STA 250: Statistics

Notes 3. Hypotheses testing

Book chapters: 9.1, 9.5

1 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 $X_i \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu, \sigma^2)$. Given the evidence in our observed data, we must decide whether to declare "$\mu > 0$" or to declare "$\mu \leq 0$".

Such inferential tasks are referred to as hypotheses testing. This is statistical inference where one has to decide between two competing statements or hypotheses about a model parameter (μ in our example) representing the quantity of interest (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). Consider three more examples.

Example (Opinion poll)**.** Out of $n = 500$ randomly chosen students from a university, $X = 200$ said they were in favor of a recent senate bill. With the model $X \sim \mathsf{Binomial}(n, p)$ where $p \in [0, 1]$ denotes the university wide support percentage, we would like to decide between "$p \leq 0.5$" and "$p > 0.5$".

Example (NSW study)**.** National Supported Work (NSW) was a US federally and privately funded program in the 1970s that aimed to provide work experience to individuals who faced economic and social hardship. Data $X_1, \cdots, X_n$ are available from $n = 185$ individuals on the difference between their annual earning before enrolling in the program and after completion of the program. With the model $X_i \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu, \sigma^2)$, $(\mu, \sigma^2) \in (-\infty, \infty) \times (0, \infty)$ we would like to test "$\mu = 0$" and "$\mu \neq 0$".

Example (TC counts)**.** Annual North Atlantic TC counts $X_1, \cdots, X_n$ from last $n = 100$ years are modeled as $X_t \overset{\text{IND}}{\sim} \mathsf{Poisson}(\alpha\beta^{t-1})$, $\alpha \in (0, \infty)$, $\beta \in (0, \infty)$. To judge whether overall TC activity has been trending over time, we would like to test "$\beta = 1$" and "$\beta \neq 1$".

2 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 bill has less than majority support, a federal program has no effect, 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 bill has more than majority support, a federal program affects earning, hurricane count is trending 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.

3 Test statistic & testing rule

A fundamental concept in classical testing is the construction of a test statistics $T(X)$, a scalar summary of data X to represent evidence against H_0. Given the status-quo nature of of the null hypothesis, one will decide against H_0 only when the test statistic is very large. Operationally, we will need to fix an a priori cut-off $c > 0$ and our testing rule is:

$$\text{for data } X = x \text{ reject } H_0 \text{ if } T(x) > c, \text{ otherwise fail to reject } H_0.$$

The test statistic can be constructed by considering any summary of data X that exhibits one type of behavior when H_0 is true and a different behavior when H_1 is true.

Example (Drug effectiveness)**.** With $X = (X_1, \cdots, X_n)$ denoting the increase in sleep hours for the n patients, our model is $X_i \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu, \sigma^2)$, $\mu \in (-\infty, \infty)$, $\sigma \in (0, \infty)$. Consider the summary $\bar{X} = (X_1 + \cdots + X_n)/n$. Large positive values of $\bar{X}$ are unlikely when $\mu \leq 0$. So we could take $T(X) = \bar{X}$ and use a rule of the type "reject $H_0 : \mu \leq 0$ iff $\bar{x} > c$". However, if σ was large, then large positive values of $\bar{X}$ could still be possible under a $\mu \leq 0$. Indeed, when $X_i \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu, \sigma^2)$, the distribution of $\bar{X} \sim \mathsf{Normal}(\mu, \sigma^2/n)$. So we could instead take $T(X) = \bar{X}/\{\hat{\sigma}(X)/\sqrt{n}\}$ where $\hat{\sigma}(X)$ is another summary of X that represents the magnitude of σ (i.e., given as estimate of σ). One such possible summary is the sample standard deviation s_X, defined by

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{x})^2 = \frac{1}{n-1}\left\{\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2\right\}.$$

So now our test statistic is $T(x) = \frac{\bar{x}}{s_x/\sqrt{n}} = \sqrt{n}\bar{x}/s_x$ and testing rule is to reject $H_0 : \mu \leq 0$ iff $\sqrt{n}\bar{x}/s_x > c$ for some pre-specified $c > 0$.

