Chapter 8:
Fitting probability models to frequency data

Goodness-of-fit

- The binomial test from the last chapter is an example of a **goodness-of-fit test**
- Goodness-of-fit test: method of comparing an observed frequency distribution with the frequency distribution expected under a probability model
- But binomial test is limited
  - Data must fit into **two** mutually exclusive outcomes (success and failure)

χ^2 goodness-of-fit test

- The **χ^2 goodness-of-fit test** compares frequency data to a probability model stated by the null
- More general than the binomial test because it **can handle more than two categories**
- Calculations also easier

3

Ex 8.1: No weekend getaways

- Are babies born at the same frequency on each day of the week?
  - 7 categories (Sunday, Monday, Tuesday, etc.)
- Random sample of 180 births from 2016

TABLE 8.1-1 Day of the week for 180 births in the U.S. in 2016.

Day	Number of births
Sunday	14
Monday	26
Tuesday	34
Wednesday	21
Thursday	27
Friday	38
Saturday	20
Total	180

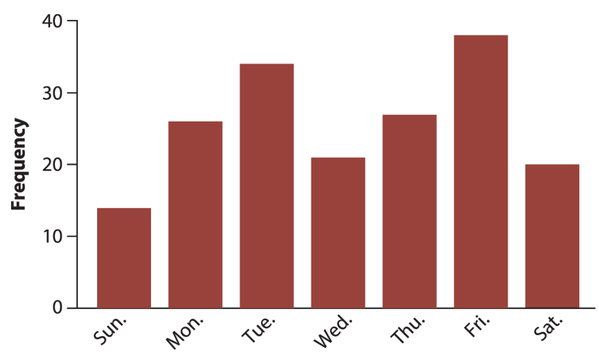


Fig 8-1.1

4

Hypotheses

- H_0: the probability of birth is the same on every day of the week
- H_A: the probability of birth is not the same on every day of the week
- We have observed frequencies based on our sample of 180 births, but we **need to calculate the expected frequency under the null hypothesis**

5

Observed and expected frequencies

TABLE 8.1-2 Expected frequency of births on each day of the week in 2016 under the proportional model.

Day	Number of days in 2016
Sunday	52
Monday	52
Tuesday	52
Wednesday	52
Thursday	52
Friday	53
Saturday	53
Sum	366

6

Observed and expected frequencies

TABLE 8.1-2 Expected frequency of births on each day of the week in 2016 under the proportional model.

Day	Number of days in 2016	Proportion of days in 2016
Sunday	52	52/366
Monday	52	52/366
Tuesday	52	52/366
Wednesday	52	52/366
Thursday	52	52/366
Friday	53	53/366
Saturday	53	53/366
Sum	366	1

7

Observed and expected frequencies

TABLE 8.1-2 Expected frequency of births on each day of the week in 2016 under the proportional model.

Day	Number of days in 2016	Proportion of days in 2016	Expected frequency of births
Sunday	52	(52/366) x 180 =	25.574
Monday	52	52/366	25.574
Tuesday	52	52/366	25.574
Wednesday	52	52/366	25.574
Thursday	52	52/366	25.574
Friday	53	53/366	26.066
Saturday	53	53/366	26.066
Sum	366	1	180

8

χ^2 test statistic

- The ***χ^2* test statistic** measures the discrepancy between observed frequencies from the data and expected frequencies from the null hypothesis

$$\chi^2 = \sum \frac{(Observed_i - Expected_i)^2}{Expected_i}$$

9

χ^2 test statistic

TABLE 8.1-3 Observed and expected numbers of births on each day of the week under the proportional model.

