Null Hypothesis Significance Testing III
Class 19, 18.05, Spring 2014
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Given hypotheses and data, be able to identify to identify an appropriate significance test from a list of common ones.
2. Given hypotheses, data, and a suggested significance test, know how to look up details and apply the significance test.

2 Introduction

In these notes we will collect together some of the most common significance tests, though by necessity we will leave out many other useful ones. Still, all significance tests follow the same basic pattern in their design and implementation, so by learning the ones we include you should be able to easily apply other ones as needed.

Designing a null hypothesis significance test (NHST):

- Specify null and alternative hypotheses.
- Choose a test statistic whose null distribution and alternative distribution(s) are known.
- Specify a rejection region. Most often this is done implicitly by specifying a significance level α and a method for computing p-values based on the tails of the null distribution.
- Compute power using the alternative distribution(s).

Running a NHST:

- Collect data and compute the test statistic.
- Check if the test statistic is in the rejection region. Most often this is done implicitly by checking if $p < \alpha$. If so, we 'reject the null hypothesis in favor of the alternative hypothesis'. Otherwise we conclude 'the data does not support rejecting the null hypothesis'.

Note the careful phrasing: when we fail to reject H_0, we do not conclude that H_0 is true. The failure to reject may have other causes. For example, we might not have enough data to clearly distinguish H_0 and H_A, whereas more data would indicate that we should reject H_0.

3 A gallery of common significance tests

We will show a number of tests. For completeness we will include the z and t tests we've already explored. You shouldn't try to memorize these tests. Rather your goal should be to be able to find the correct test when you need it. Pay attention to the types of hypotheses the tests are designed to distinguish and the assumptions about the data needed for the test to be valid.

The null distributions for all of these tests are all related to the normal distribution by explicit formulas. We will not go into the details of these distributions since, given the name of any distribution, you can easily look up the details of its construction and properties online. You can also use R to explore the distribution numerically and graphically.

3.1 z-test

- Use: Compare the data mean to an hypothesized mean.
- Data: $x_1, x_2, \ldots, x_n$.
- Assumptions: The data are independent normal samples:
  $x_i \sim N(\mu, \sigma^2)$ where μ is unknown, but σ is known.
- H_0: For a specified μ_0, $\mu = \mu_0$.
- H_A:
  Two-sided: $\mu \neq \mu_0$
  one-sided-greater: $\mu > \mu_0$
  one-sided-less: $\mu < \mu_0$
- Test statistic: $z = \dfrac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$
- Null distribution: $f(z \,|\, H_0)$ is the pdf of $Z \sim N(0,1)$.
- p-value:
  Two-sided: $p = P(|Z| > z)$ = `2*(1-pnorm(abs(z), 0, 1))`
  one-sided-greater: $p = P(Z > z)$ = `1 - pnorm(z, 0, 1)`
  one-sided-less: $p = P(Z < z)$ = `pnorm(z, 0, 1)`

Example 1. We quickly reprise our example from the class 17 notes.

IQ is normally distributed in the population according to a $\mathrm{N}(100, 15^2)$ distribution. We suspect that most MIT students have above average IQ so we frame the following hypotheses.

$$\begin{aligned} H_0 &= \text{MIT student IQs are distributed identically to the general population} \\ &= \text{MIT IQ's follow a } \mathrm{N}(100, 15^2) \text{ distribution.} \\ H_A &= \text{MIT student IQs tend to be higher than those of the general population} \\ &= \text{the average MIT student IQ is greater than 100.} \end{aligned}$$

Notice that H_A is one-sided.

Suppose we test 9 students and find they have an average IQ of $\bar{x} = 112$. Can we reject H_0 at a significance level $\alpha = .05$?

answer: Our test statistic is

$$z = \frac{\bar{x} - 100}{15/\sqrt{9}} = \frac{36}{15} = 2.4.$$

The right-sided p-value is thereofre

$$p = P(Z \geq 2.4) = \texttt{1- pnorm(2.4,0,1) = 0.0081975.}$$

Since $p \leq \alpha$ we reject the null hypothesis in favor of the alternative hypothesis that MIT students have higher IQs on average.

