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

1 Learning Goals

1. Given hypotheses and data, be able 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 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. Very 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 might indicate that we should reject H_0.

3 Population parameters and sample statistics

Example 1. If we randomly select 10 men from a population and measure their heights we say that we have sampled the heights from the population. In this case the sample mean, say $\overline{x}$, is the mean of the sampled heights. It is a statistic and we know its value explicitly. On the other hand, the true average height of the population, say μ, is unknown and we can only estimate its value. We call μ a population parameter.

The main purpose of significance testing is to use sample statistics to draw conlusions about population parameters. For example, we might test if the average height of men in a given population is greater than 70 inches.

4 A gallery of common significance tests related to the normal distribution

We will show a number of tests that all assume normal data. For completeness we will include the z and t tests we've already explored.

You shouldn't try to memorize these tests. It is a hopeless task to memorize the tests given here and even more hopeless to memorize all the tests we've left out. 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. We will work through the details of these tests in class and on homework.

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 or the arguments showing how they arise as the null distributions in our significance tests. However, the arguments are accessible to anyone who knows calculus and is interested in understanding them. 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.

When analyzing data with any of these tests one thing of key importance is to verify that the assumptions are true or at least approximately true. For example, you shouldn't use a test that assumes the data is normal unless you've checked that the data is approximately normal.

The script class19.r contains examples of using R to run some of these tests. It is posted in our usual place for R code.

4.1 z-test

- Use: Test if the population mean equals a 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: $\phi(z \mid H_0)$ is the pdf of $Z \sim N(0, 1)$.
- p-value:

Two-sided:	$p = P(\lvert Z\rvert > z)$	=	`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)`

- R code: There does not seem to be a single R function in the base R packages that runs a z-test. There are other packages you can install that have a z.test function. Of course, it is easy enough to get R to compute the z score and p-value. There is an example of this in class19.r.

Example 2. 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 $\alpha = 0.05$?

Solution: Our test statistic is

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

The right-sided p-value is therefore

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

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

4.2 One-sample t-test of the mean

- Use: Test if the population mean equals a 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: $\quad s^2 = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(x_i - \overline{x})^2$
- Null distribution: $\phi(t \mid H_0)$ is the pdf of $\; T \sim t(n-1)$.
  (Student t-distribution with $n-1$ degrees of freedom)
- p-value:

Two-sided:	$p = P(\lvert T\rvert > t)$	$=$	`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)`

- R code example: For data $x = 1, 3, 5, 7, 2$ we can run a one-sample t-test with H_0: $\mu_0 = 2.5$ using the R command:

```
t.test(x, mu = 2.5, alternative=two.sided)
```

  This will return a several pieces of information including the mean of the data, t-value and the two-sided p-value. See the help for this function for other argument settings.

Example 3. Look in the class 18 notes or slides for an example of this test. The class 19 example R code also gives an example.

4.3 Two-sample t-test for comparing means

4.3.1 The case of equal variances

We start by describing the test assuming equal variances.

- Use: Test if the population means from two populations differ by a hypothesized amount.
- 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^2 is unknown, but the same for both groups.
- H_0: For a specified $\Delta\mu$ the difference of means $\mu_x - \mu_y = \Delta\mu$
- H_A:

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

- Test statistic: $t = \dfrac{\overline{x} - \overline{y} - \Delta\mu}{s_P}$,
  where s_x^2 and s_y^2 are the sample variances of the x and y data respectively, and s_P^2 is (sometimes called) the pooled sample variance:
  $$s_p^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}\left(\frac{1}{n} + \frac{1}{m}\right) \quad \text{and} \quad df = n+m-2$$
- Null distribution: $\phi(t \mid H_0)$ is the pdf of $T \sim t(df)$, the t-distribution with $df = n + m - 2$ degrees of freedom.
- p-value:

Two-sided:	$p = P(\lvert T\rvert > t)$	=	`2*(1-pt(abs(t), df))`
one-sided-greater:	$p = P(T > t)$	=	`1 - pt(t, df)`
one-sided-less:	$p = P(T < t)$	=	`pt(t, df)`
- R code: The R function `t.test` will run a two-sample t-test. See the example code in class19.r. In `t.test` the argument `mu` is used for what we have called $\Delta\mu$.

