Caltech Department of Mathematics

Ma 3/103 KC Border
Introduction to Probability and Statistics Winter 2017

# Lecture 20: Significance Tests, I

**Relevant textbook passages:**

**Larsen–Marx [8]:** Sections 7.2, 7.4, 7.5; 7.3, Appendices

## 20.1 A simple statistical problem

Consider the issue of biasedness of our coins. Over four years we collected 110,848 tosses of which 55,515 were Heads. That is 50.0821% Heads. Is this close enough to 1/2 to warrant concluding that $p$, the probability of Heads is actually 1/2?

Let's think about why we care if $p = 1/2$. First $p = 1/2$ seems to be the default assumption—it's hard to justify believing something else, so the *hypothesis* that $p = 1/2$ seems to have a special place, and I shall refer to it as the **null hypothesis**. It would take some convincing evidence for me to *reject* the hypothesis, but can I quantify that?

Well if in truth, $p = 1/2$, then if $X$ is the number of Heads in $n$ tosses, then the random variable

$$Z = \frac{X - n/2}{\sqrt{n/4}} \sim N(0,1),$$

at least approximately. Now let me choose a probability $\alpha$ that is low enough so that if I observe a value of $T$ that is sufficiently "unlikely" then I am willing to reject my null hypothesis $p = 1/2$. For many people $\alpha = 0.05$ seems like a magic number. So let's adopt this. Now I know (and you have calculated on your homework) that for a standard Normal random variable $Z$,

$$P\left(|Z| > 1.96\right) = 0.05.$$

This suggests the following *decision rule*:

If $\left|\dfrac{x - n/2}{\sqrt{n/4}}\right| > 1.96$, then reject the null hypothesis $p = 1/2$.

Otherwise *accept* (fail to reject) the mull hypothesis.

Note that by construction, even if in truth $p = 1/2$, then with probability $\alpha = 0.05$, I will still reject the null hypothesis.

On the other hand, if in truth the probability of heads is some value $p$, then $(X-np)/\sqrt{np(1-p)}$ is distributed approximately $N(0,1)$, but we will accept the null hypothesis when $|(X-n/2)|/2\sqrt{n} < 1.96$. What is the probability of this? It is small when $p$ is near 1/2, but gets larger as $p$ becomes farther from 1/2.

## 20.2 Another simple statistical problem

Assume $X_1, \dots, X_n$ are independent and identically distributed random variables from a Normal$(\mu, \sigma^2)$ distribution? What can we conclude about $\mu$ and $\sigma^2$? Let's simplify the problem by first assuming that we know that

$$\sigma^2 = 1.$$

(We'll come back to the general problem in a bit.) What can we say about $\mu$?

We already know that given a sample $x_1, \dots, x_n$, the sample average

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n}$$

is the method of moments and the maximum likelihood estimator of $\mu$, that it is an unbiased and consistent estimator and that the 95% confidence for $\mu$ interval is given by

$$\left(\bar{x} - \frac{1.96}{\sqrt{n}\sigma}, \bar{x} + \frac{1.96}{\sqrt{n}\sigma}\right).$$

We also know that

$$\frac{\bar{X} - \mu}{\underbrace{\sigma}_{=1} / \sqrt{n}} \sim N(0, 1).$$

Let's simply the question further by assuming that either $\mu = 0$ or $\mu = 1$, and that it is our job to decide which.

Consider Figure 20.1, which shows the probability densities of the two normals, Normal$(0, 1)$ and Normal$(1, 1)$. Let's make the hypothesis that $\mu = 0$ our null hypothesis. It is clear that large values of $\bar{x}$ make it less likely that $\mu = 0$ is the parameter generating the data.

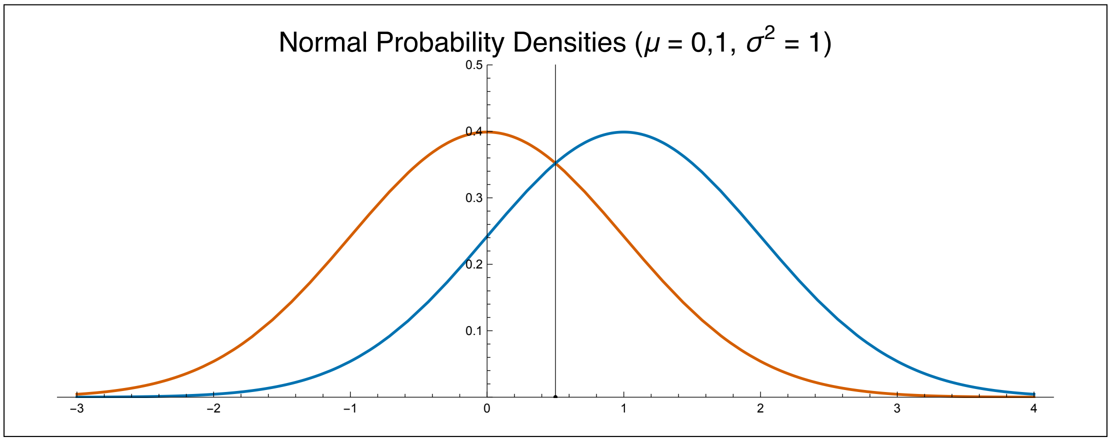


Figure 20.1. A simple hypothesis test.

If in truth $\mu = 0$, then

$$P\left(\frac{\bar{X}}{1/\sqrt{n}} > 1.96\right) = 0.025$$

So setting a critical value $c = 1.96$ and testing whether $\sqrt{n}\bar{X} > c$ ( so-called one-sided test) gives a probability of 0.025 of rejecting the null hypothesis even when it is true.

Now suppose that in fact $\mu = 1$. Then $\sqrt{n}(\bar{X} - 1) \sim N(0, 1)$. Let's compute Prob $(\sqrt{n}\bar{x} < c)$ in this case. This is the case where we decide $\mu = 0$ when in truth $\mu = 1$. Now

$$\sqrt{n}\bar{x} < c \iff \sqrt{n}(\bar{x} - 1) < c - \sqrt{n}$$

so

$$\text{Prob}\left(\sqrt{n}\bar{X} < c\right) = \text{Prob}\left(\sqrt{n}(\bar{X} - 1) < c - \sqrt{n}\right) = 1 - \Phi(c - \sqrt{n}).$$

This is the probability of deciding $\mu = 0$, when in truth $\mu = 1$.

## 20.3 Hypothesis testing in the abstract

A typical **data model** or **probability model** is a density/likelihood function $f(x;\theta)$ depending on a datum $x$ from some set $\mathcal{X}$, and parameterized by $\theta$ belonging to some set $\Theta$. Having estimated parameters is usually not enough. We know that the estimates are random, and that if we are very unlucky our estimates can be very misleading. Usually, we want to *test* some *hypothesis* about the probability model.

