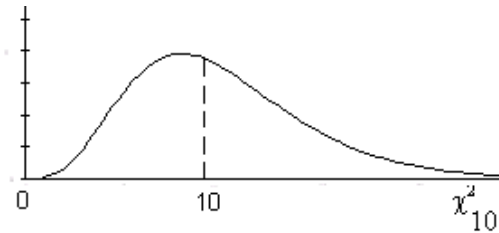# Chapter 11 χ^2 "chi-square" probability distribution

- continuous probability distribution
- shape is skewed to the right
- variable values on horizontal axis are ≥ 0
- area under the curve represents probability
- horizontal asymptote – extends to infinity along positive horizontal axis - curve gets closer to horizontal axis but does not touch it as X gets large
- shape depends on "degrees of freedom" (d.f.)
  - when d.f. = 2 only, curve is an exponential distribution
  - for higher degrees of freedom, the peak shifts to the right
  - when d.f. > 90, curve begins to look like normal distribution
    (even then χ^2 is still <u>always</u> somewhat skewed right)
- mean is located a little to the right of the peak
- Mean and Standard Deviation are determined by the degrees of freedom:

  mean = d.f.   standard deviation = $\sqrt{2\,(\text{d.f.})}$
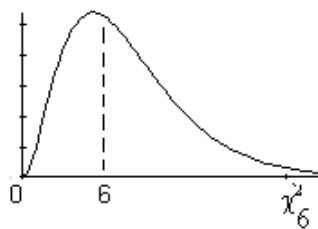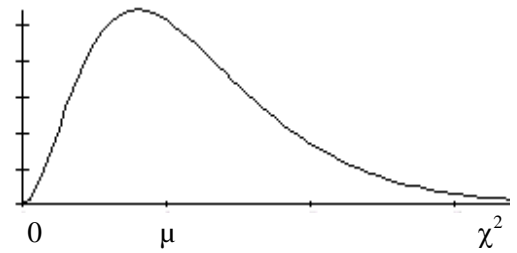
χ^2 with 6 degrees of freedom χ^2 with 10 degrees of freedom

Finding Probabilities: TI –83, 84: χ^2 cdf (lower bound, upper bound, degrees of freedom)

We will be using the right tail for hypothesis tests for Goodness of Fit orfor Independence:
TI –83, 84: χ^2 cdf (lower bound, 10^99, degrees of freedom)

χ^2 "chi-square" probability distribution is used for

Test of Goodness of Fit:
Hypothesis: Population fits an assumed distribution (theory)
Sample data is collected from a population
Hypothesis test is performed to see if the sample data supports the theory that the population fits this assumed distribution, or not.

Test of Independence:
Hypothesis: Two qualitative variables are independent of each other
Sample data is collected from a population
Hypothesis test is performed to see if the sample data supports the theory that these two variables are independent or not

Hypothesis Test (and confidence intervals) for an unknown population standard deviation.

χ²-Goodness of Fit Test

Null hypothesis states a probability distribution that we think the population follows.
We look at sample data to decide if the population follows that distribution or not.

- Write the **Null and Alternative Hypothesis** (as sentences)
  Ho: The data fit the expected (*hypothesized*) distribution
     (Describe the expected hypothesized distribution either in a sentence, by using its appropriate name, by showing a formula or referring to a nearby visible table or list)

  Ha: The data do not fit the expected (*hypothesized*) distribution

- Decide upon α

- Collect observed data AND calculate expected data based on the table of percentages given in the problem or based on the probability distribution named or described in the problem

- Find the test statistic and pvalue and draw the graph

$$\text{Test statistic} = \sum_{\substack{all\ cells \\ in\ table}} \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

Test statistic is a number measuring the size of the differences between observed sample data and expected data

   χ^2 distribution, right tailed test ; df = # of CELLS − 1

   2nd DISTR χ^2cdf (test statistic, 10^99, df)
 Draw the graph ; shade pvalue; label axis, test statistic & pvalue

- Make a Decision by comparing pvalue to significance level α
  - If test statistic is large indicating sample data differ from expected, then area in right tail is small. If pvalue < α, Reject Ho
  - If test statistic is small, indicating sample data are similar to expected, then area in right tail is large. If pvalue > α, Reject Ho
  -
- Write a conclusion in the context of the problem.

