MATH 1052H - S62 - Lecture 06

Goodness-of-Fit

Our topic for this lecture is Goodness of Fit Tests, the χ^2 (chi-square) distribution, and Contingency Tables. This is our last **categorical** statistical model.

Key Idea

Given some data, do we believe the data could have come from a particular distribution?

Do a test!

Definition

A **goodness-of-fit** test is used to test the hypothesis that an observed frequency distribution fits some claimed distribution.

Note: the data will be discrete and divided into categories.

New Notation

We will use O and E extensively this week:

- O: observed frequency
- E: expected frequency (under null hypothesis assumption)

Test Statistic

Just like the tests we did in the first few weeks of the term, we will compute a **test statistic**, and then use one of our three methods to determine the outcome of the hypothesis test. The statistic for goodness-of-fit is:

$$\chi^2_{\text{test}} = \sum \frac{(O-E)^2}{E}$$

Note: when performing the test, we will always use **right-tailed** (single tail!) testing.

Expected Frequencies?

How do we find the **expected** frequencies? We use one of two designs for our test:

- All expected frequencies are equal, that is, $E = n/k$ for n total observations, and k categories
- All expected frequencies are not equal, so $E = np_i$, for p_i the probability of observing category i, for $i = 1, 2, \cdots, k$.

Simple Example

Imagine you have a single die, faces 1 to 6. If the die is **fair**, then we expect an equal probability of getting any particular face to be up after rolling. Thus,

$$p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = \frac{1}{6}$$

and

$$E = \frac{1}{6} = 1 \cdot \frac{1}{6}$$

is the expected number of each face you expect to get in a single roll. If we rolled this fair die 30 times, then our expected number of each face would be 5:

$$E = \frac{30}{6} = 5 = 30 \cdot \frac{1}{6}.$$

Back to the Test

Obviously some variation from expected numbers is normal, so our test is really asking the question: are the differences between the actually observed values O and the theoretically expected values E statistically significant?

If O and E are similar for all the categories, then $O - E$ will be quite small, and the overall sum will be reasonably small, resulting in a small χ^2_{test} value, and a large p-value. If they are not similar, the opposite will be true.

