

STA261 - Module 3

Hypothesis Testing

Rob Zimmerman

University of Toronto

July 19-21, 2022

Initial Hypotheses

- Consider our usual setup: we collect $X_1, X_2, \dots, X_n \stackrel{iid}{\sim} f_\theta$ 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 $\stackrel{iid}{\sim} N(\mu, 1)$, $\mu \in \mathbb{R}$ unknown
Average height in Canada is 5'6.5". Maybe an eye test suggests that $\mu \neq 5'6.5"$
 - Event of voting for Candidate A $\sim \text{Bernoulli}(p)$, $p \in (0, 1)$. Suspect (maybe) that $p < 0.5$.

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" statistical model for \hat{X} (representing "nothing interesting") and set a threshold $\alpha \in [0, 1]$ for plausibility
- ② Observe $\hat{X} = \hat{x}$, and calculate the likelihood of observing $\hat{X} = \hat{x}$ under the "default" model
- ③ If that likelihood falls below α , we reject the "default" model in favour of its complement

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 U&T heights:
 $H_0: \mu = 5'0.5''$ Candidate preference: $H_0: p = 0.5$
 $H_A: \mu > 5'0.5''$ $H_A: p \neq 0.5$
- More exotic: $\Theta = \{a, b\}$ $H_0: \theta = a$ $H_A: \theta = b$ $\Theta = \{5'0.5'', \infty\}$
-
- $\Theta = \mathbb{R}$ $H_0: \theta \in \{24, 67\} \cup [10, \infty)$
 $H_A: \theta \in \mathbb{R} \setminus (\{24, 67\} \cup [10, \infty))$

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 matter for our 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.: effect of a new drug on a disease,
effect of CO₂ emissions on global temps,
effects of a scandal on a politician’s ratings)
- The “default” assumption is that there is 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. Model the coin flip as Bernoulli(p), $p \in (0, 1)$.
 $H_0: p = \frac{1}{2}$
 $H_A: p \neq \frac{1}{2}$.

- Example 3.4:

$\eta = \# \text{ of aces produced in a deck of cards produced by some company.}$

$$H_0: \eta = 4$$

$$H_A: \eta < 4 \text{ or } H_A: \eta \neq 4 \text{ etc.}$$

- 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 $|\Theta_0| = 1$, then H is a simple hypothesis. Otherwise, H is a composite hypothesis.

i.e., $\Theta_0 = \{\theta_0\}$ for some $\theta_0 \in \Theta$
i.e., $\theta \in \Theta_0$ with $|\Theta_0| > 1$

A simple hypothesis completely specifies the data-generating distribution

The Courtroom Analogy

- Consider a prosecution: the defendant is *innocent until proven guilty*
- But the whole point of the case is that the prosecutor suspects the defendant *is guilty*, and the purpose of the trial is to determine whether the evidence supports that guilt
- The jurors ask themselves: if the defendant really was innocent, how unlikely would this evidence be?
- If the evidence is overwhelmingly unlikely, the defendant is found guilty
- But if there's a *lack* of unlikely evidence, they find the defendant *not guilty*
NOT THE SAME AS INNOCENCE! Doesn't mean the defendant is innocent, just that there's not enough evidence 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 .

A Motivating Example

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

Seems unlikely!
• Large sample size (H_0 supports $\bar{X}_n \approx 5$ under $H_0: \theta = 5$ for large n)
• Observation is many standard deviations away from the mean under H_0 .
• ... etc

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

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

Under H_0 , $P_{\theta}(\bar{X}_{100} \leq -10)$

$$= P_{\theta}\left(\frac{\bar{X}-5}{\sqrt{100}} \leq \frac{-10-5}{\sqrt{100}}\right)$$

$\sim N(0, 1)$

$$= P(Z \leq -1.50) \text{ where } Z \sim N(0, 1)$$

$$= \underline{\underline{P}}(-1.50)$$

$$= 0.00002\dots$$

Lower than any reasonable threshold & plausibility!

So the 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 H_0 if $\bar{x} < 2$
 - Reject H_0 if $x_1 = 2$ or $x_1 = 4$
 - Reject H_0 if $x_{(1)} \geq 12$
- **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, $\mathbb{P}_\theta(\text{Reject } H_0) = \mathbb{P}_\theta(\mathbf{X} \in R)$
- **Example 3.7:**
 - $R = \{\bar{x} \in \mathcal{X}^n : \bar{x} < 2\}$
 - $R = \{x \in \mathcal{X}^n : x_1 = 2 \text{ or } x_1 = 4\}$
 - $R = \{\bar{x} \in \mathcal{X}^n : x_{(1)} \geq 12\} \rightarrow \mathbb{P}_\theta(\text{reject } H_0) = \mathbb{P}_\theta(\bar{x} \in R) = \mathbb{P}_\theta(X_{(1)} \geq 12) \text{ depends on } \theta !$

Poll Time!

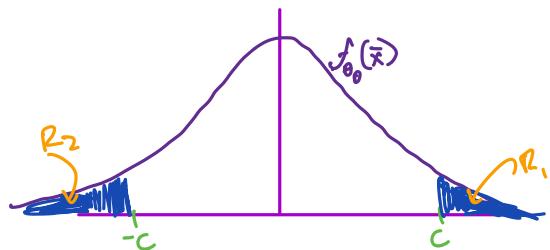
$P_0(\text{fail to reject } H_0)$

$= 1 - P_0(\text{reject } H_0)$

$= 1 - P_0(\bar{X} \in R)$

One-Tailed and Two-Tailed Tests

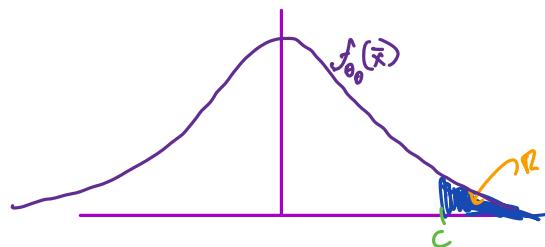
- If $\Theta \subseteq \mathbb{R}$ and H_0 is simple, then the rejection region is usually in both tails of the distribution: $H_0: \theta = \theta_0$



