STA 114: Statistics

Notes 17. Hypotheses testing

Hypotheses about model parameters

A new soporific drug is tried on $n = 10$ patients with sleep disorder, and the average increase in sleep hours is found to be 2.33 hours (with standard deviation 2 hours). Is the drug effective in increasing sleep hours?

Suppose we model the increase in sleep hours for the 10 patients as $X_i \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu, \sigma^2)$. However, we are no longer interested in a general summary of the model parameters. To ascertain whether the drug is effective, we must contrast the support lent by data to two precise statements about the parameters: "$\mu > 0$" and "$\mu \leq 0$".

Such inferential tasks are referred to as hypotheses testing. This is statistical inference where one of two well specified decisions must be made. Each decision corresponds to a specific hypothesis about the model parameter (μ in our example) and the quantity of interested characterized by it (drug efficacy).

More formally, hypotheses testing about a statistical model $X \sim f(x|\theta)$, $\theta \in \Theta$ is about deciding whether to declare $\theta \in \Theta_0$ or declare $\theta \in \Theta_1$, where Θ_0 and Θ_1 form a partition of the parameter space Θ. This means, $\Theta = \Theta_0 \cup \Theta_1$ and that Θ_0 and Θ_1 are disjoint. The two subsets Θ_0 and Θ_1 represent two contrasting scientific hypotheses about the model parameter (drug is effective or not effective).

Null and alternative hypotheses

For now we will be content to look only at the classical approach to hypotheses testing. A foundational point of this approach is that it treats the two hypotheses asymmetrically. One of the hypotheses is taken to represent the status-quo, the no-change scenario (a drug is not effective, a federal policy has 50% support, annual hurricane counts are steady over time, etc.) and is labelled *the null hypothesis* (denoted H_0, and the corresponding parameters subset is labelled Θ_0). The other hypothesis is called *the alternative hypothesis* (denote H_1 and corresponds to Θ_1), one that provides an alternative to the status-quo (a drug is effective, a federal policy is hated more than it is loved, hurricane counts are increasing with time, etc.).

The classical approach takes the stand that without any data we would accept the null hypothesis and one has to find data with substantial evidence against this hypothesis to reject it and go for the alternative (i..e, the null is innocent until proven guilty).

This stand simplifies the task at hand. One simply needs to check whether there is any support in the data toward any $\theta \in \Theta_0$. If yes, then the null hypothesis stands undefeated. Otherwise, we reject it.

ML testing

Checking whether any $\theta \in \Theta_0$ receives support from data blends nicely with the ML approach toward quantifying data support. Recall that subject to a thresholding fraction $k \in [0, 1]$, for any observation $X = x$, we dichotomize the parameter space Θ into a subsets of well and not-so-well supported theories based on whether $L_x(\theta) \geq k \max_{\theta' \in \Theta} L_x(\theta')$. Therefore, subject to the choice of this k, we can perform a test of our hypotheses by checking whether $L_x(\theta) \geq k \max_{\theta' \in \Theta} L_x(\theta')$ for any $\theta \in \Theta_0$. The obvious and important question here is how to choose this constant? More precisely, what impact does the choice of this constant have on our decision making?

This question is awfully close to one that we asked when talking about interval summaries of parameters and found an answer in the form of confidence coefficient guarantee calculations. In fact, the ML approach to testing described above is same as checking whether or not any $\theta \in \Theta_0$ belongs to the ML interval $A_k(x) = \{\theta \in \Theta : L_x(\theta) \geq kL_x(\hat{\theta}_{\text{MLE}}(x))\}$. Could we simply use guarantee calculations for $A_k(x)$ to fix k? The answer is "yes" – but we still need to look directly at guarantee calculations for a testing procedure itself. That the two guarantee calculations are related is not a surprise. The whole concept of confidence intervals was a spin off of an already mature understanding of hypotheses testing. And a direct look at hypotheses testing allows us to broaden the scope to models where confidence intervals are not easy to obtain directly.

