Lecture 20: Chi-square Tests

- Contingency tables
- Tests of goodness-of-fit
- Tests of independence

Introduction

– Today we will continue with our discussions on inference for categorical data.

– We will talk about one way of testing the null hypothesis that data comes from some distribution we have in mind. This is known as goodness-of-fit.

– Lastly, we will talk about testing for independence between two "categorical" variables.

Considering Categorical Data

Often we observe data where each unit/individual can be categorized according to two different criteria (two categorical variables). For example:

- Each person gets a drug or a placebo, and each person is cured or not;
- Letter grade in a statistics course, and major.

Example 1: For example, suppose we observe 90 students at Duke and are interested in the relationship between major and gender, then we can display the data by gender and major as below:

		Major	
	Math	**English**	**History**
Male	**10**	**20**	**15**
Female	**20**	**10**	**15**

so that 10 males study math, 20 females study math, and so on.

Contingency tables

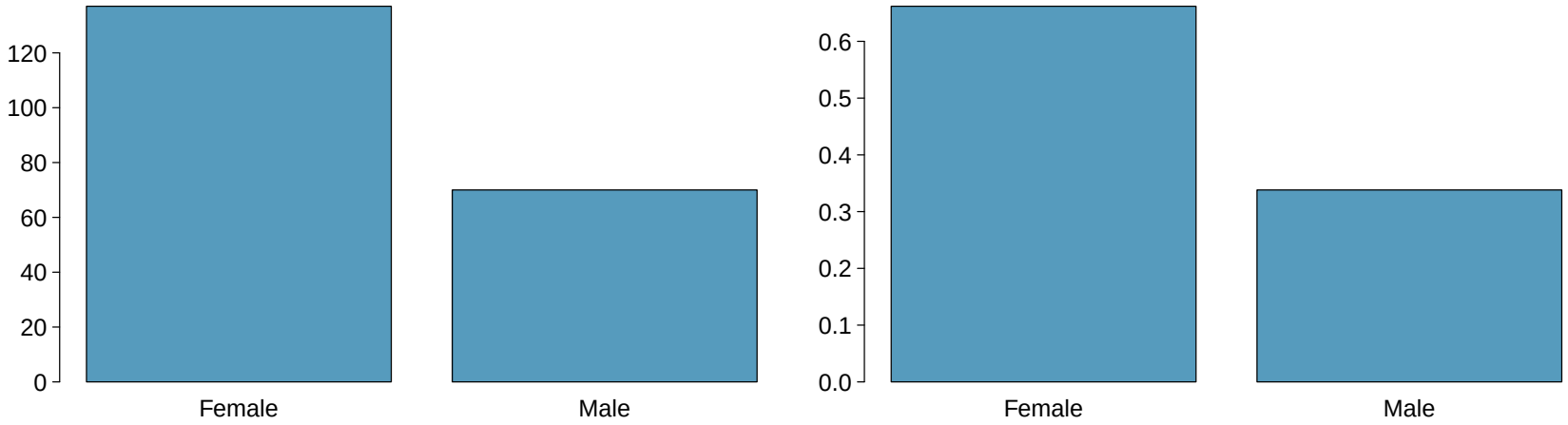
A table that summarizes data for two categorical variables is called a *contingency table*.

The contingency table below shows the distribution of students' genders and whether or not they are looking for a spouse while in college.

		looking for spouse		
		No	Yes	Total
gender	Female	86	51	137
	Male	52	18	70
	Total	138	69	207

Bar plots

A *bar plot* is a common way to display a single categorical variable. A bar plot where proportions instead of frequencies are shown is called a *relative frequency bar plot*.



How are bar plots different than histograms?

Bar plots are used for displaying distributions of categorical variables, while histograms are used for numerical variables. The x-axis in a histogram is a number line, hence the order of the bars cannot be changed, while in a bar plot the categories can be listed in any order (though some orderings make more sense than others, especially for ordinal variables.)

Choosing the appropriate proportion

Does there appear to be a relationship between gender and whether the student is looking for a spouse in college?

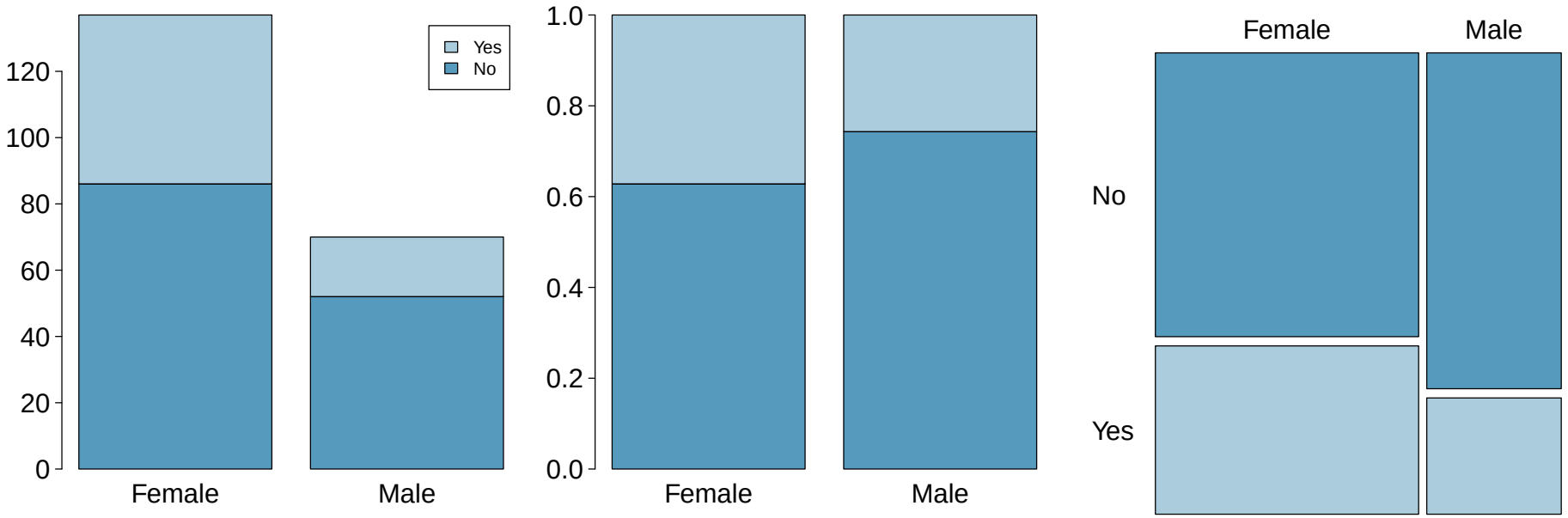
		looking for spouse		
		No	Yes	Total
gender	Female	86	51	137
	Male	52	18	70
	Total	138	69	207

To answer this question we examine the row proportions:

- % Females looking for a spouse: $51/137 \approx 0.37$
- % Males looking for a spouse: $18/70 \approx 0.26$

Segmented bar and mosaic plots

What are the differences between the three visualizations shown below?



Weldon's dice

- Walter Frank Raphael Weldon (1860 - 1906), was an English evolutionary biologist and a founder of biometry. He was the joint founding editor of Biometrika, with Francis Galton and Karl Pearson.
- In 1894, he rolled 12 dice 26,306 times, and recorded the number of 5s or 6s (which he considered to be a success).

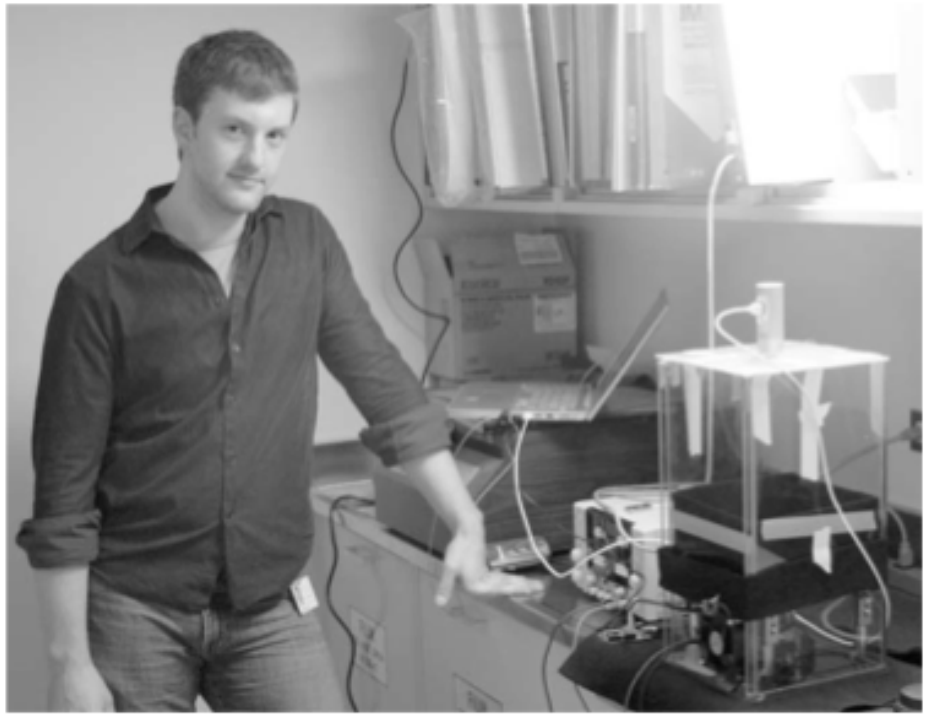
- It was observed that 5s or 6s occurred more often than expected, and Pearson hypothesized that this was probably due to the construction of the dice. Most inexpensive dice have hollowed-out pips, and since opposite sides add to 7, the face with 6 pips is lighter than its opposing face, which has only 1 pip.

