STAT 135 Lab 11
Tests for Categorical Data (Fisher's Exact test, χ^2 tests for Homogeneity and Independence) and Linear Regression

Rebecca Barter

April 20, 2015

Fisher's Exact Test

Fisher's Exact Test

A **categorical variable** is a variable that can take a finite number of possible values, thus assigning each individual to a particular group or "category".

We have seen categorical variables before, for example, our ANOVA factors:

- pen color
- gender
- beer type

However, for ANOVA we also had a continuous response variable.

Fisher's Exact Test

Purely categorical data are summarized in the form of a **contingency table**, where the entries in the table describe how many observations fall into each grouping. For example

	Right-handed	Left-handed	Total
Males	43	9	52
Females	44	4	48
Totals	87	13	100

The total number of individuals represented in the contingency table, is the number in the bottom right corner.

Fisher's Exact Test

	Right-handed	Left-handed	Total
Males	43	9	52
Females	44	4	48
Totals	87	13	100

We might want to know whether the proportion of right-handed-ness is significantly different between males and females.

We can use **Fisher's exact test** to test this hypothesis.

- It is called an exact test because we know the distribution of the test statistic *exactly* rather than just approximately or asymptotically.

Fisher's Exact Test

Suppose that we have two categorical factors, A and B, each of which has two-levels, and the number of observations in each group is given as follows:

	A_1	A_2	Total
B_1	N_{11}	N_{12}	$n_{1\cdot}$
B_2	N_{21}	N_{22}	$n_{2\cdot}$
Totals	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot\cdot}$

Then under the **null hypothesis** that there is no difference between the proportions of the levels for factor A or B, i.e. that the two factors are **independent**, then the distribution of N_{11} is **hypergeometric**.

Fisher's Exact Test

The $Hypergeometric(N, K, n)$ distribution describes the probability of k successes in n draws, without replacement, from a finite population of size N that contains exactly K successes, wherein each draw is either a success or failure.

If $X \sim Hypergeometric(N, K, n)$, then

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

and we note that

$$E(X) = \frac{nK}{N}$$

Fisher's Exact Test

	A_1	A_2	Total
B_1	N_{11}	N_{12}	$n_{1\cdot}$
B_2	N_{21}	N_{22}	$n_{2\cdot}$
Totals	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot\cdot}$

Under H_0, the dist of N_{11} is $Hypergeometric(N, K, n)$, where

$$N = n_{\cdot\cdot} \ , \quad K = n_{1\cdot} \ , \quad n = n_{\cdot 1}$$

So

$$P(N_{11} = n_{11}) = \frac{\binom{n_{1\cdot}}{n_{11}}\binom{n_{2\cdot}}{n_{21}}}{\binom{n_{\cdot\cdot}}{n_{\cdot 1}}}$$

and

$$E(N_{11}) = \frac{n_{\cdot 1} n_{1\cdot}}{n_{\cdot\cdot}}$$

Note that under the symmetry of the problem, I could transpose the matrix and get the same answer (I usually just think about what do I want to be considered a "success").

Fisher's Exact Test

	A_1	A_2	Total
B_1	N_{11}	N_{12}	$n_{1\cdot}$
B_2	N_{21}	N_{22}	$n_{2\cdot}$
Totals	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot\cdot}$

To conduct the hypothesis test, we can think about rejecting H_0 for extreme values of N_{11}

- N_{11} is our **test statistic**

Fisher's Exact Test

A **two-sided alternative** hypothesis can be stated in several equivalent ways, for example:

- the proportion of right-handedness differs between men and women
- the proportion of women differs between right-handed people and left-handed people
- right/left handedness and gender are independent

The two-sided p-value can be written as

$$P\left(|N_{11} - E(N_{11})| \geq |n_{11} - E(N_{11}|\right)$$

Fisher's Exact Test

A **one-sided alternative** hypothesis can also be stated in several equivalent ways, for example:

- the proportion of right-handedness is higher (lower) for men than for women
- the proportion of men is higher (lower) for right-handed people than for left-handed people

The one-sided p-value can be written as

$$P\left(N_{11} \geq n_{11}\right) \qquad \text{or} \qquad P\left(N_{11} \leq n_{11}\right)$$

Fisher's Exact Test

For our example

	Right-handed	Left-handed	Total
Males	43	9	52
Females	44	4	48
Totals	87	13	100

$$N_{11} \sim Hypergeoetric(100, 87, 52)$$

$$n_{11} = 43 \qquad E(N_{11}) = \frac{n_{\cdot 1} n_{1\cdot}}{n_{\cdot\cdot}} = \frac{52 \times 87}{100} = 45.24$$