Day	Observed number of births	Expected number of births	$\frac{(Observed - Expected)^2}{Expected}$
Sunday	14	25.574	5.238
Monday	26	25.574	0.007
Tuesday	34	25.574	2.776
Wednesday	21	25.574	0.818
Thursday	27	25.574	0.080
Friday	38	26.066	5.464
Saturday	20	26.066	1.412
Sum	180	180	15.795

10

Sampling distribution of χ^2

- Recall that in chap 6 the authors created a null distribution of righty vs lefty toads using computer simulation
- The χ^2 distribution is a mathematical function, and it's features have been compiled in tables (or stored in computer commands)

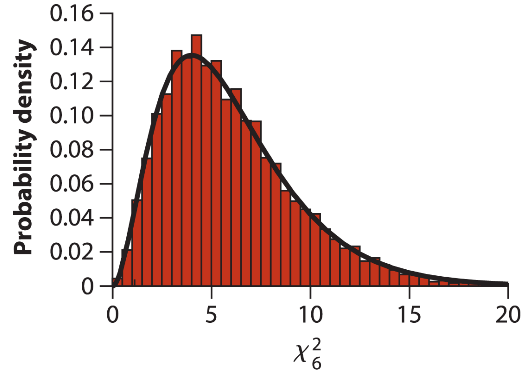


Fig 8-1.2

11

Degrees of freedom

- The χ^2 distribution is a mathematical function, and to use it we need to specify the degrees of freedom (*df*)
- The number of **degrees of freedom** of a χ^2 statistic specifies which χ^2 distribution to use as the null hypothesis
- https://www.youtube.com/watch_popup?v=rATNoxKg1yA

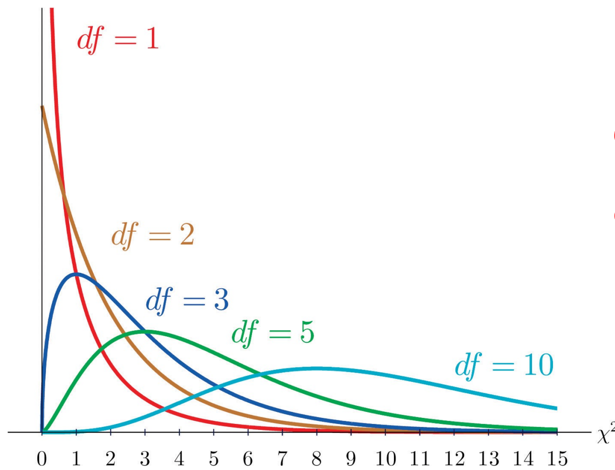
$$df = (num\ categories) - 1 - (num\ parameters\ estimated\ from\ data)$$

this last term is often zero

12

Degrees of freedom

- Degrees of freedom (*df*) are important because the shape of the χ^2 distribution, and thus the cutoff points for α, change as *df* changes



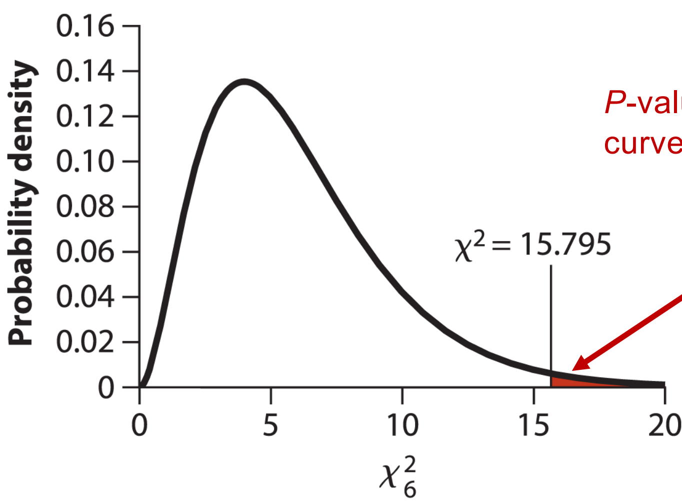
Our problem:

df = 7 – 1 = 6

13

P-value for χ^2 test

- χ^2 test statistic = 15.795; *df* = 6

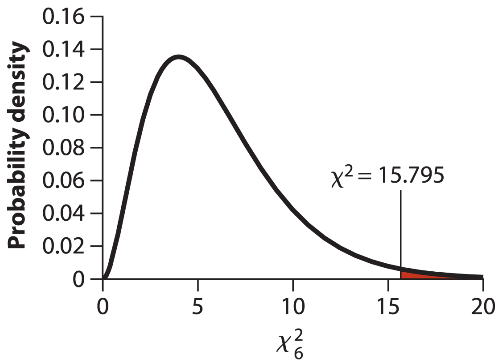


P-value = probability (area under curve) of getting $\chi^2 \geq 15.795$

Fig 8-1.3 Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

14

P-value for χ^2 test



Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

Fig 8-1.3

- Two options for determining if your χ^2 is significant
- #1: get exact *P*-value from computer software
  - *P* = 0.0149
- #2: get a critical value for the test using a table

15

Critical value

- A **critical value** is the value of a test statistic that marks the boundary of a specified area in the tail (or tails) of the sampling distribution under H_0

16

Critical value

TABLE 8.1-4 An excerpt from the table of χ^2 critical values (Statistical Table A). Numbers down the left side are the number of degrees of freedom (df). Numbers across the top are significance levels (α). The critical value for a χ^2 distribution with $df = 6$ and $\alpha = 0.05$ is 12.59 (indicated in red).

Significance level (α)

df	0.999	0.995	0.99	0.975	0.95	0.05	0.025	0.01	0.005	0.001
1	0.000002	0.00004	0.00016	0.00098	0.00393	3.84	5.02	6.63	7.88	10.83
2	0.002	0.01	0.02	0.05	0.10	5.99	7.38	9.21	10.6	13.82
3	0.02	0.07	0.11	0.22	0.35	7.81	9.35	11.34	12.84	16.27
4	0.09	0.21	0.30	0.48	0.71	9.49	11.14	13.28	14.86	18.47
5	0.21	0.41	0.55	0.83	1.15	11.07	12.83	15.09	16.75	20.52
6	0.38	0.68	0.87	1.24	1.64	12.59	14.45	16.81	18.55	22.46
7	0.60	0.99	1.24	1.69	2.17	14.07	16.01	18.48	20.28	24.32
8	0.86	1.34	1.65	2.18	2.73	15.51	17.53	20.09	21.95	26.12

17

Significant χ^2 test

TABLE 8.1-3 Observed and expected numbers of births on each day of the week under the proportional model.

Day	Observed number of births	Expected number of births	$\frac{(Observed - Expected)^2}{Expected}$
Sunday	14	25.574	5.238
Monday	26	25.574	0.007
Tuesday	34	25.574	2.776
Wednesday	21	25.574	0.818
Thursday	27	25.574	0.080
Friday	38	26.066	5.464
Saturday	20	26.066	1.412
Sum	180	180	15.795

This discrepancy is largely due to scheduled C-sections and induced labor, but these do not explain the effect completely (Ventura at al. 2001)

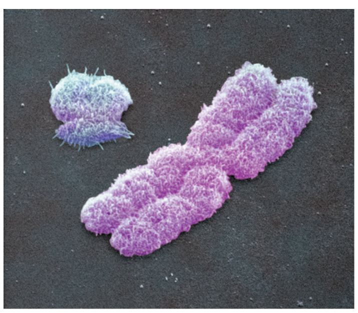
18

χ^2 goodness-of-fit test assumptions

- None of the categories should have an expected frequency less than one
- No more than 20% of the categories should have expected frequencies less than five

19

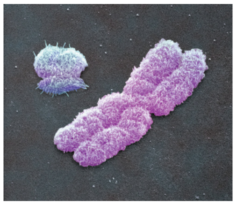
Ex 8.4: Gene content of the human X chromosome



- Human Genome Project found 19,628 genes, 839 of which were on the X chromosome
- Are the number of genes on the X chromosome as expected due to chance?
- X chromosome represents 5.2% of entire genome

20

Ex 8.4: Gene content of the human X chromosome



- What is the expected number of genes?
- X chromosome represents 5.2% of entire genome
- Version of genome used in this study has 19,628 genes
- So null expectation for genes on X:
  - 19,628 x 0.052 = 1,020.7
- But the observed number of genes on the X is 839

21

Two categories

TABLE 8.3-1 Numbers of genes on the human X chromosome and on the rest of the genome.