- There are two kinds of such hypotheses.

- The first kind is a **hypothesis about the parameter**. A typical such hypothesis is of the form: $\theta \in A$, for some subset $A$ of the parameter space $\Theta$. For instance, we may be interested in whether the mean $\mu$ of a normal is greater than zero, $\mu \in [0,\infty)$. These tests are usually referred to as **significance tests**.

- The second kind of hypothesis is whether our model is **misspecified**. Misspecification occurs when the data generating process is governed by a different functional form $g(x;\psi)$ for some possibly alternate parameter set. (Remember that we could take as our parameter space the set of *all* probability measures on $\mathcal{X}$. But this space is so large we usually call this a nonparametric model.) An example of this might be whether the data come from *any* normal distribution, or weather an entirely different family, such as the Cauchy distribution, is a "better" model. These tests are frequently usually to as**specification tests**.

We'll start with the first kind of testing, parametric hypothesis testing in the context of a fixed data model $f$.

### A bunch of definitions

**Larsen–Marx [8]:** Section 6.2

- The data model: $f\colon \mathcal{X} \times \Theta \to \boldsymbol{R}$. Typically, $\mathcal{X}$ and $\Theta$ are subsets of some finite dimensional Euclidean spaces. $X$ is a random variable (or random vector) with either a density or a mass function given by $f(x;\theta)$. For convenience, I will typically refer to $f$ as a pdf. Recall that the likelihood function $L(\theta;x)$ is just $f(x;\theta)$ so I may also refer to $f$ as the likelihood function. Finally, when it is convenient, I may also write $f_\theta(x)$ for $f(x;\theta)$.

- The **Null Hypothesis**: Denoted $H_0$, it takes the form of statement $\theta \in \Theta_0 \subset \Theta$. It could be a simple as $\theta = 0$ (more properly $\theta \in \{0\}$). The next section elaborates on the role of the null hypothesis.

Often the null hypothesis is something the researcher hopes to prove false. For instance, if you want to show that a drug improves the cure rate for a disease, the null hypothesis is probably going to be that the difference in the cure rate (over the control group) is zero. If the drug is actually useful, then you should reject the null hypothesis.

- **Alternative Hypothesis**: $H_1$ is the hypothesis that $\theta \in \Theta_1 = \Theta \setminus \Theta_0$. The two sets $\Theta_0$ and $\Theta_1$ partition the parameter space $\Theta$.

E.g., $\Theta = [0,\infty)$, $\Theta_0 = \{0\}$, and $\Theta_1 = (0,\infty)$. The null hypothesis is $\theta = 0$ and the alternative is $\theta > 0$.

- A **simple hypothesis** is that $\Theta_i$ has just one point, and a **composite hypothesis** is that $\Theta_i$ has more than one point.

Often a null hypothesis is simple and the alternative is composite, but that needn't be.

The point of a **test** is to, based on the datum (vector) $x$, either **reject** the null hypothesis in favor of the alternative, or to **fail to reject** the null hypothesis. (It is considered a faux pas to say that you accept the null hypothesis.)

How do you decide which hypothesis gets to be the null hypothesis, and which gets to be the alternative?

## 20.4 Choosing the null hypothesis

Brad Efron [2, pp. 556–557] gives a nice discussion of the role of the null and alternative hypotheses in scientific investigation. In the quotation below, the case he is referring to is known as **Bode's Law**, namely that the distance from the sun of the $n^{\text{th}}$ planet is of the form the $d_n = a + b2^n$. (Bode did not have a large sample of solar systems.) Efron is critiquing the analysis of Good [5] who was in turn critiquing Bode. In the quote. Model B is Bode's Law, model $\overline{\text{B}}$ is Good's alternative hypothesis, and C is Efron's alternative. Note this is concerned with model specification rather than parametric hypotheses, but the comments on the role of the null hypothesis are still relevant.

> The most interesting feature of this problem is the light it casts on the role of the null hypothesis in hypothesis testing. These terms are used here in Fisher's sense in which the null hypothesis is by design a hypothesis of uninteresting structure compared to that which we are considering as an alternative, though it may contain its own interesting features. [...]
>
> However it is not necessary to believe in the null hypothesis in order to use it as a test against the alternative of interest. Very often, perhaps most of the time, we do not believe in the validity of the Fisherian null hypothesis, whether or not the test based on it accepts or rejects in the usual sense. [...]
>
> The null hypothesis in the context of this discussion plays the role of devil's advocate, a competitor that an alternative of interest to us must soundly discredit in order to show its strength. [...]
>
> The conclusions of a significance test are bound to be less than completely satisfying given the indirect route of the argument. In the case at hand for instance, accepting C doesn't mean we believe C is true (Figure B mildly discourages such a belief). All we can say is that a statistical model that is relatively uninteresting compared to Bode's law would often yield data as "simple" as that actually observed, and this undermines the necessity of our belief in the law's validity. Conversely even if we had decisively rejected C we still might fear that we had overlooked some other reasonable null hypothesis which would do better.
>
> One should not be dismayed by the limitations of the Fisherian significance test since it is designed only to give us some direction toward the correct answer in situations like the present one where there is little data to work with. As more data accumulate in any given problem, significance testing becomes superfluous.
>
> [...]
>
> By definition "estimation" refers to situations where we believe we know all the possible relevant statistical models and we are simply trying to choose the correct one. Estimation is an inherently more satisfying operation than significance testing, but demands more data or more theoretical knowledge from the statistician.

## 20.5 The abstract mechanics of a statistical test

A statistical test is made up of the following pieces:

- **Test statistic**: $T$, is a function of the data, and of the null hypothesis.

E.g., $T = (\bar{x} - \theta_0)/(s/\sqrt{n})$.

When we discussed estimators, I said that a statistic must be a function of the data, and not of the unknown parameters. If we have a simple null hypothesis, say $\theta = \theta_0$, the test statistic is allowed to depend on $\theta_0$ because in its role as the null hypothesis, $\theta_0$ is not unknown to us—we know quite well what its value is because we picked it.

Even if the null hypothesis is composite, but say of the form $\theta \leqslant \theta_0$, we can allow the test statistic depend on the boundary point $\theta_0$ (or any other fixed feature of the set $\Theta_0$).

- **Critical region**: $C$. If the value of $T$ belongs to $C$, the null hypothesis is **rejected** in favor of the alternative hypothesis. Otherwise we **fail to reject** the null hypothesis.

The critical region is often either an interval (possibly infinite) or the complement of an interval. The endpoint(s) define the **critical value(s)** of the test. For example, if $C = [t^*, \infty)$, then $t^*$ is the critical value.

