STA261 - Module 3

Hypothesis Testing

Rob Zimmerman

University of Toronto

July 16-18, 2024

Initial Hypotheses

- Consider our usual setup: we collect $X_1, X_2, \ldots, X_n \overset{iid}{\sim} f_\theta$ for some unknown $\theta \in \Theta$
- In Module 2, we learned how to produce the "best" point estimators of $\tau(\theta)$
- Now, we turn things around (sort of)
- Before observing $\mathbf{X} = \mathbf{x}$, we already have some conjecture/hypothesis about which specific value (or values) of $\theta \in \Theta$ generate $\mathbf{X}$
- Example 3.1: - Heights of UofT students $\sim N(\mu, \sigma^2)$ where $\mu \in \mathbb{R}$ is unknown. The average height in Canada is 5'6.5". Are UofT students different on average? (ie., is $\mu \neq$ 5'6.5"?)

  - Event of voting for Candidate A in the election $\sim$ Bernoulli(θ) where $\theta \in (0,1)$ is unknown... Is the candidate unpopular? (ie., is $p < 0.5$?)

Questions About Plausibility

- Suppose, for example, we initially suspect that $\theta = \theta_0$
- We find a good point estimator $\hat{\theta}(\mathbf{X})$ for θ, observe $\mathbf{X} = \mathbf{x}$, and produce the estimate $\hat{\theta}(\mathbf{x})$, which turns out to equal, say, $\theta_0 + 3$
- Is this evidence in favor of our initial suspicion, or against it? It depends!
- Is the difference of 3 "significant"? Depends on what we mean by "significant"
- *Hypothesis testing* allows us to formulate this question rigorously (and answer it)

"significantly different"

"significantly lower"

⋮

etc.

The Hypotheses in Hypothesis Testing

- **Null hypothesis significance testing (NHST)** (or **null hypothesis testing** or **statistical hypothesis testing**) is a framework for testing the plausibility of a statistical model based on observed data

- For better or worse, it has become a major component of statistical inference

- *Very* roughly speaking, NHST consists of three basic steps:

  1. Assume some "default" model (or set of models) for $\vec{X}$ and set a threshold $\alpha \in [0,1]$ for plausibility
  2. Observe $\vec{X} = \vec{x}$ and calculate the likelihood of observing such data under the "default" model(s)
  3. If that likelihood falls below α, reject the default model(s) in favor of alternatives

The "Hypothesis" in Hypothesis Testing

- Definition 3.1: A **hypothesis** is a statement about the statistical model that generates the data, which is either true or false.

- The negation of any hypothesis is another hypothesis, so they come in pairs

- Usually, we already have a parametric model $\{f_\theta : \theta \in \Theta\}$ in mind, and our hypotheses relate to the possible value (or values) of the parameter θ itself
(not always the case, as we'll see in Module 4)

- The two hypotheses in this setup can be written generically as $H_0 : \theta \in \Theta_0$ versus $H_A : \theta \in \Theta_0^c$, where $\Theta_0 \subset \Theta$ is some "default" set of parameters

- Example 3.2:

For the UofT heights: $H_0: \mu = 5'6.5''$ $(\Leftrightarrow \Theta_0 = \{5'6.5''\})$
$H_A: \mu \neq 5'6.5''$ $(\Leftrightarrow \Theta_0^c = \mathbb{R} \setminus \{5'6.5''\})$

Event of voting for Candidate A: $H_0: \theta = 0.5$ $(\Leftrightarrow \Theta_0 = \{0.5\})$
$H_A: \theta < 0.5$ $(\Leftrightarrow \Theta_0^c = (0, 0.5))$

$\Theta = \{a, b\}$, $H_0: \theta = a$
$H_A: \theta = b$

$\Theta = \mathbb{R}$, $H_0: \theta \in \{2,4,6\} \cup [10, \infty)$
$H_A: \theta \in (-\infty, 10) \setminus \{2,4,6\}$

Kinds of Hypotheses

- We designate one hypothesis the **null hypothesis** (written H_0) and its negation the **alternative hypothesis** (written H_A or H_1)
- Mathematically speaking, any subjective meanings of the null and alternative hypotheses are irrelevant Only the mathematical statements are relevant for the theory.
- But in a scientific study, the null hypothesis typically represents the "status quo" or the "default" assumption
- The study is being conducted in the first place because we suspect the alternative hypothesis may be true instead

- Typically a scientific study looks for evidence of an "effect" (e.g., the effect of a new drug on a disease, the effect of CO_2 emissions on climate, the effect of getting shot in the ear on a presidential candidate's favourability)

- The "default" assumption is that there's no effect

Simple and Composite Hypotheses

- Example 3.3: We're given a coin which may be biased; we want to assess whether it is or not. If we flip the coin and model the event of H as Bernoulli(θ), $\theta \in (0,1)$, then

  $H_0: p = 1/2 \iff \Theta_0 = \{1/2\}$

  $H_A: p \neq 1/2 \iff \Theta_0^c = (0, 1/2) \cup (1/2, 1)$

- Example 3.4:

  Maybe the number of aces in a deck of cards produced by some company is Poisson(λ).

  $H_0: \lambda = 4$

  (would be a pretty bad company: the variance in the # of aces in a deck would also be λ...)

  $H_A: \lambda < 4$

- Definition 3.2: Suppose a hypothesis H can be written in the form $H : \theta \in \Theta_0$ for some non-empty $\Theta_0 \subset \Theta$. If $\underbrace{|\Theta_0| = 1}$, then H is a **simple hypothesis**. Otherwise, H is a **composite hypothesis**. i.e., $\Theta_0 = \{\theta_0\}$ for some $\theta_0 \in \Theta$

A simple hypothesis completely specifies the data-generating distribution!

The Courtroom Analogy

- Consider a prosecution: the defendent is *innocent until proven guilty*
- But the whole point of the case is that the prosecutor suspects the defendent *is* guilty, and the purpose of the trial is to determine whether the evidence supports that guilt
- The jurors ask themselves: if the defendent really was innocent, how unlikely would this evidence be?
- If the evidence is overwhelmingly unlikely, the defendent is found guilty
- But if there's a *lack* of unlikely evidence, they find the defendent *not guilty*

NOT THE SAME AS INNOCENCE!!! It doesn't mean the defendent is truly innocent, just that there's not enough data to "prove" (beyond a reasonable doubt) guilt

- In NHST, we never "accept H_0"; either we reject H_0 or we fail to reject H_0

"find guilty" "find not guilty"

A Motivating Example

- Example 3.5: Let $X_1, \ldots, X_{100} \overset{iid}{\sim} \mathcal{N}(\theta, 1)$, where $\theta \in \mathbb{R}$. Assess the plausibility that $\theta = 5$ if we observe $\bar{X} = -10$.

Seems unlikely! Heuristics:
- large sample size: the WLLN/SLLN says $\bar{X}_n \approx 5$ under $H_0: \theta = 5$
- our observed $\bar{X}_n$ is <u>many</u> standard deviations away from the mean under $H_0: \theta = 5$
- etc...

If $\theta = 5$, then $\mathbb{P}_5(\bar{X}_{100} = -10) = 0$. Doesn't help!

Instead of just -10, how about all values ≤ -10?

Under H_0, $\mathbb{P}_5(\bar{X}_{100} \leq -10)$

$$= \mathbb{P}_5\left(\frac{\bar{X}_{100} - 5}{\sqrt{1/100}} \leq \frac{-10 - 5}{\sqrt{1/100}}\right)$$

$$= \mathbb{P}(Z \leq -150) \quad \text{where } Z \sim N(0,1)$$

$$= \Phi(-150)$$

$$= 0.0000000\ldots$$

Lower than any reasonable threshold of plausibility! So the observed data provides evidence against H_0

Hypothesis Tests and Rejection Regions

- Definition 3.3: A **hypothesis test** is a rule that specifies for which sample values the decision is made to reject H_0 in favour of H_A.

- Example 3.6:
  Reject if $\bar{x}_n < 2$
  Reject if $x_1 = 2$ or $x_{10} = 4$
  Reject if $x_{(n)} \geq 12$

  "Reject if $\theta > 2$" is not a hypothesis test!

- Definition 3.4: In a hypothesis test, the subset of the sample space for which H_0 will be rejected is called the **rejection region** (or **critical region**), and its complement is called the **acceptance region**.

- Given competing hypotheses H_0 and H_A, a hypothesis test is *characterized* by its rejection region $R \subseteq \mathcal{X}^n$

- In other words, $\mathbb{P}_\theta\left(\text{Reject } H_0\right) = \mathbb{P}_\theta\left(\mathbf{X} \in R\right)$

- Example 3.7:
  $R = \{\vec{x} \in \mathcal{X}^n : \bar{x} < 2\} \Rightarrow \mathbb{P}_\theta(\text{reject } H_0) = \mathbb{P}_\theta(\vec{X} \in R) = \mathbb{P}_\theta(\bar{X}_n < 2)$
  $R = \{\vec{x} \in \mathcal{X}^n : x_1 = 2 \text{ or } x_{10} = 4\} \rightarrow \mathbb{P}_\theta(\text{reject } H_0) = \mathbb{P}_\theta(X_1 = 2 \text{ or } X_{10} = 4)$
  $R = \{\vec{x} \in \mathcal{X}^n : x_{(n)} \geq 12\} \Rightarrow \mathbb{P}_\theta(\text{reject } H_0) = \mathbb{P}_\theta(X_{(n)} \geq 12)$

  all depend on θ!

