MAST90044 Thinking and Reasoning with Data — Chapter 4

HYPOTHESIS TESTING FOR CATEGORICAL DATA

Chapter 4:

- Hypothesis Testing
  - The P-value
  - Hypothesis tests for a proportion
  - The sign test
- Contingency Tables
  - The chi-square test
  - Fisher's exact test

The P-value: an example of how we use it

- A lady who claims to be an expert in tea preparation claims that the tea tastes different if the milk is added to the tea or the tea is added to the milk.
- To test her claim, she was given 10 cups of tea to taste and assess the preparation order.
- We model the number of successful assessments she makes to be $X \stackrel{\mathrm{d}}{=} \mathrm{Bi}(10, p)$.

The P-value: an example of how we use it

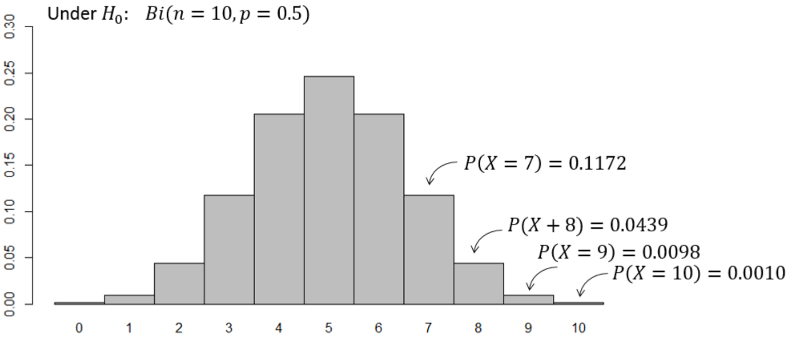
We wish to test the null hypothesis:
$H_0 : p = 0.5$ (i.e. she is just guessing) versus
Alternative hypothesis:
$H_1 : p > 0.5$ (i.e. she has some ability to tell the difference).

Suppose she makes **7 correct assessments** -
is there sufficient evidence to say that this is better than guesswork?
If $p = 1/2$ then

$$\begin{aligned}\mathbb{P}(X \geqslant 7) &= \mathbb{P}(X = 7) + \mathbb{P}(X = 8) + \mathbb{P}(X = 9) + \mathbb{P}(X = 10)\\ &\approx 0.1172 + 0.0439 + 0.0098 + 0.0010 \approx 0.172\end{aligned}$$

So the (One-sided) P-value is 0.172.

The P-value: an example of how we use it



$$\mathbb{P}(X \geqslant 7) = \mathbb{P}(X = 7) + \mathbb{P}(X = 8) + \mathbb{P}(X = 9) + \mathbb{P}(X = 10)$$
$$\approx 0.1172 + 0.0439 + 0.0098 + 0.0010 \approx 0.172$$

Not a small one !!!

The P-value: an example of how we use it

```
> dbinom(7:10,10,.5)
[1] 0.1171875000 0.0439453125 0.0097656250 0.0009765625.
> pbinom(6,10,.5)
[1] 0.828125
```

So the (One-sided) P-value is 0.172.

The P-value

The P-value is defined as the probability of obtaining a result that is as extreme or more extreme than that actually observed, assuming the null hypothesis is true.

The P-value *is not* the probability that the null hypothesis is true.

The P-value: an example of how we use it

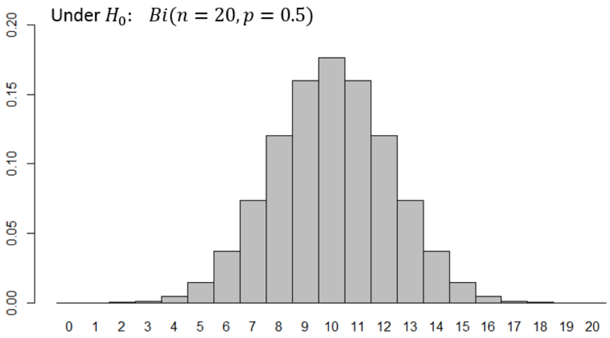
- We are interested in whether a coin in our possession is a fair one.
- So we will toss it 20 times and observe the number of heads
- $N \stackrel{\mathrm{d}}{=} \mathrm{Bi}(20, p)$.

