Chi-Square Tests

Keegan Korthauer
Department of Statistics
UW Madison

General HT for Proportions

- Hypothesis tests for proportions we've studied so far:
  - success probability p when we observe X successes in a collection of n independent Bernoulli trials (6.3)
  - difference in success probability $p_x - p_Y$ when we observe X out of n_X successes in one sample and Y out of n_Y in the other (6.6)

- In these situations we are confined to Bernoulli trials
  - only two possible outcomes for each trial
  - examples: coin flip, voting between two candidates

- What if there are more than two possible outcomes for each trial?

Chi-Square Test Motivation

- We want to **check if a die is fair** (all sides have equal chance of landing face-up)
- Experiment – throwing the die N times (e.g. N=600) and observe the number of times each side (numbered from 1 to 6) comes up:

Category	Observed
1	115
2	97
3	91
4	101
5	110
6	86
Total	600

- Generalization of a Bernoulli trial – **multinomial trial**
- Does this die seem to be fair?

Multinomial Trial

- Generalization of the Bernoulli trial to more than two possible outcomes

- **Bernoulli Trial**: one parameter *p* = success probability

- **Multinomial Trial**: *k* parameters $p_1,\ldots, p_k$ represent the probabilities of each of the *k* possible outcomes
  - Sum of $p_1+\ldots+ p_k = 1$
  - Models discrete random variables with *k* possible categories
  - Example: roll of a die has 6 possible outcomes

Hypothesis Test for the Die Example

- Want to test **H_0: the die is fair** versus **H_1: the die is not fair**

- Think of the die roll as a multinomial trial with 6 possible outcomes, and probabilities $p_1, \ldots, p_6$

- Under the null hypothesis, each side is equally likely to come up, which means:

$$H_0: \; p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = 1/6$$

- We want a test statistic that will measure the deviation what we **expect** under the null hypothesis from what we actually **observed**

Hypothesis Test for the Die Example

- Under the null hypothesis, we expect 1/6 of the total die rolls to show each number

- Out of 600 rolls, we expect 100 of each number:

Category	Observed	Expected
1	115	100
2	97	100
3	91	100
4	101	100
5	110	100
6	86	100
Total	600	600

- Idea of the test statistic: Add up the **squared** deviations of the observed and expected values

↑

without this, some of the positive deviations will cancel out some of the negative deviations

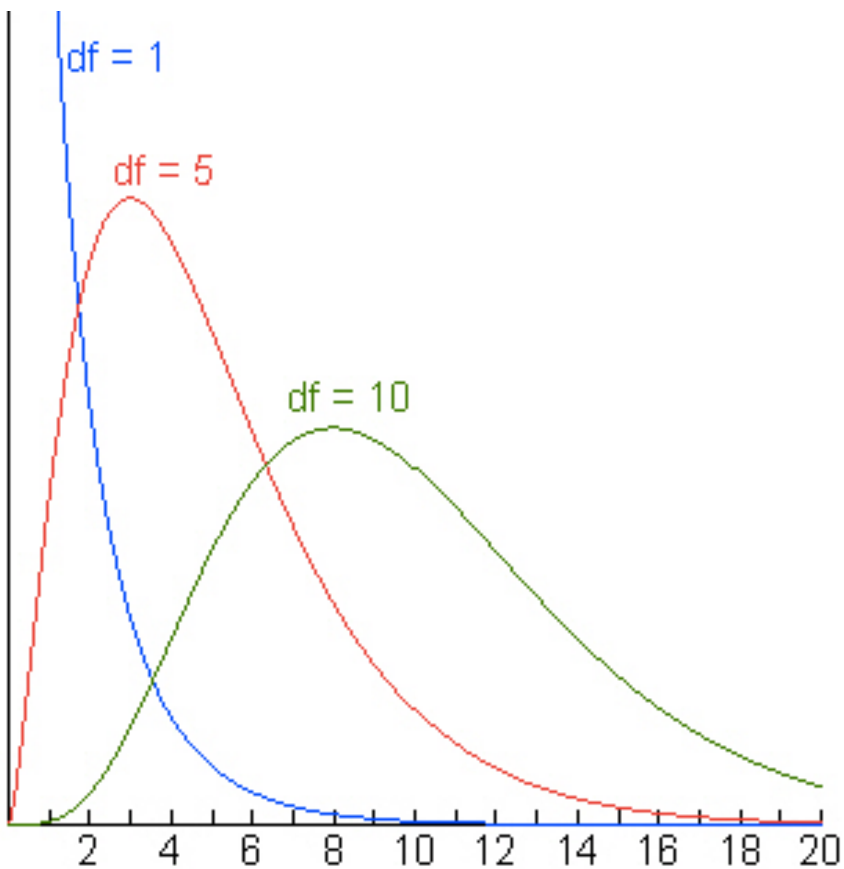
Chi-Square Test Statistic

- Let N be the total number of trials for which we have measured a categorical variable with k categories
- Measure the deviation of the expected from the observed counts:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

