Null Hypothesis Significance Testing
Gallery of Tests

18.05 Spring 2014

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 frequences.
  – Simulated $P(\text{reject}|H_0) \approx \alpha$
  – Simulated $P(\text{reject}|H_A) \approx \text{power}$
  – $P(H_0|\text{reject}$ can be anything depending on the (usually) unknown prior

Concept question

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 are interested in whether to reject H_0 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 α and find the rejection region.
- For a simple alternative H_A, use $f(x \mid H_A)$ to compute the 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

In this setting homogeneity means that the data sets are all drawn from the same distribution.

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

	Treatment 1	Treatment 2	Treatment 3
Cured	50	30	12
Not cured	100	80	18

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 = all three treatments have the same cure rate.
H_A = the three treatments have different cure rates.

Expected counts

- Under H_0 the MLE for the cure rate is
$$\text{(total cured)}/\text{(total treated)} = 92/290 = 0.317 \quad .$$
- 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 are all needed to compute the expected counts.

	Treatment 1	Treatment 2	Treatment 3	
Cured	50, 47.6	30, 34.9	12, 9.5	92
Not cured	100, 102.4	80, 75.1	18, 20.5	198
	150	110	30	290

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

- Formula: Test for homogeneity $df = (2-1)(3-1) = 2$.
- Counting: The marginal totals are fixed because they are needed to compute the expected counts. So we can freely put values in 2 of the cells and then 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 1 below. Sal records the data in row 2 himself one week.

	M	T	W	R	F	S
Owner's distribution	.1	.1	.15	.2	.3	.15
Observed # of cust.	30	14	34	45	57	20

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 both G and 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 x ppll (initial cross)
F1: PpLl x PpLl (all second generation plants were PpLl)
F2: 2132 plants (third generation)

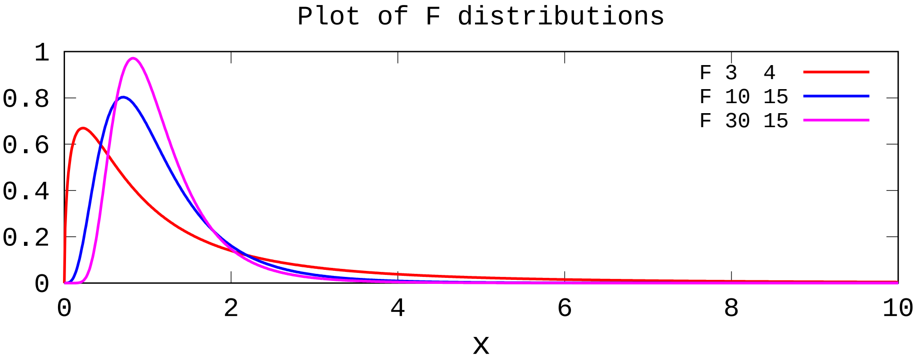
$H_0 =$ independent assortment: color and shape are independent.

	purple, long	purple, round	red, long	red, round
Expected	?	?	?	?
Observed	1528	106	117	381

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: $[0, \infty)$



F-test = one-way ANOVA

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

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

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

Test statistic is $\frac{\text{MS}_B}{\text{MS}_W}$ where:

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

MS_W = within group variance = sample mean of $s_1^2, \ldots, s_n^2$

Idea: If μ_i are equal, this ratio should be near 1.

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

$$\frac{\text{MS}_B}{\text{MS}_W} \sim F_{n-1,\ n(m-1)}$$

Note: Formulas easily generalizes 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?

T_1	T_2	T_3
6	8	13
8	12	9
4	9	11
5	11	8
3	6	7
4	8	12

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

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

Education	Married once	Married multiple times	Total
College	550	61	611
No college	681	144	825
Total	1231	205	1436

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