ECS7024 Statistics for Artificial Intelligence and Data Science

Topic 14: Hypothesis Testing using χ^2

William Marsh

Outline

- Aim: understand how to compare proportions using Chi-squared (χ^2)

- Recap
  - Principle of hypothesis testing
- Testing difference between proportions
  - χ^2 distribution
  - Degrees of freedom
- An experiment with p-values

Recap: Principles of Hypothesis Testing

Inferential Statistics

- We have a sample
- We calculate a statistic from the sample
- What do we reliably know about the population?

Inferential Statistics

- We have a sample
- We calculate a statistic from the sample
- What do we reliably know about the population?

- *We assume that the sample is unbiased*
- *Is the sample statistic an unbiased estimate?*
  - *Take account of degrees of freedom*
- *We know that there is a potential 'error' in the sample statistic*
  - *Given the error, decide what we know*

Sample Statistics & Distribution

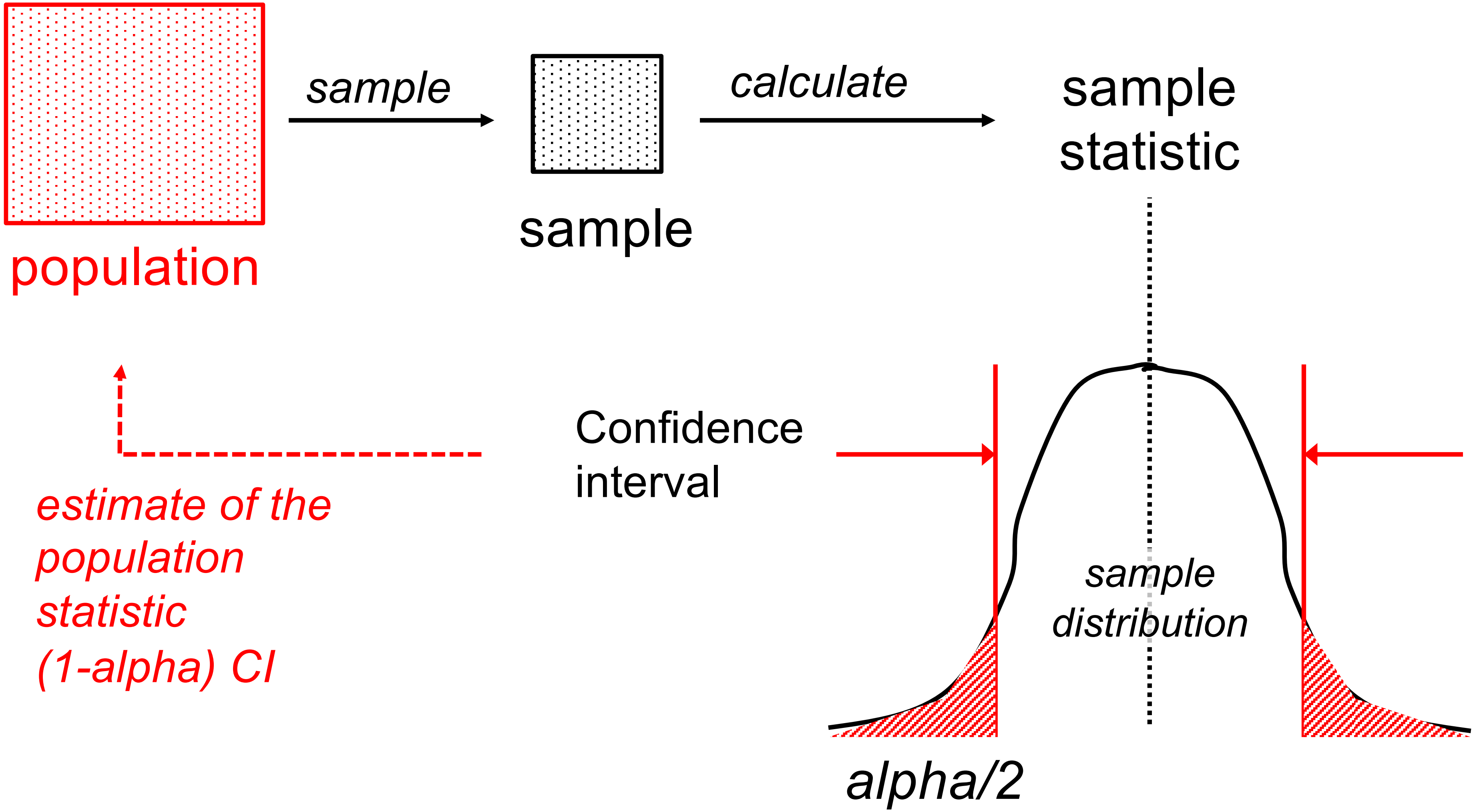
Sample Statistic

- The sample mean (when variance known)
- The sample mean when variance estimated

