Tests for Categorial Data

Vera Demberg

Saarland University

WS 2021/2022

Hypothesis tests

Topic today: hypothesis tests for nominal scale variables (*categorial data*).

Can you think of an example for data in nominal scale?

Table of Contents

How likely that we'll get exactly 9 skulls when throwing these 20 dice?

(pic from Pinterest)

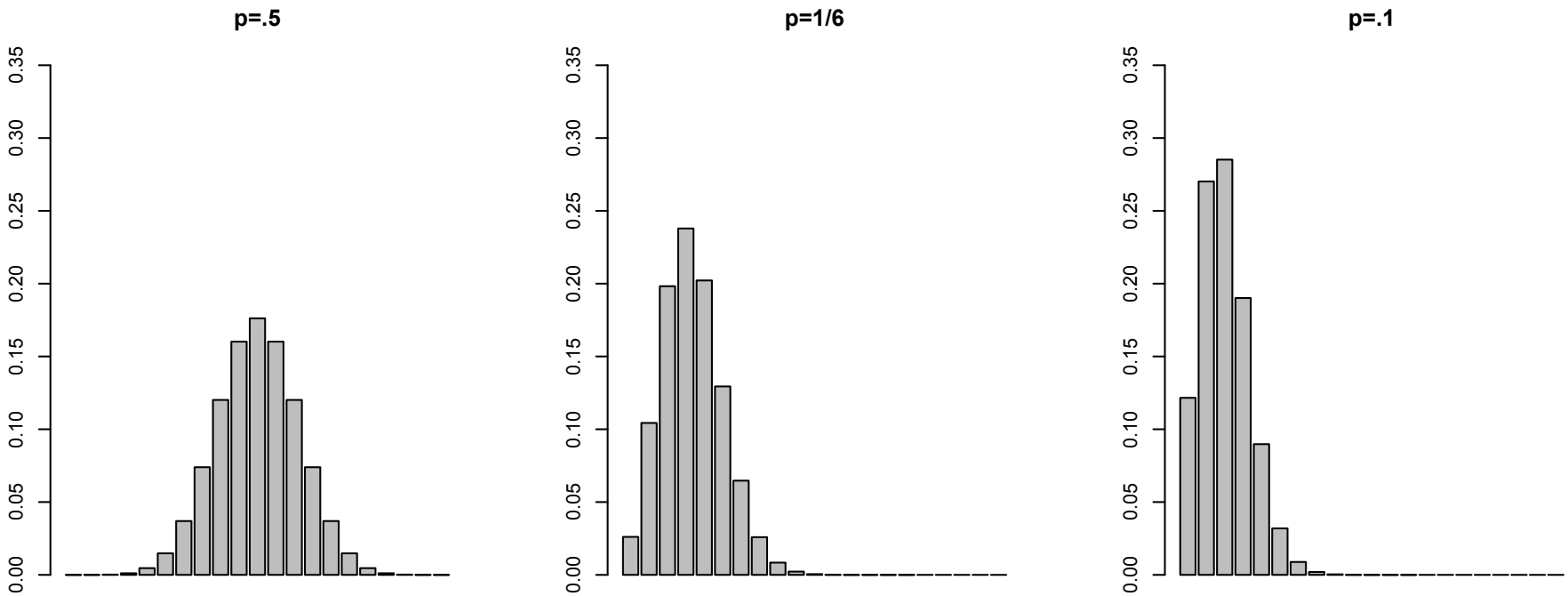
Skull dice example

In throwing these dice, we're only interested in two possible outcomes (skull or not skull), per die.

Terminology: "Bernouilli trial"

= event with two mutually exclusive outcomes

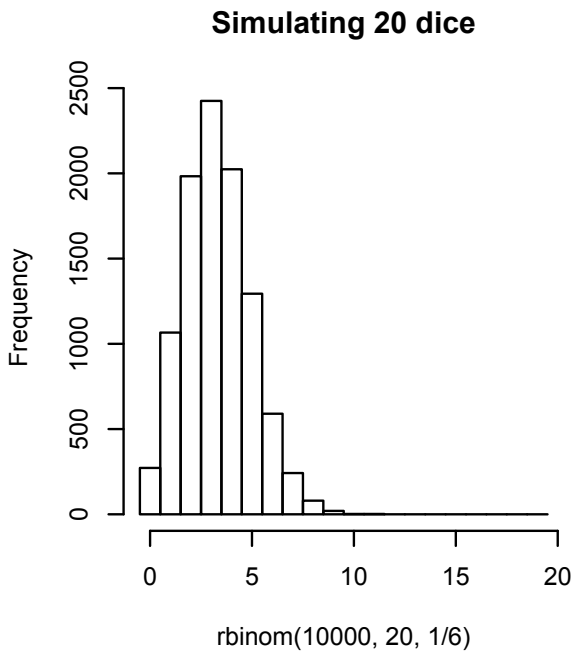
This can be described by the binomial distribution:

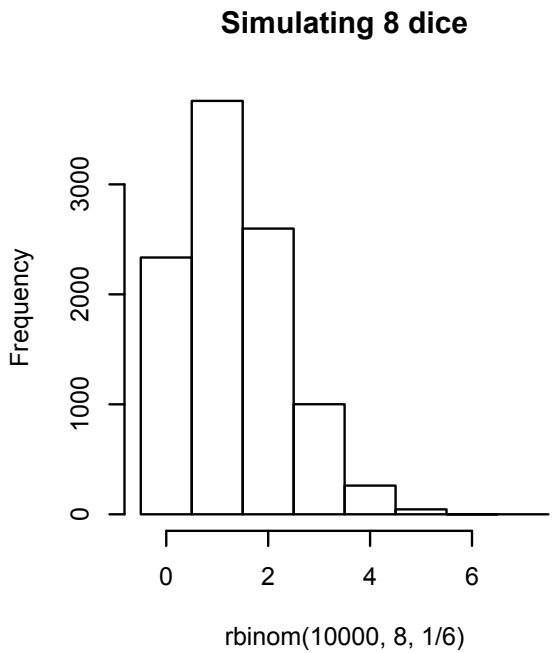


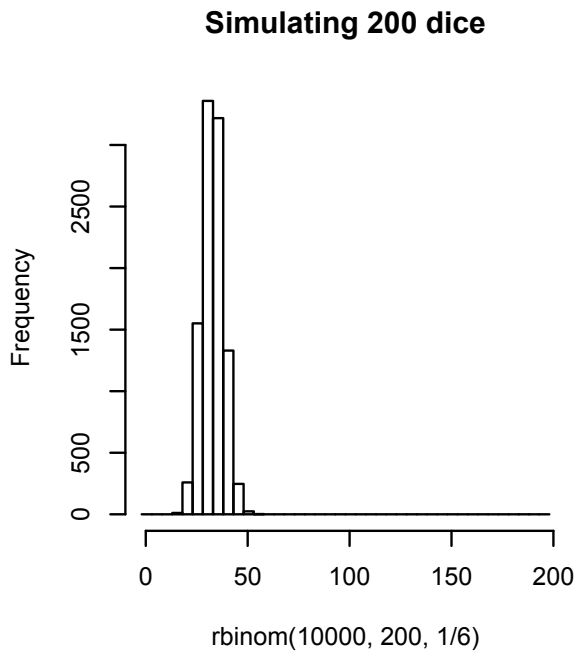
Binomial distribution

The binomial distribution is:

- discrete (you can get 4 skulls, but not 4.5 skulls)
- trials are independent of one another
- parameters
  - probability of one of the outcomes (here: likelihood of skull $\frac{1}{6}$)
  - number of trials (here: number of dice 20)







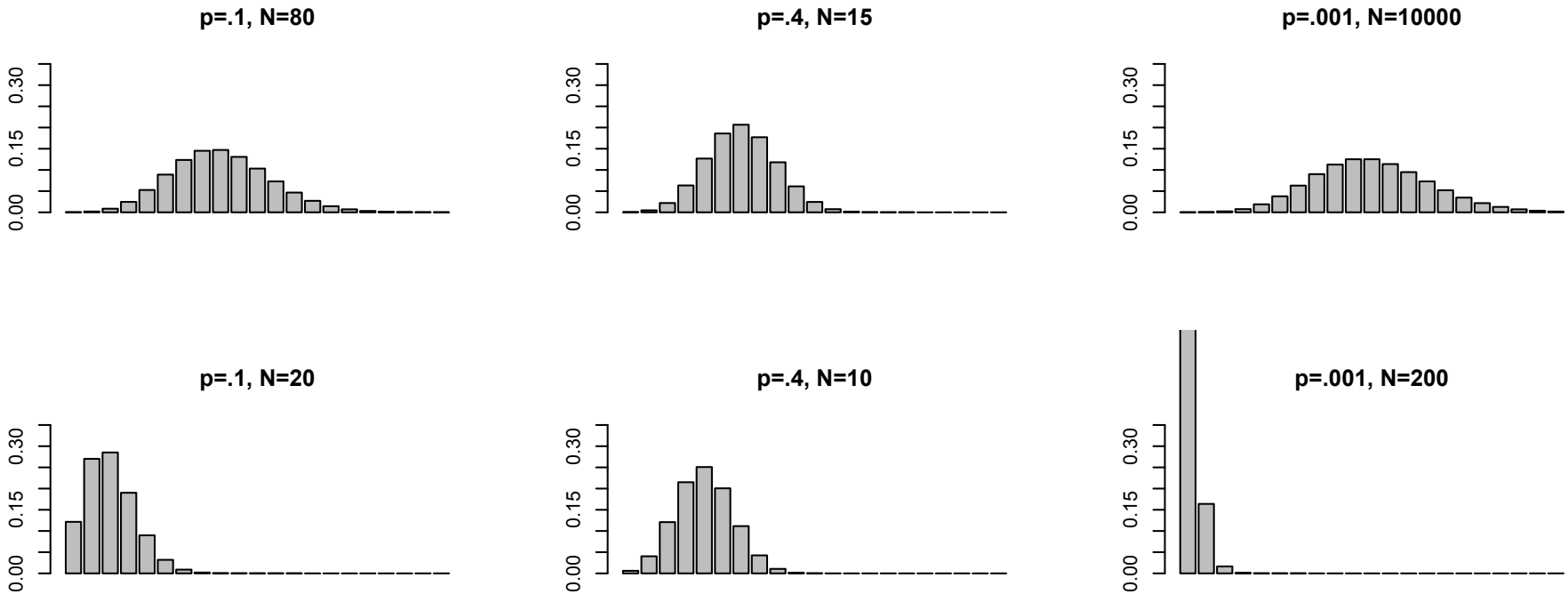
The binomial distribution

The binomial distribution can be described as follows:

mean $= Np$ (e.g. $20 \times \frac{1}{6}$)
variance $= Npq$ (e.g., $20 \times \frac{1}{6} \times \frac{5}{6}$)
standard deviation$= \sqrt{Npq}$ (e.g. $\sqrt{20 \times \frac{1}{6} \times \frac{5}{6}}$)

Binomial vs. normal distribution

The binomial distribution becomes similar to the normal distribution for larger samples.
Rule of thumb: if both Np and $Nq > 5$, it can be approximated as normal.



Using the binomial distribution for hypothesis testing

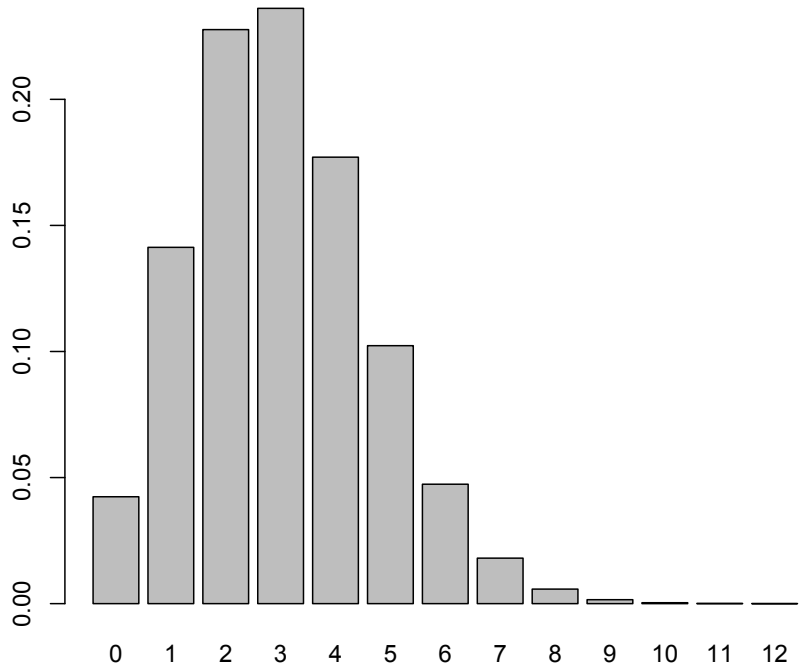
Example: perception of rapidly presented stimuli

We do an experiment where we want to test whether people can perceive a digit if it's presented to them very rapidly (just a few milliseconds). 30 digits are presented in total.
In order to do this, we present a series of digits 0-9 and ask people to press the key corresponding to the digit they think was shown.
People get it right 8 times on average[a].

[a] I made this up.

Null-hypothesis: people can't perceive the numbers. We expect performance at chance level: p=1/10.
Need to calculate probability that they manage to get 8 or more trials right by chance.

Need to calculate probability that they manage to get 8 or more trials right by chance.



pbinom(8, 30, .1, lower.tail=FALSE)
p=0.002019829
We would hence conclude that we reject the null-hypothesis in this case.

1. The Binomial distribution
2. Comparing observed to expected frequencies
3. The χ^2 distribution
4. Contingency tables
5. Assumptions of the χ^2 test
   - Fisher's Exact Test
   - McNemar's test

Table of Contents

More questions about categorial data

What if we have more than 2 categories?

Example: Rock, Paper, Scissors (= Schere-Stein-Papier)

Are people good random number generators (in terms of on average generating the same number of each category[a])?

Observation: In playing this game 75 times, children produced the following symbols with the following frequencies:
Rock – 30; Paper – 21; Scissors – 24

[a] we're ignoring order here

Is this significantly different from truly random behaviour?

Null-hypothesis

Null-hypothesis:
random distribution – all three symbols come up equally often.

Symbol	Rock	Paper	Scissors
observed	30	21	24
expected	25	25	25

Is the observed significantly different from the expected?

Is the observed significantly different from the expected?

- calculate difference between observed and expected for each category.
- but we need to make sure we're not summing up positive and negative values
  therefore we square the differences (this way, large deviations count more), and divide by expected in order to normalize (be independent of how many samples we have)

Symbol	Rock	Paper	Scissors
observed	30	21	24
expected	25	25	25

$$\sum O - E$$
$$= (30 - 25) + (21 - 25) + (24 - 25)$$
$$= 0 \text{ (so this doesn't work)}$$

Is the observed significantly different from the expected?

