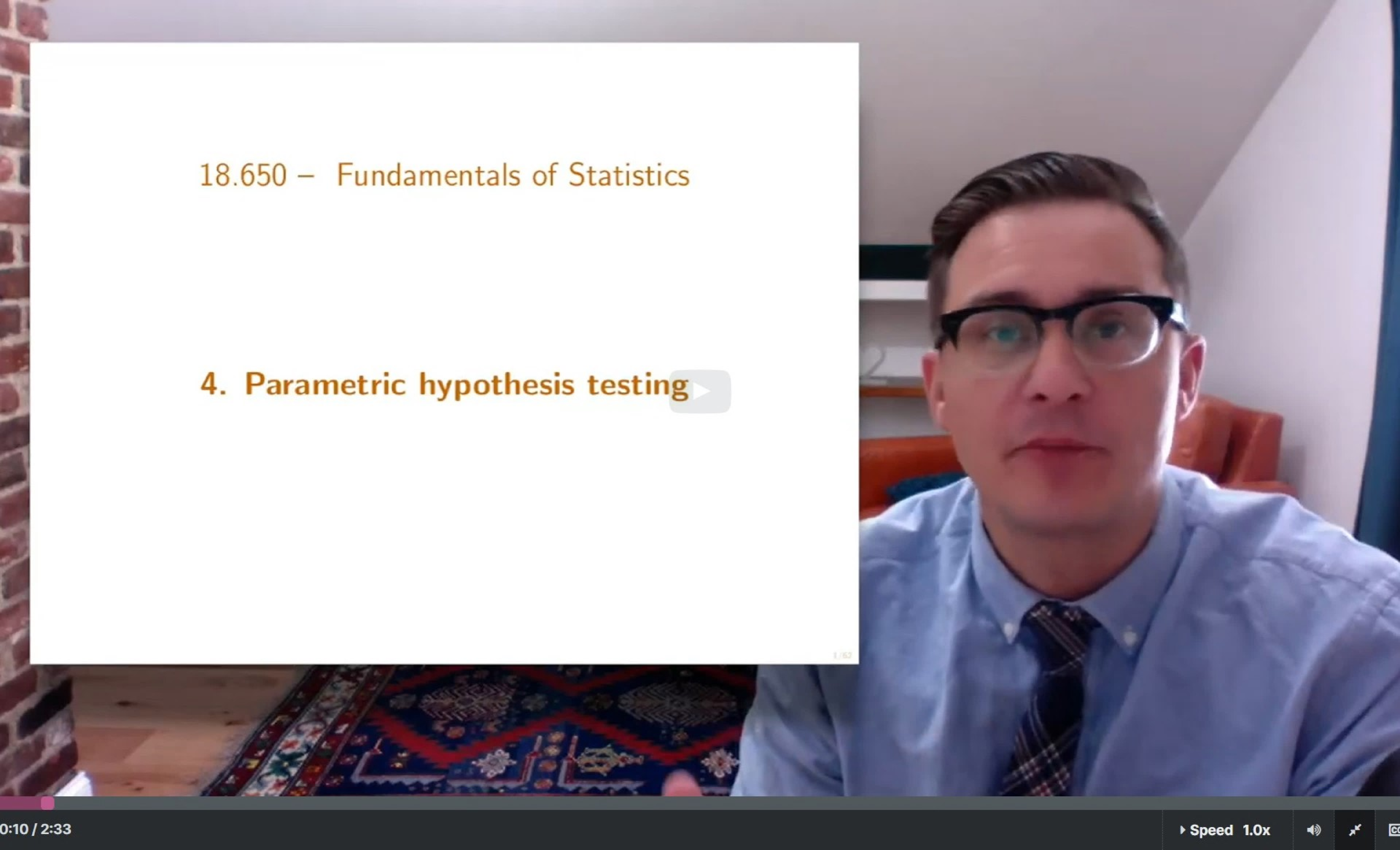

18.650 – Fundamentals of Statistics
4. Parametric hypothesis testing
0:10 / 2:33
Speed 1.0x

Goals

Recall: waiting time in the ER

$$H_0 : \mu \leq 30$$
$$H_1 : \mu > 30$$

How to perform this test based on data?