Poll Time!

$$\mathbb{P}_\theta(\text{fail to reject } H_0) = 1 - \mathbb{P}_\theta(\text{reject } H_0)$$
$$= 1 - \mathbb{P}_\theta(\vec{X} \in R)$$

On Quercus: Module 3 - Poll 1

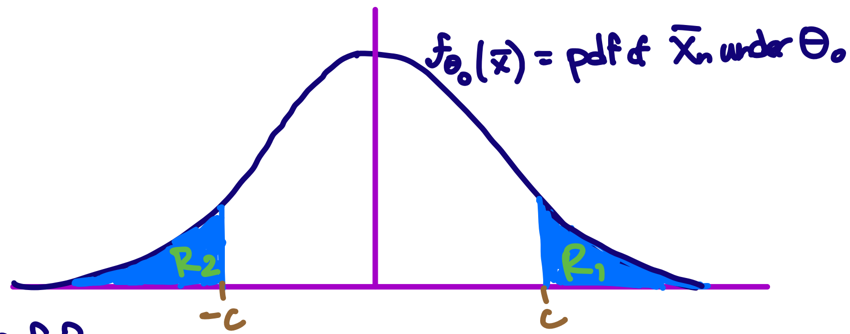
One-Tailed and Two-Tailed Tests

$H_0: \theta = \theta_0$

- If $\Theta \subseteq \mathbb{R}$ and H_0 is simple, then the rejection region is usually in both tails of the distribution:

Eg: $R = \{\vec{x} \in \mathcal{X}^n : |\bar{x}| > c\}$ for some $c > 0$

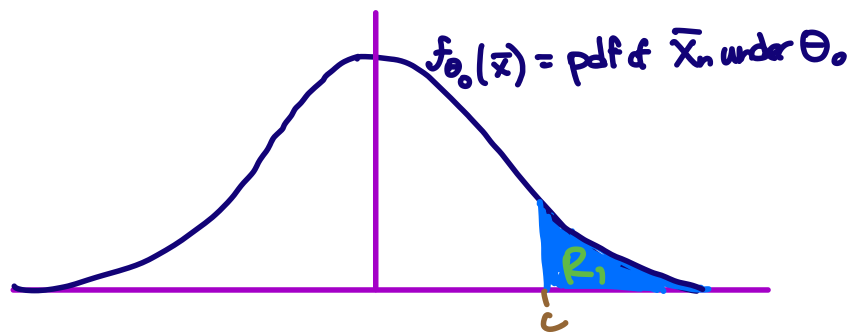
$= \underbrace{\{\vec{x} \in \mathcal{X}^n : \bar{x} > c\}}_{=: R_1} \cup \underbrace{\{\vec{x} \in \mathcal{X}^n : \bar{x} < -c\}}_{=: R_2}$



$\mathbb{P}_{\theta_0}(\vec{X} \in R) = \mathbb{P}_{\theta_0}(\bar{X}_n < -c \text{ or } \bar{X}_n > c) = $ area of R_1 + area of R_2

- But if $H_0 : \theta \leq \theta_0$, then the rejection region is only in one tail:

Eg: $R = \{\vec{x} \in \mathcal{X}^n : \bar{x} > c\}$



$\mathbb{P}_{\theta_0}(\vec{X} \in R) = \mathbb{P}_{\theta_0}(\bar{X}_n > c) = $ area of R_1

- Definition 3.5: Suppose $\Theta \subseteq \mathbb{R}$. A **two-sided test** (or **two-tailed test**) has $H_0 : \theta = \theta_0$, for some $\theta_0 \in \Theta$. A **one-sided test** (or **one-tailed test**) has $H_0 : \theta \leq \theta_0$ or $H_0 : \theta \geq \theta_0$ for some $\theta_0 \in \Theta$.

Type I and Type II Errors

i.e., a "false positive"

- Definition 3.6: A **type I error** is the rejection of H_0 when it is actually true. A **type II error** is the failure to reject H_0 when it is actually false.

  i.e., a "false negative"

- Example 3.8:

$X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, σ^2 known. $H_0: \mu = 0$ vs $H_A: \mu \neq 0$. Our test is (say) $R = \{\vec{x} \in \mathcal{X}^n : \bar{x} < -1\}$.

Suppose we observe $\overline{X}_n = -3$ and hence reject H_0.

If the data actually come from $N(0, \sigma^2)$, we've made a Type I error!

OR: under the same setup, suppose we observe $\overline{X}_n = -0.5$ and hence fail to reject H_0. If the data actually came from $N(-1, \sigma^2)$, we've made a type II error!

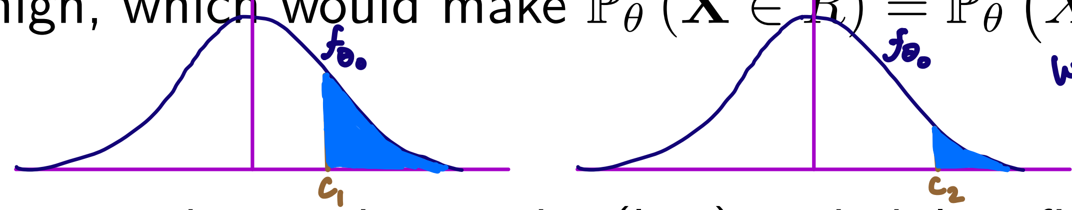
- Of course, we can never *know* if we are committing either of these errors

... because they depend on what the true μ is, which we'll never know!

The Probability of Rejection

- Suppose the rejection region looks like $R = \{\mathbf{x} \in \mathcal{X}^n : \bar{x} \geq c\}$, for some $c \in \mathbb{R}$ (i.e., we reject H_0 when $\bar{X}_n$ is large enough)

- If we demand *very* strong evidence against H_0 before we would reject it, we might set c very high, which would make $\mathbb{P}_\theta(\mathbf{X} \in R) = \mathbb{P}_\theta\left(\bar{X} \geq c\right)$ very small under H_0



we're more demanding with c_2 than with c_1

- In the standard framework, we choose the (low) probability *first*, and then calculate c based on that

- Example 3.9: $X_1, \ldots, X_{100} \overset{iid}{\sim} N(\mu, 1)$. $H_0: \mu \leq 0$ vs $H_A: \mu > 0$. Say our "threshold" is $\alpha = 0.05$.

What c do we need?

$$0.05 = \mathbb{P}_0(\bar{X}_{100} \geq c)$$

$$= \mathbb{P}_0\left(\frac{\bar{X}_{100} - 0}{\sqrt{1/100}} \geq \frac{c - 0}{\sqrt{1/100}}\right)$$

$$= \mathbb{P}(Z \geq 10c) \text{ where } Z \sim N(0,1)$$

$$= 1 - \Phi(10 \cdot c)$$

$$\Rightarrow c = \frac{\Phi^{-1}(0.95)}{10} \approx 0.1645.$$

If instead $n = 10$, we'd get $c \approx 0.5201$.

For "good" tests, a smaller n means we demand more extreme values to reject H_0

The Power Function

- Definition 3.7: The **power function** of a test with rejection region R is the function $\beta : \Theta \to [0,1]$ given by $\beta(\theta) = \mathbb{P}_\theta\left(\mathbf{X} \in R\right)$.

- Observe that

$$\beta(\theta) = \begin{cases} \mathbb{P}_\theta\left(\text{Type I error}\right), & \theta \in \Theta_0 \\ 1 - \mathbb{P}_\theta\left(\text{Type II error}\right), & \theta \in \Theta_0^c \end{cases}$$

- Definition 3.8: Let $\theta \in \Theta_0^c$. The **power** of a test at θ is defined as $\beta(\theta)$.

- ~~Example 3.10:~~ Unfortunately, the power of a test is often written as "$1-\beta$". That β is not the same as our $\beta(\theta)$!!!

The Power Function: Examples

- Example 3.11: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$ with σ^2 known. Suppose a test of has a rejection region of the form $R = \{\mathbf{x} \in \mathcal{X}^n : \bar{x} > c\}$. Calculate the power function of this test.

$$\beta(\mu) = \mathbb{P}_\mu(\vec{X} \in R)$$

$$= \mathbb{P}_\mu(\overline{X}_n > c)$$

$$= \mathbb{P}_\mu\left(\frac{\overline{X}_n - \mu}{\sqrt{\sigma^2/n}} > \frac{c-\mu}{\sqrt{\sigma^2/n}}\right)$$

$$= \mathbb{P}\left(Z > \frac{c-\mu}{\sqrt{\sigma^2/n}}\right) \text{ where } Z \sim N(0,1)$$

$$= 1 - \Phi\left(\frac{c-\mu}{\sqrt{\sigma^2/n}}\right).$$

Note: we didn't need to specify H_0 or H_A here. But $\beta(\mu)$ is only <u>useful</u> when we know which $\mu \in \Theta_0$ and which $\mu \in \Theta_0^c$.

Poll Time!

$$\mathbb{P}_{\theta_0}(\vec{X} \in R)$$

$$= \mathbb{P}_{\theta_0}(\text{reject } H_0)$$

$= \text{probability of rejecting } H_0 \text{ when } H_0 \text{ is } \underline{\text{true}}$

On Quercus: Module 3 - Poll 2

Size and the Probability of Rejection

- If we have a simple null hypothesis and $\mathbf{X}$ is continuous, we can often construct R so that $\mathbb{P}_{\theta_0}(\mathbf{X} \in R) = \alpha$, for some pre-chosen $\alpha \in (0, 1)$
- But for a more general null hypothesis $H_0 : \theta \in \Theta_0$, it's usually impossible to have $\mathbb{P}_{\theta}(\mathbf{X} \in R) = \alpha$ for all $\theta \in \Theta_0$
- Instead, we can try to ask for a "worst-case" probability
- Definition 3.9: The **size** of a test with rejection region R is a number $\alpha \in [0, 1]$ such that $\sup_{\theta\in\Theta_0} \mathbb{P}_{\theta}(\mathbf{X} \in R) = \alpha$.

  Think of this as the "maximum over all possible $\theta \in \Theta_0$"
- Example 3.12:

  $N(\mu, \sigma^2)$, σ^2 known. $H_0: \mu \leq 0$ vs. $H_A: \mu > 0$. $R = \{\vec{x} \in \mathcal{X}^n : \bar{x} > c\}$. How do we choose c to make R a size-α test? We need

$$\alpha = \sup_{\mu \leq 0} \mathbb{P}_\mu(\vec{X} \in R)$$
$$= \sup_{\mu\leq 0}\left(1 - \Phi\left(\frac{c-\mu}{\sqrt{\sigma^2/n}}\right)\right) \text{ from before}$$
$$= 1 - \inf_{\mu \leq 0} \Phi\left(\frac{c-\mu}{\sqrt{\sigma^2/n}}\right)$$
$$= 1 - \Phi\left(\frac{c}{\sqrt{\sigma^2/n}}\right)$$

$\Rightarrow$ Choose $c = \sqrt{\frac{\sigma^2}{n}} \cdot \Phi^{-1}(1-\alpha)$

Significance Levels

- A size-α test might be too much to ask for (especially when the underlying distribution is discrete)

- All we might be able to do is upper bound the worst-case probability

- Definition 3.10: The **level** (or **significance level**) of a test with rejection region R is a number $\alpha \in [0, 1]$ such that $\sup_{\theta \in \Theta_0} \mathbb{P}_\theta (\mathbf{X} \in R) \leq \alpha$.

  i.e., $\sup_{\theta \in \Theta_0} \beta(\theta) \leq \alpha$

Note: some authors use "size" and "level" interchangeably! E/R calls our size "exact size" and our level "size"

- Example 3.13: Let $X \sim \text{Bin}(5, \theta)$, $\theta \in (0,1)$. $H_0: \theta \leq 1/2$ vs $H_A: \theta > 1/2$.