- calculate difference between observed and expected for each category.
- but we need to make sure we're not summing up positive and negative values
  therefore we square the differences (this way, large deviations count more), and divide by expected in order to normalize (be independent of how many samples we have)

Symbol	Rock	Paper	Scissors
observed	30	21	24
expected	25	25	25

$$\sum O - E$$
$$= (30 - 25) + (21 - 25) + (24 - 25)$$
$$= 0 \text{ (so this doesn't work)}$$

$$\sum \frac{(O-E)^2}{E}$$
$$= \frac{(30-25)^2}{25} + \frac{(21-25)^2}{25} + \frac{(24-25)^2}{25}$$
$$= 1.68$$

Is the observed significantly different from the expected?

- calculate difference between observed and expected for each category.
- but we need to make sure we're not summing up positive and negative values
  therefore we square the differences (this way, large deviations count more), and divide by expected in order to normalize (be independent of how many samples we have)

Symbol	Rock	Paper	Scissors
observed	30	21	24
expected	25	25	25

$$\sum O - E$$
$$= (30 - 25) + (21 - 25) + (24 - 25)$$
$$= 0 \text{ (so this doesn't work)}$$

$$\sum \frac{(O-E)^2}{E}$$
$$= \frac{(30-25)^2}{25} + \frac{(21-25)^2}{25} + \frac{(24-25)^2}{25}$$
$$= 1.68$$

(this is the χ^2 value, for which we can look up significance using the χ^2 distribution, like we did with the z value for the normal distribution.)

Exercise

There are four tutorials for the same course, taught by four different instructors. The head of department suspects that some instructors are more popular with students than others. The number of students enrolled in each of the four tutorials is:

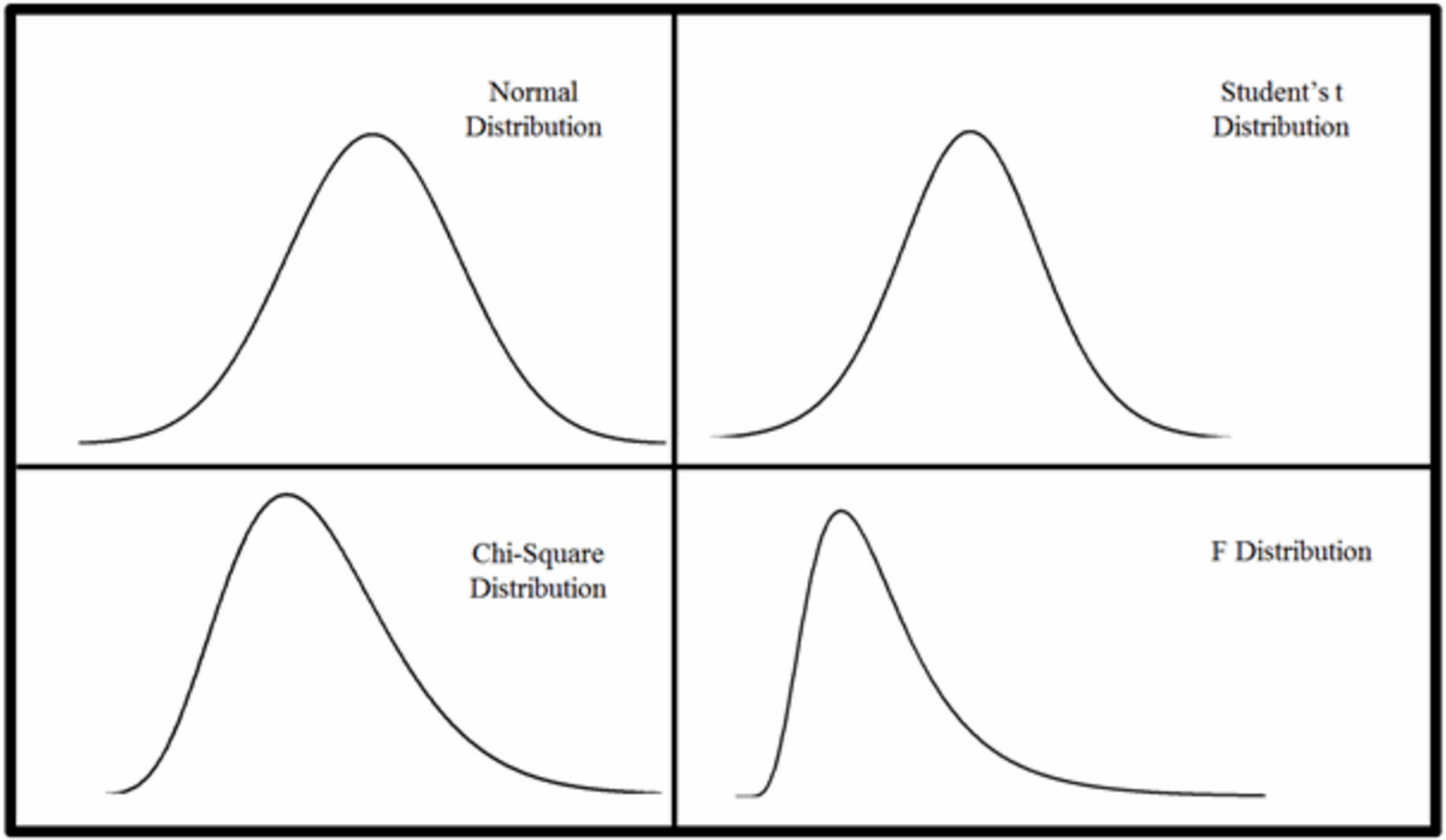
Instructor Miller	Instructor Smith	Instructor Anderson	Instructor Peters
28	25	14	13

Calculate the χ^2 value for whether this distribution of students to instructors might have happened by chance.

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Table of Contents

The χ^2 distribution



... but actually the χ^2 distribution has lots of shapes...