  We wish to test the null hypothesis
  $H_0 : p = 0.5$ (i.e. the coin is fair) versus

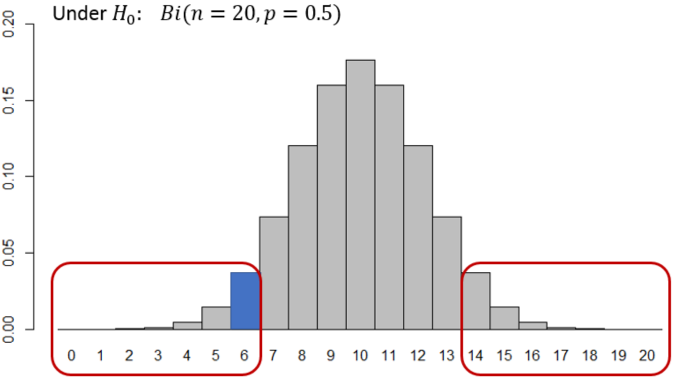
  Alternative hypothesis
  $H_1 : p \neq 0.5$ (i.e. the coin is not fair).

Suppose we **toss the coin 20 times** and **observe 6 heads**. Is there strong evidence to suggest that the coin is not fair?

The P-value: an example of how we use it



The P-value: an example of how we use it



$$H_1: p \neq 0.5$$

The P-value: an example of how we use it

Let $\hat{p} = N/20$. Assuming that $p = 0.5$ we have

$$\begin{aligned} P - value &= \mathbb{P}(\hat{p} \geqslant 14/20) + \mathbb{P}(\hat{p} \leqslant 6/20) \\ &= \mathbb{P}(N \geqslant 14) + \mathbb{P}(N \leqslant 6). \end{aligned}$$

This is approximately 0.115, so the P-value for this test is 0.115.

Hypothesis tests for a proportion: large sample result

Null hypothesis is $H_0 : p = p_0$. Estimator is $\hat{p}$.

Now What?

- Sampling variability: the calculated proportion will vary from sample to sample.
- Do we know how?
- Almost!
- $X \stackrel{\mathrm{d}}{=} \mathrm{Bi}(n,p)$ => $\bar{X} = np$ & $var(X) = np(1-p)$
- In addition, according to CLT $\bar{X} = np \stackrel{\mathrm{d}}{=} N(np, np(1-p)/n)$
- Let's get rid of the n
- Yeaha! $p \stackrel{\mathrm{d}}{=} N(p, p(1-p)/n)$
- Now we know the 'Expected' behaviour!

Hypothesis tests for a proportion: large sample result

Null hypothesis is $H_0 : p = p_0$. Estimator is $\hat{p}$.

Under H_0, $p \stackrel{\mathrm{d}}{=} N(p_0, p_0(1-p_0)/n)$

Test statistic $\quad Z = \dfrac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$

is approximated by a standard normal distribution, $\mathrm{N}(0,1)$.

Z represents how many standard deviations (assuming H_0) the observed value and hypothesised value differ by.

Hypothesis tests for a proportion: large sample result

A casino wishes to test if a newly purchased 6 sided die has probability 1/6 of rolling a 6.

They roll the die 300 times and observe the number N of 6's rolled, $N \stackrel{\mathrm{d}}{=} \mathrm{Bi}(300, p)$.

$H_0 : p = 1/6$ vs $H_1 : p \neq 1/6$

Suppose that they observe 58 sixes out of 300 rolls, $\hat{p} = 58/300 \approx 0.1933$

Hypothesis tests for a proportion: large sample result

Under H_0, sd$(\hat{p}) = \sqrt{\frac{1/6(1-1/6)}{300}} \approx 0.0215$

$Z = \frac{0.1933 - 1/6}{0.0215} \approx 1.24$

```
> 2*(1-pnorm(1.24))
[1] 0.2149754
```

Two-sided P-value ≈ 0.215. Do not reject H_0.