Calculator Instructions for test statistic and pvalue
STAT EDIT: Enter Observed Data frequencies in list L1
               Enter Expected Theoretical frequencies in list L2
                  (Do NOT put totals in either list)
               Arrow up to the very top (title line) in list L3 and input $(L1 - L2)^2 / L2$; press Enter
STAT CALC: 1 variable stats L3

   Test Statistic is ΣX = total of list L3 = $\sum_{\substack{all\ cells \\ in\ table}} \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$

2nd DISTR χ^2 cdf (lower bound, 10^99, degrees of freedom) = **pvalue**

EXAMPLE 1: <u>**Hypothesis Test for Goodness of Fit (GOF)**</u>

Student Demographics for all California Community Colleges by Age for 2014-15 *http://californiacommunitycolleges.cccco.edu/PolicyInAction/KeyFacts.aspx*			Ages for a Sample of 227 De Anza College Students	
19 or less	25%		19 or less	48
20 – 24	32%		20 – 24	107
25 – 29	14%		25 – 29	30
30 – 34	8%		30 – 34	14
35 and over	21%		35 and over	28
Total	100%		Total	227

We want to know if the age distribution of De Anza Colletge Students fits the age distribution of community college students statewide.

Null Hypothesis Ho: _____

_____

Alternate Hypothesis H_A: _____

_____

α = _____

	L1 Observed Data	L2 Expected Data	L3= $(L1-L2)^2/L2$ $\dfrac{(Observed – Expected)^2}{Expected}$
19 or less			
20 – 24			
25 – 29			
30 – 34			
35 and over			

Total =   Total =   **Test Statistic = Sum =** _____

The test statistic measures the size of the differences between the distributions

Calculate the p-value

Use_____distribution with

_____degrees of freedom

Draw the graph

pvalue = _____

Decision: _____

Conclusion: _____

_____

_____

Example 2: Are calls for emergency medical services uniformly distributed by day of week? A city is reviewing staffing levels and staffing schedules for their emergency medical response team. To determine if calls are uniformly distributed by day of the week, they analyzed data for a sample of 575 EMS calls.

Day of the Week	Sun.	Mon.	Tues.	Wed.	Thurs.	Fri.	Sat.
Number of EMS calls	88	79	84	85	77	75	87

Null Hypothesis Ho: _____

_____

Alternate Hypothesis H$_A$: _____

_____

α = _____

	L1 Observed Data	L2 Expected Data	L3= $(L1-L2)^2$/L2 $\dfrac{(Observed - Expected)^2}{Expected}$
Sunday			
Monday			
Tuesday			
Wednesday			
Thursday			
Friday			
Saturday			

Total =   Total =   **Test Statistic = Sum = _____**

The test statistic measures the size of the differences between the distributions

Calculate the p-value
Use_____distribution with   Draw the graph
_____degrees of freedom

pvalue = _____

Decision: _____

Conclusion: _____

_____

_____

Practice Examples for Goodness of Fit Test   *Reminder: If alpha is not stated in the problem, use 5%.*

Example 3. Are births uniformly distributed by day of the week?
The Health Commissioner in River County wants to know whether births in the county are uniformly distributed by day of week.
A sample of 434 randomly selected births at hospitals in the county yield the following data.

Day of the Week	Sun.	Mon.	Tues.	Wed.	Thurs.	Fri.	Sat.
Number of Births in sample	41	64	71	72	71	69	46

Based on information in the National Vital Statistics Report January 7, 2009
http://www.cdc.gov/nchs/data/nvsr/nvsr57/nvsr57_07.pdf

Example 4. At a cell phone company, in the past 5 years the frequency with which their US customers replaced their phone by a newer model is:

Phone ownership Time Period	<1 year	1 to <2 years	2 to <3 years	3 to <4 years	≥4 years
Percent Replaced	3%	25%	45%	18%	9%

Their marketing staff conducts a market research survey of 225 customers who bought a new model of phone and asks about the time period in which they plan to replace their phones

Phone ownership Time Period	<1 year	1 to <2 years	2 to <3 years	3 to <4 years	≥4 years
Number who plan to replace phone	0	60	106	42	17

Do the observed data for planned phone replacement fit the past distribution of phone replacement times or does the distribution for planned phone replacement differ from the past distribution?

