Lecture #11

- In practice, statistics is used to make inferences of the form "is the observed data strong evidence for/against a particular claim?" Mathematics allows us to formalize the notion of "strong evidence" through the language of probability theory.

- In the parametric setting, we formalize the notion of "claim" via hypotheses. That is, suppose our data is modelled as coming from a parametric class $\{f(x;\theta)\}_{\theta\in\Theta}$. Let $\Theta = \Theta_0 \cup \Theta_{01}$ be a partition (so $\Theta_0 \cap \Theta_{01} = \emptyset$)

of the parameter space. We form null and alternative hypotheses

$H_0: \theta \in \Theta_0$ } null hypothesis is the default

$H_1: \theta \in \Theta_1$ } alternative is defined as "alternative to the null"

ex: Suppose $\{x_i\}_{i=1}^n$ is an iid sample from a Gaussian (known) but with unknown parameters (μ, σ^2). A hypothesis test might state a claim on a particular choice of (μ, σ^2), e.g.

$H_0: (\mu, \sigma^2) = (0,1)$

$H_1: (\mu, \sigma^2) \neq (0,1)$

In this case $\Theta_0 = \{(0,1)\}$, $\Theta_1 = \mathbb{R}\times\mathbb{R}_+ \setminus \{(0,1)\}$. The goal of testing this hypothesis is to use the data observed ($\{x_i\}_{i=1}^n$) to determine the plausibility of $H_0: (\mu, \sigma^2) = (0,1)$.

Remark: In the previous example, μ, σ^2 are the true / population parameters, we will (at best) have access to empirical estimates on μ, σ^2, given by the data, for example the MLE estimates or variants.

So, we use the data $\{x_i\}_{i=1}^n$ in an intelligent way (we will see many methods), and make a decision to retain H_0 or reject H_0 in favor of H_1. In reality, the data does not change the truth or falsity of H_0/H_1. Indeed, the data only impacts our decision making process. This leads to four possible outcomes of a hypothesis test:

Epistemic

Our decision	H_0 True	H_0 False
Retain H_0	Correct	Type II Error
Reject H_0	Type I Error	Correct

We break down the possible mistakes, because Type I & II errors are often interpreted differently. In certain settings, avoiding one can be much more important than avoiding the other.

ex: If we are gathering data to determine "does this person have HIV," one might say Type II (testing person is HIV- but in reality they are positive) is more dangerous than a Type I [redacted]. That is because of the specific costs of running a new test relative to the costs of unanticipated transmission and delay of treatment. Very much a problem-dependent tradeoff.

- While hypothesis tests may look different depending on what H_0 and H_1 look like, they typically take the form: "if some statistic on my observed data falls into a certain region, I reject H_0. Otherwise I retain it."

- Formally, we let $R = \{ x \mid T(x) > c \}$, where:
  - X is data
  - T is a function on the data
  - c is a cut-off

- The basic scheme is:

$$X \in R \rightarrow \text{reject } H_0 \text{, in favor of } H_1$$

(observed)

$$X \notin R \rightarrow \text{retain } H_0$$

Each test has a different T and way of assigning a cutoff value.

<u>ex</u>: Suppose $\{X_i\}_{i=1}^n \overset{iid}{\sim} N(\mu, \sigma^2)$, where σ^2 is known but μ is not. Then we are interested in the hypotheses:

$$H_0: \mu \leq 0$$
$$H_1: \mu > 0$$

So, $\Theta = \{\mu\}_{\mu \in \mathbb{R}}$ and we have partitioned into $\Theta = \Theta_0 \cup \Theta_1$, where $\Theta_0 = \{\mu \leq 0\}$, $\Theta_1 = \{\mu > 0\}$. So, our test will be to reject H_0 if $T(x_1, \dots, x_n) > c$ for some constant c. What could be a good choice of T?

Well, if T is the empirical mean, then we know T is an unbiased and consistent estimator for μ. So, let $T(x_1, \dots, x_n) = \frac{1}{n} \sum_{i=1}^n X_i$.

So, our rejection region will take the form: $R = \{(x_1, \dots, x_n) \mid \frac{1}{n}\sum_{i=1}^{n} x_i \geq c\}$

for some cut-off c.

Q: What role is c playing here? It is essentially related to how we trade-off Type I and II errors. Indeed,
- c large $\Rightarrow$ hard to reject H_0 $\Rightarrow$ increases Type II risk
- c small $\Rightarrow$ easy to reject H_0 $\rightarrow$ increases Type I risk

. This motivates the following definition:

Defn: (1.) The power of a test [redacted] is $\mathbb{P}(\text{reject } H_0 \mid H_1 \text{ is true})$

If we fix a parameter θ, the power function is $\beta(\theta) = \mathbb{P}([\text{redacted}]_{X \in R|\theta})$

(2.) The size of a test is $\alpha = \sup_{\theta \in \Theta_0} \beta(\theta)$.

