11.1 Chi-Square Tests

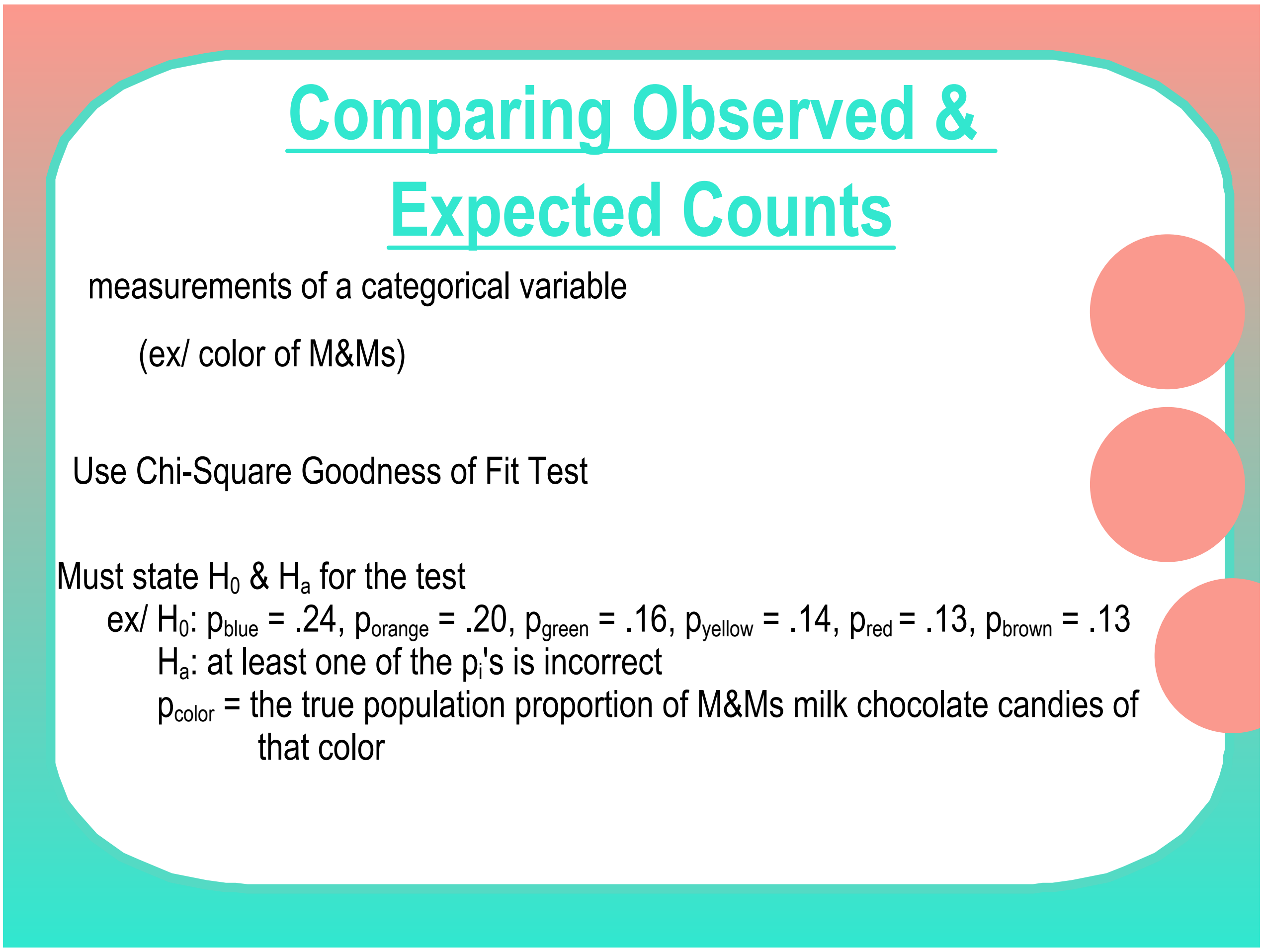
(Day 1)

vocab

new seats

examples

Objectives



Comparing Observed & Expected Counts

measurements of a categorical variable

(ex/ color of M&Ms)