- test statistic
- rejection region
- p-value

1:10 / 2:33

Speed 1.0x

How to measure the performance of a test?

- Type I and type II errors
- level
- power

Construct PARAMETRIC tests:

$$H_0 : \mu \leq 30$$
$$H_1 : \mu > 30$$

- Wald test
- T-Test

Waiting time in the ER

- The average waiting time in the Emergency Room (ER) in the US is 30 minutes according to the CDC
- Some patients claim that the new Princeton-Plainsboro hospital has a longer waiting time. Is it true?
- Collect a sample: $X_1, \ldots, X_n$ (waiting time in minutes for n random patients) with unknown expected value $\mathbb{E}[X_1] = \mu$.
- We want to know if $\mu > 30$.

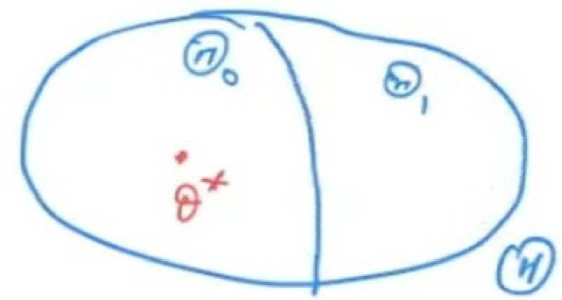
$H_0: \mu \leq 30$

$H_1: \mu > 30$

2:29 / 4:54 Speed 1.0x

Statistical formulation

- Consider a sample $X_1, \dots, X_n$ of i.i.d. random variables and a statistical model $(E, (\mathbb{P}_\theta)_{\theta \in \Theta})$.



- Let Θ_0 and Θ_1 be a *partition* of Θ.

- Consider the two hypotheses: $\begin{cases} H_0 : & \theta \in \Theta_0 \\ H_1 : & \theta \in \Theta_1 \end{cases}$

- H_0 is the *null hypothesis*, H_1 is the *alternative hypothesis*.

- We say that we *test* H_0 *against* H_1.

Testing lexicon

H_k

- For $k = 0$ (H_0) or $k = 1$ (H_1), we say that
  - Θ_k is a *simple hypothesis* if $\Theta_k = \{\theta_k\}$
  - Θ_k is a *composite hypothesis* if Θ_k is of the following three forms

$$\Theta_k = \{\theta : \theta > \theta_k\}, \quad \Theta_k = \{\theta : \theta < \theta_k\}, \quad \Theta_k = \{\theta : \theta \neq \theta_k\}$$

- A test is typically either *one-sided* or *two-sided*

Two-sided

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \neq \theta_0 \end{cases}$$

One-sided

$$\begin{cases} H_0 : \theta \leq \theta_0 \\ H_1 : \theta > \theta_0 \end{cases} \quad \text{or} \quad \begin{cases} H_0 : \theta \geq \theta_0 \\ H_1 : \theta < \theta_0 \end{cases}$$

Examples

1. Waiting time in the ER

composite

$$
\begin{aligned}
H_0 &: \quad \mu \leq 30 \\
H_1 &: \quad \mu > 30
\end{aligned}
$$

composite

One-sided test.

2. In the Kiss example, we want to test

simple $\Theta_0 = \{.5\}$

$$
\begin{aligned}
H_0 &: \quad p = .5 \\
H_1 &: \quad p \neq .5
\end{aligned}
$$

composite

two-sided test

Clinical trials

- Pharmaceutical companies use hypothesis testing to test if a new drug is efficient.
- To do so, they administer a drug to a group of patients (test group) and a placebo to another group (control group).
- We consider testing a drug that is supposed to lower LDL (low-density lipoprotein), a.k.a "bad cholesterol" among patients with a high level of LDL (above 200 mg/dL)