One-sided vs two-sided hypothesis tests

- Convention in research is to prefer two-sided tests.
- One-sided tests assume the effect can only be in one direction.
  Not often a reasonable assumption!
- Researcher needs to show that a one-sided alternative is justified.

The sign test

Effect of tyres on the fuel consumption of cars

Eight different cars were driven over a set course, once fitted with regular tyres and once with radial tyres. Fuel consumptions:

Car	1	2	3	4	5	6	7	8
Radial tyres	9.3	15.4	16.9	11.7	14.8	13.5	10.2	13.8
Regular tyres	9.8	15.2	17.3	11.8	14.8	14.3	10.5	14.1

H_0: No difference between radial and regular tyres.

H_1: A difference.

n = number of non-zero differences = 7.

x = number of positive differences = 6.

Under H_0, $X \overset{\mathrm{d}}{=} \mathrm{Bi}(7, 0.5)$.

$\mathbb{P}(X \geqslant 6) = 0.0625$, so $P = 0.125$.

The sign test

A hypothesis test which uses the binomial distribution

- Used for paired samples when the assumption of normality of the differences is not reasonable;
- Tests the null hypothesis that the median of the differences is zero;
- Uses only the number of positive (or negative) differences among the non-zero differences;
- Only assumption: the differences are independent;
- The test statistic is the number of positive (or negative) differences.

Limitations of the sign test

Creates a binary variable from a continuous variable. Ignores the magnitude of the differences, so does not use all the information. Not as powerful statistically as some other tests.

The chi-squared test

Used to:

1. Compare tables of frequencies with expected frequencies under a null hypothesis.
2. Investigate associations in cross-tables of frequencies (contingency tables). This is equivalent to comparing two or more proportions.
3. Based on the χ^2 distribution.

Example 1: 120 throws of a die

	1	2	3	4	5	6
frequency	25	18	28	20	16	13

Is the die biased?

Example 2: Are pianists more likely to play guitar than saxophonists?

Survey of musicians at a Jazz Academy:

“What is your main instrument?"

vs

“Do you play the guitar?"

		Main instrument		
		Piano	Saxophone	Total
Play guitar	Yes	37	14	51
	No	36	34	70
	Total	73	48	121

Is the main instrument associated with guitar playing?

The chi-squared(χ^2) test

The test statistic takes the form

$$X^2 = \sum \frac{(observed - expected)^2}{expected}$$

Where:

- observed is the observed frequency of a category.
- expected is the frequency that would be expected if the hypothesis being tested (H_0) is true.

For the die tossing example, all of the expected frequencies are 20 – if the die is unbiased. Under H_0, X^2 is approximately χ^2_5 distributed, where $5 = 6 - 1$ is the number of degrees of freedom.

Is the die biased?

	1	2	3	4	5	6
(observed) frequency	25	18	28	20	16	13
(expected) frequency	20	20	20	20	20	20

H_0 : the die is unbiased (all outcomes are equally likely)

H_1 : H_0 is not true (some outcomes are more likely than others)

$$\begin{aligned} & \frac{(25-20)^2}{20} + \ldots + \frac{(13-20)^2}{20} \\ = \quad & 1.25 + 0.20 + 3.20 + 0 + 0.80 + 2.45 \\ = \quad & 7.90 \end{aligned}$$

P-value is $\mathbb{P}(X^2 \geqslant 7.90)$ where $X^2 \sim \chi^2_5$.

```
> 1 - pchisq(7.9, 5)
[1] 0.1618336
```

Do not reject H_0.

Is the main instrument associated with guitar playing?

		Main instrument		
		Piano	Saxophone	Total
Play guitar	Yes	37	14	51
	No	36	34	70
	Total	73	48	121

How many pianists are expected to play guitar? (under H_0)

Overall proportion of musicians who play guitar = 51/121 = 0.421.

