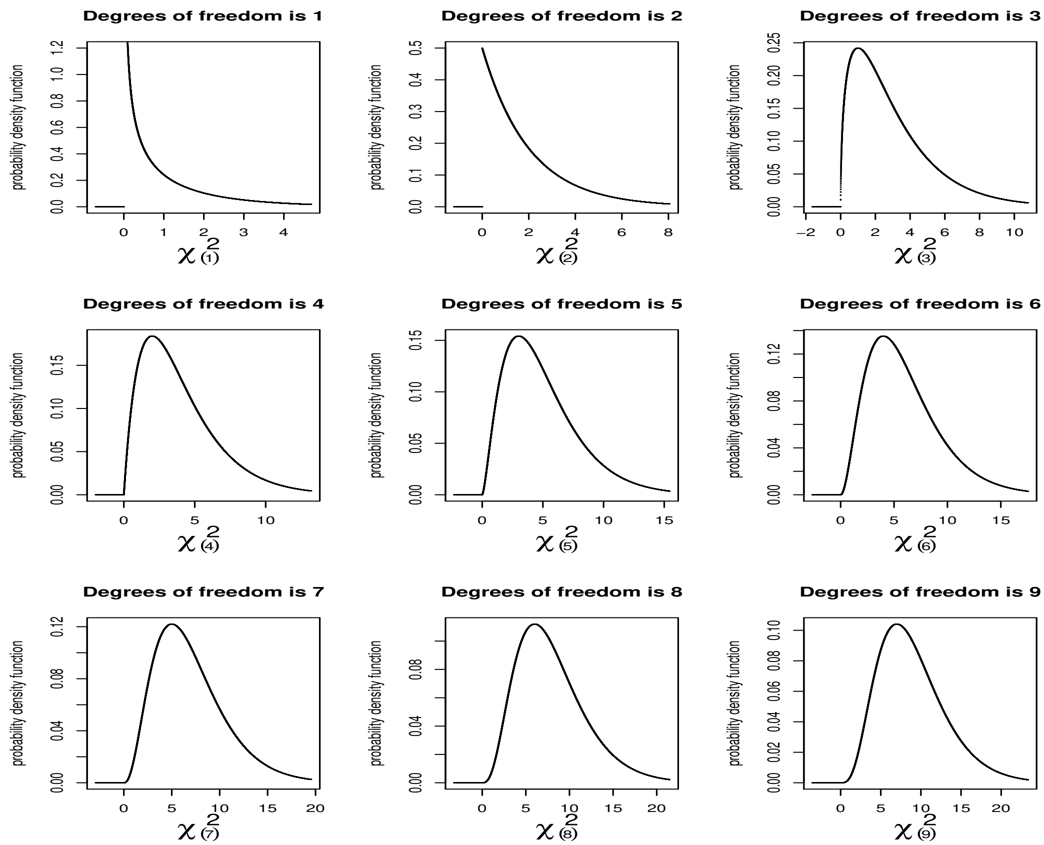
10 Tests with Qualitative Data

The Chi-Square (χ^2) Distribution

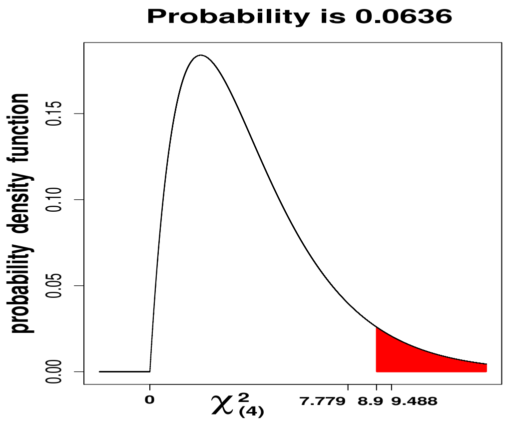
Below are the probability density functions of the χ^2 distribution with degrees of freedom equal to **1**, **2**, **3**, **4**, **5**, **6**, **7**, **8**, and **9**.



Example:

(a) Compute $P(\chi^2_4 > 8.9)$.

(b) Compute $P(\chi^2_4 > 3.8)$.