- Suppose the test statistic $T$ has the value $t$ and critical region is of the form $[c^*, \infty)$ . The probability

$$P_{\theta_0}\left(T > t\right)$$

is called the ***p*-value** of $t$.

An equivalent description of the test is to reject $H_0$ whenever the $p$-value of the statistic is less than $\alpha$.

- If critical region is of the form $(-\infty, c^*] \cup [c^*, \infty)$, the ***p*-value** of $t$ is the probability

$$P_{\theta_0}\left(|T| > |t|\right).$$

- A generalization of the critical region is the **critical function** $\phi$. This allows for randomizing when deciding whether or not to reject the null hypothesis. The value $\phi(t)$ is the probability of rejecting the null when $T(x) = t$. If the test uses a critical region $C$, then $\phi$ is the indicator function $\mathbf{1}_C$. When you would want to randomize? When the test statistic is discrete, and I largely ignore the issue, but read on to the next point.

- **Significance level**: $\alpha$. This is, roughly speaking, the probability $P\left(T \in C\right)$ *when* $H_0$ *is true*. That is, it is the probability of rejecting the null hypothesis when it is true. This also called the **size** of the test. Note: people often misspeak and use $1 - \alpha$ as the significance level. Or they might say that the test is "significant at the $1 - \alpha$ confidence level."

When the null hypothesis is simple, the meaning of $P\left(T \in C\right)$ is clear. The probability $P$ is actually $P_{\theta_0}$, the probability whose pdf is $f(\cdot; \theta_0)$. For a composite null hypothesis, we define the size $\alpha$ to be

$$\sup\{P_\theta\left(T \in C\right) : \theta \in \Theta_0\}.$$

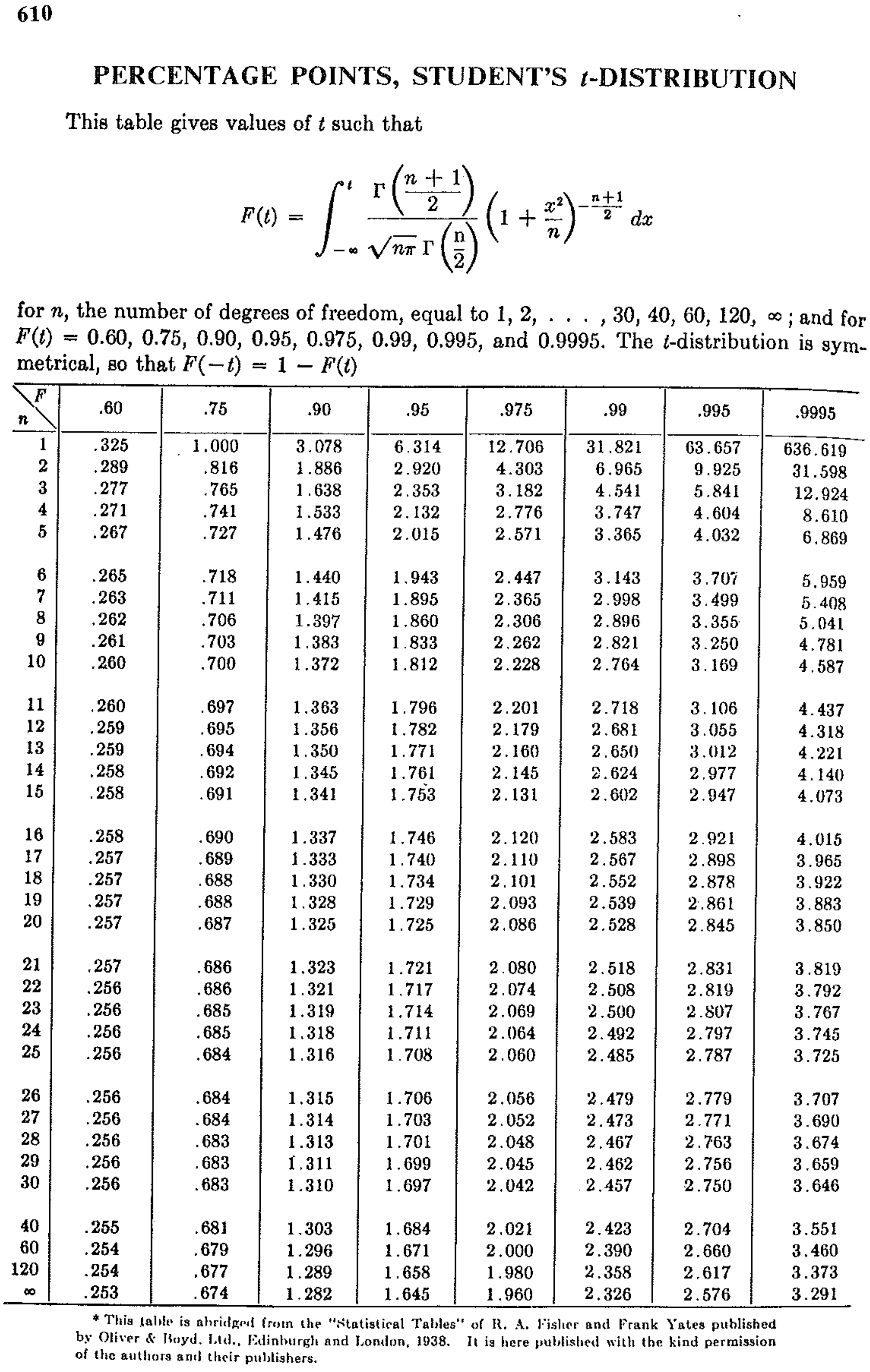
That is, for a composite null hypothesis, the probability of rejecting the hypothesis for any parameter in the null hypothesis is no greater than $\alpha$. The actual probability will depend on which $\theta \in \Theta_0$ is the "true" parameter.

Statisticians are inordinately fond of $\alpha = 0.05$ and $\alpha = 0.01$. This goes back to Ronald A. Fisher's [3, 4] dicta in 1925. It also goes back to the pre-statistical software era in which you had to look up critical values in a table in a book, which listed only a few values of $\alpha$. When I was a freshman, one of the first books that everyone bought was the Chemical Rubber Company's *CRC Standard Math Tables* [12]. You can see a typical page in Figure 20.2.

If the test statistic $T$ is a discrete random variable, for instance, a count of successes, it may be impossible to find a critical region of size exactly $\alpha$ due to the "lumpiness" of $T$. This is when you might want to randomize for a particular value $t$ of $T$, in order to get the probability of rejection exactly equal to $\alpha$. This is when to use a critical function instead of a critical region. When the null hypothesis is simple, $\theta = \theta_0$, then

$$\boldsymbol{E}_{\theta_0}\, \phi(T) = \alpha.$$

- **The significance level and critical region can only be computed if we know the distribution of $T$ given the parameters $\theta$.** This is why so much effort has been devoted to figuring out the distribution of various test statistics.