2:02 / 6:30 Speed 1.0x

Notation and modelling

- Let $\mu_\mathrm{d} > 0$ denote the expected decrease of LDL level (in mg/dL) for a patient that has used the drug.
- Let $\mu_\mathrm{c} > 0$ denote the expected decrease of LDL level (in mg/dL) for a patient that has used the placebo.
- Hypothesis testing problem:

$$H_0 : \mu_d \leq \mu_c$$
$$H_1 : \mu_d > \mu_c$$

- We observe two independent samples:
  - $X_1, \ldots, X_{\underline{n}} \overset{iid}{\sim} \mathcal{N}(\mu_d, \sigma_\mathrm{d}^2)$ from the test group and
  - $Y_1, \ldots, Y_{\underline{m}} \overset{iid}{\sim} \mathcal{N}(\mu_c, \sigma_\mathrm{c}^2)$ from the Control group.
- This is a two-*sample* test: these are very common (A/B testing).

5:53 / 6:30 Speed 1.0x

Asymmetry in the hypotheses

$H_0 : \mu_d \leq \mu_c$
$H_1 : \mu_d > \mu_c$

- We want to decide whether to *reject* H_0 (look for evidence against H_0 in the data).
- H_0 and H_1 do not play a symmetric role: the data is only used to try to disprove H_0

H_0 : status quo
H_1 : a (scientific) discovery

- In particular lack of evidence, does not mean that H_0 is true ("innocent until proven guilty")

Examples

1. Waiting time in the ER

$$H_0: \quad \mu \leq 30 \rightarrow \text{status quo}$$
$$H_1: \quad \mu > 30$$

Status quo: CDC statement. We collect data to show that Princeton-Plainsboro is different

2. Kiss

$$H_0: \quad p = .5$$
$$H_1: \quad p \neq .5$$

Status quo: our intuition tells us there should be no preference. We collect data to show that there is one.

3. Clinical trials

$$H_0: \mu_d \leq \mu_c$$
$$H_1: \mu_d > \mu_c$$

Status quo: The drug is not more effective than a placebo. We collect data to prove that the drug is effective.

What is a test?

- A *test* is a statistic $\psi \in \{0, 1\}$ that does not depend on unknown quantities and such that:
  - If $\psi = 0$, H_0 is not rejected;
  - If $\psi = 1$, H_0 is rejected.

  Important remark: Can always write $\psi = \mathbb{1}\{R\}$, where R is an *event* called rejection region.

  $\psi = \mathbb{1}(\psi = 1)$

- Waiting time in the ER:

$$\begin{aligned} H_0 &: \quad \mu \leq 30 \\ H_1 &: \quad \mu > 30 \end{aligned} \qquad \psi = \mathbb{1}\{\bar{X}_n > C\}$$

- Kiss:

$$\begin{aligned} H_0 &: \quad p = .5 \\ H_1 &: \quad p \neq .5 \end{aligned} \qquad \psi = \mathbb{1}\left\{\left|\bar{X}_n - \tfrac{1}{2}\right| > C\right\}$$

- Clinical trials

$$\begin{aligned} H_0 &: \quad \mu_{\mathrm{d}} \leq \mu_{\mathrm{c}} \\ H_1 &: \quad \mu_{\mathrm{d}} > \mu_{\mathrm{c}} \end{aligned} \qquad \psi = \mathbb{1}\{\bar{X}_n - \bar{Y}_m > C\}$$

Errors

A test can make two types of errors:

	Fail to reject Null	Reject Null
H_0 true ($\theta \in \Theta_0$)	✓	type 1
H_1 true ($\theta \in \Theta_1$)	type 2	✓

Both errors can be computed from the power function

$$\beta(\theta) = \mathbb{P}_\theta[\psi = 1]$$

- If $\theta \in \Theta_0$,

$$\beta(\theta) = \mathbb{P}_\theta[\psi \text{ makes an error of type } 1]$$

We want $\beta(\theta)$ to be small

- If $\theta \in \Theta_1$,

$$\beta(\theta) = 1 - \mathbb{P}_\theta[\psi \text{ makes an error of type } 2]$$

We want $\beta(\theta)$ to be large

The Neyman-Pearson paradigm

Recall the waiting time in the ER example

$$H_0: \quad \mu \leq 30 \qquad\qquad \psi = \mathbb{I}\{\bar{X}_n > C\}$$
$$H_1: \quad \mu > 30$$

How to choose C ?