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

$$P_{\theta_0}(\bar{X} \in R) = P_{\theta_0}(\bar{X}_n > c \text{ OR } \bar{X}_n < -c)$$

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



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

$$P_{\theta_0}(\bar{X} \in R) = P_{\theta_0}(\bar{X}_n > c)$$

- 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

- 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.

ie, a "false positive"

- Example 3.8:

$N(\mu, \sigma^2)$, σ^2 known. $H_0: \mu = 0$, $H_A: \mu \neq 0$.

Observe $\bar{X}_n = -3$ and reject H_0 .

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

Or, observe $\bar{X}_n = 1.3$ and fail to reject H_0 .

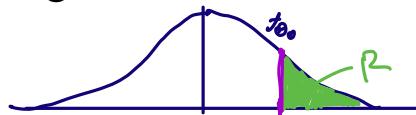
If the data actually came from $N(1, \sigma^2)$, we've made a type II error

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

Because they depend on whether $\theta \in \Theta_0$ or $\theta \in \Theta_0^c$, which we can 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 $c \in \mathbb{R}$ (ie, we reject 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 $\mathbb{P}_{\theta}(\mathbf{X} \in R) = \mathbb{P}_{\theta}(\bar{X} \geq c)$ very small under H_0



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

- Example 3.9: $N(\mu_1)$ model, $H_0: \mu \leq 0$ vs $H_A: \mu > 0$. Threshold: $\alpha = 0.05$.
Which $c \in \mathbb{R}$ do we need? Say $n=100$.

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

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

$$= \mathbb{P}(Z > c/\sqrt{100}) \quad \text{where } Z \sim N(0, 1)$$

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

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

If $n=10$, then $c \approx 0.5201$.

Usually, 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 $\beta : \Theta \rightarrow [0, 1]$ given by $\beta(\theta) = \mathbb{P}_\theta (\mathbf{X} \in R)$.

- Observe that

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

- **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 " $1 - \beta$ ". Not the same as our $\beta(\theta)$!!!

The Power Function: Examples

- **Example 3.11:** Let $X_1, X_2, \dots, X_n \stackrel{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ 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.

$$\begin{aligned}\beta(\mu) &= P_{\mu}(\bar{X} \in R) \\ &= P_{\mu}(\bar{X}_n > c) \\ &= P_{\mu}\left(\frac{\bar{X}_n - \mu}{\sqrt{\sigma^2/n}} > \frac{c - \mu}{\sqrt{\sigma^2/n}}\right) \\ &= P(Z > \frac{c - \mu}{\sqrt{\sigma^2/n}}) \quad \text{where } Z \sim N(0, 1) \\ &= 1 - \Phi\left(\frac{c - \mu}{\sqrt{\sigma^2/n}}\right)\end{aligned}$$

Technically, we didn't need to know H_0 or H_0^c here. But $\beta(\mu)$ is only useful when we have them in mind.

Poll Time!

$$P_{\theta_0}(X \in R)$$

$$= P_{\theta_0}(\text{reject } H_0)$$

= probability of rejecting H_0 when H_0 is true

= probability of a type I error

Size and the Probability of Rejection

and the distribution of \bar{X} is continuous

- If we have a simple null hypothesis, 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 as "maximum over all possible θ 's in Θ_0 "
- Example 3.12:

$N(\mu, \sigma^2)$, σ^2 known. $H_0: \mu \leq 0$ vs $H_A: \mu > 0$, $R = \{\bar{x} \in \mathcal{X}: \bar{x} > c\}$. Want a size- α test.

We want $\alpha = \sup_{\mu \leq 0} \mathbb{P}_\mu(\bar{X} > c)$

$$\begin{aligned} &= \sup_{\mu \geq 0} \left(1 - \Phi\left(\frac{c-\mu}{\sigma/\sqrt{n}}\right) \right) \\ &= 1 - \inf_{\mu \geq 0} \Phi\left(\frac{c-\mu}{\sigma/\sqrt{n}}\right) \end{aligned}$$

$\Rightarrow = 1 - \Phi\left(\frac{c}{\sigma/\sqrt{n}}\right)$

\Rightarrow Choose $c = \sqrt{n} \cdot \Phi^{-1}(1-\alpha)$ to make this a size- α test

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 $\sup_{\theta \in \Theta_0} \mathbb{P}_\theta(\mathbf{X} \in R) \leq \alpha$.

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

- Example 3.13:

Let $X \sim \text{Bin}(5, \theta)$, $\Theta \subset (0, 1)$. $H_0: \theta \leq \frac{1}{2}$ vs $H_A: \theta > \frac{1}{2}$. Suppose $R = \{5\}$.

Then
$$\begin{aligned} & \sup_{\theta \leq \frac{1}{2}} \mathbb{P}_\theta(X \in R) \\ &= \sup_{\theta \leq \frac{1}{2}} \theta^5 \\ &= \left(\frac{1}{2}\right)^5 = 0.03125. \end{aligned}$$

Generally $R \subseteq \{0, 1, 2, \dots, 5\}$

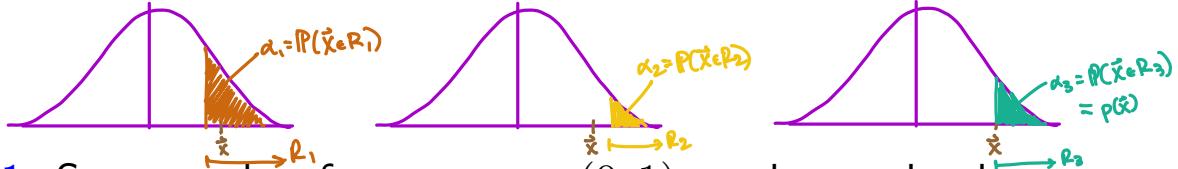
So this is a level-0.05 test and a level-0.04 test and a level-0.03125 test (but not a level-0.03 test!)