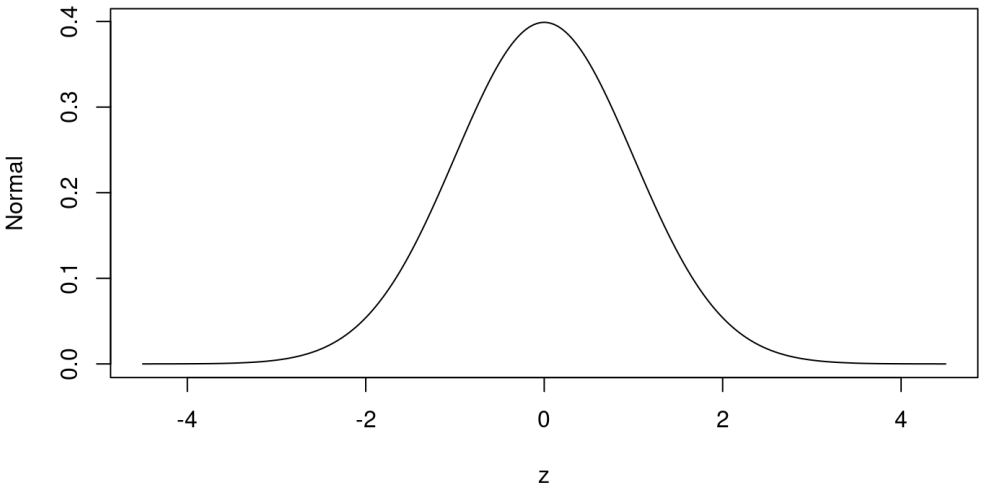
Statement of the Hypothesis

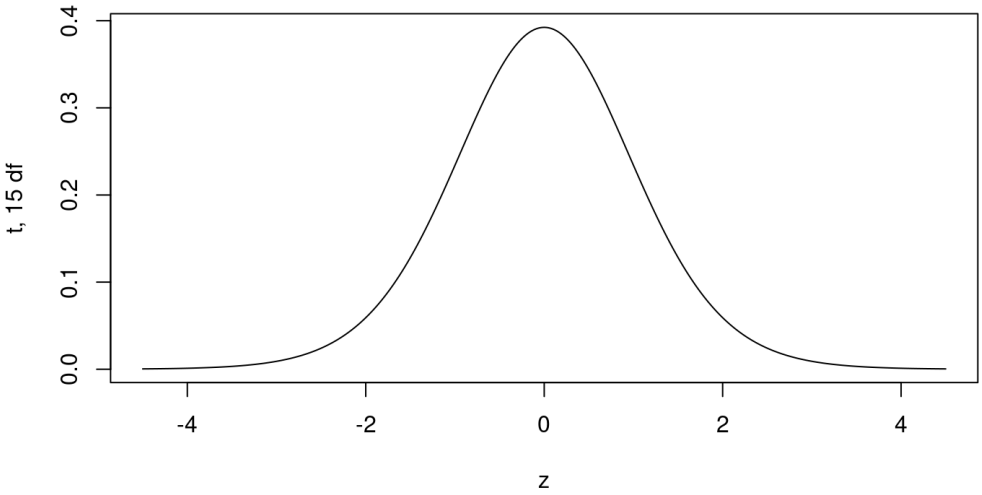
Our null hypothesis is normally “nothing to see here, move along”. In this particular case, the null hypothesis is that the observed data come from the distribution specified.

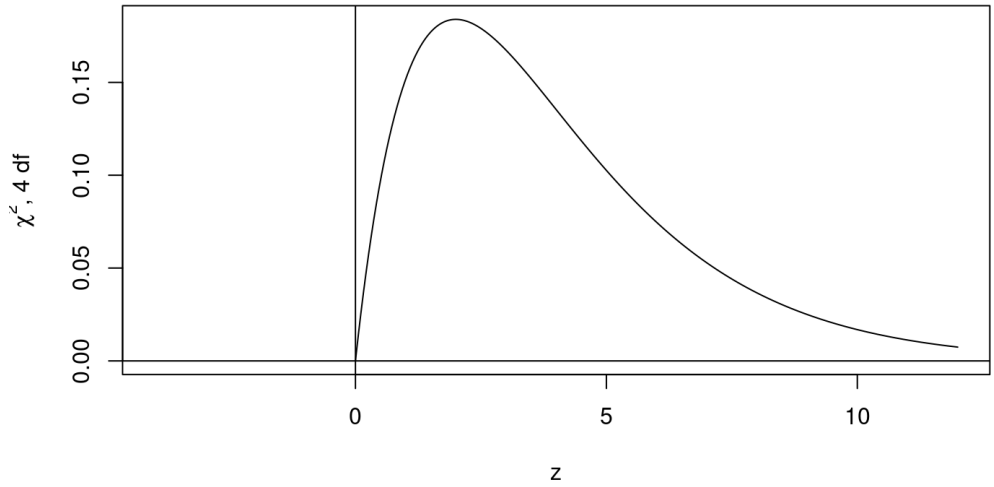
$$H_0 : \text{distribution is as claimed versus}$$
$$H_A : \text{distribution is not as claimed.}$$

χ^2: The Chi-Squared Distribution

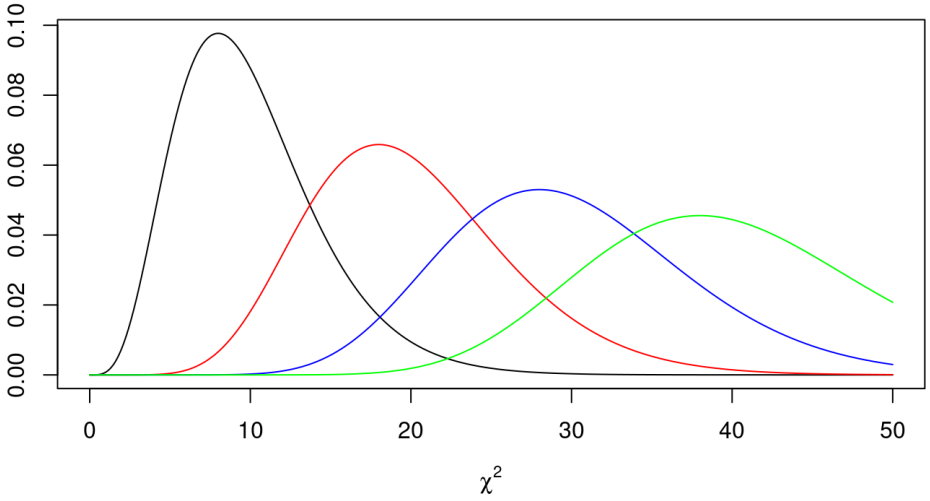
The χ^2 distribution is our second example of a **skewed** distribution. We have previously seen normal and t distributions, each of which is symmetric about 0, and the F distribution, which is also right-skewed.