If $R = \{5\}$, then $\sup_{\theta \leq 1/2} \mathbb{P}_\theta(X \in R)$

$$= \sup_{\theta \leq 1/2} \theta^5$$

$$= \left(\frac{1}{2}\right)^5 = 0.03125.$$

So this is a level-0.05 test (and a level-0.04 test...) and a level 0.03125 test.
But it's not a level-0.03 test! Can we ever get a size-α test here? Actually no!
There's no $R \subseteq \{0, 1, \ldots, 5\}$ such that $\sup_{\theta \leq 1/2} \mathbb{P}(X \in R) = 0.05$.

Test Statistics

- A **test statistic** $T(\mathbf{X})$ is a statistic which is used to specify a hypothesis test
- The rejection region specifies which values of $T(\mathbf{X})$ have low probability under H_0
- If $R = \{\mathbf{x} \in \mathcal{X}^n : T(\mathbf{x}) \geq c\}$, then $\mathbb{P}_\theta(\mathbf{X} \in R) = \mathbb{P}_\theta(T(\mathbf{X}) \geq c)$, and evaluating that requires knowing the distribution of $T(\mathbf{X})$
- So a test statistic is only useful if we know its distribution under the null hypothesis
- Example 3.14: - In the $N(\mu, \sigma^2)$ model with σ^2 known, $T(\vec{X}) = \bar{X}_n$ is a good test statistic because under $H_0: \mu = \mu_0$, we know $T(\vec{X}) \sim N(\mu_0, \sigma^2/n)$

- In the Bernoulli(θ) model, $T(\vec{X}) = \sum_{i=1}^n X_i$ is good because under $H_0: \theta = \theta_0$, $T(\vec{X}) \sim Bin(n, \theta_0)$

- In the Poisson(λ) model, $T(\vec{X}) = \frac{X_{(n)}}{X_{(1)}}$ is.... probably not that useful...

p-Values

- Definition 3.11: Suppose that for every $\alpha \in (0,1)$, we have a level-α test with rejection region R_α. For a given sample $\mathbf{X}$, the ***p*-value** is defined as

$$p(\mathbf{X}) = \inf\{\alpha \in (0,1) : \mathbf{X} \in R_\alpha\}.$$

- The idea of a p-value may be the single most misinterpreted concept in statistics

How do we <u>use</u> p-values? We <u>first</u> set $\alpha \in (0,1)$, then we observe $\vec{X} = \vec{x}$, and then we calculate our (observed) p-value $p(\vec{x})$.

- If $p(\vec{x}) < \alpha$, we "reject H_0 at the α-significance level"
- If $p(\vec{x}) \geq \alpha$, we "fail to reject H_0 at the α-significance level"

p-Values Based On Test Statistics

- In non-specialist statistics courses, the p-value for a test with observed data $\mathbf{X} = \mathbf{x}$ is often defined as "the probability of obtaining data at least as extreme as the data observed, given that H_0 is true"

- At first glance, this bears no resemblance to the previous definition; however...

- Theorem 3.1: Suppose a test has rejection region of the form $R = \{\mathbf{x} \in \mathcal{X}^n : T(\mathbf{x}) \geq c\}$, for some test statistic $T : \mathcal{X}^n \to \mathbb{R}$. If we observe $\mathbf{X} = \mathbf{x}$, then our observed p-value is $p(\mathbf{x}) = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta\left(T(\mathbf{X}) \geq T(\mathbf{x})\right)$.

  No proof (it's hard...)

- When H_0 is simple, that becomes $p(\mathbf{x}) = \mathbb{P}_{\theta_0}(T(\mathbf{X}) \geq T(\mathbf{x}))$

- Of course, the theorem also applies when the test specifies that low values of $T(\mathbf{x})$ are to be rejected

  i.e., if $R = \{\vec{x} \in \mathcal{X}^n : T(\vec{x}) \leq c\}$, then $p(\vec{x}) = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(T(\vec{X}) \leq T(\vec{x}))$

Poll Time!

p-values: none of the above!

On Quercus: Module 3 - Poll 3

Famous Examples: The Two-Sided Z-Test

- Example 3.15: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ with $\mu \in \mathbb{R}$ and σ^2 known. Construct a size-α test of $H_0 : \mu = \mu_0$ versus $H_A : \mu \neq \mu_0$ using the **Z-statistic**

i.e., $\Theta_0 = \{\mu_0\}$

$$Z(\mathbf{X}) = \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \sim N(0,1) \text{ under } \mu$$

We want $c > 0$ s.t.

$$\alpha = \sup_{\mu \in \Theta_0} \mathbb{P}_\mu(|Z(\vec{X})| > c)$$

$$= \mathbb{P}_{\mu_0}(|Z(\vec{X})| > c)$$

$$= \mathbb{P}_{\mu_0}(|Z| > c) \quad \text{where } Z \sim N(0,1)$$

$$= 1 - \mathbb{P}_{\mu_0}(-c \leq Z \leq c)$$

$$= 1 - \Phi(c) + [1 - \Phi(c)]$$

$$= 2 - 2\Phi(c)$$

$$\Rightarrow c = \Phi^{-1}(1 - \alpha/2) =: z_{\alpha/2} = \text{"cutoff point"}$$

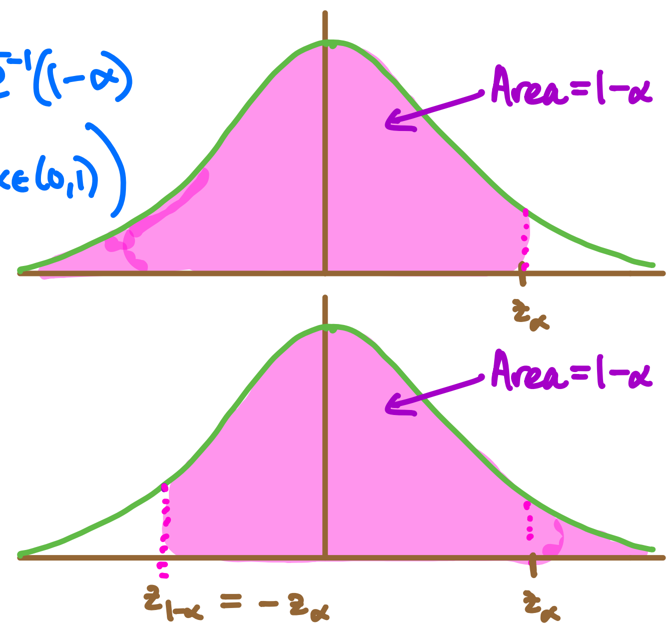
"critical value" for the $N(0,1)$ distribution

$\Rightarrow$ So our rejection region is

$$R = \{\vec{x} \in \mathcal{X}^n : |Z(\vec{x})| > z_{\alpha/2}\}.$$

$z_\alpha := \Phi^{-1}(1-\alpha)$

($z_\alpha = -z_{1-\alpha}$ for any $\alpha \in (0,1)$)



Famous Examples: The One-Sided Z-Test

- Example 3.16: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$ with $\mu \in \mathbb{R}$ and σ^2 known. Construct a size-α test of $H_0 : \mu \leq \mu_0$ versus $H_A : \mu > \mu_0$ using the Z-statistic.

We want some $c > 0$ s.t.

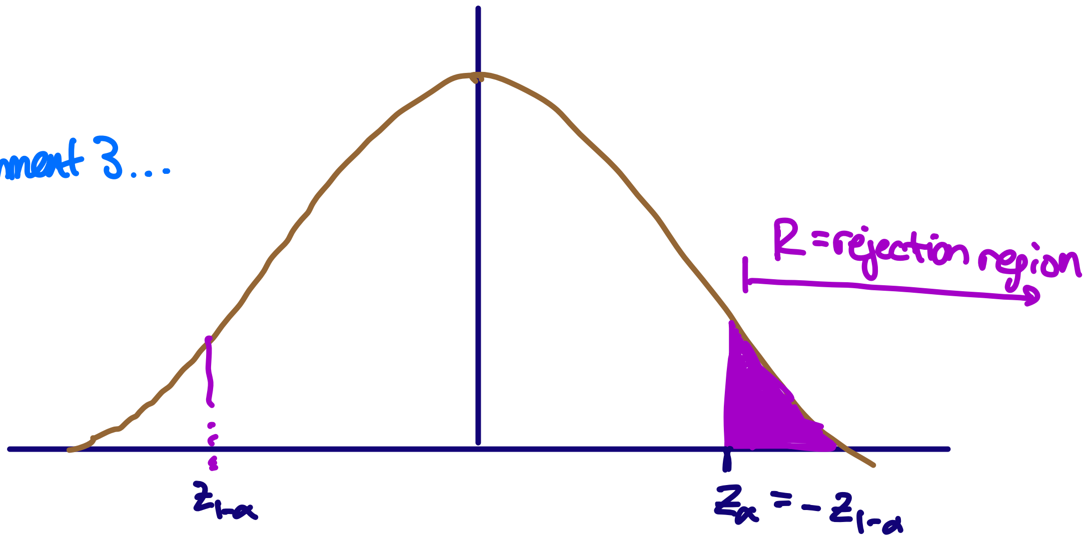
$$\alpha = \sup_{\mu \in \Theta_0} \mathbb{P}_\mu\left(\frac{\overline{X}_n - \mu}{\sqrt{\sigma^2/n}} > c\right)$$

$$= \mathbb{P}_{\mu_0}\left(\frac{\overline{X}_n - \mu_0}{\sqrt{\sigma^2/n}} > c\right)$$

Why μ_0? Assignment 3...

$$= \mathbb{P}(Z > c) \text{ where } Z \sim N(0,1)$$

$$= 1 - \Phi(c)$$



$\Rightarrow c = \Phi^{-1}(1-\alpha) = z_\alpha \; (= -z_{1-\alpha}) \Rightarrow$ Take $R = \{\underset{\sim}{x} \in \mathcal{X}^n : Z(\underset{\sim}{x}) > z_\alpha\}$