Sample Distribution

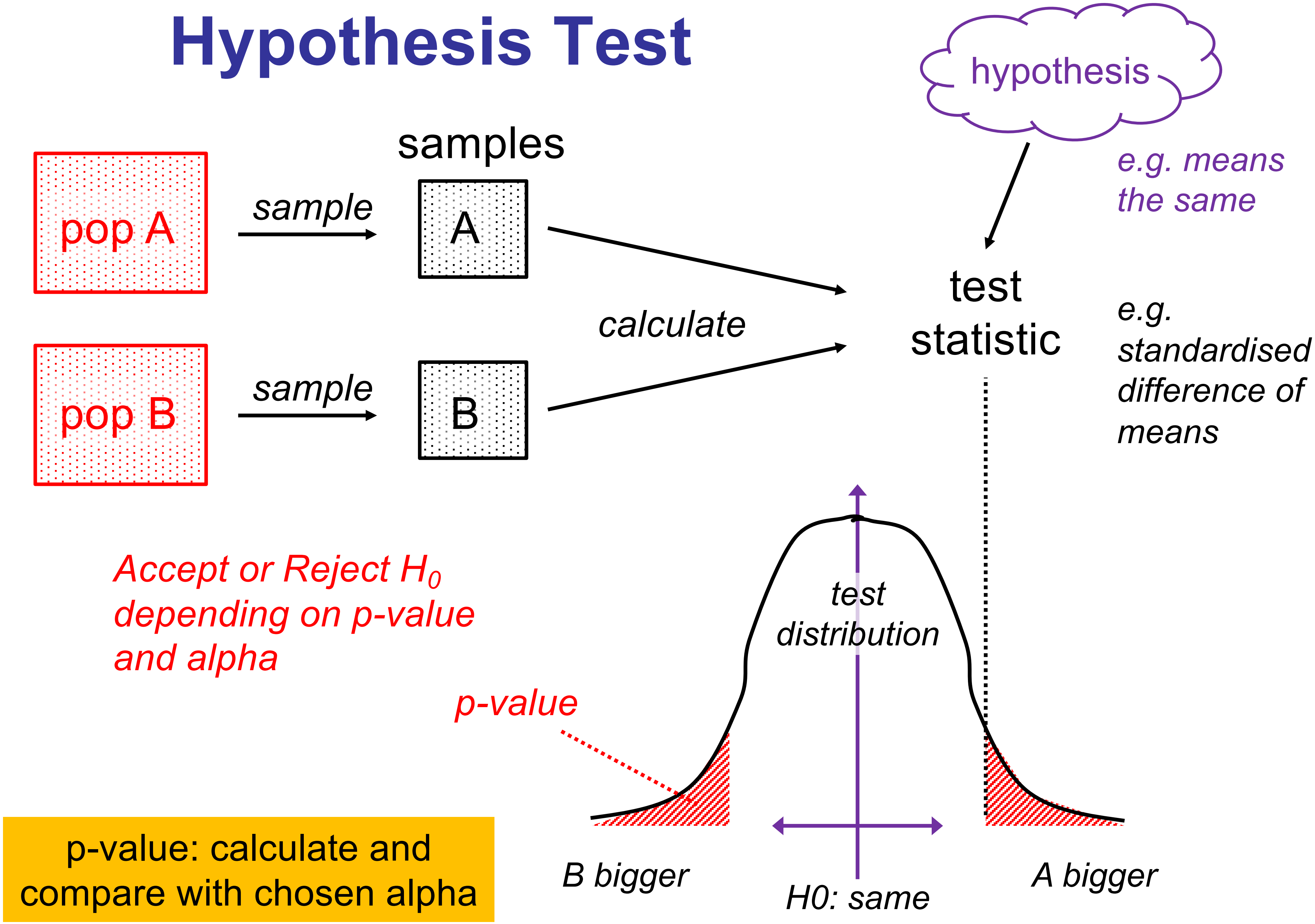
- Normal
- Student’s

Confidence Intervals



alpha is the significance threshold – choose it

Hypothesis Test



Some Issues

- You have to know
  - The test statistics
  - The correct distribution
  - The assumptions

- CIs and p-value can be mis-understood
  - p-value is not the probability you want

- Hypothesis testing does not consider effect size

Testing Proportions in a Contingency Table

Test statistics

New distribution - χ^2

Problem We Are Solving

- Contingency table
  - Type of jobs
  - City region
  - Sample

	A	B	C	D	Total
White collar	90	60	104	95	349
Blue collar	30	50	51	20	151
No collar	30	40	45	35	150
Total	**150**	**150**	**200**	**150**	650

  - Question: is the distribution of jobs the same in each region?
    - Null hypothesis: it is the same

Assuming Equal Proportions

	A	B	C	D	Total
White collar	90	60	104	95	349
Blue collar	30	50	51	20	151
No collar	30	40	45	35	150
Total	**150**	**150**	**200**	**150**	650

- Overall proportions
  - White collar: 349 / 650
  - Blue collar: 151 / 650