But we can't get a size-0.05 test! There's no R st. $\sup_{\theta \leq \frac{1}{2}} \mathbb{P}_\theta(X \in R) = 0.05$.

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 w/ σ^2 known, $T(\vec{X}) = \bar{X}_n$ is a good test statistic, because under $H_0: \mu = \mu_0$, we know $\bar{X}_n \sim N(\mu_0, \sigma^2/n)$
 - In the Bernoulli(p) model, $T(\vec{X}) = \sum_{i=1}^n X_i$ is good because under $H_0: p = p_0$, $T(\vec{X}) \sim \text{Bin}(n, p_0)$
 - 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 use *p*-values? We first set $\alpha \in (0, 1)$, then we observe $\vec{X} = \vec{x}$, and then we calculate our *p*-value $p(\vec{x})$.
If $p(\vec{x}) < \alpha$, then we reject H_0 .
If $p(\vec{x}) \geq \alpha$, then we fail to reject H_0 .

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 $T : \mathcal{X}^n \rightarrow \mathbb{R}$. If we observe $\mathbf{X} = \mathbf{x}$, then our observed p -value is $p(\mathbf{x}) = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(T(\mathbf{X}) \geq T(\mathbf{x}))$.
No proof...
- 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

$$\text{If } R = \left\{ \bar{\mathbf{x}} \in \mathcal{X}^n : T(\bar{\mathbf{x}}) \leq c \right\} \text{ then } p(\bar{\mathbf{x}}) = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(T(\bar{\mathbf{x}}) \leq T(\bar{\mathbf{x}}))$$

Poll Time!

p-values: none of the above'

Famous Examples: The Two-Sided Z -Test

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

$$\text{ie, } H_0 : \mu = \mu_0$$

$$Z(\mathbf{X}) = \frac{\bar{X}_n - \mu_0}{\sqrt{\sigma^2/n}} \sim N(0, 1) \text{ under } H_0$$

We want $c > 0$ such that

$$\begin{aligned}\alpha &= \sup_{\mu \in \mathbb{R}_0} P_\mu(|Z(\mathbf{x})| > c) \\ &= P_{\mu_0}(|Z(\mathbf{x})| > c) \\ &= P(|Z| > c) \text{ where } Z \sim N(0, 1) \\ &= 1 - P(-c \leq Z \leq c) \\ &= 1 - \Phi(c) + [1 - \Phi(c)] \\ &= 2 - 2 \cdot \Phi(c) \\ \Rightarrow c &= \Phi^{-1}(1 - \alpha_2) =: z_{1-\alpha_2} \end{aligned}$$

So our rejection region is

$$R = \{x \in \mathcal{X} : |Z(x)| > z_{1-\alpha_2}\}$$

$$z_{1-\alpha_2} = -z_{\alpha/2}$$

by symmetry **check!**

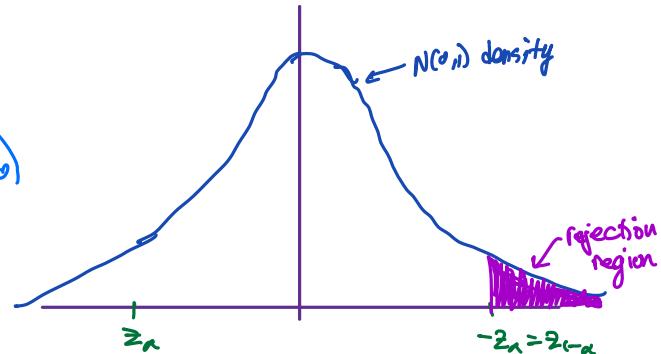
"critical value"
or "cutoff point"
for the $N(0, 1)$ distribution (basically a quantile)
WARNING: some authors instead define $z_\alpha = \Phi^{-1}(1 - \alpha) \dots$

Famous Examples: The One-Sided Z-Test

- **Example 3.16:** Let $X_1, X_2, \dots, X_n \stackrel{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ with $\mu \in \mathbb{R}$ and σ^2 known. Construct a ~~level- α~~ ^{size} 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.

$$\begin{aligned}\alpha &= \sup_{\mu \leq \mu_0} P_{\mu} \left(\frac{\bar{X}_n - \mu}{\sqrt{\sigma^2/n}} > c \right) \\ &= P_{\mu_0} \left(\frac{\bar{X}_n - \mu_0}{\sqrt{\sigma^2/n}} > c \right) \quad (\text{the sup is attained at } \mu_0) \\ &= P(Z > c) \quad \text{where } Z \sim N(0, 1) \\ &= 1 - \Phi(c) \\ \Rightarrow c &= \Phi^{-1}(1 - \alpha) = z_{1-\alpha} \quad (= -z_\alpha)\end{aligned}$$