Labby's dice

- In 2009, Zacariah Labby (U of Chicago), repeated Weldon's experiment using a homemade dice-throwing, pip counting machine.

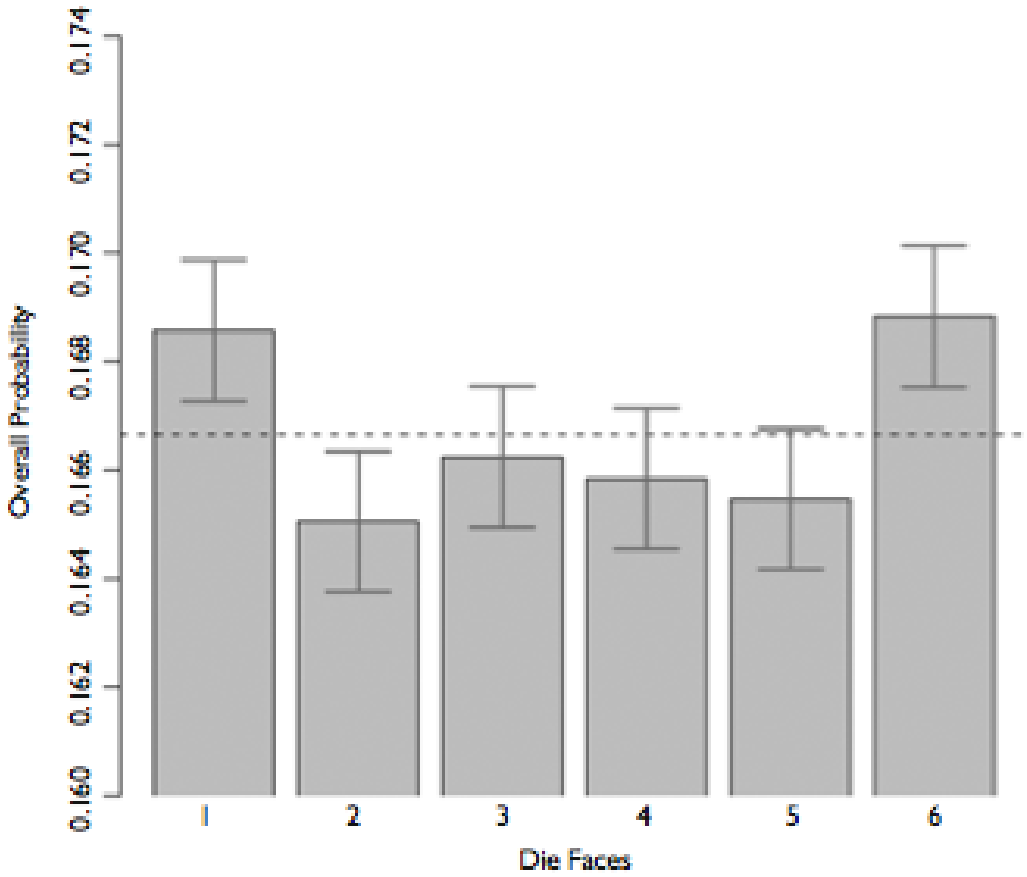
  http://www.youtube.com/watch?v=95EErdouO2w

- The rolling-imaging process took about 20 seconds per roll.



- Each day there were ~150 images to process manually.
- At this rate Weldon's experiment was repeated in a little more than six full days.
- Recommended reading: *https://galton.uchicago.edu/about/docs/2009/2009_dice_zac_labby.pdf*

- Labby did not actually observe the same phenomenon that Weldon observed (higher frequency of 5s and 6s).
- Automation allowed Labby to collect more data than Weldon did in 1894, instead of recording “successes" and “failures", Labby recorded the individual number of pips on each die.



Expected counts

Labby rolled 12 dice 26,306 times. If each side is equally likely to come up, how many 1s, 2s, $\cdots$, 6s would he expect to have observed?

(a) $\frac{1}{6}$

(b) $\frac{12}{6}$

(c) $\frac{26{,}306}{6}$

(d) $\frac{12 \times 26{,}306}{6}$

Expected counts

Labby rolled 12 dice 26,306 times. If each side is equally likely to come up, how many 1s, 2s, $\cdots$, 6s would he expect to have observed?

(a) $\frac{1}{6}$

(b) $\frac{12}{6}$

(c) $\frac{26{,}306}{6}$

(d) $\frac{12 \times 26{,}306}{6} = 52,612$

Summarizing Labby's results

The table below shows the observed and expected counts from Labby's experiment.

Outcome	Observed	Expected
1	53,222	52,612
2	52,118	52,612
3	52,465	52,612
4	52,338	52,612
5	52,244	52,612
6	53,285	52,612
Total	315,672	315,672

Why are the expected counts the same for all outcomes but the observed counts are different? At a first glance, does there appear to be an inconsistency between the observed and expected counts?

Setting the hypotheses

Do these data provide convincing evidence of an inconsistency between the observed and expected counts?

H_0: There is no inconsistency between the observed and the expected counts. *The observed counts follow the same distribution as the expected counts.*

H_A: There is an inconsistency between the observed and the expected counts. *The observed counts do not follow the same distribution as the expected counts.* There is a bias in which side comes up on the roll of a die.

Evaluating the hypotheses

- To evaluate these hypotheses, we quantify how different the observed counts are from the expected counts.
- Large deviations from what would be expected based on sampling variation (chance) alone provide strong evidence for the alternative hypothesis.
- This is called a *goodness of fit* test since we're evaluating how well the observed data fit the expected distribution.

Anatomy of a test statistic

- The general form of a test statistic is

$$\frac{\text{point estimate} - \text{null value}}{\text{SE of point estimate}}$$

- This construction is based on
  1. identifying the difference between a point estimate and an expected value if the null hypothesis was true, and
  2. standardizing that difference using the standard error of the point estimate.

  These two ideas will help in the construction of an appropriate test statistic for count data.

Chi-square statistic

When dealing with counts and investigating how far the observed counts are from the expected counts, we use a new test statistic called the *chi-square (χ^2) statistic*.

χ^2 statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{(O-E)^2}{E} \qquad \text{where } k = \text{total number of cells}$$

Calculating the chi-square statistic

Outcome	Observed	Expected	$\frac{(O-E)^2}{E}$
1	53,222	52,612	$\frac{(53{,}222-52{,}612)^2}{52{,}612} = 7.07$
2	52,118	52,612	$\frac{(52{,}118-52{,}612)^2}{52{,}612} = 4.64$
3	52,465	52,612	$\frac{(52{,}465-52{,}612)^2}{52{,}612} = 0.41$
4	52,338	52,612	$\frac{(52{,}338-52{,}612)^2}{52{,}612} = 1.43$
5	52,244	52,612	$\frac{(52{,}244-52{,}612)^2}{52{,}612} = 2.57$
6	53,285	52,612	$\frac{(53{,}285-52{,}612)^2}{52{,}612} = 8.61$
Total	315,672	315,672	24.73

Why square?

Squaring the difference between the observed and the expected outcome does two things:

- Any standardized difference that is squared will now be positive.
- Differences that already looked unusual will become much larger after being squared.

