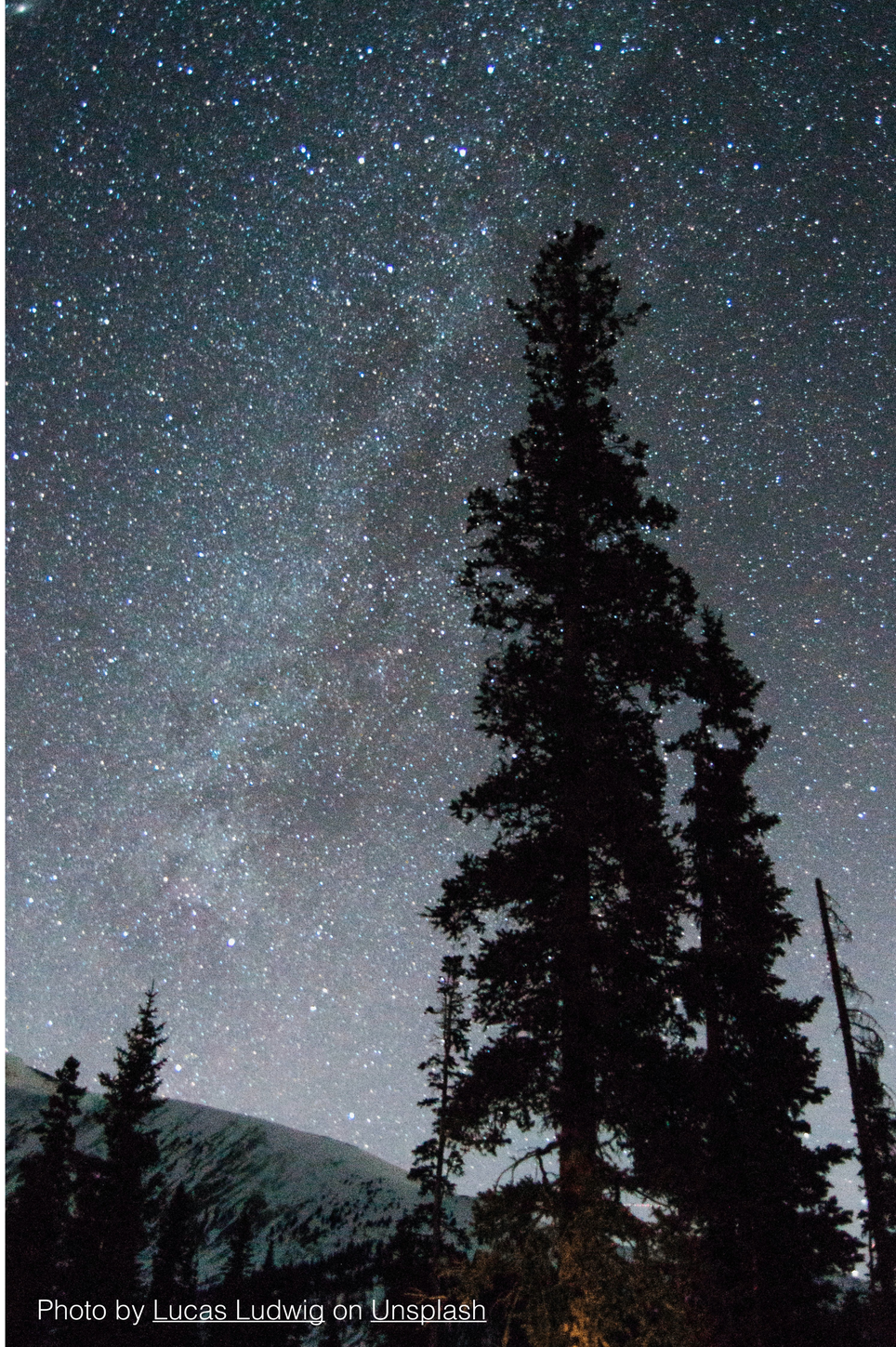
Unit #7 (a): Hypothesis Testing (One Sample Tests)

9.1, 9.2, 9.5, 9.7.1, 9.7.2, 9.9.2



At the end of this unit, students should be able to:

1. Define **research hypotheses** and **statistical hypotheses**.
2. Model research hypotheses using statistical hypotheses included in a statistical model.
3. Articulate the logic of hypothesis testing.
4. Define a test statistic and describe how it is used in a hypothesis test.
5. Perform hypothesis tests for means and proportions.
6. Derive hypothesis tests for other distributions/parameters (e.g., a test for λ from a Poisson distribution).
7. Define a rejection region, critical value, significance level, type I error, and type II error.
8. Articulate the tradeoff between the rate of type I and type II errors.
9. Describe the relationship between hypothesis testing and confidence intervals.
10. Define and properly interpret p-values. Identify some common misinterpretations.

