Null Hypothesis Significance Testing
p-values, significance level, power, t-tests

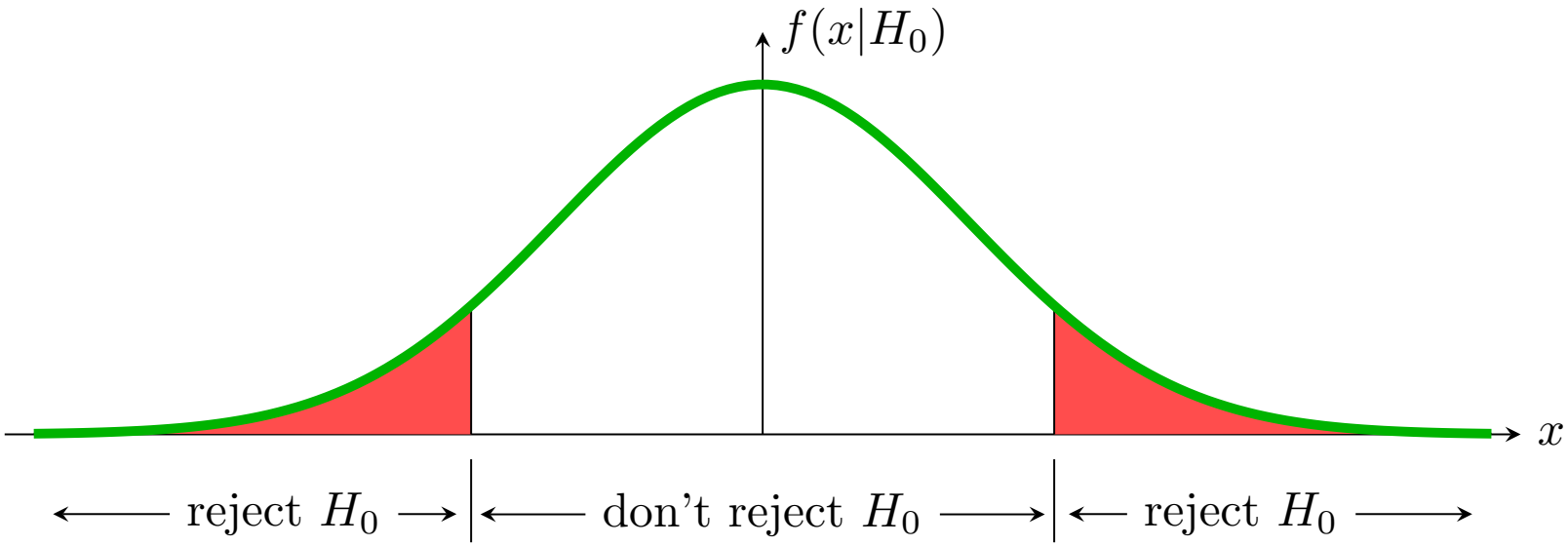
18.05 Spring 2018

NO CLASS Monday April 16 (Patriots' Day)

Problem set due Wednesday April 18

Watch class web site for RESCHEDULED OFFICE HOURS

Understand this figure



- x = test statistic
- $f(x|H_0)$ = pdf of null distribution = green curve
- Rejection region is a portion of the x-axis.
- Significance = probability over the rejection region = red 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 θ of heads. Toss it 30 times and let x be the number of heads.

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

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

Extreme data and p-values

Hypotheses: H_0, H_A.

Test statistic: value: x, computed from data.