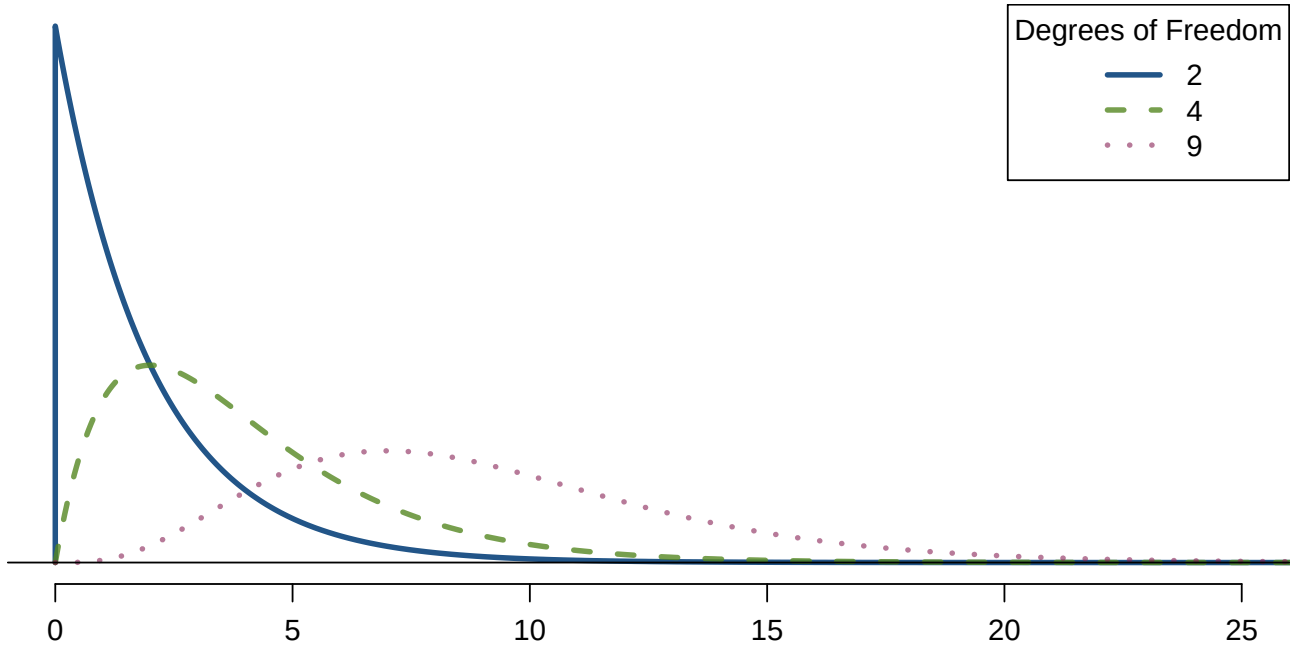
The chi-square distribution

- In order to determine if the χ^2 statistic we calculated is considered unusually high or not we need to first describe its distribution.
- The chi-square distribution has just one parameter called *degrees of freedom (df)*, which influences the shape, center, and spread of the distribution.

Three other continuous distributions:

- normal distribution: unimodal and symmetric with two parameters: mean and standard deviation
- T distribution: unimodal and symmetric with one parameter: degrees of freedom
- F distribution: unimodal and right skewed with two parameters: degrees of freedom or numerator (between group variance) and denominator (within group variance)

Which of the following is false?



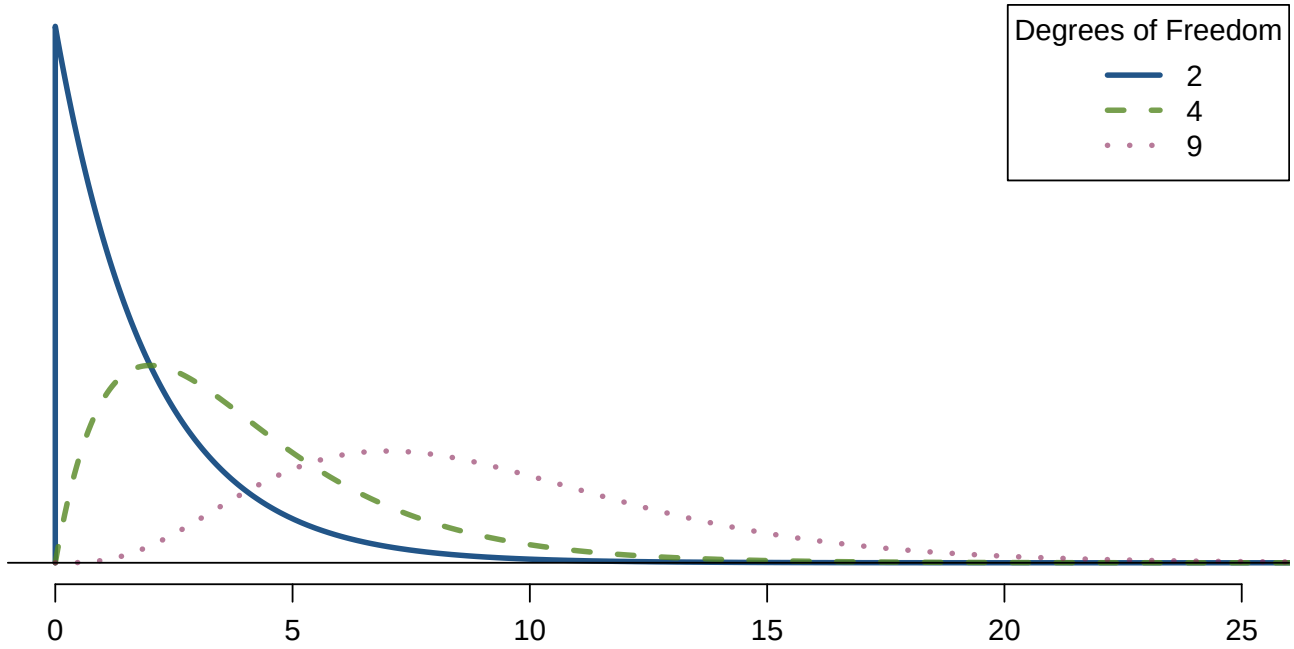
As the df increases,

(a) the center of the χ^2 distribution increases as well

(b) the variability of the χ^2 distribution increases as well

(c) the shape of the χ^2 distribution becomes more skewed (less like a normal)

Which of the following is false?



As the df increases,

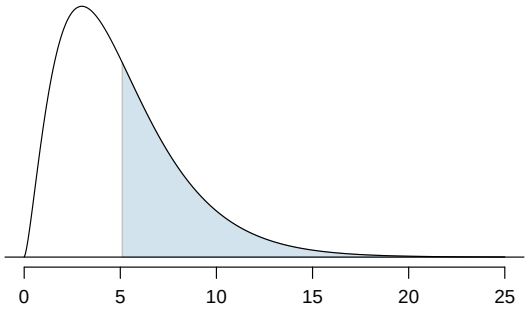
(a) the center of the χ^2 distribution increases as well

(b) the variability of the χ^2 distribution increases as well

(c) *the shape of the χ^2 distribution becomes more skewed (less like a normal)*

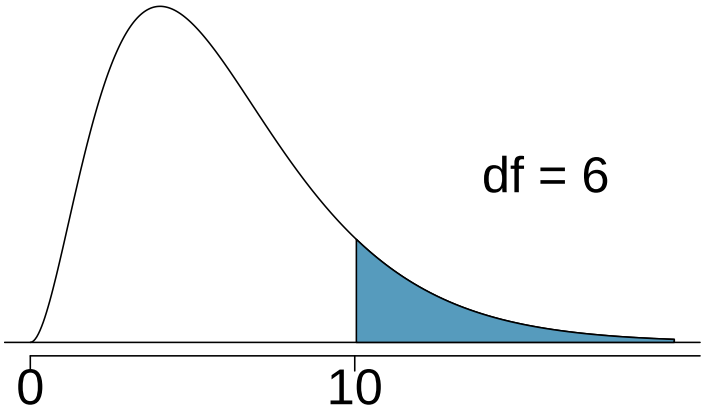
Finding areas under the chi-square curve

- p-value = tail area under the chi-square distribution (as usual)
- For this we can use technology, or a *chi-square probability table*.
- This table works a lot like the t table, but only provides upper tail values.



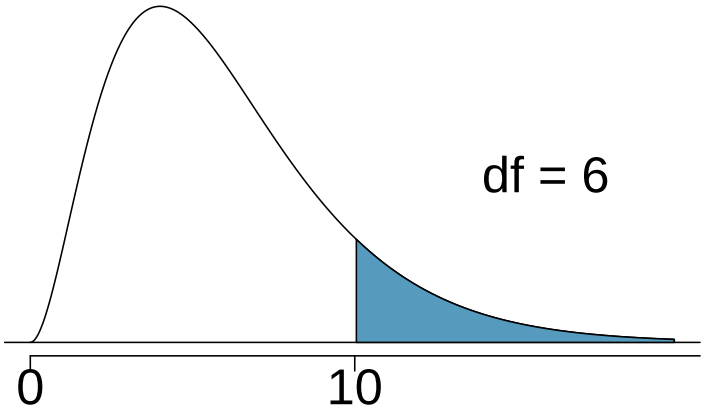
Upper tail		0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
df	1	1.07	1.64	2.71	3.84	5.41	6.63	7.88	10.83
	2	2.41	3.22	4.61	5.99	7.82	9.21	10.60	13.82
	3	3.66	4.64	6.25	7.81	9.84	11.34	12.84	16.27
	4	4.88	5.99	7.78	9.49	11.67	13.28	14.86	18.47
	5	6.06	7.29	9.24	11.07	13.39	15.09	16.75	20.52
	6	7.23	8.56	10.64	12.59	15.03	16.81	18.55	22.46
	7	8.38	9.80	12.02	14.07	16.62	18.48	20.28	24.32
	...								

Estimate the shaded area under the chi-square curve with $df = 6$.