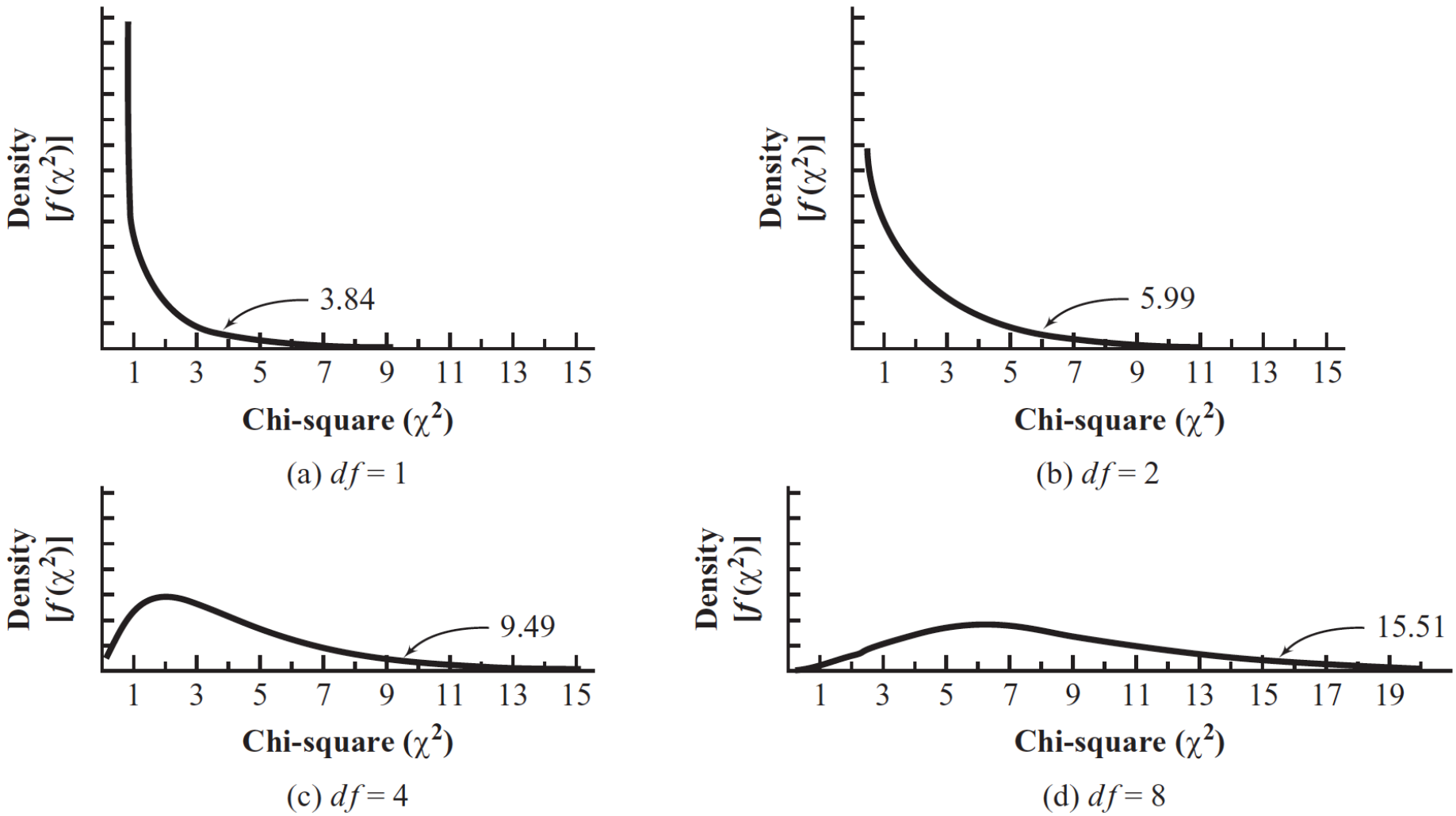
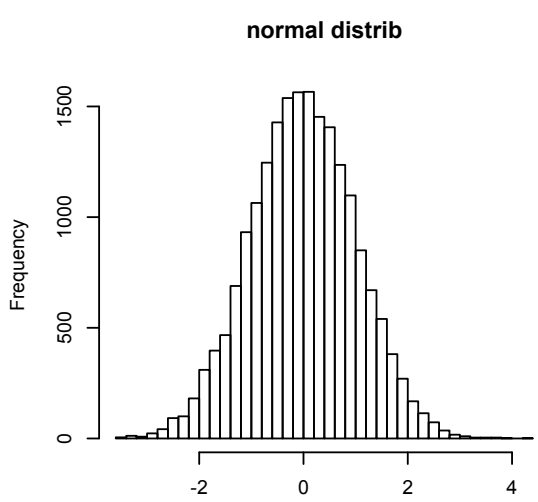
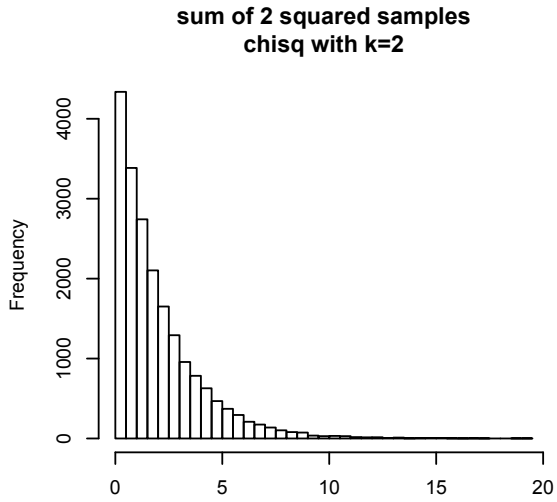
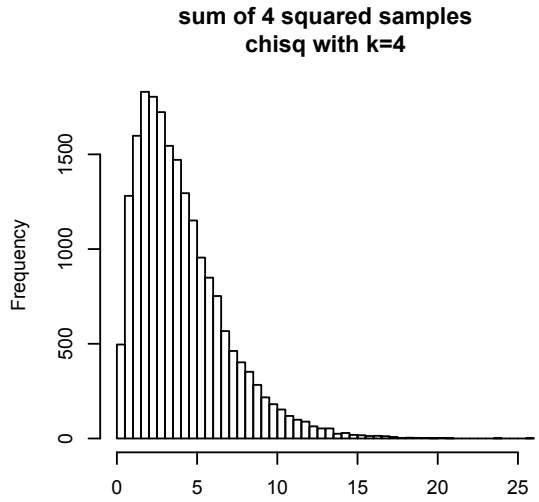


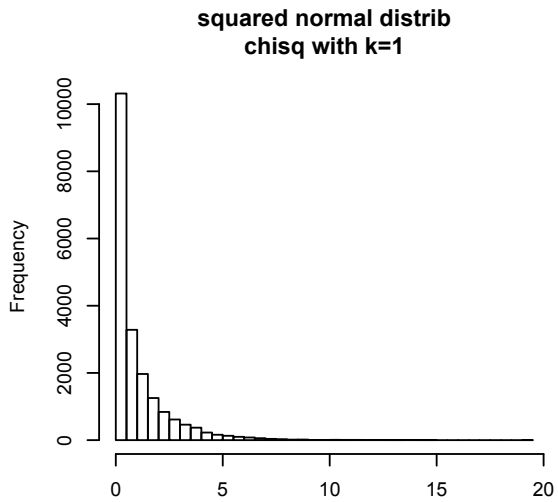
Figure 6.1 Chi-square distributions for $df = 1, 2, 4$, and 8. (Arrows indicate critical values at alpha = .05.)

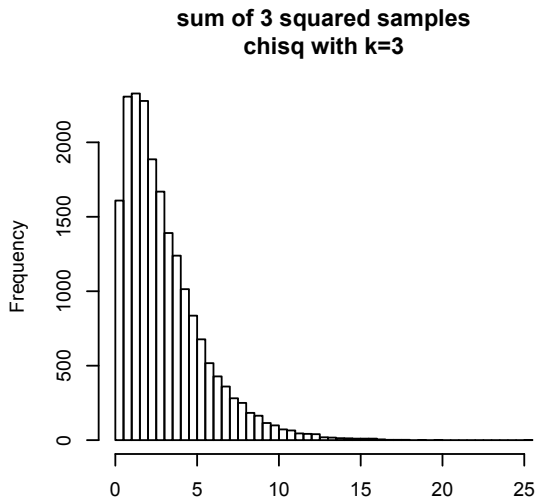
The χ^2distribution describes the distribution of summing up k normal distributions that have been squared.

