Comparing multiple proportions

February 24, 2017

psych10.stanford.edu

Announcements / Action Items

- Practice and assessment problem sets will be posted today, might be after 5 PM

- Reminder of OH switch today

Last time

- Some ways to integrate previous material
- We can make better predictions about individual observations when considering the variable *group*, and we can quantify the improvement in these predictions using r^2
- We can remove variance related to *individual differences* by using paired designs, but we must adjust our analysis approach to account for the fact that observations in each pair are not independent (single mean or single proportion)

This time

Our final cycle with **frequency (count) data** and **proportions**

Categorical variable*:* responses or are categories or groups ("levels")

Goal: handle response variables with ***any*** number of categories and grouping variables with ***any*** number of groups ***with a single statistic***
(why not compare each possible pair? we'll discuss on Monday)

	one variable (a response variable)	**two variables (one grouping, one response)**
binary variable(s)	z-test for a single proportion (1 x 2 table)	z-test for a difference in proportions (2 x 2 table)
any categorical variable(s)	chi-square test for goodness-of-fit (1 x any # table)	chi-square test for independence (any # by any # table)

This time

- How can we test whether observed proportions are consistent with expected proportions?

- How can we use proportions to test whether two variables are *associated*?

This time

- **How can we test whether observed proportions are consistent with expected proportions?**

- How can we use proportions to test whether two variables are *associated*?

Previously

converted non-binary variables to binary variables

rock-paper-scissors

rock	paper	scissors
1/3	1/3	1/3

→

rock-paper-scissors

not scissors	scissors
2/3	1/3

roll reported

1	2	3	4	5	6
1/6	1/6	1/6	1/6	1/6	1/6

→

roll reported

not 6	6
5/6	1/6

section called upon

1st	2nd	3rd	4th
1/4	1/4	1/4	1/4

→

section called upon

1st	not 1st
1/4	3/4

Today

assess fit of all categories to hypothesized distribution

rock-paper-scissors

rock	paper	scissors
1/3	1/3	1/3

→

rock-paper-scissors

not scissors	scissors
2/3	1/3

roll reported

1	2	3	4	5	6
1/6	1/6	1/6	1/6	1/6	1/6

→

roll reported

not 6	6
5/6	1/6

section called upon

1st	2nd	3rd	4th
1/4	1/4	1/4	1/4

→

section called upon

1st	not 1st
1/4	3/4

Chi square: goodness-of-fit

Goal: use observed ***sample*** proportions to test hypotheses about unobserved ***population*** proportions — looking at a **single** variable

General approach:

Generate a null hypothesis (H_0) about population proportions

Compute frequencies that we would expect if H_0 was true
✳ usually the hardest step

Summarize the discrepancy between these with a single statistic (X^2)

Use the distribution of all of the statistics that we *could have observed* if the null hypothesis was true to determine whether this statistic would be *unlikely* if the null hypothesis was true

Academic dishonesty

Are cases of academic dishonesty evenly distributed across departments?

The Mercury News

News

Stanford finds cheating — especially among computer science students — on the rise

By LISA M. KRIEGER | lkrieger@bayareanewsgroup.com |
PUBLISHED: February 6, 2010 at 1:22 pm | UPDATED: August 13, 2016 at 10:33 pm

The Stanford Daily

Stanford CS department battles honor code violations

June 3, 2014 5 Comments Tweet Like 46

Academic dishonesty

A simplified scenario. A university has four, equally sized departments: computer science, biology, sociology, and art history. They handle 100 honor code cases, as below.

observed frequencies, O_i

CS	Bio	Soc	Art	Total
35	30	20	15	100

A null hypothesis: students in each department are equally likely to be included in these cases

expected proportions

CS	Bio	Soc	Art	Total
0.25	0.25	0.25	0.25	1

Academic dishonesty

A simplified scenario. A university has four, equally sized departments: computer science, biology, sociology, and art history. They handle 100 honor code cases, as below.

observed frequencies, O_i

CS	Bio	Soc	Art	Total
35	30	20	15	100

A null hypothesis: students in each department are equally likely to be included in these cases

expected frequencies, E_i (multiply expected proportion by n)

CS	Bio	Soc	Art	Total
0.25 * 100 = 25	0.25 * 100 = 25	0.25 * 100 = 25	0.25 * 100 = 25	1 * n = 100

Academic dishonesty

Do we expect the observed frequencies to perfectly match the expected frequencies if the null hypothesis is true?

How likely is it to find a discrepancy *this extreme or more* if our null hypothesis is true?