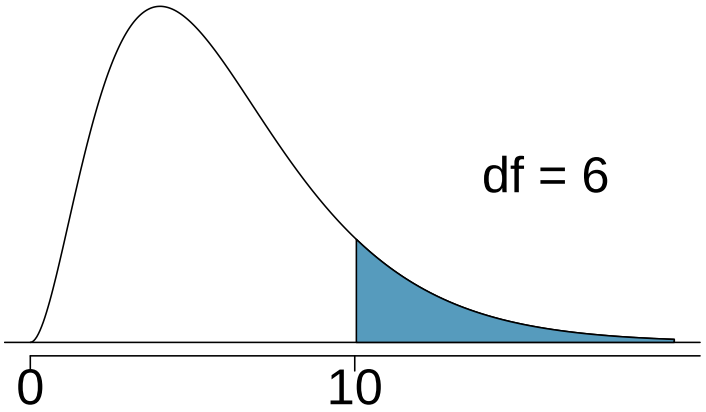
Upper tail		0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
df	1	1.07	1.64	2.71	3.84	5.41	6.63	7.88	10.83
	2	2.41	3.22	4.61	5.99	7.82	9.21	10.60	13.82
	3	3.66	4.64	6.25	7.81	9.84	11.34	12.84	16.27
	4	4.88	5.99	7.78	9.49	11.67	13.28	14.86	18.47
	5	6.06	7.29	9.24	11.07	13.39	15.09	16.75	20.52
	6	7.23	8.56	10.64	12.59	15.03	16.81	18.55	22.46
	7	8.38	9.80	12.02	14.07	16.62	18.48	20.28	24.32

Estimate the shaded area under the chi-square curve with $df = 6$.



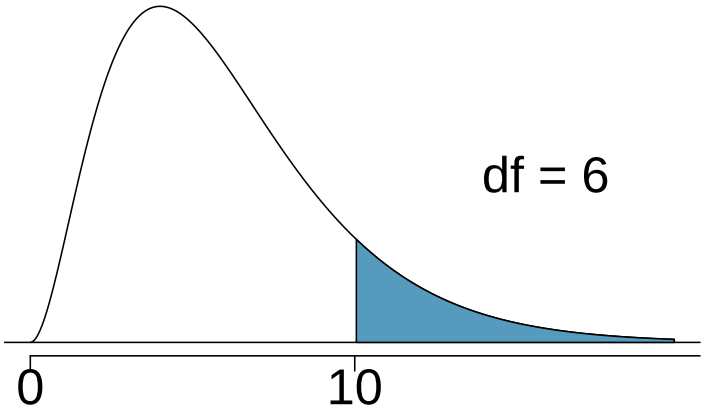
Upper tail		0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
df	1	1.07	1.64	2.71	3.84	5.41	6.63	7.88	10.83
	2	2.41	3.22	4.61	5.99	7.82	9.21	10.60	13.82
	3	3.66	4.64	6.25	7.81	9.84	11.34	12.84	16.27
	4	4.88	5.99	7.78	9.49	11.67	13.28	14.86	18.47
	5	6.06	7.29	9.24	11.07	13.39	15.09	16.75	20.52
	6	7.23	8.56	10.64	12.59	15.03	16.81	18.55	22.46
	7	8.38	9.80	12.02	14.07	16.62	18.48	20.28	24.32

Estimate the shaded area under the chi-square curve with $df = 6$.



Upper tail		0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
df	1	1.07	1.64	2.71	3.84	5.41	6.63	7.88	10.83
	2	2.41	3.22	4.61	5.99	7.82	9.21	10.60	13.82
	3	3.66	4.64	6.25	7.81	9.84	11.34	12.84	16.27
	4	4.88	5.99	7.78	9.49	11.67	13.28	14.86	18.47
	5	6.06	7.29	9.24	11.07	13.39	15.09	16.75	20.52
	6	7.23	*8.56*	*10.64*	12.59	15.03	16.81	18.55	22.46
	7	8.38	9.80	12.02	14.07	16.62	18.48	20.28	24.32

Estimate the shaded area under the chi-square curve with $df = 6$.

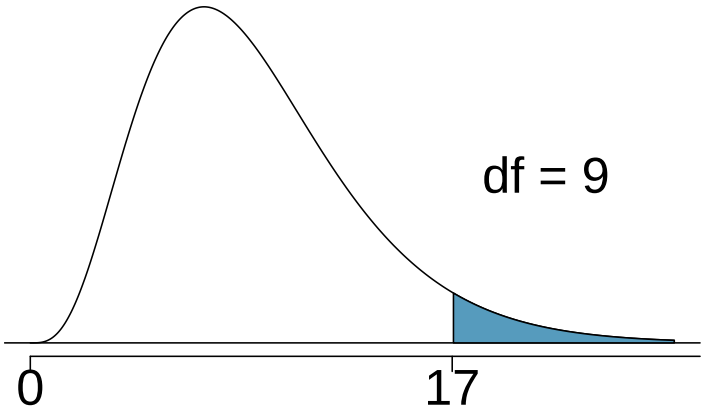


$P(\chi^2_{df=6} > 10)$
is between 0.1 and 0.2

Upper tail		0.3	*0.2*	*0.1*	0.05	0.02	0.01	0.005	0.001
df	1	1.07	1.64	2.71	3.84	5.41	6.63	7.88	10.83
	2	2.41	3.22	4.61	5.99	7.82	9.21	10.60	13.82
	3	3.66	4.64	6.25	7.81	9.84	11.34	12.84	16.27
	4	4.88	5.99	7.78	9.49	11.67	13.28	14.86	18.47
	5	6.06	7.29	9.24	11.07	13.39	15.09	16.75	20.52
	6	7.23	*8.56*	*10.64*	12.59	15.03	16.81	18.55	22.46
	7	8.38	9.80	12.02	14.07	16.62	18.48	20.28	24.32

Finding areas under the chi-square curve (cont.)

Estimate the shaded area (above 17) under the χ^2 curve with $df = 9$.

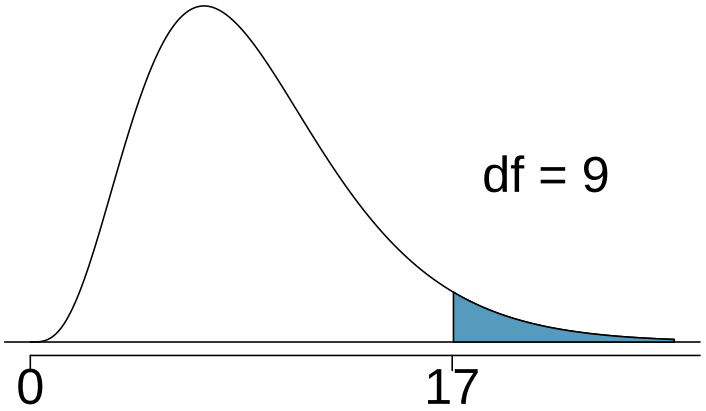


(a) 0.05
(b) 0.02
(c) between 0.02 and 0.05
(d) between 0.05 and 0.1
(e) between 0.01 and 0.02

Upper tail		0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
df	7	8.38	9.80	12.02	14.07	16.62	18.48	20.28	24.32
	8	9.52	11.03	13.36	15.51	18.17	20.09	21.95	26.12
	9	10.66	12.24	14.68	16.92	19.68	21.67	23.59	27.88
	10	11.78	13.44	15.99	18.31	21.16	23.21	25.19	29.59
	11	12.90	14.63	17.28	19.68	22.62	24.72	26.76	31.26

Finding areas under the chi-square curve (cont.)

Estimate the shaded area (above 17) under the χ^2 curve with $df = 9$.

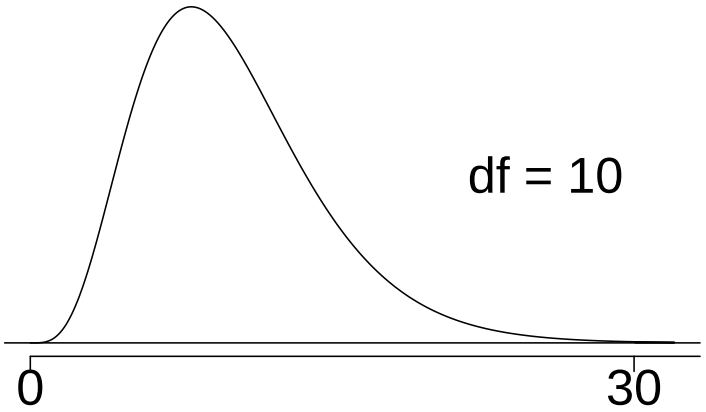


(a) 0.05
(b) 0.02
(c) *between 0.02 and 0.05*
(d) between 0.05 and 0.1
(e) between 0.01 and 0.02

Upper tail		0.3	0.2	0.1	*0.05*	*0.02*	0.01	0.005	0.001
df	7	8.38	9.80	12.02	14.07	16.62	18.48	20.28	24.32
	8	9.52	11.03	13.36	15.51	18.17	20.09	21.95	26.12
	9	10.66	12.24	14.68	*16.92*	*19.68*	21.67	23.59	27.88
	10	11.78	13.44	15.99	18.31	21.16	23.21	25.19	29.59
	11	12.90	14.63	17.28	19.68	22.62	24.72	26.76	31.26

Finding areas under the chi-square curve (one more)

Estimate the shaded area (above 30) under the χ^2 curve with $df = 10$.