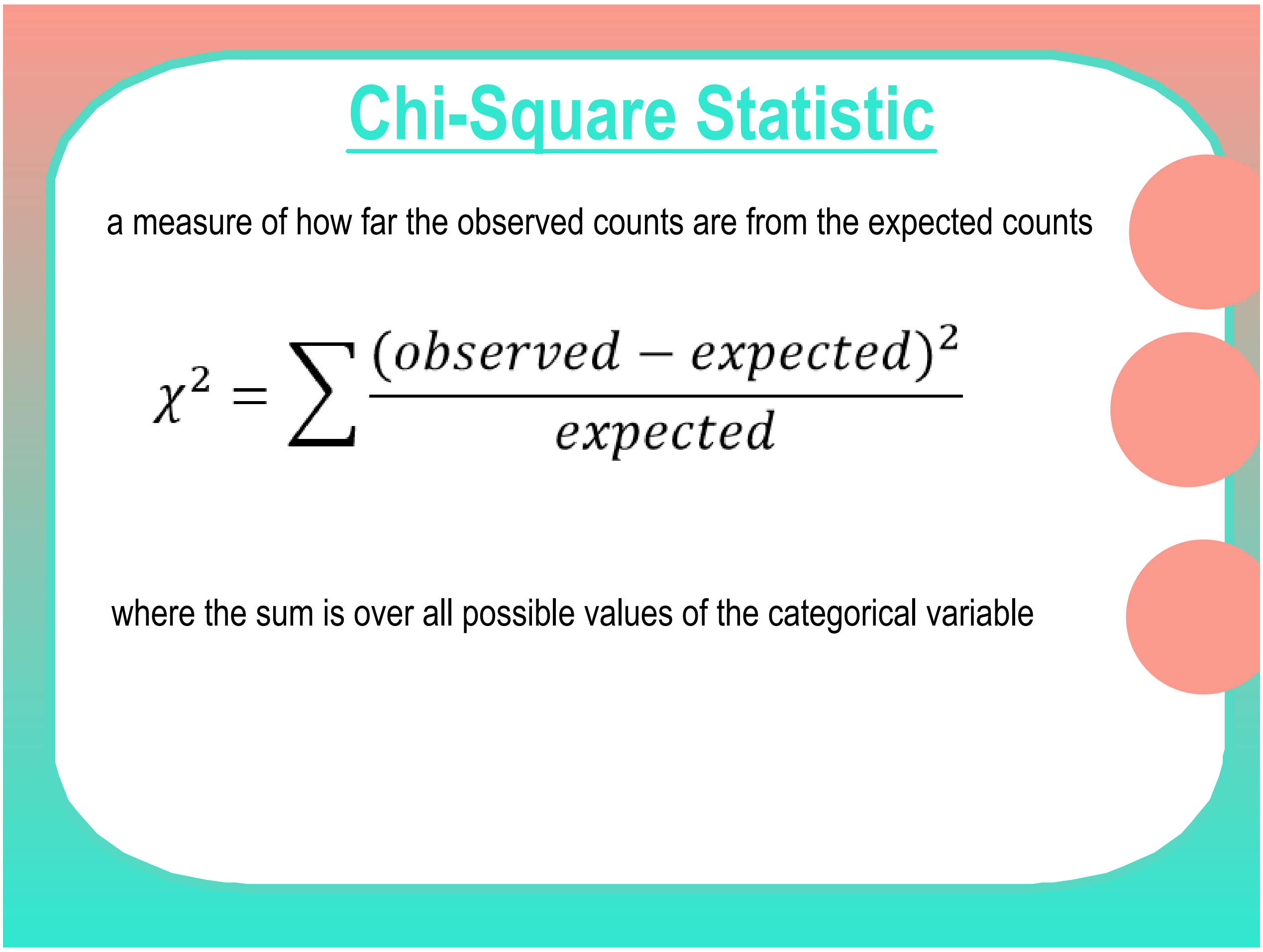
Use Chi-Square Goodness of Fit Test

Must state H_0 & H_a for the test

ex/ H_0: $p_{blue} = .24$, $p_{orange} = .20$, $p_{green} = .16$, $p_{yellow} = .14$, $p_{red} = .13$, $p_{brown} = .13$

H_a: at least one of the p_i's is incorrect

p_{color} = the true population proportion of M&Ms milk chocolate candies of that color



Chi-Square Statistic

a measure of how far the observed counts are from the expected counts

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

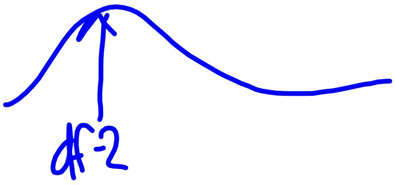
where the sum is over all possible values of the categorical variable

