alternative* H_1. Both H_0 and H_1 are claims about the parameter θ, so they are of the type $H_0 : \theta \in S_0$, $H_1 : \theta \in S_1$.

Let's look at an example. Say I have a coin that is either fair ($p = \theta = 1/2$), or it comes up heads less often than it should ($p < 1/2$). I want to test the null hypothesis $H_0 : p = 1/2$ against the alternative $H_1 : p < 1/2$. It seems natural to proceed as follows. I will observe a random sample of size n based on this distribution $P(X_1 = 1) = p$, $P(X_1 = 0) = 1 - p$, and I will reject H_0 in favor of H_1 if the observed value of $T = X_1 + \ldots + X_n$ is too small.

This set of outcomes

$$C = \{(x_1, \ldots, x_n) : T(x_1, \ldots, x_n) \leq t\}$$

that leads to the rejection of H_0 is called the *critical region* of the test. If $(x_1, \ldots, x_n) \notin C$, then we will accept H_0; in fact, for typical tests, it is better to say that we are not able to reject H_0.

There are two types of error possible here. If H_0 is true, but is rejected by the test, we speak of a *type I error;* if H_0 is false, but is not rejected, we say that a *type II error* occurred. We will be interested in the corresponding probabilities

$$\alpha = \sup_{\theta \in S_0} P_\theta((X_1, \ldots, X_n) \in C),$$
$$\gamma(\theta) = P_\theta((X_1, \ldots, X_n) \in C) \qquad (\theta \in S_1).$$

We call α the *significance* of the test and $\gamma(\theta)$ its *power function.* So the significance is the probability of committing a type I error, in a worst case scenario, and the power function at a specific $\theta \in S_1$ gives us the probability of avoiding a type II error for this θ.

Ideally, we would like to keep α small and γ close to 1, but these requirements are partly contradictory and we will have to compromise. When designing tests, one usually gives priority to the significance. Typically, one tries to keep this below a previously chosen small probability such as 0.05 or 0.01.

Let's now design concrete tests for the situation outlined above, to detect loaded coins. Let's start out with a test based on a random sample of size $n = 5$. We already said that we want to use a critical region of the type $C = \{x : T(x) \leq t\}$, with $T = X_1 + \ldots + X_n$. If I

take $t = 0$, say, then only the outcome $X_1 = \ldots = X_5 = 0$ is in C, so

$$\alpha = P(X_1 = 0, \ldots, X_5 = 0) = 2^{-5} = \frac{1}{32} = 0.03125;$$

no maximization is necessary here because the null hypothesis determines $\theta = p = 1/2$ completely. This is a reasonably small significance. Unfortunately, we are paying a price for this: the power function of this test is very unimpressive. We have

$$\gamma(p) = P_p(X_1 = 0, \ldots, X_5 = 0) = (1-p)^5.$$

A few typical values are $\gamma(0.1) \simeq 0.59$, $\gamma(0.3) \simeq 0.17$, $\gamma(0.49) \simeq 0.035$. Recall that $\gamma(p)$ is the probability of being able to reject H_0 when it is indeed false, and p is the value of the parameter. So we would like γ to be close to 1, and our test doesn't perform very well in this respect, especially for coins that only have a mild bias.

This is an important general lesson that can be extracted from this simple example. When designing a test in this way, we are really hoping to be able to reject H_0. If this doesn't happen, the test was to some extent a failure or at least not very conclusive. Indeed, suppose in the example above you observed $T = 1$, so you'd have to accept H_0. Obviously, it would be ridiculous to claim that the fact that the coin came up heads once in five coin flips is supporting evidence for what H_0 claims is true, namely, the coin is fair. All that can be said is that this outcome does not justify rejection of H_0 at the chosen significance.

When designing a test with controlled significance, take as the null hypothesis the claim you are hoping to refute. Failure to reject H_0 must not be interpreted as evidence in favor of H_0.

This is the (legal) principle *in dubio pro reo* in action. While we suspect that H_0 is false and are eager to reject it, we design our tests in such a way that false rejections will only happen a controlled percentage of the time when H_0 is actually true.

On a related note, the general remarks I made long ago, in Chapter 2, are also relevant here: ideally, we would like to know conditional probabilities of the type $P(H_0 \text{ holds} | X_1 = x_1, \ldots, X_n = x_n)$, which tell us about the null hypothesis being valid (or not), given what we just observed. These, however, are inaccessible, and what we actually control are the conditional probabilities the other way around, where the condition talks about θ and the event is about the values of the random sample.

