# Null Hypothesis Significance Testing
# Gallery of Tests

18.05 Spring 2018

# 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.

# 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

**answer:** (e) unknown. Frequentist methods only give probabilities for data under an assumed hypothesis. They do not give probabilities or odds for hypotheses. So we don't know the odds for distribution means

# 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:

| | 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
$$\text{(total cured)/(total treated)} = 92/290 = 0.317 \quad .$$
- 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 were used 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** 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.

| | 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 $X^2$.

# Solution

The total number of observed customers is 200.

The expected counts (under $H_0$) are 20 20 30 40 60 30

$$X^2 = \sum \frac{(O_i - E_i)^2|}{E_i} = 11.44$$

$df = 6 - 1 = 5$ (6 cells, compute 1 value –the total count– from the data)

$$p = \texttt{1-pchisq(11.39,5) = 0.043.}$$

So, at a significance level of 0.05 we reject the null hypothesis in favor of the alternative that the owner's distribution is wrong.

# 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.

# Solution

Since every F1 generation flower has genotype Pp we'd expect F2 to split 1/4, 1/2, 1/4 between PP, Pp, pp. For phenotype we expect F2 to have 3/4 purple and 1/4 red flowers. Similarly for LL, Ll, ll. Assuming $H_0$ that color and shape are independent we'd expect the following probabilities for F2.

| | LL | Ll | ll | |
|---|---|---|---|---|
| PP | 1/16 | 1/8 | 1/16 | 1/4 |
| Pp | 1/8 | 1/4 | 1/8 | 1/2 |
| pp | 1/16 | 1/8 | 1/16 | 1/4 |
| | 1/4 | 1/2 | 1/4 | 1 |

Genotype

| | Long | Round | |
|---|---|---|---|
| Purple | 9/16 | 3/16 | 3/4 |
| Red | 3/16 | 1/16 | 1/4 |
| | 3/4 | 1/4 | 1 |

Phenotype

Using the total of 2132 plants in F2, the expected counts come from the phenotype table:

| | purple, long | purple, round | red, long | red, round |
|---|---|---|---|---|
| Expected | 1199 | 400 | 400 | 133 |
| Observed | 1528 | 106 | 117 | 381 |