610

## PERCENTAGE POINTS, STUDENT'S $t$-DISTRIBUTION

This table gives values of $t$ such that

$$F(t) = \int_{-\infty}^{t} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}} dx$$

for $n$, the number of degrees of freedom, equal to 1, 2, . . . , 30, 40, 60, 120, ∞ ; and for $F(t)$ = 0.60, 0.75, 0.90, 0.95, 0.975, 0.99, 0.995, and 0.9995. The $t$-distribution is symmetrical, so that $F(-t) = 1 - F(t)$

| $n \backslash F$ | .60 | .75 | .90 | .95 | .975 | .99 | .995 | .9995 |
|---|---|---|---|---|---|---|---|---|
| 1 | .325 | 1.000 | 3.078 | 6.314 | 12.706 | 31.821 | 63.657 | 636.619 |
| 2 | .289 | .816 | 1.886 | 2.920 | 4.303 | 6.965 | 9.925 | 31.598 |
| 3 | .277 | .765 | 1.638 | 2.353 | 3.182 | 4.541 | 5.841 | 12.924 |
| 4 | .271 | .741 | 1.533 | 2.132 | 2.776 | 3.747 | 4.604 | 8.610 |
| 5 | .267 | .727 | 1.476 | 2.015 | 2.571 | 3.365 | 4.032 | 6.869 |
| 6 | .265 | .718 | 1.440 | 1.943 | 2.447 | 3.143 | 3.707 | 5.959 |
| 7 | .263 | .711 | 1.415 | 1.895 | 2.365 | 2.998 | 3.499 | 5.408 |
| 8 | .262 | .706 | 1.397 | 1.860 | 2.306 | 2.896 | 3.355 | 5.041 |
| 9 | .261 | .703 | 1.383 | 1.833 | 2.262 | 2.821 | 3.250 | 4.781 |
| 10 | .260 | .700 | 1.372 | 1.812 | 2.228 | 2.764 | 3.169 | 4.587 |
| 11 | .260 | .697 | 1.363 | 1.796 | 2.201 | 2.718 | 3.106 | 4.437 |
| 12 | .259 | .695 | 1.356 | 1.782 | 2.179 | 2.681 | 3.055 | 4.318 |
| 13 | .259 | .694 | 1.350 | 1.771 | 2.160 | 2.650 | 3.012 | 4.221 |
| 14 | .258 | .692 | 1.345 | 1.761 | 2.145 | 2.624 | 2.977 | 4.140 |
| 15 | .258 | .691 | 1.341 | 1.753 | 2.131 | 2.602 | 2.947 | 4.073 |
| 16 | .258 | .690 | 1.337 | 1.746 | 2.120 | 2.583 | 2.921 | 4.015 |
| 17 | .257 | .689 | 1.333 | 1.740 | 2.110 | 2.567 | 2.898 | 3.965 |
| 18 | .257 | .688 | 1.330 | 1.734 | 2.101 | 2.552 | 2.878 | 3.922 |
| 19 | .257 | .688 | 1.328 | 1.729 | 2.093 | 2.539 | 2.861 | 3.883 |
| 20 | .257 | .687 | 1.325 | 1.725 | 2.086 | 2.528 | 2.845 | 3.850 |
| 21 | .257 | .686 | 1.323 | 1.721 | 2.080 | 2.518 | 2.831 | 3.819 |
| 22 | .256 | .686 | 1.321 | 1.717 | 2.074 | 2.508 | 2.819 | 3.792 |
| 23 | .256 | .685 | 1.319 | 1.714 | 2.069 | 2.500 | 2.807 | 3.767 |
| 24 | .256 | .685 | 1.318 | 1.711 | 2.064 | 2.492 | 2.797 | 3.745 |
| 25 | .256 | .684 | 1.316 | 1.708 | 2.060 | 2.485 | 2.787 | 3.725 |
| 26 | .256 | .684 | 1.315 | 1.706 | 2.056 | 2.479 | 2.779 | 3.707 |
| 27 | .256 | .684 | 1.314 | 1.703 | 2.052 | 2.473 | 2.771 | 3.690 |
| 28 | .256 | .683 | 1.313 | 1.701 | 2.048 | 2.467 | 2.763 | 3.674 |
| 29 | .256 | .683 | 1.311 | 1.699 | 2.045 | 2.462 | 2.756 | 3.659 |
| 30 | .256 | .683 | 1.310 | 1.697 | 2.042 | 2.457 | 2.750 | 3.646 |
| 40 | .255 | .681 | 1.303 | 1.684 | 2.021 | 2.423 | 2.704 | 3.551 |
| 60 | .254 | .679 | 1.296 | 1.671 | 2.000 | 2.390 | 2.660 | 3.460 |
| 120 | .254 | .677 | 1.289 | 1.658 | 1.980 | 2.358 | 2.617 | 3.373 |
| ∞ | .253 | .674 | 1.282 | 1.645 | 1.960 | 2.326 | 2.576 | 3.291 |

* This table is abridged from the "Statistical Tables" of R. A. Fisher and Frank Yates published by Oliver & Boyd, Ltd., Edinburgh and London, 1938. It is here published with the kind permission of the authors and their publishers.

Figure 20.2. A typical statistical table from the *CRC Standard Math Tables* [12].

## 20.6 Statistics means never having to say you're certain

**Larsen–Marx [8]:** § 6.4

- A **Type I error** occurs when $H_0$ is rejected when in fact it is true. (False rejection.) This can only happen if we reject the hypothesis.

- A **Type II error** occurs when $H_0$ fails to be rejected when in fact it is false. (False acceptance.) This can only happen if we fail to reject the hypothesis.

- **We may never know if an error of either type has occurred!**

- The probability of committing a Type I error is $\alpha$, the significance level of the test. *This is a choice made by the experimenter (we hope) before the experiment is conducted.*

- Once we have chosen the critical region $C$ and the associated level of significance, with luck we can calculate the probability of a Type II error. This probability is frequently referred to as $\beta(\theta)$. It depends on the actual parameter value, which of course we do not know.

- The **complement of a type II error** is to properly reject $H_0$ when it is indeed false. The probability of this is a function of the parameter $\theta$, namely $1 - \beta(\theta)$, and is called the **power** of the test. A graph of the power vs. $\theta$ is called the **power curve** of the test. The function $\beta(\theta)$ is sometimes called the **operating characteristic** of the test [1, 6]. It is simply the probability of "accepting" the null hypothesis, as a function of the (unknown) parameter value.