Let's return to the concrete example. What happens if I make the critical region larger, in an attempt to increase the power of the test? Say I take $C = \{T \leq 1\}$. While this will increase the power, as

intended, it will unfortunately also hurt the significance: we now have

$$\alpha = P(T = 0, 1) = 2^{-5} + \binom{5}{1} 2^{-5} = 6 \cdot 2^{-5} \simeq 0.19,$$

so the probability of a type I error is now six times as large as before, and one would typically dismiss this test as having an α that is too large. If we want a better power function, we will have to increase the sample size.

So let's try our luck with $n = 10$ now. The critical region will of course again be of the form $T \leq t$. If we want a better power function than above, then we will have to take t at least equal to 1. Let's give the critical region $C : T \leq 1$ a try. Then

$$\alpha = 2^{-10} + \binom{10}{1} 2^{-10} = 11 \cdot 2^{-10} \simeq 0.011,$$

so this test achieves a rather good significance. Its power function is given by

$$\gamma(p) = (1-p)^{10} + 10p(1-p)^9;$$

a few typical values are $\gamma(0.1) \simeq 0.74$, $\gamma(0.3) \simeq 0.15$, $\gamma(0.4) \simeq 0.046$. This isn't much better than before but recall that this test has a better significance. We can sacrifice some of this and take $C : T \leq 2$. Then

$$\alpha = 2^{-10} + \binom{10}{1} 2^{-10} + \binom{10}{2} 2^{-10} = 56 \cdot 2^{-10} \simeq 0.055.$$

Exercise 6.1. Show that for this test, $\gamma(0.1) \simeq 0.93$, $\gamma(0.3) \simeq 0.38$, $\gamma(0.4) \simeq 0.17$.

For large n, we would typically use an approximation by normal distributions to work out the relevant probabilities, backed up by the Central Limit Theorem. More precisely, the random variable

$$\frac{T - np}{\sqrt{np(1-p)}}$$

is approximately $N(0, 1)$-distributed. In particular, this means that if we want to achieve a significance α with a critical region of the type $T \leq t$, then we need to choose t such that

$$P\left(Z \leq \frac{2t - n}{\sqrt{n}}\right) = \alpha$$

for a random variable $Z \sim N(0, 1)$.

Exercise 6.2. Derive this in more detail.

For example, if we want $\alpha = 0.01$, then we need

$$\frac{2t - n}{\sqrt{n}} = -2.33$$

or

$$t = \frac{n}{2} - 1.165\sqrt{n}. \tag{6.1}$$

This follows because $P(Z \leq 2.33) \simeq 0.99$, as we extract from a table. For $n = 1000$, this gives $t = 463$.

Similarly, we can approximate the power function in this style. For the t from (6.1), we have that $T \leq t$ precisely if

$$Z := \frac{T - np}{\sqrt{np(1-p)}} \leq \frac{t - np}{\sqrt{np(1-p)}} = \frac{n(1/2 - p) - 1.165\sqrt{n}}{\sqrt{np(1-p)}},$$

and, as we observed above, the random variable Z is approximately $N(0, 1)$-distributed. For example, if $n = 1000$ and $p = 0.45$, then this is the event $Z \leq 0.84$, which has probability

$$\gamma(0.45) \simeq P(Z \leq 0.84) \simeq 0.8.$$

So this test is quite good at detecting even slight irregularities of our coin. A more biased coin, say $p = 0.4$, will almost certainly be found out: $\gamma(0.4) \simeq 0.99998$

Finally, let's return to small samples, let's say $n = 10$ (though this is not essential for what I'm about to say). This time, let's work with the critical region $C : T = 9, 10$. This critical region certainly looks ridiculous: we are rejecting H_0 precisely for those outcomes that intuitively provide the most convincing evidence *against* the alternative.

However, let us just work out the relevant probabilities anyway. First of all, the significance of this test is the same as the one with critical region $T = 0, 1$, by the symmetry of the distribution for $p = 1/2$. So

$$\alpha = 2^{-10} + 10 \cdot 2^{-10} \simeq 0.011,$$

which is actually reasonably small. So from a formal point of view, our silly looking test performs quite well if our only criterion is the significance. The problem, of course, is the power function, which is unnecessarily small due to our bad choice of critical region. Indeed, it is given by

$$\gamma(p) = p^{10} + 10p^9(1 - p).$$

For example, $\gamma(0.1) \simeq 9 \cdot 10^{-9}$, $\gamma(0.45) \simeq 4.5 \cdot 10^{-3}$, so biased coins are almost never detected by this test. Another undesirable feature is that the power function gets larger as p approaches $1/2$, so the more biased a coin is, the less likely it becomes that our test will detect this. Essentially, this test will erroneously reject about 1.1% of the

fair coins, but it becomes almost impossible that H_0 will get rejected when the coin is in fact appreciably biased. So, to summarize, there are of course very concrete reasons for dismissing this test, but these will have to involve the power function; if the significance is our only concern, then this grotesque test is not inferior to the reasonable tests discussed earlier.