→ summarize discrepancy with a single statistic, X^2

observed frequencies, O_i

CS	Bio	Soc	Art	Total
35	30	20	15	100

expected frequencies, E_i

CS	Bio	Soc	Art	Total
0.25 * 100 = 25	0.25 * 100 = 25	0.25 * 100 = 25	0.25 * 100 = 25	1 * n = 100

Academic dishonesty

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

numerator: discrepancy, squared — why do we square it?

denominator: divided by expected frequency *(normalized)* — less surprised to be off by 4 if you expect 1000 than if you expect 8

sum terms from each *level* (each *category*)

observed frequencies, O_i

CS	Bio	Soc	Art	Total
35	30	20	15	100

expected frequencies, E_i

CS	Bio	Soc	Art	Total
0.25 * 100 = 25	0.25 * 100 = 25	0.25 * 100 = 25	0.25 * 100 = 25	1 * n = 100

Academic dishonesty

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

$$\chi^2 = \frac{(35-25)^2}{25} + \frac{(30-25)^2}{25} + \frac{(20-25)^2}{25} + \frac{(15-25)^2}{25}$$

$$\chi^2 = \frac{100}{25} + \frac{25}{25} + \frac{25}{25} + \frac{100}{25} = 4 + 1 + 1 + 4 = 10$$

observed frequencies, O_i

CS	Bio	Soc	Art	Total
35	30	20	15	100

expected frequencies, E_i

CS	Bio	Soc	Art	Total
0.25 * 100 = 25	0.25 * 100 = 25	0.25 * 100 = 25	0.25 * 100 = 25	1 * n = 100

Academic dishonesty

A (slightly less) simplified scenario. A university has four departments: computer science, biology, sociology, and art history, **which include 30%, 30%, 20% and 20% of the students**, respectively. They handle 100 honor code cases, as below.

observed frequencies, O_i

CS	Bio	Soc	Art	Total
35	30	20	15	100

A null hypothesis: students in each department are equally likely to be included in these cases

expected frequencies, E_i (multiply expected proportion by n)

CS	Bio	Soc	Art	Total
0.30 * 100 = **30**	0.30 * 100 = **30**	0.20 * 100 = **20**	0.20 * 100 = **20**	1 * n = 100

Academic dishonesty

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

$$\chi^2 = \frac{(35-30)^2}{30} + \frac{(30-30)^2}{30} + \frac{(20-20)^2}{20} + \frac{(15-20)^2}{20}$$

$$\chi^2 = \frac{25}{30} + \frac{0}{30} + \frac{0}{20} + \frac{25}{20} = .83 + 0 + 0 + 1.25 = 2.08$$

observed frequencies, O_i

CS	Bio	Soc	Art	Total
35	30	20	15	100

expected frequencies, E_i

CS	Bio	Soc	Art	Total
0.30 * 100 = **30**	0.30 * 100 = **30**	0.20 * 100 = **20**	0.20 * 100 = **20**	1 * n = 100

We must take the base rate into account when we specify what we expect to see

As a consumer of statistics, beware of raw counts!

The X^2 distribution(s)

Is $X^2 = 10$ unlikely?
Is $X^2 = 2.08$ unlikely?

When some assumptions are met, the potential X^2 statistics we *could* observe if the null hypothesis is true are described by a X^2 distribution (→ a **sampling distribution** of X^2 statistics)

Like the *family* of t-distributions, we have a *family* of X^2 distributions, specified by degrees of freedom *(df)*

Here, df specifies how many *cell counts* are free to vary

For a goodness-of-fit test, df = # of categories - 1

observed frequencies, O_i

CS	Bio	Soc	Art	Total
35	30	20	15	100

The X^2 distribution(s)

Is X^2 = 10 unlikely?
Is X^2 = 2.08 unlikely?

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Think about the distribution of X^2 values that we *could observe …*

What shape do you expect? (Hint: can X^2 be negative?)

How does df affect central tendency?

How does df affect variability?

How does df affect shape?

Are we interested in unusually low X^2, unusually high X^2, or both?

The X^2 distribution(s)

Is $X^2 = 10$ unlikely?
Is $X^2 = 2.08$ unlikely?

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

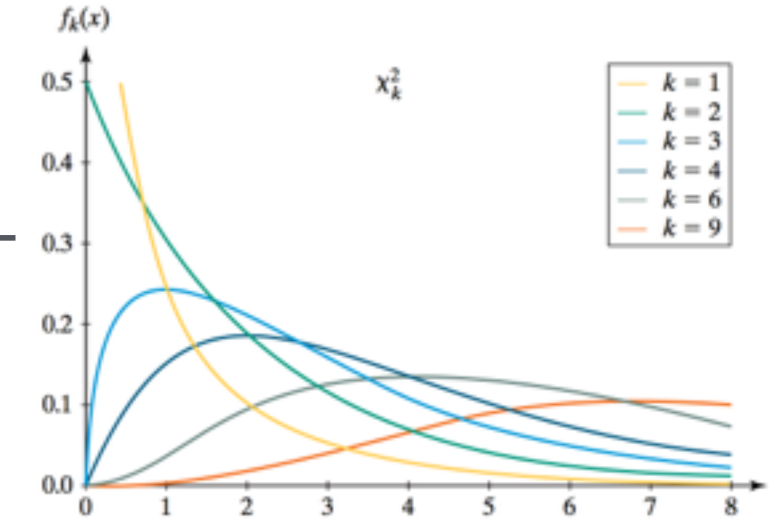


FIGURE 8.10 Chi-square distributions with various degrees of freedom (k).