So that our two-sided p-value is given by

$$\begin{aligned} P(|N_{11} - E(N_{11})| \geq |n_{11} - E(N_{11}|) &= P(|N_{11} - 45.24| \geq 2.24) \\ &= P(N_{11} - 45.24 \geq 2.24) + P(N_{11} - 45.24 \leq -2.24) \\ &= P(N_{11} \geq 47.48) + P(N_{11} \leq 43) \\ &= 0.24 \end{aligned}$$

Exercise

Exercise: Fisher's Exact Test (Rice 13.8.19)

- A psychological experiment was done to investigate the effect of anxiety on a person's desire to be alone or in company.
- A group of 30 subjects was randomly divided into two groups of sizes 13 and 17.
- The subjects were told that they would be subject to electric shocks.
  - The "high anxiety" group was told that the shocks would be quite painful
  - The "low-anxiety" group was told that they would be mild and painless.
- Both groups were told that there would be a 10-min wait before the experiment began and each subject was given the choice of waiting alone or with other subjects.

The results are as follows:

	Wait Together	Wait Alone	Total
High-Anxiety	12	5	17
Low-Anxiety	4	9	13
Total	16	14	30

Test whether there is a significant difference between the high- and low-anxiety groups.

χ^2 Test for Homogeneity

χ^2 Test for Homogeneity

Suppose now that instead of a 2×2 table, we have an arbitrary $I \times J$ table, and we want to see if the count proportions are differently distributed across different populations.

For example, suppose we are interested in whether TV show preference differs significantly between different age groups.

	18 to 30	30 to 45	Total
Game of Thrones	40	31	71
House of Cards	25	35	60
Orange is the New Black	30	23	53
Total	95	89	184

χ^2 Test for Homogeneity

Suppose that we have J multinomial distributions, each having I categories. If the probability of the ith category of the jth multinomial is denoted π_{ij}, the null hypothesis to be tested is

$$H_0 : \pi_{i1} = \pi_{i2} = ... = \pi_{i1}, \qquad i = 1, ..., I$$

However, we can view this as a **goodness-of-fit** test: Does the model prescribed by the null hypothesis fit the data?
Recall that Pearson's χ^2 statistic is given by

$$\begin{aligned} X^2 &= \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \\ &= \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - n_{i.}n_{.j}/n_{..})^2}{n_{i.}n_{j.}/n_{..}} \end{aligned}$$

χ^2 Test for Homogeneity

The degrees of freedom for the χ^2 statistic are the number of independent counts minus the number of independent parameters estimated from the data. Under the assumption of homogeneity, we have

- $J(I-1)$ independent counts, since each of the J multinomials has $I-1$ independent counts.
- $(I-1)$ independent parameters that have been estimated since the totals for each multinomial are fixed

The degrees of freedom for our test statistic is thus

$$df = J(I-1) - (I-1) = (I-1)(J-1)$$

χ^2 Test for Homogeneity

For the null hypothesis of homogeneity in the populations:

$$H_0 : \pi_{i1} = \pi_{i2} = ... = \pi_{i1}, \qquad i = 1, ..., I$$

our p-value is given by

$$\text{p-value} = P\left(\chi^2_{(I-1)(J-1)} \geq X^2\right)$$

χ^2 Test for Homogeneity

So for our example, our observed counts are given by

	18 to 30	30 to 45	Total
Game of Thrones	$O_{11} = 40$	$O_{12} = 31$	71
House of Cards	$O_{21} = 25$	$O_{22} = 35$	60
Orange is the New Black	$O_{31} = 30$	$O_{32} = 23$	53
Total	95	89	184

and our expected counts under H_0 are given by

	18 to 30	30 to 45
Game of Thrones	$E_{11} = \frac{95 \times 71}{184} = 36.7$	$E_{12} = \frac{89 \times 71}{184} = 34.3$
House of Cards	$E_{21} = \frac{95 \times 60}{184} = 31.0$	$E_{22} = \frac{89 \times 60}{184} = 29.0$
Orange is the New Black	$E_{31} = \frac{95 \times 53}{184} = 27.4$	$E_{32} = \frac{89 \times 53}{184} = 25.6$

