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

18.05 Spring 2017
Understand this figure

- $x = \text{test statistic}$
- $f(x|H_0) = \text{pdf of null distribution} = \text{green curve}$
- Rejection region is a \text{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 \( \theta \) 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 \( \theta \) is chosen.
Extreme data and \textit{p}-values

\textbf{Hypotheses}: \(H_0, H_A\).

\textbf{Test statistic}: value: \(x\), computed from data.

\textbf{Null distribution}: \(f(x|H_0)\) (assumes null hypothesis is true)

\textbf{Sides}: \(H_A\) determines if the rejection region is one or two-sided.

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

The \textit{p}-value is a tool to check if the test statistic is in the rejection region. It is also a \textbf{measure of the evidence for rejecting} \(H_0\).

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

“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 $H_0$?

\[ x \in f(x|H_0) \]

\[ c_\alpha \]

\[ 4.2 \]

\[ \begin{array}{c}
\text{don’t reject } H_0 \\
\text{reject } H_0
\end{array} \]

\[ p \text{-value: } p = P(x \geq 4.2 | H_0) = \text{blue area}. \]

\[ \alpha = P(x \text{ in rejection region } | H_0) = \text{red area}. \]

Since $p < \alpha$, 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 $H_0$?

![Diagram showing rejection region with test statistic $x = 4.2$.]

**answer:** 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 \( H_0 \)?

\[
\begin{align*}
\text{do n’t reject } H_0 \quad &\quad c_\alpha \quad &\quad 4.2 \quad &\quad \text{reject } H_0
\end{align*}
\]

\[ f(x|H_0) \]

**answer:** The test statistic is in the rejection region, so reject \( H_0 \).

**Alternatively:** blue area \( < \) red area

**Significance:** \( \alpha = P(x \text{ in rejection region} | H_0) = \text{red area} \)

**\( p \)-value:** \( p = P(\text{data at least as extreme as } x | H_0) = \text{blue area} \)

Since \( p < \alpha \) we reject \( H_0 \).
Extreme data and $p$-values

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

$$x \sim f(x|H_0)$$

$2.1$ $c_\alpha$

[Diagram: Green curve representing $f(x|H_0)$ with blue and red areas for rejection region.]

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

Alternatively: blue area $>$ red area

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

$p$-value: $p = P(x \text{ at least as extreme as } 2.1 | H_0) = \text{blue area}$.

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

April 11, 2017 6 / 21
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 $H_0$.**
Extreme data and $p$-values

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

\[ f(x|H_0) \]

\[ \text{don't reject } H_0 \quad 2.1 \quad c_\alpha \quad \text{reject } H_0 \]

**answer:** Test statistic **not** in the rejection region: don’t reject $H_0$.

**Alternatively:** blue area $\geq$ red area

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

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

Since $p > \alpha$ we **don’t reject** $H_0$. 
Critical values

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: $c_{0.1} = q_{0.9}$
- Example: for a standard normal $c_{0.025} = 1.96$ and $c_{0.975} = -1.96$.
- In R, for a standard normal $c_{0.025} = \text{qnorm}(0.975)$. 
Two-sided $p$-values

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

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

If the significance (rejection) probability is split evenly between the left and right tails then

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

\[x\] 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) $\text{pnorm}(0.1, 0, 1)$
- (c) $\text{pnorm}(0.95, 0, 1)$
- (e) $1 - \text{pnorm}(1.8, 0, 1)$
- (g) $\text{qnorm}(0.1, 0, 1)$
- (i) $\text{qnorm}(0.95, 0, 1)$

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

- (b) $\text{pnorm}(0.9, 0, 1)$
- (d) $\text{pnorm}(1.8, 0, 1)$
- (f) $\text{qnorm}(0.05, 0, 1)$
- (h) $\text{qnorm}(0.9, 0, 1)$

(iii) Should you reject the null hypothesis?

- (a) Yes
- (b) No
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) $\text{pnorm}(0.1, 0, 1)$  
(b) $\text{pnorm}(0.9, 0, 1)$  
(c) $\text{pnorm}(0.95, 0, 1)$  
(d) $\text{pnorm}(1.8, 0, 1)$  
(e) $1 - \text{pnorm}(1.8, 0, 1)$  
(f) $\text{qnorm}(0.05, 0, 1)$  
(g) $\text{qnorm}(0.1, 0, 1)$  
(h) $\text{qnorm}(0.9, 0, 1)$  
(i) $\text{qnorm}(0.95, 0, 1)$

(ii) Which of the above computes the $p$-value for this experiment?
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) $\text{pnorm}(0.1, 0, 1)$
(b) $\text{pnorm}(0.9, 0, 1)$
(c) $\text{pnorm}(0.95, 0, 1)$
(d) $\text{pnorm}(1.8, 0, 1)$
(e) $1 - \text{pnorm}(1.8, 0, 1)$
(f) $\text{qnorm}(0.05, 0, 1)$
(g) $\text{qnorm}(0.1, 0, 1)$
(h) $\text{qnorm}(0.9, 0, 1)$
(i) $\text{qnorm}(0.95, 0, 1)$

(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 level and power

<table>
<thead>
<tr>
<th>Our decision</th>
<th>True state of nature</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Reject $H_0$</td>
<td>Type I error</td>
<td>correct decision</td>
<td></td>
</tr>
<tr>
<td>Don’t reject $H_0$</td>
<td>correct decision</td>
<td></td>
<td>Type II error</td>
</tr>
</tbody>
</table>

**Significance level**

\[ P(\text{type I error}) \]
\[ = P(\text{probability we incorrectly reject } H_0) \]
\[ = P(\text{test statistic in rejection region } | H_0) \]
\[ = P(\text{false positive}) \]

**Power**

\[ P(\text{test statistic in rejection region } | H_A) \]
\[ = 1 - P(\text{type II error}) \]
\[ = P(\text{true positive}) \]

- $H_A$ 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.

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

1. Find the significance level of the test.
2. Find the power of the test for each of the two alternative hypotheses.
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) $R_1 + R_2 + R_3$
2. Which test has higher power?
Discussion question

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

What is the significance level and power of this test?
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**:

\[
t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}
\]

where

\[
s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2.
\]

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 \( t(n-1) \).

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 $\sigma^2$ is unknown and test $H_0$ against $H_A$. 
Two-sample $t$-test: equal variances

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

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

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?
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?
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.
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 $\alpha$ 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?
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?