FEEG6017 lecture: The chi-squared test and distribution

bjn1c13@ecs.soton.ac.uk

Assumption of normal error term

- In all of the techniques we've covered so far, there has been a background assumption that the noise in the data generating process was normally distributed.
- Taking ANOVA as an example, we assume that each group has a characteristic mean value and that a normal random variate is added to that mean to get the actual data point.

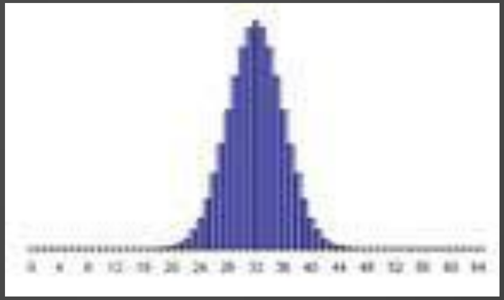
Normality assumption in ANOVA

Group A = 35

Group B = 26

Group C = 41

+



Random noise

=

Actual scores

Note the distribution

Normality assumption: Regression

Intercept

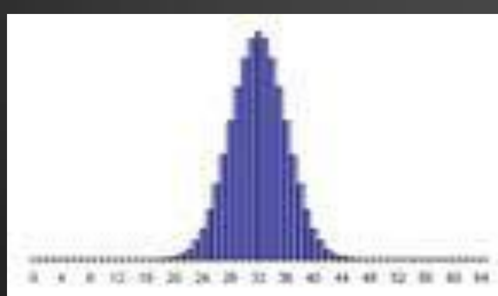
+

Slope

X

This man's height

+



Random noise

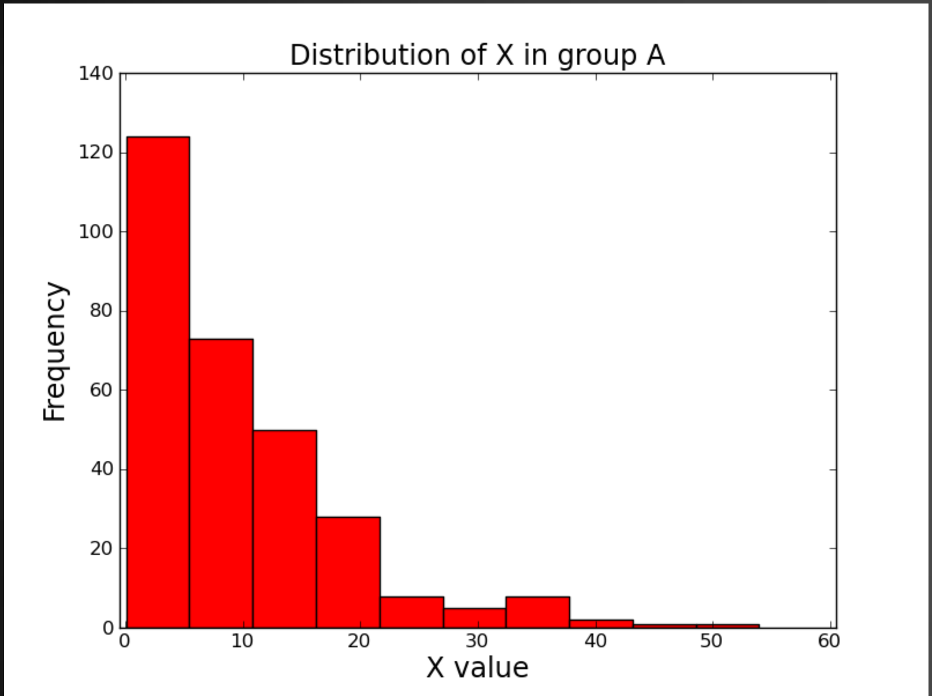
=

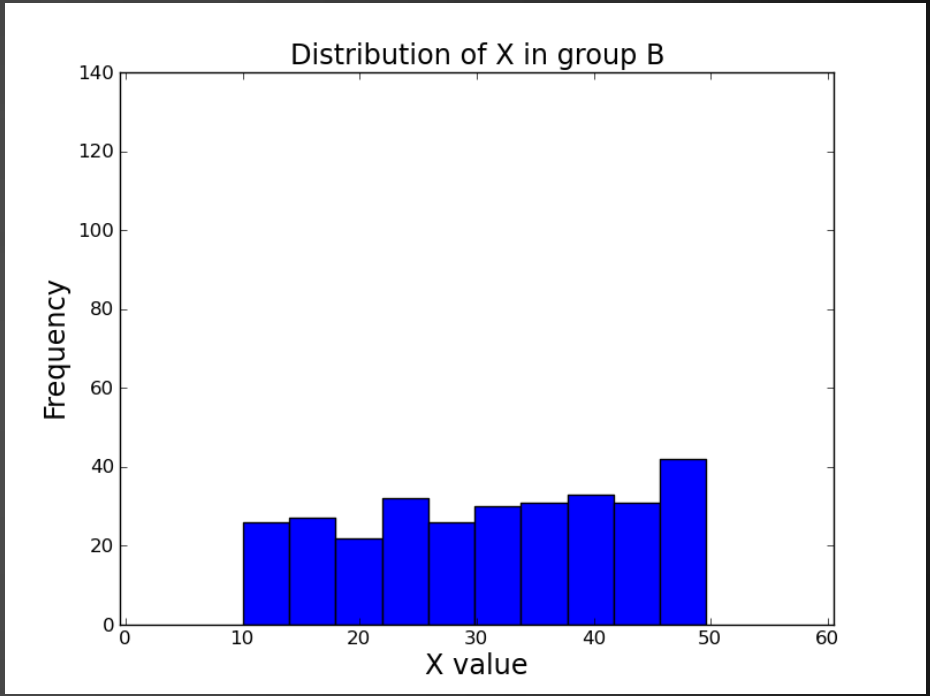
His basketball ability score

A reasonable assumption?

- Often the normal-error-term assumption is a safe one: the truth is close enough that our inferential logic works well.
- But sometimes it is obvious that the data generating process could not have been some systematic effect plus a normal error term.

Distribution of X across 2 groups





- Values in group A and B are clearly being generated by two very different -- and non-normal -- processes.