Chromosome	Observed	Expected
X	839	1,020.7
Not X	18,789	18,607.3
Total	19,628	19,628

22

Two categories

- Could use the binomial test, but calculations would be challenging
  - P = 2 x Pr[X=0] + Pr[X=1] + Pr[X=2] + … + Pr[X=839]

- Easier to calculate the χ^2 statistic if you're doing it by hand!

$$\chi^2 = \sum \frac{(Observed_i - Expected_i)^2}{Expected_i} = \frac{(839 - 1020.7)^2}{1020.7} + \frac{(18789 - 18607.3)^2}{18607.3} = 34.1$$

23

χ^2 result

- χ^2 = 34.1
- df = 2 – 1 = 1
- Using a table:
  - P < 0.001

TABLE 8.1-4 An excerpt from the table of χ^2 critical values (Statistical Table A). Numbers down the left side are the number of degrees of freedom (df). Numbers across the top are significance levels (α). The critical value for a χ^2 distribution with $df = 6$ and $\alpha = 0.05$ is 12.59 (indicated in red).

	Significance level (α)									
df	0.999	0.995	0.99	0.975	0.95	0.05	0.025	0.01	0.005	0.001
1	0.000002	0.00004	0.00016	0.00098	0.00393	3.84	5.02	6.63	7.88	10.83
2	0.002	0.01	0.02	0.05	0.10	5.99	7.38	9.21	10.6	13.82
3	0.02	0.07	0.11	0.22	0.35	7.81	9.35	11.34	12.84	16.27
4	0.09	0.21	0.30	0.48	0.71	9.49	11.14	13.28	14.86	18.47
5	0.21	0.41	0.55	0.83	1.15	11.07	12.83	15.09	16.75	20.52
6	0.38	0.68	0.87	1.24	1.64	12.59	14.45	16.81	18.55	22.46
7	0.60	0.99	1.24	1.69	2.17	14.07	16.01	18.48	20.28	24.32
8	0.86	1.34	1.65	2.18	2.73	15.51	17.53	20.09	21.95	26.12

24

χ^2 result

- χ^2 = 34.1
- *df* = 2 – 1 = 1
- Using a table:
  – *P* < 0.001
- Using computer:
  – *P* = 2.03 x 10^{-9}

TABLE 8.1-4 An excerpt from the table of χ^2 critical values (Statistical Table A). Numbers down the left side are the number of degrees of freedom (df). Numbers across the top are significance levels (α). The critical value for a χ^2 distribution with $df = 6$ and $\alpha = 0.05$ is 12.59 (indicated in red).

	Significance level (α)									
df	0.999	0.995	0.99	0.975	0.95	0.05	0.025	0.01	0.005	0.001
1	0.000002	0.00004	0.00016	0.00098	0.00393	3.84	5.02	6.63	7.88	10.83
2	0.002	0.01	0.02	0.05	0.10	5.99	7.38	9.21	10.6	13.82
3	0.02	0.07	0.11	0.22	0.35	7.81	9.35	11.34	12.84	16.27
4	0.09	0.21	0.30	0.48	0.71	9.49	11.14	13.28	14.86	18.47
5	0.21	0.41	0.55	0.83	1.15	11.07	12.83	15.09	16.75	20.52
6	0.38	0.68	0.87	1.24	1.64	12.59	14.45	16.81	18.55	22.46
7	0.60	0.99	1.24	1.69	2.17	14.07	16.01	18.48	20.28	24.32
8	0.86	1.34	1.65	2.18	2.73	15.51	17.53	20.09	21.95	26.12

25

Which test when there are two categories?

- If you're doing it by hand: χ^2 (easier calculation)
- If you're using a computer?
- Binomial test is recommended (will give exact *P*-value)
  – Sampling distribution of the χ^2 statistic is an approximation of the χ^2 distribution
  – Approximation is excellent, but does not generate an exact *P*-value

26

Notes

- Skipping sections 8.4