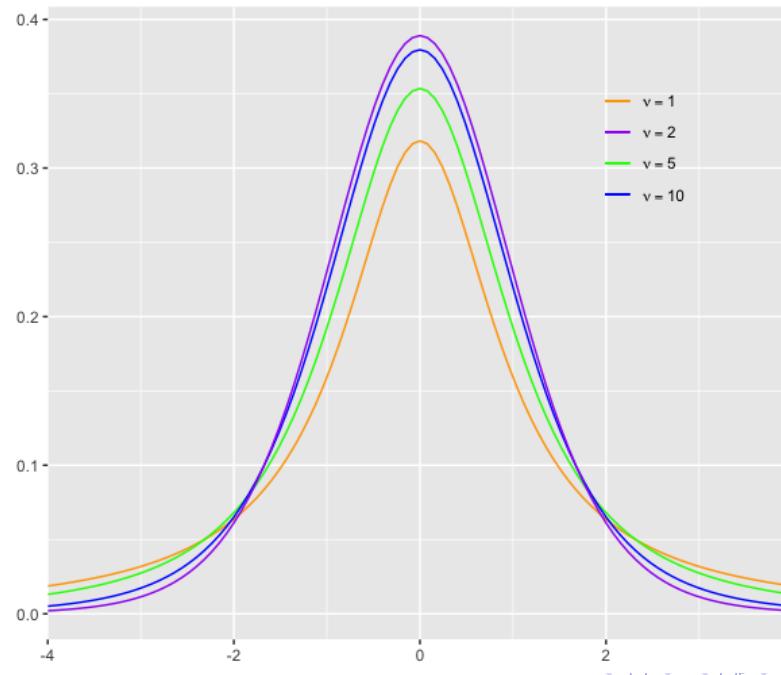
Exercise: construct a level- α test of $H_0: \mu \geq \mu_0$ vs $H_A: \mu < \mu_0$.

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 $Y, X_1, X_2, \dots, X_n \stackrel{iid}{\sim} \mathcal{N}(0, 1)$. Then

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

~~Exercise!~~

num, too hard...

- Equivariantly, $T \stackrel{d}{=} \frac{Y}{\sqrt{Q/n}}$ where $Y \sim N(0, 1)$, $Q \sim \chi_{(n)}^2$ and $Y \perp Q$
- Theorem 3.3: Let $T_n \sim t_n$. Then $T_n \xrightarrow{d} Z$ as $n \rightarrow \infty$, where $Z \sim N(0, 1)$.

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

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

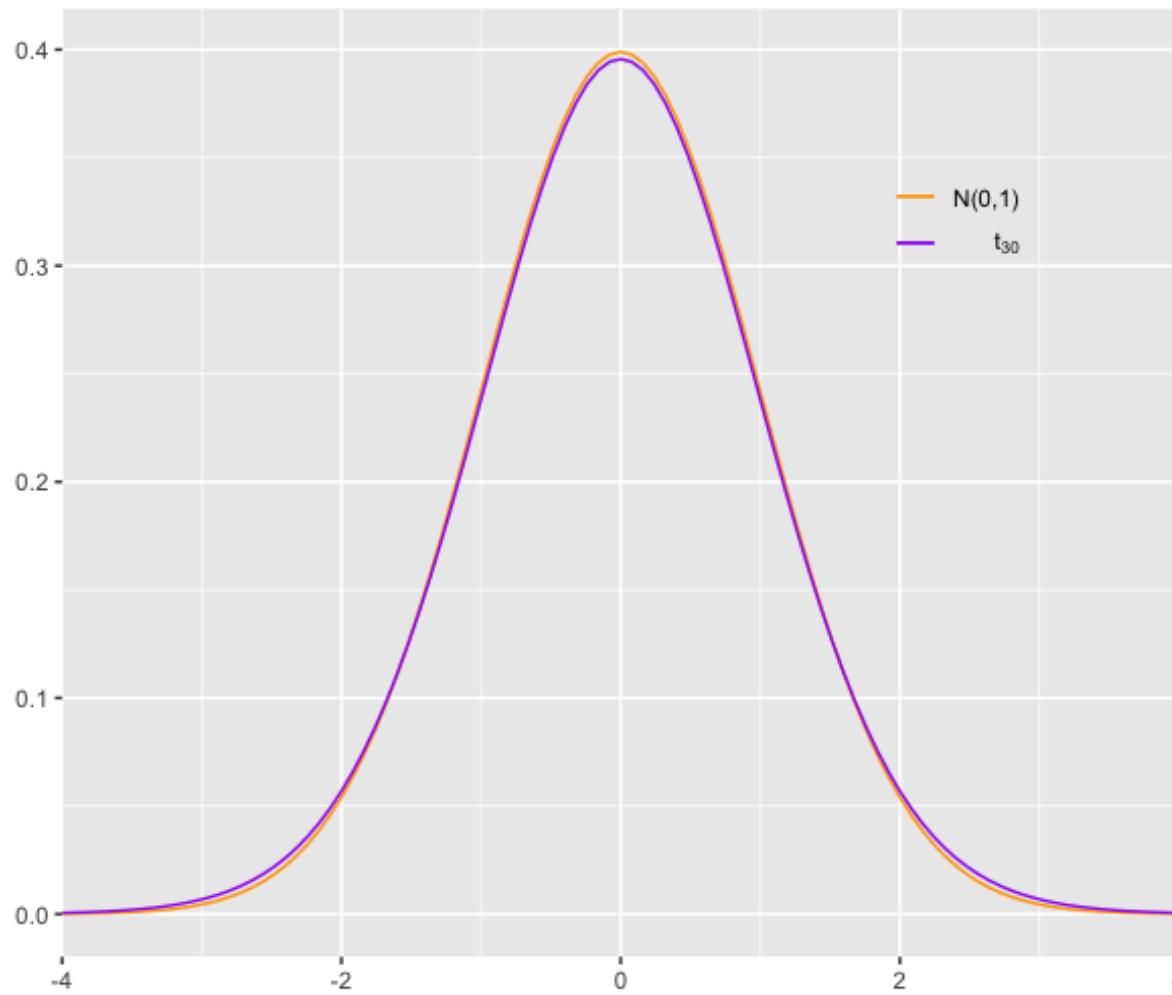
Continuous
Mapping
Theorem

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

By Slutsky's theorem, $\frac{Y}{\sqrt{\frac{1}{n} \sum_{i=1}^n X_i^2}} \xrightarrow{P} N(0, 1)$. \square