Notes: 1. Most often the test is done with $\Delta\mu = 0$. That is, the null hypothesis is the the means are equal, i.e. $\mu_x - \mu_y = 0$.

2. If the x and y data have the same length $n = m$, then the formula for s_p^2 becomes simpler:
$$s_p^2 = \frac{s_x^2 + s_y^2}{n}$$

Example 4. Look in the class 18 notes or slides for an example of the two-sample t-test.

4.3.2 The case of unequal variances

There is a form of the t-test for when the variances are not assumed equal. It is sometimes called Welch's t-test.

This looks exactly the same as the case of equal except for a small change in the assumptions and the formula for the pooled variance:

- Use: Test if the population means from two populations differ by a hypothesized amount.
- 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:
  $$x_i \sim N(\mu_x, \sigma_x^2)$$
  $$y_j \sim N(\mu_y, \sigma_y^2)$$
  where both μ_x and μ_y are unknown and possibly different. The variances σ_x^2 and σ_y^2 are unknown and not assumed to be equal.
- H_0, H_A: Exactly the same as the case of equal variances.

- Test statistic: $t = \dfrac{\overline{x} - \overline{y} - \Delta\mu}{s_P}$,
  where s_x^2 and s_y^2 are the sample variances of the x and y data respectively, and s_P^2 is (sometimes called) the pooled sample variance:
  $$s_p^2 = \frac{s_x^2}{n} + \frac{s_y^2}{m} \quad \text{and} \quad df = \frac{(s_x^2/n + s_y^2/m)^2}{(s_x^2/n)^2/(n-1) + (s_y^2/m)^2/(m-1)}$$
- Null distribution: $\phi(t\,|\,H_0)$ is the pdf of $T \sim t(df)$, the t distribution with df degrees of freedom.
- p-value: Exactly the same as the case of equal variances.
- R code: The function `t.test` also handles this case if you set the argument `var.equal=FALSE`.

Notes. 1. In truth, the null distribution given above only approximates the exact null distribution.

2. Notice that the degrees of freedom are unlikely to be a whole number.

3. Some people recommend always using Welch's t-test, even if the variances are believed to be equal. This avoids making the assumption that the variances are equal and has very little downside if they are equal.

4.3.3 The paired two-sample t-test

When the data naturally comes in pairs (x_i, y_i), we can use the paired two-sample t-test. (After checking the assumptions are valid!)

Example 5. To measure the effectiveness of a cholesterol lowering medication we might test each subject before and after treatment with the drug. So for each subject we have a pair of measurements:

$$\begin{aligned} x_i &= \text{cholesterol level before treatment} \\ y_i &= \text{cholesterol level after treatment.} \end{aligned}$$

Example 6. To measure the effectiveness of a cancer treatment we might pair each subject who received the treatment with one who did not. In this case we would want to pair subjects who are similar in terms of stage of the disease, age, sex, etc.

- Use: Test if the average difference between paired values in a population equals a hypothesized value.
- Data: $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$ must have the same length.
- Assumptions: The differences $w_i = x_i - y_i$ between the paired samples are independent draws from a normal distribution $\mathrm{N}(\mu, \sigma^2)$, where 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: $t = \dfrac{\overline{w} - \mu_0}{s/\sqrt{n}}$,

  where s^2 is the sample variance: $s^2 = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(w_i - \overline{w})^2$

- Null distribution: $\phi(t \,|\, H_0)$ is the pdf of $T \sim t(n-1)$.
  (Student t-distribution with $n-1$ degrees of freedom)

- p-value:

Two-sided:	$p = P(\lvert T\rvert > t)$	=	`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)`

- R code: The R function `t.test` will do a paired two-sample test if you set the argument `paired=TRUE`. You can also run a one-sample t-test on $x - y$. There are examples of both of these in class19.r

Notes. 1. This is just a one-sample t-test using w_i.

2. Another way to write the assumption is that we assume a relation between x_i and y_i of the form $y_i = x_i + \mu + e$. Here μ is some (unknown) constant, and e is random error (noise) of mean 0 and (unknown) variance σ^2.