Example 5. A genetics lab experiment investigates feather color in a species of bird. Genetic theory predicts that when mating pairs of green birds of this species, the expected outcomes for probabilities of the offspring's feather colors will follow the pattern for a "dihybrid cross":
    56.25% green, 18.75% yellow, 18.75% blue, 6.25% white.
For a sample of pairs of green birds mated in the lab, the offspring's observed color distribution was:
    41 green, 12 blue, 20 yellow, 7 white.
Conduct a hypothesis test to determine if the feather color distribution fits or does not fit the hypothesized distribution: 56.25% green, 18.75% yellow, 18.75% blue, 6.25% white.

Example 6. In the fourth week of class we used the binomial distribution
B(4, 0.70) to find the probabilities for X = the number of students receiving financial aid in a group of 4 students, if 70% of students at the college get financial aid.

X = Number of Students getting financial aid	0	1	2	3	4
Probability	0.0081	0.0756	0.2646	0.4116	0.2401

   Probabilities are found as: binompdf(4, 0.70, x)
   In chapter 4 notes, we also calculated them using probability rules

A statistics student decides to sample many groups of 4 students to see if the binomial distribution is a good fit for this situation. She and her friends collect data for 150 groups of 4 students.
The table below shows the number of groups of students in which 0,1,2,3 or 4 students receive financial aid.

X = Number of Students getting financial aid	0	1	2	3	4
Frequency	3	15	36	55	41

Perform a goodness of fit test to determine if the observed data are consistent with the binomial distribution B(4, 0.70)

χ^2 Hypothesis Test for Independence
Review: Understanding Independence in a Contingency Table:

EXAMPLE 7: Lina is a statistics student doing a project to compare ice cream flavor preferences at 3 ice cream stores in different cities. She wants to determine if customer preferences are related to store location or if they are independent. She will select a sample of customers, and categorize each customer by store location and flavor preference.

Assume that there are only 3 flavors

A.

	Fremont(F)	Gilroy(G)	Hayward (H)	Total
Chocolate (C)	25	25	50	100
Vanilla (V)	15	15	30	60
Mint Chip (M)	10	10	20	40
Total	50	50	100	200

Are flavor preference and store location INDEPENDENT?

B:

	Fremont (F)	Gilroy(G)	Hayward (H)	Total
Chocolate (C)	20	10	70	100
Vanilla (V)	20	10	20	50
Mint Chip (M)	10	30	10	50
Total	50	50	100	200

Are flavor preference and store location INDEPENDENT?

C:

	Fremont (F)	Gilroy(G)	Hayward (H)	Total
Chocolate (C)	27	24	49	100
Vanilla (V)	13	15	32	60
Mint Chip (M)	10	11	19	40
Total	50	50	100	200

Are flavor preference and store location INDEPENDENT?

D.

	Fremont (F)	Gilroy(G)	Hayward (H)	Total
Chocolate (C)	22	30	48	100
Vanilla (V)	10	12	38	60
Mint Chip (M)	18	8	14	40
Total	50	50	100	200

Are flavor preference and store location INDEPENDENT?

How far can our data vary from the independent data in part A and still be considered "similar"?
How far must our data vary from the independent data in part A in order to be considered "different"?

χ² Hypothesis Test for Independence

We use a hypothesis Test for Independence to help us decide whether two qualitative variables are independent or not.

We calculate what the data is <u>expected</u> to look like if the variables are independent.

We compare the <u>observed</u> data to the <u>expected</u> data
        Is the "observed data" sufficiently different from "expected" to decide that Ho is not true.
        OR is the "observed data" reasonably close to "expected"

Writing the Hypothesis Test for Independence
This is an overview of the process.
Calculator instructions are on the last page of the notes for this chapter.

- Write the hypotheses
       Null Hypothesis: Ho: The variables (*DESCRIBE THEM*) are independent
       Alternate Hypothesis: H$_A$: The variables (*DESCRIBE THEM*) are NOT independent

- Decide upon significance level α

- Collect observed data. Calculate expected data.