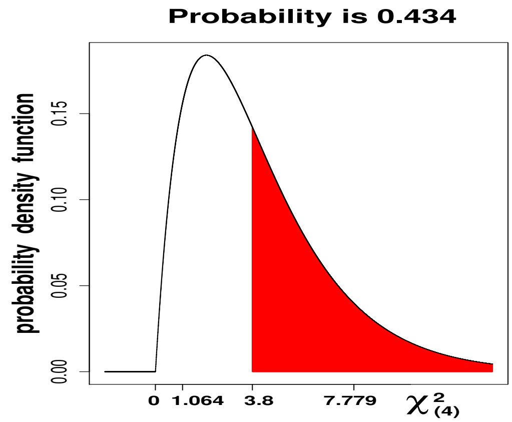
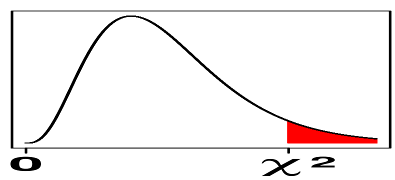


Table A.4 Critical Values for the $\chi 2$ Distribution, p. A-9



Degrees of Freedom	Area in Right Tail 0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

□

10.2 Tests for Independence and Homogeneity

In this chapter, all observations are independent or nearly independent under the null hypothesis.

Comparing Percentages

Example: The *Titanic* collided with an iceberg April 14, 1912.

The data below are from the *British Board of Trade Inquiry Report* (1990), written originally on July 30, 1912. Did the different classes of passengers have equal chances of survival?

Observed table	First	Second	Third	Crew	total
Alive	203	118	178	212	711
Dead	122	167	528	673	1490
total	325	285	706	885	2201

Determine the **conditional probabilities** of **survival**, **given** the **class** of the passenger.

(a) Determine the probability that a randomly selected passenger **survived**, **given** that the passenger was **first**-class. Alternatively, determine the proportion of **first**-class passengers who **survived**.

Let $S = \{\text{Passenger survived}\}$ and $D = \{\text{Passenger died}\}$.

(b) Determine the probability that a randomly selected passenger **survived**, **given** that the passenger was **second**-class. Alternatively, determine the proportion of **second**-class passengers who **survived**.

(c) Determine the probability that a randomly selected passenger **survived**, **given** that the passenger was **third**-class. Alternatively, determine the proportion of **third**-class passengers who **survived**.

(d) Determine the probability that a randomly selected passenger **survived**, **given** that the passenger was a member of the **crew**. Alternatively, determine the proportion of **crew**-members who **survived**.

(e) Do the **class** of the passenger and the **survival** status seem to be **independent** or **dependent**? In other words, do the discrepancies among the answers to parts (a), (b), (c), and (d) seem to be caused by **random chance**, or do the discrepancies seem to be caused by discrimination among the four **classes** of passengers?

(f) Determine the probability that a randomly selected passenger **died**, **given** that the passenger was **first**-class. Alternatively, determine the proportion of **first**-class passengers who **died**.

(g) Determine the probability that a randomly selected passenger **survived**. Alternatively, determine the proportion of passengers who **survived**.

(h) Determine the probability that a randomly selected passenger **died**.

Alternatively, determine the proportion of passengers who **died**.

□

Testing Whether Categorical Variables Are Independent or Dependent

What Do We Expect for Cell Counts If the Variables Are Independent?

Example: *Revisit the Titanic.*

(a) How many **first**-class passengers do we **expect** to have **survived** if the **class** and the **survival status** of the passengers were **independent**?

(b) How many **first**-class passengers do we **expect** to have **died** if the **class** and the **survival status** of the passengers were **independent**?

(c) How many **second**-class passengers do we **expect** to have **survived** if the **class** and the **survival status** of the passengers were **independent**?

(d) What is the formula for the **expected count** under the assumption of **independence** between the **class** and the **survival status** of the passengers?

(e) Complete the table below for **expected counts** under H_0 (i.e., "The class and survival status of the passengers were **independent**.").

Expected table under H_0	First	Second	Third	Crew	total
Alive					
Dead					
total					

□

The **chi-squared statistic** is defined as

$$\chi^2 = \sum \frac{(\text{observed count} \; - \; \text{expected count})^2}{\text{expected count}}$$

The statistic $\boldsymbol{\chi^2}$ has a distribution which is approximated **chi-squared**, denoted $\boldsymbol{\chi^2}$, with **degrees of freedom** $=$ [(Number of rows) $-$ 1] $\times$ [(Number of columns) $-$ 1], if the **expected count** is **at least 5** in all cells under the assumption that the two variables are **independent**.