Distribution of X across 2 groups

- If we were trying to test for a difference between the means of samples A and B, we can't really trust the answer we'd get from a 2-sample t-test.
- We need some statistical tests that don't make the normally-distributed-errors assumption.

Non-parametric statistics

- The term "non-parametric" refers to statistical tools that do not assume normality in the error term of the generation process.
- There are non-parametric equivalents for most of the techniques we've discussed.
- Why not use them all the time? The mathematics are trickier, and we lose some power by dropping the normality assumption.

The chi-squared test

- This is a versatile non-parametric test.
- **Goodness of Fit**: whether a sample fits an expected distribution (of arbitrary shape).
- **Test for Independence**: are paired observations on two categorical variables independent.
  - Variable X: Gender [Male,Female]
  - Variable Y: Voting Preference [Lab,Con,Lib]
  - Does gender affect voting preference?

Chi-squared test: Goodness of Fit

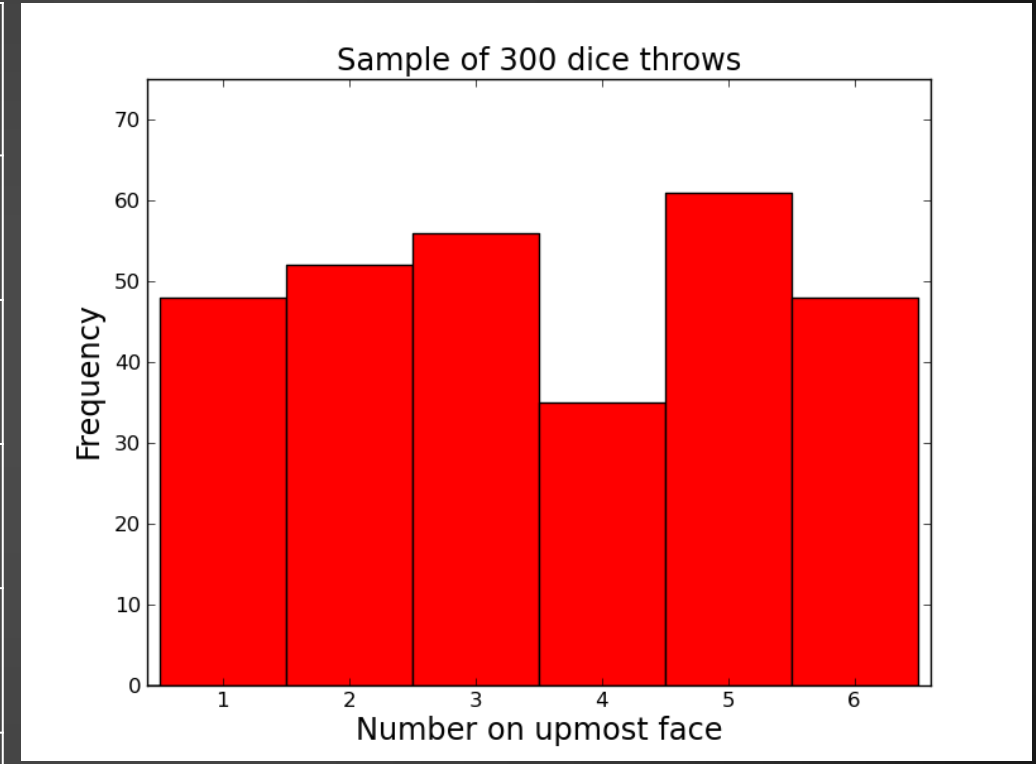
- Consider throwing a die many times to see whether or not it's fair.
- We get a frequency distribution.
- We expect this to be a uniform distribution: an equal number of 1s, 2s, 3s, etc.