3.2 One-sample t-test of the mean

- Use: Compare the data mean to an hypothesized mean.
- Data: $x_1, x_2, \ldots, x_n$.
- Assumptions: The data are independent normal samples:
  $x_i \sim N(\mu, \sigma^2)$ where both μ and σ are unknown.
- H_0: For a specified μ_0, $\mu = \mu_0$
- H_A:
  Two-sided: $\mu \neq \mu_0$
  one-sided-greater: $\mu > \mu_0$
  one-sided-less: $\mu < \mu_0$
- Test statistic: $t = \dfrac{\overline{x} - \mu_0}{s/\sqrt{n}}$,
  where s^2 is the sample variance: $s^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$
- Null distribution: $f(t \,|\, H_0)$ is the pdf of $T \sim t(n-1)$.
  (Student t-distribution with $n-1$ degrees of freedom)
- p-value:
  Two-sided: $p = P(|T| > t)$ = `2*(1-pt(abs(t), n-1))`
  one-sided-greater: $p = P(T > t)$ = `1 - pt(t, n-1)`
  one-sided-less: $p = P(T < t)$ = `pt(t, n-1)`

Example 2. Look in the class 18 notes or slides for an example of this test.

3.3 Two-sample t-test for comparing means (assuming equal variance)

- Use: Compare the data means from two groups.
- Data: $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_m$.
- Assumptions: Both groups of data are independent normal samples:
  $$x_i \sim N(\mu_x, \sigma^2)$$
  $$y_j \sim N(\mu_y, \sigma^2)$$
  where both μ_x and μ_y are unknown and possibly different. The variance σ is unknown, but the same for both groups.

- H_0: $\mu_x = \mu_y$
- H_A:

Two-sided:	$\mu_x \neq \mu_y$
one-sided-greater:	$\mu_x > \mu_y$
one-sided-less:	$\mu_x < \mu_y$

- Test statistic: $t = \dfrac{\overline{x} - \bar{y}}{s_P}$,

  where s_x^2 and s_y^2 are the sample variances and s_P^2 is (sometimes called) the pooled sample variance:
  $$s_p^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}\left(\frac{1}{n} + \frac{1}{m}\right)$$
- Null distribution: $f(t \mid H_0)$ is the pdf of $T \sim t(n+m-2)$.
- p-value:

Two-sided:	$p = P(\lvert T\rvert > t)$	$=$	`2*(1-pt(abs(t), n+m-2))`
one-sided-greater:	$p = P(T > t)$	$=$	`1 - pt(t, n+m-2)`
one-sided-less:	$p = P(T < t)$	$=$	`pt(t, n+m-2)`

Notes: 1. There is a form of the t-test for when the variances are not assumed equal. It is sometimes called Welch's t-test.
2. When the data naturally comes in pairs (x_i, y_i), one uses the *paired two-sample t-test.* For example, in comparing two treatments, each patient receiving treatment 1 might be paired with a patient receiving treatment 2 who is similar in terms of stage of disease, age, sex, etc.

Example 3. Look in the class 18 notes or slides for an example of this test.

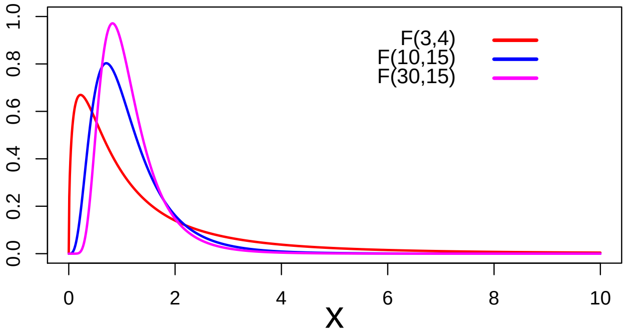
3.4 One-way ANOVA (F-test for equal means)

- Use: Compare the data means from n groups with m data points in each group.
- Data:
  $$\begin{array}{cccc} x_{1,1}, & x_{1,2}, & \ldots, & x_{1,m} \\ x_{2,1}, & x_{2,2}, & \ldots, & x_{2,m} \\ & & \ldots & \\ x_{n,1}, & x_{n,2}, & \ldots, & x_{n,m} \end{array}$$
- Assumptions: Data for each group is an independent normal sample drawn from distributions with (possibly) different means but the same variance:
  $$\begin{array}{ll} x_{1,j} & \sim N(\mu_1, \sigma^2) \\ x_{2,j} & \sim N(\mu_2, \sigma^2) \\ & \ldots \\ x_{n,j} & \sim N(\mu_n, \sigma^2) \end{array}$$
  The group means μ_i are unknown and possibly different. The variance σ is unknown, but the same for all groups.
- H_0: All the means are identical $\mu_1 = \mu_2 = \ldots = \mu_n$.
- H_A: Not all the means are the same.