(Module 5)

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

$$\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

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

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

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

Proof.

$$\frac{\bar{X}_n - \mu}{\sqrt{S^2/n}} = \frac{\frac{\bar{X}_n - \mu}{\sqrt{\sigma^2/n}}}{\sqrt{\frac{(n-1)S^2}{(n-1)\sigma^2}}} \stackrel{\text{def}}{=} \frac{Z}{Q} \quad \text{from Theorem 1.7}$$

$$\stackrel{\text{def}}{=} \frac{Z}{\sqrt{\frac{1}{n-1} \cdot Q}} \quad \text{where } Z \sim N(0, 1) \quad Q \sim \chi^2_{n-1} \quad Z \perp Q$$

$$\stackrel{\text{def}}{=} t_{n-1} \text{ by Theorem 3.2} \quad \square$$

Famous Examples: The Two-Sided t -Test

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

$$T(\mathbf{X}) = \frac{\bar{X} - \mu_{\text{o}}}{\sqrt{S^2/n}}.$$

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

$$\alpha = P_{y_0}(|T(x)| > c)$$

$$= 1 - P_{\mu_0}(-c \leq \frac{\bar{X}_n - \mu_0}{\sqrt{S^2/n}} \leq c)$$

$$= 1 - P_{\mu_0} \left(\frac{\bar{X} - \mu_0}{\sqrt{s^2/n}} \leq c \right) + P_{\mu_0} \left(\frac{\bar{X} - \mu_0}{\sqrt{s^2/n}} \leq -c \right)$$

$$= \underbrace{1 - P(t_{n+1} \leq c)}_{= P(t_{n+1} > c)} + \underbrace{P(t_{n+1} \leq -c)}_{= F_{t_{n+1}}(-c)}$$

$\stackrel{?}{=} \mathbb{P}(t_{n-1} \leq -c)$ by symmetry

$$\stackrel{=}{\rightarrow} F_{t_{n+1}}(-c)$$

↳ cdf & a t_{n-1} random variable

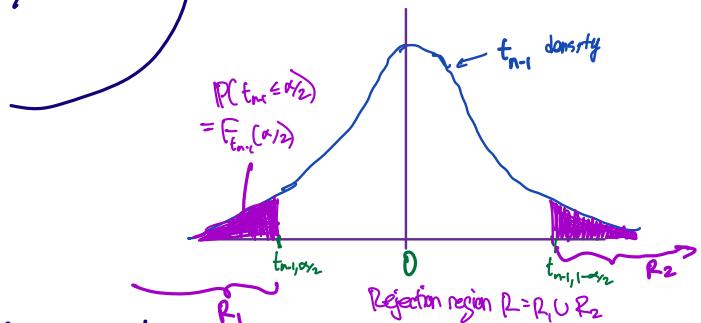
$$S_0 R = \{x \in X^n : |\pi(x)| > -t_{n-1, \alpha_2}\}$$

$$r = 2 \cdot F_{t_{\max}}(-c)$$

$$\Rightarrow c = -F_{t_{n-1}}^{-1}(\alpha_2) =: -t_{n-1, \alpha_2}$$

$$\left(= t_{n-r, 1-\alpha/2} \right)$$

by symmetry



Famous Examples: The One-Sided t -Test

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

EXERCISE!

$H_0: \mu \leq \mu_0$ vs $H_A: \mu > \mu_0$ too!

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 
- Moreover, 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 $X_1, X_2, \dots, X_n \stackrel{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ where $\mu \in \mathbb{R}$ and σ^2 is known, and we want to test $H_0 : \mu \leq \mu_0$ 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 $\mu \geq \mu_0 + \sigma$?

Power function: $\beta(\mu) = P_{\mu}\left(\frac{\bar{X}-\mu_0}{\sqrt{\sigma^2/n}} > c\right) = P_{\mu}\left(\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}} > c + \frac{\mu_0-\mu}{\sqrt{\sigma^2/n}}\right) = P\left(Z > c + \frac{\mu_0-\mu}{\sqrt{\sigma^2/n}}\right)$ where $Z \sim N(0, 1)$

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

$\sup_{\mu \geq \mu_0}$ attained at μ_0

subtract $\frac{\mu_0-\mu}{\sqrt{\sigma^2/n}}$ from both sides

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

$$\Rightarrow \Phi(c) = 0.9 \Leftrightarrow c = \Phi^{-1}(0.9) \approx 1.2816 \quad (\text{regardless of } n)$$

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

$$\begin{aligned} \Rightarrow 0.8 &\leq \beta(\mu_0 + \sigma) \\ &= 1 - \Phi(c - \sqrt{n}) \approx 1 - \Phi(1.2816 - \sqrt{n}) \end{aligned}$$

$$\Rightarrow 1.2816 \leq \Phi^{-1}(0.2) + \sqrt{n}$$

$$\Rightarrow n \approx 4.507$$

Choose $c \approx 1.2816$
and $n = 5$.

The Problems With the p 's

- 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!

The Problems With the p 's

<u>P-VALUE</u>	<u>INTERPRETATION</u>
0.001	
0.01	
0.02	HIGHLY SIGNIFICANT
0.03	
0.04	
0.049	SIGNIFICANT
0.050	OH CRAP. REDO CALCULATIONS.
0.051	
0.06	ON THE EDGE OF SIGNIFICANCE
0.07	
0.08	HIGHLY SUGGESTIVE, SIGNIFICANT AT THE $p < 0.10$ LEVEL
0.09	
0.099	
≥ 0.1	HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS

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

If: start w/ $\alpha = 0.05$, observe $\bar{X} = \bar{x}$, calculate $p(\bar{x}) = 0.07$,

declare results "statistically 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

"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."

