Applied Statistics

Lectures10

David Barton & Sabine Hauert

Department of Engineering Mathematics

Outline

- Categorical data
- Pearson's χ^2 test

OpenIntro Statistics

Chapter 6, particularly §6.3 and §6.4

Categorical data

Categorical data is data that has been broken into categories of some description.

Consider data from Home Accident Surveillance System (HASS) —
`www.hassandlass.org.uk` 16-18 Uk hospitals
1978-2002

What makes a typical toddler?

Mechanism	Count
Fall	6348
Struck — static object	1259
Pinch/crush (blunt)	595
Cut/tear (sharp)	323
Foreign body	821
(Suspected) poisoning	658
Total	10004

(Boys aged 0-4, data submitted from 18 hospitals for 2002)

What makes a typical toddler?

If there were two categories then we could use a Binomial distribution with a given number of events (say 40) to determine whether a particular toddler is “normal”

Probabilities

Fall	Struck
$\frac{6348}{6348+1259} = 0.83$	$\frac{1259}{6348+1259} = 0.17$

mathematical model describing normal toddler accidents

Expected

Fall	Struck
$40 \times 0.83 = 33.2$	$40 \times 0.17 = 6.8$

Observed

Fall	Struck
36	4

actual data

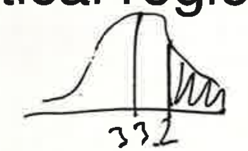
can I reject this null hypothesis or not?

H_0: the proportion of Falls to Struck follows a Binomial distribution with probability $p = 0.83$; i.e., $\text{Falls} \sim B(0.83, 40)$ follows this distribution

Predictions of categorical data (Binomial)

Use a significance level of 5% with our test statistic being the number of Falls (assume we know the total number of accidents)

$\boxed{\text{Calculate the p-value}}$ (can't find the critical region for this — discrete variable)



number falls

$$p = 2\min(P(F \geqslant 36), P(F \leqslant 36))$$
$$= 2P(F \geqslant 36)$$
$$= 2\left(\binom{40}{36}0.83^{36}0.17^{4} + \binom{40}{37}0.83^{37}0.17^{3} + \cdots\right)$$
$$= ?!$$

very large   very small

→ imprecise calculation - although can be done!
& difficult by hand

In principle this can be calculated but in practice it's difficult. . .

Predictions of categorical data (Binomial)

However we can use the normal approximation to the Binomial distribution!

discrete value

same mean and variance

When n is large ($n \geq 30$) $p \approx 0.5$ → normal distribution approximates binomial distribution

$$\mu = \underbrace{np}_{40,\ 0.83} = 33.2, \qquad \sigma^2 = \underbrace{np(1-p)}_{40,\ 0.83,\ 0.17} = 5.644$$

Hence $F \sim N(np, np(1-p)) = N(33.2, 5.644)$ (approximately)

continuous values

$p = 2P(F_B \geq 36) \approx 2P(F_N \geq 35.5)$

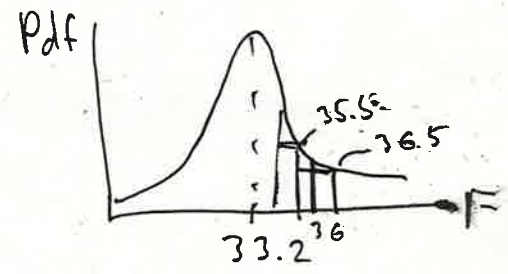
$= 2P\left(Z \geq \frac{35.5 - 33.2}{\sqrt{5.644}}\right) = 2P(Z \geq 0.968)$

$= 2 \cdot 0.166 = 0.332 > 0.05$

Can not reject the null hypothesis!

continuity correction

$P(F_B = 36) \approx P(35.5 \leq F_N \leq 36.5)$



Hence we are well above the 5% significance level and cannot reject H_0 (the exact answer is $p = 0.3334$)

Predictions of categorical data

How can we generalise to multiple categories? Use a multinomial (or categorical) distribution.

This gives the *Fisher Exact test* but it's a pain to use. . . software

Do use it when there are few observations though!

Exact means that no approximations have been made. However, the normal approximation is quite useful and more convenient!

Pearson's χ^2 test is an approximate test that works well for large (and not so large) numbers of measurements. In the limit as $n \to \infty$ it gives the exact answer.

At least 5 observations in each category — years bs hand

(Note: many statistical tests are approximate in this way because the exact versions are too difficult to work with!)

Pearson's χ^2 test

Label observed outcomes as O_i (random variables!); $n = \sum_i O_i$.

Label expected outcomes as $E_i = nP_i$ (*not* random variables!) where P_i are the probabilities of each outcome.