Normal
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
z


t, 15 df
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
z


χ^2, 4 df
0.00
0.05
0.10
0.15
0
5
10
z

Some examples of the χ^2 distribution for increasing degrees-of-freedom. The higher the degrees-of-freedom, the closer this distribution gets to a normal density.



Example 1: People Lie About Their Weight

The book contains several data sets with health information for males and females. When obtaining weights of subjects, it is critical that the subjects actually be weighed, rather than asking for a self-reported number.

What would you answer if someone asked you for your weight as part of a health survey?

Real Data versus Rounded/Reported

People tend to round. Or lie. Or “shave” off pounds.

If we actually **measure** people’s weight in pounds (accurate to 0.1 pounds), the numbers should be fairly random. In particular, the final digit of the weight should be random, 0 through 9.

From the 80 weights listed in the appendix of the book, we wish to determine if these weights were actually measured, or simply reported (and thus, more likely to be rounded).

The Frequencies

Last Digit	0	1	2	3	4	5	6	7	8	9
Frequency	7	14	6	10	8	4	5	6	12	8

Test the claim that the sample is from a population of weights in which the last digits do **not** occur with the same frequency.

The Hypothesis

$$H_0 : \text{the last digits occur with the same frequency versus}$$
$$H_A : \text{the last digits do not occur with the same frequency}$$

We can also write this in the more complicated form as:

$$H_0 : p_0 = p_1 = p_2 = \cdots = p_8 = p_9 \text{ versus}$$
$$H_A : \text{At least one probability differs from the others.}$$

Performing the test

No significance level is specified, so let $\alpha = 0.05$. The observed frequencies are listed in the table, and the expected frequencies are equal, so

$$E = \frac{80}{10} = 8.$$

Thus, the test statistic is

$$\chi^2_{\text{test}} = \sum \frac{(O - E)^2}{E}.$$

Compute the statistic

Put it all together manually ...

$$\begin{aligned}\chi^2_{\text{test}} &= \frac{(7-8)^2}{8} + \frac{(14-8)^2}{8} + \frac{(6-8)^2}{8} + \frac{(10-8)^2}{8} + \frac{(8-8)^2}{8} \\ &+ \frac{(4-8)^2}{8} + \frac{(5-8)^2}{8} + \frac{(6-8)^2}{8} + \frac{(12-8)^2}{8} + \frac{(8-8)^2}{8} \\ &= \frac{1}{8}[1+36+4+4+0+16+9+4+16+0] \\ &= \frac{90}{8} = 11.25.\end{aligned}$$

(this test statistic has $k-1 = 10-1 = 9$ degrees-of-freedom)

Gross math to do manually!

p-value

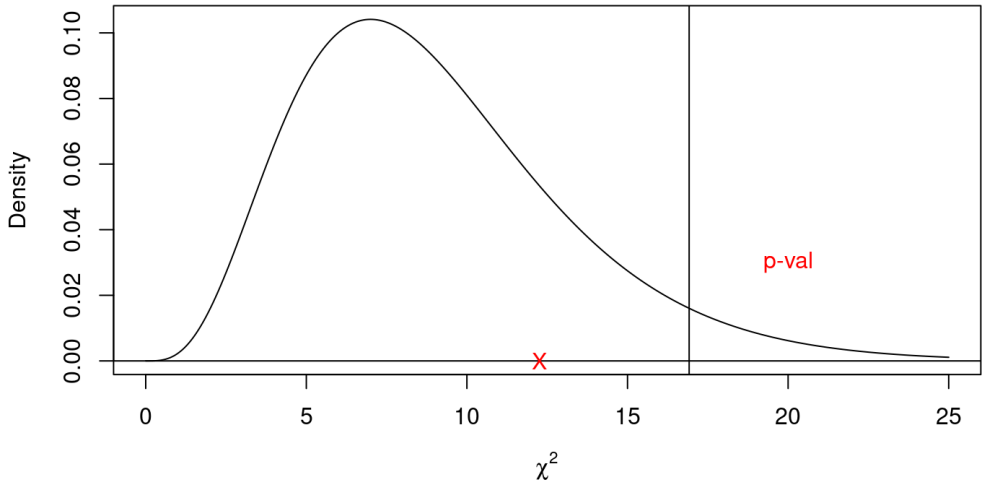
How do we compute the p-value? Well, above we said always go right … and this is the χ^2 distribution. Any guesses as to what function we use to find **p**-values for chisquare?

p-value

```
pchisq(q = 11.25, df = 10 - 1, lower.tail = FALSE)
```

```
## [1] 0.2589613
```

On a graph



Conclusion

Since our test statistic has p-value is greater than α ($0.26 > 0.05$), we **do not have evidence** at level $\alpha = 0.05$ to conclude that there is a difference in the probability of obtaining each terminal digit in the weights.