For our study suppose we consider $c = 1.83$. For our data, $n = 10$, $\bar{x} = 2.33$ and $s_x = 2$ and hence $T(x) = \sqrt{10} \times 2.33/2 = 3.684 > 1.83$. And hence for this test statistic and this cut-off we will reject $H_0 : \mu \leq 0$. We will later justify our choices of both $T(x)$ and c. □

Example (Comparative drug effectiveness)**.** Suppose we knew that an existing drug already provides an extra hour of sleep on an average. In that case we would want to test $H_0 : \mu \leq 1$

against $H_1 : \mu > 1$. In this case we could modify our test statistic to $T(x) = \sqrt{n}(\bar{x} - 1)/s_x$. For observed data $T(x) = \sqrt{10} \times 1.33/2 = 2.103$. So with $c = 1.83$ we still reject H_0 and accept the drug as having more effect than the existing one. We will later justify why we chose to work with the same cut-off as in the previous example. □

Example (NSW study)**.** In the NSW study $X = (X_1, \cdots, X_n)$ denotes the earning differences of n recipients of the training program. We have the same model as in the drug effectiveness study: $X_i \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu, \sigma^2)$, $\mu \in (-\infty, \infty)$, $\sigma \in (0, \infty)$. But since we are testing for $H_0 : \mu = 0$ against $H_1 : \mu \neq 0$, a more reasonable test statistic should be $T(x) = \sqrt{n}|\bar{x}|/s_x$ because a large value of $|\bar{x}|$ of either sign is unlikely under H_0. We will take our cut-off to be $c = 1.97$ and our testing rule is to reject H_0 if $\sqrt{n}|\bar{x}|/s_x > 0.145$. Recorded data has $n = 185$, $\bar{x} = 4253.57$, $s_x = 8926.985$ and so $T(x) = \sqrt{185} \times 4253.57/8926.985 = 6.481 > 1.97$. And so we reject H_0 and conclude that the program had some effects on earning. Again, we will later see the justification of choosing $c = 1.97$.

Example (Opinion poll)**.** With $X = \#\{\text{in favor}\}$ modeled as $X \sim \mathsf{Binomial}(n, p)$, $p \in [0, 1]$, we will expect X/n to be close to $1/2$ if p were $1/2$. Indeed, when $p = 1/2$, the pmf of X/n is approximately $\mathsf{Normal}(1/2, 1/(4n))$. So a large value of

$$T(X) = \frac{X/n - 1/2}{\sqrt{1/(4n)}} = 2\sqrt{n}(X/n - 1/2)$$

is unlikely if p were $1/2$. If p were smaller than $1/2$, then X/n is expected to be even smaller. So we could take the above $T(X)$ as our test statistic and for a pre-defined cut-off take our testing rule to be "reject $H_0 : p \leq 1/2$ if $T(x) > c$". We will work with $c = 1.64$ (justified later). If we observed $X = 200$ among $n = 500$, our test statistic will be $T(x) = 2\sqrt{500} \times (0.4 - 0.5) = -4.472 \leq 1.64$ and hence we will fail to reject $H_0 : p \leq 1/2$.

If instead we were testing $H_0 : p = 1/2$ against $H_1 : p \neq 1/2$, then a reasonable choice is to take $T(x) = 2\sqrt{n}|X/n - 1/2|$, as it is unlikely for X/n to be much away from $1/2$ on either side if H_0 were true. If we used a cut-off $c = 1.96$, then our test rule will give $T(x) = 4.472 > 1.96$ and hence we will reject $H_0 : p = 1/2$.

Example (Annual TC counts)**.** Unlike the above examples, it is not so obvious to think of a test statistic $T(X)$ that quantifies evidence against $H_0 : \beta = 1$. One possibility is to count in $S(X)$ the number of years in which the count was greater than the count in the previous year, i.e.,

