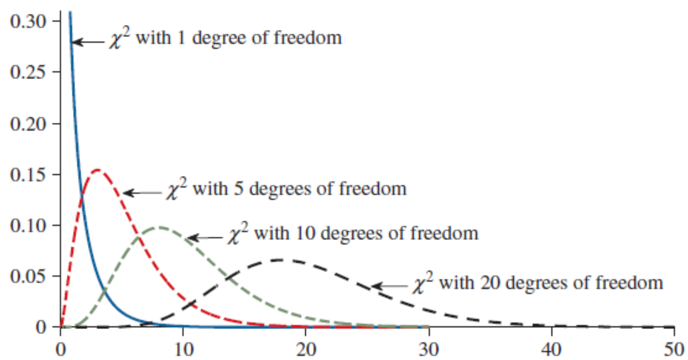
OBJECTIVES

1. Find critical values of the chi-square distribution
2. Perform goodness-of-fit tests

OBJECTIVE 1
FIND CRITICAL VALUES OF THE CHI-SQUARE DISTRIBUTION

THE CHI-SQUARE DISTRIBUTION

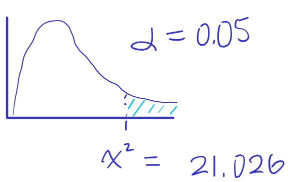
Hypothesis tests for qualitative data, also called categorical data, are based on the **chi-square distribution**. The test statistic used for these tests is called the **chi-square statistic**, denoted χ^2. There are actually many different chi-square distributions, each with a different number of degrees of freedom. The figure below shows several examples of chi-square distributions.



Recall that the chi-square distributions are skewed to the right, and the values of the χ^2 statistic are always greater than or equal to 0. They are never negative.

EXAMPLE: Find the α = 0.05 critical value for the chi-square distribution with 12 degrees of freedom.

SOLUTION:



df	Area = 0.05
1	↓
⋮	
12	→ $\chi^2 = 21.026$

OBJECTIVE 2
PERFORM GOODNESS-OF-FIT TESTS

Imagine that a gambler wants to test a die to determine whether it is fair. The roll of a die has six possible outcomes: 1, 2, 3, 4, 5, and 6; and the die is fair if each of these outcomes is equally likely. The gambler rolls the die 60 times, and counts the number of times each number comes up. The results are presented below. These counts are called the **observed frequencies**.

Outcome	**1**	**2**	**3**	**4**	**5**	**6**
Observed	12	7	14	15	4	8

The null hypothesis for this test says that the die is fair; in other words, it says that each of the six outcomes has probability 1/6 of occurring. Let p_1 be the probability of rolling a 1, p_2 be the probability of rolling a 2, and so on. Then the null hypothesis is

$$H_0: p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = 1/6$$

The alternate hypothesis says that the roll of a die does not follow the distribution specified by H_0; in other words, it states that not all of the p_i are equal to 1/6.

binomial dist

EXPECTED FREQUENCIES

$\mu = E(x) = n \cdot p$

To test the gambler's null hypothesis, H_0, we begin by computing expected frequencies. The **expected frequencies** are the mean counts that would occur if H_0 were true.
If the probabilities specified by H_0 are p_1, p_2, ..., and the total number of trials is n, the expected frequencies are

$$E_1 = np_1, \; E_2 = np_2, \text{ and so on}$$

EXAMPLE: Compute the expected frequencies for the gambler rolling a die.

SOLUTION:

Outcome	1	2	3	4	5	6
Observed	12	7	14	15	4	8
Expected	10	10	10	10	10	10

$n = 60$

$E_1 = n \cdot p_1 = 60(\frac{1}{6}) = 10$

$\vdots \qquad \vdots$

$E_6 = n \cdot p_6 = 60(\frac{1}{6}) = 10$

THE CHI-SQUARE STATISTIC

The statistic that measures how large the differences are between the observed and expected frequencies is called the **chi-square statistic**.

Let k be the number of categories, let O_1, ..., O_k be the observed frequencies, and let E_1, ..., E_k be the expected frequencies. The **chi-square statistic** is:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

When H_0 is true, the chi-square statistic has approximately a chi-square distribution, provided that all the expected frequencies are 5 or more.

PERFORMING A GOODNESS-OF-FIT TEST

Step 1: State the null and alternate hypotheses.

Step 2: Compute the expected frequencies and check to be sure that all of them are 5 or more. If they are, then proceed.

Step 3: Choose a significance level α.