Example: *Revisit the Titanic.* Test at level 0.01 if the **class** and the **survival status** of the passengers were **dependent**, according to the following steps.

(a) State the null and alternative hypotheses.

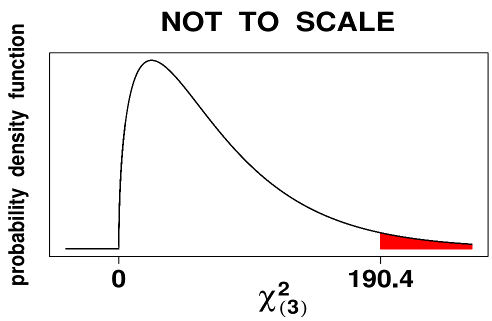
(b) Determine the value of the **test statistic**.

$$\chi^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}} = \frac{(203-105.0)^2}{105.0} + \frac{(118-92.1)^2}{92.1}$$
$$+\frac{(178-228.1)^2}{228.1} + \frac{(212-285.9)^2}{285.9} + \frac{(122-220.0)^2}{220.0} + \frac{(167-192.9)^2}{192.9} + \frac{(528-477.9)^2}{477.9} + \frac{(673-599.1)^2}{599.1}$$
$$= 91.50 + 7.31 + 10.99 + 19.10 + 43.66 + 3.49 + 5.24 + 9.11 = 190.4$$

(c) Determine the **degrees of freedom**.

(d) Is the chi-squared approximation valid?

(e) Determine the P-value.



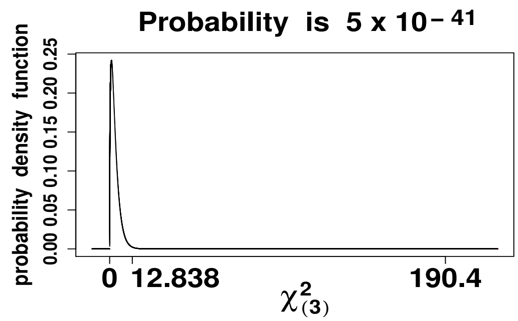
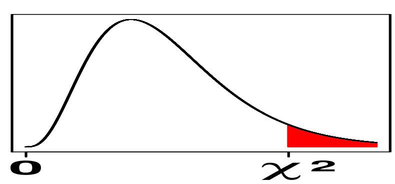


Table A.4 Critical Values for the χ2 Distribution, p. A-9



Degrees of Freedom	Area in Right Tail 0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

(f) State the conclusion in statistical terms and in regular English.

□

Remark: The **rows** and **columns** are interchangeable.

□

Example: Are **tattoos** and **hepatitis C** infections **dependent**? Test at level $\alpha = 0.01$.

Hepatitis C is a potentially fatal disease that attacks the liver, and causes 10,000 to 20,000 deaths in the U.S. each year from cirrhosis and liver cancer. Hepatitis C affects an estimated 4 million people in the U.S. The following data are based on a study in the University of Texas Southwestern Medical Center at Dallas. In 1991–1992, 626 participants in the study were patients of an orthopedic spinal clinic, a setting that provided a large volume of patients seeing a physician for reasons unrelated to blood-born infection. Participants unaware of their hepatitis status were examined, interviewed for risk factors, and tested for hepatitis C. Below is the **observed** table.

Population #	status	hepatitis C	no hepatitis C	total
1	tattoo	25	88	113
		( )	( )	
2	no tattoo	22	491	513
		( )	( )	
	total	47	579	626

(a) Determine the proportion of people **with** tattoos who **have** hepatitis C.

(b) Determine the proportion of people **without** tattoos who **have** hepatitis C.

(c) State the null and alternative hypotheses.

(d) Determine the **expected count** of people with both tattoos and hepatitis C under H_0.

(e) In the above table, list the **expected counts** in the parentheses.