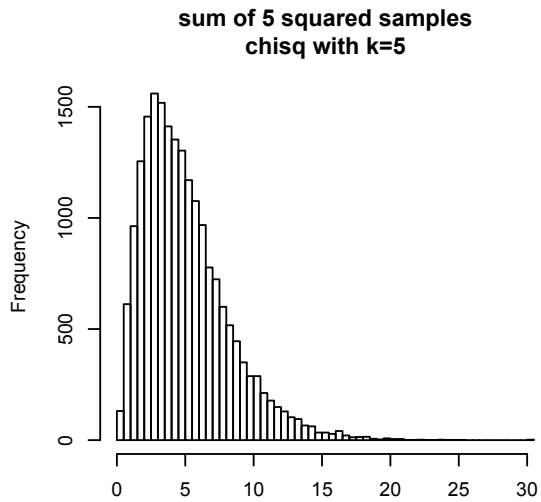




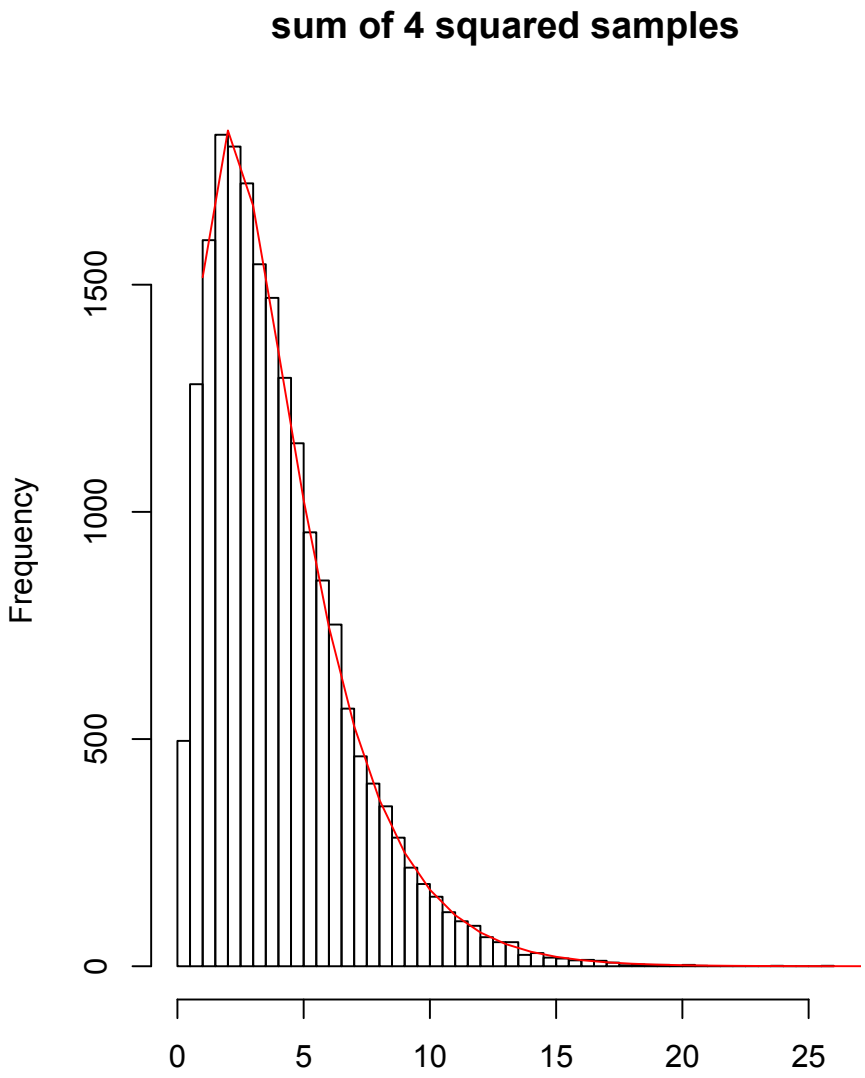


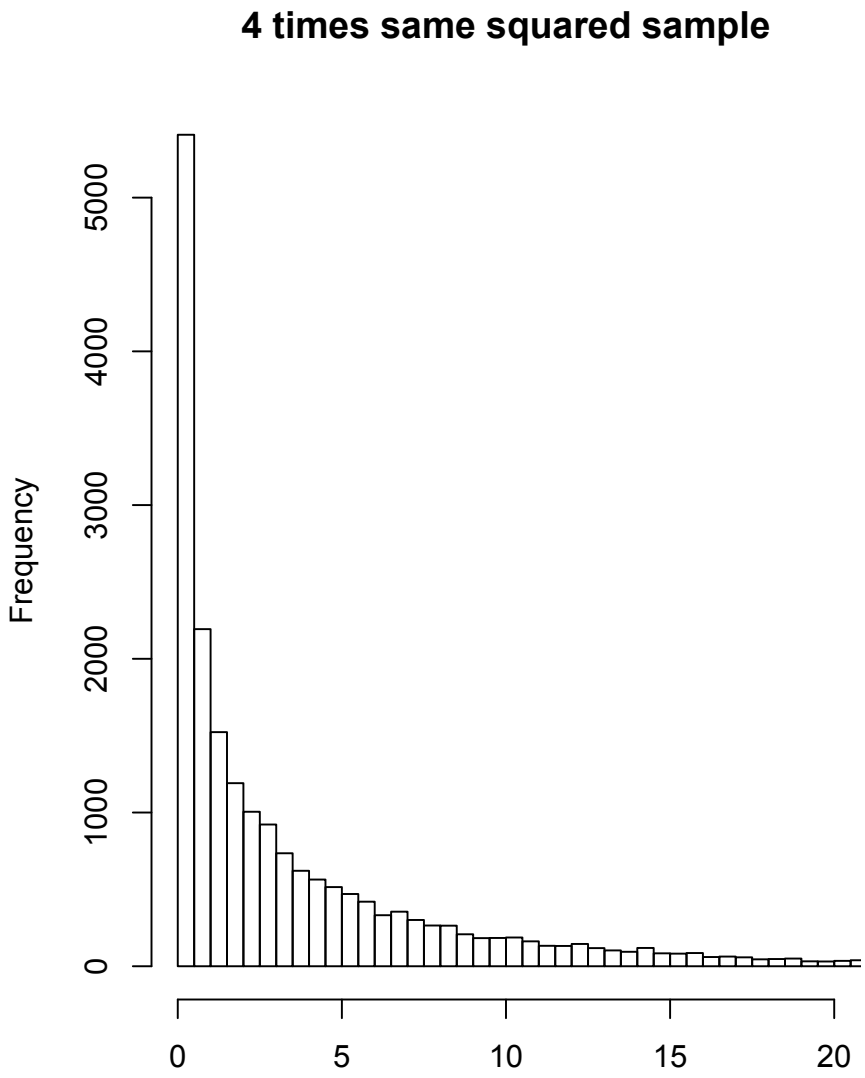






If you want to repeat this at home, make sure you draw a fresh distribution sample each time!





... back to our example

we obtained a χ^2value of 1.68 for the rock-paper-scissors example.
But which version of the χ^2distributions do we need to look at?

