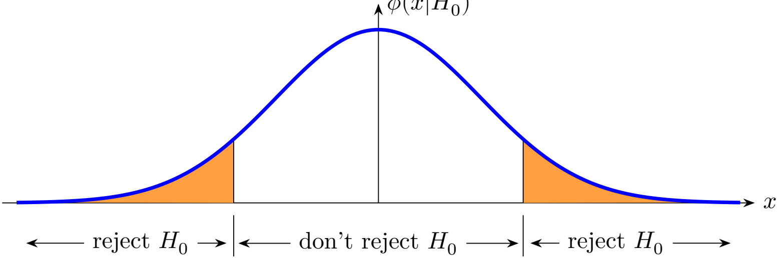
# Null Hypothesis Significance Testing
# $p$-values, significance level, power, $t$-tests
# 18.05 Spring 2022



# Announcements/Agenda

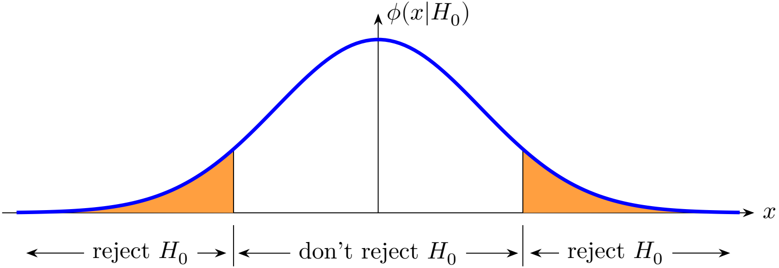
## Announcements

- Studio 6: In R studio 6: Most people hardwired the value of where to look for the secret path to be the value found from the test data. The grading data was different, so produced a different value.
- Next pset due on Tuesday, April 19

## Agenda

- Simple and compund hypotheses
- p-values and extreme data
- Critical values
- Errors, significance, power
- t-tests

# Understand this figure




- $x =$ test statistic
- $\phi(x|H_0) =$ pdf of null distribution = blue curve
- Rejection region is a portion of the $x$-axis.
- Significance = probability of rejection = orange shaded area.

# Simple and composite hypotheses

**Simple hypothesis**: the sampling distribution is fully specified. Usually the parameter of interest has a specific value.

**Composite hypotheses**: the sampling distribution is not fully specified. Usually the parameter of interest has a range of values.

**Example.** A coin has probability $\theta$ of heads. Toss it 30 times and let $x$ be the number of heads.

(i) $H$: $\theta = 0.4$ is simple. $x \sim$ binomial$(30, 0.4)$.

(ii) $H$: $\theta > 0.4$ is composite. $x \sim$ binomial$(30, \theta)$ depends on which value of $\theta$ is chosen.

# Extreme data and $p$-values

**Hypotheses:** $H_0$, $H_A$.

**Test statistic:** value: $x$, computed from data, random.

**Null distribution:** $\phi(x|H_0)$ (assumes null hypothesis is true)

**Sides:** $H_A$ determines if the rejection region is one or two-sided.

**Rejection region/Significance:** $P(x \text{ in rejection region} \,|\, H_0) = \alpha$.

---

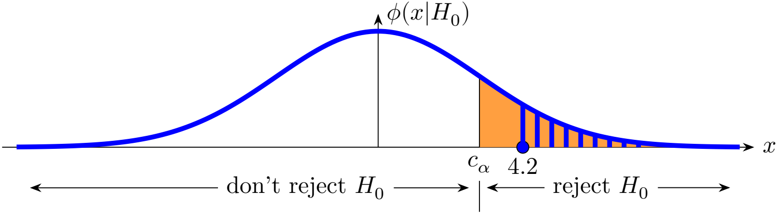
The $p$-value is a computational tool to check if the test statistic is in the rejection region. It is also a measure of the evidence for rejecting $H_0$.