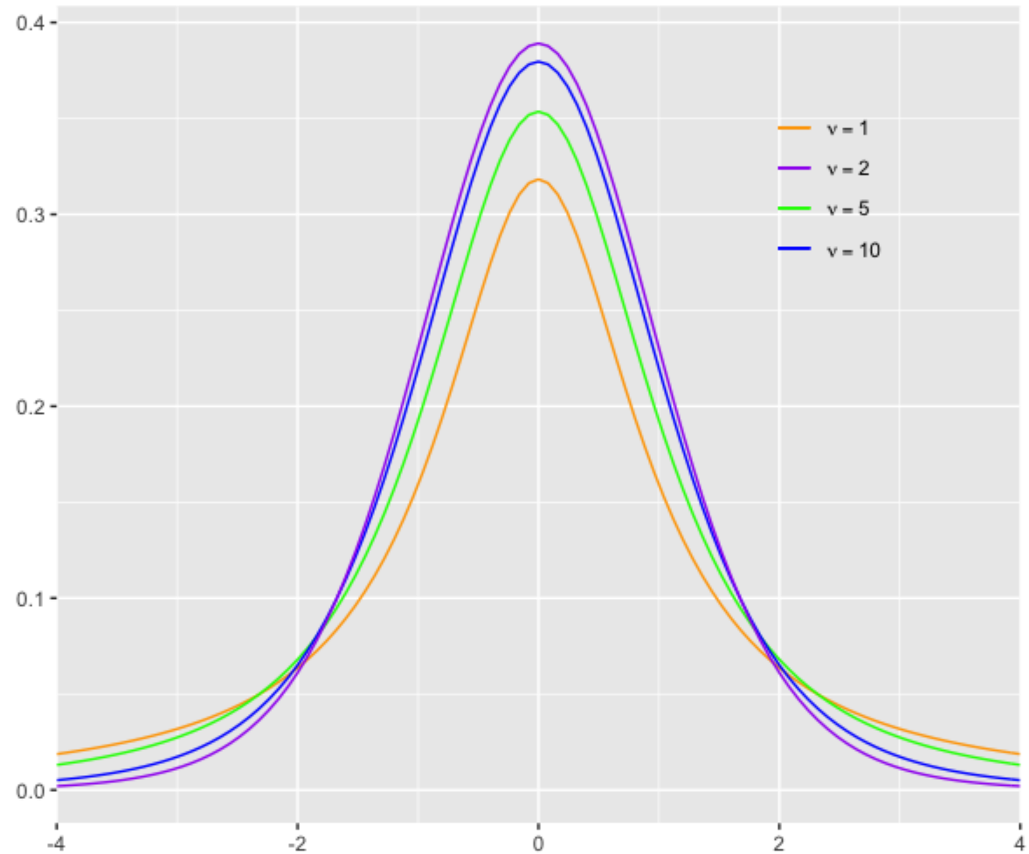
EXERCISE: construct a size-α test of $H_0: \mu \geq \mu_0$ vs $H_A: \mu < \mu_0$.

The t-Distribution

- Definition 3.12: A real-valued random variable T is said to follow a **Student's t-distribution** with $\nu > 0$ degrees of freedom if its pdf is given by

$$f_T(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \quad x \in \mathbb{R}.$$

We write this as $T \sim t_\nu$.



The t-Distribution: Important Properties

- Theorem 3.2: Let $Y, X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(0,1)$. Then

$$T = \frac{Y}{\sqrt{(X_1^2 + \cdots + X_n^2)/n}} \sim t_n.$$

- Equivalently, $T \overset{d}{=} \dfrac{Y}{\sqrt{Q/n}}$ where $Q \sim \chi^2_{(n)}$, $Y \sim N(0,1)$, $Q \perp\!\!\!\perp Y$

- Theorem 3.3: Let $T_n \sim t_n$. Then $T_n \xrightarrow{d} Z$ as $n \to \infty$, where $Z \sim \mathcal{N}(0,1)$.

Proof. By the WLLN, $\frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{P} \mathbb{E}[X_i^2] = 1$

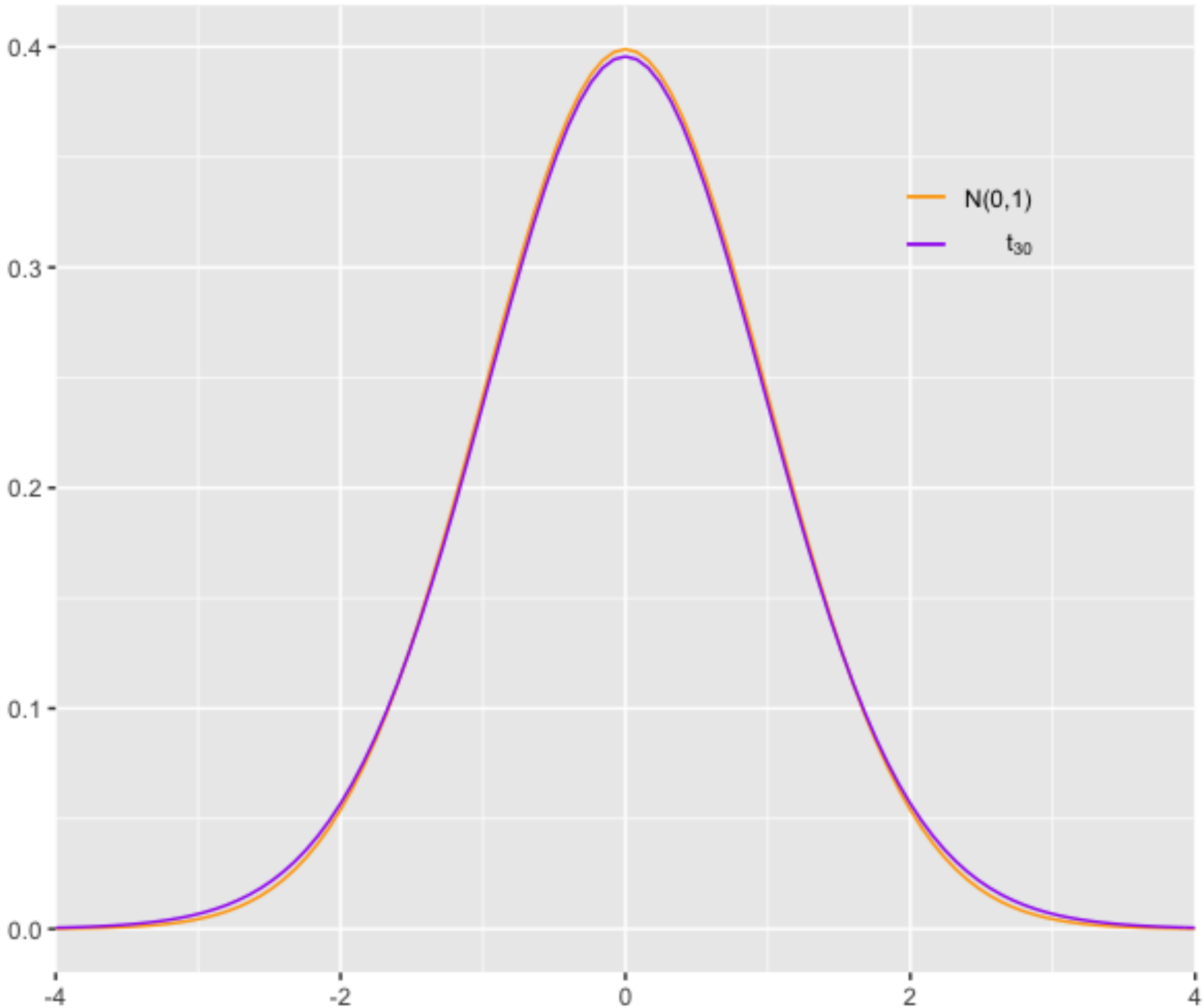
By the CMT, $\sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2} \xrightarrow{P} 1$

Clearly $Y \xrightarrow{P} N(0,1)$

By $\underbrace{\text{Slutsky's theorem}}_{\text{Module 5, if you haven't seen it already}}$, $\dfrac{Y}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2}} \xrightarrow{P} N(0,1)$ $\left(\text{and hence } \dfrac{Y}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i}} \xrightarrow{d} N(0,1) \text{ too.}\right)$

$\square$

A Great Approximation For Even Moderate n



The t-Distribution: More Important Properties

- The t-distribution is mainly used when we have $\mathcal{N}(\mu, \sigma^2)$ data and we're interested in μ, but σ^2 is unknown

- What happens if we swap σ^2 with S_n^2 in the Z-statistic?

- Theorem 3.4: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ with $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. Then

$$\frac{\bar{X}_n - \mu}{\sqrt{S_n^2/n}} \sim t_{n-1}.$$

Proof.

$$\frac{\bar{X}_n - \mu}{\sqrt{S_n^2/n}} = \frac{\left.\frac{\bar{X}_n - \mu}{\sqrt{\sigma^2/n}}\right\} \overset{d}{=} N(0,1)}{\sqrt{\left.\frac{(n-1)S_n^2}{(n-1)\sigma^2}\right\} \overset{d}{=} \frac{1}{n-1} \cdot \chi^2_{(n-1)} \text{ by Theorem 1.7}}}$$

$$\overset{d}{=} \frac{Z}{\sqrt{Q/n-1}}$$ where $Z \sim N(0,1)$, $Q \sim \chi^2_{(n-1)}$ are independent by Module 1 stuff ($\bar{X}_n \perp\!\!\!\perp S_n^2$ for the normal model)

$$\overset{d}{=} t_{n-1} \text{ by Theorem 3.2.} \quad \square$$

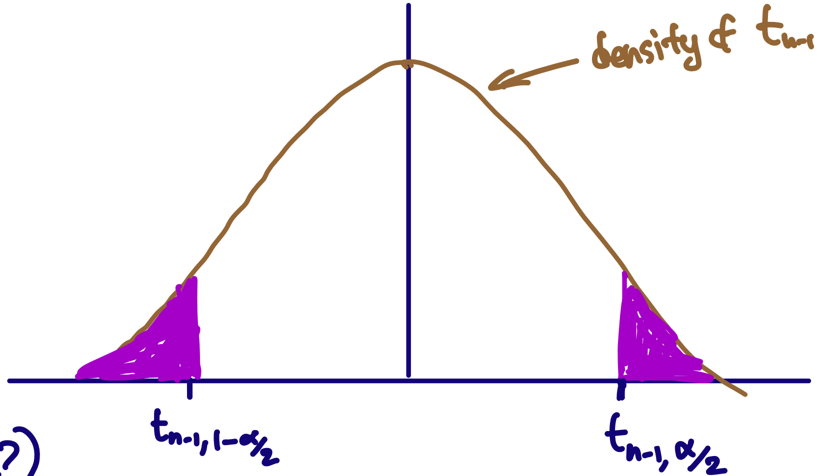
Famous Examples: The Two-Sided t-Test

- Example 3.17: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ with $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. Construct a size-α test of $H_0 : \mu = \mu_0$ versus $H_A : \mu \neq \mu_0$ using the ***t*-statistic**

$$T(\mathbf{X}) = \frac{\bar{X} - \mu}{\sqrt{S^2/n}}. \sim t_{n-1} \text{ under } \mu_0$$