Null distribution: $f(x|H_0)$ (assumes null hypothesis is true)

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

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

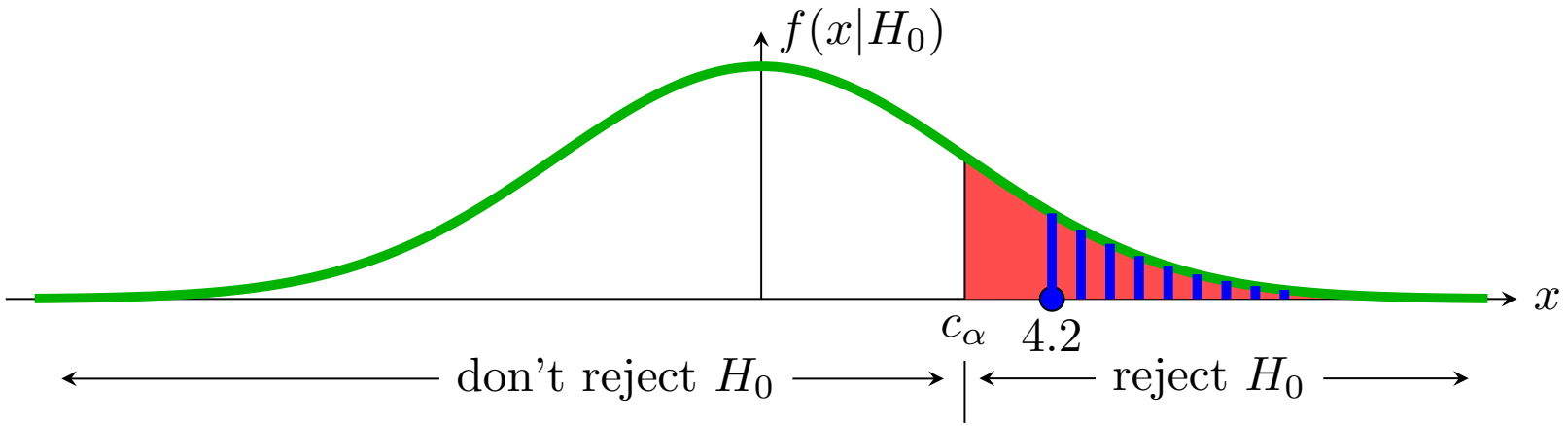
The p-value is a tool to check if the test statistic is in the rejection region. It is also a measure of the evidence for rejecting H_0.

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

"Data at least as extreme" is defined 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?



answer: The test statistic is in the rejection region, so reject H_0.

Alternatively: blue area $<$ red area

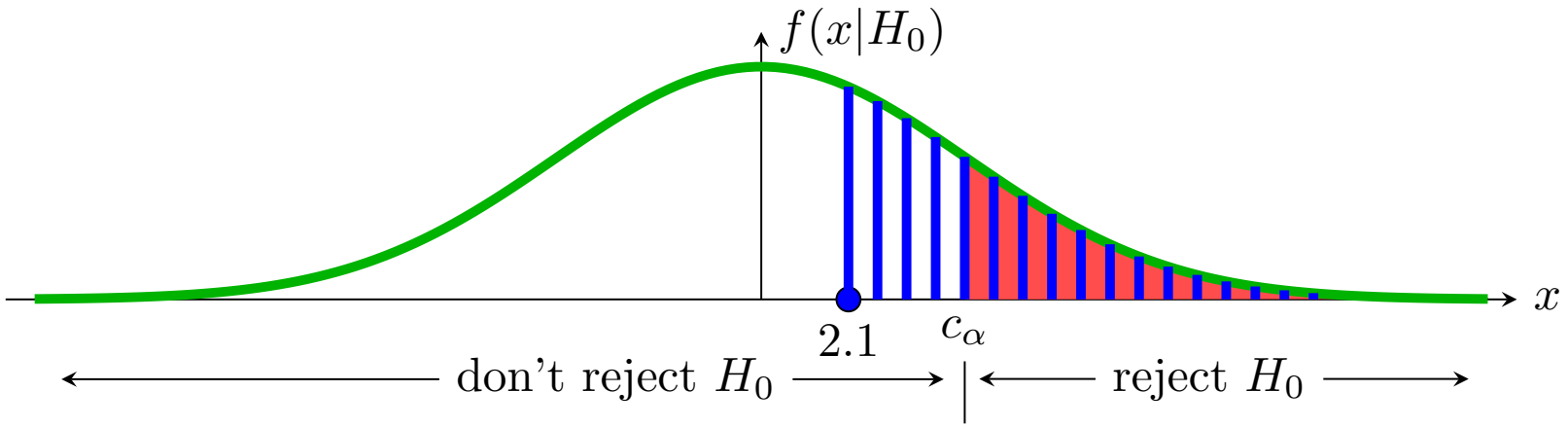
Significance: $\alpha = P(x \text{ in rejection region} \mid H_0) =$ red area.