Think about the distribution of X^2 values that we *could observe …*

What shape do you expect? (Hint: can X^2 be negative?)
positively skewed

How does df affect central tendency?
as df increases, central tendency increases

How does df affect variability?
as df increases, variability increases

How does df affect shape?
as df increases, the shape becomes less skewed

Are we interested in unusually low X^2, unusually high X^2, or both?
only unusually high values, unusually low values are consistent with H_0 (imagine $X^2 = 0$, what would this mean?) → only interested in the upper tail, which includes deviations in any direction

The X^2 distribution(s)

Is $X^2 = 10$ unlikely?
Is $X^2 = 2.08$ unlikely?

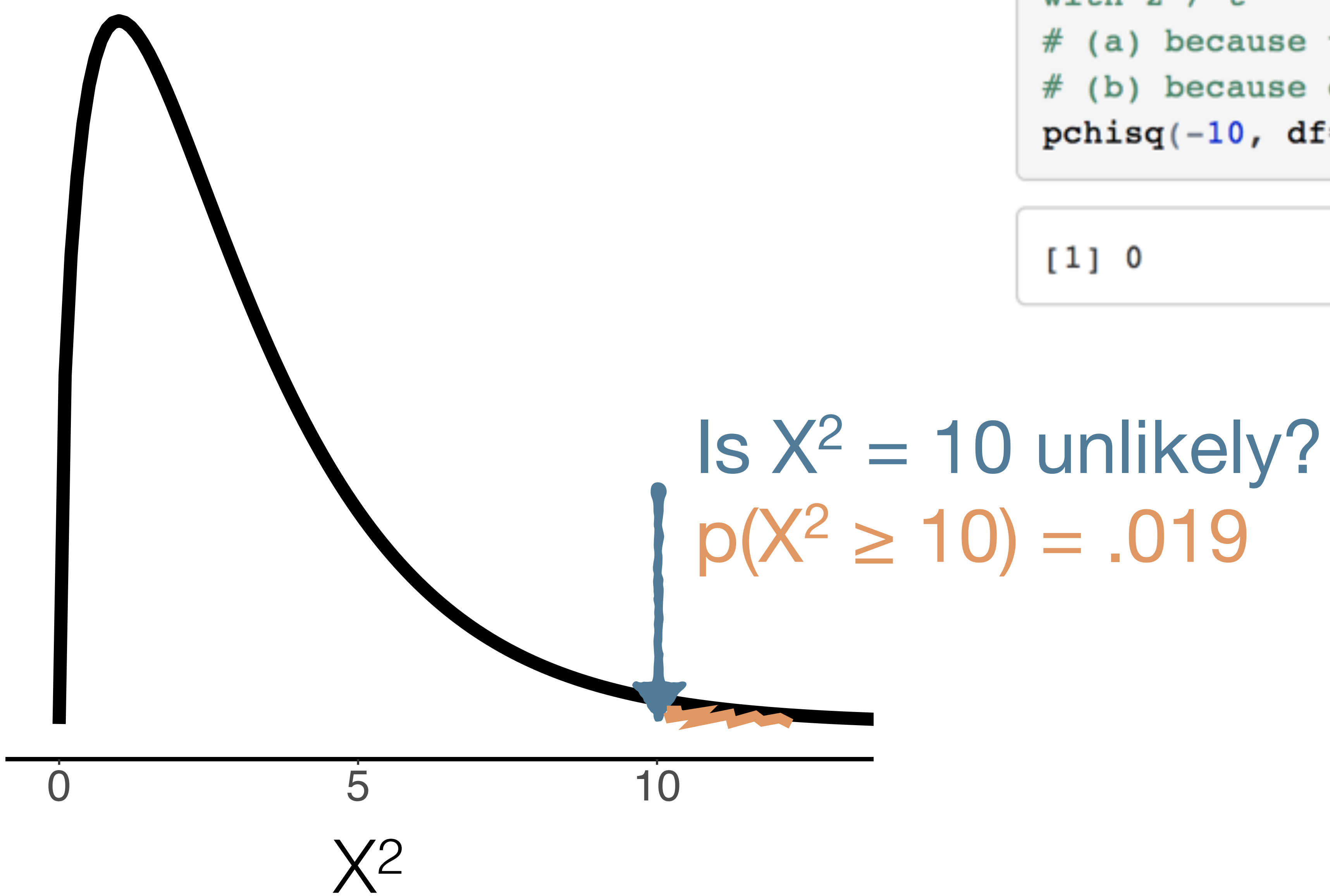
```
pchisq(10, df=3, lower.tail=FALSE)
```

```
[1] 0.01856614
```