Example 7. The following example is taken from Rice [1]

To study the effect of cigarette smoking on platelet aggregation Levine (1973) drew blood samples from 11 subjects before and after they smoked a cigarette and measured the extent to which platelets aggregated. Here is the data:

Before	25	25	27	44	30	67	53	53	52	60	28
After	27	29	37	56	46	82	57	80	61	59	43
Difference	2	4	10	12	16	15	4	27	9	-1	15

The null hypothesis is that smoking had no effect on platelet aggregation, i.e. that the difference between before and after should have mean $\mu_0 = 0$. We ran a paired two-sample t-test to test this hypothesis. Here is the R code: (It's also in class19.r.)

```
before.cig = c(25,25,27,44,30,67,53,53,52,60,28)
after.cig = c(27,29,37,56,46,82,57,80,61,59,43)
mu0 = 0
result = t.test(after.cig, before.cig, alternative="two.sided", mu=mu0, paired=TRUE)
print(result)
```

Here is the output:

```
    Paired t-test
data: after.cig and before.cig
t = 4.2716, df = 10, p-value = 0.001633
alternative hypothesis: true difference in means is not equal to 0
mean of the differences: 10.27273
```

We got the same results with the one-sample t-test:

```
t.test(after.cig - before.cig, mu=0)
```

[1] John Rice, *Mathematical Statistics and Data Analysis*, 2nd edition, p. 412. This example references P.H Levine (1973) An acute effect of cigarette smoking on platelet function. *Circulation, 48, 619-623.*

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

- Use: Test if the population means from n groups are all the same.
- Data: (n groups, m samples from each group)

  $$\begin{array}{llll} x_{1,1}, & x_{1,2}, & \ldots, & x_{1,m} \\ x_{2,1}, & x_{2,2}, & \ldots, & x_{2,m} \\ & & \ldots & \\ x_{n,1}, & x_{n,2}, & \ldots, & x_{n,m} \end{array}$$

- Assumptions: Data for each group is an independent normal sample drawn from distributions with (possibly) different means but the same variance:

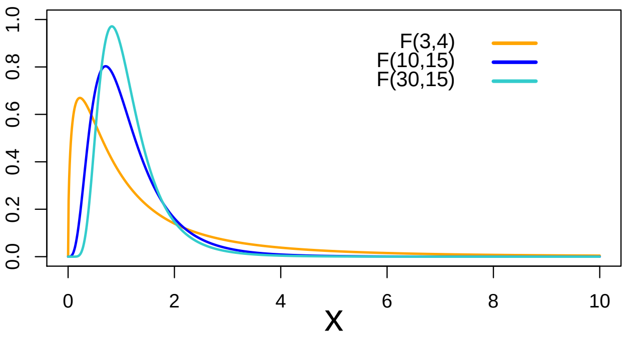
  $$\begin{array}{ll} x_{1,j} & \sim N(\mu_1, \sigma^2) \\ x_{2,j} & \sim N(\mu_2, \sigma^2) \\ & \ldots \\ x_{n,j} & \sim N(\mu_n, \sigma^2) \end{array}$$

  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: $f = \frac{\text{MS}_B}{\text{MS}_W}$, where

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

- Idea: If the μ_i are all equal, test statistic f, which is a 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 f should be larger. We won't give a proof of this.
- Null distribution: $\phi(f \mid H_0)$ is the pdf of $F \sim F(n-1, n(m-1))$.
  This is the F-distribution with $(n-1)$ and $n(m-1)$ degrees of freedom. Several F-distributions are plotted below.
- p-value: $p = P(F > f) =$ `1- pf(f, 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.
5. See: https://en.wikipedia.org/wiki/F-test

Example 8. 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?

Solution: 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 that the average pain level is the same for all three treatments..