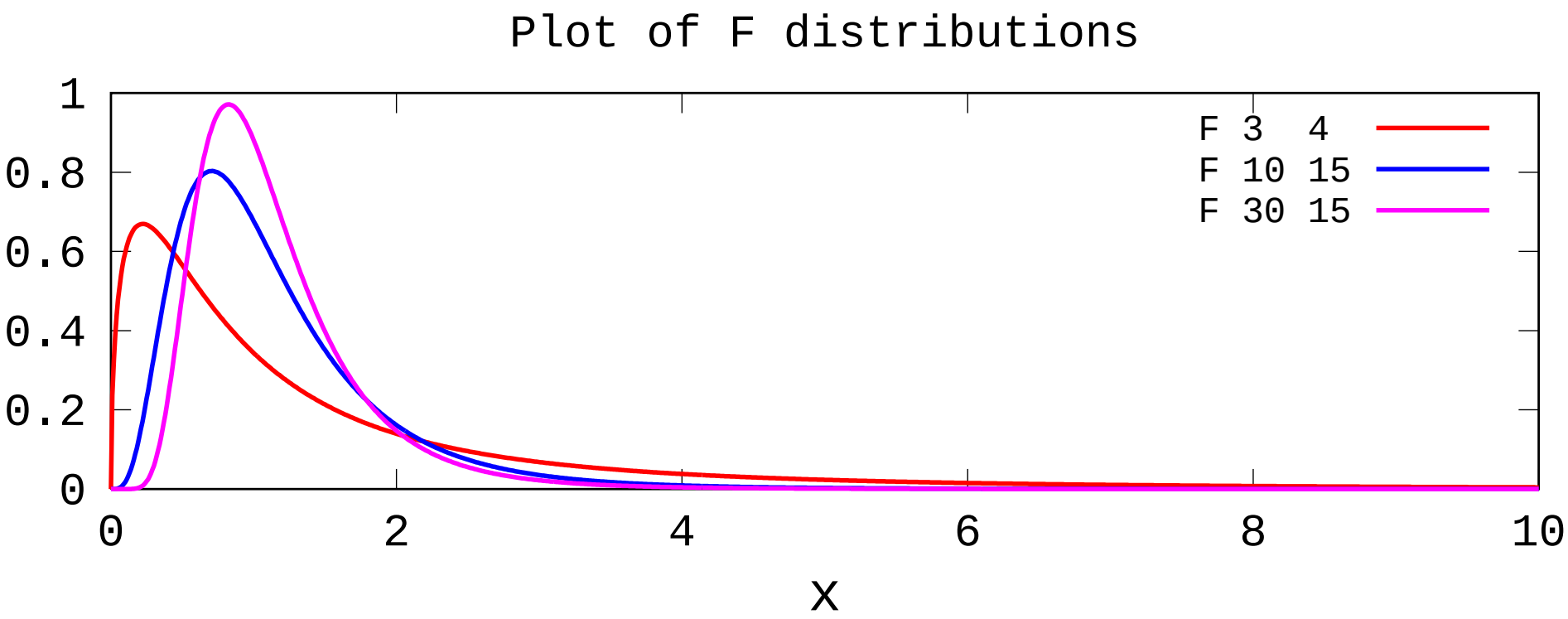
# Continued

Using R we compute: $G = 972.0$, $X^2 = 966.6$.

The degrees of freedom is 3 (4 cells - 1 cell needed to make the total work out). The $p$-values for both statistics is too close to zero for R to compute. (For example, $1 - \texttt{pchisq}(70, 3) = 4 \times 10^{-15}$.) With such a small $p$-value we reject $H_0$ in favor of the alternative that the genes are not independent.

# $F$-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: $\mu_1 = \cdots = \mu_n$.

Group variances $\sigma^2$ assumed equal. Test statistic is $\frac{\text{MS}_B}{\text{MS}_W}$ where:

$$\text{MS}_B = \text{between group variance} = \frac{m}{n-1}\sum_{i=1}^{n}(\bar{y}_i - \bar{y})^2$$

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

Idea: If $\mu_i$ are equal, $\text{MS}_B \approx \sigma^2$, so ratio should be near 1.

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

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

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?

| $T_1$ | $T_2$ | $T_3$ |
|---|---|---|
| 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.

# Solution

$H_0$ is that the means of the 3 treatments are the same. $H_A$ is that they are not.

Our test statistic $w$ is computed following the procedure from a previous slide. We get that the test statistic $w$ is approximately 9.25. The $p$-value is approximately 0.0024. We reject $H_0$ in favor of the hypothesis that the means of three treatments are not the same.

# 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?

**answer:** *Solution on next slide.*

# Solution

**answer:** **1.** 6 choose $2 = 15$.

**2.** **answer:** (d) Greater than 0.25.

Under $H_0$ the probability of rejecting for any given pair is 0.05. Because the tests aren't independent, i.e. if the group1-group2 and group2-group3 comparisons fail to reject $H_0$, then the probability increases that the group1-group3 comparison will also fail to reject.

We can say that the following 3 comparisons: group1-group2, group3-group4, group5-group6 are independent. The number of rejections among these three follows a binom$(3, 0.05)$ distribution. The probability the number is greater than 0 is $1 - (0.95)^3 \approx 0.14$.

Even though the other pairwise tests are not independent they do increase the probability of rejection. In simulations of this with normal data the false rejection rate was about 0.36.

# 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.

# Solution

The null hypothesis is that the cell probabilities are the product of the marginal probabilities. Assuming the null hypothesis we estimate the marginal probabilities in red and multiply them to get the cell probabilities in blue.

| Education | Married once | Married multiple times | Total |
|---|---|---|---|
| College | 0.365 | 0.061 | 611/1436 |
| No college | 0.492 | 0.082 | 825/1436 |
| Total | 1231/1436 | 205/1436 | 1 |

We then get expected counts by multiplying the cell probabilities by the total number of women surveyed (1436). The table shows the observed, expected counts:

| Education | Married once | Married multiple times |
|---|---|---|
| College | 550, 523.8 | 61, 87.2 |
| No college | 681, 707.2 | 144, 117.8 |

# Solution continued

We then have

$$G = 16.55 \qquad \text{and} \qquad X^2 = 16.01$$

The number of degrees of freedom is $(2-1)(2-1) = 1$. We could count this: we needed the marginal probabilities to compute the expected counts. Now setting any one of the cell counts determines all the rest because they need to be consistent with the marginal probabilities. We get

$$p = \texttt{1-pchisq(16.55,1)} = 0.000047$$

Therefore we reject the null hypothesis in favor of the alternate hypothesis that number of marriages and education level are not independent