(3.) A test has level α if it has size $\leq \alpha$.

. So, what does it mean if the level of a test is small? It means that if $\theta \in \Theta_0$ (i.e. a parameter under H_0), then $\sup_{\theta \in \Theta_0} \beta(\theta)$ is small,

i.e. [redacted] $\sup_{\theta \in \Theta_0} \mathbb{P}(\text{rejecting } H_0 \mid \theta \text{ is true parameter})$ is small.

$\Downarrow$ (usually)

[redacted] $\mathbb{P}(\text{rejecting } H_0 \mid H_0 \text{ is true})$.

$\mathbb{P}(\text{Type I})$

- So, if we keep the level of a test low, it guards against committing Type I errors. This is like a knob we can turn when running the test.

Ex: Let's return to our earlier example: $\{X_i\}_{i=1}^n \overset{iid}{\sim} N(\mu, \sigma^2)$ with σ^2 known.

$H_0: \mu \leq 0$

$H_1: \mu > 0$

$T(x_1, \dots, x_n)$ = $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$

Then let us fix c. The test is: reject H_0 if $\bar{X} > c$. How does this work from the standpoints of power and level? Well, recall

$$\beta(\mu) = \mathbb{P}(\text{reject } H_0 \mid \mu \text{ is true parameter})$$

$$= \mathbb{P}(\bar{X} > c \mid \mu \text{ is true mean})$$

$$\overset{(*)}{=} \mathbb{P}\left(\frac{\sqrt{n}[\bar{X}-\mu]}{\sigma} > \frac{\sqrt{n}[c-\mu]}{\sigma} \;\middle|\; \mu \text{ is true mean}\right).$$

Now, if $\{X_i\}_{i=1}^n \overset{iid}{\sim}$ samples from $N(\mu, \sigma^2)$, then $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$, so that

$\frac{\sqrt{n}[\bar{X}-\mu]}{\sigma} \sim Z \sim N(0,1)$. This is what happens when μ is the true parameter. So, $(*) = \mathbb{P}\left(Z > \frac{\sqrt{n}[c-\mu]}{\sigma}\right)$

can be computed exactly since n, c, μ, σ are all known in this calculation and Z is

the standard normal. To summarize, $\beta(\mu) = \mathbb{P}\left(Z_1 > \frac{\sqrt{n}[c-\mu]}{\sigma}\right)$.

Thus, the size of the test is $\text{size} = \sup_{\mu \in \Theta_0} \mathbb{P}\left(Z_1 > \frac{\sqrt{n}[c-\mu]}{\sigma}\right)$

$$= \sup_{\mu \leq 0} \mathbb{P}\left(Z_1 > \frac{\sqrt{n}[c-\mu]}{\sigma}\right)$$

$$= \mathbb{P}\left(Z_1 > \frac{\sqrt{n}\,c}{\sigma}\right)$$

So, how can we choose c to achieve a desired level α (i.e. turn the level knob)? Well, can we solve $\alpha = \mathbb{P}\left(Z_1 > \frac{\sqrt{n}c}{\sigma}\right)$ for c? Sure, if we let F be the cdf of Z_1 and notice

$$\alpha = \mathbb{P}\left(Z_1 > \frac{\sqrt{n}c}{\sigma}\right)$$

$$\Leftrightarrow \alpha = 1 - \mathbb{P}\left(Z_1 < \frac{\sqrt{n}c}{\sigma}\right)$$

$$\Leftrightarrow \alpha = 1 - F\left(\frac{\sqrt{n}c}{\sigma}\right)$$

$$\Leftrightarrow 1-\alpha = F\left(\frac{\sqrt{n}c}{\sigma}\right)$$

Since F is monotonic increasing, F^{-1} exists and we get

$$\frac{\sqrt{n}c}{\sigma} = F^{-1}(1-\alpha)$$

$$\Rightarrow c = \frac{\sigma}{\sqrt{n}} F^{-1}(1-\alpha).$$

To summarize: if we use cut-off $\frac{\sigma}{\sqrt{n}} F^{-1}(1-\alpha)$, our test will have level α, i.e. the $\mathbb{P}(\text{Type I}) \leq \alpha$. This gets stricter as $\alpha \to 0^+$.