6.2. **Neyman-Pearson tests.** We now analyze these issues more systematically. We would like to develop tests that are in some sense optimal. To do this, we focus for now on a simplified situation: the parameter takes only two values $\theta = \theta_0$ (and this is the null hypothesis) or $\theta = \theta_1$ (which serves as the alternative). It will also be useful to consider *randomized tests,* to gain some extra flexibility when designing tests.

Definition 6.1. A *randomized test* works as follows: Let C, D be disjoint sets of possible values of the random sample, and let $0 \le q \le 1$. Then reject H_0 if $(x_1, \dots, x_n) \in C$, and if $(x_1, \dots, x_n) \in D$, reject H_0 with probability q.

In other words, we reject H_0 unconditionally if the outcome is in the critical region C, and if it is in D, we perform an independent random experiment (we could draw a point uniformly from $[0, 1]$) to decide between rejection and acceptance. This latter part looks strange at first (if my test has a random component, maybe I should have thought about that part a little harder in advance), but it will come in handy when we build tests to order. If $q = 0$ or $q = 1$, the randomness disappears.

We also introduce the function

$$\varphi(x) = \begin{cases} 1 & x \in C \\ q & x \in D \\ 0 & \text{otherwise} \end{cases};$$

here, we have again abbreviated $x = (x_1, \dots, x_n)$. We can now run the test as follows: draw a point u uniformly from $[0, 1]$, and observe the random sample $X_1 = x_1, \dots, X_n = x_n$; reject H_0 if $\varphi(x_1, \dots, x_n) > u$.

We can also express the significance and power in terms of φ. We have

$$\alpha = P_0(X \in C) + qP_0(X \in D) = E_0\varphi(X);$$

the subscript 0 reminds us that we are using the distribution with $\theta = \theta_0$ here. Since the alternative now just states that $\theta = \theta_1$, the power function is no longer a function but just a single probability, and it is similarly given by $\gamma = E_1\varphi(X)$. Finally, note that if, conversely,

we are given a step function φ that takes the three values $0, q, 1$, then this determines a unique randomized test, by just setting $C = \varphi^{-1}(\{1\})$ and $D = \varphi^{-1}(\{q\})$.

In fact, we can go further and also admit arbitrary test statistics $0 \leq \psi(X) \leq 1$. As before, we then reject H_0 if a random number u drawn uniformly from $u \in [0,1]$ satisfies $u \leq \psi(X)$. We have no real use for such tests here, but for example Theorem 6.4 is true (and will be proved) in this generality.

Definition 6.2. We say that a test φ is *most powerful* of significance α if $E_0\varphi(X) = \alpha$ and $E_1\varphi(X) \geq E_1\psi(X)$ for all tests ψ with $E_0\psi(X) \leq \alpha$.

So, as the terminology suggests, a most powerful test achieves the best possible power at a given significance. Notice that things are done in this order: we first insist on a specific value for the significance and then optimize the power under this constraint.

Definition 6.3. A test of the form

$$\varphi(x) = \begin{cases} 1 & L_0(x) < kL_1(x) \\ q & L_0(x) = kL_1(x) \\ 0 & L_0(x) > kL_1(x) \end{cases}$$

is called a *Neyman-Pearson test.*

Here L refers to the likelihood function. Recall that $L(x_1, \ldots, x_n) = P(X_1 = x_1, \ldots, X_n = x_n)$ in the discrete case, and in the continuous case, we work with the product of the individual densities instead.

The NP test is very reminiscent of maximum likelihood estimation: we reject H_0 if what we actually observed was, before it happened, not sufficiently likely under H_0, compared with its probability under H_1. We can also say that we make the *likelihood ratio* $T = L_0(X)/L_1(X)$ our test statistic, and we will reject H_0 if $T < k$.