Photo by Jan Huber on Unsplash

A **research**, **scientific**, or **business hypothesis** is…

A hypothesis is given in the language of the relevant discipline.

Examples:

1. Is the a difference in gene expression across people w/ asthma and w/ out asthma.
2.

Photo by Jan Huber on Unsplash

A **statistical hypothesis** is a claim about the value of a parameter in a statistical model.

$(\underline{X}, f(\underline{x};\theta))$

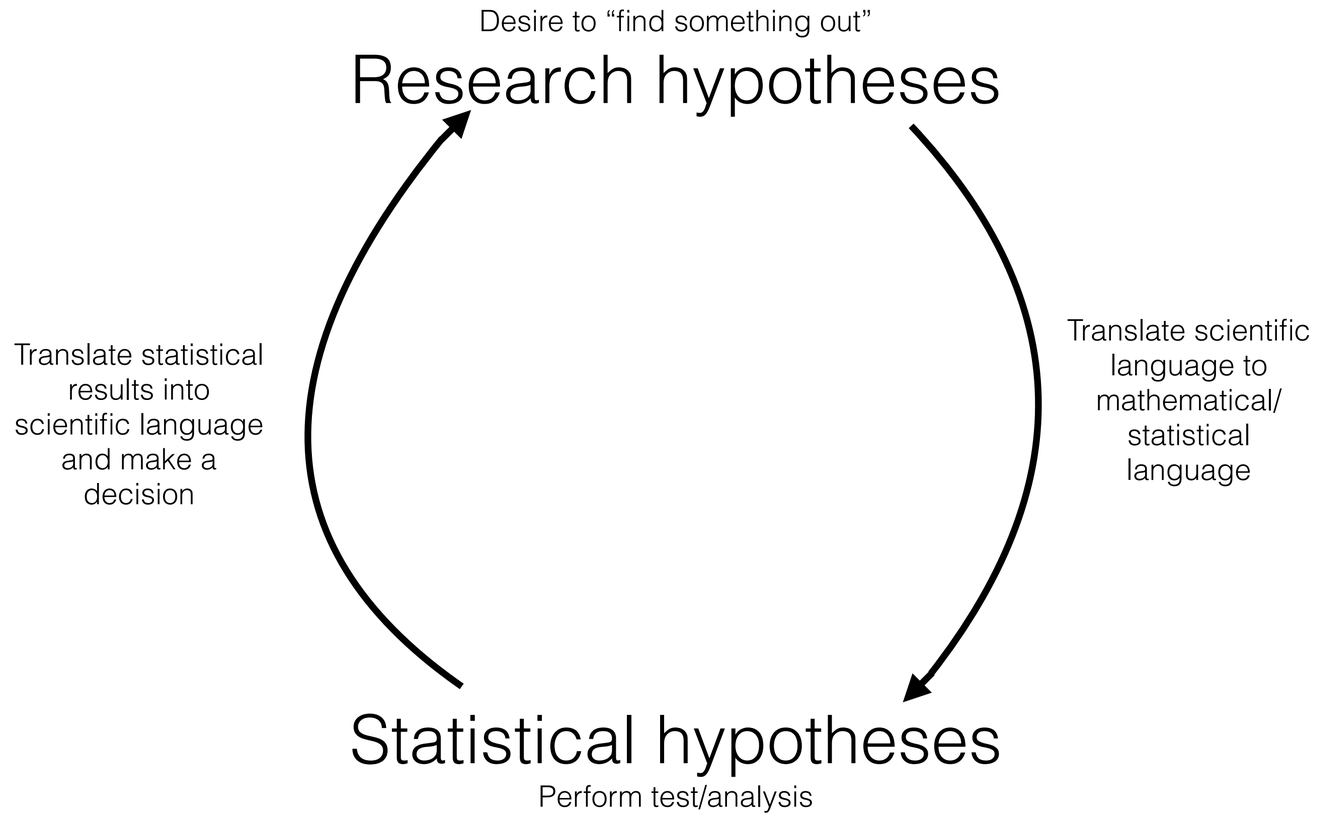
Note: parameters model population characteristics, e.g., quantities referred to in research hypotheses.