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)

procedure = c(rep('T1',length(T1)),rep('T2',length(T2)),rep('T3',length(T3)))
pain = c(T1,T2,T3)
data.pain = data.frame(procedure,pain)
aov.data = aov(pain~procedure,data=data.pain) # do the analysis of variance
print(summary(aov.data)) # show the summary table

# class19.r also shows code to compute the ANOVA by hand.
```

The summary shows a p-value (shown as `Pr(>F)`) of 0.729. Therefore we do not reject the null hypothesis that all three group population means are the same.

4.5 Chi-square test for goodness of fit

This is a test of how well a hypothesized probability distribution fits a set of data. The test statistic is called a chi-square statistic and the null distribution associated to the chi-square statistic 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	...	ω_n
Probabilities	p_1	p_2	...	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 reasonably have 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 hypothesis $X^2 \approx G$ 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: For chi-square tests the number of degrees of freedom can be a bit tricky. In this case $df = n - 1$. It is computed as the number of cell counts that can be freely set under H_A consistent with the statistics needed to compute the expected cell counts assuming H_0.
- Null distribution: Assuming H_0, both statistics (approximately) follow a chi-square distribution with df degrees of freedom. That is both $\phi(G \,|\, H_0)$ and $\phi(X^2 \,|\, H_0)$ have (approximately) the same pdf as $Y \sim \chi^2(df)$.
- p-value: Extreme data means large values of X^2, i.e. large differences between the observed and expected counts. So,

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

- R code: The R function `chisq.test` can be used to do the computations for a chi-square test use X^2. For G you either have to do it by hand or find a package that has a function. (It will probably be called `likelihood.test` or `G.test`.

Notes. 1. When the likelihood ratio statistic G is used the test is also called a G-test or a likelihood ratio test.

Example 9. First chi-square example. Suppose we have an experiment that produces numerical data. For this experiment the possible outcomes are 0, 1, 2, 3, 4, 5 or more. We run 51 trials and count the frequency of each outcome, getting the following data:

Outcomes	0	1	2	3	4	≥ 5
Observed counts	3	10	15	13	7	3

Suppose our null hypothesis H_0 is that the data is drawn from 51 trials of a binomial(8, 0.5) distribution and our alternative hypothesis H_A is that the data is drawn from some other distribution. Do all of the following:

1. Make a table of the observed and expected counts.
2. Compute both the likelihood ratio statistic G and Pearson's chi-square statistic X^2.
3. Compute the degrees of freedom of the null distribution.
4. Compute the p-values corresponding to G and X^2.

Solution: All of the R code used for this example is in class19.r.

1. Assuming H_0 the data truly comes from a binomial(8, 0.5) distribution. We have 51 total observations, so the expected count for each outcome is just 51 times its probability. We computed the binomial(8, 0.5) probabilities and expected counts in R:

Outcomes	0	1	2	3	4	≥ 5
Observed counts	3	10	15	13	7	3
H_0 probabilities	0.0039	0.0313	0.1094	0.2188	0.2734	0.3633
Expected counts	0.20	1.59	5.58	11.16	13.95	18.53

2. Using the formulas above we compute that $X^2 = 116.41$ and $G = 66.08$

3. The only statistic used in computing the expected counts was the total number of observations 51. So, the degrees of freedom is 5, i.e we can set 5 of the cell counts freely and the last is determined by requiring that the total number is 51.

4. The p-values are $pG =$`1 - pchisq(G, 5)` and $pX2 =$ `1 - pchisq(`X^2`, 5)`. Both p-values are effectively 0. For almost any significance level we would reject H_0 in favor of H_A.

4.5.1 Degrees of freedom in chi-square tests

We alreay gave a quick definition of degrees of freedom for a chi-square test. Here we will try to go a little slower in showing how to compute degrees of freedom.

To start, recall that in a chi-square test, our table has n observed counts. Then, we use observed counts and the null hypothesis to compute n expected counts. This is typically done by computing some statistics and using them estimate the parameters needed to compute the expected counts. For example, in the previous example the statistic computed was the total number of counts.

