Null Hypothesis Significance Testing

Gallery of Tests

18.05 Spring 2017
Discussion of Studio 8 and simulation

- What is a simulation?
  - Run an experiment with pseudo-random data instead of real-world real random data.
  - By doing this many times we can estimate the statistics for the experiment.

- Why do a simulation?
  - In the real world we are not omniscient.
  - In the real world we don’t have infinite resources.

- What was the point of Studio 8?
  - To simulate some simple significance tests and compare various frequencies.
  - Simulated $P(\text{reject}|H_0) \approx$ significance $\alpha$
  - Simulated $P(\text{reject}|H_A) \approx$ power
  - $P(H_0|\text{reject})$ can be anything; depends on distribution of hypotheses which we almost never know.
We run a two-sample $t$-test for equal means, with $\alpha = 0.05$, and obtain a $p$-value of 0.04. What are the odds that the two samples are drawn from distributions with the same mean?

(a) 19/1  (b) 1/19  (c) 1/20  (d) 1/24  (e) unknown
General pattern of NHST

You wish to decide whether to reject $H_0$ (perhaps in favor of $H_A$).

Design:

- **Design experiment** to collect data relevant to hypotheses.
- **Choose text statistic** $x$ with known null distribution $f(x \mid H_0)$.
- **Choose the significance level** $\alpha$ and compute the rejection region.
- Perhaps, for simple alternative $H_A$, use $f(x \mid H_A)$ to compute power.

Alternatively, you can choose both the significance level and the power, and then compute the necessary sample size.

Implementation:

- **Run the experiment** to collect data.
- **Compute the statistic** $x$ and the corresponding $p$-value.
- If $p < \alpha$, reject $H_0$. 
Chi-square test for homogeneity

**Setting:** We have several data sets (for example results of applying several different treatments; or polling data from several different states).

**Homogeneity** (the null hypothesis) means that the data sets are all drawn from the same distribution: that all the treatments are equally effective, or that Connecticut voters have the same political opinions as those in New York.

Three treatments for a disease are compared in a clinical trial, yielding the following data:

<table>
<thead>
<tr>
<th></th>
<th>Treatment 1</th>
<th>Treatment 2</th>
<th>Treatment 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cured</td>
<td>50</td>
<td>30</td>
<td>12</td>
</tr>
<tr>
<td>Not cured</td>
<td>100</td>
<td>80</td>
<td>18</td>
</tr>
</tbody>
</table>

Use a **chi-square test** to compare the cure rates for the three treatments, i.e., to test if all three cure rates are the same.
Solution

\[ H_0 = \text{all three treatments have the same cure rate.} \]
\[ H_A = \text{the three treatments have different cure rates.} \]

**Expected counts**
- Under \( H_0 \) the MLE for the cure rate is \[ \frac{\text{total cured}}{\text{total treated}} = \frac{92}{290} = 0.317 \].
- Assuming \( H_0 \), the expected number cured for each treatment is the number treated times 0.317.
- This gives the following table of observed and expected counts (observed in black, expected in blue).
- We include the marginal values (in red). These were used to compute the expected counts.

<table>
<thead>
<tr>
<th></th>
<th>Treatment 1</th>
<th>Treatment 2</th>
<th>Treatment 3</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cured</td>
<td>50, 47.6</td>
<td>30, 34.9</td>
<td>12, 9.5</td>
<td>92</td>
</tr>
<tr>
<td>Not cured</td>
<td>100, 102.4</td>
<td>80, 75.1</td>
<td>18, 20.5</td>
<td>198</td>
</tr>
<tr>
<td></td>
<td>150</td>
<td>110</td>
<td>30</td>
<td>290</td>
</tr>
</tbody>
</table>

*continued*
Solution continued

Likelihood ratio statistic: \[ G = 2 \sum O_i \ln(O_i/E_i) = 2.12 \]

Pearson’s chi-square statistic: \[ X^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 2.13 \]

**Degrees of freedom** means how many choices describe the data.

- **Formula:** degrees of freedom \( df = (2 - 1)(3 - 1) = 2 \).
- **Counting:** Marginal totals are fixed because they are needed to compute expected counts. So we can choose 2 of the 6 cells and all the others are determined: degrees of freedom = 2.

**p-value**

\[ p = 1 - pchisq(2.12, 2) = 0.346 \]

The data does not support rejecting \( H_0 \). We do not conclude that the treatments have differing efficacy.
Board question: Khan’s restaurant

Sal is thinking of buying a restaurant and asks about the distribution of lunch customers. The owner provides row one below. Sal records the data in row two himself one week.

<table>
<thead>
<tr>
<th></th>
<th>M</th>
<th>T</th>
<th>W</th>
<th>R</th>
<th>F</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>Owner’s distribution</td>
<td>.1</td>
<td>.1</td>
<td>.15</td>
<td>.2</td>
<td>.3</td>
<td>.15</td>
</tr>
<tr>
<td>Observed # of cust.</td>
<td>30</td>
<td>14</td>
<td>34</td>
<td>45</td>
<td>57</td>
<td>20</td>
</tr>
</tbody>
</table>

Run a chi-square **goodness-of-fit test** on the null hypotheses:

- $H_0$: the owner’s distribution is correct.
- $H_A$: the owner’s distribution is not correct.