(f) Determine the value of the **test statistic.**

$$\chi^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$
$$= \frac{(25-8.48)^2}{8.48} + \frac{(88-104.52)^2}{104.52} + \frac{(22-38.52)^2}{38.52} + \frac{(491-474.48)^2}{474.48}$$
$$= 32.152 + 2.610 + 7.082 + 0.575 = 42.42$$

(g) Determine the **degrees of freedom**.

(h) Is the chi-squared approximation valid?

(i) Determine the P-value.

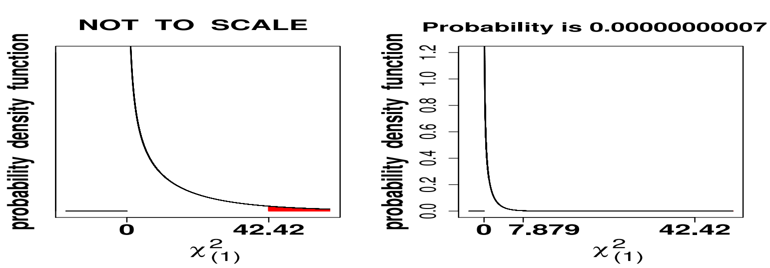
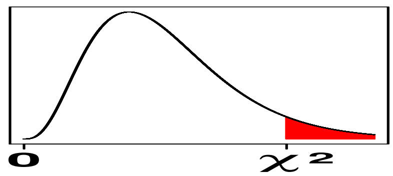


Table A.4 Critical Values for the $\chi 2$ Distribution, p. A-9



Degrees of Freedom	Area in Right Tail 0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

(j) State the conclusion in statistical terms and in regular English.

□

Remark: Association does NOT imply causation.

Remark: For a 2×2 table, the χ^2 test for independence produces the same P-value and same conclusion as a two-sided Z-test on the difference between two proportions, since $Z^2 = \chi_1^2$.

Example: *Revisit tattoos and hepatitis C.* Are **tattoos** and **hepatitis C** infections **dependent**? Test at level $\alpha = 0.01$ using the Z-test on the difference between two proportions.

Population #	status	hepatitis C	no hepatitis C	total
1	tattoo	25	88	113
2	no tattoo	22	491	513
	total	47	579	626

(a) Define your notation.

Let $\boldsymbol{p_1}$ be the *unknown* **population** proportion of people **with tattoos** who **have hepatitis C**.

Alternatively, let $\boldsymbol{p_1}$ be the *unknown* **population** proportion of **tattooed** people who **have hepatitis C**.

Let $\boldsymbol{p_2}$ be the *unknown* **population** proportion of people **withOUT tattoos** who **have hepatitis C**.

Alternatively, let $\boldsymbol{p_2}$ be the *unknown* **population** proportion of **NON-tattooed** people who **have hepatitis C**.

(b) State the null and alternative hypotheses in terms of your notation.

(c) Define your notation for the **samples**.

Let $\boldsymbol{\hat{p}_1}$ be the **sample** proportion of people **with tattoos** who **have hepatitis C**.

Alternatively, let $\boldsymbol{\hat{p}_1}$ be the **sample** proportion of **tattooed** people who **have hepatitis C**.

Let $\boldsymbol{\hat{p}_2}$ be the **sample** proportion of people **withOUT tattoos** who **have hepatitis C**.

Alternatively, let $\boldsymbol{\hat{p}_2}$ be the **sample** proportion of **NON-tattooed** people who **have hepatitis C**.

Let $\boldsymbol{\hat{p}}$ be the **pooled sample** proportion of **people** who **have hepatitis C**.

(d) Evaluate $\hat{\boldsymbol{p}}_{\mathbf{1}}$, $\hat{\boldsymbol{p}}_{\mathbf{2}}$, and $\hat{\boldsymbol{p}}$.

(e) Determine the **point estimate** of $(\boldsymbol{p}_{\mathbf{1}} - \boldsymbol{p}_{\mathbf{2}})$.

(f) Interpret the above **point estimate** in regular English.

We estimate that for 17.8% of patients, having a tattoo results in having hepatitis C instead of not having hepatitis C.

(g) Check the rule of thumb, and discuss any other necessary assumptions.

(h) Determine the value of the **standardized test statistic.**

(i) Determine the P-value.