Examples:

1. $X_1, \ldots, X_n \overset{iid}{\sim} Bin(1, p)$. $H_0 : p = 0.5$ $H_1 : p > 0.5$
2. $Y_1, \ldots, Y_n \sim f(y;\theta)$. $H : \theta = \theta_0$

θ_0 is a constant

Photo by Jan Huber on Unsplash



Data, $\mathbf{x}$, are assumed to be realizations of the random variables, $\mathbf{X} = (X_1, \ldots, X_n)$, defined by a statistical model.

A statistical model will have a set of unknown **parameters**, or constants that define the probability distribution that generates the data.

Hypotheses are claims about parameters, e.g., $H : \mu = 5$.

In any hypothesis-testing problem, there are always two competing hypotheses under consideration:

(null) $H_0 : \mu = 5$

(alternative) $H_1 : \mu \neq 5$

The objective of **hypothesis testing** is to decide, based on sample information, if the alternative hypotheses is actually "supported" by the data.

Photo by Jan Huber on Unsplash

A (**test**) **statistic** is a function of random variables ($\mathbf{X}$) or data ($\mathbf{x}$).

In hypothesis testing, sample information is summarized by a **test statistic.** The value of the test statistic provides support (or lack thereof) for the alternative hypotheses.

If the statistical model is a reasonable model for the research area *then* the result of the statistical hypothesis test is applicable to the research question.


Photo by Tingey Injury Law Firm on Unsplash

Analogy: Jury in a criminal trial.

When a defendant is accused of a crime, the jury (is supposed to) presumes that she is not guilty (not guilty; that's the "null hypothesis").

Then, we gather evidence. If the evidence is seems implausible under the assumption of non-guilt, we might reject non-guilt and claim that the defendant is guilty.

Photo by Jan Huber on Unsplash

Important Question: Is there evidence for the alternative? The burden of proof is placed on those who support the alternative claim.

The initially favored claim, the null hypothesis H_0, will not be rejected in favor of the alternative hypothesis, H_1 (or H_A), unless the sample evidence provides a lot of support for the alternative.

The two possible conclusions: (for now)

1. There is some evidence for H_1

2. There is not evidence for H_1

Photo by Jan Huber on Unsplash

Why *assume* the null hypothesis?

Logic of Hypothesis Testing

$H_0 : \mu = 0$

$H_1 : \mu > 0$

		Your Decision	
		Fail to Reject H_0	Reject H_0
Reality	H_0 is true	No error!	**Type I error (false positive)** (rate : α)
	H_0 is false	**Type II error (false negative)** (rate : β)	no error!

For a simple null hypothesis, the **significance level** or **size** of the test is the probability of a Type I error.

$H_0: \theta = 1$

$H_0: \mu = 0$

For a composite null hypothesis, the **significance level** or **size** of the test is the max probability of a Type I error.

$H_1: \theta > 1$

$H_1: \mu < 0$

In both cases, the significance level is denoted by α:

$$\alpha = \max_{\theta \in H_0} P(\text{reject } H_0 \; ; \; \theta \in H_0)$$

in terms of test stat

$\mu > 0$

Photo by Mika Baumeister on Unsplash

Example: Suppose a company is considering putting a new type of coating on bearings that it produces.

$\mu_c = 1000$

The true average wear life with the current coating is known to be 1000 hours. With μ denoting the true average life for the new coating, the company would not want to make any (costly) changes unless evidence strongly suggested that μ exceeds 1000.

$H_0 : \mu \leq \mu_c = 1000$

$H_1 : \mu > \mu_c = 1000$

Photo by Mika Baumeister on Unsplash

An appropriate problem formulation would involve testing:

$$H_0 : \mu \leq \mu_c = 1000$$

$$H_1 : \mu > \mu_c = 1000$$

The conclusion that a change is justified is identified with H_1, and it would take conclusive evidence to justify rejecting H_0 and switching to the new coating.

Scientific research often involves trying to decide whether a current theory should be replaced, or "elaborated upon."