  - No collar: 150 / 650
- What if we assume each region has these proportions?

Assuming Equal Proportions

	A	B	C	D	Total
White collar	90	60	104	95	349
Blue collar	30	50	51	20	151
No collar	30	40	45	35	150
Total	**150**	**150**	**200**	**150**	650

349 / 650
151 / 650
150 / 650

- Region A – 150 people
- Expect
  - 150 x 349 / 650 white collar
  - 150 x 151 / 650 ...
  - ....

	A	B	C	D
White collar	80.5	80.5	107.4	80.5
Blue collar	34.8	34.8	46.5	34.8
No collar	34.6	34.6	46.2	34.6

Test Statistic

- Observed

	A	B	C	D
White collar	90	60	104	95
Blue collar	30	50	51	20
No collar	30	40	45	35

- Expected
  - Assuming null hypothesis

	A	B	C	D
White collar	80.5	80.5	107.4	80.5
Blue collar	34.8	34.8	46.5	34.8
No collar	34.6	34.6	46.2	34.6

$$\sum_{All\ cells} \frac{(Observed\ - Expected)^2}{Expected}$$

Test Statistic

- Observed

	A	B	C	D
White collar	90	60	104	95
Blue collar	30	50	51	20
No collar	30	40	45	35

- Expected
  - Assuming null hypothesis

	A	B	C	D
White collar	80.5	80.5	107.4	80.5
Blue collar	34.8	34.8	46.5	34.8
No collar	34.6	34.6	46.2	34.6