Chi-squared test: Goodness of Fit

- Most of the time we won't get a *perfectly* uniform distribution though.
- How far away from uniform would the results have to be before we would suspect the die was not fair?
- How might we measure deviation from the expected values?

Sample of 300 throws

1	48
2	52
3	56
4$_4$	35
5	61
6	48



Difference from what was expected?

	Obs.	Exp.	O - E	(O-E)^2
1	48	50	-2	4
2	52	50	2	4
3	56	50	6	36
4	35	50	-15	225
5	61	50	11	121
6	48	50	-2	4

Difference from what was expected?

- We're "expecting" to see each number 50 times.
- We can tally up how far from the expected value each observation is.
- We then try squaring those values to deal with negative numbers.
- We're on our way to some measure of deviation from the expected values.

Scaling our measure

- The particular size of these numbers depends on our sample size: if we'd thrown the dice 30 times or 3000 times we'd have different numbers.
- We scale the "observed minus expected, squared" term somewhat by dividing through by the expected value.

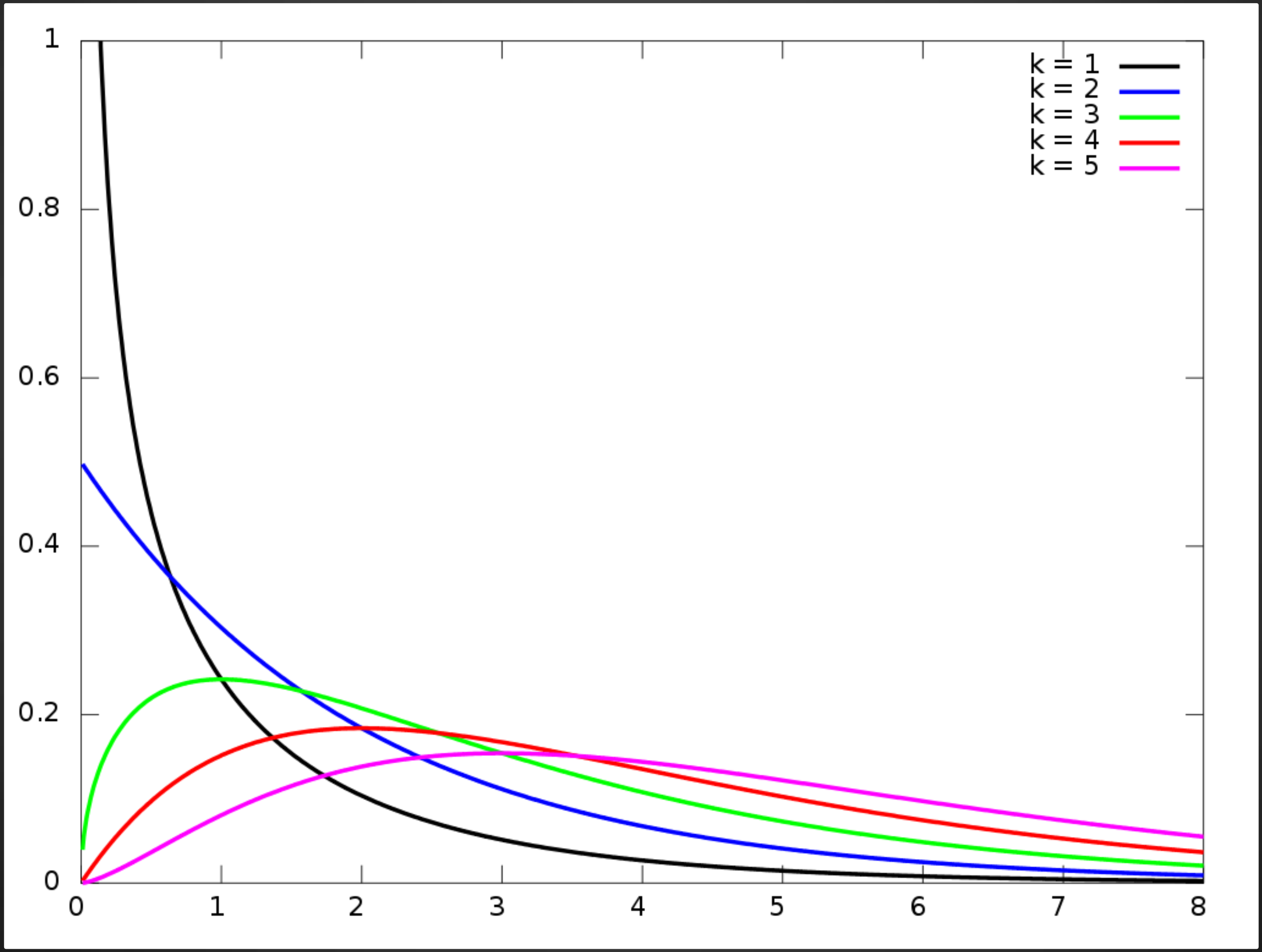
Scaling our measure

	Obs.	(O-E)^2	(O-E)^2 / E
1	48	4	0.08
2	52	4	0.08
3	56	36	0.72
4	35	225	4.5
5	61	121	2.42
6	48	4	0.08

The chi-squared test

- The final step is to ***sum*** the (O-E)^2 / E terms.
- The total value is the chi-squared statistic. In our example this number is 7.88.
- But what distribution should we expect this to have? How do we translate this number to a p-value?
- Chi-squared distribution… (hint is in the name)

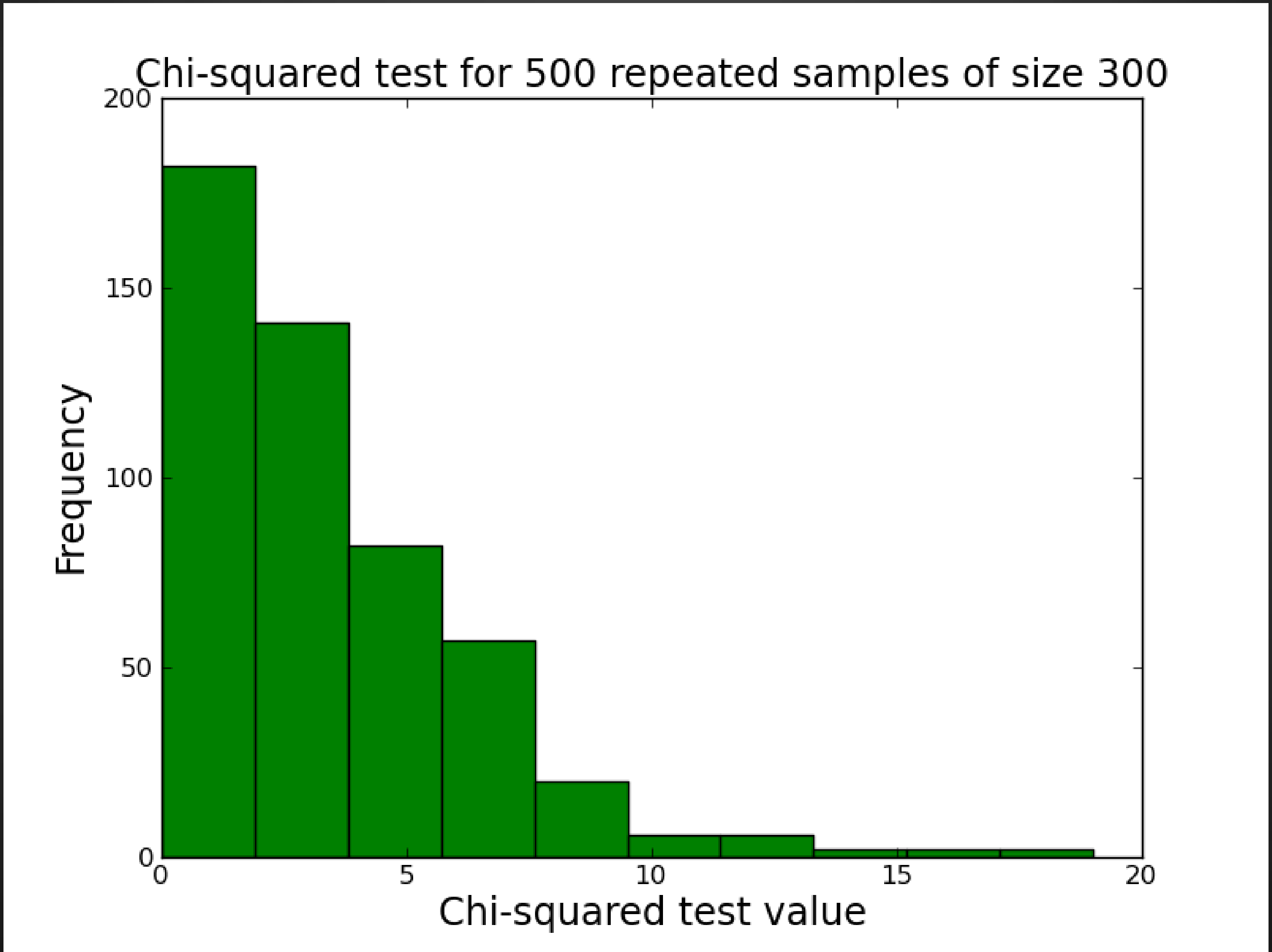
The chi-squared distribution



Repeated sampling experiment

- Let's generate 500 samples of 300 die throws, and calculate the chi-squared statistic each time.
- That should give us an idea of how often extreme deviations from uniformity (and thus large values on the chi-squared statistic) come up by chance.