- Find the test statistic and pvalue: Test statistic $\displaystyle\sum_{\substack{all\ cells \\ in\ table}} \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$

   Find the pvalue using the χ^2 distribution , right tailed test
       df = (# of Rows – 1) (# of Columns -1)
       χ^2cdf (test statistic, 10^99, df)
   Draw the graph ; shade the pvalue; label axis, test statistic & pvalue

- Compare pvalue to the significance level α and make a decision.

- Write a conclusion in the context of the problem.

Calculator Instructions for Test of Independence:
Entering the data:
2nd MATRIX EDIT; select matrix [A] and press enter.
Enter matrix size: number of rows **x** number of columns, pressing enter after each is entered
NOTE: Look only at data when figuring out size; do NOT include row or column totals in the matrix.
Matrix on calculator will change shape on the screen to match the stated size.
Enter observed data into matrix A, press enter after each number is entered.

Performing the hypothesis test to find the pvalue and test statistic
STAT TESTS: χ^2 Test
    Observed: 2nd Matrix 1: [A] Enter   *(This is the input to the test)*

    Expected: 2nd Matrix 2: [B] Enter   *(This is output the calculator creates when doing the test)*

OUTPUT SCREEN will show pvalue and test statistic and degrees of freedom

<u>ON HOME SCREEN:</u> To see matrices containing observed and expected data
    Observed: 2nd Matrix 1: [A] Enter
    Expected: 2nd Matrix 2: [B] Enter
You may need to scroll to the right to see part of the matrix if it is too wide to fit on y our calculator screen

EXAMPLE 8: χ^2 Hypothesis Test for Independence

Use the following sample of data from 3 ice cream store locations to determine whether flavor preference and store location are independent.

OBSERVED DATA	Fremont (F)	Gilroy(G)	Hayward (H)	Total
Chocolate (C)	22	30	48	100
Vanilla (V)	10	12	38	60
Mint Chip (M)	18	8	14	40
Total	50	50	100	200

Ho: _____

Ha: _____

Calculating the EXPECTED DATA (assumes independence)

EXPECTED DATA	Fremont (F)	Gilroy(G)	Hayward (H)	Total
Chocolate (C)				100
Vanilla (V)				60
Mint Chip (M)				40
Total	50	50	100	200

Compare OBSERVED Data to how the data is EXPECTED to look IF INDEPENDENT.

How close is actual "observed data (Obs)" to "expected data (E)"?

Measure of the difference for each cell inside the table: $\dfrac{(\text{Observed}-\text{Expected})^2}{\text{Expected}}$

Add them all up to get: Test Statistic $= \displaystyle\sum_{\substack{all\ cells \\ in\ table}} \dfrac{(\text{Observed}-\text{Expected})^2}{\text{Expected}}$

Test Statistic = $(22-25)^2/25 + (30-25)^2/25 + (48-50)^2/50$

$+ (10-15)^2/15 + (12-15)^2/15 + (38-30)^2/30$

$+ (18-10)^2/10 + (8-10)^2/10 + (14-20)^2/20$ Test Statistic = 14.44

Draw, shade and label the graph

Degrees of freedom = (number of rows − 1)(number of columns − 1); df =(___ − 1)(____ − 1) = _____

Probability Distribution to use to calculate pvalue: _____ with df = _____

pvalue = P($\chi^2 \geq$ ____) = _____ (_____, _____, ___) = _____

Decision:

Conclusion: At a 5% level of significance, the data show sufficient evidence to conclude that customer ice cream flavor preference and store location are NOT independent. (Statistically Significant: Flavor preference depends on store location.)

EXAMPLE 9: χ^2 Hypothesis Test for Independence

A professor believes that giving extra time on an exam won't affect exam grades. He divided a random sample of 300 students into two groups and gave the same test. The first group had 1hour for the exam; the second group had unlimited time. At a 5% level of significance, can we conclude that time given for exam and grade are related?