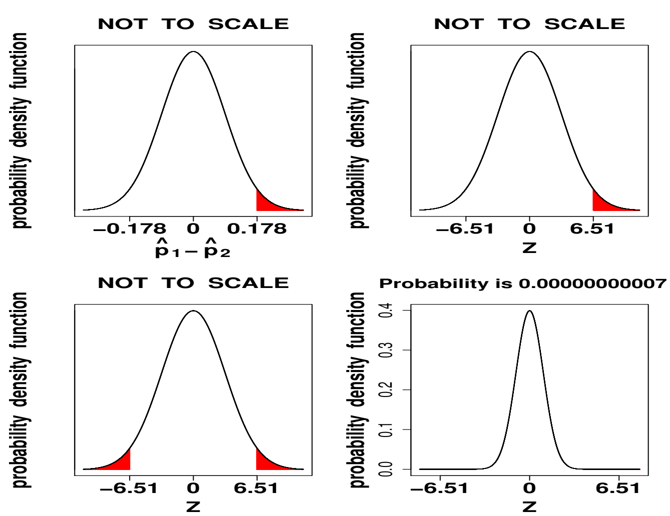
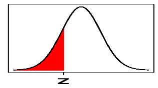
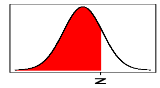


Table A.2 Cumulative Normal Distribution, pp. A-6 and A-7





z	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
-3.7 or less	.0001									
-3.6	.0002	.0002	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
-3.5	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

(j) State the conclusion in statistical terms and in regular English.

We conclude that the **population** proportion of people **with tattoos** who **have hepatitis C differs** from the **population** proportion of people **withOUT tattoos** who **have hepatitis C**.

□

Remark: The χ^2 test does **NOT** provide results for **one**-sided tests, whereas the $\boldsymbol{Z}$ test on the difference between two proportions **DOES** provide results for **one**-sided tests.

10.1 Testing Goodness of Fit

H_0 : The data are from a specified given population.

H_1 : The data are from a different population.

The test is called a **goodness-of-fit** test, since we are testing if the specified population provides a good fit to the data.

Example: *Hypothetical data.* Are the three sections of Spanish 101 (all taught at the same time) equally likely to be selected by the students? Test at level 0.05.

	Section #		
	1	2	3
Frequency	137	156	109

(a) Define your notation.

Let $\boldsymbol{p_1}$ be the probability that a Spanish 101 student will enroll in Section **#1** (or the population proportion of all Spanish 101 students who would enroll in Section **#1**).

Let $\boldsymbol{p_2}$ be the probability that a Spanish 101 student will enroll in Section **#2**.

Let $\boldsymbol{p_3}$ be the probability that a Spanish 101 student will enroll in Section **#3**.

(b) State the null and alternative hypotheses.

(c) Determine the **expected count** of students in Section **#1** under H_0.

(d) Determine the value of the **test statistic.**

(e) Determine the **degrees of freedom**.

(f) Is the chi-squared approximation valid?

(g) Determine the P-value.

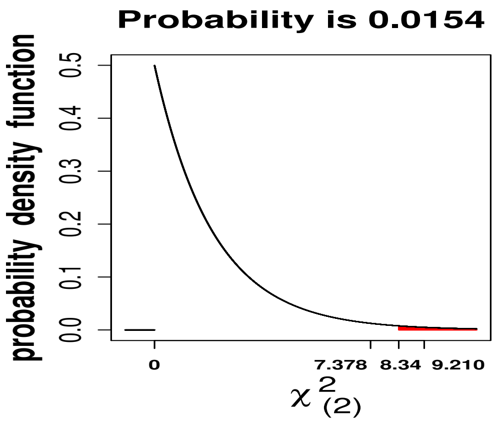
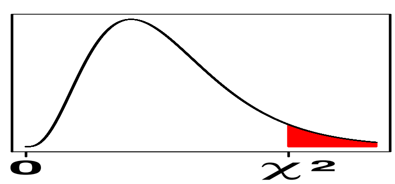


Table A.4 Critical Values for the $\chi 2$ Distribution, p. A-9



Degrees of Freedom	Area in Right Tail 0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

(h) State the conclusion in statistical terms and in regular English.