Photo by Mika Baumeister on Unsplash

The alternative to the null hypothesis $H_0 : \theta = \theta_0$ will look like one of the following three assertions:

$H_1 : \theta > \theta_0$ or

$H_1 : \theta < \theta_0$ or

$H_1 : \theta \neq \theta_0$

The equality sign is *always* with the null hypothesis.

The alternate hypothesis is the claim for which we are seeking statistical evidence.

Photo by Jan Huber on Unsplash

Definition: A *test statistic* is ~~a~~ a quantity derived based on sample data and calculated under the null hypothesis. It is used in a decision about whether to reject H_0.

We can think of a test statistic as our evidence. Next, we need to quantify whether we think our evidence is “rare” under the null hypothesis.

(Does rare under the null hypothesis imply not rare under some alternative?)

Example: Company A produces circuit boards, but $10\,\%$ of them are defective. Company B claims that they produce fewer defective circuit boards. What are the null and alternative hypotheses?

$H_0 : p_B \geq 0.1$

$H_1 : p_B < 0.1$

Our data is a random sample of $n = 200$ boards from company B. What test procedure (or rule) could we devise to decide if the null hypothesis should be rejected?

$x : 0, 0, 0, 1, 0, 0 \ldots 0, 1, 1$

$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \hat{p} < K$ then reject H_0

Photo by Jan Huber on Unsplash

Which test statistic is “best”?

There are an infinite number of *possible* tests that could be devised, so we have to limit this in some way.

We want test statistics built from estimators that have nice properties, and for which we can derive sampling distributions.

In the previous example, we might use:

$$\frac{X_1 + X_n}{2}$$

Photo by Jan Huber on Unsplash

How would we know when the test statistic is "sufficiently rare" under the null hypothesis such that we might regard the null as false?

We could define a **rejection region**—a range of values that leads a researcher to *reject* the null hypothesis.

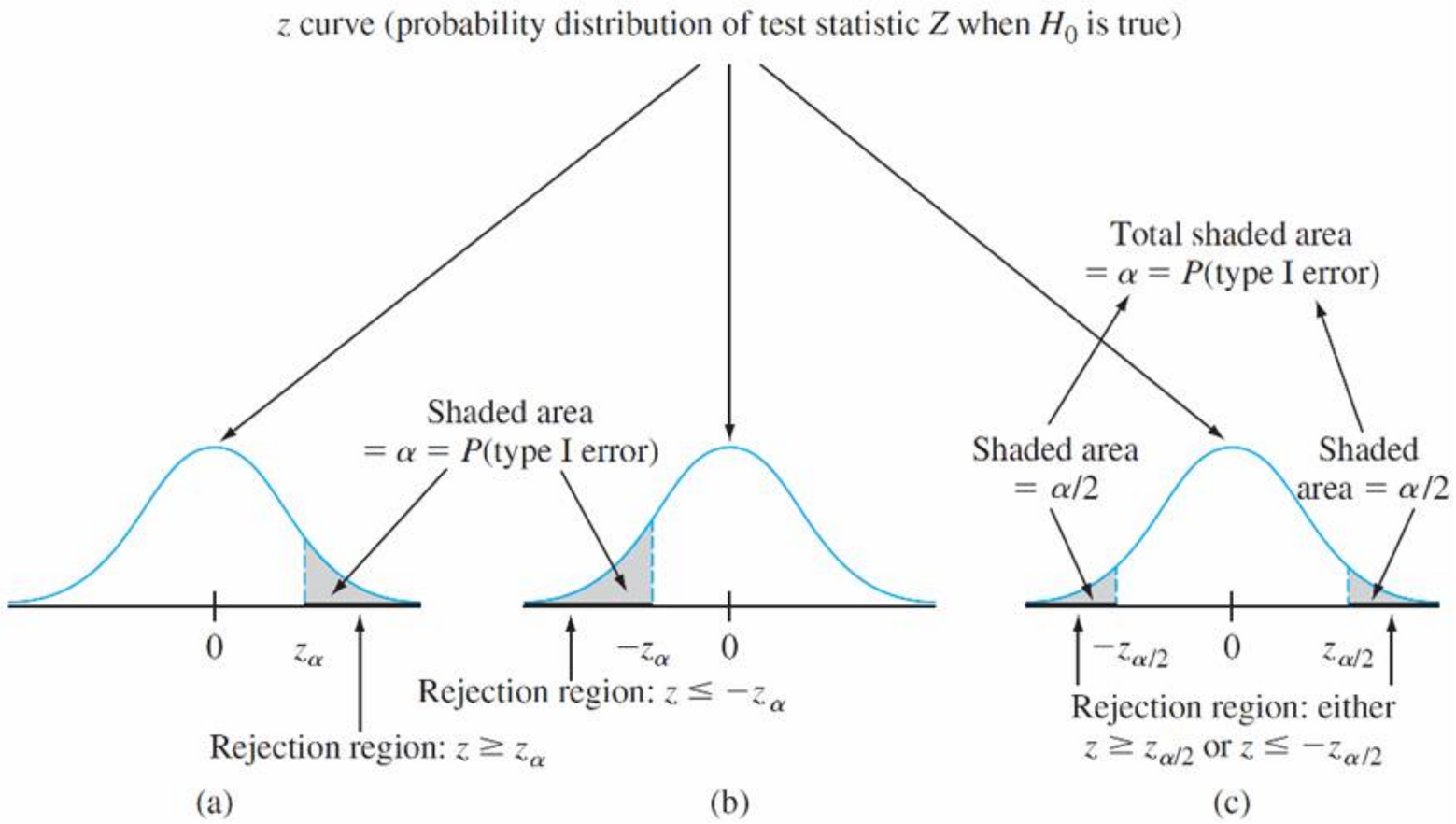
Ex.