- **Warning:** The classic bible of hypothesis testing by E. L. Lehmann [9, p. 61] uses $\beta$ to denote the power. Thus his $\beta$ is Larsen and Marx's $1 - \beta$ and vice versa. A brief inquiry shows that my colleague Bob Sherman and [6, 7, 10] agree with Larsen and Marx.

- If we have two tests with the same significance level $\alpha$, if one is always more powerful than the other, then it is a (statistically) better test. (One way to increase the power is to get more data, which may be expensive.)

- The power of a test is influenced by the shape of the critical region. For instance, why do we usually take the critical region to be the regions where the density of $T$ is smallest? Because it makes the test more powerful. This is the essence of the **Neyman–Pearson Lemma**.

- Every test is characterized by its pair $(T, C)$ of test statistic and $T$ and critical region $C$. Let $\beta_{T,C}(\theta)$ denote the test's probability of Type II error (false acceptance of the null) when the parameter is $\theta$,

$$\beta_{T,C}(\theta) = P_\theta\left(T \notin C\right).$$

If a test $(T^*, C^*)$ has size (significance level) $\alpha$, and if for every $\theta \in \Theta_1$,

$$1 - \beta_{T^*,C^*}(\theta) \geqslant \max\{1 - \beta_{T,C}(\theta) : \text{test } (T, C) \text{ has size } \alpha\},$$

then we say that $(T^*, C^*)$ is **uniformly most powerful test (UMP)**.

- UMPs are good to have, but they don't always exist. We shall discuss situations where UMPs do exist, and what they look like, next time.

- When a UMP does not exist, many practitioners argue that a test should be selected to minimize a weighted average of $\alpha$ and $\beta$, where the weights reflect the tester's concerns about the two types of errors.

## 20.7 ⋆ Likelihood Ratio Tests for simple hypotheses

**20.7.1 Example (Likelihood ratio test for the mean of a Normal($\mu, 1$))** Let us return to the simple example of Section 20.2 to make sense out of all this jargon and abstraction. We know that the sample average $\bar{X}$ is either a Normal $N(0, 1/n)$ random variable or a Normal $N(1, 1/n)$ random variable. On the basis of the datum $\bar{x}$ we have to decide which.

The parameter set $\Theta$ for the mean $\mu$ has only two points, $\Theta = \{0, 1\}$. To avoid needless confusion, let me say that the null hypothesis $H_0$, and the alternative $H_1$ are

$$H_0 \colon \mu = 0, \qquad H_1 \colon \mu = 1.$$

Let me also write $f_0(\bar{x})$ for the pdf of $\bar{X}$ under the null hypothesis, and $f_1(\bar{x})$ for the pdf under the alternative.

The same intuition that motivated our maximum likelihood estimation suggest that perhaps we ought to use a test like the following:

Let the test statistic $T(\bar{x})$ just be $\bar{x}$ itself.

Define the **likelihood ratio**

$$\lambda(x) = \frac{f_1(\bar{x})}{f_0(\bar{x})}$$

**Warning:** I am writing the ratio the way that Lehmann [9, p. 64] writes it. Larsen–Marx [8, p. 380] invert the ratio. It's merely a convention, but you have to know which one is being used or your inequalities will be reversed. I took an informal survey of my colleagues, to ask how they write likelihood ratios, and most of them agreed with Lehmann. However, Wikipedia agrees with Larsen and Marx.

For our Normal case, the sample average $\bar{X}$ has variance $1/n$, so the likelihood ratio for $\mu < \mu'$ is

$$\lambda(\bar{x}) = \frac{e^{-(\bar{x}-1)^2/(2/n)}}{e^{-(\bar{x}-0)^2/(2/n)}} = e^{\frac{-1}{2n}(1+2\bar{x})},$$

which is increasing in $\bar{x}$.

A **likelihood ratio test** takes the form:

- Choose a cutoff $k > 0$,
- and reject $H_0$ if $\lambda(\bar{x}) \geqslant k$,
- otherwise fail to reject $H_0$, or accept $H_0$ over $H_1$.

Note that if you invert the likelihood ratio the way that Larsen and Marx do, you want to reject $H_0$ if $\lambda \leqslant 1/k$.

So there is a one-parameter family of likelihood ratio tests parametrized by $k$. It is this freedom that lets you use the Likelihood Principle, but still accommodate cost/benefit considerations.

The likelihood ratio test is equivalent to

- Choose a cutoff $c$,
- and reject $H_0$ if $\bar{x} \geqslant c$,
- otherwise fail to reject $H_0$, or accept $H_1$ over $H_0$.

In order to keep the significance level the same as the sample size varies, we have to scale the cutoff point by $1\sqrt{n}$. The **size** or **significance level** $\alpha$ of the test is

$$P_0\left(\bar{X} \geqslant c/\sqrt{n}\right) = P_0\left(\frac{\bar{X}}{1/\sqrt{n}} \geqslant c\right) = 1 - \Phi(c)$$

since under the null hypothesis, $\bar{X}/(1/\sqrt{n}) \sim N(0,1)$.

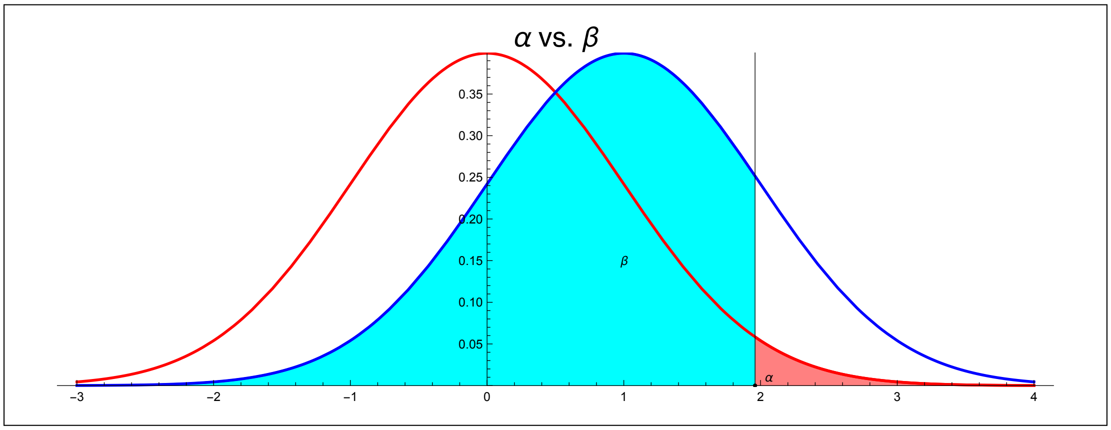
Setting $c = 1.64485$ and using the rule "reject the null if $\bar{x} > c/\sqrt{n}$" yields $\alpha = 0.05$ independent of $n$. (This is a one-sided test.)