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



answer: Test statistic not in the rejection region: don't reject H_0.

Alternatively: blue area $>$ red area

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

p-value: $p = P(\text{data at least as extreme as } x \mid H_0) =$ blue 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: $c_p = q_{1-p}$.
- Example: for a standard normal $c_{0.025} = 1.96$ and $c_{0.975} = -1.96$.
- In R, for a standard normal $c_{0.025}$ = `qnorm(0.975)`.

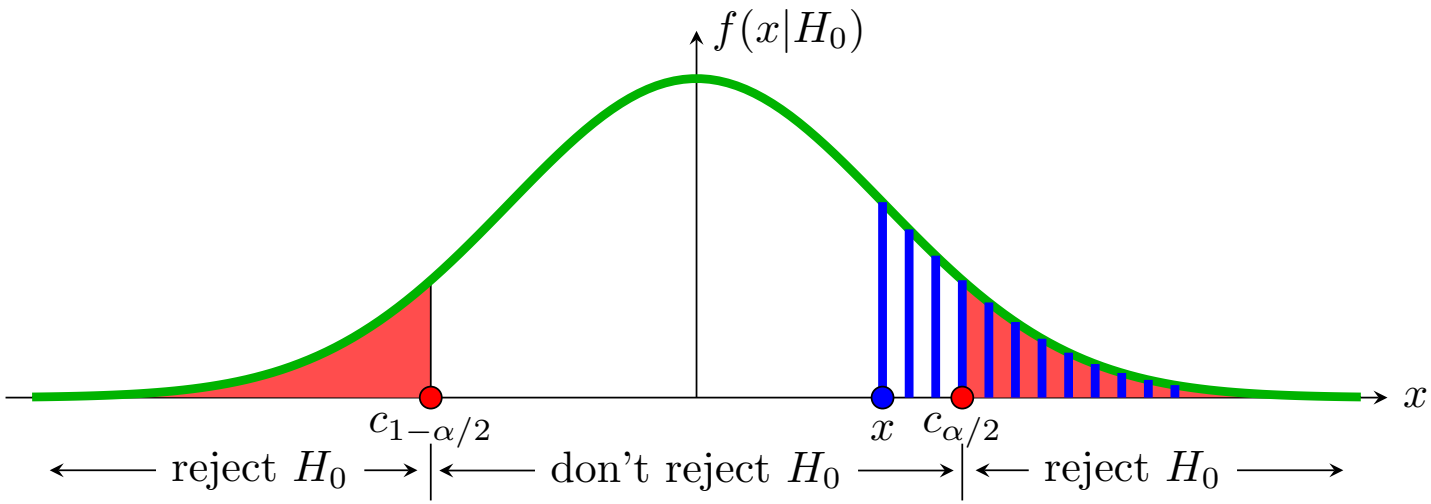
Two-sided p-values

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

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

If the null distribution is symmetric around zero 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

1. 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

answer: **(i)** g. **(ii)** d. **(iii)** No. (Draw a picture!)

Error, significance and power: a tale of a President

Our decision		H_0 is true	H_A is true
	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

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

x	0	1	2	3	4	5	6	7	8	9	10
$H_0: p(x\|\theta = 0.5)$	.001	.010	.044	.117	.205	.246	.205	.117	.044	.010	.001
$H_A: p(x\|\theta = 0.6)$	.000	.002	.011	.042	.111	.201	.251	.215	.121	.040	.006
$H_A: p(x\|\theta = 0.7)$	.000	.0001	.001	.009	.037	.103	.200	.267	.233	.121	.028

1. Find the significance level of the test.

2. Find the power of the test for each of the two alternative hypotheses.

answer:

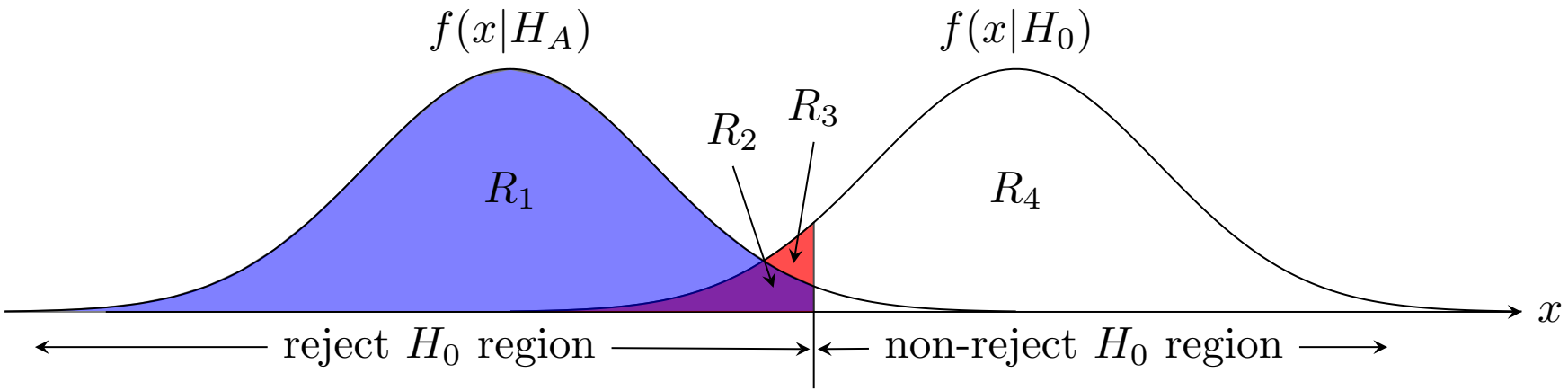
1. Significance level $= P(x \text{ in rejection region} \mid H_0) = 0.11$