- $\hat{p} < k$
- $\bar{x} > 5$
- $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} < z_\alpha$

$X_1, \ldots, X_n \overset{iid}{\sim} f(x; \theta)$; $H_0 : \theta = \theta_0$, $H_1 : \theta \neq \theta_0$; $T(\underline{X})$

$\underline{X}$

Rejection Regions



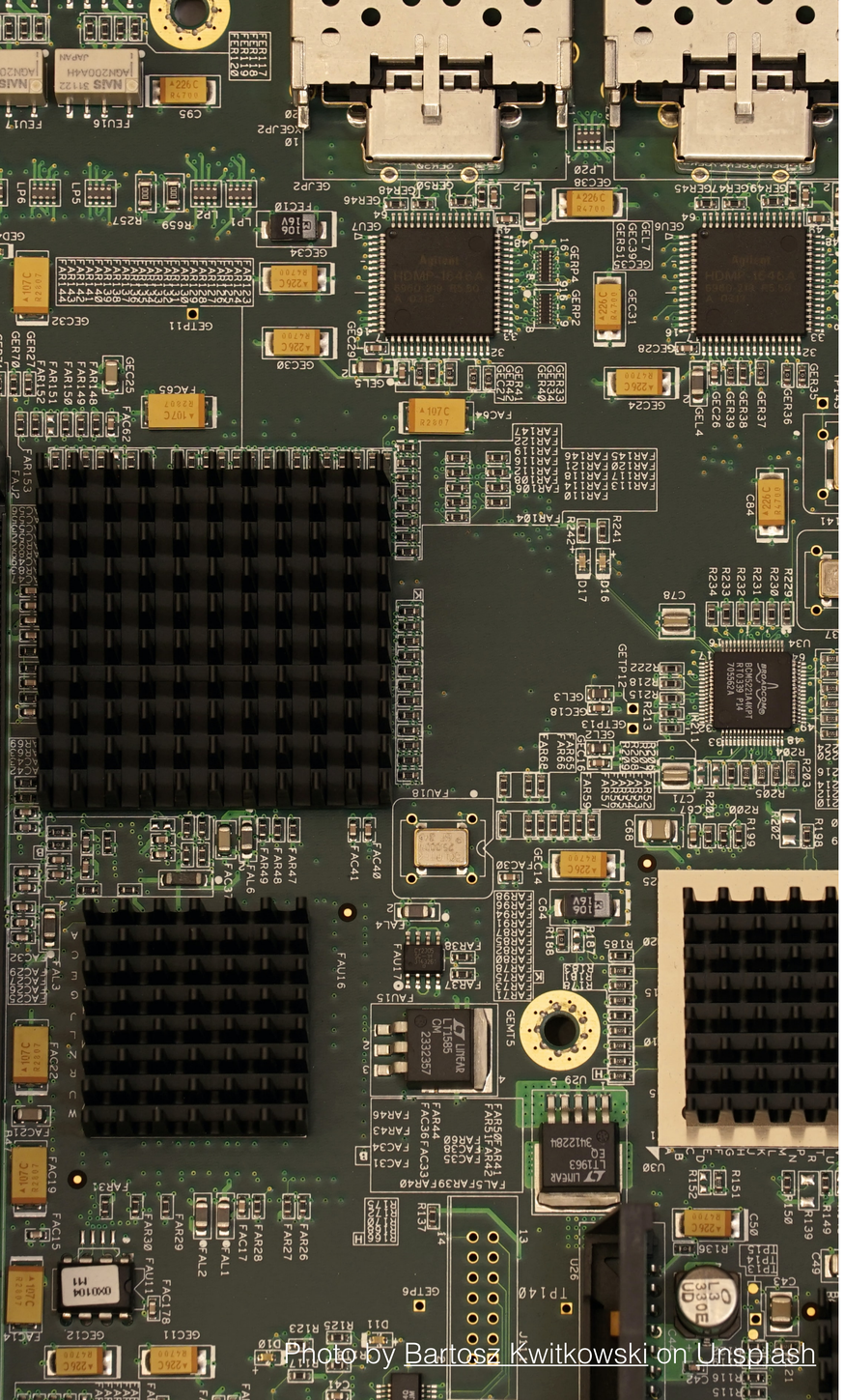


Photo by Bartosz Kwitkowski on Unsplash

Example (continued): Example: Company A produces circuit boards, but $10\,\%$ of them are defective. Company B claims that they produce fewer defective circuit boards. What are the null and alternative hypotheses?

$H_0: p_B \geq 0.1$

$H_1: p_B < 0.1$

p_{H_0}

$$Z = \frac{(\hat{p} - p_{H_0})}{\sqrt{p_{H_0}(1-p_{H_0})/n}} \overset{a}{\sim} N(0,1)$$

Suppose that in the sample of $n = 200$, 7 circuit boards from Company B were defective. Calculate the test statistic, and decide, based on a lower-tailed test with significance level $\alpha = 0.05$, whether we should reject the null hypothesis.

- $\hat{p} = \frac{7}{200} \approx 0.035 \Rightarrow Z = \frac{0.035 - 0.1}{\sqrt{\frac{(0.1)(0.9)}{200}}} \approx -3.06$ ($\hat{p} = \bar{x}$)
- RR: $Z < -Z_\alpha = -Z_{0.05} = -1.64$

So, $Z = z = -3.06 < -Z_\alpha = -1.64$. So z is in the RR. Thus, we have evidence against H_0.

Consider testing a claim about a population mean μ. Let $H_0 : \mu = \mu_0$ and assume that σ^2 is known.

- For $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$; or
- for a random sample $X_1, \ldots, X_n$ from a population with finite mean μ, finite variance σ^2, and large sample size (roughly $n > 30$):

