Lecture #13

- Recall our testing set-up.
- Before looking at any data, we make claims: $H_0: \theta \in \Theta_0$
  $H_1: \theta \in \Theta_1$,

where $\Theta_0 \cup \Theta_1 = \Theta$ and $\Theta_0 \cap \Theta_1 = \emptyset$, i.e. $\{\Theta_0, \Theta_1\}$ is a partition of the parameter space. We then set a significance level α, which determines a rejection region $R\ (= R_\alpha)$. This is the set-up.

- Then, we compute a test statistic based on observed data $X_1, \dots, X_n$, $T(X_1, \dots, X_n) = T$. The outcome of the test is determined by T:

$T \in R \Rightarrow$ reject H_0; $T \notin R \Rightarrow$ retain H_0.

ex: Suppose we observe data $X_1, \dots, X_n$ from an unknown r.v. and we are interested in estimating the mean μ. Form hypotheses $H_0: \mu = 0$
$H_1: \mu \neq 0$

and a significance level $\alpha = .05$. Let's say $n = 100$, $\bar{X}_{100} = \frac{1}{100}\sum_{i=1}^{n} X_i$

$= .3$

$\hat{\sigma}^2_{100} = \frac{1}{99}\sum_{i=1}^{n}[X_i - \bar{X}_{100}]^2$

$= 1$

Is this enough to reject the null at the specified significance level? Well, since the sample/empirical mean $\frac{1}{n}\sum_{i=1}^{n} X_i$ is asymptotically normal (CLT) with standard error given by $\hat{\sigma}_n$, we may run a Wald test:

$W = \dfrac{\bar{X}_n - \mu_0}{\hat{se}}$, where (1) $\bar{X}$, $\hat{se}$ are determined from the observed data as $\bar{X} = \bar{X}_{100} = .3$

$$\hat{se} = \frac{\hat{\sigma}_{100}}{\sqrt{100}} = \frac{[illegible]1}{\sqrt{100}} = [illegible].1$$

(2.) μ_0 is from H_0

So, our test statistic to validate or reject H_0 is $W = \dfrac{.3 - 0}{.1}$

$$= 3$$

We then set the rejection region as $R = \{(x_1, \ldots, x_n) \mid [illegible]W > [illegible]z_{\alpha/2}\}$, where the cut-off $z_{\alpha/2}$ is determined based on the standard normal distribution so that $\mathbb{P}(|Z_1| > z_{\alpha/2}) = \alpha$, $Z_1 \sim N(0,1)$. In this case of $\alpha = .05$, this is (famously) $z_{\alpha/2} [illegible]\approx 1.96\ (\approx 2)$. So, we check: $W [illegible] > z_{\alpha/2}$

i.e. $3 [illegible] > 1.96$

$\Rightarrow$ reject H_0.

- This is all very by-the-rules of the scientific method. In particular, we pre-register the hypotheses and significance level.

Remarks: (1.) Clearly if we had chosen a significance level $\alpha > .05$, we would have still rejected H_0

(2.) Making the significance level smaller is more subtle. Indeed, as $\alpha \to 0^+$, we will eventually reach the point where our

evidence is insignificant to reject H_0. When exactly do we cross that precipice? That is the p-value.

- Indeed, in the scientific literature, running a hypothesis test the way we just did would be considered a little blasé. What we really want is not just to reject or retain H_0 at a fixed level, but to develop a quantitative measure of the strength of the observed data for rejecting H_0.

Defn: The p-value for data $X_1, \dots, X_n$ and testing scheme with statistic T is

(1.) (Informal) the smallest significance level at which H_0 is rejected based on statistic $T(X_1, \dots, X_n)$.

(2.) (Formal) $\inf_{\alpha \in (0,1)} \left\{ \alpha \mid T(X_1, \dots, X_n) \in R_\alpha \right\}$, where R_α is the rejection region at level α.

ex: Let's return to our earlier example of running a Wald test, where we computed $W = 3$. What is the p-value for this test? Well, the smallest significance level at which $W > Z_{\alpha/2}$ is when $W = Z_{\alpha/2}$,

So we solve for α in the equation

$$3 = Z_{\alpha/2}$$

(definition of $Z_{\alpha/2}$)

$$\iff \mathbb{P}(|Z| > 3) = \alpha$$

This can be computed using a computer (preferred method in 2021) and with a look-up table (preferred method in 1921): $P(|Z| > 3) \approx .0027$

So, the p-value for this problem is $p = .0027$.

Remark: The interpretation of p-values is context-dependent. They are commonly used in biological and social science empirical work to assert "statistical significance." Common standards include $p = .05 \Rightarrow$ "good evidence against H_0"

more recent $\longrightarrow p = .01 \Rightarrow$ " "

"in the air" $\longrightarrow p = .005 \Rightarrow$ " "

Much research in the social sciences (particularly psychology) has been criticized for "p-hacking" and not being reproducable. See the work of e.g. Ioannidis, Lazar on this topic for state-of-art criticism of p-values.