The probability of False Acceptance

$$\beta(\mu_1) = P_1\left(\bar{X} < c/\sqrt{n}\right) = P_1\left((\bar{X}-1)/(1/\sqrt{n}) < c - \sqrt{n}\mu\right) = \Phi\bigl(c - \sqrt{n}\mu\bigr)$$

since under the alternative hypothesis $\mu = 1$, $(\bar{X}-1)/\sqrt{n} \sim N(0,1)$. So the power is

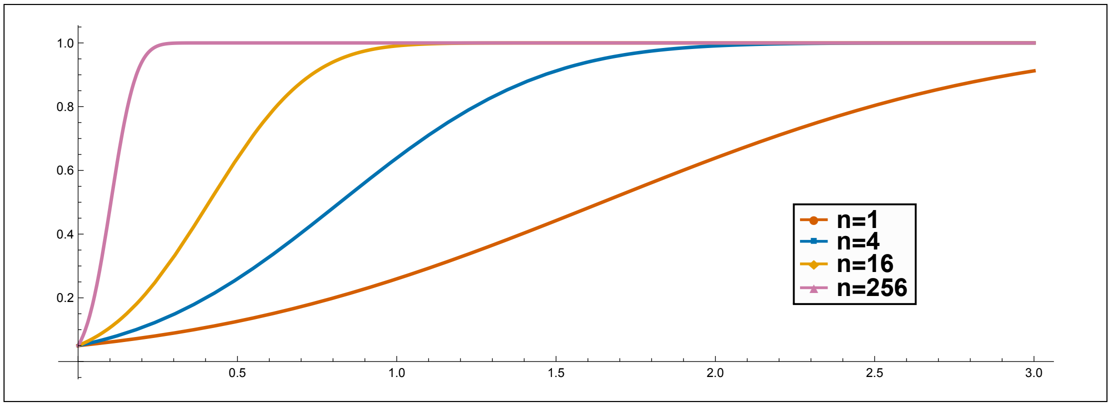
$$1 - \Phi\bigl(c - \sqrt{n}\mu\bigr).$$



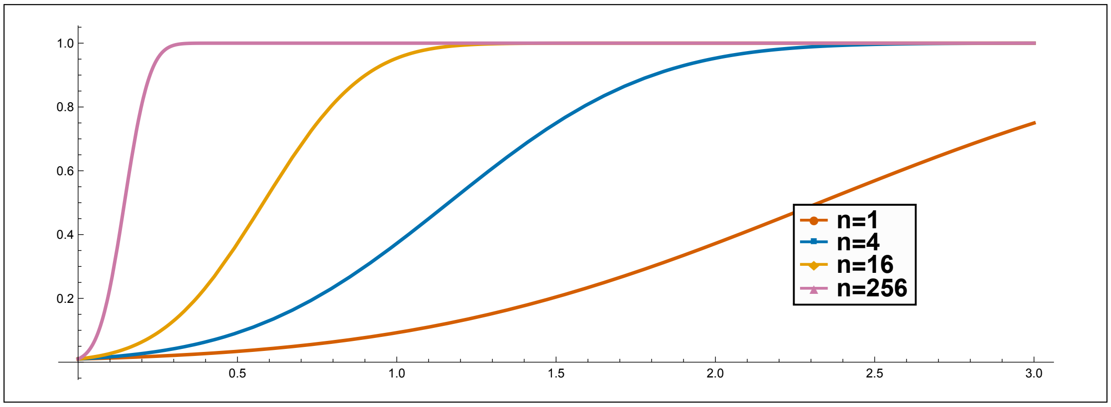
More generally if the alternative hypothesis is $\mu > 0$, the **power function** is given by

$$1 - \beta(\mu) = 1 - \Phi\bigl(c - \sqrt{n}\mu\bigr).$$

Here is the graph of the power function for various $n$.



What if I want a smaller value of $\alpha$? To get $\alpha = 0.01$, I need to set $c = 2.32635$. Here are the new power curves:



□

Can we get more power? No, this is the best we can do. This was first proven by Jerzy Neyman and Egon Pearson [11].[1]

## 20.8 Neyman–Pearson Fundamental Lemma

**20.8.1 Neyman–Pearson Fundamental Lemma** *For testing a simple null versus a simple alternative, a likelihood ratio test maximizes the power, given the size.*

Here is a reasonably convincing proof for the case of absolutely continuous test statistics, but if you want to dot all your *i*s and cross all your *t*s, you need to use critical functions, not critical regions. See Lehmann [9, pp. 65–67] for a more complete proof.

*Proof*: Pick a point $c$ in the critical region, and a point $d$ in the non-critical region and imagine swapping tiny intervals about them:

$$\begin{aligned}\Delta\alpha &\approx -f_0(c)\delta + f_0(d)\varepsilon,\\ \Delta\beta &\approx f_1(c)\delta - f_1(d)\varepsilon,\end{aligned}$$

where $\delta$ and $\varepsilon$ are the widths of the intervals around $c$ and $d$. Then

$$\Delta\alpha = 0 \implies \varepsilon = \frac{f_0(c)}{f_0(d)}\delta.$$

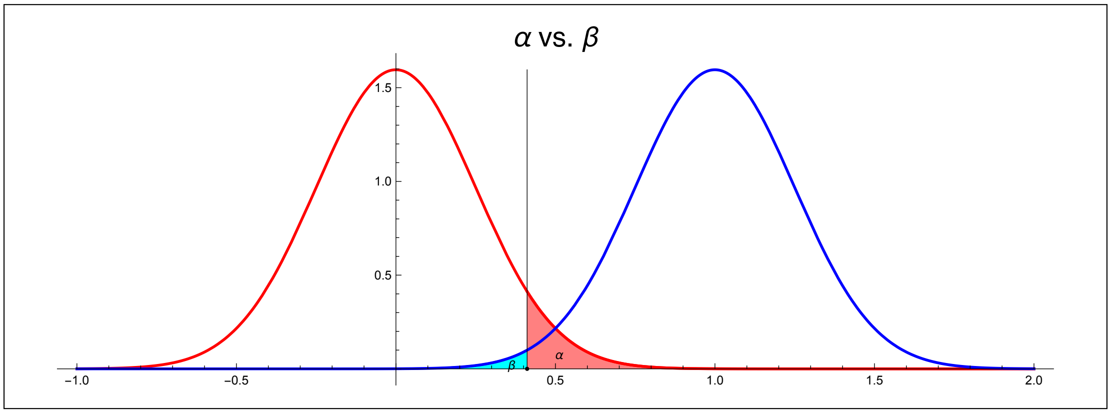
So

$$\begin{aligned}\Delta\beta &= \left[f_1(c) - f_1(d)\frac{f_0(c)}{f_0(d)}\right]\delta\\ &= \Big[\underset{(\geqslant k^*)}{\frac{f_1(c)}{f_0(c)}} - \underset{(\leqslant k^*)}{\frac{f_1(d)}{f_0(d)}}\Big] f_0(c)\delta \geqslant 0.\end{aligned}$$

That is any small change to the critical region that keeps the significance (size) $\alpha$ constant (or reduces it) must increase $\beta$, the probability of Type II error, reducing the power. ∎

So what can we do to get a more powerful test? Increasing the sample size reduces the standard deviation, so the power increases. The next chart shows the effect of reducing the standard deviation by a factor of 4 (increasing the sample size by a factor of 16) for our toy problem. Of course, increasing the sample size can be expensive, especially if your experiment involves medical testing, colliding large hadrons, or mapping the climate on assorted planets.