(Observed – Expected)^2 / Expected

$\sum$

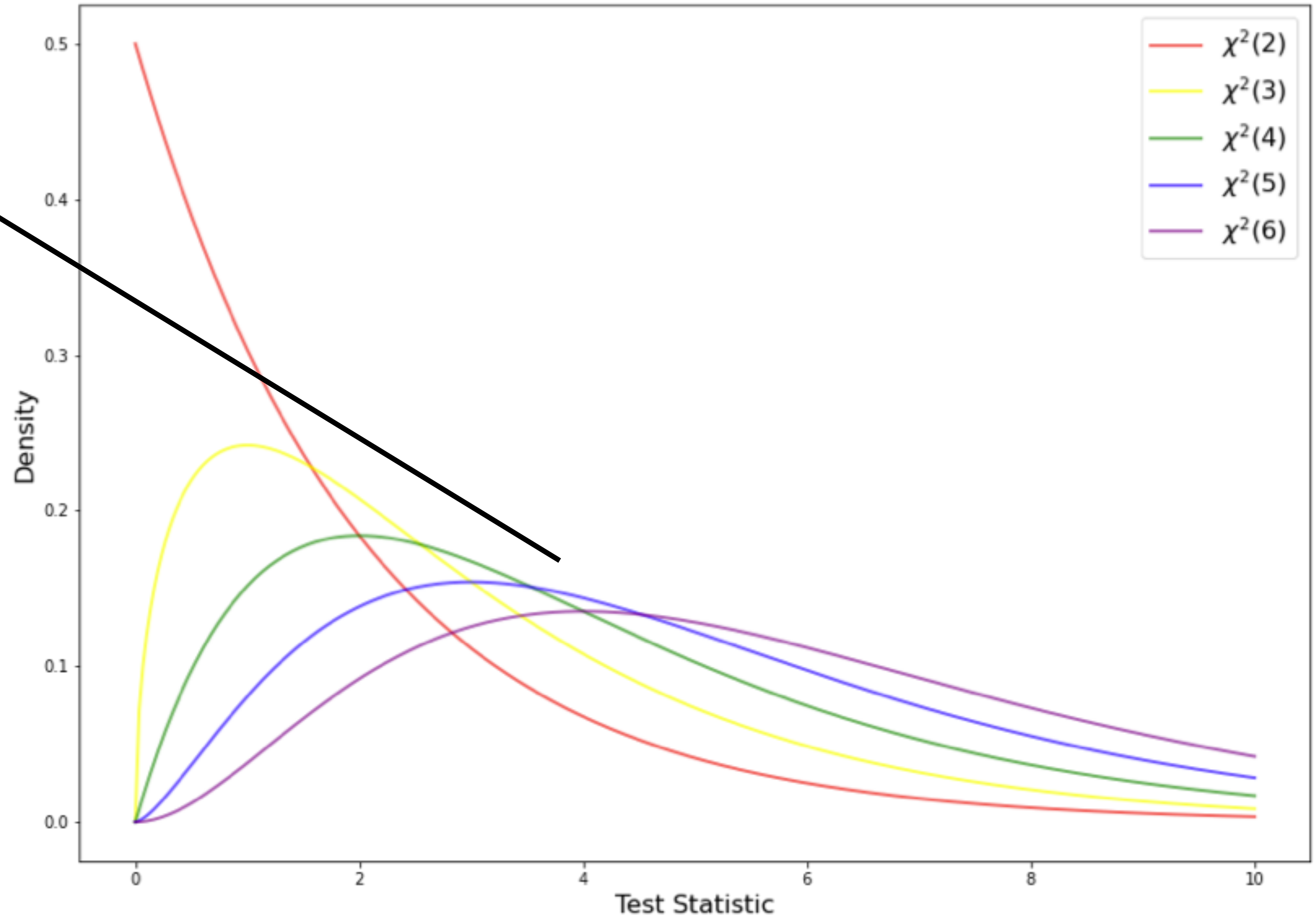
	A	B	C	D
White collar	1.11	5.24	0.11	2.60
Blue collar	0.67	6.59	0.44	6.33
No collar	0.62	0.84	0.03	0.00

= 24.57

Chi-Squared Distribution

- Parameter: degrees of freedom
  - (number of rows – 1) * (number of columns – 1)