2. $\theta = 0.6$: power $= P(x \text{ in rejection region} \mid H_A) = 0.18$

$\theta = 0.7$: power $= P(x \text{ in rejection region} \mid H_A) = 0.383$

Concept question

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

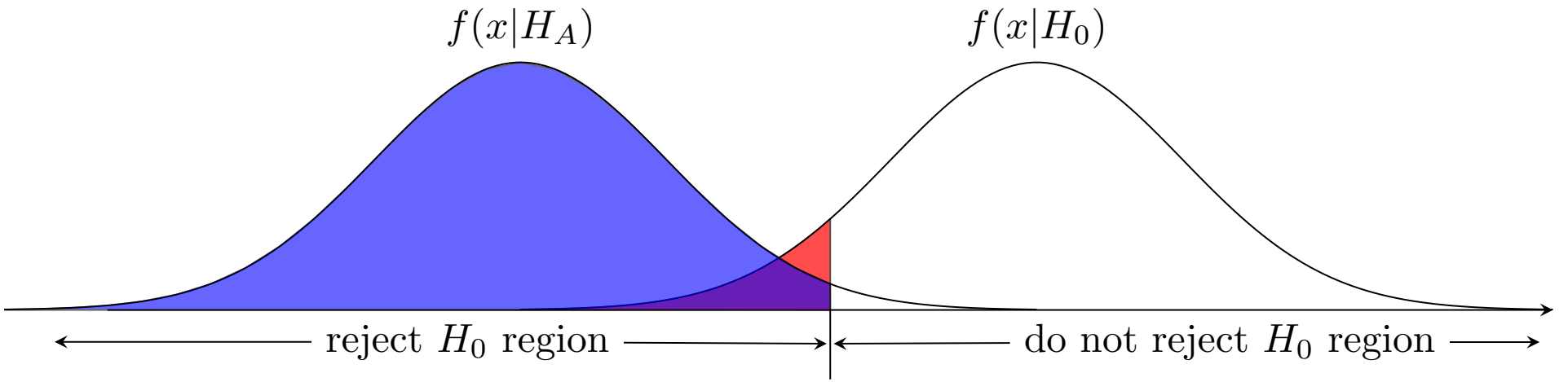


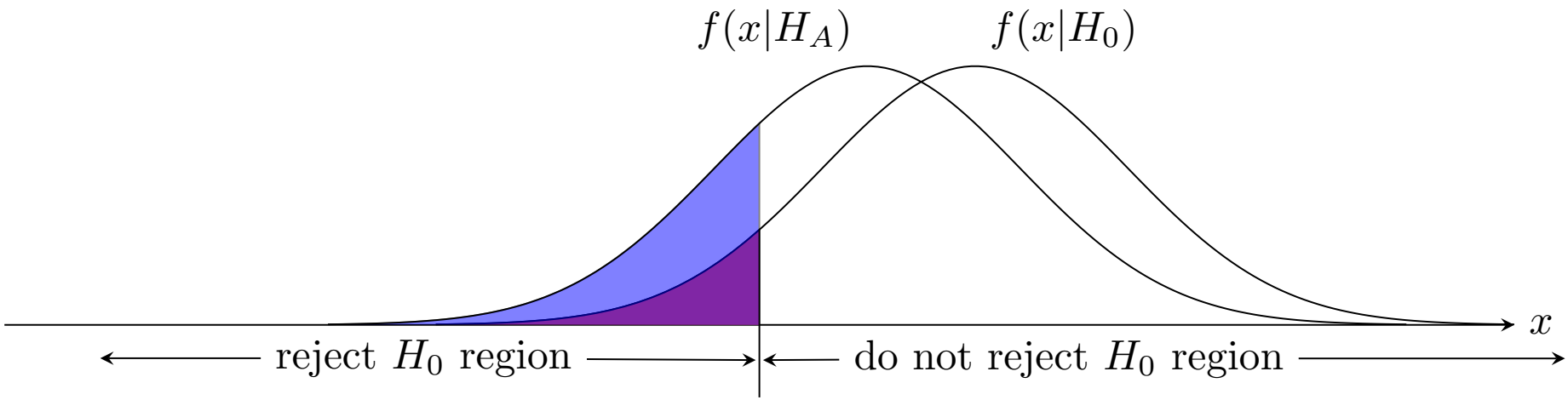
(a) R_1 (b) R_2 (c) $R_1 + R_2$ (d) $R_1 + R_2 + R_3$

answer: (c) $R_1 + R_2$. Power $= P(\text{rejection region} \mid H_A) = \text{area } R_1 + R_2$.

Concept question

2. Which test has higher power?





(a) Top graph (b) Bottom graph

Solution

answer: (a) The top graph.

Power $= P(x \text{ in rejection region} \mid H_A)$. In the top graph almost all the probability of H_A is in the rejection region, so the power is close to 1.

Discussion question

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?

answer: 20 is two standard deviations above the mean of 4. Thus,

$$\text{significance} = P(x \geq 20 | H_0) \approx 0.025$$

The question about power was a trick: we can't compute the power without an alternative distribution.

One-sample t-test

- Data: we assume normal data with both μ and σ unknown:

$$x_1, x_2, \ldots, x_n \sim N(\mu, \sigma^2).$$

- Null hypothesis: $\mu = \mu_0$ for some specific value μ_0.
- Test statistic:

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

  Here t is the Studentized mean and s^2 is the sample variance.
- Null distribution: $f(t \mid H_0)$ is the pdf of $T \sim t(n-1)$, the t distribution with $n-1$ degrees of freedom.
- Two-sided p-value: $p = P(|T| > |t|)$.
- R command: `pt(x,n-1)` is the cdf of $t(n-1)$.
- `http://mathlets.org/mathlets/t-distribution/`

Board question: z and one-sample t-test

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

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

Suppose H_0: $\mu = 0$; H_A: $\mu \neq 0$.

1. Is the test one or two-sided? If one-sided, which side?

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

3. Now assume σ^2 is unknown and test H_0 against H_A.

Answer on next slide.