  - k is the number of possible outcomes in the multinomial trial
  - O_i is the number of **observed** samples in category i
  - E_i is the **expected** number of samples in category i
- The larger the value χ^2, the stronger the evidence against H_0
- Under certain conditions, χ^2 has a **chi-square distribution** which we can utilize to obtain p-values

Chi-square Distribution



- Parameterized by the degrees of freedom (just like the t)
- Right-skewed distribution
- Defined for nonnegative numbers
- Under H_0 the χ^2 test statistic is approximately chi-square distributed with *k*-1 df when **expected counts are large**
- Find p-value using Table A.7
- Only **one-sided** tests (only care if test statistic is large – right tail)

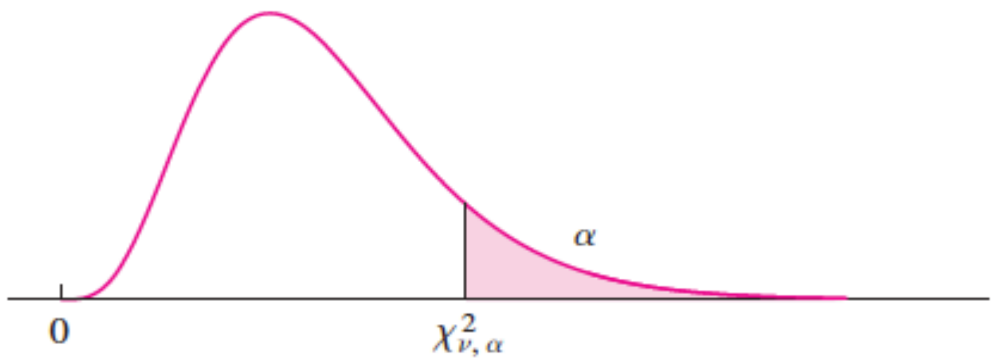
Null Distribution of χ^2 Test Statistic

- We do not have an **exact** distribution for χ^2 – there is only a good **approximate distribution when the expected counts are large**:
  - Rule of thumb: **expected** counts in each category are greater than or equal to 5

- Note the abuse of notation $\chi^2 \sim \chi^2_{k-1}$
  - "the chi-square test statistic has a chi-square distribution with k-1 degrees of freedom"
  - We will use χ^2 to denote the test statistic
  - We will use χ^2_{k-1} to denote the distribution

How to Use a χ^2 Table (A.7)

TABLE A.7 Upper percentage points for the χ^2 distribution



	α									
ν	**0.995**	**0.99**	**0.975**	**0.95**	**0.90**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188

Chi-square Test for the Die Example

Category	Observed	Expected
1	115	100
2	97	100
3	91	100
4	101	100
5	110	100
6	86	100
Total	600	600

$$\chi^2 = \frac{(115-100)^2}{100} + \frac{(97-100)^2}{100} + \frac{(91-100)^2}{100} + \frac{(101-100)^2}{100} + \frac{(110-100)^2}{100} + \frac{(86-100)^2}{100}$$

$$= 6.12$$

- The expected number in each category is at least 5, so χ^2 has a chi-square distribution with *k*-1 degrees of freedom

- The p-value (using the table) is: $P(\chi_5^2 > 6.12) > 0.10$

  Or using R can get a more precise estimate:

```
> pchisq(6.12, df=5, lower.tail=FALSE)
[1] 0.2947169
```