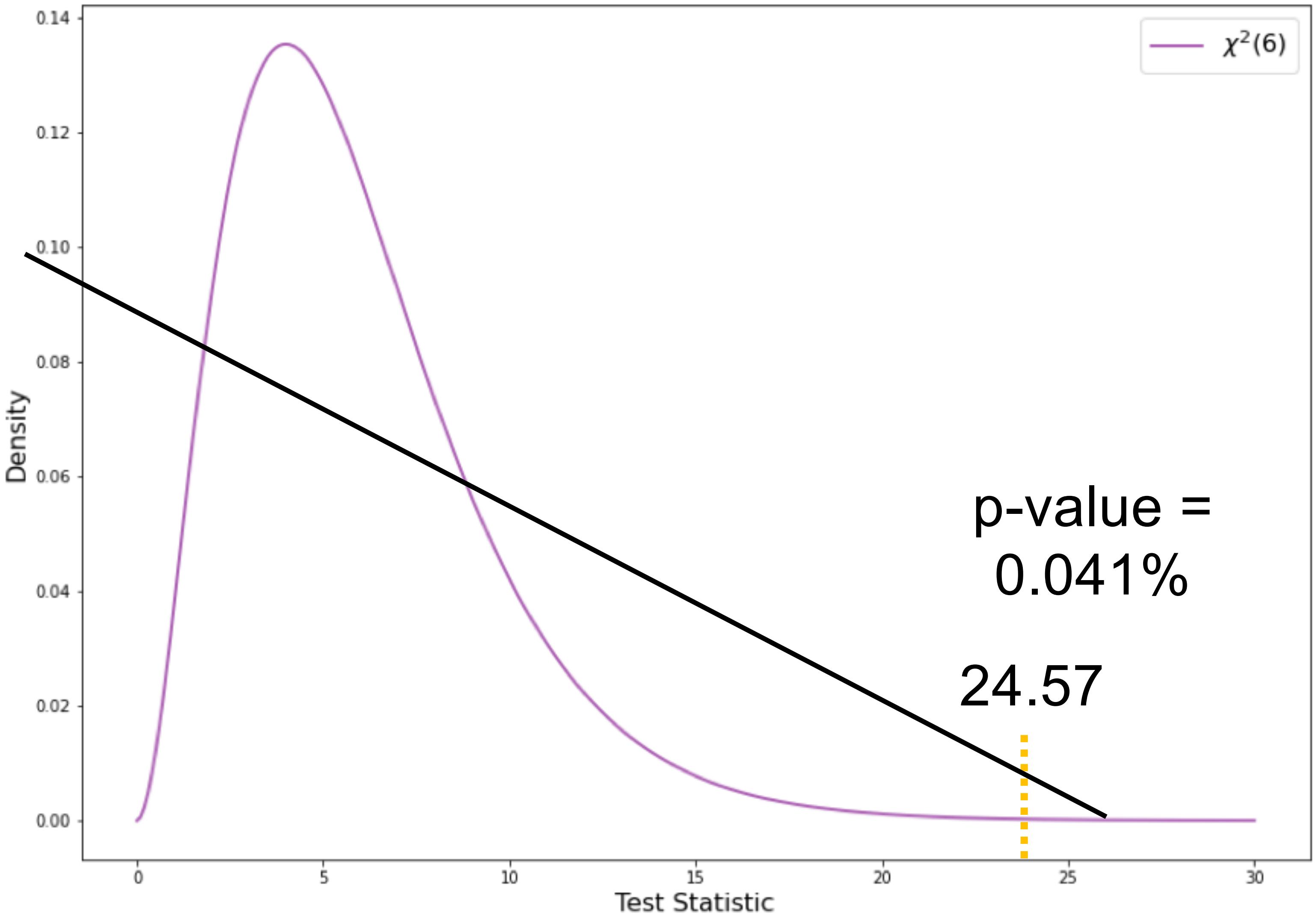
More symmetric as dof increases



Our Test has 6 Degrees of Freedom

p-value is area under curve beyond the test statistics.

It is obtained from the CDF (rather than PDF shown here)



Issues and Limitations

- We have described Pearson's χ^2 test
- It's an approximation
  - Acceptable when numbers in most cells >= 5
- Many alternatives
  - E.g. different test statistic: G-test

	A	B	C	D
White collar	80.5	80.5	107.4	80.5
Blue collar	34.8	34.8	46.5	34.8
No collar	34.6	34.6	46.2	34.6

Expected values not possible!!