Example 6.1. Let's return to the coin flip example, with a coin that is either fair (= null hypothesis), or is biased with $\theta = p = 1/4$. Then $L_0(x_1, \ldots, x_n) = 2^{-n}$ is actually independent of the outcome, and

$$L_1(x) = 4^{-x_1-\ldots-x_n}(3/4)^{n-x_1-\ldots-x_n}.$$

We can simplify this by introducing the statistic $T = X_1 + \ldots + X_n$, as above. Then $L_1(x) = 4^{-t}(3/4)^{n-t} = 3^{n-t}4^{-n}$, where $t = T(x)$. So we are going to reject H_0 when $2^{-n} < k3^{n-t}4^{-n}$, or, equivalently, when $t < t_0$. (If $t = t_0$, there might be a randomized decision.) Here k and t_0 are related by $k = 2^n 3^{t_0-n}$; we won't need this here, but we will

briefly refer to this formula in the next example. We have essentially the recovered the common sense test from the previous section.

Theorem 6.4 (Neyman-Pearson)**.** *For any $\alpha \in [0,1]$, there exist $k = k(\alpha)$ and $q = q(\alpha)$ so that the corresponding NP test has significance α. This test is a most powerful test at this significance.*

This can be made more concrete. If we took $k = q = 0$, then this would give the (trivial) test that always accepts, and this achieves significance 0. We now make the critical region larger by increasing k until we are at the given significance, if this can be done. In other words, if there exists a k such that $P_0(L_0 < kL_1) = \alpha$, then we take this k and set $q = 0$.

Notice that $P_0(L_0 < kL_1)$ should approach 1 when we send $k \to \infty$ because the event gets larger and larger. However, if our distribution is discrete, then the value of P_0 will jump at certain values of k, and there is no guarantee that we will be able to hit α exactly. In this case, we fix the value of k at which we jump past α, that is,

$$P_0(L_0 < kL_1) < \alpha \leq P_0(L_0 \leq kL_1), \tag{6.2}$$

and we will then have to assign a suitable value to q to make the significance exactly equal to α anyway. We will see that the correct value is

$$q = \frac{\alpha - P_0(L_0 < kL_1)}{P_0(L_0 = kL_1)}. \tag{6.3}$$

Notice that the denominator can not be zero here, or otherwise we couldn't have (6.2). Also, if we go through this argument again, we see that k, q are essentially unique. More precisely, any choice of k, q that achieves the desired significance leads to the same $\varphi(X)$; even more precisely, any two such functions will be equal to one another with probability one.

Example 6.2. Let's continue the discussion of Example 6.1. I would like to construct the NP test for $\alpha = 0.05$ and random samples of size $n = 5$. As we observed above, the event $L_0 < kL_1$ can be rewritten as $T < t_0$, where $T = X_1 + \ldots + X_5$. When t_0 increases through 0, the significance becomes $P_0(T = 0) = 2^{-5} = 1/32$, which is still smaller than the desired significance $\alpha = 0.05$. However, when t_0 increases past 1, then we obtain $P_0(T = 0, 1) = 6 \cdot 2^{-5} = 3/16$, which is too large. So we are in the second case above. We must take the k that makes $t_0 = 1$. So $k = 2^5 3^{1-5} = 32/81 \simeq 0.4$. Moreover, we need the q from

(6.3):

$$q = \frac{0.05 - P_0(T=0)}{P_0(T=1)} = \frac{0.05 - 2^{-5}}{5 \cdot 2^{-5}} = \frac{1.6-1}{5} = 0.12$$

Let's summarize: we reject H_0 if $T = 0$, and if $T = 1$, we reject randomly with probability $q = 0.12$.

Proof of Theorem 6.4. We want to check that the test constructed above, in the paragraphs following the statement of the theorem, is a most powerful test at significance α. I'll denote the corresponding step function by φ. Let's first check that the significance is indeed equal to α. This is clear if we are in the first case. If our test was constructed via (6.2), (6.3), then

$$E_0\varphi(X) = P_0(L_0 < kL_1) + qP_0(L_0 = kL_1) = \alpha,$$

by these equations, as desired.