Observed

Mechanism	Count	Probability	Expected	Outcome
Fall	6348	$\frac{6348}{10009}$ 0.635	$\cancel{10009}$ 200 · 0.635 = 127	140
Struck — static object	1259	0.126	25.2	20
Pinch/crush (blunt)	595	0.059	11.8	20
Cut/tear (sharp)	323	0.032	6.4	4
Foreign body	821	0.082	16.4	16
(Suspected) poisoning	658	0.066	13.2	0
Total	10004	1.000	200	200

Same

Pearson's χ^2 test

Pearson's χ^2 test states that

observed expected

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

degrees of freedom

where m is the number of categories

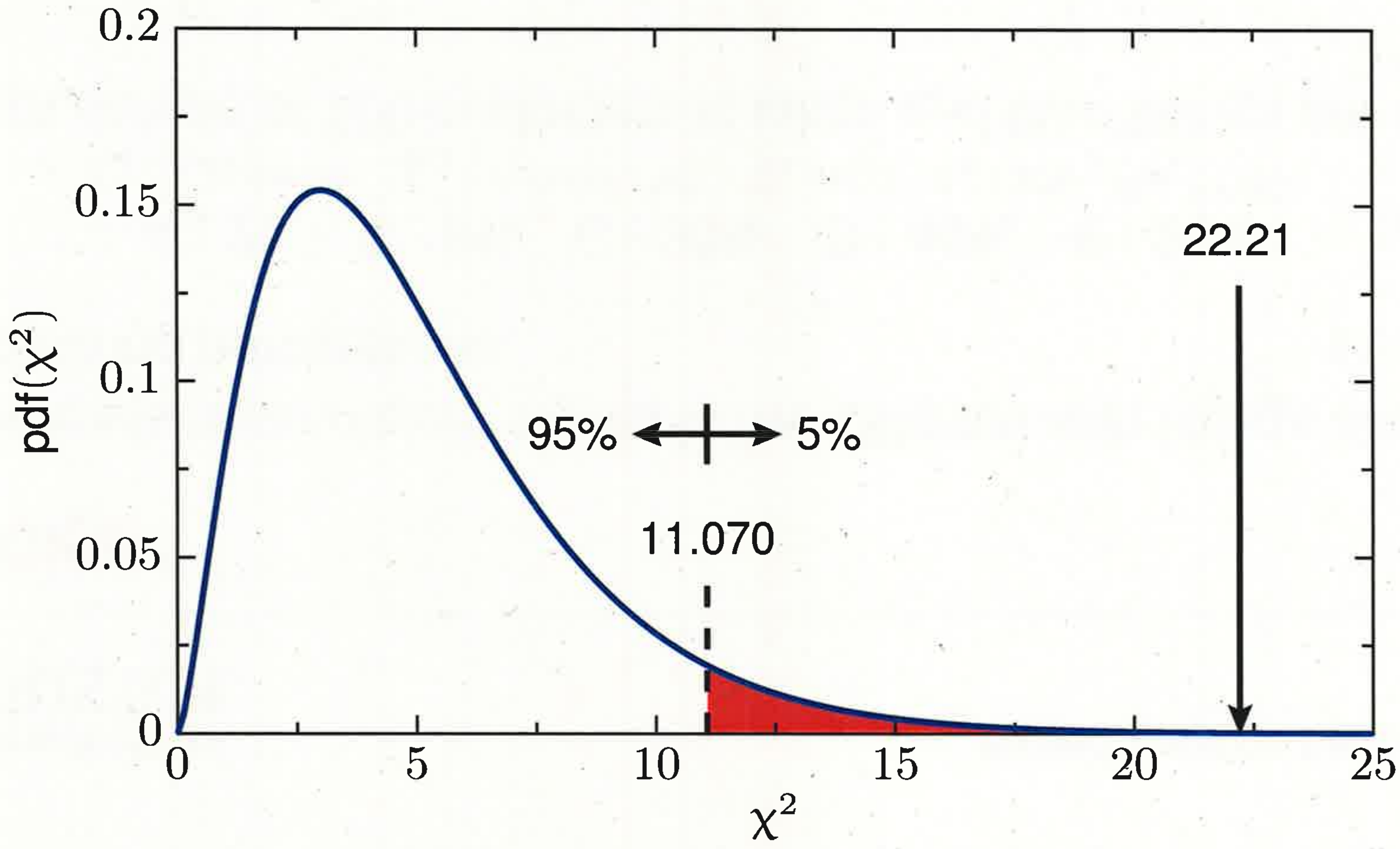
H_0: the toddler is "typical" (i.e., their distribution of accidents follows the nationwide statistics).

$$\chi^2 = \frac{(140-127)^2}{127} + \frac{(20-25.2)^2}{25.2} + \frac{(20-11.8)^2}{11.8} + \frac{(4-6.4)^2}{6.4} + \frac{(16-16.4)^2}{16.4} + \frac{(0-13.2)^2}{13.2} = 22.21$$

Pearson's χ^2 test

6 categories and so $6 - 1 = 5$ degrees-of-freedom. Tables give a critical value of $\chi^2_5(0.05) = 11.070$ and so we reject the null hypothesis.

degrees of freedom



Exercise

A boot manufacturer makes moves in five different with fittings according to the following percentages.

A : 2% B : 8% C : 30% D : 40% E : 20%

$E_A = 0.02 \cdot 500 = 10$ $E_B = 0.08 \cdot 500 = 40$ $E_C = 150$ $E_D = 200$ $E_E = 100$

$$\chi^2 = \frac{(12-10)^2}{10} + \frac{(42-40)^2}{40} + \frac{(171-150)^2}{150} + \frac{(178-200)^2}{200} + \frac{(93-100)^2}{100} = 7.15$$

A random sample of 500 customers is taken and their fittings are as follows:

Observed $\{$ A : 12 B : 46 C : 171 D : 178 E : 93

Does this sample suggests that the proportions of the five width fittings are different from the model assumed by the boot manufacturer?

$$\chi_1^2 = 3.841, \quad \chi_2^2 = 5.991, \quad \chi_3^2 = 7.815,$$
$$\chi_4^2 = 9.488, \quad \chi_5^2 = 11.070$$

$\chi_4^2 = 9.488 > 7.15$ Can not reject H_0