Quiz 1

Initial Review of C/W 1

Some Areas for Improvement I

Commenting on descriptive statistics tables and plots such as histograms (part 2):

- ***Less good answers***: tend to describe what we can already see in the table or graphs, repeating numerical values, telling us what the table or graph is
- ***Better answers***: try to understand the situation that results in these tables and graphs instead of describing what's seen in the table or graph

Examples

- Which is better?

Most tube stations have a similar amount of exits in the morning but a small number of stations have a much larger percentage of their overall exits in the morning. The most intuitive explanation for this is that stations with a high percentage of overall exits in the morning are located near to areas of work.

AM Peak Proportion has a mean of 0.21 and a median of 0.17. That means the average AM Peak Proportion is 0.21, but half of the stations have a proportion of 0.17 or under.

PM Peak Proportion has a mean of 0.30 and a median of 0.31. That means the average PM Peak Proportion is 0.30, and half of the stations have a proportion of 0.31 or under.

Some Areas for Improvement II

- Title Formatting
  - Inconsistent numbering of headings
  - No table of contents

- Classification thresholds (part 3):
  - ***Less good***: thresholds without justification
  - ***Better***: thresholds selected based on the descriptive statistics of the parameter used (difference or ratio), e.g. first and third quartiles

Example: Headings

```
### Content
### **Section 1: Creating a relevant dataframe with selected vari
#### Section 1.A. Finding AM & PM hour exit counts for each stati
#### Section 1.B. Calculating AM & PM proportions
#### Section 1.C. Describing the newly created dataframe and its
### **Section 2: Plotting and Analysing Distributions**
#### Section 2.A. Histogram of AM_proportion and Statistics
#### Section 2.B. Histogram of PM_proportion and Statistics
#### Section 2.C. Assumptions about stations
### Section 3: A Simple Classification of Stations
#### Section 3.A. Categories for classifying stations
#### Section 3.B. Justifying the thresholds of categories
#### Section 3.C. Reporting classification on a set of Northern L
```

Example: Thresholds

Those stations that have a higher AM_proportion than the 75% Quartile (=0.26) and those that have a lower PM_proportion than the 25% Quartile (=0.24). However, the former was rounded up to 0.3 (due to skewed distribution; more on this under 'Residential' category') and round down the latter to 0.2 (more on this under 'Residential' category.)

(Assumed ratio > 1.1)

(Assumed ratio < 0.9)

(Assumed ratio between 0.9 and 1.1)

An Experiment with p-values

Simulated a Fair Dice

- Create many samples with 12 rolls

Test Num	Counts						
	1	2	3	4	5	6	
1	4	1	1	2	1	3	= 12
2	2	2	2	1	3	2	= 12
3	1	2	2	3	2	2	= 12
4	2	1	4	2	1	2	= 12
5	1	2	0	3	2	4	= 12

- Null hypothesis “dice fair”
- Test, with alpha 1%
- As expected, 1% of tests reject null hypothesis

Question?

- What happens if we increase the number of rolls from 12 to 120 or 1200?