we can use the test statistic

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1).$$

based on α

If $Z = z$ falls within the rejection region (defined by the alternative hypothesis), then we reject H_0 (or have some evidence against H_0). Otherwise, we fail to reject H_0.

Consider testing a claim about a population mean μ. Let $H_0 : \mu = \mu_0$ and assume σ^2 is unknown.

- For $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ where $n > 30$; or
- for a random sample $X_1, \ldots, X_n$ from a population with finite mean μ, finite variance σ^2, and large sample size (roughly $n > 30$):

we can use the test statistic

$$Z = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \overset{approx}{\sim} N(0,1).$$

If $Z = z$ falls within the rejection region (defined by the alternative hypothesis), then we reject H_0 (or have some evidence against H_0). Otherwise, we fail to reject H_0.

$\sum X_i \sim Bin(n,p)$. $\sum X_i > k \Rightarrow$ reject

Consider testing a claim about a population proportion p. Let $H_0 : p = p_0$.

For $X_1, \ldots, X_n \overset{iid}{\sim} Bin(1,p)$, where $np > 5$ and $np(1-p) > 5$, we can use the test statistic $Z = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}} \overset{a}{\sim} N(0,1)$.

If $Z = z$ falls within the rejection region (defined by the alternative hypothesis), then we reject H_0 (or have some evidence against H_0). Otherwise, we fail to reject H_0.