Repeated sampling experiment



Repeated sampling experiment

- In practice, a p-value of 0.05 (i.e., the top-most 5% of the distribution) equates to a chi-squared statistic of around 10.
- The exact value is 11.07.
- For p = 0.01, the critical chi-squared value is 15.09.
- Our observed value of 7.88 is not unusual.

How is the chi-squared statistic distributed?

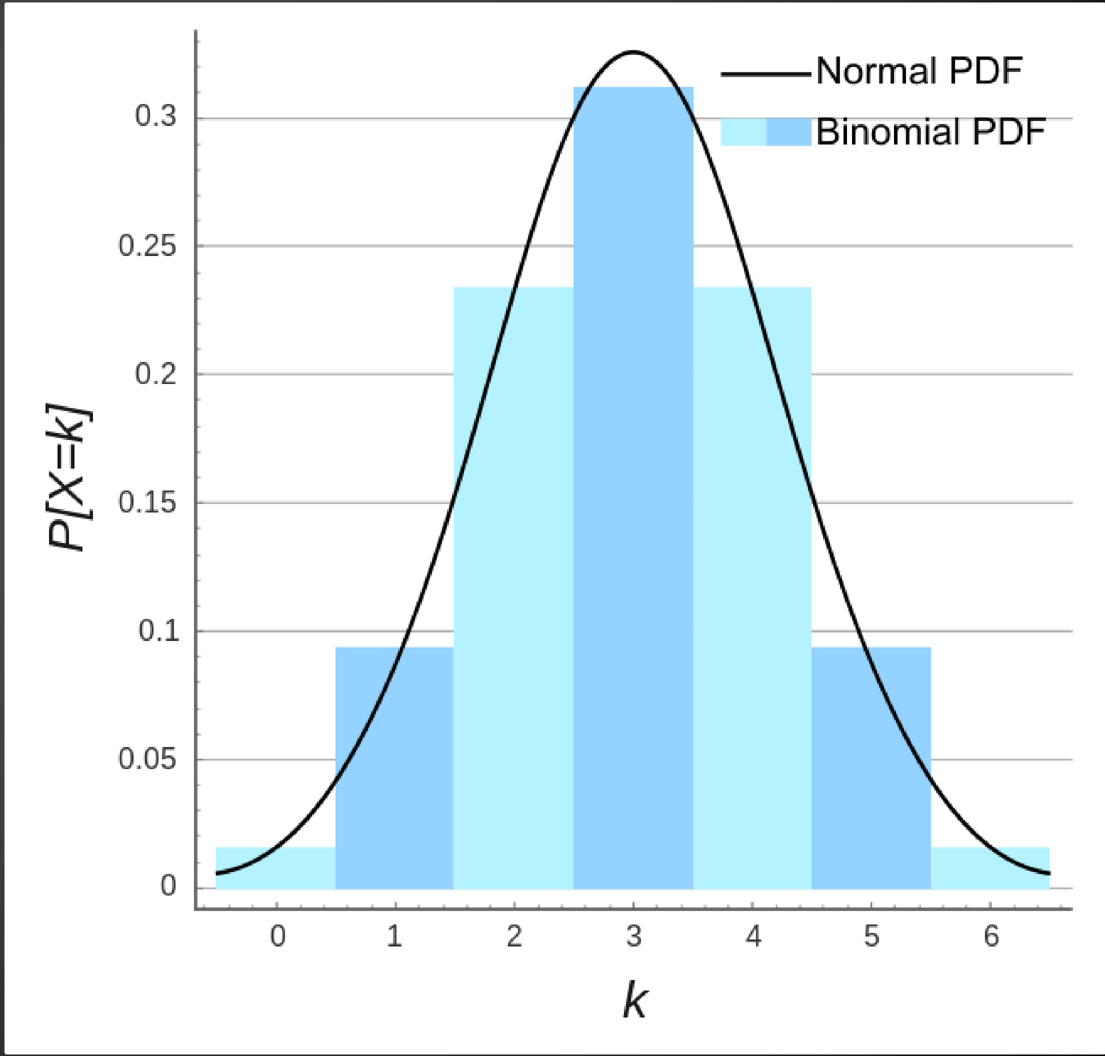
- Imagine that we were doing the same kind of test not with die-rolling but with coin tossing.
- If we toss a fair coin 100 times, clearly the expected values for heads and tails are 50 each.
- In reality we'll get combinations like 48/52, 53/47, 45/55, etc.
- What's the probability distribution of getting different numbers of heads?