Compute $X^2$. 
Board question: genetic linkage

In 1905, William Bateson, Edith Saunders, and Reginald Punnett were examining flower color and pollen shape in sweet pea plants by performing crosses similar to those carried out by Gregor Mendel.

Purple flowers (P) is dominant over red flowers (p).
Long seeds (L) is dominant over round seeds (l).

F0: PPLL × ppll (initial cross)
F1: PpLl × PpLl (all second generation plants were PpLl)
F2: 2132 plants (third generation)

$H_0 = \text{independent assortment: color and shape are independent.}$

<table>
<thead>
<tr>
<th></th>
<th>purple, long</th>
<th>purple, round</th>
<th>red, long</th>
<th>red, round</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expected</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>Observed</td>
<td>1528</td>
<td>106</td>
<td>117</td>
<td>381</td>
</tr>
</tbody>
</table>

Determine the expected counts for $F_2$ under $H_0$ and find the $p$-value for a Pearson chi-square test. Explain your findings biologically.
$F$-distribution

- Notation: $F_{a,b}$, $a$ and $b$ degrees of freedom
- Derived from normal data
- Range: $[0, \infty)$

*Plot of $F$ distributions:
  - $F_{3,4}$
  - $F_{10,15}$
  - $F_{30,15}$*
$F$-test = one-way ANOVA

Like $t$-test but for $n$ groups of data with $m$ data points each.

$$y_{i,j} \sim N(\mu_i, \sigma^2), \quad y_{i,j} = j^{\text{th}} \text{ point in } i^{\text{th}} \text{ group}$$

Null hypothesis is that means are all equal: $\mu_1 = \cdots = \mu_n$.

Group variances $\sigma^2$ assumed equal. Test statistic is $\frac{MS_B}{MS_W}$ where:

$MS_B = \text{between group variance} = \frac{m}{n-1} \sum (\bar{y}_i - \bar{y})^2$

$MS_W = \text{within group variance} = \text{sample mean of } s_1^2, \ldots, s_n^2 \approx \sigma^2$.

Idea: If $\mu_i$ are equal, $MS_B \approx \sigma^2$, so ratio should be near 1.

Null distribution is F-statistic with $n-1$ and $n(m-1)$ d.o.f.:

$$\frac{MS_B}{MS_W} \sim F_{n-1, n(m-1)}$$

Note: Formulas easily generalize to unequal group sizes: 
http://en.wikipedia.org/wiki/F-test
Board question

The table shows recovery time in days for three medical treatments.

1. Set up and run an F-test testing if the average recovery time is the same for all three treatments.

2. Based on the test, what might you conclude about the treatments?

<table>
<thead>
<tr>
<th></th>
<th>$T_1$</th>
<th>$T_2$</th>
<th>$T_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
<td>8</td>
<td>13</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>9</td>
<td>11</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>11</td>
<td>8</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>8</td>
<td>12</td>
</tr>
</tbody>
</table>

For $\alpha = 0.05$, the critical value of $F_{2,15}$ is 3.68.
Concept question: multiple-testing

1. Suppose we have 6 treatments and want to know if the average recovery time is the same for all of them. If we compare two at a time, how many two-sample $t$-tests do we need to run?

(a) 1  (b) 2  (c) 6  (d) 15  (e) 30
Concept question: multiple-testing

1. Suppose we have 6 treatments and want to know if the average recovery time is the same for all of them. If we compare two at a time, how many two-sample $t$-tests do we need to run?

   (a) 1  (b) 2  (c) 6  (d) 15  (e) 30

2. Suppose we use the significance level 0.05 for each of the 15 tests. Assuming the null hypothesis, what is the probability that we reject at least one of the 15 null hypotheses?

   (a) Less than 0.05  (b) 0.05  (c) 0.10  (d) Greater than 0.25
Concept question: multiple-testing

1. Suppose we have 6 treatments and want to know if the average recovery time is the same for all of them. If we compare two at a time, how many two-sample $t$-tests do we need to run?

   (a) 1   (b) 2   (c) 6   (d) 15   (e) 30

2. Suppose we use the significance level 0.05 for each of the 15 tests. Assuming the null hypothesis, what is the probability that we reject at least one of the 15 null hypotheses?

   (a) Less than 0.05   (b) 0.05   (c) 0.10   (d) Greater than 0.25

Discussion: Recall that there is an $F$-test that tests if all the means are the same. What are the trade-offs of using the $F$-test rather than many two-sample $t$-tests?
Board question: chi-square for independence

(From Rice, *Mathematical Statistics and Data Analysis*, 2nd ed. p.489)

Consider the following contingency table of counts

<table>
<thead>
<tr>
<th>Education</th>
<th>Married once</th>
<th>Married multiple times</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>College</td>
<td>550</td>
<td>61</td>
<td>611</td>
</tr>
<tr>
<td>No college</td>
<td>681</td>
<td>144</td>
<td>825</td>
</tr>
<tr>
<td>Total</td>
<td>1231</td>
<td>205</td>
<td>1436</td>
</tr>
</tbody>
</table>

Use a chi-square test with significance level 0.01 to test the hypothesis that the number of marriages and education level are independent.