(a) greater than 0.3

(b) between 0.005 and 0.001

(c) less than 0.001

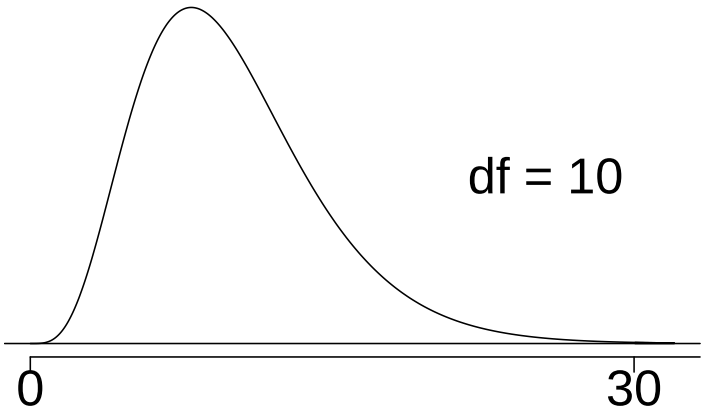
(d) greater than 0.001

(e) cannot tell using this table

Upper tail		0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
df	7	8.38	9.80	12.02	14.07	16.62	18.48	20.28	24.32
	8	9.52	11.03	13.36	15.51	18.17	20.09	21.95	26.12
	9	10.66	12.24	14.68	16.92	19.68	21.67	23.59	27.88
	10	11.78	13.44	15.99	18.31	21.16	23.21	25.19	29.59
	11	12.90	14.63	17.28	19.68	22.62	24.72	26.76	31.26

Finding areas under the chi-square curve (one more)

Estimate the shaded area (above 30) under the χ^2 curve with $df = 10$.



(a) greater than 0.3

(b) between 0.005 and 0.001

(c) *less than 0.001*

(d) greater than 0.001

(e) cannot tell using this table

Upper tail		0.3	0.2	0.1	0.05	0.02	0.01	0.005	*0.001*	→
df	7	8.38	9.80	12.02	14.07	16.62	18.48	20.28	24.32	
	8	9.52	11.03	13.36	15.51	18.17	20.09	21.95	26.12	
	9	10.66	12.24	14.68	16.92	19.68	21.67	23.59	27.88	
	10	11.78	13.44	15.99	18.31	21.16	23.21	25.19	*29.59*	→
	11	12.90	14.63	17.28	19.68	22.62	24.72	26.76	31.26	

Finding the tail areas using computation

- While probability tables are very helpful in understanding how probability distributions work, and provide quick reference when computational resources are not available, they are somewhat archaic.
- Using R:

```
pchisq(q = 30, df = 10, lower.tail = FALSE)
# 0.0008566412
```

- Using a web applet:
  http://bitly.com/dist_calc

Back to Labby's dice

- The research question was: Do these data provide convincing evidence of an inconsistency between the observed and expected counts?
- The hypotheses were:

  H_0: There is no inconsistency between the observed and the expected counts. The observed counts follow the same distribution as the expected counts.

  H_A: There is an inconsistency between the observed and the expected counts. The observed counts *do not* follow the same distribution as the expected counts. There is a bias in which side comes up on the roll of a die.
- We had calculated a test statistic of $\chi^2 = 24.67$.
- All we need is the df and we can calculate the tail area (the p-value) and make a decision on the hypotheses.

Degrees of freedom for a goodness of fit test

- When conducting a goodness of fit test to evaluate how well the observed data follow an expected distribution, the degrees of freedom are calculated as the number of cells (k) minus 1.

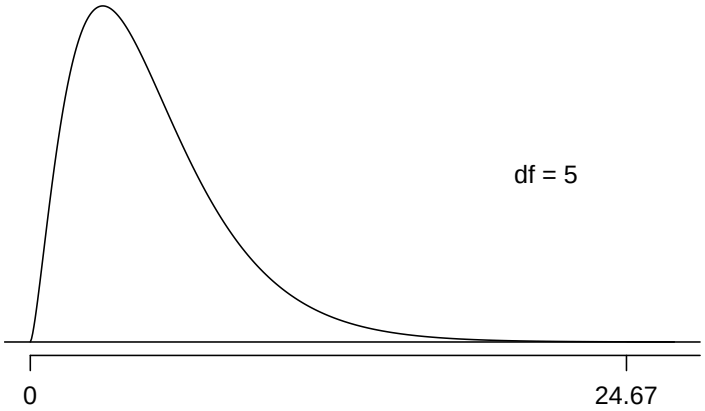
$$df = k - 1$$

- For dice outcomes, $k = 6$, therefore

$$df = 6 - 1 = 5$$

Finding a p-value for a chi-square test

The *p-value* for a chi-square test is defined as the *tail area above the calculated test statistic*.



p-value = $P(\chi^2_{df=5} > 24.67)$
is less than 0.001

Upper tail		0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001	→
df	1	1.07	1.64	2.71	3.84	5.41	6.63	7.88	10.83	
	2	2.41	3.22	4.61	5.99	7.82	9.21	10.60	13.82	
	3	3.66	4.64	6.25	7.81	9.84	11.34	12.84	16.27	
	4	4.88	5.99	7.78	9.49	11.67	13.28	14.86	18.47	
	5	6.06	7.29	9.24	11.07	13.39	15.09	16.75	20.52	→

Conclusion of the hypothesis test

We calculated a p-value less than 0.001. At 5% significance level, what is the conclusion of the hypothesis test?

(a) Reject H_0, the data provide convincing evidence that the dice are fair.

(b) Reject H_0, the data provide convincing evidence that the dice are biased.

(c) Fail to reject H_0, the data provide convincing evidence that the dice are fair.

(d) Fail to reject H_0, the data provide convincing evidence that the dice are biased.

Conclusion of the hypothesis test

We calculated a p-value less than 0.001. At 5% significance level, what is the conclusion of the hypothesis test?

(a) Reject H_0, the data provide convincing evidence that the dice are fair.

(b) *Reject H_0, the data provide convincing evidence that the dice are biased.*

(c) Fail to reject H_0, the data provide convincing evidence that the dice are fair.

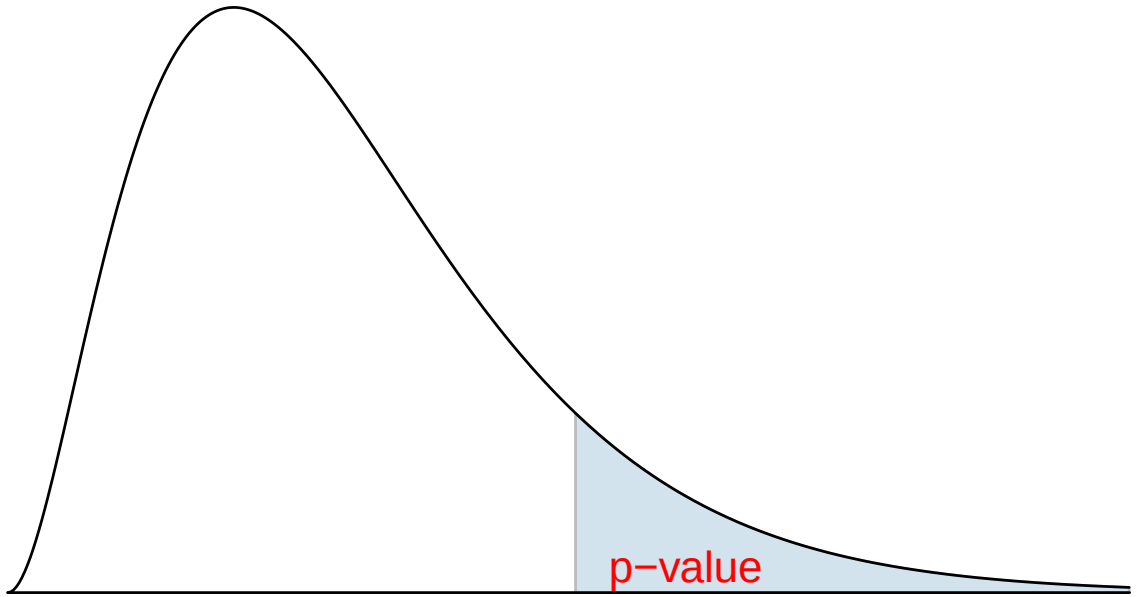
(d) Fail to reject H_0, the data provide convincing evidence that the dice are biased.

Turns out...

- The 1-6 axis is consistently shorter than the other two (2-5 and 3-4), thereby supporting the hypothesis that the faces with one and six pips are larger than the other faces.
- Pearson's claim that 5s and 6s appear more often due to the carved-out pips is not supported by these data.
- Dice used in casinos have flush faces, where the pips are filled in with a plastic of the same density as the surrounding material and are precisely balanced.

Recap: p-value for a chi-square test

- The p-value for a chi-square test is defined as the tail area *above* the calculated test statistic.
- This is because the test statistic is always positive, and a higher test statistic means a stronger deviation from the null hypothesis.