73 pianists, so expected number is

$73 \times \frac{51}{121} = 30.77 = \frac{73 \times 51}{121}$

Expected frequency in row i and column j =

$$\frac{(\text{row } i \text{ total}) \times (\text{column } j \text{ total})}{\text{grand total}}$$

Observed and expected values

		Main instrument		
		Piano	Saxophone	Total
Play guitar	Yes	37	14	51
	No	36	34	70
	Total	73	48	121

		Main instrument		
		Piano	Saxophone	Total
Play guitar	Yes	30.77	20.23	51
	No	42.23	27.77	70
	Total	73	48	121

Contingency tables

2-way contingency table: Two factors, which determine the 'rows' and 'columns' of the table, with r and c levels, respectively.

The hypothesis tested is

H_0 : no association between the two factors;

H_1 : H_0 is not true.

The degrees of freedom for the test are given by

$$(\#\, rows - 1) \times (\#\, columns - 1) = (r-1)(c-1)$$

Is the main instrument associated with guitar playing?

		Main instrument				
		Piano		Saxophone		Total
		O	E	O	E	
Play	Yes	37	30.77	14	20.23	51
guitar	No	36	42.23	34	27.77	70
	Total	73		48		121

$$
\begin{aligned}
x^2 &= \frac{(37-30.77)^2}{30.77} + \frac{(14-20.23)^2}{20.23} + \frac{(36-42.23)^2}{42.23} + \frac{(34-27.77)^2}{27.77} \\
&= 5.50 \ ,
\end{aligned}
$$

compared with χ^2_1 distribution gives $P = 0.019$, so strong evidence against H_0.

Requirements of the chi-squared test

1. Cell contents must be counts.
2. Categories must not overlap.
3. All expected frequencies must be $\geqslant 1$.
4. At least 80% of expected frequencies must be $\geqslant 5$.

To overcome insufficient expected frequencies: combine categories.

Fisher's exact test for a 2×2 table

Survey of arts students on gender and handedness:

H_0: there is no difference between the proportions of males and females that are left handed.

	Left	Right	Total
Female	1	30	31
Male	4	12	16
Total	5	42	47

	Left	Right	Total
Female	3.3	27.7	31
Male	1.7	14.3	16
Total	5	42	47

$x^2 = 5.2757$, compared with χ_1^2 distribution gives $P = 0.022$.

```
> 1-pchisq(5.2757,1)
[1] 0.02162509
```

But two cells have expected values less than 5.

Fisher's exact test: marginal totals

	Left	Right	Total
Female			31
Male			16
Total	5	42	47

Need to find all possible tables with these marginal totals.

Fisher's exact test: marginal totals

	L	R	L	R	L	R	L	R	L	R	L	R
Female	0	31	1	30	2	29	3	28	4	27	5	26
Male	5	11	4	12	3	13	2	14	1	15	0	16
			observed									
Probability	0.0028		0.0368		0.1698		0.3516		0.3282		0.1108	

P-value $= 0.0368 + 0.0028 = 0.040$

Here we have added all probabilities less than or equal to the probability for the observed data.

Fisher's exact test

Where do the probabilities come from? Hypergeometric distribution.

Suppose we have a bin with N balls of which K are good and $N-K$ are bad. If we take a random sample of size n from the N balls without replacement, the probability that exactly k of them are good is:

$$\frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}.$$

In the above example, we have 47 people, of whom 31 are female. If we choose a random sample (without replacement) of size 5, what is the probability that exactly 1 of them is female? This is

$$\frac{\binom{31}{1}\binom{16}{4}}{\binom{47}{5}} \approx 0.0368.$$

Hypothesis tests: some cautionary remarks

- Testing at the 5% level is standard, and extremely common, but is still essentially arbitrary.
- Accepting a (null) hypothesis does not mean you've proved it's true; rather, the data do not give sufficient evidence to refute it.
- A non-significant test can occur simply as a result of not having enough data to detect a significant difference.
- Statistical significance is a different issue from practical significance.