[1] Egon Pearson is the son of Karl Pearson, who originated the chi-square test, among other things.



## 20.9 ⋆ The monotone likelihood ratio property

For composite hypotheses, likelihood ratio tests work best when the data model $f(x;\theta)$ satisfies an additional property, known as the monotone likelihood ratio property. In this case, the Neyman–Pearson Fundamental Lemma generalizes, and likelihood ratio tests are UMP, they are characterized by critical values, and the notion of a $p$-value applies.

When $\Theta \subset \boldsymbol{R}$, the probability model $f(x;\theta)$ satisfies the **Monotone Likelihood Ratio Property (MLRP)** if there is real valued (one-dimensional) test statistic $T(x)$, such that for every pair $\theta, \theta'$, with

$$\theta < \theta'$$

the likelihood ratio

$$\frac{f(x;\theta')}{f(x;\theta)} \text{ is nondecreasing in } T(x).$$

We then say the likelihood ratio is monotone in $T(x)$.

Another way to rewrite the MLRP is

$$\big(\theta < \theta' \ \&\ T(x) < T(x')\big) \implies \frac{f(x;\theta')}{f(x;\theta)} \leqslant \frac{f(x';\theta')}{f(x';\theta)}.$$

**20.9.1 Example (MLRP and the Normal Family)** Consider the Normal family $N(\mu, 1)$. Here the one-dimensional parameter is $\mu$ and the parameter space is $\Theta = \boldsymbol{R}$. If we a sample $n$ independent observations $x_1, \dots, x_n$, the sample mean $\bar{x}$ is a sufficient statistic for $\mu$ with a Normal distribution with mean $\mu$ and variance $1/n$. Its density is

$$f(\bar{x};\mu) = \frac{1}{\sqrt{2\pi}} e^{-(\bar{x}-\mu)^2/(2/n)}.$$

So for $\mu < \mu'$, the likelihood ratio is

$$\frac{e^{-(\bar{x}-\mu')^2/(2/n)}}{e^{-(\bar{x}-\mu)^2/(2/n)}} = e^{\frac{n}{2}\left((\bar{x}-\mu)^2-(\bar{x}-\mu')^2\right)} = e^{\frac{n}{2}\left(\mu^2-\mu'^2+2(\mu'-\mu)\bar{x}\right)}$$

which is a strictly increasing function of $\bar{x}$. This remains true for $\sigma^2 \neq 1$. Thus the Normal family has the MLRP for $\mu$ with respect to $\bar{x}$ for fixed $\sigma^2$. □

**20.9.2 Example (MLRP and the Poisson distribution)** The Poisson($\mu$) pmf is

$$f(k;\mu) = e^{-\mu}\frac{\mu^k}{k!} \quad (K = 0, 1, 2, \dots).$$

The sample mean $\bar{x} = (k_1 + \cdots + k_n)/n$ of an independent sample of size $n$ has pmf

$$f(\bar{x};\mu) = \frac{1}{k_1!\cdots k_n!} e^{-n\mu}\mu^{n\bar{x}}$$

For $\mu < \mu'$, the likelihood ratio is

$$\lambda(\bar{x}) = \frac{\frac{1}{k_1!\cdots k_n!} e^{-n\mu'}{\mu'}^{n\bar{x}}}{\frac{1}{k_1!\cdots k_n!} e^{-n\mu}\mu^{n\bar{x}}} = e^{-n(\mu'-\mu)} \underbrace{\left(\frac{\mu'}{\mu}\right)}_{(>1)}^{n\bar{x}},$$

which is a strictly increasing function of $\bar{x}$.

□

## 20.10★ UMP Tests for MLRP Densities

When $\theta$ is a one-dimensional parameter, we say that the null hypothesis is **one-sided** if it is of the form

$$H_0\colon \theta \leqslant \bar{\theta} \quad \text{so} \quad H_1\colon \theta > \bar{\theta},$$

(or if we reverse the sense of the inequalities).

When the density has the MLRP for the statistic $T$, then a Uniformly Most Powerful Test exists. The next result may be found in Lehmann [9, Theorem 2, p. 68].

**20.10.1 Theorem** *Let $\Theta$ be one-dimensional, and assume the probability model $f(x;\theta)$ has monotone likelihood ratio in $T(x)$. For testing the null hypothesis $H_0\colon \theta \leqslant \theta_0$ against the alternative $H_1\colon \theta > \theta_0$, there is a UMP test with a critical region of the form $[c^*, \infty)$. That is, there is a critical value $c^*$, so that the test*

$$\textit{rejects } H_0 \textit{ if } T(x) > c^*.$$

*The size of the test, by definition*

$$\alpha = \sup\{P_{\theta_0}(T(X) > c^*) : \theta \leqslant \theta_0\}$$

*is achieved for $\theta = \theta_0$, that is,*

$$P_{\theta_0}(T(X) > c^*) = \alpha.$$

*In addition, the power function (Larsen and Marx's $1 - \beta(\theta)$) is strictly increasing in $\theta$ (up to the point where it becomes 1, and then it is constant).*

*Sketch of proof*: • If we test the simple null $\theta = \theta_0$ against a simple alternative $\theta = \theta'$, where $\theta' > \theta$ the Neyman–Pearson Lemma tells us the most powerful test is a likelihood ratio test.

• Because the likelihood ratio is monotone in $T$, the test takes the form:

$$\text{Reject } \theta = \theta_0 \text{ against the alternative } \theta = \theta' \text{ if } T > c^*$$

for some critical value $c^*$.

• Find $c^*$ to give you the desired level of significance $\alpha$.

• Now observe that because of the MLRP the same test specified by $c^*$ is also a likelihood ratio test of the null $\theta = \theta_0$ against the simple alternative $\theta = \theta''$ for any $\theta'' > \theta_0$, and it also has significance level $\alpha$. So by the Neyman–Pearson Lemma, it is the most powerful such test.