Chi-Square Distribution

a family of distributions that take only positive values & are skewed to the right

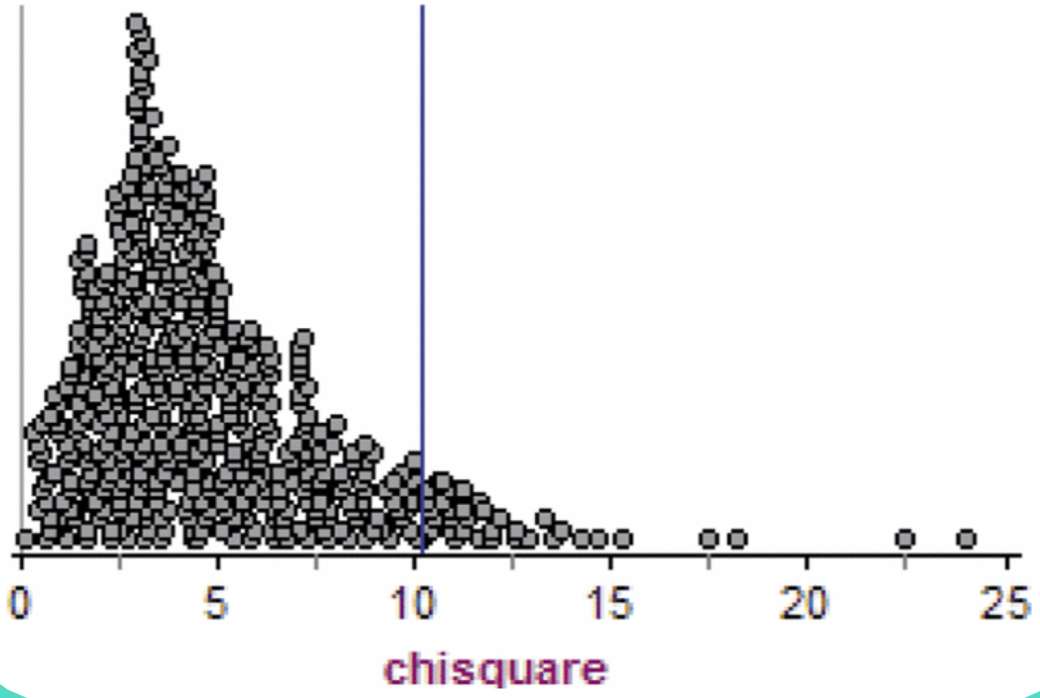
specified by its degrees of freedom = # of categories - 1

mean = df sd = $\sqrt{2df}$ peak at: df - 2



* expected counts must be at least 5

Chi-Square Distribution



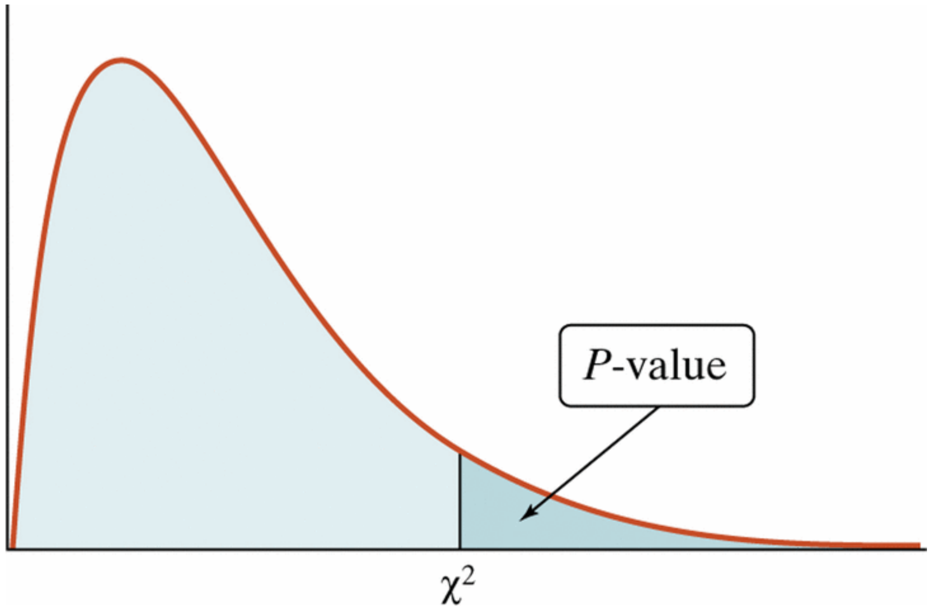
Calculating a P-value

use table C (P-value will fall between two values)

use calculator χ^2cdf(min, max, df)