Hide

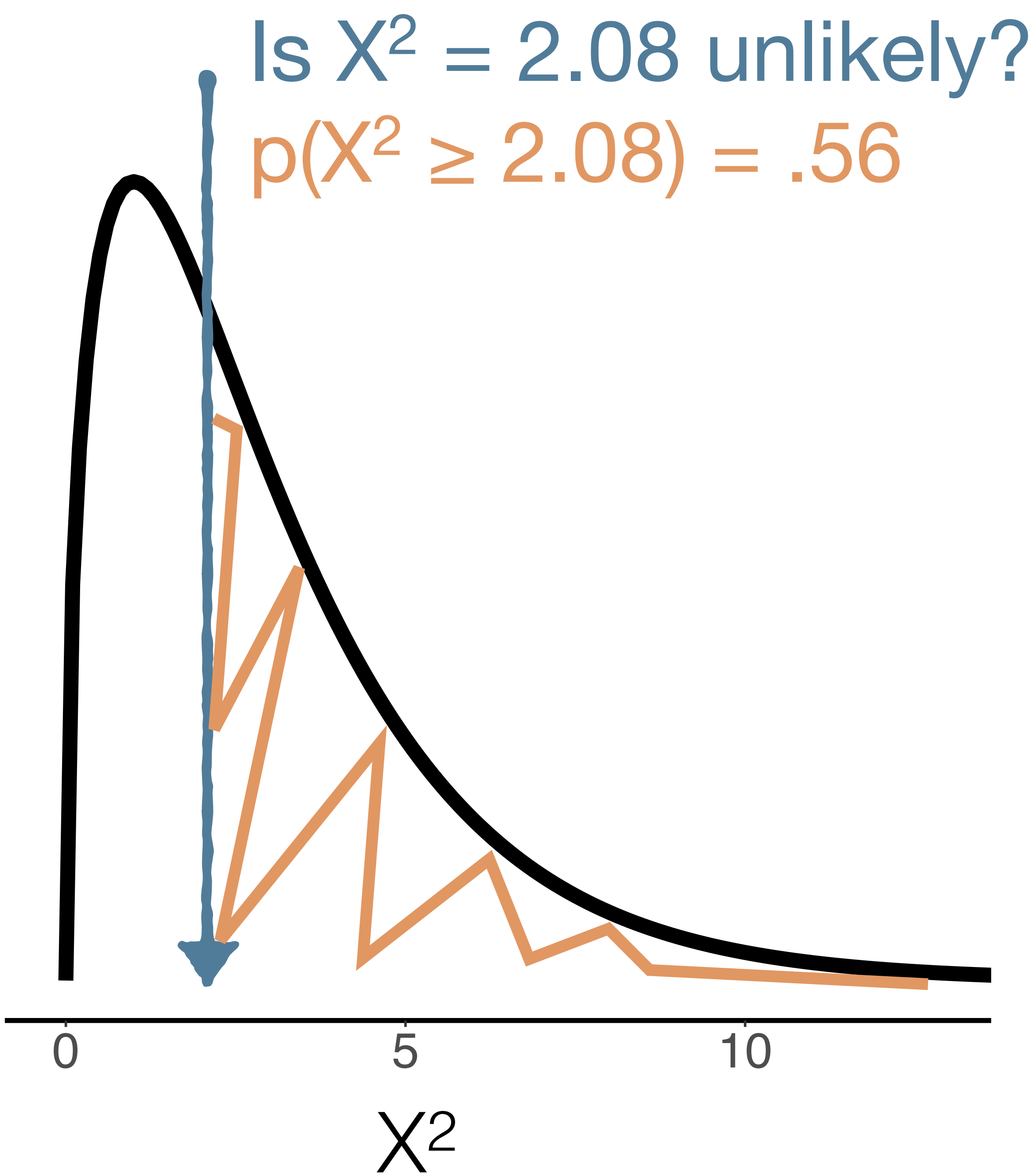
```
# it doesn't work to take our previous shortcut with z / t
# (a) because the distribution is not symmetric
# (b) because chisquare cannot be negative
pchisq(-10, df=3)
```

```
[1] 0
```



The X^2 distribution(s)

Is $X^2 = 10$ unlikely?
Is $X^2 = 2.08$ unlikely?




```
pchisq(2.08, df=3, lower.tail=FALSE)
```

```
[1] 0.5559695
```

Hide

```
# it doesn't work to take our previous shortcut with z / t
# (a) because the distribution is not symmetric
# (b) because chisquare cannot be negative
pchisq(-2.08, df=3)
```

```
[1] 0
```

Hypothesis test in *simplified* scenario

Two hypotheses:

- **H_0**: students in each department are equally likely to be included in honor code cases
  → the proportions of students involved across departments are .(25, .25, .25, .25) for CS, biology, sociology, and art history
- **H_A**: the students in each department are not equally likely to be included in honor code cases
  → it is not the case that the proportions of students involved across departments are (.25, .25, .25, .25) for CS, biology, sociology, and art history
  (this is not the same as saying that *none of these* proportions is .25)

Hypothesis test in *simplified* scenario

The probability of observing frequencies as or more extreme as our sample frequencies *if the null hypothesis was true* is .019

- if $\alpha = .05$, reject the null hypothesis and infer that the true proportions are not (.25, .25, .25, .25)

We still do not know …

- exactly *which* proportions are significantly different from what we expect
  - can perform follow-up tests
- *why* the proportions are not what we expect
  - this is *observational*
  - this is a measure of *people who are "caught" and reported* — why might this differ across departments?