The binomial distribution

- The binomial distribution describes the number of heads we expect to get from throwing a coin multiple times.
- The binomial describes the outcome of multiple *Bernoulli trials*,
- Bernoulli trial is any situation where there's a probability p of one outcome and 1-p for the other.
- What does it look like?

Binomial ≈ normal distribution

- n=6
- k=#heads
- As n gets larger, the binomial approximates the normal distribution very closely.



From binomial to chi-squared

- So if the number of heads we get in multiple coin tosses is binomially distributed...
- And the binomial is approximately normal...
- As the number of trials (n) increases…
- Then the chi-squared procedure of squaring the differences from the expected value, then dividing by the expected value, will give us what?

Difference from what was expected?

	Obs.	Exp.	O - E	(O-E)^2
H	48	50	-2	4
T	52	50	2	4

Normally Distributed

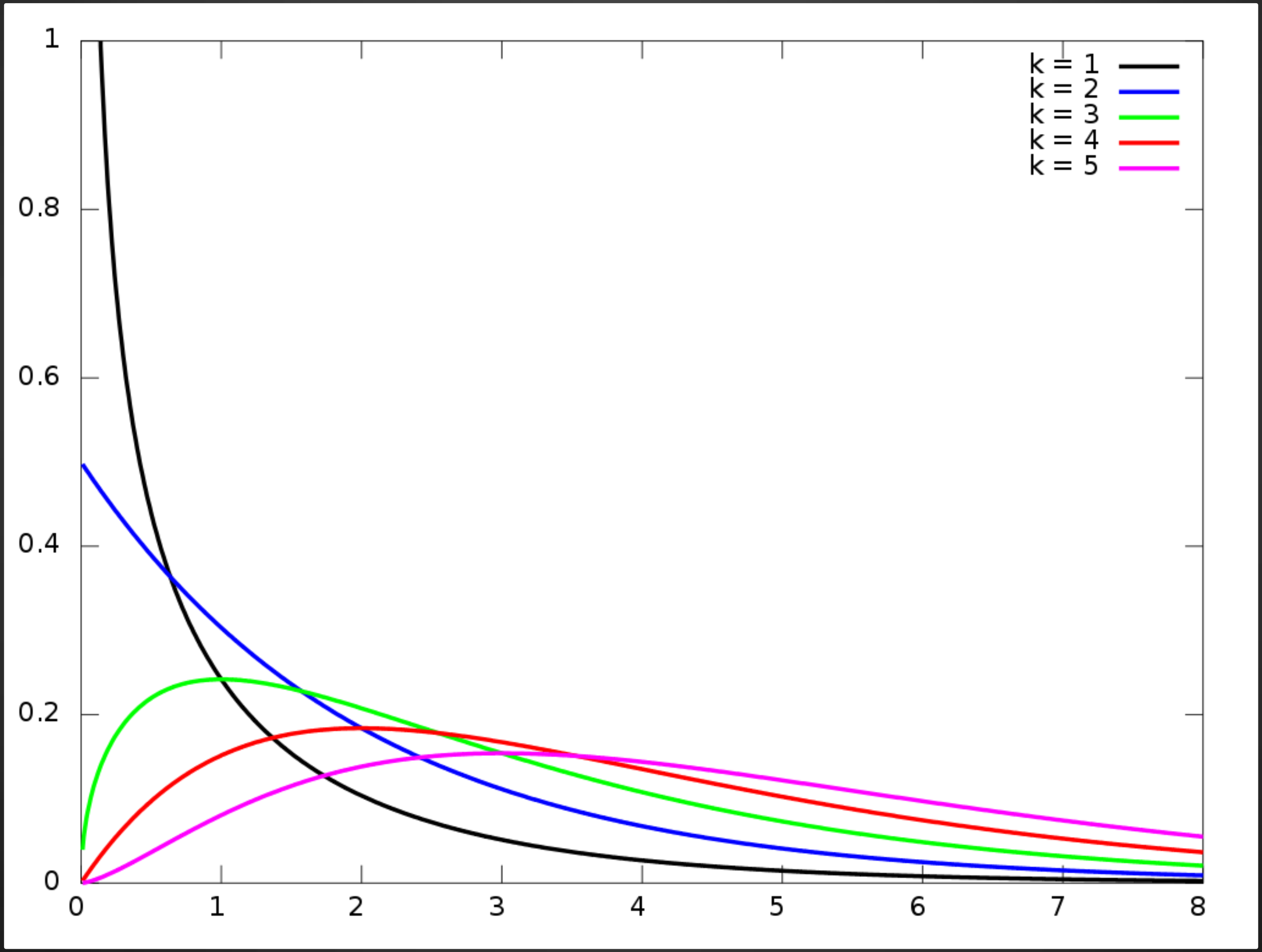
From binomial to chi-squared

- Each term in our chi-squared procedure is taking an approximately normally distributed value and squaring it.
- The chi-squared distribution is in fact the sum of K squared-standard-normal deviates (K is the degrees of freedom of the test).
- Ironic that we've returned to the normal distribution in an effort to avoid assuming it was present.