$$S(X) = \sum_{t=1}^{n-1} I(X_{t+1} > X_t)$$

where $I(A)$ denotes the indicator function: $I(A) = 1$ if A happens and $I(A) = 0$ otherwise. If β were 1, we will expect $S(X)$ to be close to $(n-1)/2$. It's tempting to approximate the distribution of $S(X)$ under $\beta = 1$ by $\mathsf{Binomial}(n - 1, 1/2)$ and that of $S(X)/(n-1)$ by $\mathsf{Normal}(\frac{1}{2}, \frac{1}{4(n-1)})$. Notice the similarity with the opinion poll example. Analogous to our treatment in the opinion poll, we could take our test statistic as $T(X) = 2\sqrt{n-1}|S(X)/(n-1) - 1/2|$ and reject $H_0 : \beta = 1$ if $T(x) > c$ for a pre-defined c. We will work with $c = 1.96$. For the observed data $n = 100$ and $S(x) = 45$ and so $T(x) = 0.905 < 1.96$. So we fail to reject $H_0 : \beta = 1$.

The above test based on $S(X)$ is "wrong"; we will later see why and discuss a "correct" test rule based on it. We will also see a "better" test statistic and testing rule for this problem. You should also be warned that our model is unrealistic as it ignores the fact that not all TCs were counted 100 year back. We won't do a full analysis here, rather use the data for illustration only. Do not take to heart the conclusions drawn by these analyses.

4 Choosing $T(X)$ and c, prelude to classical theory

From the examples above, two important questions remain unanswered: how do we decide what test statistic $T(X)$ to use and how do we decide on a cut-off c? In some cases, there is a natural choice of test statistic, such as the sample average for the normal data or the sample proportion for the binomial data. But the exact details need more justification. In other cases, as in the TC count study, the choice of $T(X)$ appears an open question. Even when a test statistic $T(X)$ has been chosen, the choice of the cut-off c appears quite arbitrary at the moment, though you may have noticed that the values were not very different (1.83, 1.97, 1.64 and 1.96).

We will resolve these questions to an extent by studying the properties of the associated testing rule "reject H_0 if $T(x) > c$". Properties will be studied by taking the rule to our Stat Lab where we can run a bunch of random experiments to study the performance of the rule.

Take for example, for the drug effectiveness study, where we want to see how the rule "reject $H_0 : \mu \leq 0$ if $T(x) = \sqrt{n}\bar{x}/s_x > 1.83$" performs across different scenarios. To create a "scenario" we will first have to fix values for μ and σ. Say we first pick $\mu = 0$ and $\sigma = 3$. We then generate $n = 10$ random numbers $x^L = (x_1^L, \cdots, x_n^L)$ from $\mathsf{Normal}(0, 3^2)$ (superscript L indicates data generated in the lab) and calculate $T(x^L)$ and note down whether we have $T(x^L) > 1.83$. We then repeat this many many times (all with $\mu = 0$ and $\sigma = 3$) and note down in what proportion of repeats did we end up having $T(x^L) > 1.83$. Since our choice of $\mu = 0$ matches with H_0, the proportion of times we had $T(x^L) > 1.83$ is roughly the probability of incorrectly rejecting H_0. We would hope this number to be small.

We could then replicate the above experiment with other choices of μ and σ. Let's say we took $\mu = 1$ and $\sigma = 3$. So now in each repeat of the experiment we will generate random numbers $x^L = (x_1^L, \cdots, x_n^L)$ from $\mathsf{Normal}(1, 3^2)$ distribution. In this case, the proportion of times we end up having $T(x^L) > 1.83$ is actually the probability of correctly rejecting H_0. So we would hope this number to be large.

You can probably see that if we used a much large cut-off, say $c = 200$ instead of $c = 1.83$, then we will end up with a small fraction in either of the above two experiments. Which would not be a good news because we want a large fraction in the second case. On the other hand, if we had used a very small cut-off, say $c = 0.01$, then we will end up with a large fraction in either experiment, again a bad news for the first scenario.

Classical statistics thrives on identifying a good pair of $(T(X), c)$ for which we will get a small fraction in the first experiment but a large fraction in the second. In fact we want a small fraction in any experiment that we run with a μ matching H_0 (i.e., a non-positive value for μ) and any σ. And we want a large fraction in any experiment with μ matching H_1 (i.e., a positive number) and any σ. Toward this, we will first see how to calculate analytically what this fraction should be for any experiment. And then we will see how a special class of procedures, known as maximum likelihood procedures, will give us a good $(T(X), c)$ pair.