- Test statistic: $w = \frac{\text{MS}_B}{\text{MS}_W}$, where

$$\begin{aligned}
\bar{x}_i &= \text{mean of group } i \\
&= \frac{x_{i,1} + x_{i,2} + \ldots + x_{i,m}}{m}. \\
\overline{x} &= \text{grand mean of all the data.} \\
s_i^2 &= \text{sample variance of group } i \\
&= \frac{1}{m-1}\sum_{j=1}^{m}(x_{i,j} - \bar{x}_i)^2. \\
\text{MS}_B &= \text{between group variance} \\
&= m \times \text{sample variance of group means} \\
&= \frac{m}{n-1}\sum_{i=1}^{n}(\bar{x}_i - \overline{x})^2. \\
\text{MS}_W &= \text{average within group variance} \\
&= \text{sample mean of } s_1^2, \ldots, s_n^2 \\
&= \frac{s_1^2 + s_2^2 + \ldots + s_n^2}{n}
\end{aligned}$$

- Idea: If the μ_i are all equal, this ratio should be near 1. If they are not equal then MS_B should be larger while MS_W should remain about the same, so w should be larger. We won't give a proof of this.
- Null distribution: $f(w \mid H_0)$ is the pdf of $W \sim F(n-1, n(m-1))$. This is the F-distribution with $(n-1)$ and $n(m-1)$ degrees of freedom. Several F-distributions are plotted below.
- p-value: $p = P(W > w) =$ `1- pf(w, n-1, n*(m-1)))`



Notes: 1. ANOVA tests whether all the means are the same. It does not test whether some subset of the means are the same.
2. There is a test where the variances are not assumed equal.
3. There is a test where the groups don't all have the same number of samples.
4. R has a function `aov()` to run ANOVA tests. See:
https://personality-project.org/r/r.anova.html#oneway
http://en.wikipedia.org/wiki/F-test

Example 4. The table shows patients' perceived level of pain (on a scale of 1 to 6) after 3 different medical procedures.

T_1	T_2	T_3
2	3	2
4	4	1
1	6	3
5	1	3
3	4	5

(1) Set up and run an F-test comparing the means of these 3 treatments.

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

answer: Using the code below, the F statistic is 0.325 and the p-value is 0.729 At any reasonable significance level we will fail to reject the null hypothesis.

Note, it is not reasonable to conclude the the null hypothesis is true. With just 5 data points per procedure we might simply lack the power to distinguish different means.

R code to perform the test

```
#DATA ----
T1 = c(2,4,1,5,3)
T2 = c(3,4,6,1,4)
T3 = c(2,1,3,3,5)

# One way ANOVA by hand ----
# Make sure all the groups are the same size
if (length(T1) != length(T2) || length(T2) != length(T3))
    stop("lengths are not equal", call.= F)
m = length(T1)
n = 3
group.means = c(mean(T1), mean(T2), mean(T3))
group.vars = c(var(T1), var(T2), var(T3))
betweenGroupVar = m*var(group.means) aveWithinGroupVar = mean(group.vars)
fstat = betweenGroupVar/aveWithinGroupVar
p = 1-pf(fstat, n-1,n*(m-1))
print(fstat)
print(p)
```

3.5 Chi-square test for goodness of fit

The distribution associated to this test is the chi-square distribution. It is denoted by $\chi^2(df)$ where the parameter df is called the degrees of freedom.

Suppose we have an unknown probability mass function given by the following table.

Outcomes	ω_1	ω_2	$\dots$	ω_n
Probabilities	p_1	p_2	$\dots$	p_n

In the chi-square test for goodness of fit we hypothesize a set of values for the probabilities. Typically we will hypothesize that the probabilities follow a known distribution with certain parameters, e.g. binomial, Poisson, multinomial. The test then tries to determine if this set of probabilities could have reasonably generated the data we collected.

- Use: Test whether discrete data fits a specific finite probability mass function.
- Data: An observed count O_i for each possible outcome ω_i.
- Assumptions: None
- H_0: The data was drawn from a specific discrete distribution.
- H_A: The data was drawn from a different distribution

- Test statistic: The data consists of observed counts O_i for each ω_i. From the null hypothesis probability table we get a set of expected counts E_i. There are two statistics that we can use:

$$\text{Likelihood ratio statistic } G = 2 * \sum O_i \ln\left(\frac{O_i}{E_i}\right)$$

$$\text{Pearson's chi-square statistic } X^2 = \sum \frac{(O_i - E_i)^2}{E_i}.$$

  It is a theorem that under the null hypthesis $X^2 \approx G$ and both are approximately chi-square. Before computers, X^2 was used because it was easier to compute. Now, it is better to use G although you will still see X^2 used quite often.

- Degrees of freedom *df*: Significance tests check whether the observed data is extreme compared to what is expected under the null hypothesis. For chi-square tests this requires considering all the possible cell counts consistent with the way we compute the expected cell counts. The degrees of freedom is the number of cell counts we can freely set and maintain consistency. More precisely, the degrees of freedom is the number of cells minus the number of parameters computed from the data. This will become more clear as we work examples.