Degrees of freedom?

- If we toss a coin 100 times, it's easy to see how many degrees of freedom there are: if there are 53 heads, there must be 47 tails. Thus there's 1 degree of freedom for the chi-squared test here.
- Similarly with the die: once we know how many 1s, 2s, 3s, 4s, and 5s, the number of 6s is determined. So there are 5 degrees of freedom.

The chi-squared distribution



Another use of chi-squared tests

- We've seen how to use the test to look at whether a one-dimensional distribution fits its expected values reasonably closely.
- What about two-dimensional distributions?
- When we have pairs of observations, like our gender and voting data set.

Recap: when to use which test?

- If we have a continuous outcome measure, and a binary predictor variable, we use a two-sample t-test.
- If we have a continuous outcome measure, and one or more categorical predictor variables, we use ANOVA.
- If we have a continuous outcome measure, and a single continuous predictor variable, we use simple linear regression.

Recap: when to use which test?

- If we have a continuous outcome measure and multiple continuous and categorical predictor variables, we use multiple regression.
- If we have a binary categorical outcome measure and multiple predictor variables, we use logistic regression.

When to use a chi-squared test?

- If we have a categorical outcome measure, and one (or more) categorical predictor variables, it turns out we can use a chi-squared test.
- We're asking whether the observed incidence of outcomes for each level of the predictor variable could have happened by chance.

Voting example: Independence Test

- Suppose we're asking whether sex has any relevance to predicting the way people will vote.
- We ask 50 randomly selected men and 50 randomly selected women which party they voted for at the last election.
- Note that both variables are categorical.

Voting example

	Labour	Tory	Lib dem	
Men	16	23	11	50
Women	28	17	5	50
	44	40	16	

- Number of men and women voting for each party shown, and the marginal totals.

Voting example: expected values

	Labour	Tory	Lib dem	
Men	22	20	8	50
Women	22	20	8	50
	44	40	16	

- Using the marginal totals, we can easily calculate the expected number of counts in each cell if there were no relationship between sex and voting preference.

Voting example: chi-squared

- We now have six places to calculate "observed minus expected".
- We can use the same logic of the chi-squared test as outlined earlier.
- Sum of: (O - E) ^ 2 / E.

Voting example: O-E

	Labour	Tory	Lib dem	
Men	16-22	23-20	11-8	
Women	28-22	17-20	5-8	

Voting example: O-E

	Labour	Tory	Lib dem	
Men	-6	3	3	
Women	6	-3	-3	

Voting example: (O-E)^2

	Labour	Tory	Lib dem	
Men	36	9	9	
Women	36	9	9	

Voting example: (O-E)^2/E

	Labour	Tory	Lib dem	
Men	36/22	9/20	9/8	
Women	36/22	9/20	9/8	

Voting example: (O-E)^2/E

	Labour	Tory	Lib dem	
Men	1.64	0.45	1.125	
Women	1.64	0.45	1.125	

$$chisquared\ stat = \sum (O - E)^2 = 6.43$$

How many degrees of freedom?

- How many degrees of freedom?
- You might think 5, because we have 6 observed counts.
- But there are not 5 numbers here that are free to vary if we are to hit the marginal totals.
- DF = num columns minus one, times num rows minus one.
- DF=(c-1)(r-1)
- In this case: 2 x 1 = 2.

Voting example

- The calculation for our voting example works out at a chi-squared value of 6.42.
- The critical value for a chi-squared distribution with 2 DF, $p = 0.05$, is 5.99.
- Our test statistic is larger than the critical value, so we reject the null hypothesis of no link between sex and voting (assuming we're happy with the $p = 0.05$ threshold).