**p-value:** $P(\textbf{data at least as extreme as } x \,|\, H_0)$

**"Data at least as extreme":** determined by the sidedness of the rejection region.

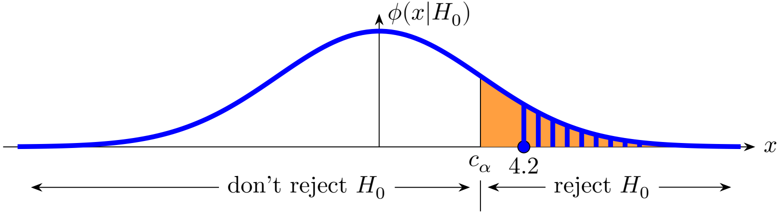
# Extreme data and $p$-values

**Example.** Suppose we have the right-sided rejection region shown below. Also suppose we see data with test statistic $x = 4.2$. Should we reject $H_0$?



# Extreme data and $p$-values

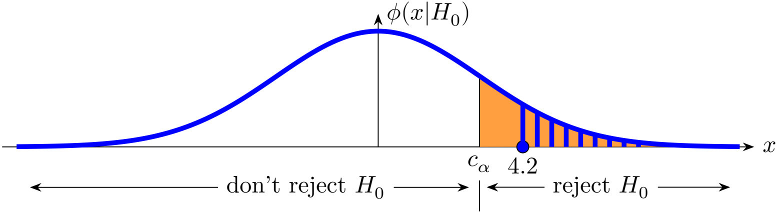
**Example.** Suppose we have the right-sided rejection region shown below. Also suppose we see data with test statistic $x = 4.2$. Should we reject $H_0$?




**Solution:** The test statistic is in the rejection region, so reject $H_0$.

# Extreme data and $p$-values

**Example.** Suppose we have the right-sided rejection region shown below. Also suppose we see data with test statistic $x = 4.2$. Should we reject $H_0$?




**Solution:** The test statistic is in the rejection region, so reject $H_0$.

**Alternatively:** blue striped area $<$ orange shaded area

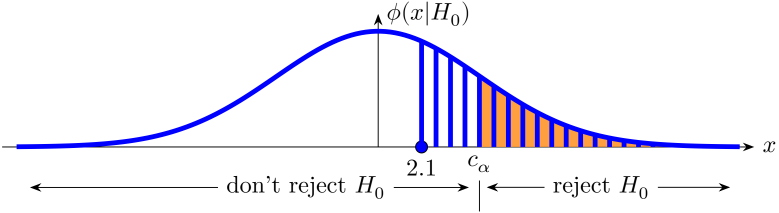
Significance: $\alpha = P(x \text{ in rejection region} \mid H_0) =$ orange shaded area.

p-value: $p = P(\text{data at least as extreme as } x \mid H_0) =$ blue striped area.

Since $p < \alpha$ we reject $H_0$.

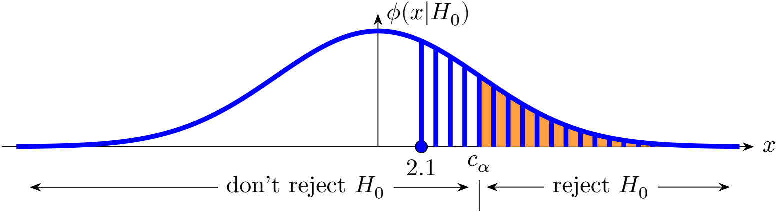
# Extreme data and $p$-values

**Example.** Now suppose $x = 2.1$ as shown. Should we reject $H_0$?



# Extreme data and $p$-values

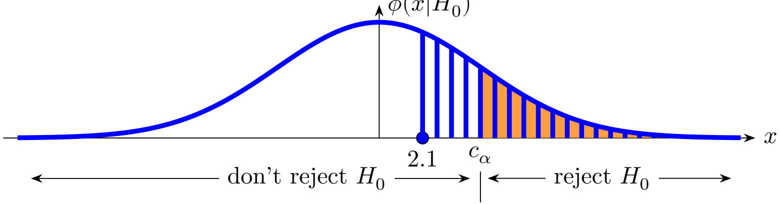
**Example.** Now suppose $x = 2.1$ as shown. Should we reject $H_0$?




**Solution:** Test statistic not in the rejection region: don't reject $H_0$.

# Extreme data and $p$-values

**Example.** Now suppose $x = 2.1$ as shown. Should we reject $H_0$?




**Solution:** Test statistic not in the rejection region: don't reject $H_0$.

**Alternatively:** blue area $>$ orange shaded area

Significance: $\alpha = P(x \text{ in rejection region} \mid H_0) =$ orange shaded area.

p-value: $p = P(\text{data at least as extreme as } x \mid H_0) =$ blue striped area.