Now let ψ be an arbitrary test with $E_0\psi(X) \leq \alpha$. In the remainder of this proof, I'll assume that the distributions are discrete, but a completely analogous argument works in the continuous case. We split

$$E_1(\varphi(X) - \psi(X)) = \sum_x (\varphi(x) - \psi(x))L_1(x)$$

intro three parts, according to $L_0(x) < kL_1(x)$, $L_0(x) = kL_1(x)$, and $L_0(x) > kL_1(x)$. If we denote the corresponding (partial) sums by S_1, S_2, and S_3, respectively, then I claim that in all three cases, we have

$$S_j \geq \frac{1}{k} \sum{}' (\varphi(x) - \psi(x))L_0(x); \tag{6.4}$$

the prime at the sum sign reminds us that the summation is only over the outcomes consistent with the condition we imposed. Now (6.4) clearly holds for $j = 2$, with equality in fact, because in this case $L_0 = kL_1$. If $j = 1$, then $\varphi(x) = 1$, so $\varphi(x) - \psi(x) \geq 0$ and $L_1 > L_0/k$, so again the inequality holds. Finally, if $j = 3$, then $\varphi(x) = 0$, so this time we have $\varphi(x) - \psi(x) \leq 0$, and since also $L_1 < L_0/k$, we again obtain (6.4). So by putting the individual parts back together, we see that

$$E_1(\varphi(X) - \psi(X)) \geq \frac{1}{k} E_0(\varphi(X) - \psi(X)).$$

Now $E_0\varphi(X) = \alpha$ and $E_0\psi(X) \leq \alpha$ by assumption, so this is ≥ 0, as claimed. □

Example 6.3. Suppose we have $N(0, \sigma)$-distributed data, and we want to test $H_0 : \sigma = 1$ against the alternative $H_1 : \sigma = 2$. Let's set up a NP test with significance $\alpha = 0.05$.

Since $f(x) = (2\pi\sigma^2)^{-1/2}e^{-x^2/(2\sigma^2)}$, the ratio of the likelihood functions is given by

$$\frac{L_0}{L_1} = 2^n e^{-3(x_1^2+\ldots+x_n^2)/8}. \tag{6.5}$$

We will reject H_0 if $L_0/L_1 < k$, where k will be chosen such that the probability of this event (under H_0) equals $\alpha = 0.05$. Note that this will be possible since the distributions are continuous; no randomness is required in the test design.

Now, by (6.5), the condition that $L_0/L_1 < k$ is equivalent to $T > c$, where T denotes the test statistic $T = X_1^2 + \ldots + X_n^2$ (and $c = c(k)$ is a suitable constant that could in principle be obtained from k). Recall that $T \sim \chi^2(n)$. So we can obtain c directly as the unique value for which $P(Y > c) = \alpha = 0.05$ for a $\chi^2(n)$-distributed random variable Y. For example, if $n = 20$, then $c = 31.41$, as we read off from a table. To summarize: the NP test for $n = 20$ at significance $\alpha = 0.05$ will reject $H_0 : \sigma = 1$ if

$$T = X_1^2 + X_2^2 + \ldots + X_{20}^2 > 31.41.$$

In other words, we will be able to reject H_0 if T is at least about 1.5 times as large as its expected value $E_0T = 20$.

By Theorem 6.4, this is the most powerful test (at this significance). What is its power equal to? To answer this, notice that under H_1, the random variables $X_j/2$ are iid and standard normal. Thus, still assuming H_1, we have $T/4 \sim \chi^2(20)$. So if Y again denotes a fixed $\chi^2(20)$-distributed random variable, then we have

$$\gamma = P_1(T > 31.41) = P(Y > 31.41/4) \simeq P(Y > 7.85) > 0.99.$$

For large n, we can work with normal approximations of the distribution of T.

Exercise 6.3. Use the CLT to show that (under H_0), $(T-n)/(2n)^{1/2}$ is approximately $N(0,1)$-distributed for large n.

Exercise 6.4. Show that the (approximate) NP test (for large n) of significance $\alpha = 0.01$ will reject H_0 if

$$\frac{1}{n}\left(X_1^2 + \ldots + X_n^2\right) > 1 + \frac{2.33\sqrt{2}}{\sqrt{n}} = 1 + \frac{3.3}{\sqrt{n}}.$$

The statistic T/n could be used as an estimator for σ^2; in fact, we showed earlier that this is the MVUE for an $N(0,\sigma)$ distribution with parameter $\theta = \sigma^2$. So for large n we only need T/n to be slightly larger than its expected value 1 under H_0 to be able to reject H_0; the excess needed decays at the rate $\sim n^{-1/2}$.

6.3. **Uniformly most powerful tests.** We now drop (some of) the artificial assumptions of the previous section. We again allow *composite* hypotheses $H_j : \theta \in S_j$. As above, we consider randomized tests, and it will again be convenient to describe these in terms of the associated step function $\varphi(X)$. The following analog of Definition 6.2 suggests itself:

Definition 6.5. We say that a test φ is *uniformly most powerful* at significance α if: (1) $\sup_{\theta\in S_0} E_\theta\varphi(X) = \alpha$; (2) if ψ is another test with $\sup_{\theta\in S_0} E_\theta\psi(X) \leq \alpha$, then $E_\theta\varphi(X) \geq E_\theta\psi(X)$ for all $\theta \in S_1$.