Step 4: Compute the test statistic: $\chi^2 = \sum \frac{(O-E)^2}{E}$ and the P-value using technology

Step 5: State a conclusion.

$df = k - 1$

k = # of categories

EXAMPLE: Using the observed frequencies for the die example, can you conclude at the $\alpha = 0.05$ level that the die is not fair?

Outcome	1	2	3	4	5	6
Observed	12	7	14	15	4	8

$k = 6$

SOLUTION: $H_0: P_1 = P_2 = P_3 \ldots = P_6 = 1/6$

H_1: some or all the P_i differ from $1/6$

$df = n - 1 = 5$

L1 observed	L2 expected
12	10
7	10
14	10
15	10
4	10
8	10

Using TI-84, STAT TEST D: χ^2 GOF-Test (L1, L2, df)

Test statistic: $\chi^2 = 9.4$

p-value = $p = 0.0941$

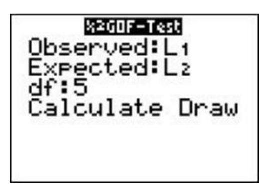
Since the p-value > α, we do not reject H_0

There is enough evidence to conclude that the die is fair

or there is not enough evidence to conclude that the die is not fair

HYPOTHESIS TESTS ON THE TI-84 PLUS

The **χ²-GOF-Test** command will perform a goodness of fit hypothesis test. This command is accessed by pressing **STAT** and highlighting the **TESTS** menu.



Enter the location of the observed frequencies in the **Observed** field and the location of the expected frequencies in the **Expected** field. Enter the number of degrees of freedom in the **df** field.

YOU SHOULD KNOW …

- How to find critical values of the chi-square distribution
- How to perform goodness-of-fit tests

OBJECTIVES

1. Interpret contingency tables
2. Perform tests of independence
3. Perform tests of homogeneity

OBJECTIVE 1
INTERPRET CONTINGENCY TABLES

CONTINGENCY TABLES

Do some college majors require more studying than others? The table below shows the results from the 2009 National Survey of Student Engagement survey where a sample of 1000 college freshmen were asked their major and the average number of hours per week spent studying. The table is called a **contingency table**.

Hours Studying Per Week	Major			
	Humanities	**Social Science**	**Business**	**Engineerin g**
0–10	68	106	131	40
11–20	119	103	127	81
More than 20	70	52	51	52

A contingency table relates two qualitative variables. One of the variables, called the **row variable**, has one category for each row of the table. The other variable, called the **column variable**, has one category for each column of the table. The intersection of a row and a column is called a **cell**. The total number of individuals in the table, 1000, is called the **grand total**.

OBJECTIVE 2
PERFORM TESTS OF INDEPENDENCE

DEPENDENT VERSUS INDEPENDENT

We are interested in determining whether the distribution of one variable differs, depending on the value of the other variable. If so, the variables are **dependent**. If the distribution of one variable is the same for all values of the other variable, the variables are **independent**.
In the survey about major and hours spent studying, if these variables are independent, then the distribution of hours studied will be the same for all majors, and the distribution of majors will be the same for all categories of hours studied.

Hours Studying Per Week	Major			
	Humanities	**Social Science**	**Business**	**Engineering**
0–10	68	106	131	40
11–20	119	103	127	81
More than 20	70	52	51	52

If we are to conduct a test to determine if the two variables in the previous survey about student majors and hours spent studying are independent, the null and alternate hypotheses are

H_0: Hours studying and major are independent
H_1: Hours studying and major are not independent

We will use the chi-square statistic to test the null hypothesis that hours studying and major are independent. If we reject H_0, we will conclude that the variables are dependent.

EXPECTED FREQUENCY

The null hypothesis states that hours studying and major are independent. By using the multiplication rule for independent events, we can compute the expected frequencies.

Hours Studying Per Week	Major				Row Total
	Humanities	Social Science	Business	Engineering	
0–10	68	106	131	40	**345**
11–20	119	103	127	81	**430**
More than 20	70	52	51	52	**225**
Column Total	**257**	**261**	**309**	**173**	**1000**

For example, P(Studying 0 – 10 hours and Humanities)
= P(Studying 0 – 10 hours) · P(Humanities major) = $\frac{345}{1000} \cdot \frac{257}{1000}$

We obtain the expected frequency for those who study 0 – 10 hours and are humanities majors by multiplying this probability by the grand total.