χ^2 Test for Homogeneity

So for our example, our observed (expected) counts are given by

	18 to 30	30 to 45	Total
Game of Thrones	40 (36.7)	31 (34.3)	71
House of Cards	25 (31.0)	35 (29.0)	60
Orange is the New Black	30 (27.4)	23 (25.6)	53
Total	95	89	184

so our test statistic is given by

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \frac{(40 - 36.7)^2}{36.7} + \frac{(25 - 31)^2}{31}$$
$$+ \frac{(30 - 27.4)^2}{27.4} + \frac{(31 - 34.3)^2}{34.3}$$
$$+ \frac{(35 - 29)^2}{29} + \frac{(23 - 25.6)^2}{25.6}$$
$$\approx 3.528$$

χ^2 Test for Homogeneity

Our test statistic is given by

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 3.528$$

and thus the p-value for the test is

$$P\left(\chi^2_{(I-1)(J-1)} \geq X^2\right) = P\left(\chi^2_2 \geq 3.528\right) = 0.17$$

which is larger than 0.05, so we fail to reject our null hypothesis and conclude that there is no significant difference between TV preferences for the different age groups.

χ^2 Test for Independence

χ^2 Test for Independence

The χ^2 test for independence is eerily similar to (read *exactly the same as*) the χ^2 test for homogeneity, but is aimed at answering a slightly different question.

Suppose that a psychologist wants to test whether there is a relationship between personality and color preference. Then they are testing the null hypothesis that color preference and personality are independent. They have observed the following counts

	Blue	Red	Yellow	Total
Extroverted	5	20	5	30
Introverted	10	5	5	20
Total	15	25	10	50

We again use the X^2 test statistic.

χ^2 Test for Independence

The degrees of freedom for the χ^2 statistic are the number of independent counts minus the number of independent parameters estimated from the data. Under the assumption of independence, we have

- $IJ - 1$ independent counts, since we have IJ cells, with any one cell entirely determined by the sum of the others.
- $(I - 1) + (J - 1)$ independent parameters that have been estimated to give the marginal probabilities that determine the expected counts

The degrees of freedom for our test statistic is thus

$$df = (IJ - 1) - [(I - 1) + (J - 1)] = (I - 1)(J - 1)$$

χ^2 Test for Independence

For the null hypothesis of indpendence:

H_0 : The row factor is independent of the column factor

our p-value is given by

$$\text{p-value} = P\left(\chi^2_{(I-1)(J-1)} \geq X^2\right)$$

which is **exactly the same** as the χ^2 test for homogeneity.

Exercise

Exercise: χ^2 Test for Independence

Determine whether there is a relationship between color preference and personality in the psychologists's experiment

	Blue	Red	Yellow	Total
Extroverted	5	20	5	30
Introverted	10	5	5	20
Total	15	25	10	50

Simple Linear Regression

Simple Linear Regression

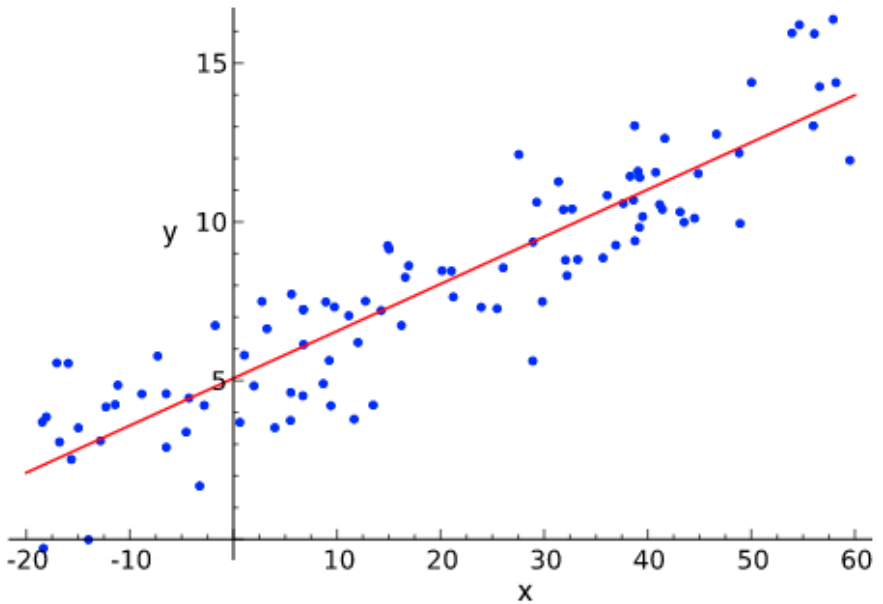
We have spent the last couple of weeks identifying if variables of interest have an effect on some response. Now we turn to asking the question of *how* our variables of interest affect the response.