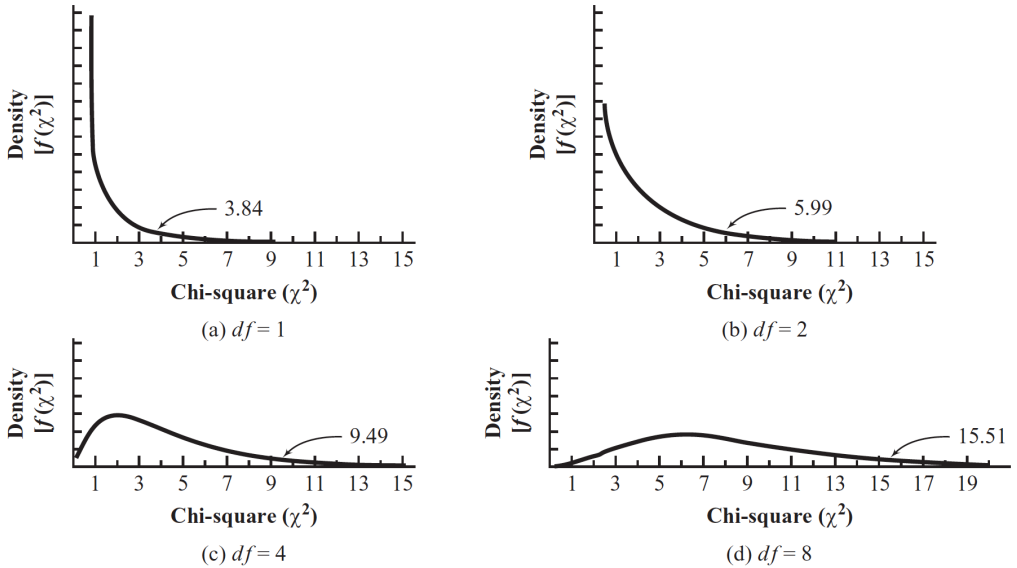


Figure 6.1 Chi-square distributions for $df = 1, 2, 4$, and 8. (Arrows indicate critical values at alpha = .05.)

$\rightarrow$ how many "degrees of freedom" do we have?

... back to our example

we obtained a χ^2value of 1.68 for the rock-paper-scissors example.
But which version of the χ^2distributions do we need to look at?

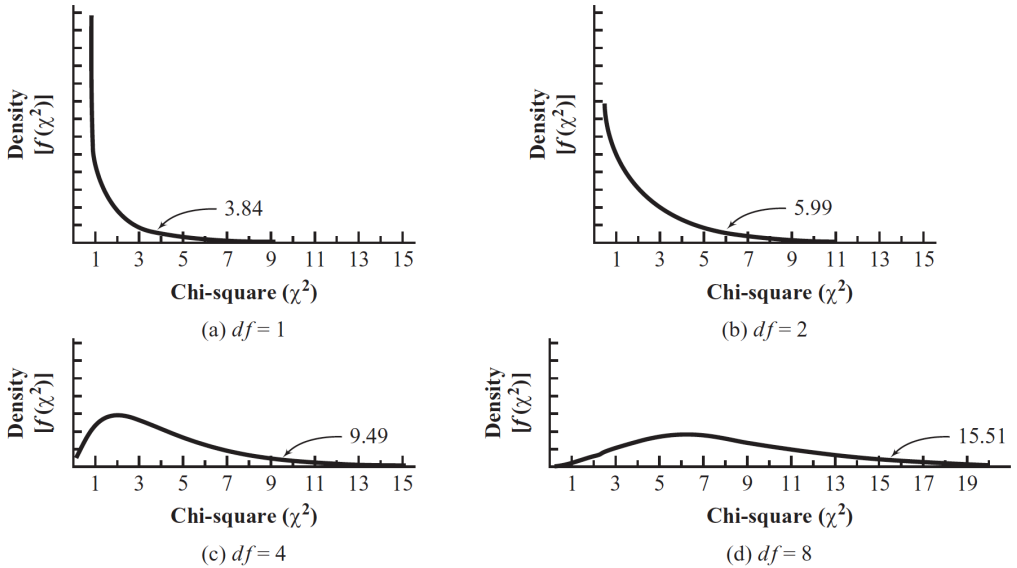


Figure 6.1 Chi-square distributions for $df = 1, 2, 4$, and 8. (Arrows indicate critical values at alpha = .05.)

$\rightarrow$ how many "degrees of freedom" do we have?

= How many squared normal distributions do we need to add, and why do we do anything with normal distributions in the first place here??????

how many degrees of freedom?

The normal distribution is used here as an approximation of the binomial distribution.

how many degrees of freedom?

The normal distribution is used here as an approximation of the binomial distribution.

And we can think of our rock-stone-paper experiment as two bernouilli trials:
1) rock or not-rock ($p = 0.333$)
2) if not-rock: paper or scissors ($p = 0.5$)

Therefore, the χ^2distribution for two degrees of freedom is chosen.

how many degrees of freedom?

The normal distribution is used here as an approximation of the binomial distribution.

And we can think of our rock-stone-paper experiment as two bernouilli trials:
1) rock or not-rock ($p = 0.333$)
2) if not-rock: paper or scissors ($p = 0.5$)

Therefore, the χ^2distribution for two degrees of freedom is chosen.
(You can remember: number of categories minus 1.)