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**)

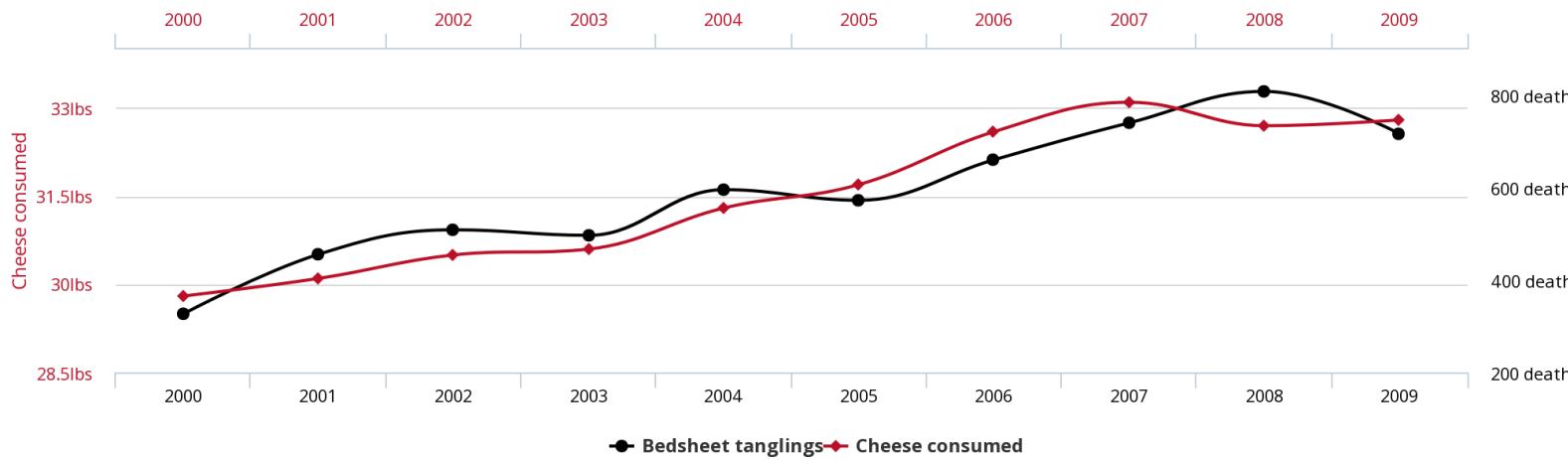
Should We Be Eating Less Cheese?

Per capita cheese consumption

correlates with

$\rho \approx 0.96$

Number of people who died by becoming tangled in their bedsheets



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

Poll Time!

Examples of *p*-Hacking

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

Like circular reasoning! It's like observing $X=12$ and then claiming $P(X>11)$ 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

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

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, $\log(\lambda(\mathbf{x})) = \sup_{\theta \in \Theta_0} l(\theta | \mathbf{x}) - \sup_{\theta \in \Theta} l(\theta | \mathbf{x})$

Poll Time!

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

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

Choosing $c=1$ means we reject when $\lambda(\vec{x}) \leq 1$ (always true!)

If $c=0$, we'd always fail to reject!

So we usually care about $c \in (0, 1)$.

LRTs: Examples

$X_1, \dots, X_n \stackrel{iid}{\sim} N(\mu, \sigma^2)$.
 H₀: $\mu = \mu_0$.
 H_A: $\mu \neq \mu_0$.

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

$$\lambda(\vec{x}) = \frac{L(\mu_0 | \vec{x})}{L(\bar{\mu} | \vec{x})} = \frac{\exp\left(-\frac{\sum(x_i - \mu_0)^2}{2\sigma^2}\right)}{\exp\left(-\frac{\sum(x_i - \bar{\mu})^2}{2\sigma^2}\right)} = \exp\left(\frac{1}{2\sigma^2} \left[-\sum(x_i - \mu_0)^2 + \sum(x_i - \bar{\mu})^2 \right]\right) = \exp\left(\frac{-n}{2\sigma^2} \cdot (\bar{x} - \mu_0)^2\right) \quad \text{Check!}$$

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

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

$$\Leftrightarrow -\frac{n}{2\sigma^2} (\bar{x} - \mu_0)^2 \leq \log(c)$$

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

So we reject 'when $|Z(\vec{x})| > c'$ for some $c > 0 \Rightarrow$ The 2-sided Z-test is an LRT!

LRTs: Examples

- **Example 3.21:** Let X_1, X_2, \dots, 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 $H_A : \theta > \theta_0$.

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

Unrestricted MLE? $L(\theta | \vec{x})$ is clearly increasing in θ until $\theta = x_{(1)}$, and then = 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 is θ_0 .

$$\text{So } \lambda(\vec{x}) = \begin{cases} 0, & X_{(1)} \leq \theta_0 \\ e^{-n(x_{(1)} - \theta_0)}, & X_{(1)} > \theta_0 \end{cases}$$

So the LRT has rejection region $R = \{ \vec{x} \in \mathcal{X}^n : e^{-n(x_{(1)} - \theta_0)} \leq c \text{ OR } X_{(1)} > \theta_0 \}$
 $= \{ \vec{x} \in \mathcal{X}^n : X_{(1)} \geq \theta_0 - c' \text{ OR } X_{(1)} < \theta_0 \}$ for some c' .