We are facing a dilemma: both errors should be small!

- To make Type I error $\to 0$, take $C \to +\infty$
- To make Type II error $\to 0$, take $C \to -\infty$ (enough to have $C = -\infty$)

Cannot make both small at the same time.

The *Neyman-Pearson paradigm*:

- Make sure that $\mathbb{P}[\text{Type I error}] \leq \alpha$ (e.g., $\alpha = 5\%, 1\%, \dots$)
- Minimize $\mathbb{P}[\text{Type II error}]$ subject to this constraint

Level

The value of $\alpha \in (0,1)$ chosen in the Neyman-pearson paradigm is called level of a test

For which $\theta \in \Theta_0$ should we compute $\mathbb{P}_\theta[\psi = 1]$ (probability of Type 1 error)?

- A test ψ has *level* α if

$$\mathbb{P}_\theta[\psi = 1] \le \alpha, \qquad \forall \theta \in \Theta_0.$$

$$\iff \max_{\theta \in \Theta_0} \mathbb{P}_\theta[\psi = 1] \le \alpha$$

- A test $\psi = \psi_n$ has *asymptotic level* α if

$$\lim_{n \to \infty} \max_{\theta \in \Theta_0} \mathbb{P}_\theta[\psi_n = 1] \le \alpha,$$

Building a test from a confidence interval

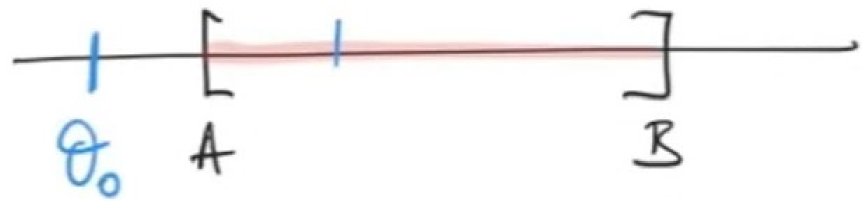
Given a confidence interval, we can often build a test (and vice versa).

- Let $I = [A, B]$ be a confidence interval at level $1 - \alpha$ for a parameter θ:

$$\mathbb{P}_\theta(\theta \in [A, B]) \geq 1 - \alpha$$

- We want to use this I to build a test at level α for

$$\begin{aligned} H_0 &: \quad \theta = \theta_0 \\ H_1 &: \quad \theta \neq \theta_0 \end{aligned}$$



- Natural candidate:

$$\psi = \mathbb{1}\{\theta_0 \notin [A, B]\}$$

- Level of test:

$$\mathbb{P}_{\theta_0}[\psi = 1] = \mathbb{P}_{\theta_0}[\theta_0 \notin I] = 1 - \mathbb{P}_{\theta_0}[\theta_0 \in I] \leq 1 - (1 - \alpha) = \alpha$$

- Therefore ψ is a test with level α

A test for the Kiss example

We want to test:

$$\begin{aligned} H_0 &: \quad p = 0.5 \\ H_1 &: \quad p \neq 0.5 \end{aligned}$$

We observe $R_1, \ldots, R_n \overset{iid}{\sim} \mathsf{Ber}(p)$.

- Recall that

$$\mathcal{I}_{\text{conserv}} = \left[\bar{R}_n - \frac{1.96}{2\sqrt{n}}, \bar{R}_n + \frac{1.96}{2\sqrt{n}} \right]$$

  is a confidence interval of asymptotic level $1 - \alpha$ for p.