• This means that the test with critical value $c^*$ is Uniformly Most Powerful for testing the simple null $\theta = \theta_0$ against the *composite* hypothesis $\theta > \theta^*$.

- The MLRP also implies the test is UMP for the composite null. For $\theta < \theta_0$, we have $P_\theta\,(T > c^*) \leqslant P_{\theta_0}\,(T > c^*)$.

■

The details are spelled out in [9, p. 69], but you can probably work them out yourself.
In this setting, there is another quantity of interest.

- Given such a test we reject $H_0$ if $T > c^*$, where $c^*$ is chosen so that

$$P_{\theta_0}\,(T > c^*) = \alpha.$$

Suppose the test statistic $T$ has the value $t$. The probability

$$P_{\theta_0}\,(T > t)$$

is called the ***p*-value** of $t$.

An equivalent description of the test is to reject $H_0$ whenever the $p$-value of the statistic is less than $\alpha$.

## 20.11 Likelihood Ratio Tests for composite hypotheses without MLRP

**Larsen–Marx [8]:** Section 6.5

Likelihood ratio tests can also be used with composite hypotheses even in the absence of the MLRP. For testing the Null Hypothesis $H_0\colon \theta \in \Theta_0$ versus the Alternative Hypothesis $H_1 : \theta \in \Theta_1$, let $\hat\theta_0$ be the maximum likelihood estimator of $\theta$ over $\Theta_0$ and let $\hat\theta_1$ be the maximum likelihood estimator of $\theta$ over $\Theta_1$. That is,

$$L(\hat\theta_0(\boldsymbol{x}); \boldsymbol{x}) = \max\{L(\theta; \boldsymbol{x}) : \theta \in \Theta_0\}$$

and

$$L(\hat\theta_1(\boldsymbol{x}); \boldsymbol{x}) = \max\{L(\theta; \boldsymbol{x}) : \theta \in \Theta_1\}$$

**Warning:** Because it is easier to warn you than to change the following section, and make sure that I change it correctly everywhere, in this section I am going to write the likelihood ratio the way Larsen and Marx do, not the way Lehmann does. Then the ratio

$$\lambda(\boldsymbol{x}) = \frac{L(\hat\theta_0(\boldsymbol{x}); \boldsymbol{x})}{L(\hat\theta_1(\boldsymbol{x}); \boldsymbol{x})} = \frac{\max\{L(\theta; \boldsymbol{x}) : \theta \in \Theta_0\}}{\max\{L(\theta; \boldsymbol{x}) : \theta \in \Theta_1\}}$$

may serve as a test of the null hypothesis $H_0\colon \theta \in \Theta_0$ versus the alternative $H_1 : \theta \in \Theta_1$.

Now $\lambda(\boldsymbol{x})$ depends on the sample $\boldsymbol{x}$, and so is a random variable, which L&M call $\Lambda$. According to Larsen–Marx [8, Definition 6.5.1, p. 381] you should choose a critical value $\lambda^*$ so that the null hypothesis is rejected if

$$0 \leqslant \lambda(\boldsymbol{x}) \leqslant \lambda^*.$$

They assert the significance level of such a test is given by the $\alpha$ such that

$$P\left(\Lambda \leqslant \lambda^* \,\middle|\, H_0 \text{ is true}\right) = \alpha. \qquad (\star)$$

**20.11.1 Remark** When $\Theta_0$ consists of just a single just a single parameter value $\theta_0$, I know how to make sense of L&M's statement $(\star)$: Compute the probability using $\theta_0$. When $\Theta_0$ has more than one possible parameter, then $(\star)$ is ambiguous at best, and meaningless at worst. Here is what they mean. For each $\theta$, there is a density $f(\boldsymbol{x}; \theta)$ of $\boldsymbol{x}$. Then for each $\theta$,

$$P_\theta\left(\lambda(\boldsymbol{x}) \leqslant \hat\lambda\right) = \int \mathbf{1}_{\lambda(\boldsymbol{x}) \leqslant \hat\lambda}(\boldsymbol{x}) f(\boldsymbol{x}; \theta)\, d\boldsymbol{x} = \alpha. \qquad (\star\star)$$

defines a critical value $\hat{\lambda}^*(\theta)$ that makes $(\star\star)$ true. The critical value that should be used for the test is most stringent one,

$$\lambda^* = \min\{\hat{\lambda}^*(\theta) : \theta \in \Theta_0\}.$$

That way, for every $\theta \in \Theta_0$ the probability of a Type I error is *no more* than $\alpha$. Remember the role of the null hypothesis is the hypothesis that you want to thoroughly trounce before you are willing to give it up.

Now if Larsen and Marx were Bayesians, $(\star)$ would be less problematic. We could use the posterior density on $\theta$ to average over the $P_\theta$s.

Larsen and Marx called this a **generalized likelihood ratio test**, but others may drop the "generalized." Again, the usefulness of such test is not because we attach magical properties to the likelihood function, but because test constructed in this way usually have desirable properties.

## Bibliography

[1] L. Breiman. 1973. *Statistics: With a view toward applications*. Boston: Houghton Mifflin Co.

[2] B. Efron. 1971. Does an observed sequence of numbers follow a simple rule? (another look at Bode's law). *Journal of the American Statistical Association* 66(335):552–559. http://www.jstor.org/stable/2283522

[3] R. A. Fisher. 1925. *Statistical methods for research workers*. Edinburgh: Oliver and Boyd. http://psychclassics.yorku.ca/Fisher/Methods/

[4] ——— . 1970. *Statistical methods for research workers*, 14th ed. Darien, Conn.: Hafner.

[5] I. J. Good. 1969. A subjective evaluation of Bode's law and an 'objective' test for approximate numerical rationality. *Journal of the American Statistical Association* 64:23–66.

[6] I. Guttman, S. S. Wilks, and J. S. Hunter. 1971. *Introductory engineering statistics*, second ed. New York: John Wiley & Sons.

[7] J. L. Hodges, Jr. and E. L. Lehmann. 2005. *Basic concepts of probability and statistics*, 2d. ed. Number 48 in Classics in Applied Mathematics. Philadelphia: SIAM.

[8] R. J. Larsen and M. L. Marx. 2012. *An introduction to mathematical statistics and its applications*, fifth ed. Boston: Prentice Hall.

[9] E. L. Lehmann. 1959. *Testing statistical hypotheses*. Wiley Series in Probability and Mathematical Statistics. New York: John Wiley and Sons.

[10] I. Miller and M. Miller. 2014. *John E. Freund's mathematical statistics with applications*, 8th. ed. Boston: Pearson.

[11] J. Neyman and E. S. Pearson. 1933. On the problem of the most efficient test of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences* 231:289–337.

[12] S. Selby, ed. 1971. *CRC standard math tables*. Cleveland, Ohio: The Chemical Rubber Company.