Classical theory: Errors in decision making and frequentist guarantees

As with interval summarization, the classical theory looks at every data analytic task as an application of a statistical procedure to the observed data. Accordingly, from a classical perspective, deciding between $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_1$ given data $X = x$, is simply the application of a decision rule $\delta(x)$ to the observed data x. The decision rule is any map from the sample space S to the binary decision space $\{\textit{reject } H_0, \textit{accept } H_0\}$. Such a decision rule is called a test procedure, or simply a test for the pair H_0, H_1.

Classical theory therefore tries to quantify the performance of a test procedure under various possibilities about the true value of θ. To assess the performance of a test procedure $\delta(x)$, one needs to consider two types of decision errors. First, the true value could belong to Θ_0, and we might declare *reject* H_0 based on $\delta(x)$, committing a *Type I error*. Or, the true value could belong to Θ_1 and we might declare *accept* H_0, committing a *Type II error*. These two errors and also the corresponding non-errors are reported in the following table:

		Truth	
		$\theta \in \Theta_0$	$\theta \in \Theta_1$
$\delta(x)$	*accept* H_0	✓	Type II error
	reject H_0	Type II error	✓

To quantify a test's tendency to make either kind of error, we introduce a new quantity, *the power function* of the test. For any test $\delta(x)$ for testing $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$, its power function is defined as

$$\pi(\theta; \delta) = P_{[X|\theta]}(\delta(X) = \textit{reject } H_0), \quad \theta \in \Theta.$$

Therefore, for any $\theta \in \Theta_0$, $\pi(\theta; \delta)$ denotes the corresponding type I error probability (which might be different for two different values of θ from Θ_0). And for a $\theta \in \Theta_1$, the corresponding type II error probability is given by $1 - \pi(\theta; \delta)$.

It is sometimes useful to represent a test $\delta(x)$ by its associated critical or rejection region defined as the set of all $x \in S$ for which $\delta(x) = \textit{reject } H_0$. If a test $\delta(x)$ has critical region C then,

$$\pi(\theta : \delta) = P_{[X|\theta]}(\delta(X) = \textit{reject } H_0) = P_{[X|\theta]}(X \in C).$$

The size of a test

Recall that H_0 plays the squatter's role in classical testing. Therefore an error of type I must be considered much more serious than an error of type II (declaring an ineffective drug effective is more harmful than failing to declare a drug effective when it actually is effective). For this reason performance measurements of a test is driven primarily by the type I error probabilities. The *size* of a test $\delta(x)$, denoted $\alpha(\delta)$ is defined to be its maximum type I error probability,

$$\alpha(\delta) = \max_{\theta \in \Theta_0} \pi(\theta; \delta).$$

This is the worst possible chance the test has in erroneously rejecting the null hypothesis.

Size calculation for ML tests of a point null

A null hypothesis $H_0 : \theta \in \Theta_0$ is called a point null (or a simple null) if Θ_0 contains a single point θ_0. In this case the size of a test is simply its power at θ_0:

$$\alpha(\delta) = \max_{\theta \in \Theta_0} \pi(\theta; \delta) = \pi(\theta_0; \delta).$$

For point nulls, size calculation is often a simple task. We will see this for ML tests.

An ML test $\delta(x)$ for $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$, based on a threshold $k \in [0, 1]$ can be written as:

$$\delta(x) = \begin{cases} \textit{accept } H_0 & \text{if } \frac{L_x(\theta_0)}{\max_{\theta \in \Theta} L_x(\theta)} \geq k, \quad \text{i..e, if } \theta_0 \in A_k(x) \\ \textit{reject } H_0 & \text{if } \frac{L_x(\theta_0)}{\max_{\theta \in \Theta} L_x(\theta)} < k, \quad \text{i..e, if } \theta_0 \notin A_k(x) \end{cases}$$

where $A_k(x)$ is the ML interval procedure based on k: $A_k(x) = \{\theta : L_x(\theta) \geq k \max_{\theta \in \Theta} L_x(\theta)\}$. Therefore

$$\alpha(\delta) = \pi(\theta_0; \delta) = P_{[X|\theta_0]}(\theta_0 \notin A_k(X)) = 1 - \gamma(\theta_0; A_k),$$

one minus the coverage of A_k at θ_0.