OBSERVED DATA	Grade Earned on Exam				
	A	**B**	**C**	**F**	TOTAL
1 hour	23	42	65	12	142
Unlimited time	17	48	85	8	158
TOTAL	40	90	150	20	300

Ho: _____

Ha: _____

Use the _____ test on your calculator to find the following:

   Test Statistic = _____   pvalue = _____

   Show calculation for Degrees of freedom = _____

   What probability distribution is used to calculate the pvalue:_____

Draw, shade and label the graph

Decision: _____

Reason for Decision: _____

Is the result Statistically Significant (circle one):   YES   NO
Conclusion:

Find the output matrix with the EXPECTED DATA and fill it in

EXPECTED DATA	Grade Earned on Exam				
	A	**B**	**C**	**F**	TOTAL
1 hour					142
Unlimited time					158
TOTAL	40	90	150	20	300

Show all values in the expected data accurate to two decimal places
By hand, show the calculation that the calculator has performed to find the expected number of
a) students earning an A when given 1 hour for the exam

b) students earning a C when given unlimited time for the exam

EXAMPLE 10: χ^2 Hypothesis Test for Independence

We can compare proportions for several populations using the methods of a Test for Independence. Since there are separate samples drawn from each population, instead of one sample drawn from one population and then categorized by two variables, this is actually called a "Test of Homogeneity" but the procedures of this hypothesis test are the same as the Test of Independence.

Suppose that we want to determine the proportion of students at a college in different majors who intend to continue on to graduate school.

- For a sample of 150 students in major A, 47 intend to continue on to graduate school.
- For a sample of 100 students in major B, 24 intend to continue on to graduate school.
- For a sample of 100 students in major C, 29 intend to continue on to graduate school.
- For a sample of 150 students in major D, 26 intend to continue on to graduate school.

Can we conclude that the proportions of all students in these majors who intend to continue to graduate school are the different or are the proportions who intend to continue to graduate school independent of the student's choice of major?

1. Calculate the sample proportions
2. Create a contingency table with columns for the majors and with rows for the events "intend to continue on to graduate school" and "not intend to continue on to graduate school"
3. Use the techniques for the test for independence, using a hypotheses about whether intent to continue to graduate school is independent of major.

	Major				
OBSERVED DATA					TOTAL
intend to continue on to graduate school					
not intend to continue on to graduate school					
TOTAL					

Ho: _____

Ha: _____

Use the _____ test on your calculator to find the following:

 Test Statistic = _____   pvalue = _____

 Show calculation for Degrees of freedom = _____

 What probability distribution is used to calculate the pvalue:_____

Draw, shade and label the graph

Decision: _____

Reason for Decision: _____

Is the result Statistically Significant (circle one):   YES   NO

Conclusion:

Practice Examples for Test of Independence
Reminder: If alpha is not stated in the problem, use 5%.

Example 11. A random sample of 1343 US adults age 25 & over were surveyed and asked about their educational status and whether they were employed or unemployed. Are employment status and educational status independent?

SOURCE: Based on information in the Current Population Survey Feb 2013
http://www.bls.gov/webapps/legacy/cpsatab4.htm

http://data.bls.gov/pdq/SurveyOutputServlet

OBSERVED DATA	Not a high school Graduate	High School Graduate	Some College or Associate Degree	Bachelors Degree or Higher	TOTAL
Employed	100	333	348	476	1257
Unemployed	13	29	25	19	86
TOTAL	113	362	373	495	1343

Example 12. A personal trainer at a gym wants to compare three exercise plans (Plans A, B, C) to see if they are equally effective in improving the fitness of the participants.

	A	B	C
Moderate or large improvement	41	55	30
No or little improvement	17	12	15

A sample of people who train at the gym are randomly assigned to one of the exercise plans and the trainer evaluates whether fitness improvement occurs or not.
Perform a hypothesis test to determine if the improvement in fitness depends on the fitness plan or if improvement is independent of the particular fitness plan the person participates in.

Example 13. *http://www.people-press.org/files/2015/06/6-4-15-Immigration-release.pdf*
The Pew Research Center conducted a poll of US adults' opinions on immigration. One of the questions asked was whether undocumented immigrants should be permitted a way to stay in the US legally, if requirements are met. The responses are summarized below

		Political Party Affiliation			
		Democrat	Republican	Unaffiliated Or Other	Total
Response to Immigration Question	Yes	509	283	576	1368
	No	121	218	174	513
	Don't Know	6	5	8	19
	Total	636	506	758	1900

At a 5% level of significance, are resident's opinions about illegal immigration independent of their political party affiliation or is there a relationship between their opinions and their political party affiliation?