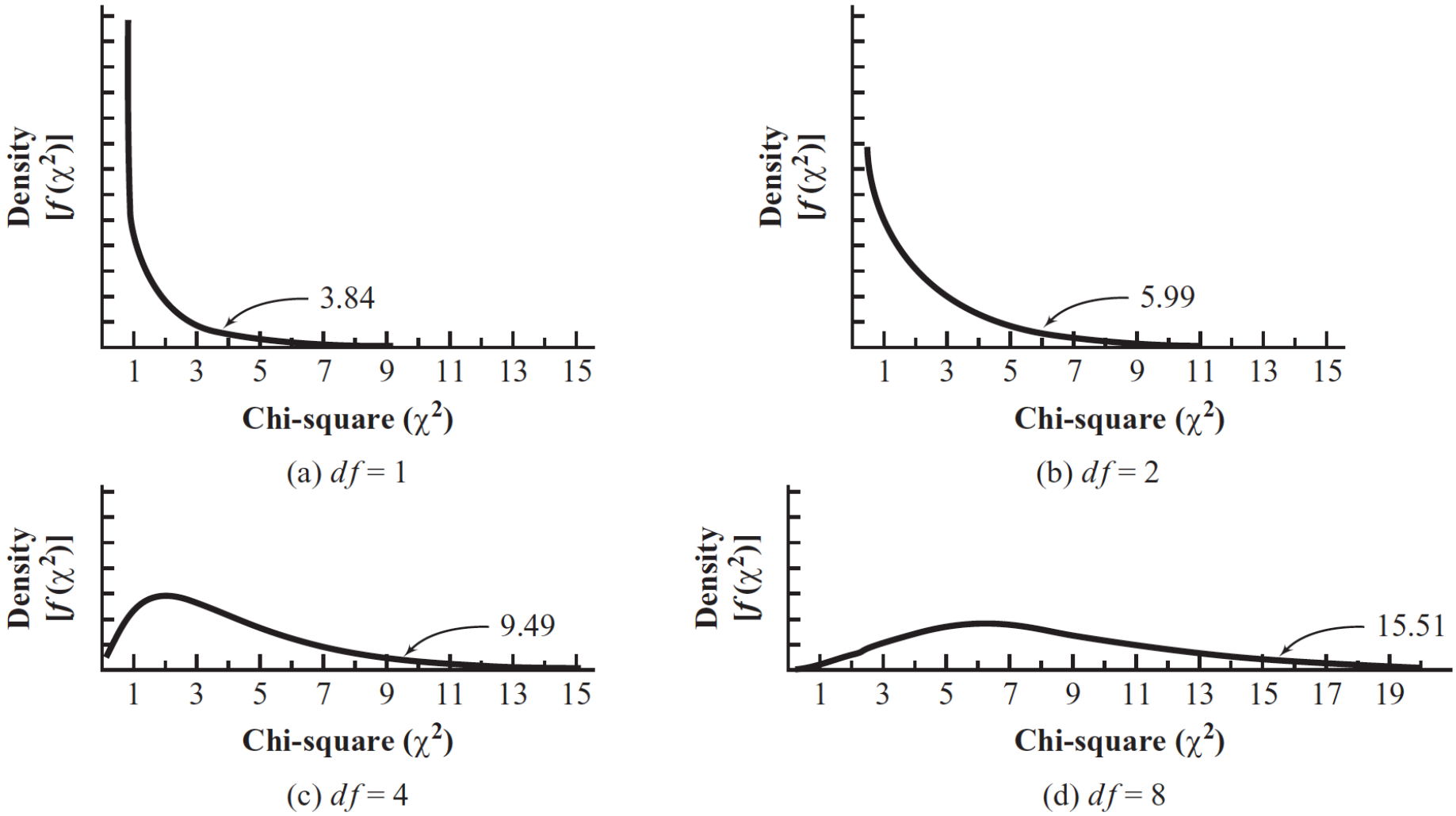


Figure 6.1 Chi-square distributions for $df = 1, 2, 4$, and 8. (Arrows indicate critical values at alpha = .05.)

Our value of 1.68 is not anywhere near the rejection region.

We can look this up in the χ^2 table

Table 6.2 Upper percentage points of the χ^2 distribution

df	.995	.990	.975	.950	.900	.750	.500	.250	.100	.050	.025	.010	.005
1	0.00	0.00	0.00	0.00	0.02	0.10	0.45	1.32	2.71	**3.84**	5.02	6.63	7.88
2	0.01	0.02	0.05	0.10	0.21	0.58	1.39	2.77	4.61	5.99	7.38	9.21	10.60
3	0.07	0.11	0.22	0.35	0.58	1.21	2.37	4.11	6.25	7.82	9.35	11.35	12.84
4	0.21	0.30	0.48	0.71	1.06	1.92	3.36	5.39	7.78	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	1.61	2.67	4.35	6.63	9.24	11.07	12.83	15.09	16.75
6	0.68	0.87	1.24	1.64	2.20	3.45	5.35	7.84	10.64	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	2.83	4.25	6.35	9.04	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	5.07	7.34	10.22	13.36	15.51	17.54	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	5.90	8.34	11.39	14.68	16.92	19.02	21.66	23.59
...	...	...	...	...	...	...	...	...	...	...	...	...	...

... or use R

The (non-central) Chi-Squared Distribution

Description

Density, distribution function, quantile function and random generation for the chi-squared (*chi^2*) distribution with `df` degrees of freedom and optional non-centrality parameter `ncp`.

Usage

```
dchisq(x, df, ncp = 0, log = FALSE)
pchisq(q, df, ncp = 0, lower.tail = TRUE, log.p = FALSE)
qchisq(p, df, ncp = 0, lower.tail = TRUE, log.p = FALSE)
rchisq(n, df, ncp = 0)
```

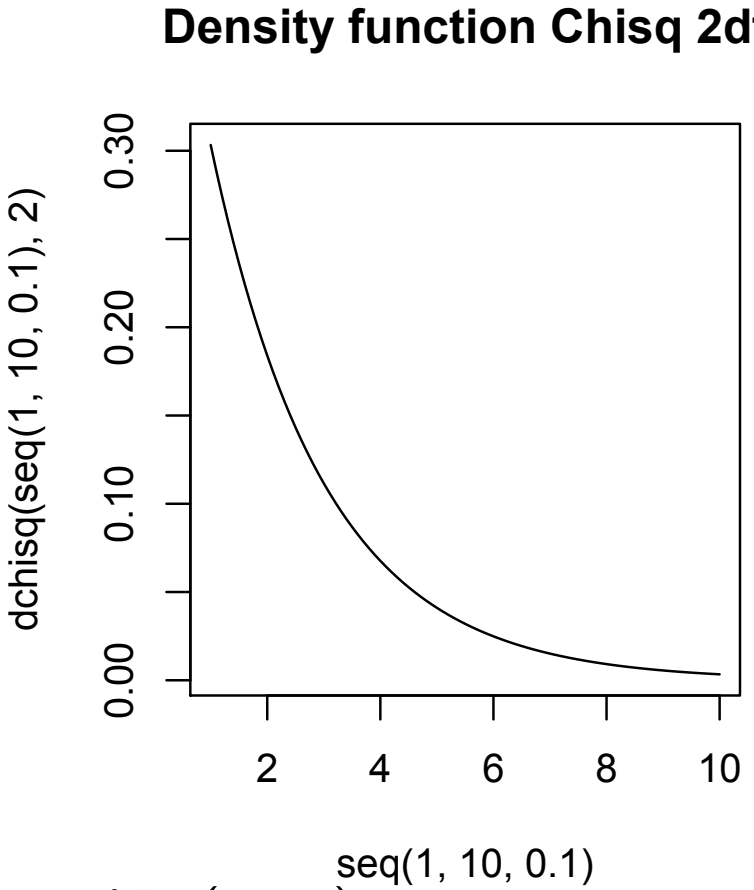
Arguments

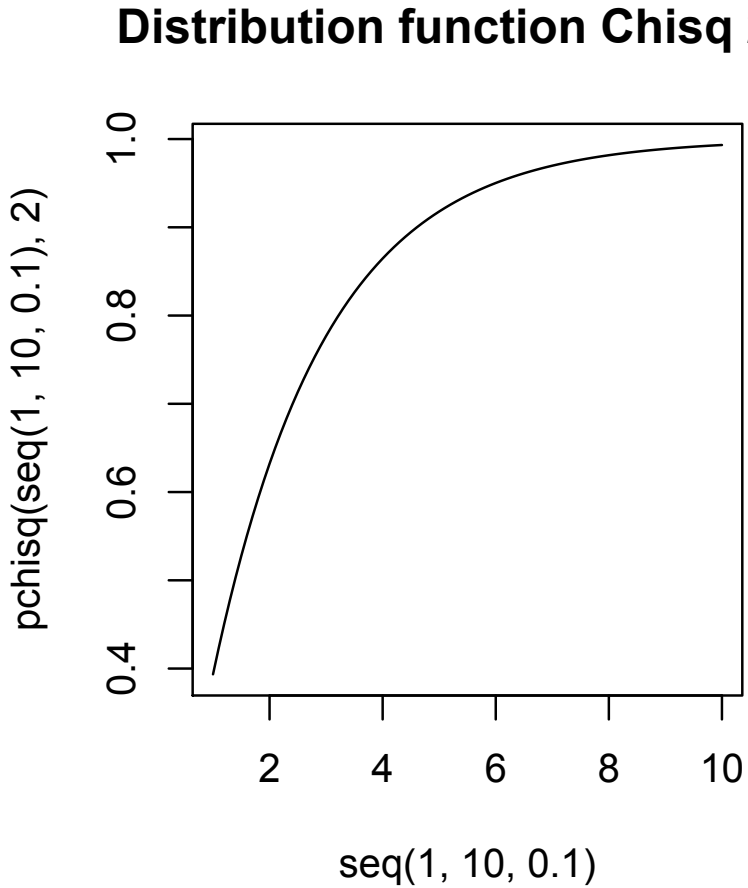
`x, q`
: vector of quantiles.

`p`
: vector of probabilities.

`n`
: number of observations. If `length(n) > 1`, the length is taken to be the number required.