Expected frequency = $1000 \cdot \frac{345}{1000} \cdot \frac{257}{1000} = \frac{345 \cdot 257}{1000}$ = 88.665

From the previous example and applying some arithmetic, we can see that each expected frequency, E, can be obtained by the following:

$$E = \frac{\text{Row total} \cdot \text{column total}}{\text{Grand total}}$$

The expected frequency for a cell represents the number of individuals we would expect to find in that cell under the assumption that the two variables are independent. If the differences between the observed and expected frequencies tend to be large, we will reject the null hypothesis of independence.

PERFORMING A TEST OF INDEPENDENCE

Step 1: State the null and alternate hypotheses.
Step 2: Compute the row and column totals.
Step 3: Compute the expected frequencies.

$$E = \frac{row\ total\ \cdot column\ total}{Grand\ total}$$

Check to be sure that all the expected frequencies are at least 5

Step 4: Choose a level of significance α, and compute the test statistic

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

And the P-value using technology

Step 5: State a conclusion.

EXAMPLE: Perform a test of the null hypothesis that major and hours studying are independent. Use α = 0.01.

Hours Studying Per Week	Major			
	Humanities	**Social Science**	**Business**	**Engineering**
0–10	68	106	131	40
11–20	119	103	127	81
More than 20	70	52	51	52

SOLUTION: Assumptions ✓

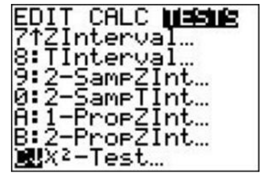
H_0: hours studying and major are independent
H_1: hours studying and major are not independent
are dependent

$\alpha = 0.01$

STAT TEST C: χ^2 Test observed : [A]
Expected : [B]

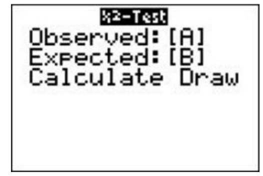
Test statistic: $\chi^2 = 34.64$
P-value : $p = 0.000005064$

Since $p < \alpha$, we reject H_0. We conclude that the choice of major and number of hours spent studying are not independent: the number of hours that students study varies among majors.



HYPOTHESIS TESTS ON THE TI-84 PLUS

The χ^2**-Test** will perform a test for independence. This command is accessed by pressing **STAT** and highlighting the **TESTS** menu.



This procedure requires that the observed frequencies be entered into a matrix.

OBJECTIVE 3
PERFORM TESTS OF HOMOGENEITY

TEST OF HOMOGENEITY

In the contingency table we have seen so far, the individuals in the table were sampled from a single population. For each individual, the values of both the row and column variables were random. In some cases, values of one of the variables (say, the row variable) are assigned by the investigator, and are not random.

In these cases, we consider the rows as representing separate populations, and we are interested in testing the hypothesis that the distribution of the column variable is the same for each row. This is known as a **test of homogeneity**.

The drugs telmisartan and ramipril are designed to reduce high blood pressure. In a clinical trial to compare the effectiveness of these drugs in preventing heart attacks, 25,620 patients were divided into three groups.

One group took one telmisartan tablet each day, another took one ramipril tablet each day, and the third group took one tablet of each drug each day. The patients were followed for 56 months, and the numbers who suffered fatal and nonfatal heart attacks were counted.

	Fatal Heart Attack	Nonfatal Heart Attack	No Heart Attack
Telmisartan only	598	431	7513
Ramipril only	603	400	7573
Both drugs	620	424	7458

In this experiment, the patients were assigned a row category depending on which drug was administered, so only the column variable is random.

We are interested in performing a test of homogeneity, to test the hypothesis that the distribution of outcomes is the same for each row.

The method for performing a test of homogeneity is:

Identical to the method for performing a test of homogeneity

EXAMPLE: Test the hypothesis that the distribution of outcomes is the same for all three treatment groups in the example about the drugs telmisartan and rampiral. Use the $\alpha = 0.05$ level.

	Fatal Heart Attack	Nonfatal Heart Attack	No Heart Attack
Telmisartan only	598	431	7513
Ramipril only	603	400	7573
Both drugs	620	424	7458

SOLUTION: Assumptions ✓

H_0: The distribution of outcome is the same for all drug treatments

H_1: The distribution is not the same

χ^2 - test : matrix 3 x 3

Test statistic : $\chi^2 = 2.2590$

p-value : p 0.6882

Since $p > \alpha$, we do not reject H_0.

There is not enough evidence to conclude that the distribution of outcomes is ~~difference~~ is different for different drug treatments.

YOU SHOULD KNOW …

- How to interpret contingency tables
- How to perform tests of independence
- How to perform tests of homogeneity