# Null Hypothesis Significance Testing
# Signifcance Level, Power, $t$-Tests

18.05 Spring 2014

Jeremy Orloff and Jonathan Bloom

# 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.

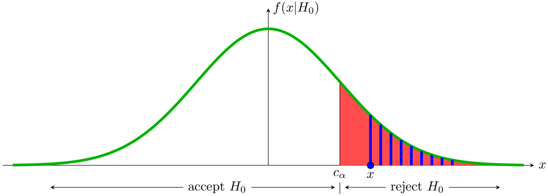
**Example.** A coin has probability $\theta$ of heads. Toss it 30 times and let $x$ be the number of heads.

(i) $H$: $\theta = .4$ is simple. $x \sim \text{binomial}(30, .4)$.

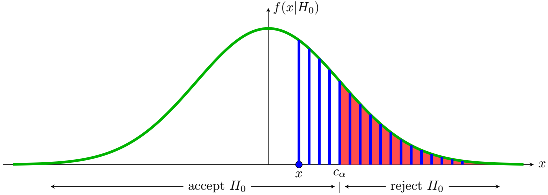
(ii) $H$: $\theta > .4$ is composite. $x \sim \text{binomial}(30, \theta)$ depends on which value of $\theta$ is chosen.

# Extreme data and $p$-values

Area in red $= P(\text{rejection region} \,|\, H_0) = \alpha$.

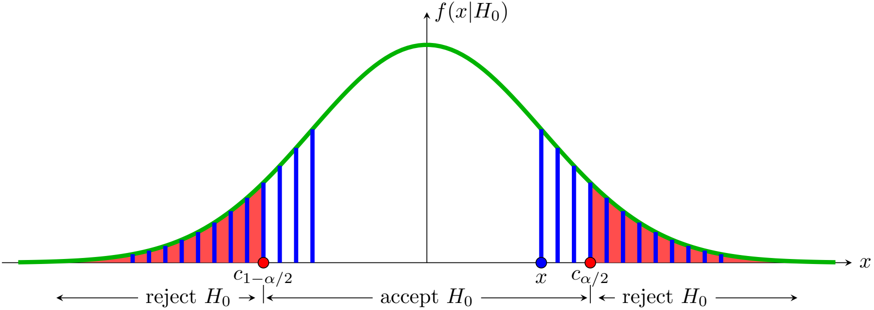



Statistic $x$ inside rej. region $\Leftrightarrow$ $p < \alpha$ $\Leftrightarrow$ reject $H_0$




Statistic $x$ outside rej. region $\Leftrightarrow$ $p > \alpha$ $\Leftrightarrow$ do not reject $H_0$

# Two-sided $p$-values




$p > \alpha$: do not reject $H_0$

**Critical values:**

- The boundary of the rejection region are called critical values.
- Critical values are labeled by the probability to their *right*.
- They are complementary to quantiles: $c_{.1} = q_{.9}$
- Example: for a standard normal $c_{.025} = 2$ and $c_{.975} = -2$.

# Error, significance level and power

| | | True state of nature | |
|---|---|---|---|
| | | $H_0$ | $H_A$ |
| Our decision | Reject $H_0$ | Type I error | correct decision |
| | 'Accept' $H_0$ | correct decision | Type II error |

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

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

****Want significance level near 0 and power near 1.****

# Board 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 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $H_0 : p(x\|\theta = .5)$ | .001 | .010 | .044 | .117 | .205 | .246 | .205 | .117 | .044 | .010 | .001 |
| $H_A : p(x\|\theta = .6)$ | .000 | .002 | .011 | .042 | .111 | .201 | .251 | .215 | .121 | .040 | .006 |
| $H_A : p(x\|\theta = .7)$ | .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.

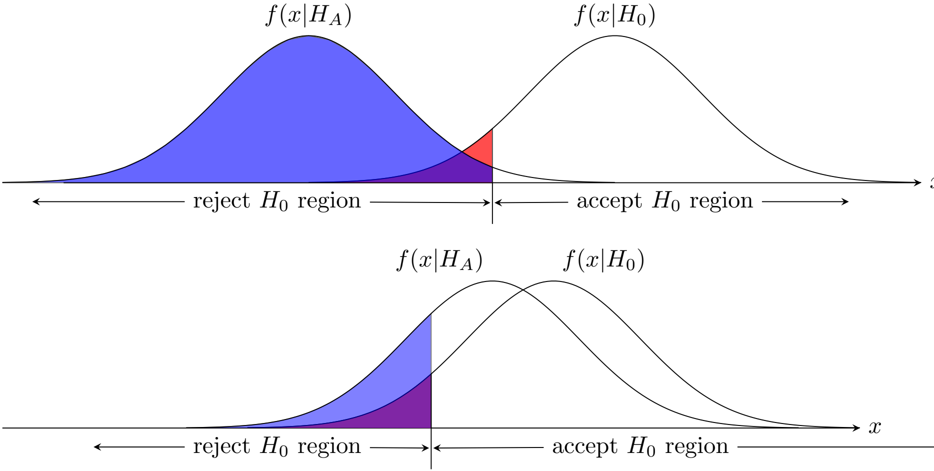
**1.** Significance level $= P(\text{rejection region} \mid H_0) = .11$