`df`
: degrees of freedom (non-negative, but can be non-integer)






```
> qchisq(.95,2)
5.991465

> rchisq(10,2)
1.58964669 3.70070340 3.13125961 0.53948551 1.39359956 3.52782730
0.24961532 0.95064728 0.06941609 0.80893519
```

Table of Contents

Another example: Death penalty study

Question: when the victim was white, were nonwhite people more likely to receive a death sentence than white people?

	Death Sentence		
Defendant's Race	Yes	No	Total
Nonwhite	33 (22.72)	251 (261.28)	284
White	33 (43.28)	508 (497.72)	541
Total	66	759	825

(from Howell book)

This is an example of a "contingency table analysis"
Are the variables (race and death sentence) conditional (= contingent) on one another?
or are they independent (this is why the χ^2test is also sometimes called the χ^2test of independence)?

one vs. two samples

What we saw so far was an analysis where we only had one variable (in rock-paper-scissors and instructor popularity)

For the death sentence example, we have two samples.

Calculating expected values

	Death Sentence		
Defendant's Race	Yes	No	Total
Nonwhite	33 (22.72)	251 (261.28)	284
White	33 (43.28)	508 (497.72)	541
Total	66	759	825

numbers in brackets are expected values.

Calculating expected values

	Death Sentence		
Defendant's Race	Yes	No	Total
Nonwhite	33 (22.72)	251 (261.28)	284
White	33 (43.28)	508 (497.72)	541
Total	66	759	825

numbers in brackets are expected values.

But where do these come from?
How can we calculate our expectation from the observed data?
$\rightarrow$ calculate the marginals:

$$E_{ij} = \frac{R_i \times C_j}{N} \qquad E_{11} = \frac{(33 + 251) \times (33 + 33)}{825} = \frac{284 \times 66}{825} = 22.72$$

Calculating χ^2

	Death Sentence		
Defendant's Race	Yes	No	Total
Nonwhite	33 (22.72)	251 (261.28)	284
White	33 (43.28)	508 (497.72)	541
Total	66	759	825

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

$$= \frac{(33 - 22.72)^2}{22.72} + \frac{(251 - 261.28)^2}{261.28} + \frac{(33 - 43.28)^2}{43.28} + \frac{(508 - 497.82)^2}{497.72}$$

$$= 7.71$$

Degrees of freedom: $df = (R - 1)(C - 1) = 1$

We can look this up in the χ^2 table

Table 6.2 Upper percentage points of the χ^2 distribution

df	.995	.990	.975	.950	.900	.750	.500	.250	.100	.050	.025	.010	.005
1	0.00	0.00	0.00	0.00	0.02	0.10	0.45	1.32	2.71	**3.84**	5.02	6.63	7.88
2	0.01	0.02	0.05	0.10	0.21	0.58	1.39	2.77	4.61	5.99	7.38	9.21	10.60
3	0.07	0.11	0.22	0.35	0.58	1.21	2.37	4.11	6.25	7.82	9.35	11.35	12.84
4	0.21	0.30	0.48	0.71	1.06	1.92	3.36	5.39	7.78	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	1.61	2.67	4.35	6.63	9.24	11.07	12.83	15.09	16.75
6	0.68	0.87	1.24	1.64	2.20	3.45	5.35	7.84	10.64	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	2.83	4.25	6.35	9.04	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	5.07	7.34	10.22	13.36	15.51	17.54	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	5.90	8.34	11.39	14.68	16.92	19.02	21.66	23.59
...	...	...	...	...	...	...	...	...	...	...	...	...	...

Calculating significance in R

```
pchisq(7.71,1)
0.9945084
pchisq(7.71,1, lower.tail=FALSE)
0.005491575
```

Conclusion: we will reject the null hypothesis.

We conclude that whether or not a death sentence is imposed is related to the race of the defendant. When the victim was white, nonwhite defendants were more likely to receive the death penalty than white defendants.

```
     Chi-square test of categorical association

Variables:   race, death_sent

Hypotheses:
   null:        variables are independent of one another
   alternative: some contingency exists between variables

Observed contingency table:
          death_sent
race        no yes
  nonwhite 251  33
  white    508  33

Expected contingency table under the null hypothesis:
          death_sent
race        no  yes
  nonwhite 261 22.7
  white    498 43.3

Test results:
   X-squared statistic:  6.978
   degrees of freedom:  1
   p-value:  0.008

Other information:
   estimated effect size (Cramer's v):  0.092
   Yates' continuity correction has been applied
```

Table of Contents

Assumptions

There are certain assumptions you need to know about when applying the χ^2 test.

- expected frequencies should not be too small
- independent observations

too small expected frequencies

All the values in the cells have to be "sufficiently large" (usually > 5).

Why?

too small expected frequencies

All the values in the cells have to be "sufficiently large" (usually > 5).

Why?
Remember the χ^2 distribution is used on the grounds that it's a sum of (squared) binomials, where the binomials are approximated by the normal distribution. The binomial distribution however only looks similar to the normal distribution if the values Np and Nq are "large enough". ($Np > 5$ and $Nq > 5$)

too small expected frequencies

All the values in the cells have to be "sufficiently large" (usually > 5).

Why?
Remember the χ^2 distribution is used on the grounds that it's a sum of (squared) binomials, where the binomials are approximated by the normal distribution. The binomial distribution however only looks similar to the normal distribution if the values Np and Nq are "large enough". ($Np > 5$ and $Nq > 5$)

...uhm, so what if there are cells with values smaller than 5?

Fisher's Exact Test

Idea:

- take all possible 2×2 tables that could be formed from the fixed set of marginal totals.
- determine the sum of the probabilities of those tables whose results were as extreme or more extreme than the table obtained in the data.
- if this sum is less than α, reject the null hypothesis

It's a *conditional* test, because it conditions on knowing the number of marginal totals.

Fisher's exact test in R

```
> table(ds)
           death_sent
race          no yes
  nonwhite   251  33
  white      508  33
> fisher.test(table(ds))

    Fisher's Exact Test for Count Data

data:  table(ds)
p-value = 0.006809
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.2884159 0.8477621
sample estimates:
odds ratio
 0.4945367
```

Independence of observations

Our observations aren't always independent of one another.

Examples:

- your experiment asks someone what brand of car they prefer before and after watching a certain movie
- you have created different models for assessing how well nouns and verbs *fit* together, and want to see which model is better
- you study whether children seek help before and after an intervention program that was supposed to increase their help-seeking behaviour.

You want to know how things *changed* for each movie viewer / noun-verb pair / child.

McNemar's test

Core idea: tabulate only changes!

Table 6.6 Help-seeking behavior in fall and spring

		Spring		
		Yes	No	Total
Fall	Yes	38	4	42
	No	12	18	30
	Total	50	22	72

Table 6.7 Results of experiment on help-seeking behavior in children

	No → Yes	Yes → No	Total
Observed	12	4	16
Expected	8.0	8.0	16

Summary

- Tests for categorial data
- binomial test is exact (but only for 2 outcomes)
- Pearson's χ^2 test is approximation, but for more categories
- χ^2 test has as a parameter the degrees of freedom
- χ^2 can test independence of one or two variables
- important assumptions: sufficiently many observations in all cells, independence of observations ($\rightarrow$ Fisher's test and McNemar's test)

Chai-squared test

	Variable A		Total
Variable B	 Latte	 Vanilla	Afernoon Tea
	 Cinnamon	 De-caf	Bedtime
Total	Special Treat	Tea Break	A good Cuppa!