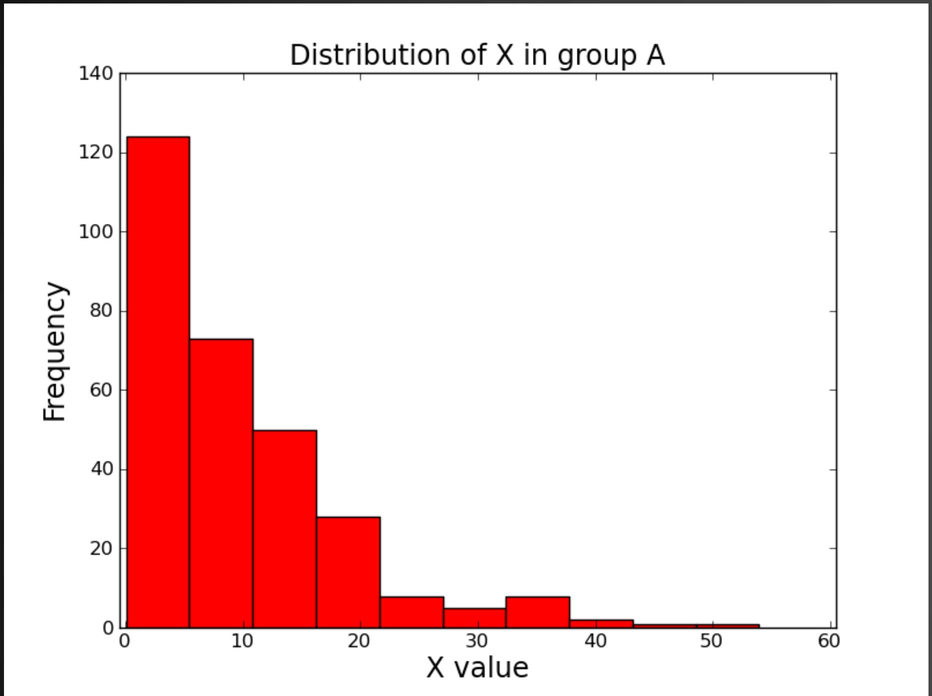
How to do this in R

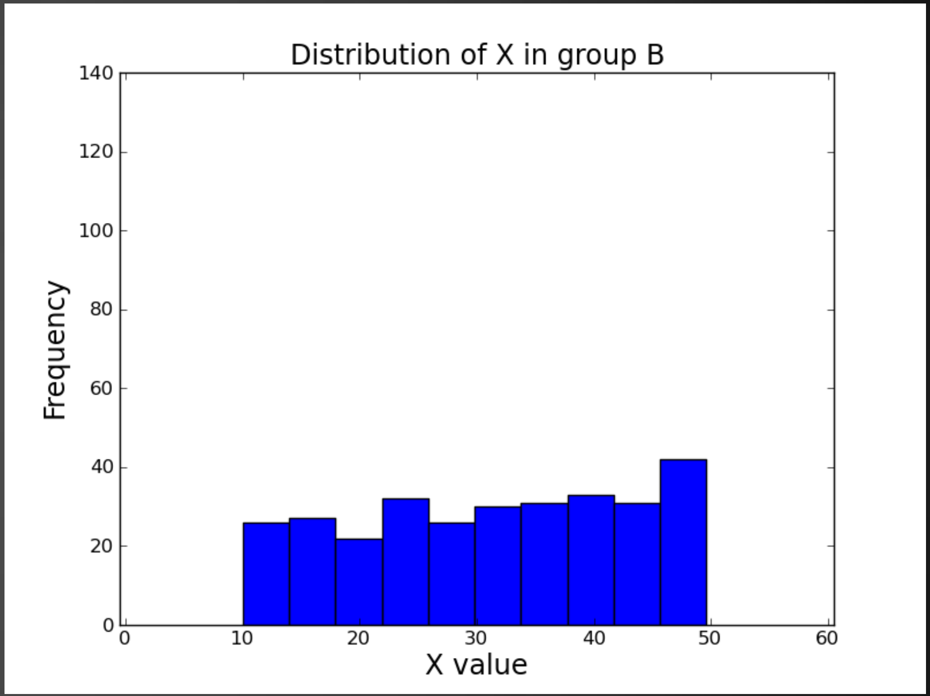
- In R the easiest way to run a chi-squared test is to first produce a table.
- For the example:`chisq.test(table(Sex,Voting))`
- The output includes the test statistic, the degrees of freedom, and the associated p-value.

Further non-parametric tests

- No time to cover them all in this lecture, but there are many tests that do not assume a normally distributed error term in the data generation process.
- The most useful is the rank-sum test (also known as the Wilcoxon rank-sum test and the Mann-Whitney U test).

Further non-parametric tests





- We would use the rank-sum test to investigate whether these two samples come from distributions with the same mean.

Further non-parametric tests

- It's based on the logic that if distribution A has a higher mean than distribution B, then values from sample A should be in the top half of the joint sample more often than those from sample B.
- Used in the same situations as a two-sample t-test.
- `wilcox.test(AValues,BValues)`

Additional material

- There is no R script for this lecture; you just need the commands `chisq.test`, `table`, and `wilcox`.
- Here is the Python program for generating the histograms and sampling experiment.