- Consider the test:

$$\psi = \mathbb{1}\{0.5 \notin \mathcal{I}_{\text{conserv}}\}$$

  if $\alpha = 95\%$

  $\mathcal{I}_{\text{conserv}} = [0.56, 0.73]$

  $\hookrightarrow$ reject.

- We have

$$\lim_{n\to\infty} \mathbb{P}_{.5}[\psi = 1] = 1 - \lim_{n\to\infty} \mathbb{P}_{.5}[.5 \in \mathcal{I}_{\text{conserv}}] \leq 1 - (1-\alpha) = \alpha$$

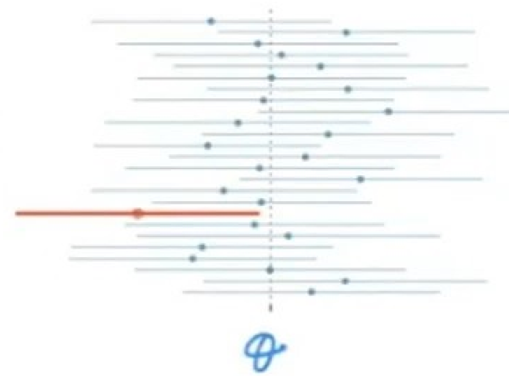
- Therefore ψ is a test with asymptotic level α

Meaning of the level

- Recall that

  $\mathcal{I}$ *is a CI at level* 95% *for* θ

  means that if we repeat the experiment many times, at least 95% confidence intervals will contain the true parameter θ.



- Similarly:

  ψ *is a test at level* 5% *for* H_0 *vs* H_1

  means that if we repeat the experiment many times, at most 5% of the tests will make an error of type 1

What if we change the level?

$\mathcal{I}_{conserv} = \left[\bar{R}_n \pm \frac{q_{\alpha/2}}{2\sqrt{n}}\right]$

level 95%

With our data $\mathcal{I}_{\text{conserv}} = [0.56, 0.73]$ so we reject H_0 at level 5%

α	$q_{\alpha/2}$	$\mathcal{I}_{\text{conserv}}$	decision
10%	1.64	$[0.57, 0.72]$	Reject
5%	1.96	$[0.56, 0.73]$	Reject
1%	2.76	$[0.52, 0.77]$	Reject
.1%	3.29	$[0.497, 0.79]$	Fail to reject
.01%	3.89	$[0.47, 0.82]$	Fail to reject

The value of α across which we switch from "reject" to "fail to reject" is called the p-value

4:56 / 8:56 Speed 1.0x

p-value

Definition

The (asymptotic) *p-value* of a test ψ is the smallest (asymptotic) level α at which ψ rejects H_0.

Golden rule

p-value $\leq \alpha \quad \Leftrightarrow \quad H_0$ is rejected by ψ, at the (asymptotic) level α.

p-value $> \alpha \Leftrightarrow H_0$ is not rejected by ψ, at the (asymptotic) level α.

Kiss example: we need to find α_0 such that $\bar{R}_n - \frac{q_{\alpha_0/2}}{2\sqrt{n}} = 0.5$

If $\bar{R}_n = .645$, $n = 124$ we get $q_{\alpha_0/2} = 3.23$. To find α_0:

$$\frac{\alpha_0}{2} = \mathbb{P}[Z > q_{\frac{\alpha_0}{2}}] = \mathbb{P}[Z > 3.23] = 1 - .9994 = 0.06\% \Rightarrow \alpha_0 = 0.12\%$$

where $Z \sim \mathcal{N}(0,1)$ and $\mathbb{P}(Z \leq 3.24) = 0.9994$ (read from table).

The evidence scale

- Statisticians, and more generally researchers, are used to communicating directly in terms of p-values rather than "reject/fail to reject at level..."
- The mental conversion is as follows:

p-value	evidence against H_0
$> 10\%$	almost none
$[5\%, 10\%]$	weak
$[1\%, 5\%]$	strong
$[.1\%, 1\%]$	very strong
$< .1\%$	undisputable