Photo by Jan Huber on Unsplash

The p-value measures the "extremeness" of the test statistic.

is a function of n

A **p-value** is the probability, *under the null hypothesis*, that we would get a test statistic *at least as extreme as the one we calculated.*

So, the smaller the p-value, the more evidence there is in the sample data against the null hypothesis (so the story goes…).

So what constitutes "sufficiently small" and "extreme enough" to make a decision about the null hypothesis?

Select a significance level α (as before, the desired *type I error probability*). Then α defines the rejection region. Then the decision rule is:

p-value $< \alpha$

$(\Leftrightarrow z \text{ in } RR)$

Thus if the *p*-value exceeds the chosen significance level, the null hypothesis cannot be rejected at that level.

Note, the *p*-value can be thought of as the <u>smallest significance level</u> at which H_0 can be rejected.

The p-value measures the "extremeness" of the test statistic.

Note:

$P(Z > z_0 ; H_0)$

1. This probability is calculated assuming that the null hypothesis is true.

2. Beware: The *p*-value is not the probability that H_0 is true, nor is it an error probability!

↳ ~~$P(H_0 | Z)$~~

3. The *p*-value does not provide information on the "effect size". That is, small *p*-values do not mean large deviations from the null.

$H_0 : \mu = 0$

$H_1 : \mu > 0$

$p = 0.000001$

The calculation of the *p*-value depends on whether the test is upper-, lower-, or two-tailed. If we have a normal test with test statistic z:

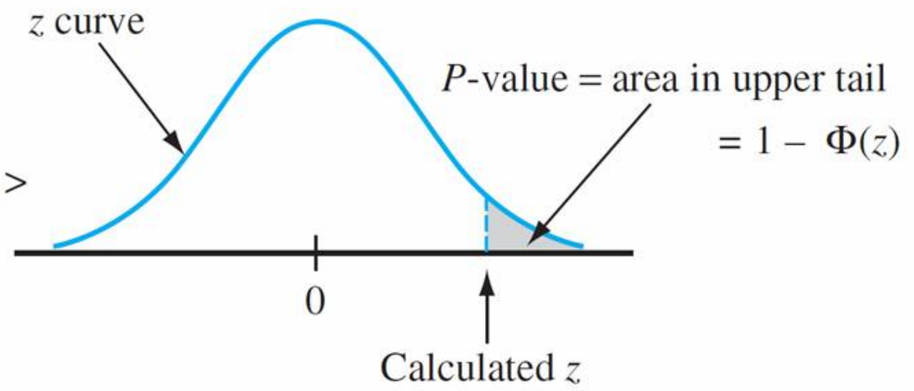
$$p\text{-value} = \begin{cases} 1 - \text{pnorm}(z), & \text{if } H_1 : \mu > \mu_0 \\ \text{pnorm}(z) & \text{if } H_1 : \mu < \mu_0 \\ 2\left(1 - \text{pnorm}(|z|)\right) & \text{if } H_1 : \mu \neq \mu_0 \end{cases}$$

Each of these is the probability of getting a value at least as extreme as what was obtained (assuming H_0 true).