□

Example: The article "Linkage Studies of the Tomato" (*Transactions of the Royal Canadian Institute* [1931]: 1–19) reported the accompanying data on phenotypes resulting from crossing tall cut-leaf tomatoes with dwarf potato-leaf tomatoes. There are four possible phenotypes: (1) tall cut-leaf, (2) tall potato leaf, (3) dwarf cut-leaf, and (4) dwarf potato-leaf. Mendel's laws of inheritance imply that the population proportions should be 9/16, 3/16, 3/16, and 1/16 for groups 1, 2, 3, and 4, respectively. Test at level 0.05 if these data below are consistent with Mendel's laws of inheritance.

	Phenotype			
	1	2	3	4
Frequency	926	288	293	104

(a) Define your notation.

Let $\boldsymbol{p_1}$ be the probability that the tomato will be in group **#1**.
Let $\boldsymbol{p_2}$ be the probability that the tomato will be in group **#2**.
Let $\boldsymbol{p_3}$ be the probability that the tomato will be in group **#3**.
Let $\boldsymbol{p_4}$ be the probability that the tomato will be in group **#4**.

(b) State the null and alternative hypotheses.

(c) Determine the **expected count** of tomatoes in each of the four groups under H_0.

(d) Determine the value of the **test statistic.**

(e) Determine the **degrees of freedom**.

(f) Is the chi-squared approximation valid?

(g) Determine the P-value.

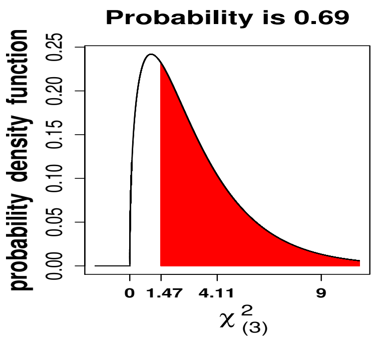
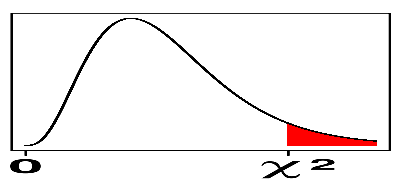


Table A.4 Critical Values for the $\chi 2$ Distribution, p. A-9



Degrees of Freedom	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
	Area in Right Tail									
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

(h) State the conclusion in statistical terms and in regular English.

□

Example: *How accurate were Germany's flying-bombs?* During World War II, Germany sent 537 flying-bombs in an area consisting of 144 km^2 of Southern London. This area was subdivided into 576 square **regions** of equal size. The table below summarizes the data. Did Germany drop the bombs in random **regions** of Southern London, or was Germany targeting specific **regions**? Test at level 0.05.

Number (x) of bomb hits per area	Number of **regions** with x bomb hits
0	229
1	211
2	93
3	35
4 or more	8
sum	576

(a) State the null and alternative hypotheses.

(b) Letting x be the number of flying-bombs to hit any one particular **region** of Southern London, show that x has the following probability distribution under H_0.

x	$P(x)$
0	0.3933
1	0.3673
2	0.1712
3	0.0531
4 or more	0.0150
sum	1

(c) Determine the **expected** number of **regions** to receive 0 bomb hits, 1 hit, 2 hits, 3 hits, and 4 or more hits, under H_0.

(d) Determine the value of the **test statistic.**

(e) Determine the **degrees of freedom**.

(f) Is the chi-squared approximation valid?

(g) Determine the P-value.

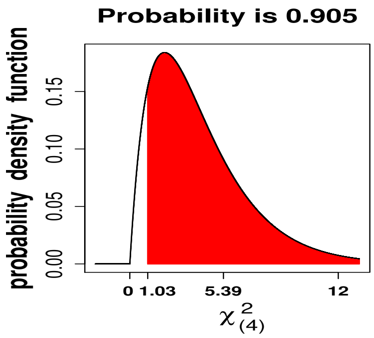
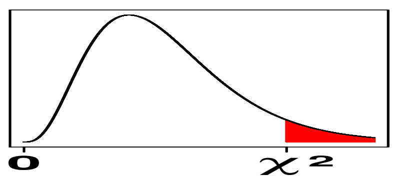


Table A.4 Critical Values for the $\chi 2$ Distribution, p. A-9



Degrees of Freedom	Area in Right Tail 0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

(h) State the conclusion in statistical terms and in regular English.

□