Benford's law

Forbes

MAY 30, 2014 @ 07:47 AM 14,434 VIEWS

The Simple Mathematical Law That Financial Fraudsters Can't Beat

Daniel Fisher, FORBES STAFF

I cover finance, the law, and how the two interact. FULL BIO

https://www.forbes.com/sites/danielfisher/2014/05/30/the-simple-mathematical-law-that-financial-fraudsters-cant-beat/#56886f5b4612

How hard is it to ferret out securities fraud? It might be as easy as looking for how many times the digit `1' appears in a company's financial entries instead of '9.'

Benford's Law has been used in a large number of forensic applications, including voter fraud, Greece's effort to hide its debt, and determining whether digital photographs have been altered. It's also been in the toolkit of auditors for years, said Amiram, a former auditor, but only at the level of operating accounts. He said his paper is the first to apply the law to company-level financial reports accessible through databases like Compustat.

Benford's law

Class survey collected the *first digit* of your previous address

Does the distribution of digits follow Benford's law? (Does it deviate from the law *beyond* what we would expect by random fluctuations?).

H_0: the distribution of first digits follows Benford's law

H_A: the distribution of first digits does not follow Benford's law

What could it mean if we rejected the null hypothesis?

inaccuracy in reporting?

something non-random about Stanford student addresses?

Type I Error?

something else?

a small cautionary note about this example, it is best to have at least 10 observations per cell

Benford's law

Benford's law states that *first digits* follow a distribution of:

	1	2	3	4	5	6	7	8	9	total
expected probabilities	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046	1

Our class data of *first digits* of previous addresses:

	1	2	3	4	5	6	7	8	9	total
O_i	51	16	11	19	15	7	9	6	7	**n = 141**

Benford's law

Benford's law states that *first digits* follow a distribution of:

	1	2	3	4	5	6	7	8	9	total
expected probabilities	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046	1
multiply by n → E_i	42.44	24.82	17.62	13.68	11.14	9.45	8.18	7.19	6.49	141

Our class data of *first digits* of previous addresses:

	1	2	3	4	5	6	7	8	9	total
O_i	51	16	11	19	15	7	9	6	7	**n = 141**

Benford's law

Benford's law states that *first digits* follow a distribution of:

	1	2	3	4	5	6	7	8	9	total
expected probabilities	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046	1
multiply by n → E_i	42.44	24.82	17.62	13.68	11.14	9.45	8.18	7.19	6.49	141

Our class data of *first digits* of previous addresses:

	1	2	3	4	5	6	7	8	9	total
O_i	51	16	11	19	15	7	9	6	7	**n = 141**
$(O_i - E_i)^2 / E_i$	1.73	3.13	2.49	2.07	1.34	0.63	0.08	0.20	0.04	**X^2 = 11.71**

pchisq(11.71, df = 8, lower.tail = FALSE) → p = .16

Benford's law

Class survey collected the *first digit* of your previous address

Does the distribution of digits follow Benford's law? (Does it deviate from the law *beyond* what we would expect by random fluctuations?).

H_0: the distribution of first digits follows Benford's law

H_A: the distribution of first digits does not follow Benford's law

Fail to reject the null hypothesis, conclude that we do **not** have evidence that the distribution of first digits is inconsistent with Benford's law

The role of sample size

law of large numbers → ask, *where is n in this calculation*?

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

$$= \sum \frac{(n\hat{p}_i - n\pi_i)^2}{n\pi_i}$$

$$= \sum \frac{n^2(\hat{p}_i - \pi_i)^2}{n\pi_i}$$

$$= \sum \frac{n(\hat{p}_i - \pi_i)^2}{\pi_i}$$

$$= \boxed{n} \boxed{\sum \frac{(\hat{p}_i - \pi_i)^2}{\pi_i}}$$

= sample size * effect size

Interim summary

We've extended the ideas behind asking whether an observed distribution of a *binary* variable is consistent with a hypothesized distribution to asking whether an observed distribution of *any categorical* variable is consistent with a hypothesized distribution!

	one variable (a response variable)	two variables (one grouping, one response)
binary variable(s)	z-test for a single proportion (1 x 2 table)	z-test for a difference in proportions (2 x 2 table)
any categorical variable(s)	**chi-square test for goodness-of-fit (1 x any # table)**	chi-square test for independence (any # by any # table)

This time

- How can we test whether observed proportions are consistent with expected proportions?