Now, imagine that we are allowed to fabricate the n observed counts, but we demand that our made up observations produce the same statistics as the true observed counts. That is, our imaginary observed counts need to produce the same expected counts as the true data. The degrees of freedom is the number of fake observed counts we can freely choose. The rest will be determined by our constraint that they produce the same statistics.

Example 10. (Degrees of freedom.) Suppose we have the following observed counts

Outcomes	0	1	2	3	4	5
Observed counts	6	9	13	12	7	3

Suppose our null hypothesis H_0 is that the data producing these counts was drawn from 50 trials of a binomial(5, θ) distribution. Our alternative hypothesis H_A is that the data is drawn from some other distribution. Run a chi-square test of these hypotheses with significance 0.05.

Solution: To compute expected counts we need a value of θ. Since it is not known, we have to use the data to estimate it.

The total number of observations is 50. So, the mean of the data is

$$m = \frac{6 \cdot 0 + 9 \cdot 1 + 13 \cdot 2 + 12 \cdot 3 + 7 \cdot 4 + 3 \cdot 5}{50} = 2.28.$$

The expected value of a binom(5,θ) distribution, is 5θ. The maximum likelihood estimate for θ is $\hat{\theta} = m/5 = 0.456$.

Now, just like in the previous example, we can compute expected counts for each possible outcome. The expected count of outcome k is $50 \cdot p(k)$. In R this is `50*dbinom(k, 50,` $\hat{\theta}$`)`. We have the following table

Outcomes	0	1	2	3	4	5
Observed counts	6	9	13	12	7	3
Expected counts	2.38	9.98	16.74	14.03	5.88	0.99

To determine the degrees of freedom:
(i) We have 6 observed counts.

(ii) To compute the expected counts we need the total number of counts = 50 and our estimate of $\hat{\theta} = m/5$ That is, we have two constraints: the total number of counts is 50, and mean $m = 2.28$.

So, to get the same expected counts, we could choose 4 of the observed counts freely and then set last two counts so the constraints are met. Thus, there are 4 degrees of freedom.

More briefly: 6 observed counts - 2 constraints = 4 degrees of freedom.

Using R we can compute $G = 8.01$, with p-value 0.09. Thus, at significance level 0.05 we would not reject the null hypothesis.

Note, the X^2 statistic is 11.05 with p-value 0.026. Clearly this example is a borderline case, since we reject H_0 when using X^2 and we don't when using G.

It bears repeating, for reasons like this, we never say H_0 is false. The most we can say is that the data does not support rejecting H_0.

4.5.2 More examples

Example 11. 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 looked at 2 genetic trait pairs: smooth/wrinkled and yellow/green. Symbolically we label a smooth gene S and a wrinkled gene s. Likewise we use Y and y for yellow and green respectively.

Mendel started by selecting a parent generation of homozygous plants. They were either smooth/yellow (genes $SSYY$) and wrinkled/green (genes $ssyy$). He crossed the smooth/yellow with the wrinkled/green peas creating the, so called, F_1 generation consisting of plants with genes $SsYy$. Since smooth (S) and yellow Y are both dominant traits, all these plants were smooth/yellow.

He then crossed 556 pairs of the F_1 generation to create the F_2 generation. We would expect 1/4 of the F_2 generation to have two smooth genes (SS), 1/4 to have two wrinkled genes (ss), and the remaining 1/2 to 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

$$\begin{aligned} G &= 2 * \sum O_i \ln\left(\frac{O_i}{E_i}\right) \\ &= 2 * \left(315 \ln\left(\frac{315}{412.75}\right) + 108 \ln\left(\frac{108}{104.25}\right) + 102 \ln\left(\frac{102}{104.25}\right) + 31 \ln\left(\frac{31}{34.75}\right)\right) \\ &= 0.618 \end{aligned}$$

Pearson's chi-square statistic is

$$X^2 = \sum \frac{(O_i - E_i)^2}{E_i} = \frac{2.75^2}{312.75} + \frac{3.75^2}{104.25} + \frac{2.25^2}{104.25} + \frac{3.75^2}{34.75} = 0.604$$

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.

The degrees of freedom is 3, because there are 4 observed quantities and one relation between them, i.e. they sum to 556. So, under the null hypothesis G follows a $\chi^2(3)$ distribution. Using R to compute the p-value we get

$$p = \texttt{1- pchisq(0.618, 3) = 0.8923}$$

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 0.8955 –nearly identical.

The script class19.r shows these calculations and also how to use `chisq.test` to run a chi-square test directly.