χ^2 1E99

Calculating a P-value



Mini M&Ms

Joey has a bag of mini M&Ms. The company claims that the colors are produced in the following ratios:

Blue (.23), Orange (.23), Green (.15), Red (.12), Brown (.12)

COLOR	OBSERVED	EXPECTED
Blue	12	10.58
Orange	7	10.58
Green	13	6.9
Yellow	4	6.9
Red	8	5.52
Brown	2	5.52

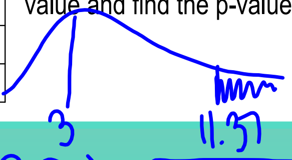
(a) Are the expected counts large enough? What are the degrees of freedom? yes, exp > 5 ✓

df = 6 - 1 = 5

(b)Calculate the Chi Square statistic

11.37

(c) Sketch a curve that shades the p-value and find the p-value



$$\chi^2 cdf(\underset{min}{11.37}, \underset{max}{1E99}, \underset{df}{5}) = \boxed{.0445}$$

Mini M&Ms

Joey has a bag of mini M&Ms. The company claims that the colors are produced in the following ratios:

Blue (.23), Orange (.23), Green (.15), Red (.12), Brown (.12)

COLOR	OBSERVED	EXPECTED
Blue	12	10.58
Orange	7	10.58
Green	13	6.9
Yellow	4	6.9
Red	8	5.52
Brown	2	5.52

(a) Are the expected counts large enough? What are the degrees of freedom? yes, all are greater than 5

(b)Calculate the Chi Square statistic

11.21

(c) Sketch a curve that shades the p-value and find the p-value

.9439

Chi Square Test

3 conditions:

Random: The data come from a random sample or a randomized experiment.

Large Sample Size: All expected counts are at least 5.

Independent: Individual observations are independent. When sampling without replacement, check that the population is at least 10 times as large as the sample (the 10% condition).

Chi Square Test

To determine whether a categorical variable has a specified distribution, expressed as the proportion of individuals falling into each possible category, perform a test of:

H_0: p_1 = _____, p_2 = _____, ... , p_k = ______

H_a: At least one of the p_i's is incorrect

Landlines

According to the 2000 census, of all U.S. residents aged 20 and older, 19.1% are in their 20s, 21.5% are in their 30s, 21.1% are in their 40s, 15.5% are in their 50s, and 22.8% are 60 and older. The table below shows the age distribution for a sample of U.S. residents aged 20 and older. Members of the sample were chosen by randomly dialing landline telephone numbers. Do these data provide convincing evidence that the age distribution of people who answer landline telephone surveys is not the same as the age distribution of all U.S. residents?

Category	Count
20–29	141
30–39	186
40–49	224
50–59	211
60+	286
Total	1048

Landlines

According to the 2000 census, of all U.S. residents aged 20 and older, 19.1% are in their 20s, 21.5% are in their 30s, 21.1% are in their 40s, 15.5% are in their 50s, and 22.8% are 60 and older. The table below shows the age distribution for a sample of U.S. residents aged 20 and older. Members of the sample were chosen by randomly dialing landline telephone numbers. Do these data provide convincing evidence that the age distribution of people who answer landline telephone surveys is not the same as the age distribution of all U.S. residents?

Category	Count
20–29	141
30–39	186
40–49	224
50–59	211
60+	286
Total	1048

STATE: $H_0: P_{20s} = .191,\ P_{30s} = .215,\ P_{40s} = .211,\ P_{50s} = .155,\ P_{60s} = .228$

H_a: at least one of the P_i's is incorrect

P_{age} = true population proportion of people in given age range who answer landline survey.