So again a UMP test achieves the best power function at a given significance, but this condition has become even stronger than before: the UMP test beats all other tests at all individual parameter values compatible with the alternative. The first important thing to realize about UMP tests is that they don't always exist.

Example 6.4. This is an artificial example, just designed to make this point. The parameter θ takes three values, let's say $\theta = 0, 1, 2$. We will work with a random sample of size $n = 1$, and the corresponding random variable $X = X_1$ also takes three values $X = 0, 1, 2$, with these probabilities:

X	0	1	2
$\theta = 0$	0.1	0.8	0.1
$\theta = 1$	0	0.1	0.9
$\theta = 2$	0.9	0.1	0

I want to test $H_0 : \theta = 0$ against the alternative $H_1 : \theta = 1, 2$ at significance $\alpha = 0.1$, and I claim that there is no UMP test φ for this. Indeed, among the competitors ψ that we have to compare φ with are the NP tests that test H_0 against the modified alternatives $\theta = 1$ and $\theta = 2$, respectively. In the first case, the NP test has the critical region $X = 2$. This is clear because the likelihood ration $L_0(X)/L_1(X)$ gets minimal at $X = 2$, and we achieve exactly the right significance $\alpha = 0.1$ if we include this value $X = 2$ and nothing else. Let's call this test ψ_1. So $\psi_1(2) = 1$, $\psi_1(x) = 0$ for $x = 0, 1$. Its power is given by $E_1\psi_1(X) = 0.9$.

Similarly, the NP test of H_0 against the alternative $\theta = 2$ is given by $\psi_2(0) = 1$, $\psi_2(x) = 0$ for $x = 1, 2$, and this also has power $E_2\psi_2(X) = 0.9$.

Thus, a hypothetical UMP test φ would have to satisfy $E_\theta\varphi(X) \geq 0.9$ for $\theta = 1, 2$. For example, for $\theta = 1$, this says that

$$0.1\varphi(1) + 0.9\varphi(2) \geq 0.9.$$

Since $0 \le \varphi \le 1$, this implies that $\varphi(2) \ge 8/9$. Similarly, the inequality on the power function at $\theta = 2$ will show that $\varphi(0) \ge 8/9$. But then $E_0\varphi(X) \ge 0.1 \cdot 8/9 + 0.1 \cdot 8/9 > 0.1$, and it now turns out that φ cannot achieve the requested significance. There is no UMP test at significance α. This is not really surprising. The alternative consists of two parts, and what we would really like to do at one these two values of the parameter will not work well at all at the other parameter value, so there is no procedure that is always best. (However, recall that the definition of the MVUE involved a similar *uniform* optimality: the variance of the MVUE beats any other unbiased estimator at each *individual* value of the parameter. As we have discussed at some length, MVUEs exist in a wide variety of situations. So it is certainly not outright ridiculous to hope for such a property.)

Intuitively, we'd probably go with the (randomized) test $\varphi(0) = \varphi(2) = 1/2$, $\varphi(1) = 0$ as a compromise. This sensibly rejects H_0 (some of the time) when the extreme values $X = 0, 2$ occur, which are much more compatible with one of the two possibilities the composite alternative has to offer than with the null hypothesis. However, for either value $\theta = 1, 2$, this test is beaten hands down by the corresponding NP test: for example, $E_1\varphi(X) = 0.45$, while the NP test had power 0.9.

Let us now look at a situation where there are no such conflicts between different θ values from the alternative and UMP tests do exist.

Definition 6.6. A family of distributions (indexed by θ, as usual) is said to have *monotone likelihood ratio* in the statistic T if for all $\theta_1 > \theta_2$, there exists a strictly increasing function $F = F_{\theta_1,\theta_2}$ such that

$$\frac{L(x,\theta_1)}{L(x,\theta_2)} = F(T(x)). \tag{6.6}$$

For example, the coin flip distribution $P(X_1 = 1) = \theta$, $P(X_1 = 0) = 1 - \theta$ has MLR in $T = X_1 + \ldots + X_n$ because $L = \theta^T(1-\theta)^{n-T}$, so

$$\frac{L_1(x)}{L_2(x)} = \left(\frac{\theta_1}{\theta_2}\right)^T \left(\frac{1-\theta_1}{1-\theta_2}\right)^{n-T},$$

and the right-hand side is an increasing function of T when $\theta_1 > \theta_2$ because then θ_1/θ_2 and $(1-\theta_2)/(1-\theta_1)$ are both > 1.