Conditions for the chi-square test

1. *Independence:* Each case that contributes a count to the table must be independent of all the other cases in the table.
2. *Sample size:* Each particular scenario (i.e. cell) must have at least 5 *expected* cases.
3. *df > 1:* Degrees of freedom must be greater than 1.

Failing to check conditions may unintentionally affect the test's error rates.

2009 Iran Election

There was lots of talk of election fraud in the 2009 Iran election. We'll compare the data from a poll conducted before the election (observed data) to the reported votes in the election to see if the two follow the same distribution.

Candidate	Observed # of voters in poll	Reported % of votes in election
(1) Ahmedinajad	338	63.29%
(2) Mousavi	136	34.10%
(3) Minor candidates	30	2.61%
Total	504	100%
	↓ *observed*	↓ *expected distribution*

Hypotheses

What are the hypotheses for testing if the distributions of reported and polled votes are different?

Hypotheses

What are the hypotheses for testing if the distributions of reported and polled votes are different?

H_0: *The observed counts from the poll follow the same distribution as the reported votes.*

H_A: *The observed counts from the poll do not follow the same distribution as the reported votes.*

Calculation of the test statistic

Candidate	Observed # of voters in poll	Reported % of votes in election	Expected # of votes in poll
(1) Ahmedinajad	338	63.29%	$504 \times 0.6329 = 319$
(2) Mousavi	136	34.10%	$504 \times 0.3410 = 172$
(3) Minor candidates	30	2.61%	$504 \times 0.0261 = 13$
Total	504	100%	504

$$\frac{(O_1 - E_1)^2}{E_1} = \frac{(338 - 319)^2}{319} = 1.13$$

$$\frac{(O_2 - E_2)^2}{E_2} = \frac{(136 - 172)^2}{172} = 7.53$$

$$\frac{(O_2 - E_2)^2}{E_2} = \frac{(30 - 13)^2}{13} = 22.23$$

$$\chi^2_{df=3-1=2} = 30.89$$

Conclusion

Based on these calculations what is the conclusion of the hypothesis test?

(a) p-value is low, H_0 is rejected. The observed counts from the poll do not follow the same distribution as the reported votes.

(b) p-value is high, H_0 is not rejected. The observed counts from the poll follow the same distribution as the reported votes.

(c) p-value is low, H_0 is rejected. The observed counts from the poll follow the same distribution as the reported votes

(d) p-value is low, H_0 is not rejected. The observed counts from the poll do not follow the same distribution as the reported votes.

Conclusion

Based on these calculations what is the conclusion of the hypothesis test?

(a) *p-value is low, H_0 is rejected. The observed counts from the poll do not follow the same distribution as the reported votes.*

(b) p-value is high, H_0 is not rejected. The observed counts from the poll follow the same distribution as the reported votes.

(c) p-value is low, H_0 is rejected. The observed counts from the poll follow the same distribution as the reported votes

(d) p-value is low, H_0 is not rejected. The observed counts from the poll do not follow the same distribution as the reported votes.

Geometric Distribution

If the probability of a success in one trial is p and the probability of a failure is $1 - p$, then the probability of finding the first success in the n-th trial is given by

$$(1-p)^{n-1}p.$$

The mean (i.e. expected value), variance and standard deviation of this wait time are given by

$$\mu = \frac{1}{p}, \quad \sigma^2 = \frac{1-p}{p^2}, \quad \sigma = \sqrt{\frac{1-p}{p^2}}$$

S&P500 Stock Data (1990-2011)

– Evaluating whether a certain statistical model fits a data set

– Daily stock returns from S&P500 Stock Data (1990-2011) can be used to assess whether stock activity each day is independent of the stock's behavior on previous days

– This sounds like a very complex question, and it is, but a **chi-square test** can be used to study the problem

– We will label each day as *Up* or *Down* (*D*) depending on whether the market was up or down that day. For example, consider the following changes in price, their new labels of up and down, and then the number of days that must be observed before each *Up* day:

Change in price	2.52	-1.46	0.51	-4.07	3.36	1.10	-5.46	-1.03	-2.99	1.71
Outcome	Up	D	Up	D	Up	Up	D	D	D	Up
Days to Up	1	-	2	-	2	1	-	-	-	4

– If the days really are independent, then the number of days until a positive trading day should follow a geometric distribution.

– The geometric distribution describes the probability of waiting for the k^{th} trial to observe the first success.

– Here each up day (Up) represents a success, and down (D) days represent failures. In the data above, it took only one day until the market was up, so the first wait time was 1 day. It took two more days before we observed our next *Up* trading day, and two more for the third *Up* day. We would like to determine if these counts (1, 2, 2, 1, 4, and so on) follow the geometric distribution.

– Table below shows the number of waiting days for a positive trading day during 1990-2011 for the S&P500.

Days	1	2	3	4	5	6	7+	Total
Observed	1532	760	338	194	74	33	17	2948

Table: Observed distribution of the waiting time until a positive trading day for the S&P500, 1990-2011.

We consider how many days one must wait until observing an *Up* day on the S&P500 stock index. If the stock activity was independent from one day to the next and the probability of a positive trading day was constant, then we would expect this waiting time to follow a geometric distribution. We can organize this into a hypothesis framework:

H_0: The stock market being up or down on a given day is independent from all other days. We will consider the number of days that pass until an *Up* day is observed. Under this hypothesis, the number of days until an *Up* day should follow a geometric distribution.

H_A: The stock market being up or down on a given day is not independent from all other days. Since we know the number of days until an *Up* day would follow a geometric distribution under the null, we look for deviations from the geometric distribution, which would support the alternative hypothesis.

Observed Data v.s. Model

There are important implications in our result for stock traders: if information from past trading days is useful in telling what will happen today, that information may provide an advantage over other traders.
The S&P500 was positive on 53.2% of those days.

Days	1	2	3	4	5	6	7+	Total
Observed	1532	760	338	194	74	33	17	2948
Geometric Model	1569	734	343	161	75	35	31	2948

Table: Distribution of the waiting time until a positive trading day. The expected counts based on the geometric model are shown in the last row.

To find each expected count, we identify the probability of waiting D days based on the geometric model ($P(D) = (1-0.532)^{D-1}(0.532)$) and multiply by the total number of streaks, 2948. For example, waiting for three days occurs under the geometric model about $0.468^2 \times 0.532 = 11.65\%$ of the time, which corresponds to $0.1165 \times 2948 = 343$.

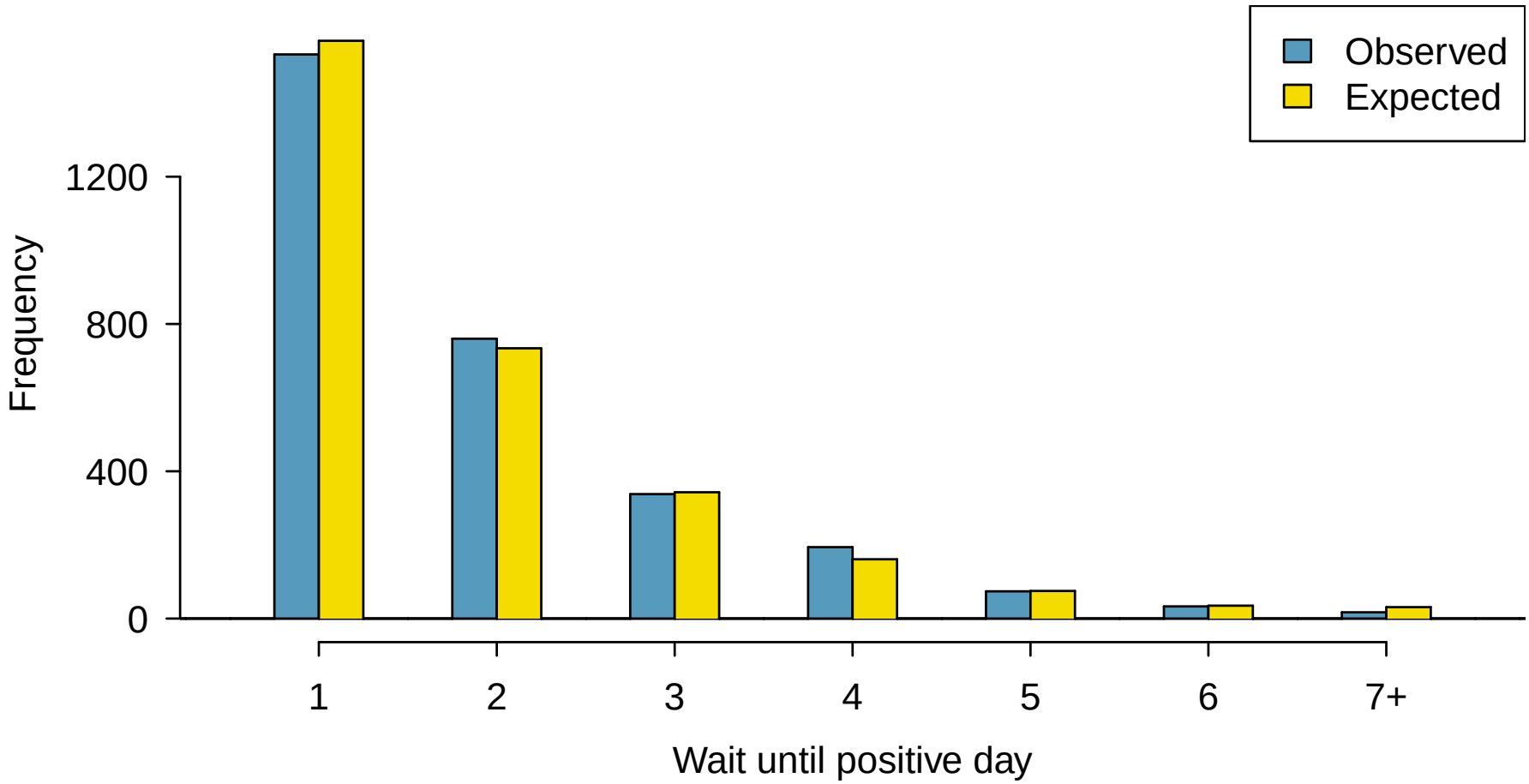


Figure: Side-by-side bar plot of the observed and expected counts for each waiting time.