$\alpha = .05$

Landlines

According to the 2000 census, of all U.S. residents aged 20 and older, 19.1% are in their 20s, 21.5% are in their 30s, 21.1% are in their 40s, 15.5% are in their 50s, and 22.8% are 60 and older. The table below shows the age distribution for a sample of U.S. residents aged 20 and older. Members of the sample were chosen by randomly dialing landline telephone numbers. Do these data provide convincing evidence that the age distribution of people who answer landline telephone surveys is not the same as the age distribution of all U.S. residents?

Category	Count
20–29	141
30–39	186
40–49	224
50–59	211
60+	286
Total	1048

PLAN: Random - yes

Independent - assuming at least 10480 people who answer land line surveys

Large sample size

expected counts:

20-29	30-39	40-49	50-59	60+
200.17	225.32	221.13	162.44	238.94

Yes, all expected counts ≥ 5

We will use a χ^2 goodness of fit test

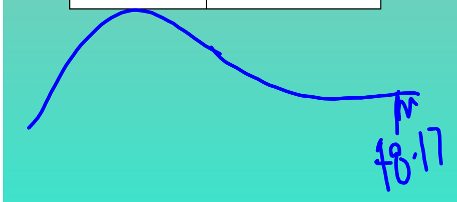
Landlines

According to the 2000 census, of all U.S. residents aged 20 and older, 19.1% are in their 20s, 21.5% are in their 30s, 21.1% are in their 40s, 15.5% are in their 50s, and 22.8% are 60 and older. The table below shows the age distribution for a sample of U.S. residents aged 20 and older. Members of the sample were chosen by randomly dialing landline telephone numbers. Do these data provide convincing evidence that the age distribution of people who answer landline telephone surveys is not the same as the age distribution of all U.S. residents?

Category	Count
20–29	141
30–39	186
40–49	224
50–59	211
60+	286
Total	1048

DO:

$$\chi^2 = \frac{(141-200.17)^2}{200.17} + \frac{(186-225.32)^2}{225.32} + \frac{(224-221.13)^2}{221.13} + \frac{(211-162.44)^2}{162.44} + \frac{(286-238.94)^2}{238.94} = \boxed{48.17}$$



χ^2cdf(48.17, 1E99, 4)
min max df

$\boxed{\approx 0}$

Landlines

According to the 2000 census, of all U.S. residents aged 20 and older, 19.1% are in their 20s, 21.5% are in their 30s, 21.1% are in their 40s, 15.5% are in their 50s, and 22.8% are 60 and older. The table below shows the age distribution for a sample of U.S. residents aged 20 and older. Members of the sample were chosen by randomly dialing landline telephone numbers. Do these data provide convincing evidence that the age distribution of people who answer landline telephone surveys is not the same as the age distribution of all U.S. residents?

Category	Count
20–29	141
30–39	186
40–49	224
50–59	211
60+	286
Total	1048

CONCLUDE:

p-value $\approx 0 < .05$ → We Reject H_0
→ We conclude that at least 1 of the true proportions of people who answer the land line survey in a given age group differs from the claimed proportions.
The Results are stat. sig. @ the 5% level.

On the calculator...

newer vs. older calculators

χ^2 GOF Test

Use L1 for observed, calculate expected in L2

Follow-up Analysis

If the results are statistically significant, it's best to look at each deviation separately to see which contributes most to the chi-square statistic:

$$\frac{(\text{observed-expected})^2}{\text{expected}}$$

and discuss if they were higher or lower than expected

On AP, only do a follow-up if the question asks you to

Birthdays

In his book *Outliers,* Malcolm Gladwell suggests that a hockey player's birth month has a big influence on his chance to make it to the highest levels of the game. Specifically, since January 1 is the cutoff date for youth leagues in Canada (where many National Hockey League players come from), players born in January will be competing against players up to 12 months younger. The older players tend to be bigger, stronger, and more coordinated and hence get more playing time and more coaching and have a better chance of being successful. To see if birth date is related to success (judged by whether a player makes it into the NHL), a random sample of 80 NHL players from the 2009–2010 season was selected and their birthdays were recorded. Overall, 32 were born in the first quarter of the year, 20 in the second quarter, 16 in the third quarter, and 12 in the fourth quarter. **Do these data provide convincing evidence that the birthdays of NHL players are not uniformly distributed throughout the entire year?**

Birthdays

To see if birth date is related to success (judged by whether a player makes it into the NHL), a random sample of 80 NHL players from the 2009–2010 season was selected and their birthdays were recorded. Overall, 32 were born in the first quarter of the year, 20 in the second quarter, 16 in the third quarter, and 12 in the fourth quarter. **Do these data provide convincing evidence that the birthdays of NHL players are not uniformly distributed throughout the entire year?**