- **How can we use proportions to test whether two variables are *associated*?**

Previously

analyzed independence of 2 x 2 table, comparing difference in observed proportions to difference we would expect by random chance

	group 1	group 2
response A		
response B		

Today

analyzed independence of any size table, comparing difference in observed proportions to difference we would expect by random chance

	group 1	group 2	group 3	…
response A				
response B				
response C				
…				

Chi-square: independence

Goal: use observed ***sample*** proportions to test hypotheses about unobserved ***population*** proportions — specifically, are **two** variables *associated* or are they *independent?*

General approach:

Null hypothesis (H_0) is that our variables are *independent* (no *association)* between them → we expect population proportions to follow certain rules due to this *independence*

Compute frequencies that we would **expect** if H_0 was true
✳ usually the hardest step

Summarize the **discrepancy** between these with a single statistic (X^2)

Use the distribution of all of the statistics that we *could have observed* if the null hypothesis was true to determine whether this statistic would be *unlikely* if the null hypothesis was true

Hand washing and city

hand washing study:

a 2005 study observed hand-washing behavior in public restrooms in four major cities

http://www.cleaninginstitute.org/assets/1/AssetManager/2005_Hand_Washing_Findings_rev.pdf

Observed	**Atlanta**	**Chicago**	**NY**	**SF**	**Total**
Washed	1175	1329	1169	1521	**5194**
Did not wash	413	180	334	215	**1142**
Total	**1588**	**1509**	**1503**	**1736**	**6336**

Marginal probabilities

ignoring city

- probability of *washing,* 5194 / 6336 = .82
- probability of *not washing,* 1142 / 6336 = .18
- *caution of self-report, in a parallel telephone survey 91% of people *reported* washing

ignoring *washing*

- probability of *Atlanta*, 1588 / 6336 = .25
- probability of *Chicago*, 1509 / 6336 = .24
- probability of *New York*, 1503 / 6336 = .24
- probability of *San Francisco*, 1736 / 6336 = .27

Observed	**Atlanta**	**Chicago**	**NY**	**SF**	**Total**
Washed	1175	1329	1169	1521	**5194**
Did not wash	413	180	334	215	**1142**
Total	**1588**	**1509**	**1503**	**1736**	**6336**

Conditional probabilities

p(washing | Atlanta) = 1175 / 1588 = .74

p(washing | Chicago) = 1329 / 1509 = .88

p(washing | NY) = 1169 / 1503 = .78

p(washing | SF) = 1521 / 1736 = .88

comparing distributions
all cities have a mode of 'washing'
Chicago and SF have lower variability (higher relative freq. at the mode)

Observed	**Atlanta**	**Chicago**	**NY**	**SF**	**Total**
Washed	1175	1329	1169	1521	**5194**
Did not wash	413	180	334	215	**1142**
Total	**1588**	**1509**	**1503**	**1736**	**6336**

Independence of city and washing

two hypotheses:

H_0: there is no association between **city** and **washing**

H_A: there is an association between **city** and **washing**

Observed	**Atlanta**	**Chicago**	**NY**	**SF**	**Total**
Washed	1175	1329	1169	1521	**5194**
Did not wash	413	180	334	215	**1142**
Total	**1588**	**1509**	**1503**	**1736**	**6336**

Reminder: independence of A and B

If A and B are independent then p(A) * p(B) = p(A and B)

→ p(A | B) = p(A and B) / p(B)

= p(A) * p(B) / p(B)

= p(A)

→ p(A | not B) = p(A and not B) / p(not B)

= p(A) * p(not B) / p(not B)

= p(A)

Here, “independent” means there is “no relationship between the grouping variable and the response variable”

Independence of city and washing

what frequencies would we expect to see if city and washing were independent? *p(A and B) = p(A) * P(B)*

consider **frequency** of (*Atlanta* and *washed*)
= n * p(*Atlanta* and *washed*)
= n * p(*Atlanta*) * p(*washed*) = n * (# Atlanta / n) * (# washed / n)
= (# Atlanta * # washed) / n
= (5194 * 1588) / 6336 = 1301.78

	Atlanta	Chicago	NY	SF	Total
Washed					**5194**
Did not wash					**1142**
Total	**1588**	**1509**	**1503**	**1736**	**6336**

Independence of city and washing

what frequencies would we expect to see if city and washing were independent? *p(A and B) = p(A) * P(B)*

more generally, **frequency** of (*A* and *B*)
= (# A * # B) / n
= (# row * # column) / n

	Atlanta	Chicago	NY	SF	Total
Washed					**5194**
Did not wash					**1142**
Total	**1588**	**1509**	**1503**	**1736**	**6336**