4.6 Chi-square test for homogeneity

This is a test to see if several independent sets of random data are all drawn from the same distribution. (The meaning of homogeneity in this case is that all the distributions are the same.)

- Use: Test whether m different independent sets of discrete data are drawn from the same distribution.
- Outcomes: ω_1, ω_2, …, ω_n are the possible outcomes. These are the same for each set of data.
- Data: We assume m independent sets of data giving counts for each of the possible outcomes. That is, for data set i we have an observed count $O_{i,j}$ for each possible outcome ω_j.
- Assumptions: None
- H_0: Each data set is drawn from the same distribution. (We don't specify what this distribution is.)
- H_A: The data sets are not all drawn from the same distribution.
- Test statistic: See the example below. There are mn cells containing counts for each outcome for each data set. Using the null distribution we can estimate expected counts for each of the data sets. The statistics X^2 and G are computed exactly as above.
- Degrees of freedom df: $(m-1)(n-1)$. (See the example below.)
- The null distribution $\chi^2(df)$. The p-values are computed just as in the chi-square test for goodness of fit.
- R code: The R function `chisq.test` can be used to do the computations for a chi-square test use X^2. For G you either have to do it by hand or find a package that has a function. (It will probably be called `likelihood.test` or `G.test`.

Example 12. Someone claims to have found a long lost work by William Shakespeare. They ask you to test whether or not the play was actually written by Shakespeare .

You go to https://www.opensourceshakespeare.org and pick a random 12 pages from *King Lear* and count the use of common words. You do the same thing for the 'long lost work'. You get the following table of counts.

Word	a	an	this	that
King Lear	150	30	30	90
Long lost work	90	20	10	80

Using this data, set up and evaluate a significance test of the claim that the long lost book is by William Shakespeare. Use a significance level of 0.1.

Solution: The null hypothesis H_0: For the 4 words counted the long lost book has the same relative frequencies as the counts taken from *King Lear*.

The total word count of both books combined is 500, so the the maximum likelihood estimate of the relative frequencies assuming H_0 is simply the total count for each word divided by the total word count.

Word	a	an	this	that	Total count
King Lear	150	30	30	90	300
Long lost work	90	20	10	80	200
totals	240	50	40	170	500
rel. frequencies under H_0	240/500	50/500	40/500	170/500	500/500

Now the expected counts for each book under H_0 are the total count for that book times the relative frequencies in the above table. The following table gives the counts: (observed, expected) for each book.

Word	a	an	this	that	Totals
King Lear	(150, 144)	(30, 30)	(30, 24)	(90, 102)	(300, 300)
Long lost work	(90, 96)	(20, 20)	(10, 16)	(80, 68)	(200, 200)
Totals	(249, 240)	(50, 50)	(40, 40)	(170, 170)	(500, 500)

The chi-square statistic is

$$\begin{aligned} X^2 &= \sum \frac{(O_i - E_i)^2}{E_i} \\ &= \frac{6^2}{144} + \frac{0^2}{30} + \frac{6^2}{24} + \frac{12^2}{102} + \frac{6^2}{96} + \frac{0^2}{20} + \frac{6^2}{16} + \frac{12^2}{68} \\ &\approx 7.9 \end{aligned}$$

There are 8 cells and all the marginal counts are fixed because they were needed to determine the expected counts. To be consistent with these statistics we could freely set the values in 3 cells in the table, e.g. the 3 blue cells, then the rest of the cells are determined in order to make the marginal totals correct. Thus $df = 3$. (Or we could recall that $df = (m-1)(n-1) = (3)(1) = 3$, where m is the number of columns and n is the number of rows.)

Using R we find `p = 1-pchisq(7.9,3) = 0.048`. Since this is less than our significance level of 0.1 we reject the null hypothesis that the relative frequencies of the words are the same in both books.

If we make the further assumption that all of Shakespeare's plays have similar word frequencies (which is something we could check) we conclude that the book is probably not by Shakespeare.

4.7 Other tests

There are far too many other tests to even make a dent. We will see some of them in class and on psets. Again, we urge you to master the paradigm of NHST and recognize the importance of choosing a test statistic with a known null distribution.