For example, we might ask "if I increase the value of variable x by one unit, what happens to the response, y?".

Simple linear regression refers to the case when we have only **one explanatory variable**, x.

Simple Linear Regression

The idea behind linear regression by least squares is to identify the *line of best fit* when we plot y against x by minimizing the sum of the squared distances from the points to the line.



The red line above will have the form

$$y = \beta_0 + \beta_1 x$$

and our goal is to estimate the intercept β_0 and the slope β_1 using our observed data.

Simple Linear Regression

The statistical model corresponding to our data is given by

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where ϵ_i corresponds to the unobserved random noise which explains the deviation from the line $\beta_0 + \beta_1 x_i$. ϵ_i satisfies $E(\epsilon_i) = 0$ and $Var(\epsilon_i) = \sigma^2$ (x_i is considered fixed).

We estimate β_0 and β_1 by finding the values, $\hat{\beta}_0$ and $\hat{\beta}_1$, that minimize the following equation

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

which corresponds to minimizing the vertical distance from each observed point, y_i, to the corresponding point on the fitted line $\hat{\beta}_0 + \hat{\beta}_1 x_i$ (i.e. we want to minimize the **residuals**).

Exercise

Exercise: Simple Linear Regression (Rice 14.9.10)

Show that the values of β_0 and β_1 that minimize

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

are given by

$$\boxed{\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}}$$

$$\boxed{\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}}$$

Simple Linear Regression

Note that these estimators can be written as

$$\boxed{\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}}$$

$$\boxed{\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{Cov(x,y)}{Var(x)} = \frac{s_{xy}}{s_x} = \rho\sqrt{\frac{s_y}{s_x}}}$$

where $\rho = \frac{s_{xy}}{\sqrt{s_x s_y}}$ is the **correlation coefficient** between x and y. Note that

$$|\rho| \leq 1$$

The correlation describes how much of a linear relationship there is between x and y. A strong linear relationship will have $|\rho|$ close to 1 (however the converse is not necessarily true – always make plots to check!)

Simple Linear Regression

Some theoretical results about these estimators:
$\hat{\beta}_0$ and $\hat{\beta}_1$ are **unbiased**:

$$E(\hat{\beta}_j) = \beta_j$$

Moreover, we can calculate the **variance and covariance** of these estimators to find that

$$Var(\hat{\beta}_0) = \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2}$$

$$Var(\hat{\beta}_1) = \frac{n\sigma^2}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2}$$

$$Cov(\hat{\beta}_0, \hat{\beta}_1) = \frac{-\sigma^2 \sum_{i=1}^n x_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2}$$

Question: where does the randomness come from?

Simple Linear Regression

To estimate the variance of the estimators, we need to know σ^2, which we rarely do. We can, however, estimate σ^2. Recall that our model is

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma^2)$.

Then we can rearrange as follows:

$$\epsilon_i = y_i - \beta_0 - \beta_1 x_i$$

so perhaps we can estimate σ^2 by looking at the variances of the ϵ_i... but we don't observe ϵ_i!

Simple Linear Regression

$$\epsilon_i = y_i - \beta_0 - \beta_1 x_i$$

We can estimate ϵ_i by substituting $\widehat{\beta}_0$ and $\widehat{\beta}_1$, so we define the i^{th} **residual** to be

$$\boxed{e_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i}$$

so let's estimate σ^2 by the variance of the residuals:

$$\boxed{\widehat{\sigma}^2 = \frac{RSS}{n-p} = \frac{\sum_{i=1}^{n} e_i^2}{n-p}}$$

where p is the number of variables we have ($p = 2$ for simple linear regression)

Exercise

Exercise: Simple Linear Regression

The `faithful` dataset in `R` contains data on (1) the waiting time (in mins) between eruptions and the duration of the eruption (in mins) for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA. We have 272 observations.

Exercise: Simple Linear Regression

The `faithful` dataset in `R` contains data on (1) the waiting time (in mins) between eruptions and the duration of the eruption (in mins) for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA. We have 272 observations.

- Plot duration versus waiting time.
- Apply the simple linear regression model where we are regressing duration on waiting time.
- Report the fitted regression line, as well as the variance and covariance of the parameter estimates.
- Estimate the next eruption duration if the waiting time since the last eruption has been 80 minutes ago.