Check Condition

1. Because applying the chi-square framework requires expected counts to be at least 5, we have binned together all the cases where the waiting time was at least 7 days to ensure each expected count is well above this minimum.
2. The actual data can be compared to the expected counts from the Geometric Model row. In general, the expected counts are determined by
   1. identifying the null proportion associated with each bin, then
   2. multiplying each null proportion by the total count to obtain the expected counts
3. That is, this strategy identifies what proportion of the total count we would expect to be in each bin.

Chi-square Test

– **Question:** Do you notice any unusually large deviations in the graph? Can you tell if these deviations are due to chance just by looking?

– **Answer:** It is not obvious whether differences in the observed counts and the expected counts from the geometric distribution are significantly different. That is, it is not clear **whether these deviations might be due to chance** or **whether they are so strong that the data provide convincing evidence against the null hypothesis**. However, we can perform a chi-square test using the counts in the table.

– **Computing the chi-square test statistic:** The table provides a set of count data for waiting times ($O_1 = 1532$, $O_2 = 760$, ...) and expected counts under the geometric distribution ($E_1 = 1569$, $E_2 = 734$, ...). Compute the chi-square test statistic, χ^2.

$$\chi^2 = \frac{(1532-1569)^2}{1569} + \frac{(760-734)^2}{734} + \cdots + \frac{(17-31)^2}{31} = 15.08$$

– **Degrees of freedom:** Because the expected counts are all at least 5, we can safely apply the chi-square distribution to χ^2. However, how many degrees of freedom should we use? There are $k = 7$ groups, so we use $df = k - 1 = 6$.

– **Compute the p-value:** If the observed counts follow the geometric model, then the chi-square test statistic $\chi^2 = 15.08$ would closely follow a chi-square distribution with $df = 6$. Using this information, compute a p-value.

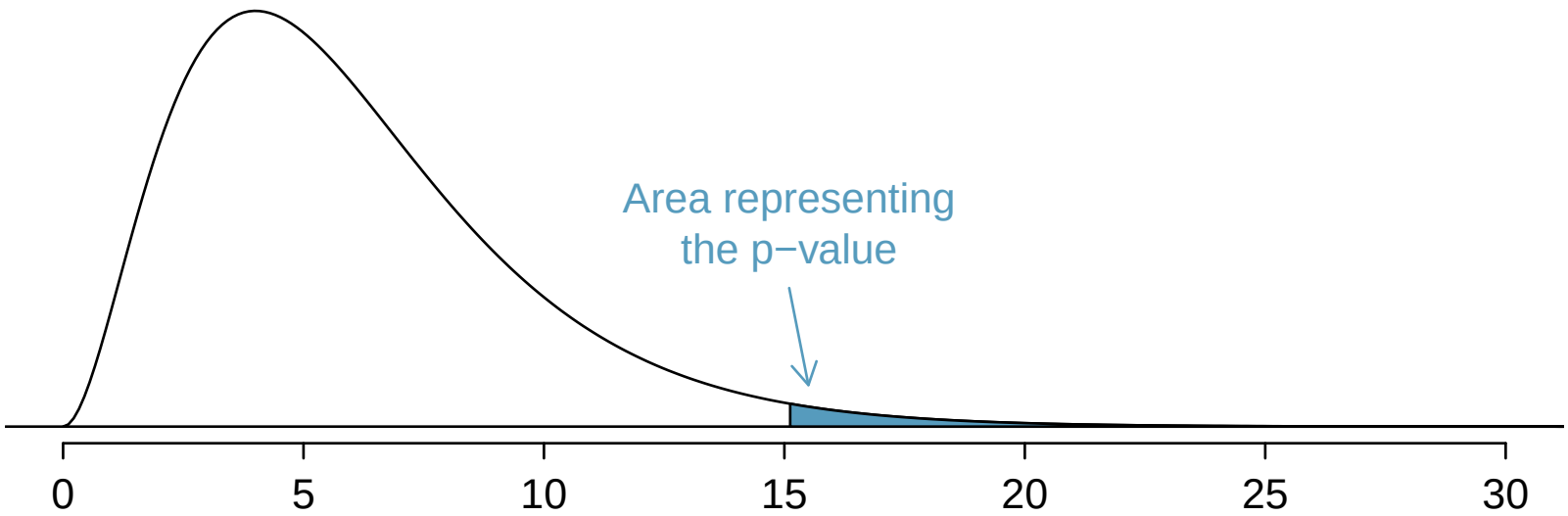


Figure: Chi-square distribution with 6 degrees of freedom. The p-value for the stock analysis is shaded.

– If we look up the statistic $\chi^2 = 15.08$ in the Chi-square probability table, we find that the p-value is between 0.01 and 0.02.

– In other words, we have sufficient evidence to reject the notion that the wait times follow a geometric distribution, i.e. trading days are not independent and past days may help predict what the stock market will do today.

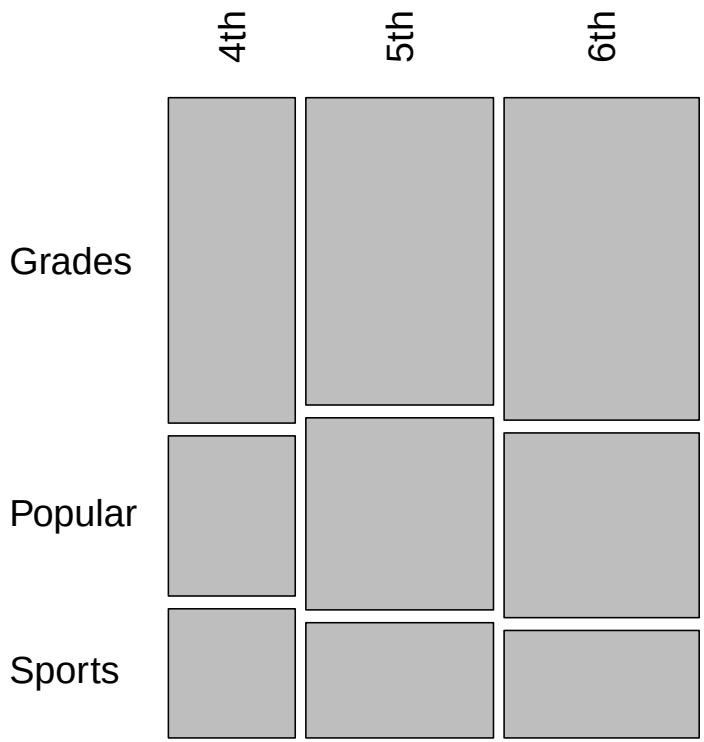
– We rejected the null hypothesis that the trading days are independent. Why is this so important? Because the data provided strong evidence that the geometric distribution is not appropriate, we reject the claim that trading days are independent.

– While it is not obvious how to exploit this information, it suggests there are some hidden patterns in the data that could be interesting and possibly useful to a stock trader.

Popular kids

In the dataset `popular`, students in grades 4-6 were asked whether good grades, athletic ability, or popularity was most important to them. A two-way table separating the students by grade and by choice of most important factor is shown below. Do these data provide evidence to suggest that goals vary by grade?

	Grades	Popular	Sports
4^{th}	63	31	25
5^{th}	88	55	33
6^{th}	96	55	32



Chi-square test of independence

- The hypotheses are:

  H_0: Grade and goals are independent. Goals do not vary by grade.

  H_A: Grade and goals are dependent. Goals vary by grade.

- The test statistic is calculated as

$$\chi^2_{df} = \sum_{i=1}^{k} \frac{(O-E)^2}{E} \quad \text{where} \quad df = (R-1) \times (C-1),$$

  where k is the number of cells, R is the number of rows, and C is the number of columns.

Note: We calculate df differently for one-way and two-way tables.

- The p-value is the area under the χ^2_{df} curve, above the calculated test statistic.