  Here are two quick examples. Both use the following data, showing a set of outcomes and observed counts for these outcomes.

Outcomes	0	1	2	3	4	≥ 5
Observed counts	8	12	12	8	6	5

  Example 5. Suppose H_0 specifies that the counts are from a binomial(8, .5) distribution. There are 6 cells and we always need the total expected count to equal the total observed count. Nothing else is computed from the data so there are $6 - 1 = 5$ degrees of freedom. That is, we could set 5 of the cell counts freely and then the last one must be set to give the correct total.

  Example 6. This time we'll have H_0 be that the observations are all drawn from a binomial(8,θ) distribution. We then have to estimate θ from the data. (We'll use the MLE.). This time we compute two values from the data: the total number of counts and the estimate of θ. So, the degrees of freedom is $6 - 2 = 4$.

- Null distribution: Assuming H_0, both statistics (approximately) follow a chi-square distribution with *df* degrees of freedom. That is both $f(G \mid H_0)$ and $f(X^2 \mid H_0)$ have the same pdf as $Y \sim \chi^2(df)$.

- p-value:
  $p \;=\; P(Y > G) \;=\;$ `1 - pchisq(G, df)`
  $p \;=\; P(Y > X^2) \;=\;$ `1 - pchisq(`X^2`, df)`

Example 7. Mendel's genetic experiments (Adapted from Rice *Mathematical Statistics and Data Analysis, 2nd ed.*, example C, p.314)

In one of his experiments on peas Mendel crossed 556 smooth, yellow male peas with wrinkled green female peas. Assuming the smooth and wrinkled genes occur with equal frequency we'd expect 1/4 of the pea population to have two smooth genes (SS), 1/4 to have two wrinkled genes (ss), and the remaining 1/2 would be heterozygous Ss. We also expect these fractions for yellow (Y) and green (y) genes. If the color and smoothness genes are inherited independently and smooth and yellow are both dominant we'd expect the following table of frequencies for phenotypes.

	Yellow	Green	
Smooth	9/16	3/16	3/4
Wrinkled	3/16	1/16	1/4
	3/4	1/4	1

Probability table for the null hypothesis

So from the 556 crosses the expected number of smooth yellow peas is $556 \times 9/16 = 312.75$. Likewise for the other possibilities. Here is a table giving the observed and expected counts from Mendel's experiments.

	Observed count	Expected count
Smooth yellow	315	312.75
Smooth green	108	104.25
Wrinkled yellow	102	104.25
Wrinkled green	31	34.75

The null hypothesis is that the observed counts are random samples distributed according to the frequency table given above. We use the counts to compute our statistics

The likelihood ratio statistic is

$$\begin{aligned} G &= 2 * \sum O_i \ln\left(\frac{O_i}{E_i}\right) \\ &= 2 * \left(315 \ln\left(\frac{315}{412.75}\right) + 108 \ln\left(\frac{108}{104.25}\right) + 102 \ln\left(\frac{102}{104.25}\right) + 31 \ln\left(\frac{31}{34.75}\right)\right) \\ &= .618 \end{aligned}$$

Pearson's chi-square statistic is

$$X^2 = \sum \frac{(O_i - E_i)^2}{E_i} = \frac{2.75}{312.75} + \frac{3.75}{104.25} + \frac{2.25}{104.25} + \frac{3.75}{34.75} = .604$$

You can see that the two statistics are very close. This is usually the case. In general the likelihood ratio statistic is more robust and should be preferred.

Under the null hypothesis G follows a $\chi^2(3)$ distribution*. Using R to compute the p-value we get

$p =$ `1- pchisq(.618, 3) = .892`

Assuming the null hypothesis we would see data at least this extreme almost 90% of the time. We would not reject the null hypothesis for any reasonable significance level.

The p-value using Pearson's statistic is .985 –nearly identical.

Here is the R-code we used to do the computations in this example.

```
n = 556
prob = c(9,3,3,1)/16
observed = c(315,108,102,31)
expected = n*prob

# Likelihood ratio statistic
likRatioStat = 2*sum(observed*log(observed/expected))
p_likRatio = 1-pchisq(likRatioStat,3)
print(likRatioStat)
print(p_likRatio)


# Pearsons X^2 statistic
X2 = sum((observed-expected)^2/expected)
p_pearson = 1- pchisq(X2,3)
print(observed)
print(expected)
print(X2)
print(p_pearson)
```

*The number of degrees of freedom comes because there are four cells (one for each phenotype) and one relation among them (they sum to the total number of 556).