Exercise 6.5. Show that the $N(0,\theta)$ distribution has MLR in $T = X_1^2 + \ldots + X_n^2$.

Exercise 6.6. Consider the exponential distribution $f(x) = e^{-x/\theta}/\theta$, $x > 0$. Show that this distribution has MLR in $T = X_1 + \ldots + X_n$.

Exercise 6.7. Show that the Poisson distribution $P(X_1 = x) = e^{-\theta}\theta^x/x!$ has MLR in $T = X_1 + \ldots + X_n$.

Suppose now that the distribution under consideration has MLR in a statistic T. Consider the null hypothesis $H_0 : \theta \geq \theta_0$, and we test this against the alternative $H_1 : \theta < \theta_0$ (we focus on this situation, but of course a similar analysis is possible for the reverse inequalities). We now design our test as follows:

$$\varphi(X) = \begin{cases} 1 & T(X) < k \\ q & T(X) = k \\ 0 & T(X) > k \end{cases} \tag{6.7}$$

The parameters k and $0 \leq q \leq 1$ are again adjusted so that the test achieves the desired significance α. This condition essentially determines k, q uniquely, in the sense that $\varphi_1 = \varphi_2$ with probability one for any two such functions.

Notice that this test generalizes NP tests in a natural way. In fact, if we use such a φ to test $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$, then it becomes exactly an NP test.

Theorem 6.7. *The test φ is a UMP test.*

Proof. Consider the NP test φ for $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$ for a fixed $\theta_1 < \theta_0$. This test is of the form (6.7) because the NP test has $\varphi = 1$ when $L_0/L_1 < c$, but by (6.6), this condition is equivalent to $F(T(x)) < c$ and thus also to $T(x) < k$, for suitable k, since F is strictly increasing. Also note that φ is independent of θ_1; indeed, we find k, q from the condition

$$P_{\theta_0}(T < k) + qP_{\theta_0}(T = k) = \alpha,$$

which does not involve θ_1.

By Theorem 6.4, we have $E_{\theta_1}\varphi(X) \geq E_{\theta_1}\psi(X)$ for any test ψ with $E_{\theta_0}\psi(X) \leq \alpha$. Moreover, as pointed out, this holds for all $\theta_1 < \theta_0$. So φ is a UMP test.

It remains to show that φ achieves the desired significance. So far, we have only shown that $E_{\theta_0}\varphi(X) = \alpha$, but we need this, as an inequality with $\leq$, for all $\theta \geq \theta_0$.

Fix such a $\theta' > \theta_0$ and consider the test φ as a test for $H_0 : \theta = \theta'$, $H_1 : \theta = \theta_0$. Since φ has the correct general structure, it is an NP test at the significance $\alpha' = E_{\theta'}\varphi(X)$ it achieves. Now by construction, its power is given by $E_{\theta_0}\varphi(X) = \alpha$. This implies that $\alpha \geq \alpha'$, as desired, because an NP test can never have a power that is smaller than its significance. Indeed, if this were the case, then the test $\psi(X) = \alpha'$

(that is, ignore the data and randomly reject with probability equal to the significance) would outperform φ, which is impossible because φ, being an NP test, is most powerful. □

Example 6.5. Consider the $N(\theta, 1)$ distribution

$$f(x) = (2\pi)^{-1/2} e^{-(x-\theta)^2/2}.$$

Let's fix $\theta_1 > \theta_2$ and let's look at the likelihood ratio

$$\begin{aligned} \frac{L_1}{L_2} &= \exp\left(\frac{1}{2} \sum_{j=1}^{n} \left((x_j - \theta_2)^2 - (x_j - \theta_1)^2 \right) \right) \\ &= \exp\left((\theta_1 - \theta_2) \sum_{j=1}^{n} x_j + \frac{1}{2} \left(\theta_2^2 - \theta_1^2 \right) \right). \end{aligned}$$

This formula shows that this distribution has MLR in $T = X_1 + \dots X_n$.