Chi-square Test for the Die Example

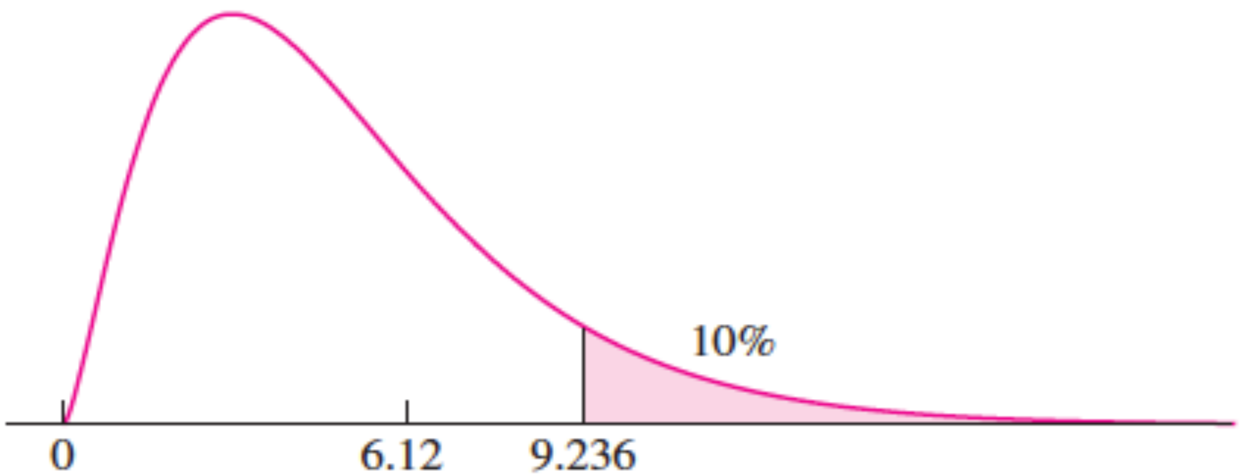


FIGURE 6.20 Probability density function of the χ^2_5 distribution. The observed value of the test statistic is 6.12. The upper 10% point is 9.236. Therefore the P-value is greater than 0.10.

Conclusion: Do not reject H_0- We do not have evidence to suggest that the die is not fair.

Chi-Square Test for Independence- Motivation

- What if a sampled item can fall into one of several categories for **two** variables?

- Example - Survey a random sample of N=200 students at UW
  1. Have you read the 'Hunger Games' series?
     a) Yes, I have read the entire series
     b) Yes, but only part of it
     c) No, I have not read any of it
  2. What is your gender?
     a) Male
     b) Female

- Place the results of the survey in a 2 x 3 table:

	Yes, I have read the entire series	Yes, but only part of it	No, I haven't read any of it	Totals
Male				
Female				
Totals				200

Chi-Square Test for Independence - Motivation

- We want to test the null hypothesis that the proportion of students who have read all, part of, or none of the 'Hunger Games' series is **independent** of gender

- Say we observe that in our sample of size N=200 there are
  - 100 males and 100 females
  - 25 who read the entire series, 50 who read part of it, and 125 who read none

- Under the **null hypothesis**, how many males do we expect have read the entire series?

Recall that if X and Y are independent, then P(X,Y)=P(X)*P(Y)

	Yes, I have read the entire series	Yes, but only part of it	No, I haven't read any of it	Totals
Male				100
Female				100
Totals	25	50	125	200

Chi-Square Test for Independence - Motivation

Recall that if X and Y are independent, then P(X,Y)=P(X)*P(Y)

Then, under the null hypothesis

P(Read entire AND Male) = P(Read entire) * P(Male)

= (25/200) * (100/200)

= 0.125*0.5 = 0.0625

So the expected **count** is 0.0625*N = 0.0625*200 = 12.5

	Yes, I have read the entire series	Yes, but only part of it	No, I haven't read any of it	Totals
Male	12.5	25	62.5	100
Female	12.5	25	62.5	100
Totals	25	50	125	200

General Formula for Expected Counts:

$$E_{ij} = \frac{\text{Row } i \text{ total} \times \text{Column } j \text{ total}}{N}$$

Chi-Square Test for Independence - Idea

We want a test statistic that measures the deviation of the observed from the expected counts:

Expected

	Yes, I have read the entire series	Yes, but only part of it	No, I haven't read any of it	Totals
Male	12.5	25	62.5	100
Female	12.5	25	62.5	100
Totals	25	50	125	200

Observed

	Yes, I have read the entire series	Yes, but only part of it	No, I haven't read any of it	Totals
Male	8	16	76	100
Female	17	34	49	100
Totals	25	50	125	200

Chi-Square Test Statistic for Independence

- Let N be the total number of trials for which we have measured **two** categorical variables (correspond to rows and columns)

- Null hypothesis: Row variable is independent of Column variable

- Test statistic: $$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

  - I and J are the number of rows and columns, respectively
  - O_{ij} is the number of **observed** samples in row i, column j
  - E_{ij} is the **expected** number of samples in row i, column j

- When the expected count of each cell is at least 5, under the null hypothesis of independence $\chi^2 \sim \chi^2_{(I-1)*(J-1)}$