We want to reject H_0 when $|T(\vec{x})| > c$. We need

$$\alpha = \mathbb{P}_{\mu_0}(|T(\vec{X})| > c)$$
$$= 1 - \mathbb{P}_{\mu_0}(-c \leq T(\vec{X}) \leq c)$$
$$= 1 - \mathbb{P}(t_{n-1} \leq c) + \mathbb{P}(t_{n-1} \leq -c)$$
$$= 2 \cdot \mathbb{P}(t_{n-1} > c) \text{ by symmetry (check?)}$$
$$\Rightarrow c = F^{-1}_{t_{n-1}}(1 - \alpha/2) =: t_{n-1, \alpha/2}$$



$t_{n-1,\alpha}$ works the same way as z_α.

Eg: $t_{n-1,\alpha} = -t_{n-1,1-\alpha}$, etc.

So $R = \{\vec{x} \in \mathcal{X}^n : |T(\vec{x})| > t_{n-1, \alpha/2}\}$.

Famous Examples: The One-Sided t-Test

- Example 3.18: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$ with $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. Construct a size-α test of $H_0 : \mu \geq \mu_0$ versus $H_A : \mu < \mu_0$ using the t-statistic.

EXERCISE!

Sample Size Calculations

- Usually, increasing our sample size increases the power of a test
- In real-world studies, obtaining a sample of independent data is typically quite expensive
- Whoever's paying for the study doesn't want experimenters collecting more data than necessary, since that costs money 💸💸💸
- Moreoever, the larger the sample, the higher the chances of problems (errors in data entry, non-independence of some samples, etc.)
- So if we have demands for the power of our test at certain alternative parameters $\theta \in \Theta_0^c$, it's often useful to find the *minimum* sample size n that will give us that power

Sample Size Calculations

- Example 3.19: Suppose $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ where $\mu \in \mathbb{R}$ and σ^2 is known, and we want to test $H_0 : \mu \le \mu_0$ versus $H_A : \mu > \mu_0$ using a test that rejects H_0 when $(\bar{X}_n - \mu_0)/\sqrt{\sigma^2/n} > c$, for some $c \in \mathbb{R}$. How can we choose c and n to obtain a size-0.1 test with a maximum Type II error probability of 0.2 if $\mu \ge \mu_0 + \sigma$?

where $Z \sim N(0,1)$.

Power function: $\beta(\mu) = \mathbb{P}_\mu\left(\frac{\bar{X}_n - \mu_0}{\sqrt{\sigma^2/n}} > c\right) = P\left(Z > c + \frac{\mu_0 - \mu}{\sqrt{\sigma^2/n}}\right) = 1 - \Phi\left(c + \frac{\mu_0 - \mu}{\sqrt{\sigma^2/n}}\right)$

We want $0.1 = \sup_{\mu \le \mu_0}\left(1 - \Phi\left(c + \frac{\mu_0 - \mu}{\sqrt{\sigma^2/n}}\right)\right)$

$= 1 - \Phi(c) \implies c = \Phi^{-1}(0.9) \simeq 1.2816$ (regardless of n)

We also want $1 - \beta(\mu_0 + \sigma) \le 0.2$

$\implies 0.8 \le \beta(\mu_0 + \sigma) = 1 - \Phi(c - \sqrt{n})$

$\implies c \le \Phi^{-1}(0.2) + \sqrt{n}$

$\implies n \ge \left(1.2816 - \Phi^{-1}(0.2)\right)^2 \simeq 4.507$

$\implies$ Choose $n = 5$.

Why plug in $\mu_0 + \sigma$? We want $1 - \beta(\tilde{\mu}) \le 0.2$ for all $\tilde{\mu} \ge \mu_0 + \sigma$. That's the same as $\Phi\left(c + \frac{\mu_0 - \tilde{\mu}}{\sqrt{\sigma^2/n}}\right) \le 0.2$ for all $\tilde{\mu} \ge \mu_0 + \sigma$. Since $\Phi(\cdot)$ is increasing, we should make sure it holds for $\tilde{\mu}$ as small as possible, subject to the constraint $\tilde{\mu} \ge \mu_0 + \sigma$. So the inequality must hold for $\tilde{\mu} = \mu_0 + \sigma$.

The Problems With the p's

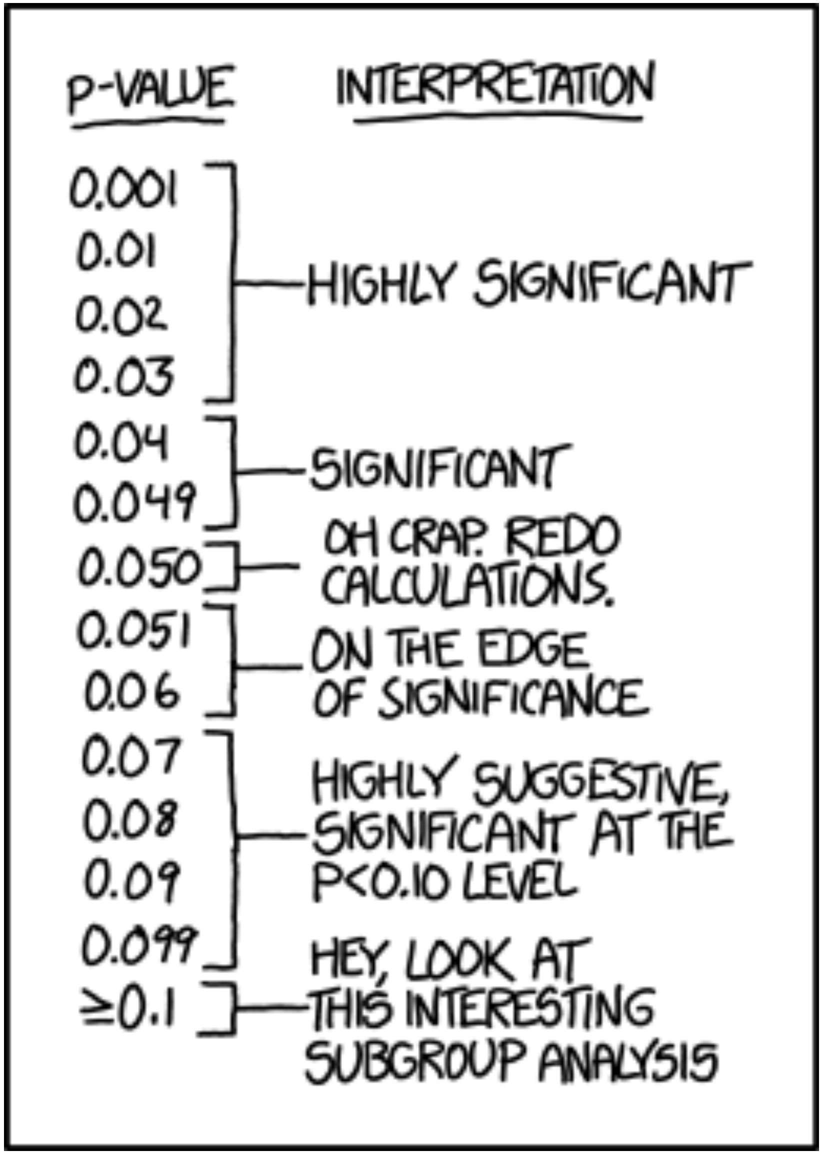
$1-\beta(\tilde{\mu}) \leq 0.2$

$\Phi\left(c+\frac{\mu_0-\tilde{\mu}}{\sigma/\sqrt{n}}\right) \leq 0.2$

- Almost every scientific study that uses statistics will feature p-values somewhere
- The "strength" of a scientific conclusion often wrests upon those p-values
- Ronald Fisher suggested 5% as a reasonable significance level, and it's been widely adopted
- But it's completely arbitrary!
- If every published study used significance levels of 5%, then on average, 1 out of every 20 studies make a type I error
- Think about how many scientific studies are published every day

Thousands! Tens of thousands?

The Problems With the p's



Source: https://xkcd.com/1478/

The Problems With the p's

- p-values lead to publication bias; the $p < 0.05$ threshold is so entrenched that a study result with $p = 0.06$ is considered a "negative" study

- Journals with limited space want to publish new, interesting, "positive" findings

- A study with $p > 0.05$ may contain important new information, but is far less likely to be published

- This pressure leads to ***p*-hacking**: "the misuse of data analysis to find patterns in data that can be presented as statistically significant, thus dramatically increasing and understating the risk of false positives."