Solution

We have $\bar{x} = 5, \quad s^2 = \frac{9+1+1+25}{3} = 12$

1. Two-sided. A standardized sample mean far above or below 0 is evidence against H_0, and consistent with H_A.

2. We'll use the standardized mean z for the test statistic (we could also use $\bar{x}$). The null distribution for z is $\mathrm{N}(0, 1)$. This is a two-sided test so the rejection region is

$$(z \leq z_{0.975} \text{ or } \ z \geq z_{0.025}) \ = \ (-\infty, -1.96] \cup [1.96, \infty)$$

Since $z = (\bar{x} - 0)/(4/2) = 2.5$ is in the rejection region we reject H_0 in favor of H_A.
Repeating the test using a p-value:

$$p = P(|z| \geq 2.5 \mid H_0) = 0.012$$

Since $p < \alpha$ we reject H_0 in favor of H_A.

Continued on next slide.

Solution continued

3. We'll use the Studentized $t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$ for the test statistic. The null distribution for t is t_3. For the data we have $t = 5/\sqrt{3}$. This is a two-sided test so the p-value is

$$p = P(|t| \geq 5/\sqrt{3} \mid H_0) = 0.06318$$

Since $p > \alpha$ we do not reject H_0.

Two-sample t-test: equal variances

Data: we assume normal data with μ_x, μ_y and (same) σ unknown:

$$x_1, \ldots, x_n \sim \mathrm{N}(\mu_x, \sigma^2), \quad y_1, \ldots, y_m \sim \mathrm{N}(\mu_y, \sigma^2)$$

Null hypothesis H_0: $\mu_x = \mu_y$.

Pooled variance: $$s_p^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}\left(\frac{1}{n} + \frac{1}{m}\right).$$

Test statistic: $$t = \frac{\bar{x} - \bar{y}}{s_p}$$

Null distribution: $f(t \mid H_0)$ 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 $s^2 = 4.95$

1. Set up and run a two-sample t-test to investigate whether the duration differs for the two groups.

2. What assumptions did you make?

Solution

The pooled variance for this data is

$$s_p^2 = \frac{774(7.77) + 632(4.95)}{1406}\left(\frac{1}{775} + \frac{1}{633}\right) = .0187$$

The t statistic for the null distribution is

$$\frac{\bar{x} - \bar{y}}{s_p} = -3.8064$$

Rather than compute the two-sided p-value using `2*tcdf(-3.8064,1406)` we simply note that with 1406 degrees of freedom the t distribution is essentially standard normal and 3.8064 is almost 4 standard deviations. So

$$P(|t| \geq 3.8064) = P(|z| \geq 3.8064)$$

which is very small, much smaller than $\alpha = .05$ or $\alpha = .01$. Therefore we reject the null hypothesis in favor of the alternative that there is a difference in the mean durations.

Continued on next slide.

Solution continued

2. We assumed the data was normal and that the two groups had equal variances. Given the big difference in the sample variances this assumption might not be warranted.

Note: there are significance tests to see if the data is normal and to see if the two groups have the same variance.

Table discussion: Type I errors Q1

1. 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?

Table discussion: Type I errors Q1

1. 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?

answer: With the information given we can't know this. **The percentage could be anywhere from 0 to 100!**
See the next two questions.

Table discussion: Type I errors Q2

2. 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 α he publishes a paper claiming the treatment is significantly better than a placebo.

(a) Assuming that his treatments are never 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?

answer: **(a)** Since in all of his experiments H_0 is true, roughly 5%, i.e. the significance level, of his experiments will have $p < 0.05$ and be published.

(b) Since he's always wrong, all of his published papers contain type I errors.

Table discussions: Type I errors: Q3

3. Efrat is a genius at designing treatments, so all of her proposed treatments are effective. She's 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 no better than placebo?

answer: 3. (a) The percentage that get published depends on the power of her treatments. If they are only a tiny bit more effective than placebo then roughly 5% of her experiments will yield a publication. If they are a lot more effective than placebo then as many as 100% could be published.
(b) None of her published papers contain type I errors.