Simple Tests Have Simple LRTs

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

$$\lambda(\mathbf{X}) = \frac{L(\theta_0 | \mathbf{X})}{L(\hat{\theta} | \mathbf{X})}, \quad \text{Proof: } \underline{\text{Exercise.}}$$

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

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

The MLE of θ is $\hat{\theta}(\vec{x}) = X_{(n)}$.

Numerator: $L(\theta_0 | \vec{x}) = \theta_0^{-n} \cdot \prod_{x_{(n)} \leq \theta_0}$

Denominator: $L(X_{(n)} | \vec{x}) = X_{(n)}^{-n} \cdot \prod_{x_{(n)} \leq X_{(n)}} = X_{(n)}^{-n}$

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

$$\text{So } \lambda(\vec{x}) = \frac{\theta_0^{-n} \cdot \prod_{x_{(n)} \leq \theta_0}}{X_{(n)}^{-n}} \Rightarrow R = \left\{ \vec{x} \in \mathcal{X}^n : \left(\frac{X_{(n)}}{\theta_0} \right)^n \cdot \prod_{x_{(n)} \leq \theta_0} \leq c \right\} \text{ for some } c$$

LRTs: Examples

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

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

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

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

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

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 $L^*(\theta | t) = g_\theta(t)$
- But would this change our conclusions?
- Theorem 3.6: Suppose ~~$T(\mathbf{X})$~~ is sufficient for θ . If $\lambda(\mathbf{x})$ and ~~$\lambda^*(\mathbf{x})$~~ 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,

$$\begin{aligned}\lambda(\vec{x}) &= \frac{\sup_{\theta \in \Theta} L(\theta | \vec{x})}{\sup_{\theta \in \Theta} L(\theta | \vec{x})} = \frac{\sup_{\theta \in \Theta} f_\theta(\vec{x})}{\sup_{\theta \in \Theta} f_\theta(\vec{x})} = \frac{\sup_{\theta \in \Theta} h(\vec{x}) \cdot g_\theta(T(\vec{x}))}{\sup_{\theta \in \Theta} h(\vec{x}) \cdot g_\theta(T(\vec{x}))} \\ &= \frac{\sup_{\theta \in \Theta} g_\theta(T(\vec{x}))}{\sup_{\theta \in \Theta} g_\theta(T(\vec{x}))} = \frac{\sup_{\theta \in \Theta} L^*(\theta | T(\vec{x}))}{\sup_{\theta \in \Theta} L^*(\theta | T(\vec{x}))} = \lambda^*(T(\vec{x})).\end{aligned}$$

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

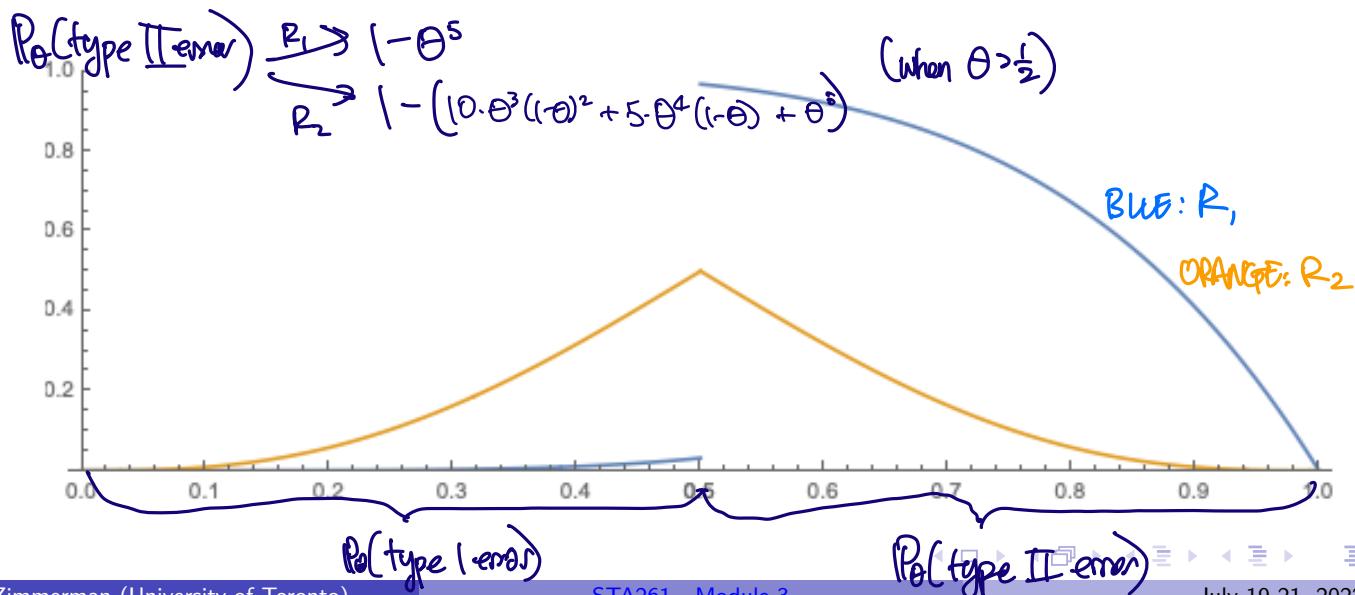
You Can't Get the Perfect Power Function

- 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 $R_2 = \{3, 4, 5\}$

Power functions: $\beta_1(\theta) = P_\theta(X=5) = \theta^5$

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

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



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:
 - Medical study: test for a 100% fatal disease. Definitely want to minimize false negatives (ie, Type II errors)
 - Courtroom setting: a conviction means the death penalty. A type I error means convicting an innocent person to death.
 - Hypothetical test for a heart disorder: if a patient has the disorder, the only treatment is a heart transplant. If untreated, there's a 50% chance of death.
 - Type I error: a donor heart is wasted, and patient is (needlessly) on anti-rejection meds for life
 - Type II error: 50% chance of the patient dying

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 level- α test will do the right thing (ie, correctly reject H_0) more often than any other level- α test of H_0 vs H_A .