Example (Normal with known variance)**.** Consider $X_1, \cdots, X_n \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu, \sigma^2)$, where σ^2 is fixed. We want to test $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$, where μ_0 is a special value of interest (in the drug example, $\mu_0 = 0$, reflecting no improvement on an average). Because ML intervals for μ are of the form $\bar{x} \mp c\sigma/\sqrt{n}$, an ML test for H_0 is of the form:

$$\textit{reject } H_0 \text{ if and only if } \mu_0 \notin \bar{x} \mp \frac{\sigma}{\sqrt{n}} \iff \frac{|\bar{x} - \mu_0|}{\sigma/\sqrt{n}} > c,$$

with size $1 - \{2\Phi(c) - 1\} = 2\{1 - \Phi(c)\}$. So with $c = 1.96$, the corresponding ML test has size 5%. More generally, with $c = z(\alpha)$ the corresponding ML test has size α.

Example (Models with asymptotically normal MLE)**.** We have seen many instances of a scalar parameter model $X \sim f(x|\theta)$, $\theta \in \Theta$, for which $\hat{\theta}_{\text{MLE}}(x) \mp c/\sqrt{I_x}$ has approximately $2\Phi(c)-1$ coverage at all θ. For any such model, an ML test for $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$ is of the form

$$\textit{reject } H_0 \text{ if and only if } \sqrt{I_x}|\hat{\theta}_{\text{MLE}}(x) - \theta_0| > c$$

with size approximately $2\{1 - \Phi(c)\}$. Again, with $c = z(\alpha)$, the size is approximately α.

Non-ML tests based on point estimates

The performance quantification concept clearly applies to any test procedure, not necessarily only those derived from a likelihood function. For example, consider again the normal model $X_1, \cdots, X_n \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu, \sigma^2)$ where σ^2 is known and we want to test $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$. We could do this based on the median x_{med} – large values of the difference $|x_{\text{med}} - \mu_0|$ shows evidence against H_0. "Large" of course is relative to how spread out x_{med} is likely to be around μ_0 if $\mu = \mu_0$ was the truth. We know that when $X_i \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu, \sigma^2)$, X_{med} is approximately $\mathsf{Normal}(\mu, \frac{\pi\sigma^2}{2n})$, therefore we can construct a test for H_0 against H_1 of the form:

$$\textit{reject } H_0 \text{ if and only if } \frac{|x_{\text{med}} - \mu_0|}{\sigma\sqrt{\pi/2n}} > c$$

with approximate size $2\{1 - \Phi(c)\}$.

Use ML tests if you can

Although we could construct and measure the size of any test procedure for a given model, it is generally wise to use an ML test whenever we can. This is because between an ML test and a non-ML test of the same size, the ML test usually enjoys a smaller type II probability error (more power under the alternative) than its rival. We won't state a very general result to this end, but discuss the classic Neyman-Pearson lemma that started this whole discussion.

Lemma 1 (Neyman-Pearson). *Consider a model $X \sim f(x|\theta)$, $\theta \in \Theta$, where the parameter space contains only two points: $\Theta = \{\theta_0, \theta_1\}$. Any ML test for $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$ is given by:*

$$\textit{reject } H_0 \textit{ if and only if } \frac{L_x(\theta_1)}{L_x(\theta_0)} > b$$

for some $b > 0$. Denote this test by $\delta_b(x)$. Then for any test $\delta(x)$ with size equal or smaller than that of $\delta_b(x)$ [i.e., $\pi(\theta_0; \delta) \leq \pi(\theta_0; \delta_b)$] one must have $\pi(\theta_1; \delta_b) > \pi(\theta_1; \delta)$.