Independence of city and washing

what frequencies would we expect to see if city and washing were independent? *p(A and B) = p(A) * P(B)*

more generally, **frequency** of (*A* and *B*)
= (# A * # B) / n
= (# row * # column) / n

Expected	**Atlanta**	**Chicago**	**NY**	**SF**	**Total**
Washed	5194 * 1588 / 6366	5194 * 1509 / 6336	5194 * 1503 / 6336	5194 * 1736 / 6336	**5194**
Did not wash	1142 * 1588 / 6336	1142 * 1509 / 6336	1142 * 1503 / 6336	1142 * 1736 / 6336	**1142**
Total	**1588**	**1509**	**1503**	**1736**	**6336**

Independence of city and washing

what frequencies would we expect to see if city and washing were independent? *p(A and B) = p(A) * P(B)*

more generally, **frequency** of (*A* and *B*)
= (# A * # B) / n
= (# row * # column) / n

Expected	**Atlanta**	**Chicago**	**NY**	**SF**	**Total**
Washed	1301.78	1237.02	1232.10	1423.10	**5194**
Did not wash	286.22	271.98	270.90	312.90	**1142**
Total	**1588**	**1509**	**1503**	**1736**	**6336**

Independence of city and washing

$X^2 = \Sigma((O_i - E_i)^2/E_i) = (1175 - 1301.78)^2 / 1301.78 + \ldots + (215 - 312.90)^2 / 312.90$

Observed	**Atlanta**	**Chicago**	**NY**	**SF**	**Total**
Washed	1175	1329	1169	1521	**5194**
Did not wash	413	180	334	215	**1142**
Total	**1588**	**1509**	**1503**	**1736**	**6336**

Expected	**Atlanta**	**Chicago**	**NY**	**SF**	**Total**
Washed	1301.78	1237.02	1232.10	1423.10	**5194**
Did not wash	286.22	271.98	270.90	312.90	**1142**
Total	**1588**	**1509**	**1503**	**1736**	**6336**

Independence of city and washing

$X^2 = \Sigma((O_i-E_i)^2/E_i) = (1175 - 1301.78)^2 / 1301.78 + \ldots + (215-312.90)^2 / 312.90$

$X^2 =$

$\Sigma((O_i-E_i)^2/E_i) =$

$(1175 - 1301.78)^2 / 1301.78 +$

$(413 - 286.22)^2 / 286.22 +$

$(1329 - 1237.02)^2 / 1237.02 +$

$(180 - 271.98)^2 / 271.98 +$

$(1169 - 1232.10)^2 / 1232.10 +$

$(334 - 270.90)^2 / 270.90 +$

$(1521 - 1423.10)^2 / 1423.10 +$

$(215 - 312.90)^2 / 312.90 =$

161.74

Degrees of freedom

how many cells are free to vary?

df = (# rows - 1) * (# columns - 1)
= (2 - 1) * (4 - 1)
= 1 * 3
= 3

```
> pchisq(161.74, df = 3, lower.tail = FALSE)
[1] 7.719954e-35
```

Observed	**Atlanta**	**Chicago**	**NY**	**SF**	**Total**
Washed	1175	1329	1169	*constraint*	**5194**
Did not wash	*constraint*	*constraint*	*constraint*	*constraint*	**1142**
Total	**1588**	**1509**	**1503**	**1736**	**6336**

Independence of city and washing

two hypotheses:

H_0: there is no association between **city** and **washing**

H_A: there is an association between **city** and **washing**

but we don't know which cities have different proportions from each other → inference or estimation with *pairwise* proportion differences

Observed	**Atlanta**	**Chicago**	**NY**	**SF**	**Total**
Washed	1175	1329	1169	1521	**5194**
Did not wash	413	180	334	215	**1142**
Total	**1588**	**1509**	**1503**	**1736**	**6336**

Placebic information

A researcher attempts to cut someone in line for the photocopier to make copies of 5 pages with three strategies (**adapted** from Langer et al., 1978)

(1) no info — ‘may I use the Xerox machine’

(2) real info — ‘may I use the Xerox machine, because I’m in a rush’

(3) ‘placebic’ info — ‘may I use the Xerox machine, because I need to make copies’

is *form of request* associated with *rate of compliance*?

Observed	**none**	**real**	**placebic**	**Total**
Yes	50	90	87	**227**
No	40	10	15	**65**
Total	**90**	**100**	**102**	**292**

Marginal probabilities

ignoring *form of request*

probability of *yes,* 227 / 292 = .78

probability of *no,* 65 / 292 = .22

ignoring *compliance*

probability of *none*, 90 / 292 = .31

probability of *real*, 100 / 292 = .34

probability of *placebic*, 102 / 292 = .35

Observed	**none**	**real**	**placebic**	**Total**
Yes	50	90	87	**227**
No	40	10	15	**65**
Total	**90**	**100**	**102**	**292**