In other words, there is not sufficient evidence to support the claim that the last digits do not occur with the same relative frequency.

Example 2: Pennies from Cheques (11-2:7)

When considering effects from eliminating the penny, the author of a popular statistics textbook randomly selected 100 cheques and recorded the cents portion of those cheques. The table below lists those cents portions, categorized as listed.

Cents portion of cheque	0-24	25-49	50-74	75-99
Number	61	17	10	12

Using significance level $\alpha = 0.05$, test the claim that the four categories are equally likely. What conclusion do you reach, and what explanation might there be?

Compute Test Statistic

We have four categories, $k = 4$, so we will have $4 - 1 = 3$ degrees-of-freedom. Our null hypothesis is that the four categories are equally likely. Thus,

$$E = \frac{100}{4} = 25.$$

$$\begin{aligned}
\chi^2_{\text{test}} &= \sum \frac{(O - E)^2}{E} \\
&= \frac{(61 - 25)^2}{25} + \frac{(17 - 25)^2}{25} + \frac{(10 - 25)^2}{25} + \frac{(12 - 25)^2}{25} \\
&= \frac{1}{25}\left[36^2 + 8^2 + 15^2 + 13^2\right] \\
&= 70.16.
\end{aligned}$$

What if we … didn't do that

```
dat <- c(61, 17, 10, 12)
probs <- rep(25, 4) / 100
chisq.test(x = dat, p = probs, correct = FALSE)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  dat
## X-squared = 70.16, df = 3, p-value = 3.945e-15
```

```
pchisq(q = 70.16, df = 3, lower.tail = FALSE)
```

```
## [1] 3.944548e-15
```

Conclusion

We do have evidence at level $\alpha = 0.05$ to conclude that the different "cents portions" of the cheques are different in frequency; from simple visual examination, it appears clear that the first category, 0–24, has significantly more cases than the others.

This may be due to the fact that many cheques are written for round numbers of **dollars**, inflating the case of $\$.00$ on the cheques.

Outliers

How do outliers affect goodness-of-fit tests? Two ways:

- First, if the outlier isn't accounted for in the "total" (that is, you say "there are n cases", but the numbers don't add up to n), the test is completely meaningless.
- Second, if the outlier is accounted for, then what do we expect to have happen?

Just like Example 2! One significant divergence from the average will produce a really, really "significant" (extreme) χ^2_{test} value.

Contingency Tables

We now move to a slightly more complicated version of our goodness-of-fit tests. Instead of having a table with a single row of observations, we now have at least two rows and at least two columns of observations. We will show how to examine two hypotheses on data like this:

- Tests of Independence
- Tests of Homogeneity (same-ness)

Note: heterogeneity is the state of being different or diverse kinds; homogeneity is the state of being the same kind.

Definition

A **contingency table** (also known as a two-way frequency table) is a table in which frequencies correspond to two variables (one variable used to categorize rows, the other columns).

Testing Independence

A **test of independence** tests the null hypothesis that in a contingency table, the row and column variables are independent.

Again, O is the observed frequency for a given cell, and E is the expected frequency in a given cell, under the null assumption that the row and column variables are independent. The test statistic is the same as before:

$$\chi^2 = \sum \frac{(O-E)^2}{E}.$$

Difference to Goodness-of-Fit

The expected value is somewhat more complicated than in the goodness-of-fit case:

$$E = \frac{(\text{row total})(\text{column total})}{(\text{grand total})}.$$

Degrees-of-FreedomA

The degrees-of-freedom used in a contingency table test for independence of row and column are $(r-1)\cdot(c-1)$ where r is the number of rows and c is the number of columns. The tests are always one-tailed (upper).

Example 3: Echinacea

The following table contains frequencies of patients involved in a study on the efficacy of echinacea. Each patient was given either a placebo, a 20% extract of echinacea, or a 60% extract of echinacea, and observed to determine if they became infected with the seasonal rhinovirus. What does the result indicate about the effectiveness of achinacea as a treatment for colds?