Expected counts in two-way tables

Expected counts in two-way tables

$$\text{Expected Count} = \frac{(\text{row total}) \times (\text{column total})}{\text{table total}}$$

	Grades	Popular	Sports	Total
4^{th}	*63*	*31*	25	119
5^{th}	88	55	33	176
6^{th}	96	55	32	183
Total	247	141	90	478

$$E_{row\ 1, col\ 1} = \frac{119 \times 247}{478} = 61 \qquad E_{row\ 1, col\ 2} = \frac{119 \times 141}{478} = 35$$

Expected counts in two-way tables

What is the expected count for the highlighted cell?

	Grades	Popular	Sports	Total
4^{th}	63	31	25	119
5^{th}	88	*55*	33	176
6^{th}	96	55	32	183
Total	247	141	90	478

(a) $\frac{176 \times 141}{478}$

(b) $\frac{119 \times 141}{478}$

(c) $\frac{176 \times 247}{478}$

(d) $\frac{176 \times 478}{478}$

Expected counts in two-way tables

What is the expected count for the highlighted cell?

	Grades	Popular	Sports	Total
4^{th}	63	31	25	119
5^{th}	88	*55*	33	176
6^{th}	96	55	32	183
Total	247	141	90	478

(a) $\frac{176 \times 141}{478}$

(b) $\frac{119 \times 141}{478}$

(c) $\frac{176 \times 247}{478}$

(d) $\frac{176 \times 478}{478}$

$\rightarrow$ *52*

more than expected # of 5th graders have a goal of being popular

Calculating the test statistic in two-way tables

Expected counts are shown in *blue* next to the observed counts.

	Grades	Popular	Sports	Total
4^{th}	63 *61*	31 *35*	25 *23*	119
5^{th}	88 *91*	55 *52*	33 *33*	176
6^{th}	96 *95*	55 *54*	32 *34*	183
Total	247	141	90	478

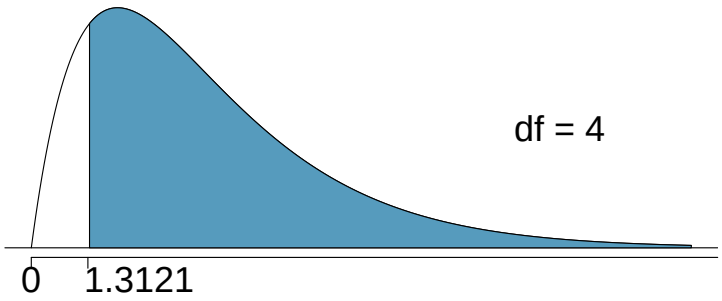
$$\chi^2 = \sum \frac{(63-61)^2}{61} + \frac{(31-35)^2}{35} + \cdots + \frac{(32-34)^2}{34} = 1.3121$$

$$df = (R-1)\times(C-1) = (3-1)\times(3-1) = 2\times 2 = 4$$

Calculating the p-value

Which of the following is the correct p-value for this hypothesis test?

$$\chi^2 = 1.3121 \qquad df = 4$$



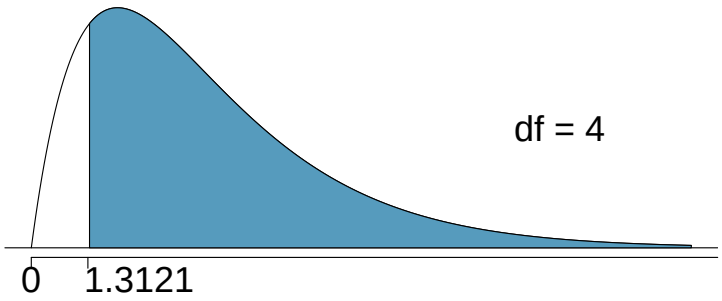
(a) more than 0.3
(b) between 0.3 and 0.2
(c) between 0.2 and 0.1
(d) between 0.1 and 0.05
(e) less than 0.001

Upper tail		0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
df	1	1.07	1.64	2.71	3.84	5.41	6.63	7.88	10.83
	2	2.41	3.22	4.61	5.99	7.82	9.21	10.60	13.82
	3	3.66	4.64	6.25	7.81	9.84	11.34	12.84	16.27
	4	4.88	5.99	7.78	9.49	11.67	13.28	14.86	18.47
	5	6.06	7.29	9.24	11.07	13.39	15.09	16.75	20.52

Calculating the p-value

Which of the following is the correct p-value for this hypothesis test?

$$\chi^2 = 1.3121 \qquad df = 4$$



(a) *more than 0.3*

(b) between 0.3 and 0.2

(c) between 0.2 and 0.1

(d) between 0.1 and 0.05

(e) less than 0.001

Upper tail		0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
df	1	1.07	1.64	2.71	3.84	5.41	6.63	7.88	10.83
	2	2.41	3.22	4.61	5.99	7.82	9.21	10.60	13.82
	3	3.66	4.64	6.25	7.81	9.84	11.34	12.84	16.27
	4	4.88	5.99	7.78	9.49	11.67	13.28	14.86	18.47
	5	6.06	7.29	9.24	11.07	13.39	15.09	16.75	20.52

Conclusion

Do these data provide evidence to suggest that goals vary by grade?

H_0: Grade and goals are independent. Goals do not vary by grade.

H_A: Grade and goals are dependent. Goals vary by grade.

Conclusion

Do these data provide evidence to suggest that goals vary by grade?

H_0: Grade and goals are independent. Goals do not vary by grade.

H_A: Grade and goals are dependent. Goals vary by grade.

Since p-value is high, we fail to reject H_0. The data do not provide convincing evidence that grade and goals are dependent. It doesn't appear that goals vary by grade.

Example 2: Suppose one took U.S. states and classified them as to whether they supported Romney or Obama, and how many executions each state had in the last five years (e.g., 0, 1-5, more than 5). You might get a **contingency table** that looks like this:

	Obama	Romney	
0	20	1	21
1-5	5	8	13
>5	2	14	16
	27	23	50

Here there are 20 states that supported Obama and had no executions, 1 state that supported Romney and had no executions, and so forth.

Usually, we are only interested in testing if there is some relationship or dependence between the two categorical variables.

Thus we can set up the general null and alternative hypotheses as:

H_0: The two variables are independent.
H_A: Some dependence exists between them.

For a given situation, it is always better to be clear and specific to the context of the problem. For the previous example, the hypotheses are:

H_0: Voting preference has nothing to do with execution rates.
H_A: There is a relationship between voting choice and executions.

Unlike before, there is only one choice for the null and alternative hypothesis. But as with all of our hypothesis tests, there are three parts. We now we need to get a test statistic and a critical value/p-value.

Test Statistic

The test statistic is

$$ts = \sum_{\textbf{all cells}} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

The O_{ij} is the observed count for the cell in row i, column j.

The E_{ij} uses the following formula:

$$E_{ij} = \frac{(\textbf{ith row sum}) * (\textbf{jth column sum})}{\textbf{total}}$$

Example 2

For example 2, we find:

$$\begin{aligned} E_{11} &= 21 * 27/50 = 11.34 \\ E_{12} &= 21 * 23/50 = 9.66 \\ E_{21} &= 27 * 13/50 = 7.02 \\ E_{22} &= 23 * 13/50 = 5.98 \\ E_{31} &= 27 * 16/50 = 8.64 \\ E_{32} &= 23 * 16/50 = 7.36 \end{aligned}$$

Then the test statistic is:

$$\begin{aligned} ts &= \frac{(20-11.34)^2}{11.34} + \frac{(1-9.66)^2}{9.66} + \ldots + \frac{(14-7.36)^2}{7.36} \\ &= 26.734. \end{aligned}$$

P-value

We compare the test statistic to the value from a chi-squared distribution with degrees of freedom equal to

$$k = \textbf{(number of rows - 1)} * \textbf{(number of columns - 1)}.$$

For our example, $k = (3-1) * (2-1) = 2$.

The p-value is the chance of getting a chi-squared random variable greater than or equal to the observed test statistic, or

$$P\textbf{-value} = \mathbf{P}[W \geq ts]$$

where W has the chi-squared distribution with k degrees of freedom.

For a chi-squared random variable with 2 degrees of freedom, the table shows that the chance of getting a value bigger than 26.734 is less than 0.01.

So the p-value is much less than 0.01. We strongly reject the null hypothesis at even $\alpha = 0.01$. There is major evidence that political preference and execution rates are somehow connected.

But the connection can be very subtle. We cannot infer causation, and the apparent relationship may not be at all what we expect. For example, one might argue that voting preferences reflect economic hardship, and states with economic hardship experience more violent crime and thus use the death penalty more often.

Sometimes there are hidden confounders that are more interesting than the relationship between the two classification criteria. It can even happen that the hidden confounder can reverse the apparent relationship in the data. When this happens, it is called **Simpson's Paradox**. This is something we already talked about briefly.

Recap

Today we learned about chi-square **tests of goodness-of-fit** and **tests of independence**.

You should be able to set-up such testing problems, calculate p-values, and make the appropriate conclusions.

Suggested reading:

- D.S. Sec. 10.1, 10.2, 10.3
- OpenIntro3: Sec. 3.3.2, 6.3, 6.4