Conditional probabilities

p(yes | none) = 50 / 90 = .56

p(yes | real) = 90 / 100 = .90

p(yes | placebic) = 87 / 102 = .85

comparing distributions
all request types have a mode of ‘yes’
‘no information’ has greatest variability (lowest relative freq. at the mode)

Observed	**none**	**real**	**placebic**	**Total**
Yes	50	90	87	**227**
No	40	10	15	**65**
Total	**90**	**100**	**102**	**292**

Independence of request and compliance

two hypotheses:

H_0: there is no association between **request type** and **compliance**

H_A: there is an association between **request type** and **compliance**

Observed	**none**	**real**	**placebic**	**Total**
Yes	50	90	87	**227**
No	40	10	15	**65**
Total	**90**	**100**	**102**	**292**

Independence of request and compliance

what frequencies would we expect to see if request type and compliance were independent? *$p(A \text{ and } B) = p(A) * P(B)$*

more generally, **frequency** of (*A* and *B*)
= (# A * # B) / n
= (# row * # column) / n

	none	**real**	**placebic**	**Total**
Yes				**227**
No				**65**
Total	**90**	**100**	**102**	**292**

Independence of request and compliance

what frequencies would we expect to see if request type and compliance were independent? *p(A and B) = p(A) * P(B)*

more generally, **frequency** of (*A* and *B*)
= (# A * # B) / n
= (# row * # column) / n

Expected	**none**	**real**	**placebic**	**Total**
Yes	227 * 90 / 292 = **69.97**	227 * 100 / 292 = **77.74**	227 * 102 / 292 = **79.29**	**227**
No	65 * 90 / 292 = **20.03**	65 * 100 / 292 = **22.26**	65 * 102 / 292 = **22.71**	**65**
Total	**90**	**100**	**102**	**292**

Independence of request and compliance

$X^2 = \Sigma((O_i - E_i)^2/E_i) = (50 - 69.97)^2 / 69.97 + \ldots + (15 - 22.71)^2 / 22.71$

Observed	**none**	**real**	**placebic**	**Total**
Yes	50	90	87	**227**
No	40	10	15	**65**
Total	**90**	**100**	**102**	**292**

Expected	**none**	**real**	**placebic**	**Total**
Yes	69.97	77.74	79.29	**227**
No	20.03	22.26	22.71	**65**
Total	**90**	**100**	**102**	**292**

Independence of request and compliance

$X^2 = \Sigma((O_i-E_i)^2/E_i) = (50 - 69.97)^2 / 69.97 + \ldots + (15 - 22.71)^2 / 22.71$

$X^2 =$

$\Sigma((O_i-E_i)^2/E_i) =$

$(50 - 69.97)^2 / 69.97 +$

$(40 - 20.03)^2 / 20.03 +$

$(90 - 77.74)^2 / 77.74 +$

$(10 - 22.26)^2 / 22.26 +$

$(87 - 79.29)^2 / 79.29 +$

$(15 - 22.71)^2 / 22.71 =$

37.66

Independence of request and compliance

how many cells are free to vary?

df = (# rows - 1) * (# columns - 1)
= (2 - 1) * (3 - 1)
= 1 * 2
= 2

```
> pchisq(37.66, df = 2, lower.tail = FALSE)
[1] 6.641022e-09
```

Observed	**none**	**real**	**placebic**	**Total**
Yes	50	90	87	**227**
No	40	10	15	**65**
Total	**90**	**100**	**102**	**292**

Independence of request and compliance

two hypotheses:

H_0: there is no association between **request type** and **compliance**

H_A: there is an association between **request type** and **compliance**

but we don't know (*for* sure) which request types have different proportions from each other → inference or estimation on *pairwise* proportion differences

Observed	**none**	**real**	**placebic**	**Total**
Yes	50	90	87	**227**
No	40	10	15	**65**
Total	**90**	**100**	**102**	**292**

Interim summary

We've extended the ideas behind asking whether there is an association between two ***binary*** variables to asking whether there is an association between ***any categorical*** variables!

	one variable (a response variable)	**two variables (one grouping, one response)**
binary variable(s)	z-test for a single proportion (1 x 2 table)	z-test for a difference in proportions (2 x 2 table)
any categorical variable(s)	chi-square test for goodness-of-fit (1 x any # table)	**chi-square test for independence (any # by any # table)**

If you're curious …

- A X^2 distribution with k degrees of freedom describes the the distribution of k independent observations that come from normal distributions with a mean of 0 and a standard deviation of 1 (a **z-distribution**) that are **squared** and **summed**

- The distribution of $SS = \Sigma(x - \mu)^2$ can be linked to a X^2 distribution (after some maneuvering)

- The t-distribution and F-distribution (coming up) are derived in part by using the X^2 distribution

Recap

- If we have a single categorical variable, we can use a X^2 test for goodness-of-fit to compare the proportions of responses in each category to a hypothesized model

- If we have multiple categorical variables, we can use a X^2 test for independence to ask whether the two variables are associated (not independent) or not associated (independent)

Questions