Alternatively: it's a size- α (or 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)

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! Use the factorization theorem!

The Neyman-Pearson Lemma: Examples

- **Example 3.25:** Let $X_1, X_2, \dots, X_n \stackrel{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ 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 $T(\bar{x}) = \bar{X}_n$, which is sufficient for μ .

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

$$\Rightarrow \log(k_0) < \frac{1}{2\sigma^2/n} [\mu_0^2 - \mu_1^2 + 2\bar{x}(\mu_1 - \mu_0)]$$

> 0 by assumption

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

$=: c$

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 So by Theorem 3.8, the one-sided Z -test for $\mathbb{H} = \{\mu_0, \mu_1\}$ is a UMP level- α test, where $\alpha = P_{\mu_0}(\bar{X}_n < c)$.

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 $\theta_2 \geq \theta_1$.

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

Discussion?

Then a test with rejection region $R = \{T > c_0\}$ is a UMP level- α test, where $\alpha = \mathbb{P}_{\theta_0}(T > c_0)$.

No proof.

- 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 directly from Karlin-Rubin
(multiply everything through by -1)

The Neyman-Pearson Lemma: Examples

$$H_0: \mu \leq \mu_0 \quad v \quad H_A: \mu > \mu_0$$

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

$T(\bar{X}) = \bar{X}_n$ is sufficient for μ , with pdf

$$\begin{aligned} g_{\mu}(t) &= \left(2\pi\sigma^2/n\right)^{-\frac{1}{2}} \cdot \exp\left(-\frac{1}{2\sigma^2/n}(t-\mu)^2\right) \\ &= \left(2\pi\sigma^2/n\right)^{-\frac{1}{2}} \cdot \exp\left(-\frac{1}{2\sigma^2/n}(t^2 - 2\mu t + \mu^2)\right). \end{aligned}$$

Let $\mu_2 \geq \mu_1$. Then

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

is monotone non-decreasing in t .

By Karlin-Rubin, the test w/ rejection region $R = \{ \bar{X}_n > c_0 \}$ is a level- α test
where $\alpha = P_{\mu_0}(\bar{X}_n > c_0)$.

$$= \left\{ \frac{\bar{X}_n - \mu_0}{\sqrt{\sigma^2/n}} > c_0' \right\}$$

That's a one-sided Z -test!

The Neyman-Pearson Lemma: Examples

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

$T = T(\vec{x}) = \sum X_i$ is sufficient for λ , where $T \sim \text{Poisson}(n\lambda)$, with pmf
 $P_\lambda(t) = \frac{(n\lambda)^t e^{-n\lambda}}{t!}$. Let $\lambda_2 \geq \lambda_1$. Then

$$\frac{P_{\lambda_2}(t)}{P_{\lambda_1}(t)} = \frac{(n\lambda_2)^t e^{-n\lambda_2}}{(n\lambda_1)^t e^{-n\lambda_1}} = \left(\frac{\lambda_2}{\lambda_1} \right)^t e^{n(\lambda_1 - \lambda_2)} \text{ is increasing in } t. \text{ By Karlin-Rubin,}$$

$\underbrace{\quad}_{\geq 1}$

a test w/ rejection region $R = \{ \vec{x} \in \mathcal{X}^n : \sum X_i > c_0 \}$ is a level- α test, where

$$\alpha = P_{\lambda_0}(\sum X_i > c_0). \text{ How do we find } [c_0]? \text{ Well } \alpha = 1 - P_{\lambda_0}(\sum X_i \leq c_0)$$
$$= 1 - \sum_{j=0}^{c_0} \underbrace{\frac{(n\lambda_0)^j e^{-n\lambda_0}}{j!}}_{P_{\lambda_0}(j)}$$

Keep subtracting $P_{\lambda_0}(j)$'s from 1 until $\leq \alpha$.

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 $X_1, X_2, \dots, X_n \stackrel{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ 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 $\frac{\bar{X} - \mu_0}{\sqrt{\sigma^2/n}} < -z_\alpha$, which is UMP level- α for $H_A' : \mu = \mu_1$. Test 2 rejects H_0 if $\frac{\bar{X} - \mu_0}{\sqrt{\sigma^2/n}} > z_\alpha$, which is UMP level- α for $H_A'' : \mu = \mu_2$.

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

$$\begin{aligned}\beta_2(\mu_2) &= P_{\mu_2} \left(\frac{\bar{X} - \mu_2}{\sqrt{\sigma^2/n}} > z_\alpha + \frac{\mu_0 - \mu_2}{\sqrt{\sigma^2/n}} \right) \\ &= P \left(Z > z_\alpha + \frac{\mu_0 - \mu_2}{\sqrt{\sigma^2/n}} \right) \quad \text{where } Z \sim N(0, 1) \\ &> P(Z > z_\alpha) \quad \underbrace{\text{where } Z \sim N(0, 1)}_{> 0} \\ &= P(Z < -z_\alpha) \quad \text{by symmetry}\end{aligned}$$

$$\begin{aligned}&= P_{\mu_2} \left(\frac{\bar{X} - \mu_2}{\sqrt{\sigma^2/n}} < -z_\alpha + \frac{\mu_0 - \mu_2}{\sqrt{\sigma^2/n}} \right) \\ &= P_{\mu_2} \left(\frac{\bar{X} - \mu_0}{\sqrt{\sigma^2/n}} < -z_\alpha \right) \\ &= \beta_1(\mu_2).\end{aligned}$$

So Test 2 has strictly higher power at μ_2 than Test 1 does. Contradiction!
So no UMP level- α test exists here.

EXERCISE: try to generalize that to general location families!