Since $p > \alpha$ we don't reject $H_0$.

# Critical values

- The boundaries of the rejection region are called critical values.
- Critical values are labeled by the probability to their right.
- They are complementary to quantiles: e.g., $c_{0.1} = q_{0.9}$
- Example: for a standard normal $c_{0.025} = 1.96$ and $c_{0.975} = -1.96$.
  For standard normal we will usually use $z_{0.025}$ instead of $c_{0.025}$.
- In R, for a standard normal $c_{0.025}$ `= qnorm(0.975)`.

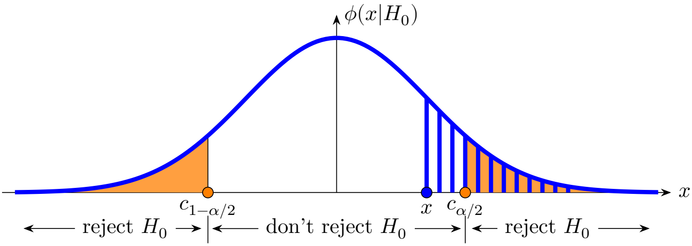
# Two-sided $p$-values

These are trickier: what does 'at least as extreme' mean in this case?

Remember the $p$-value is a tool for deciding if the test statistic is in the region.

Best to look at each test individually. Here is a somewhat general rule: If the **rejection region is equally split between left and right tails** then

$$p = 2\text{min}(\text{left tail prob. of } x, \text{ right tail prob. of } x)$$




$x$ is outside the rejection region, so $p > \alpha$: do not reject $H_0$

# Concept question: NHST

You collect data from an experiment and do a left-sided $z$-test with significance 0.1. You find the $z$-value is 1.8

**(i)** Which of the following computes the critical value for the rejection region?

(a) `pnorm(0.1, 0, 1)`
(b) `pnorm(0.9, 0, 1)`
(c) `pnorm(0.95, 0, 1)`
(d) `pnorm(1.8, 0, 1)`
(e) `1 - pnorm(1.8, 0, 1)`
(f) `qnorm(0.05, 0, 1)`
(g) `qnorm(0.1, 0, 1)`
(h) `qnorm(0.9, 0, 1)`
(i) `qnorm(0.95, 0, 1)`

# Concept question: NHST

You collect data from an experiment and do a left-sided $z$-test with significance 0.1. You find the $z$-value is 1.8

**(i)** Which of the following computes the critical value for the rejection region?

(a) `pnorm(0.1, 0, 1)`
(b) `pnorm(0.9, 0, 1)`
(c) `pnorm(0.95, 0, 1)`
(d) `pnorm(1.8, 0, 1)`
(e) `1 - pnorm(1.8, 0, 1)`
(f) `qnorm(0.05, 0, 1)`
(g) `qnorm(0.1, 0, 1)`
(h) `qnorm(0.9, 0, 1)`
(i) `qnorm(0.95, 0, 1)`

**(ii)** Which of the above computes the $p$-value for this experiment?

# Concept question: NHST

You collect data from an experiment and do a left-sided $z$-test with significance 0.1. You find the $z$-value is 1.8

**(i)** Which of the following computes the critical value for the rejection region?

(a) `pnorm(0.1, 0, 1)`
(b) `pnorm(0.9, 0, 1)`
(c) `pnorm(0.95, 0, 1)`
(d) `pnorm(1.8, 0, 1)`
(e) `1 - pnorm(1.8, 0, 1)`
(f) `qnorm(0.05, 0, 1)`
(g) `qnorm(0.1, 0, 1)`
(h) `qnorm(0.9, 0, 1)`
(i) `qnorm(0.95, 0, 1)`

**(ii)** Which of the above computes the $p$-value for this experiment?

**(iii)** Should you reject the null hypothesis?

(a) Yes (b) No

# Error, significance and power

| | | True state of nature | |
|---|---|---|---|
| | | $H_0$ is true | $H_A$ is true |
| Our decision | Reject $H_0$ | Type I error | correct decision |
| | Don't reject $H_0$ | correct decision | Type II error |

$$\begin{aligned}\text{Significance level} &= P(\text{type I error}) \\ &= \text{probability we incorrectly reject } H_0 \\ &= P(\text{test statistic in rejection region} \mid H_0) \\ &= P(\text{false positive})\end{aligned}$$

$$\begin{aligned}\text{Power} &= \text{probability we correctly reject } H_0 \\ &= P(\text{test statistic in rejection region} \mid H_A) \\ &= 1 - P(\text{type II error}) \\ &= P(\text{true positive})\end{aligned}$$

- $H_A$ determines the power of the test.
- Significance and power are both probabilities of the rejection region.
- Want significance level near 0 and power near 1.