Examples of p-Hacking

- Changing α after seeing the data to declare the results statistically significant

Eg: start with $\alpha = 0.05$, observe $\vec{X} = \vec{x}$, calculate $p(\vec{x}) = 0.07$,
declare the results significant at the 0.1-significance level

- Increasing the size of the study population to produce a result that is statistically significant, but not *practically* significant

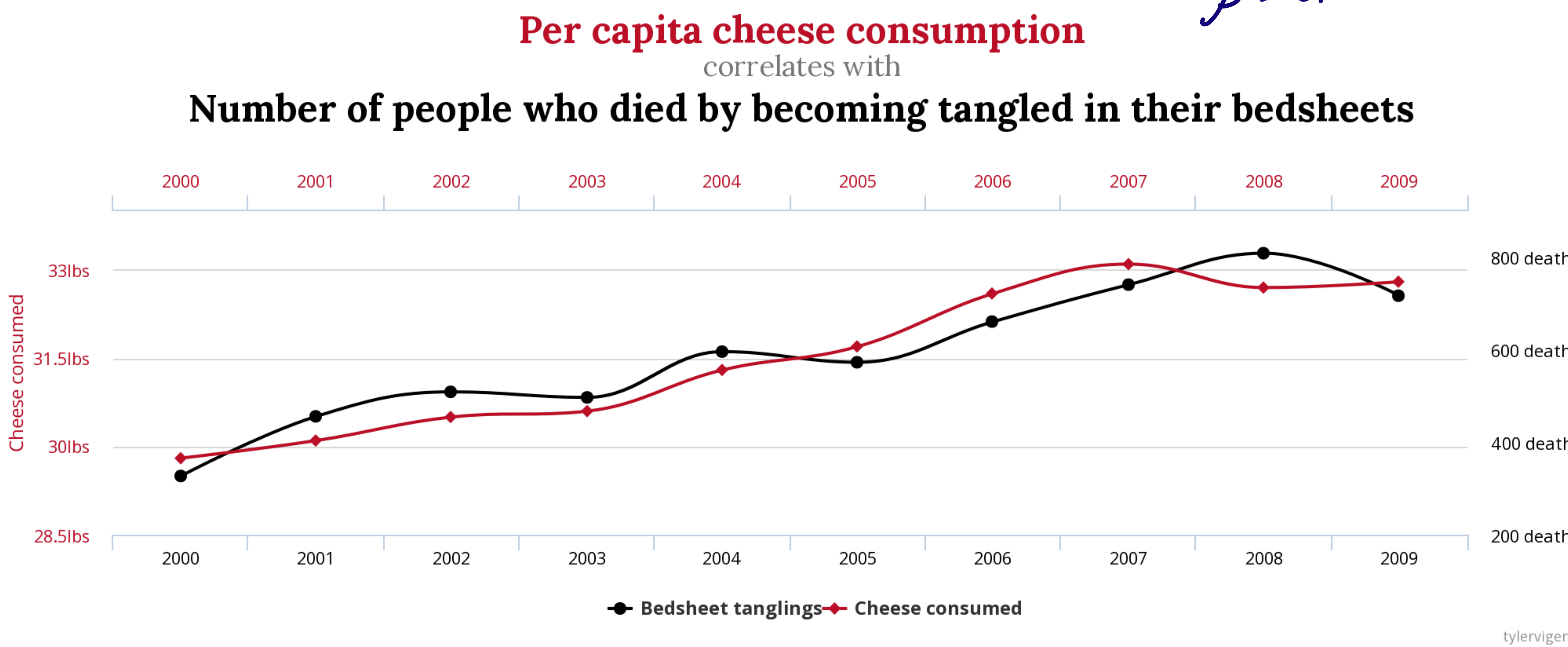
Eg: the time to achieve a normal body temperature was 19.5 hours with Drug A,
versus 19.8 hours with Drug B... a statistically significant difference. But who would wait so long anyway?!

Drug A advertisement: "Expensive new Drug A reduces fever significantly faster than cheap old Drug B!"

- Conducting multiple studies on the same data and "choosing" the one with significant results (this is called the **multiple comparisons problem**)

Should We Be Eating Less Cheese? No!

$\rho \approx 0.96$



Source: https://www.tylervigen.com/

Poll Time!

On Quercus: Module 3 - Poll 4

Examples of p-Hacking

- Post-hoc analyses (i.e., testing hypotheses suggested by a given dataset)

This is basically circular reasoning! It's like observing $X = 12$ and then claiming that $\mathbb{P}(X > 11)$ is very high!

- Outright fraud (such as "editing out" data points that sway the results away from the hoped-for conclusion, or simply lying about the p-value calculation in the hopes that no one will check)

- See also: the Replication Crisis 😬

Bringing Back the Likelihood

- In Module 2, we saw that many common point estimators turned out to be MLEs

- It turns out that many common hypothesis tests are examples of an important kind of test based on the likelihood

- Definition 3.13: The **likelihood ratio test statistic** for testing $H_0 : \theta \in \Theta_0$ versus $H_A : \theta \in \Theta_0^c$ is defined as

$$\lambda(\mathbf{X}) = \frac{\sup_{\theta \in \Theta_0} L(\theta \mid \mathbf{X})}{\sup_{\theta \in \Theta} L(\theta \mid \mathbf{X})}.$$

  A **likelihood ratio test (LRT)** is any test that has a rejection region of the form $R = \{\mathbf{x} \in \mathcal{X}^n : \lambda(\mathbf{x}) \leq c\}$ for some $c \in [0, 1]$.

equivalently, $\log(\lambda(\vec{X})) = \sup_{\theta \in \Theta_0} \ell(\theta \mid \vec{X}) - \sup_{\theta \in \Theta} \ell(\theta \mid \vec{X})$

Poll Time!

$$R = \{\vec{x} \in \mathcal{X}^n : \lambda(\vec{x}) \leq 1\}$$

$$0 \leq \lambda(\vec{x}) = \frac{\sup_{\theta \in \Theta_0} L(\theta | \vec{x})}{\sup_{\theta \in \Theta} L(\theta | \vec{x})} \leq \frac{\sup_{\theta \in \Theta} L(\theta | \vec{x})}{\sup_{\theta \in \Theta} L(\theta | \vec{x})} = 1$$

On Quercus: Module 3 - Poll 5

Choosing $c = 1$ means we reject H_0 when $\lambda(\vec{x}) \leq 1$, which is <u>always</u> true. If we chose $c = 0$, we'd always fail to reject H_0.

So we really care when $c \in (0,1)$.

LRTs: Examples

$X_1, \dots, X_n \sim N(\mu, \sigma^2)$, σ^2 known

$H_0: \mu = \mu_0$
$H_A: \mu \neq \mu_0$

- Example 3.20: Show that the two-sided Z-test is an LRT.

$$\lambda(\vec{x}) = \frac{L(\mu_0 \mid \vec{x})}{L(\bar{x} \mid \vec{x})} = \frac{\exp\left(-\frac{\sum (x_i - \mu_0)^2}{2\sigma^2}\right)}{\exp\left(-\frac{\sum (x_i - \bar{x})^2}{2\sigma^2}\right)} = \exp\left(\frac{-(\bar{x} - \mu_0)^2}{2\sigma^2/n}\right)$$

check!

The LRT rejects when $\lambda(\vec{x}) \leq c$ for some $c \in (0,1)$

$$\iff \exp\left(\frac{-(\bar{x} - \mu_0)^2}{2\sigma^2/n}\right) \leq c$$

$$\iff \frac{|\bar{x} - \mu_0|}{\sqrt{2\sigma^2/n}} \geq \underbrace{\sqrt{-\log(c)}}_{=: c' > 0}$$

observed Z-statistic!

So we reject when $|Z(\vec{x})| > c'$ for some $c' > 0$. That's the two-sided Z-test!

So the two-sided Z-test is indeed an LRT.

LRTs: Examples

- Example 3.21: Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with pdf $f_\theta(x) = e^{-(x-\theta)} \cdot \mathbb{1}_{x \geq \theta}$, where $\theta \in \mathbb{R}$. Determine the LRT for testing $H_0 : \theta \leq \theta_0$ versus $H_A : \theta > \theta_0$.

$L(\theta|\vec{x}) = e^{-\sum x_i + n\theta} \cdot \mathbb{1}_{x_{(1)} > \theta}$.

Unrestricted MLE? $L(\theta|\vec{x})$ is clearly increasing in θ until $\theta = x_{(1)}$, and then equals 0 for $\theta > x_{(1)}$. So $\hat{\theta}_{MLE}(\vec{X}) = X_{(1)}$.

Restricted MLE? Depends on θ_0... If $x_{(1)} \leq \theta_0$, then same as before. If $\theta_0 \leq x_{(1)}$, then we can't go higher than θ_0 anyway, so the MLE over Θ_0 is θ_0.

So $\lambda(\vec{x}) = \begin{cases} 1, & x_{(1)} \leq \theta_0 \\ e^{-n(x_{(1)} - \theta_0)}, & x_{(1)} > \theta_0. \end{cases}$

So $R = \{\vec{x} \in \mathcal{X}^n : e^{-n(x_{(1)} - \theta_0)} \leq c \text{ OR } x_{(1)} \leq \theta_0\}$

$= \{\vec{x} \in \mathcal{X}^n : x_{(1)} \geq \theta_0 - c' \text{ OR } x_{(1)} \leq \theta_0\}$ for some c'

Simple Tests Have Simple LRTs

- Theorem 3.5: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} f_\theta$. Suppose we want to test $H_0 : \theta = \theta_0$ versus $H_A : \theta \neq \theta_0$ using an LRT. Then

$$\lambda(\mathbf{X}) = \frac{L(\theta_0 \mid \mathbf{X})}{L(\hat{\theta} \mid \mathbf{X})},$$

Proof: EXERCISE!

where $\hat{\theta}$ is the (unrestricted) MLE of θ based on $\mathbf{X}$.

- Example 3.22: Suppose $X_1, X_2, \ldots, X_n \overset{iid}{\sim}$ Unif $(0, \theta)$ where $\theta > 0$. Determine the LRT for testing $H_0 : \theta = \theta_0$ versus $H_A : \theta \neq \theta_0$.

$L(\theta_0 | \vec{x}) = \theta_0^{-n} \cdot \mathbb{1}_{x_{(n)} \leq \theta_0} \cdot \mathbb{1}_{x_{(1)} \geq 0}$

$L(x_{(n)} | \vec{x}) = x_{(n)}^{-n} \cdot \mathbb{1}_{x_{(n)} \leq x_{(n)}} \cdot \mathbb{1}_{x_{(1)} \geq 0} = x_{(n)}^{-n} \cdot \mathbb{1}_{x_{(1)} \geq 0}$

EXERCISE: find c that makes this a size-α test!

So $\lambda(\vec{x}) = \dfrac{\theta_0^{-n} \cdot \mathbb{1}_{x_{(n)} \leq \theta_0}}{x_{(n)}^{-n}} \longrightarrow R = \left\{ \vec{x} \in \mathcal{X}^n : \left(\frac{x_{(n)}}{\theta_0}\right)^n \cdot \mathbb{1}_{x_{(n)} \leq \theta_0} \leq c \right\}$

for some $c \in (0,1)$.