Let's now design the UMP test described above, for $H_0 : \theta \geq 0$, $H_1 : \theta < 0$. Since T is a continuous random variable, we don't need randomization to achieve a requested significance α. Rather, we must choose k so that $P_0(T < k) = \alpha$. Now if $\theta = 0$, then $T \sim N(0, \sqrt{n})$, or we can say that $Z = T/\sqrt{n} \sim N(0, 1)$, and thus we need k with $P(Z < k/\sqrt{n}) = \alpha$ for a standard normal random variable Z. We will then reject H_0 if $T < k$, or maybe it's more convenient to say this in terms of the sample mean: we will reject if $\overline{X} < k/n$.

This completes the general description of this test. For example, if we want $\alpha = 0.05$, then, since $P(Z < -1.645) \simeq 0.05$, we will take $k = -1.645\sqrt{n}$, so we will reject if $\overline{X} < -1.645/\sqrt{n}$. Since $\overline{X}$ is the MVUE for θ, one would naturally be inclined to favor H_1 whenever $\overline{X} < 0$, and indeed, for large n, this test will be in a position to reject H_0 if $\overline{X}$ is just slightly smaller than zero.

Example 6.6. Finally, let's return to the urn with an unknown number N of balls in it. We would like to test $H_0 : N \geq N_0$ against the alternative $H_1 : N < N_0$. We draw balls according to the distribution $P(X_1 = x) = (1/N)\chi_{\{1,\dots,N\}}(x)$. To check whether the theory we discussed above applies, let's fix $N_1 > N_2$ and let's look at the likelihood ratio

$$\frac{L_1(x)}{L_2(x)} = \begin{cases} \left(\frac{N_2}{N_1}\right)^n & \max x_j \leq N_2 \\ \text{undefined} & \max x_j > N_2 \end{cases}. \tag{6.8}$$

The quotient is undefined in the second case because $L_2 = 0$ then. Perhaps we can with some artistic license set $T = \max X_j$ and "define" a function $F(t)$ as $F(t) = (N_2/N_1)^n$ for $t \leq N_2$ and $F(t) = \infty$ otherwise,

so that (perhaps) the current situation would still be close, at least in spirit, to the one of Definition 6.6.

Be that as it may, we can also just set up such a test, without directly referring to any general theory. So we are looking for k, q so that the test φ with $\varphi = 1$ for $T < k$ and $\varphi = q$ for $T = k$ has significance α. Now for integer k,

$$P_{N_0}(T < k) = \left(\frac{k-1}{N_0}\right)^n, \quad P_{N_0}(T = k) = \left(\frac{k}{N_0}\right)^n - \left(\frac{k-1}{N_0}\right)^n,$$

so to achieve the desired significance α, we will have to take the k with $((k-1)/N_0)^n < \alpha \le (k/N_0)^n$ or, equivalently,

$$\alpha^{1/n} N_0 \le k < \alpha^{1/n} N_0 + 1.$$

For large n, since $\alpha < 1$, $\alpha^{1/n} \to 1$, this will give $k = N_0$ and then

$$q = \frac{\alpha - ((N_0-1)/N_0)^n}{1 - ((N_0-1)/N)^n} = \frac{\alpha N_0^n - (N_0-1)^n}{N_0^n - (N_0-1)^n}.$$

This completes the design of our test. We always reject $H_0 : N \ge N_0$ if $T < N_0$, and we reject it with probability q if $T = N_0$. This last part seems strange because if $T = \max X_j$ took the value N_0, we know for sure that H_0 is true and we are deliberately committing a type I error. However, we can afford to do this because $P_{N_0}(T < N_0) = ((N_0-1)/N_0)^n$ is smaller than α (this will certainly be true for fixed $\alpha > 0$ and N_0 and large enough n), and the general design mechanism erroneously thinks it gains extra power in return for this folly. This, however, is not the case: if the alternative $N < N_0$ is valid, then $T < N_0$ also with certainty, so no power is gained by also rejecting (some of the time) for $T = N_0$.

So, to sum this up, the reasonable thing to do is to reject H_0 when $T < N_0$ and (always) accept otherwise. This achieves a better significance than what we asked for (for large n, that is), and it is trivially a UMP test at its significance because its power satisfies $P_N(T < N_0) = 1$. For example, if $N_0 = 10$ and $n = 100$, then we obtain power 1 in this way with a significance of $\alpha = (9/10)^{100} \simeq 2.7 \cdot 10^{-5}$.

Of course, this whole test ("reject H_0 in favor of H_1 precisely if H_1 has not been refuted by the data") looks just like the obvious thing to be doing. So while careful study of sophisticated theory may provide additional insight into the inner workings of things (it is hoped), common sense is not to be disdained either.