# Table question: significance level and power

Our data $x$ follows a binomial($\theta$, 10) distribution with $\theta$ unknown.

The rejection region is boxed in orange. The corresponding probabilities for different hypotheses are shaded below it.

| $x$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $H_0 : p(x\|\theta = 0.5)$ | 0.001 | 0.010 | 0.044 | 0.117 | 0.205 | 0.246 | 0.205 | 0.117 | 0.044 | 0.010 | .001 |
| $H_A : p(x\|\theta = 0.6)$ | 0.000 | 0.002 | 0.011 | 0.042 | 0.111 | 0.201 | 0.251 | 0.215 | 0.121 | 0.040 | 0.006 |
| $H_A : p(x\|\theta = 0.7)$ | 0.000 | 0.000 | 0.001 | 0.009 | 0.037 | 0.103 | 0.200 | 0.267 | 0.233 | 0.121 | 0.028 |

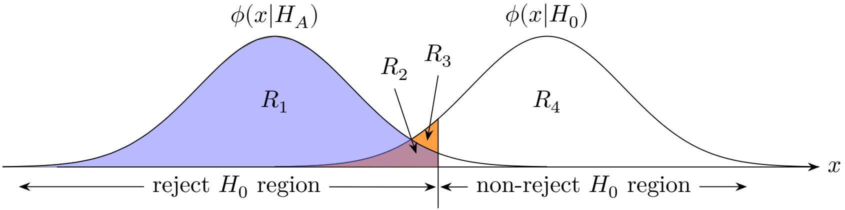
**(a)** Find the significance level of the test.

**(b)** Find the power of the test for each of the two alternative hypotheses.

**(c)** What is the probability of a type I error? type II?

# Concept question: Power

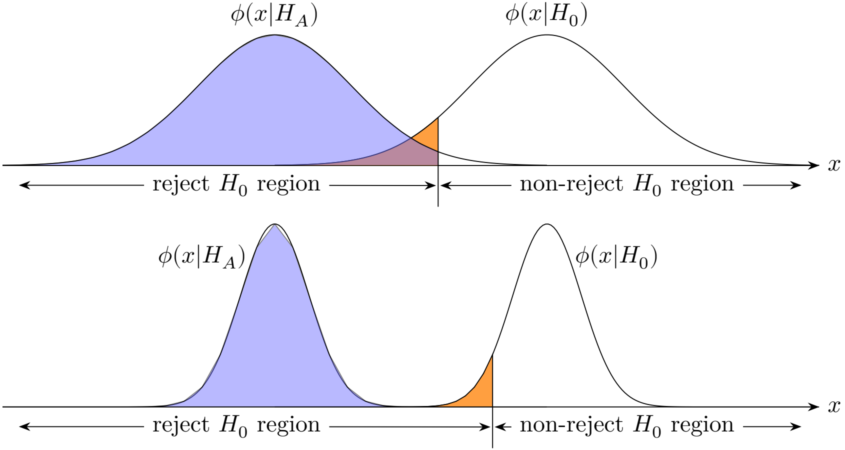
The power of the test in the graph is given by the area of




(a) $R_1$ (b) $R_2$ (c) $R_1 + R_2$ (d) $R_1 + R_2 + R_3$

# Concept question: Higher power

Which of the tests below has higher power?




(1) Top graph (2) Bottom graph

# Discussion question: significance and power

The null distribution for test statistic $x$ is $N(4, 8^2)$. The rejection region is $\{x \geq 20\}$.

What is the significance level and power of this test?

(Full solution posted with solutions to today's problems.)

# $z$-test

- Data: $x_1, x_2, \dots, x_n$.
- Assume $x_i$ are independently drawn from $N(\mu, \sigma^2)$.
- Called a sample.
- Null hypothesis: $\mu = \mu_0$ for some specific value $\mu_0$.
- $z$-test: $\mu$ unknown, $\sigma$ **known**.
- Test statistic (standardized mean): $z = \dfrac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$
- Null distribution $z \sim N(0, 1)$.