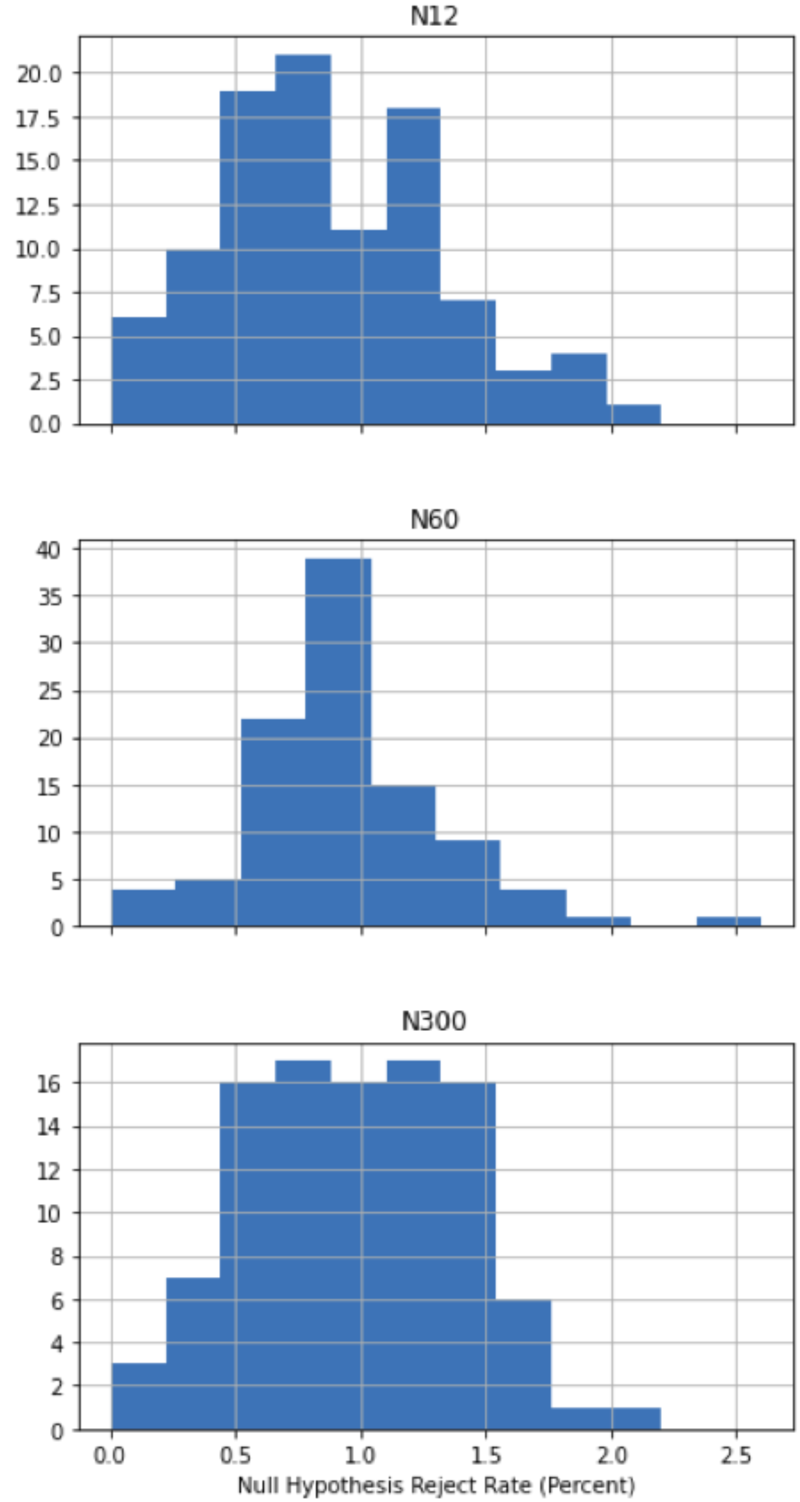
Test Num	Observed Frequency					
	1	2	3	4	5	6
1	33.3%	8.3%	8.3%	16.7%	8.3%	25.0%
2	16.7%	16.7%	16.7%	8.3%	25.0%	16.7%
3	8.3%	16.7%	16.7%	25.0%	16.7%	16.7%
4	16.7%	8.3%	33.3%	16.7%	8.3%	16.7%
5	8.3%	16.7%	0.0%	25.0%	16.7%	33.3%

- We expect the observed frequencies to be closer to 16.66%
- What about the reject rate?

Answer: Reject Rate Same

- Distribution of the reject rate
  - Different numbers of rolls

- *This what we should expect!*



See Notebook

Discussion

- A statistically significant difference is not (necessarily) a significant (*i.e. large, important*) difference
- Hypothesis testing does not consider 'effect size'
- As the sample size grows
  - The sample variance reduces
  - I.e. the sample distribution becomes narrower
  - Small difference (from the null hypothesis) become 'statistically significant'
- If you use α = 1%, then the null hypothesis <u>should be</u> incorrectly rejected 1% of the the time

Danger of misinterpretation, especially with big data

Summary

- Use chi-squared test for proportions in a contingency table
  - Other uses as well
- Test statistics depends on differences between observed and expected (assuming uniformity)
- Hypothesis testing does not consider effect size