**2.** $\theta = .6$: power $= P(\text{rejection region} \mid H_A) = .18$

$\theta = .7$: power $= P(\text{rejection region} \mid H_A) = .383$

# Concept question

**1.** Which test has higher power?



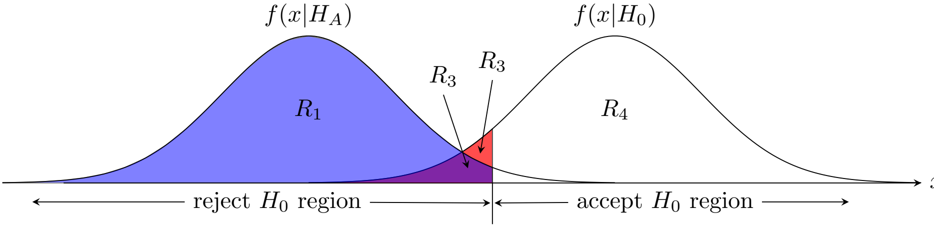

(a) Top graph (b) Bottom graph

# Solution

**answer:** (a) Power = $P(\text{rejection region} \mid H_A)$. In the top graph almost all the probability of $H_A$ is in the rejection region.

# Concept question

**2.** 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) $R_1 + R_2 + R_3$

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

# 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?

<u>**answer:**</u> 20 is two standard deviations above the mean of 4. Thus,

$$P(x \geq 20 | H_0) \approx .025$$

We can't compute the power without an alternative distribution.

## One-sample $t$-test

- Data: we assume normal data with both $\mu$ and $\sigma$ unknown:
$$x_1, x_2, \ldots, x_n \sim N(\mu, \sigma^2).$$
- Null hypothesis: $\mu = \mu_0$ for some specific value $\mu_0$.
- Test statistic:
$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$
where
$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2.$$
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://ocw.mit.edu/ans7870/18/18.05/s14/applets/t-jmo.html

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

For both problems use significance level $\alpha = .05$.

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

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

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

**2.** Now assume $\sigma^2$ 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. We'll use $\bar{x}$ for the test statistic (we could also use $z$). The null distribution for $\bar{x}$ is $\mathrm{N}(0, 4^2/4)$. This is a two-sided test so the rejection region is

$$(\bar{x} \leq \sigma_{\bar{x}} z_{.975} \text{ or } \ \bar{x} \geq \sigma_{\bar{x}} z_{.025}) \ = \ (-\infty, -3.9199] \cup [3.9199, \infty)$$

Since our sample mean $\bar{x} = 5$ is in the rejection region we reject $H_0$ in favor of $H_A$. Repeating the test using a $p$-value:

$$p = P(|\bar{x}| \geq 5 \mid H_0) = P\left(\frac{|\bar{x}|}{2} \geq \frac{5}{2} \mid H_0\right) = P(z \geq 2.5) = .012$$

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

*Continued on next slide.*

# Solution continued

2. We'll use $t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$ 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

$$p = P(|t| \geq 5/\sqrt{3}|H_0) = .06318$$

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

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

Data: we assume normal data with $\mu_x, \mu_y$ and (same) $\sigma$ 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 question

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.

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

What percentage of his published papers describe treatments that are better than placebo?

**answer:** Since in all of his experiments $H_0$ is true, 5% of his experiments will have $p < .05$ and be published.

Since he's always wrong, none of his published papers describe effective treatments.

# Table question

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

How could you determine what percentage of his experiments result in publications?

What percentage of his published papers describe effective treatments?

**answer:** The percentage that get published depends on the power of his treatments. If they are only a tiny bit more effective than placebo then roughly 5% of his experiments will yield a publication. If they are a lot more effective than placebo then as many as 100% could be published. All of his published papers describe effective treatments

18.05 Introduction to Probability and Statistics
Spring 2014