# One-sample $t$-test

- Data: $x_1, x_2, \ldots, x_n$.
- Assume $x_i$ are independently drawn from $N(\mu, \sigma^2)$.
- Null hypothesis: $\mu = \mu_0$ for some specific value $\mu_0$.
- $t$-test: $\mu$ unknown, $\sigma$ **unknown**.
- Test statistic (Studentized mean):

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}, \text{ where } s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

  $s^2$ is the *sample variance*.

- Null distribution: $\phi(t \mid H_0)$ is the pdf of $T \sim t(n-1)$, the $t$ distribution with $n-1$ degrees of freedom.

# Board question: $z$ and one-sample $t$-test

For both problems use significance level $\alpha = 0.05$.

Assume the data 2, 4, 4, 10 are independently drawn from a $N(\mu, \sigma^2)$.

The hypotheses are: $H_0$: $\mu = 0$ and $H_A$: $\mu \neq 0$.

**(a)** Is the test one or two-sided? If one-sided, which side?

**(b)** Assume $\sigma^2 = 16$ is known and test $H_0$ against $H_A$.

**(c)** Now assume $\sigma^2$ is unknown and test $H_0$ against $H_A$.

## Two-sample $t$-test: equal variances

- Data: $x_1, \dots, x_n \quad y_1, \dots, y_m$
- Assume $x_i$ are independently drawn from $N(\mu_x, \sigma^2)$.
- Assume $y_j$ are independently drawn from $N(\mu_y, \sigma^2)$. (Same $\sigma$)
- Assume $x_i$ and $y_j$ are independent.
- Null hypothesis $H_0$: $\mu_x = \mu_y$.
- $t$-test: $\mu_x$, $\mu_y$, $\sigma$ all unknown.
- Pooled variance: $s_p^2 = \dfrac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}\left(\dfrac{1}{n} + \dfrac{1}{m}\right)$.
- Test statistic: $t = \dfrac{\bar{x} - \bar{y}}{s_p}$
- Null distribution: $\phi(t \,|\, H_0)$ is the pdf of $T \sim t(n+m-2)$
- In general (so we can compute power) we have

$$\frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{s_p} \sim t(n+m-2)$$

- Note: there are more general formulas for unequal variances.

# Board question: two-sample $t$-test

Real data from 1408 women admitted to a maternity hospital for (i) medical reasons or through (ii) unbooked emergency admission. The duration of pregnancy is measured in complete weeks from the beginning of the last menstrual period.

Medical: 775 obs. with $\bar{x} = 39.08$ and $s^2 = 7.77$.

Emergency: 633 obs. with $\bar{x} = 39.60$ and $s^2 = 4.95$

**(a)** Set up and run a two-sample $t$-test to investigate whether the duration differs for the two groups.

**(b)** What assumptions did you make?

# Class discussion: Type I errors Q1

Suppose a journal will only publish results that are statistically significant at the 0.05 level. What percentage of the papers it publishes contain type I errors?

# Class discussion: Type I errors Q2

Jerry desperately wants to cure diseases but he is terrible at designing effective treatments. He is however a careful scientist and statistician, so he randomly divides his patients into control and treatment groups. The control group gets a placebo and the treatment group gets the experimental treatment. His null hypothesis $H_0$ is that the treatment is no better than the placebo. He uses a significance level of $\alpha = 0.05$. If his $p$-value is less than $\alpha$ he publishes a paper claiming the treatment is significantly better than a placebo.

**(a)** Since his treatments are never, in fact, effective what percentage of his experiments result in published papers?

**(b)** What percentage of his published papers contain type I errors, i.e. describe treatments that are no better than placebo?

# Class discussion: Type I errors: Q3

Jen is a genius at designing treatments, so all of her proposed treatments are effective. She is also a careful scientist and statistician, so she too runs double-blind, placebo controlled, randomized studies. Her null hypothesis is always that the new treatment is no better than the placebo. She also uses a significance level of $\alpha = 0.05$ and publishes a paper if $p < \alpha$.

**(a)** How could you determine what percentage of her experiments result in publications?

**(b)** What percentage of her published papers contain type I errors, i.e. describe treatments that are, in fact, no better than placebo?