	Placebo	Echinacea: 20%	Echinacea: 60%
Infected	88	48	42
Not Infected	15	4	10

Hypotheses

This is actually a test for dependence, so the hypotheses are:

$$H_0 : \text{getting an infection is independent of the treatment versus}$$
$$H_A : \text{getting an infection and the treatment are dependent}$$

Finding E

To find the expected value of each of the cells, we must sum across each row, down each column, and find the total number of elements in all of the cells together (the “grand total”).

Semi-Completed Table

	Placebo	Echinacea: 20%	Echinacea: 60%	Row Total
Infected	88	48	42	
Not Infected	15	4	10	
Column Total				

And then … E Table

	Placebo	Echinacea: 20%	Echinacea: 60%	Row Total
Infected				178
Not Infected				29
Column Total	103	52	52	207

Which becomes …

	Placebo	Echinacea: 20%	Echinacea: 60%	Row Total
Infected	88.57	44.71	44.71	178
Not Infected	14.43	7.29	7.29	29
Column Total	103	52	52	207

Note: these rounded numbers add to 207. They won't always be exact, based on the rounding: you might occasionally be off by 0.1 or 0.2. That's ok.

Compute χ^2_{test}

Now, compute:

$$\begin{aligned}\chi^2_{\text{test}} &= \frac{(88-88.57)^2}{88.57} + \frac{(44.71-48)^2}{44.71} + \frac{(42-44.71)^2}{44.71} \\ &\quad + \frac{(15-14.43)^2}{14.43} + \frac{(4-7.29)^2}{7.29} + \frac{(10-7.29)^2}{7.29} \\ &= 2.925.\end{aligned}$$

We could use p-value ...

```
pchisq(q = 2.925, df = 2, lower.tail = FALSE)
```

```
## [1] 0.2316564
```

Therefore, with $p > 0.05$, we **do not have evidence** to conclude that getting an infection and the treatment are dependent.

Alternatively, we fail to reject the null hypothesis of independence between getting an infection and treatment.

Test of Homogeneity

The first test was a test on the independence of row and column variables, where the sample data are from one population. However, sometimes we use samples drawn from **different** populations, and we want to determine whether those populations have the same proportions of characteristics.

This case uses the **test for homogeneity** (having the same quality).

Definition

In a **test of/for homogeneity**, we test the claims that *different populations* have the same proportions of some characteristics.

Example 4: Acceptable versus Correct

When asked about an issue that is considered sensitive, people sometimes change their answers based on the gender or race or the interviewer. Correspondingly, people are sometimes more open with interviewers who are "like them", assuming (sometimes correctly!) that someone who looks like them also thinks like them, and not filtering themselves.

800 men were interviewed by male interviewers and 400 men were interviewed by female interviewers, and asked to agree/disagree with the following statement: "Abortion is a private matter that should be left to a woman to decide, without government intervention."

Results

	Man	Woman
Men who agree	560	308
Men who disagree	240	92

This table lists the gender of the interviewer, and the number of agree/disagree responses.

Computing the Test Statistic

We compute the test statistic in the same way as the test for independence.

	Man	Woman	Total
Men who agree	560	308	868
Men who disagree	240	92	332
Total	800	400	1200

Compute Expected

	Man	Woman	Total
Men who agree	578.67	289.33	868
Men who disagree	221.33	110.67	332
Total	800	400	1200

Put it together

$$\begin{aligned}\chi^2_{\text{test}} &= \frac{(560-578.67)^2}{578.67} + \frac{(308-289.33)^2}{289.33} + \frac{(240-221.33)^2}{221.33} \\ &\quad + \frac{(92-110.67)^2}{110.67} \\ &= 6.529\end{aligned}$$

p-value

```
pchisq(q = 6.529, df = 1, lower.tail = FALSE)
```

```
## [1] 0.01061296
```

Therefore we do have evidence to conclude that the proportions are not the same.

Conclusion

Since we do have evidence to conclude that the proportions are not the same, this means we conclude that, at least for this case, men change their answer based on whether a man or woman is asking them the question.

Example of the previous in R

```r
dat <- data.frame(agree = c(560, 308), disagree = c(240, 92))
chisq.test(dat, correct = FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  dat
## X-squared = 6.5293, df = 1, p-value = 0.01061
```