STATE: H_0 : p_{1st} = .25, p_{2nd} = .25, p_{3rd} = .25, p_{4th} = .25 $\alpha = .05$

H_a: at least one of the p_i's is incorrect

$p_{quarter}$ is the true proportion of hockey players who were born in the given quarter

Birthdays

To see if birth date is related to success (judged by whether a player makes it into the NHL), a random sample of 80 NHL players from the 2009–2010 season was selected and their birthdays were recorded. Overall, 32 were born in the first quarter of the year, 20 in the second quarter, 16 in the third quarter, and 12 in the fourth quarter. **Do these data provide convincing evidence that the birthdays of NHL players are not uniformly distributed throughout the entire year?**

PLAN: Random - yes Independent - assuming at least 800 hockey players

Large Sample Size -

Expected:	**1st**	**2nd**	**3rd**	**4th**
	20	**20**	**20**	**20**

Yes, all expected are at least 5

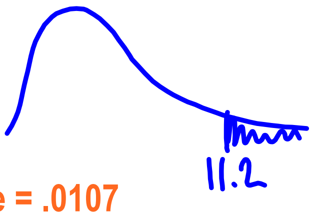
We will use a Chi-Square goodness of fit test

Birthdays

To see if birth date is related to success (judged by whether a player makes it into the NHL), a random sample of 80 NHL players from the 2009–2010 season was selected and their birthdays were recorded. Overall, 32 were born in the first quarter of the year, 20 in the second quarter, 16 in the third quarter, and 12 in the fourth quarter. **Do these data provide convincing evidence that the birthdays of NHL players are not uniformly distributed throughout the entire year?**

DO:

QUARTER	1st	2nd	3rd	4th
OBSERVED	32	20	16	12
EXPECTED	20	20	20	20
	↑	=	↓	↓



χ^2 GOF - TEST df = 3 χ^2 = 11.2 p-value = .0107

1st 12 higher 3,4th lower

2nd equal

Birthdays

To see if birth date is related to success (judged by whether a player makes it into the NHL), a random sample of 80 NHL players from the 2009–2010 season was selected and their birthdays were recorded. Overall, 32 were born in the first quarter of the year, 20 in the second quarter, 16 in the third quarter, and 12 in the fourth quarter. **Do these data provide convincing evidence that the birthdays of NHL players are not uniformly distributed throughout the entire year?**

CONCLUDE: .0107 < .05 --> We reject H_0 --> We can conclude that at least one of the quarters does not have 25% of the NHL players born in it. The results are statistically significant at the 5% level.

Birthdays

In his book *Outliers,* Malcolm Gladwell suggests that a hockey player's birth month has a big influence on his chance to make it to the highest levels of the game. Specifically, since January 1 is the cutoff date for youth leagues in Canada (where many National Hockey League players come from), players born in January will be competing against players up to 12 months younger. The older players tend to be bigger, stronger, and more coordinated and hence get more playing time and more coaching and have a better chance of being successful. To see if birth date is related to success (judged by whether a player makes it into the NHL), a random sample of 80 NHL players from the 2009–2010 season was selected and their birthdays were recorded. Overall, 32 were born in the first quarter of the year, 20 in the second quarter, 16 in the third quarter, and 12 in the fourth quarter. **If the results are significant, perform a follow-up analysis.**

Birthdays

To see if birth date is related to success (judged by whether a player makes it into the NHL), a random sample of 80 NHL players from the 2009–2010 season was selected and their birthdays were recorded. Overall, 32 were born in the first quarter of the year, 20 in the second quarter, 16 in the third quarter, and 12 in the fourth quarter. **If the results are significant, perform a follow-up analysis.**

It seems that the theory is correct as there were 12 more players than expected the first quarter, and 4 fewer from the third quarter, and 8 fewer from the 4th quarter than expected.

Our Class M&M Data

Mars Candy Company Claims:

Orange 20%, Red 13%, Yellow 14%, Green 16%, Blue 24%, Brown 13%

COLOR	**OBSERVED**
Orange	290
Red	64
Yellow	143
Green	270
Blue	247
Brown	174
TOTAL	1188

Is there convincing evidence that our M&Ms differ from the claimed proportions?