Hunger Games Example

- Compute the test statistic:

$$\chi^2 = \frac{(8-12.5)^2}{12.5} + \frac{(16-25)^2}{25} + \frac{(76-62.5)^2}{62.5} +$$

$$\frac{(17-12.5)^2}{12.5} + \frac{(34-25)^2}{25} + \frac{(49-62.5)^2}{62.5}$$

$$= 15.552$$

- Under the null hypothesis that reading Hunger Games is independent of gender, $\chi^2 \sim \chi^2_2$
  - there are two degrees of freedom because (I-1)*(J-1) = (2-1)*(3-1) = 2
- Using Table A.7: P-value = $P(\chi^2_2 > 15.552) < 0.005$

  Using R: P-value = 0.00042

Chi-square Test for Homogeneity

- In the previous setting, the row and column totals were both **random**
  - We set out to sample 200 students; didn't know in advance how many males/females we would get, or how many had read the entire series

- Sometimes, either the row totals or column totals are **fixed**
  - if we had decided beforehand to sample 100 males and 100 females, the row totals would have been fixed

- If the row totals are fixed and the column totals are random:
  - We want to test the null hypothesis that the proportion in each column category is the same for each row category **(Homogeneity)**
  - Not quite the same as independence, but **we can test for it in exactly the same way!** (*they are mathematically equivalent under the null*)

Example 6.21 - Steel Pins

- Steel pins are sampled from four different machines
- The number of pins in each category ("Too Thin", "OK", or "Too Thick") is counted

TABLE 6.4 Observed numbers of pins in various categories with regard to a diameter specification

	Too Thin	OK	Too Thick	Total
Machine 1	10	102	8	120
Machine 2	34	161	5	200
Machine 3	12	79	9	100
Machine 4	10	60	10	80
Total	66	402	32	500

Row totals are **fixed**

- H_0: the proportion of pins that are too thin, OK, or too thick are the same for all machines (homogeneity)

Steel Pin Example Continued

H_0: For each column j (j=1, 2, 3), $p_{1j}=p_{2j}=p_{3j}$

Each cell has an expected count of at least 5

TABLE 6.4 Observed numbers of pins in various categories with regard to a diameter specification

	Too Thin	OK	Too Thick	Total
Machine 1	10	102	8	120
Machine 2	34	161	5	200
Machine 3	12	79	9	100
Machine 4	10	60	10	80
Total	66	402	32	500

Expected values for Table 6.4

	Too Thin	OK	Too Thick	Total
Machine 1	15.84	96.48	7.68	120.00
Machine 2	26.40	160.80	12.80	200.00
Machine 3	13.20	80.40	6.40	100.00
Machine 4	10.56	64.32	5.12	80.00
Total	66.00	402.00	32.00	500.00

$$\chi^2 = \frac{(10-15.84)^2}{15.84} + \cdots + \frac{(10-5.12)^2}{5.12}$$

$$= \frac{34.1056}{15.84} + \cdots + \frac{23.8144}{5.12}$$

$$= 15.5844$$

Degrees of freedom = (4-1)*(3-1) = 6

Table: p-value = $P(\chi^2_6 > 15.5844)$
so $0.01 < \text{p-value} < 0.025$

Using R: p-value = 0.0162

Chi-Square Test Summary

Let I be the number of rows and J be the number of columns in a table where the rows and columns represent categories of two variables of interest. Let O_{ij} be the observed count for row i and column j (out of N).

1. Set up the null and alternative hypotheses:
   a) For a test of **independence**- H_0: Row variable is independent of column variable
   b) For a test of **homogeneity**- Let the row totals be fixed; H_0: proportion in each column category is the same for each row category
2. State the level of significance α you will use
3. Calculate the test statistic:

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

Where E_{ij} is the expected count in row i, column j under H_0

4. Assume H_0 is true and find the P-value: $P(\chi^2_{(I-1)*(J-1)} > \chi^2)$
5. Make a conclusion based on the P-value

Example – Titanic Survival Rate

- There were 2201 people on board the *Titanic*
- We want to know if we can conclude that **ticket type was dependent on survival** using a significance level of 0.01

	Ticket Type				
Survival	**Crew**	**First**	**Second**	**Third**	**Totals**
Alive	212	202	118	178	710
Dead	673	123	167	528	1491
Totals	885	325	285	706	2201

Next

- Hypothesis Tests for Variances: F-test
- Power and Type I error
- Multiple Testing Issues