LRTs: Examples

- Example 3.23: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim}$ Bernoulli (θ) with $\theta \in (0,1)$. Determine the LRT for testing $H_0 : \theta = \theta_0$ versus $H_A : \theta \neq \theta_0$.

$$L(\theta_0 | \vec{x}) = \theta_0^{\sum x_i} (1-\theta_0)^{n-\sum x_i}$$

$$L(\bar{x} | \vec{x}) = \bar{x}^{\sum x_i} (1-\bar{x})^{n-\sum x_i}$$

$$\text{So } \lambda(\vec{x}) = \left(\frac{\theta_0}{\bar{X}_n}\right)^{\sum x_i} \left(\frac{1-\theta_0}{1-\bar{X}_n}\right)^{n-\sum X_i}$$

So the LRT has rejection region $R = \left\{ \vec{x} \in \mathcal{X}^n : \left(\frac{\theta_0}{\bar{x}}\right)^{\sum x_i} \left(\frac{1-\theta_0}{1-\bar{x}}\right)^{n-\sum x_i} \leq c \right\}$

for some $c \in (0,1)$.

Making Life Easier With Sufficiency

- If $T(\mathbf{X})$ is some sufficient statistic with pdf/pmf $g_\theta(t)$, we might be interested in constructing an LRT based on its likelihood function $L^*(\theta \mid t) = g_\theta(t)$

- But would this change our conclusions?

- Theorem 3.6: Suppose $T(\mathbf{X})$ is sufficient for θ. If $\lambda(\mathbf{x})$ and $\lambda^*(T(\mathbf{x}))$ are the LRT statistics based on $\mathbf{X}$ and $T(\mathbf{X})$, respectively, then $\lambda^*(T(\mathbf{x})) = \lambda(\mathbf{x})$ for every $\mathbf{x} \in \mathcal{X}^n$.

Proof. By the factorization theorem, $f_\theta(\vec{x}) = h(\vec{x}) \cdot g_\theta(T(\vec{x}))$. Therefore,

$$\lambda(\vec{x}) = \frac{\sup_{\theta\in\Theta_0} L(\theta \mid \vec{x})}{\sup_{\theta\in\Theta} L(\theta \mid \vec{x})} = \frac{\sup_{\theta\in\Theta_0} f_\theta(\vec{x})}{\sup_{\theta\in\Theta} f_\theta(\vec{x})} = \frac{\sup_{\theta\in\Theta_0} g_\theta(T(\vec{x}))}{\sup_{\theta\in\Theta} g_\theta(T(\vec{x}))} = \frac{\sup_{\theta\in\Theta_0} L^*(\theta \mid T(\vec{x}))}{\sup_{\theta\in\Theta} L^*(\theta \mid T(\vec{x}))} = \lambda^*(T(\vec{x}))$$

$\square$

Optimal Hypothesis Testing

- We have seen that there can be many tests of two competing hypotheses, with each test characterized by a rejection region

- What makes one test "better" than another?

- A natural idea is to try minimizing the probabilities of type I and type II errors

- Unfortunately, it's usually impossible to get both of these arbitrarily low

You Can't Get the Perfect Power Function

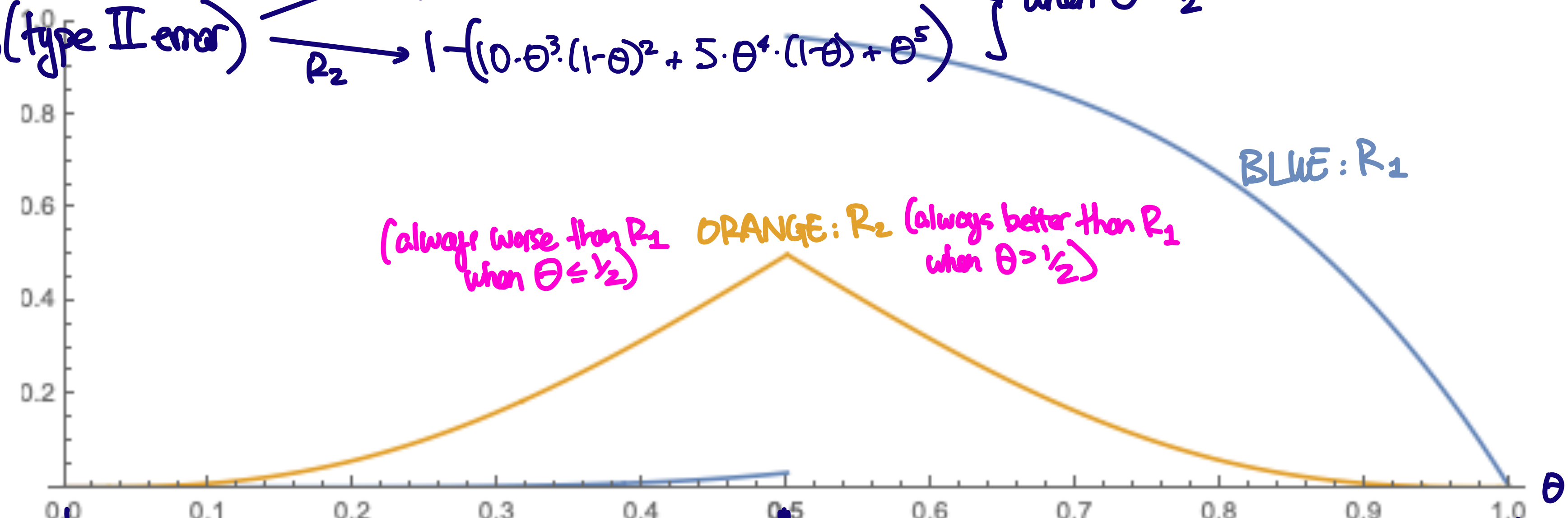
- Let $X \sim \text{Bin}(5, \theta)$, where $\theta \in (0,1)$, and suppose we want to test $H_0 : \theta \leq \frac{1}{2}$ versus $H_A : \theta > \frac{1}{2}$; consider two different tests characterized by the following rejection regions: $R_1 = \{5\}$ and $R_2 = \{3, 4, 5\}$

$\beta_1(\theta) = P_\theta(X=5) = \theta^5$

- $\beta_2(\theta) = P_\theta(X \in \{3,4,5\}) = 10 \cdot \theta^3 \cdot (1-\theta)^2 + 5 \cdot \theta^4 \cdot (1-\theta) + \theta^5$

- $P_\theta(\text{type 1 error}) \xrightarrow{R_1} \theta^5$, $\xrightarrow{R_2} 10 \cdot \theta^3 \cdot (1-\theta)^2 + 5 \cdot \theta^4 \cdot (1-\theta) + \theta^5$ } when $\theta \leq 1/2$

$P_\theta(\text{type II error}) \xrightarrow{R_1} 1 - \theta^5$, $\xrightarrow{R_2} 1 - (10 \cdot \theta^3 \cdot (1-\theta)^2 + 5 \cdot \theta^4 \cdot (1-\theta) + \theta^5)$ } when $\theta > 1/2$



BLUE : R_1

ORANGE : R_2 (always better than R_1 when $\theta > 1/2$)

(always worse than R_1 when $\theta \leq 1/2$)

θ

$y = \beta(\theta) = P_\theta(\text{type I error})$ | $y = 1 - \beta(\theta) = P_\theta(\text{type II error})$

A Compromise

- We have to settle on minimizing either type I error or type II error
- We will settle on the latter; that is, we fix a level α, and among all level-α tests, we try to find the one with the lowest probability of type II error
- This compromise isn't ideal for every real-life situation; sometimes, we care more about minimizing the probability of type I error
- Example 3.24:

- In a medical study: test for a disease which is 100% fatal unless treated. We definitely want to minimize false negatives (i.e., type II errors)
- In a courtroom: a conviction means the death penalty. A type I error means putting an innocent person to death!
- Hypothesis test for a heart disorder: if a patient has the disorder, the only treatment is a heart transplant. If left untreated, there's a 50% chance of death.
  - A type I error means a donor heart is wasted and a patient is (needlessly) on anti-rejection drugs for life
  - A type II error means letting a patient die with probability 1/2

Uniformly Most Powerful Tests

- Definition 3.14: A size-α (or level-α) test for testing $H_0 : \theta \in \Theta_0$ versus $H_A : \theta \in \Theta_0^c$ with power function $\beta(\cdot)$ is called a **uniformly most powerful (UMP) size-α (or level-α) test** if $\beta(\theta) \geq \beta'(\theta)$ for all $\theta \in \Theta_0^c$, where $\beta'(\cdot)$ is the power function of any other size-α (or level-α) test of the same hypotheses.

So regardless of which $\theta \in \Theta_0^c$ generated the data, a UMP size/level-α test will do the right thing (i.e., correctly reject H_0) more often than any other size/level-α test of H_0 vs H_A

(Equivalently: it's a size/level-α test for every simple alternative $H_A: \theta = \theta_A \in \Theta_0^c$)

- UMP tests usually don't exist

- But when they do, how do we actually find them? How do we know that a test is UMP?

The Neyman-Pearson Lemma

- Theorem 3.7 (**Neyman-Pearson Lemma**): Consider testing $H_0 : \theta = \theta_0$ versus $H_A : \theta = \theta_1$. Consider a test whose rejection region R satisfies

$$\mathbf{x} \in R \text{ if } \frac{f_{\theta_1}(\mathbf{x})}{f_{\theta_0}(\mathbf{x})} > c_0 \quad \text{and} \quad \mathbf{x} \in R^c \text{ if } \frac{f_{\theta_1}(\mathbf{x})}{f_{\theta_0}(\mathbf{x})} < c_0$$

  for some $c_0 \geq 0$, and let $\alpha = \mathbb{P}_{\theta_0}(\mathbf{X} \in R)$. Then the test is a UMP level-α test. Moreover, *any* existing UMP level-α test has a rejection region that satisfies the above conditions. No proof...

- Why is the rejection region stated so strangely here? Why not just write $R = \left\{\mathbf{x} \in \mathcal{X}^n : \frac{f_{\theta_1}(\mathbf{x})}{f_{\theta_0}(\mathbf{x})} > c_0\right\}$?

Because of what happens on the "boundary" $\left\{\vec{x} \in \mathcal{X}^n : \frac{f_{\theta_1}(\vec{x})}{f_{\theta_0}(\vec{x})} = c_0\right\}$

We can have different tests that do different things on the "boundary" (not an issue when $\vec{X}$ is continuous, of course)

A Useful Corollary

- Theorem 3.8: Consider testing $H_0 : \theta = \theta_0$ versus $H_A : \theta = \theta_1$. Suppose $T(\mathbf{X}) \sim g_\theta$ is sufficient for θ. Then any test based on $T = T(\mathbf{X})$ with rejection region S is a UMP level-α test if it satisfies

$$t \in S \text{ if } \frac{g_{\theta_1}(t)}{g_{\theta_0}(t)} > k_0 \quad \text{and} \quad t \in S^c \text{ if } \frac{g_{\theta_1}(t)}{g_{\theta_0}(t)} < k_0$$

for some $k_0 \geq 0$, where $\alpha = \mathbb{P}_{\theta_0}(T(\mathbf{X}) \in S)$.