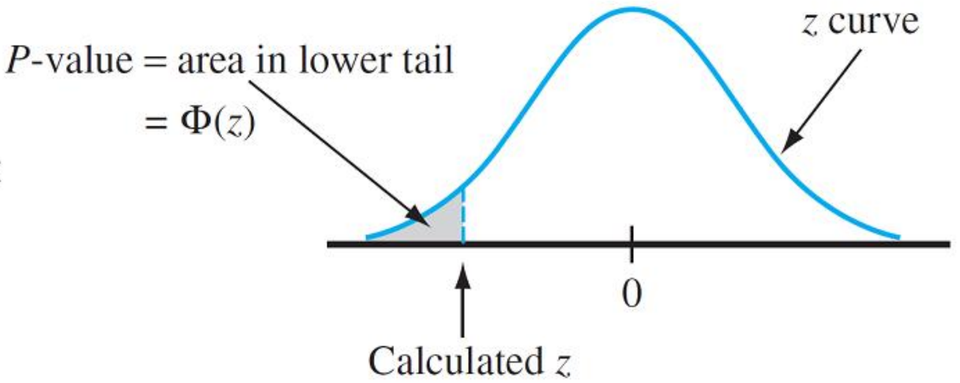
$$H_0 : \mu = \mu_0$$

P-Values for Z Tests

1. **Upper-tailed test**
   H_a contains the inequality >



2. **Lower-tailed test**
   H_a contains the inequality <



P-Values for Z Tests

3. **Two-tailed test**

   H_a contains the inequality $\neq$

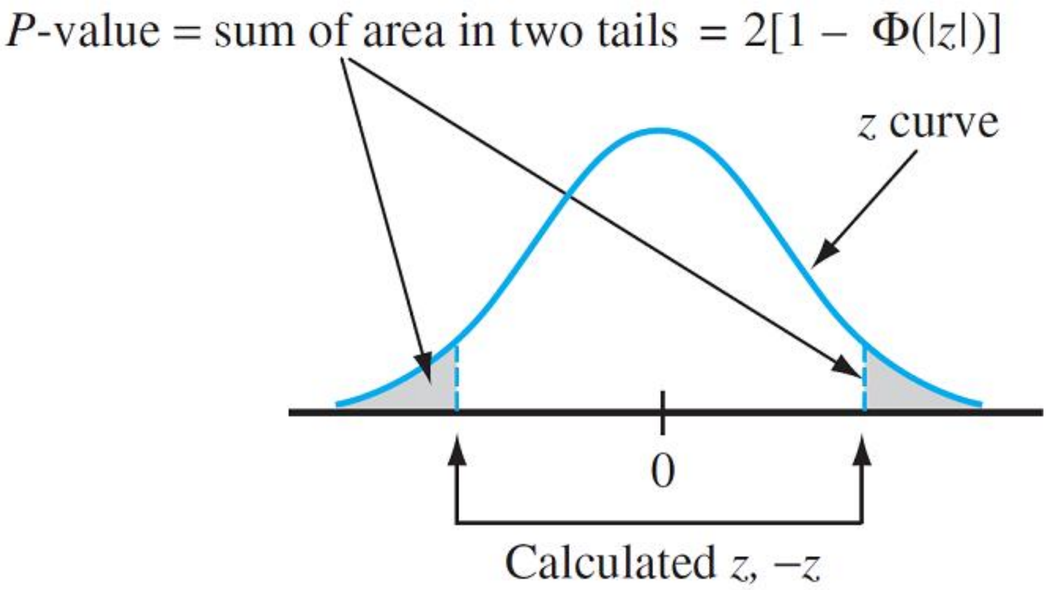


Photo by Brandon Morgan on Unsplash

$P(\text{reject } H_0 ; H_1)$

The **power** of a test is the probability it will reject H_0 when H_0 is false.

That is, the power is the probability that we *correctly reject* H_0.

$\text{power} = 1 - \beta$

β: Pr. of type II error

Photo by Brandon Morgan on Unsplash

Example: Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, 1)$. Let's calculate the power for a size $\alpha = 0.05$ test of $H_0 : \mu = 0 \quad vs \quad H_1 : \mu > 0$, for the specific value of $0.8 \in H_1$. Let $n = 5$.

$$Z = \frac{\bar{X} - \mu_0}{1/\sqrt{n}} = \sqrt{n}\,\bar{X} \; ; \; z_\alpha = z_{0.05} = 1.64$$

power function

$$\gamma(0.8) = P(\text{reject } H_0 \; ; H_1 : \mu = 0.8)$$

$$= P(\sqrt{n}\bar{X} > 1.64 \; ; H_1 : \mu = 0.8)$$

$$= P(\bar{X} > 1.64/\sqrt{5} \; ; \; H_1 : \mu = 0.8)$$

$$\bar{X} \sim N\left(0.8, \frac{1}{5}\right)$$

$$= P(\bar{X} > 0.73 \; ; H_1 : \mu = 0.8)$$

$$= 1 - \text{pnorm}\left(0.73, 0.8, \frac{1}{\sqrt{5}}\right) \approx 0.562$$

Photo by Jan Huber on Unsplash

Consider testing a claim about a population mean μ. Let $H_0 : \mu = \mu_0$ and assume σ^2 is unknown.

- For $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ where $n < 30$:

we can use the test statistic

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}.$$

If $T = t$ falls within the rejection region (defined by the alternative hypothesis), then we reject H_0 (or have some evidence against H_0). Otherwise, we fail to reject H_0.