Proof: EXERCISE (hint: factorization theorem!)

The Neyman-Pearson Lemma: Examples

- Example 3.25: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ with $\mu \in \{\mu_0, \mu_1\}$ and σ^2 known. Find a UMP level-α test of $H_0 : \mu = \mu_0$ versus $H_A : \mu = \mu_1$, where $\mu_1 < \mu_0$.

Let's use $T(\vec{X}) = \bar{X}_n$, which is sufficient for μ.

We reject H_0 when $k_0 < \frac{g_{\mu_1}(\bar{x})}{g_{\mu_0}(\bar{x})} = \frac{\exp\left(\frac{-(\bar{x}-\mu_1)^2}{2\sigma^2/n}\right)}{\exp\left(\frac{-(\bar{x}-\mu_0)^2}{2\sigma^2/n}\right)} = \cdots = \exp\left(\frac{1}{2\sigma^2/n}\left[\mu_0^2 - \mu_1^2 + 2\bar{x}(\mu_1 - \mu_0)\right]\right)$

$$\Rightarrow \log(k_0) < \frac{1}{2\sigma^2/n}\left[\mu_0^2 - \mu_1^2 + 2\bar{x}\underbrace{(\mu_1 - \mu_0)}_{<0 \text{ by assumption}}\right]$$

$$R = \{\vec{x} \in \mathcal{X}^n : \bar{x} < c\}$$

$$\Rightarrow \frac{\frac{2\sigma^2}{n} \cdot \log(k_0) - (\mu_0^2 - \mu_1^2)}{2(\mu_1 - \mu_0)} > \bar{x}$$

So we reject when $\bar{X}_n < c$ for some c

$\Leftrightarrow$ reject when $\frac{\bar{X}_n - \mu_0}{\sqrt{\sigma^2/n}} < c'$ for some c'

$\rightarrow$ That's a one-sided z-test! By Theorem 3.8, the one-sided z-test for $H_0: \theta = \theta_0$ vs $H_A: \theta = \theta_1$ is a UMP test, where $\alpha = \mathbb{P}_{\mu_0}(\bar{X}_n < c)$

Making Neyman-Pearson Useful

- There's one thing that keeps the Neyman-Pearson lemma from being useful in practice

- In real life, almost no one needs to test two simple hypotheses!

- On the other hand, one-sided tests are used in abundance

- Luckily, there's a way extend Neyman-Pearson that makes plenty of one-sided tests into UMP level-α tests

- We'll just look at a special case of this, which works when we have a sufficient statistic in an exponential family

The Karlin-Rubin Theorem

- Theorem 3.9 (**Karlin-Rubin**): Consider testing $H_0 : \theta \leq \theta_0$ versus $H_A : \theta > \theta_0$. Suppose $T = T(\mathbf{X}) \sim g_\theta$ is an $\mathbb{R}$-valued sufficient statistic for θ such that $g_{\theta_2}(t)/g_{\theta_1}(t)$ is monotone non-decreasing in t whenever $\theta_2 \geq \theta_1$. Then a test with rejection region $R = \{T > c_0\}$ is a UMP level-α test, where $\alpha = \mathbb{P}_{\theta_0}(T > c_0)$.

  $= \{\vec{x} \in \mathcal{X}^n : T(\vec{x}) > c_0\}$

  $= \{t \in \mathcal{T} : t > c_0\} = (c_0, \infty)$

  No proof...

  i.e., T has a "monotone likelihood ratio" (MLR).

  Discussion?

- By suitably restricting the entire parameter space, this also holds for a test of the form $H_0 : \theta = \theta_0$ versus $H_A : \theta > \theta_0$

- The analogous result holds when we want to test $H_0 : \theta \geq \theta_0$ versus $H_A : \theta < \theta_0$; then $g_{\theta_2}(t)/g_{\theta_1}(t)$ must be monotone non-increasing in t and the rejection region looks like $R = \{T < c_0\}$ EXERCISE! Prove using Karlin-Rubin...

The Neyman-Pearson Lemma: Examples

$H_0: \mu \le \mu_0$ vs. $H_A: \mu > \mu_0$

- Example 3.26: Show that the one-sided Z-test is a UMP level-α test.

$T(\vec{x}) = \bar{X}_n$ is sufficient for μ, with pdf $g_\mu(t) = (2\pi\sigma^2/n)^{-\frac{1}{2}} \cdot \exp\left(\frac{-(t-\mu)^2}{2\sigma^2/n} \right)$

$(2\pi\sigma^2/n)^{-\frac{1}{2}} \cdot \exp\left(\frac{-(t^2 - 2\mu t + \mu^2)}{2\sigma^2/n} \right)$

Let $\mu_2 \ge \mu_1$. Then

$$\frac{g_{\mu_2}(t)}{g_{\mu_1}(t)} = \frac{\exp\left(\frac{-(t^2 - 2\mu_2 t + \mu_2^2)}{2\sigma^2/n} \right)}{\exp\left(\frac{-(t^2 - 2\mu_1 t + \mu_1^2)}{2\sigma^2/n} \right)} = \exp\left(\frac{1}{2\sigma^2/n} \left(2t \cdot \underbrace{(\mu_2 - \mu_1)}_{\ge 0} - \underbrace{(\mu_2^2 - \mu_1^2)}_{\ge 0} \right) \right)$$

is monotone non-decreasing in t.

By Karlin-Rubin, the test with rejection region $R = \{ \vec{x} \in \mathcal{X}^n : \bar{X}_n > c_0 \}$ is a level-α test, where $\alpha = \mathbb{P}_{\mu_0}(\bar{X}_n > c_0)$.

That is, indeed, a one-sided Z-test!

The Neyman-Pearson Lemma: Examples (filled in after lecture)

- Example 3.27: Let $X_1, X_2, \dots, X_n \overset{iid}{\sim}$ Poisson (λ), where $\lambda > 0$. Explain how to produce a UMP level-α LRT for testing $H_0 : \lambda = \lambda_0$ versus $H_A : \lambda > \lambda_0$.

We know $T(\vec{x}) = \sum_{i=1}^{n} x_i$ is sufficient for λ. $T \sim \text{Poisson}(n\lambda)$ with pmf

$f_\lambda(t) = \dfrac{(n\lambda)^t e^{-n\lambda}}{t!}$. Let $\lambda_2 \geq \lambda_1$. Then

$$\frac{f_{\lambda_2}(t)}{f_{\lambda_1}(t)} = \frac{(n\lambda_2)^t e^{-n\lambda_2}}{(n\lambda_1)^t e^{-n\lambda_1}} = \underbrace{\left(\frac{\lambda_2}{\lambda_1}\right)}_{\geq 1}^{t} e^{n(\lambda_1 - \lambda_2)}$$ is increasing in t.

By Karlin-Rubin, a test with rejection region $R = \{\vec{x} \in \mathcal{X}^n : \sum x_i > c_0\}$ is a UMP level-α test, where $\alpha = \mathbb{P}_{\lambda_0}\left(\sum_{i=1}^{n} X_i > c_0\right)$.

How do we actually find c_0? Or $\lceil c_0 \rceil$, since $\sum X_i$ is an integer? By definition of "level", we must have $\alpha \geq 1 - \mathbb{P}_{\lambda_0}\left(\sum_{i=1}^{n} \leq c_0\right) \geq 1 - \sum_{j=0}^{\lceil c_0 \rceil} \underbrace{\frac{(n\lambda_0)^j e^{-n\lambda_0}}{j!}}_{=: p_j}$. So keep subtracting $p_0, p_1, p_2, \dots$ from 1 until the result is $\leq \alpha$.

UMP Tests: Nonexistence

- Sadly, UMP tests usually don't always exist for a given pair of complementary hypotheses (especially for two-sided tests)

- Example 3.28: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ with $\mu \in \mathbb{R}$ and σ^2 known. Show there exists no UMP level-α test for $H_0 : \mu = \mu_0$ versus $H_A : \mu \neq \mu_0$.

Let $\mu_1 < \mu_0 < \mu_2$. Consider 2 tests: Test 1 rejects H_0 if $\frac{\bar{X}_n - \mu_0}{\sqrt{\sigma^2/n}} < -z_{1-\alpha}$ (which is a UMP level-α test of $H_A' : \mu = \mu_1$ by K-R)

Test 2 rejects H_0 if $\frac{\bar{X}_n - \mu_0}{\sqrt{\sigma^2/n}} > z_{1-\alpha}$ (which is a UMP level-α test of $H_A'' : \mu = \mu_2$ by K-R)

We know that Test 1 has highest power at $\mu = \mu_1$ (out of all level-α tests). So if a UMP level-α test does exist for $H_A : \mu \neq \mu_0$, it must be Test 1. However...

$$\beta_2(\mu_2) = \mathbb{P}_{\mu_2}\left(\frac{\bar{X}_n - \mu_2}{\sqrt{\sigma^2/n}} > z_{1-\alpha} + \frac{\mu_0 - \mu_2}{\sqrt{\sigma^2/n}}\right)$$

$$= \mathbb{P}\left(Z > z_{1-\alpha} + \frac{\mu_0 - \mu_2}{\sqrt{\sigma^2/n}}\right) \text{ where } Z \sim N(0,1)$$

$$> \mathbb{P}(Z > z_{1-\alpha})$$

$$= \mathbb{P}(Z < -z_{1-\alpha})$$

$$> \mathbb{P}_{\mu_2}\left(\frac{\bar{X}_n - \mu_2}{\sqrt{\sigma^2/n}} < -z_{1-\alpha} + \frac{\mu_0 - \mu_2}{\sqrt{\sigma^2/n}}\right)$$

$$= \mathbb{P}_{\mu_2}\left(\frac{\bar{X}_n - \mu_0}{\sqrt{\sigma^2/n}} < -z_{1-\alpha}\right)$$

$= \beta_1(\mu_2)$. So Test 2 has strictly higher power at μ_2 than Test 1 does. Contradiction!

$\therefore$ No UMP level-α test exists here.