Sec. 12.1: Goodness-of-Fit Test

Sec. 12.2: Tests for Independence & The Homogeneity of Proportions

<u>Sec. 12.1</u>: Goodness-of-Fit Test

Idea Behind a Goodness-of-Fit Test

I believe that the numbers in the bag have the following distribution:

x	P(X=x)
1	40%
3	25%
7	15%
8	20%

Ummm, NO!

Idea Behind a Goodness-of-Fit Test

I believe that the numbers in the bag have the following distribution:

x	P(X=x)
[-7, -4]	10%
[0, 9]	65%
[12, 80]	25%

Idea Behind a Goodness-of-Fit Test

I believe that the numbers in the bag have the following distribution:

Normal

$\mu = 15\,, \quad \sigma = 4$

x	P(X=x)
$(-\infty\ , 7]$	2.28%
[7, 13]	28.57%
[13, 16]	29.02%
[16, 19]	26.26%
$[19,\ \infty\ )$	13.87%

Idea Behind a Goodness-of-Fit Test

How do you settle the dispute?

1. If categories are not given, the king and the peasant need to come up with a few categories (the categories will always be given in this class)
2. The categories need to be mutually exclusive and collectively exhaustive
3. Either the probability for each category will be given, or you'll need to find the probabilities for each category using the probability theory learned in this class

Idea Behind a Goodness-of-Fit Test

How do you settle the dispute?

4. Take a sample
5. Count how many numbers in your sample land in each category (these are called the Observed Frequencies O)
6. Calculate the Expected Frequencies E (this is how many numbers in your sample you expected to land in each category assuming the king is correct)

If the observed frequencies and expected frequencies are close together, believe the king!

If the observed frequencies and expected frequencies are far apart, believe the peasant!

Goodness-of-Fit Test Formulas & Info

Sample size: n

# of categories: k

Probability of landing in a specific category: p
(each category has a probability associated with it)

Observed frequency: O

Expected frequency: $E = np$

(each category has an expected frequency associated with it)

Goodness-of-Fit Test Formulas & Info

Probability Distribution: χ^2

Degrees of Freedom: $df = k - 1$

Test Statistic Formula: $\chi^2 = \sum \frac{(O - E)^2}{E}$

(tells you how far apart the observed and expected frequencies are)

Conditions:

1. All expected counts (E's) are greater than or equal to 1
2. No more than 20% of the expected counts (E's) are less than 5

Goodness-of-Fit Test Formulas & Info

Note:

- All Goodness-of-fit tests are RIGHT TAILED tests

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

Ex 1: (Sec. 12.1 Goodness-of-Fit) In order to test to see if a die is "fair", it is rolled 600 times. The results are in the table below. Perform a goodness-of-fit test at the $\alpha = 0.025$ significance level to test the claim that the die is "fair".

Categories	Observed Frequencies (O)
1	71
2	116
3	105
4	113
5	101
6	94

a) Use the P-value method

b) Use the rejection region method

Ex 2 (Sec. 12.1, Hw #12, pg. 595): **Peanut M&Ms** According to the manufacturer of M&Ms, 12% of the peanut M&Ms in a bag should be brown, 15% yellow, 12% red, 23% blue, 23% orange, and 15% green. A student randomly selected a bag of peanut M&Ms. He counted the number of M&M's that were each color and obtained the results shown in the table below. Test whether peanut M&Ms follow the distribution stated by M&M/Mars at the $\alpha = 0.05$ level of significance.

a) Use the P-value method

b) Use the rejection region method

<u>Ex 2 (Sec. 12.1, Hw #12, pg. 595)</u>: **Peanut M&Ms**

Color	Frequency
Brown	53
Yellow	66
Red	38
Blue	96
Orange	88
Green	59

Ex 3 (Sec. 12.1, Hw #13, pg. 595): **Benford's Law, Part I**

Our number system consists of the digits 0, 1, 2, 3, 4, 5, 6, 7, 8, and 9. The first significant digit in any number must be 1, 2, 3, 4, 5, 6, 7, 8, or 9 because we do not write numbers such as 12 as 012. Although we may think that each digit appears with equal frequency so that each digit has a 1/9 probability of being the first significant digit, this is not true. In 1881, Simon Newcomb discovered that first digits do not occur with equal frequency. This same result was discovered again in 1938 by physicist Frank Benford.

…

Ex 3 (Sec. 12.1, Hw #13, pg. 595): **Benford's Law, Part I** … After studying much data, he was able to assign probabilities of occurrence to the first digit in a number as shown in the table first table below. The probability distribution is now known as Benford's Law and plays a major role in identifying fraudulent data on tax returns and accounting books. For example, the second table below represents the first digits in 200 allegedly fraudulent checks written to a bogus company by an employee attempting to embezzle funds from his employer. Using a $\alpha = 0.01$ significance level, test whether the first digits in the allegedly fraudulent checks obey Benford's Law.

a) Use the P-value method

b) Use the rejection region method

Ex 3 (Sec. 12.1, Hw #13, pg. 595): **Benford's Law, Part I** …

Benford's Law									
Digit	1	2	3	4	5	6	7	8	9
Probability	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046

Data									
First Digit	1	2	3	4	5	6	7	8	9
Frequency	36	32	28	26	23	17	15	16	7

Sec. 12.2: Tests for Independence

&

The Homogeneity of Proportions

Idea Behind Tests for Independence

- There are two variables X & Y you are interested in (think of them as two questions)
- You want to know if the variables X & Y are independent
  1. Take a sample and ask everyone in the sample these 2 questions (rows represent variable X and the columns represent variable Y)
  2. Form a contingency table (the numbers in the table are the observed frequencies O)
  3. Assuming the rows and columns are independent, find the expected frequencies E

If the O's and E's are "close together", believe that the rows and columns are INDEPENDENT

If the O's and E's are "far apart", believe that the rows and columns are NOT INDEPENDENT

Tests for Independence Formulas & Info

- H_0 will always be

$$H_0\text{:}\ \ The\ rows\ and\ columns\ are\ independent$$

- H_1 will always be

$$H_1\text{:}\ \ The\ rows\ and\ columns\ are\ not\ independent$$

Tests for Independence Formulas & Info

Sample size (same as grand total): n

# of rows: r

# of columns: c

Observed frequency: O

Expected frequency: $E = \frac{(row\ total)(column\ total)}{grand\ total}$

Tests for Independence Formulas & Info

Probability Distribution: χ^2

Degrees of Freedom: $df = (r - 1)(c - 1)$

Test Statistic Formula: $\chi^2 = \sum \frac{(O - E)^2}{E}$

(tells you how far apart the observed and expected frequencies are)

Conditions:

1. All expected counts (E's) are greater than or equal to 1
2. No more than 20% of the expected counts (E's) are less than 5

Tests for Independence Formulas & Info

Note:

- All tests for independence are RIGHT TAILED tests

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

Ex 4 (Sec. 12.2, Ex 2, pg. 604): Is there a relationship between marital status and happiness? The data in the table below show the marital status and happiness of individuals who participated in the General Social Survey. Does the sample evidence suggest that one's happiness depends on one's marital status? Use the $\alpha = 0.05$ level of significance.

a) Use the P-value method

b) Use the rejection region method

Ex 4 (Sec. 12.2, Ex 2, pg. 604):

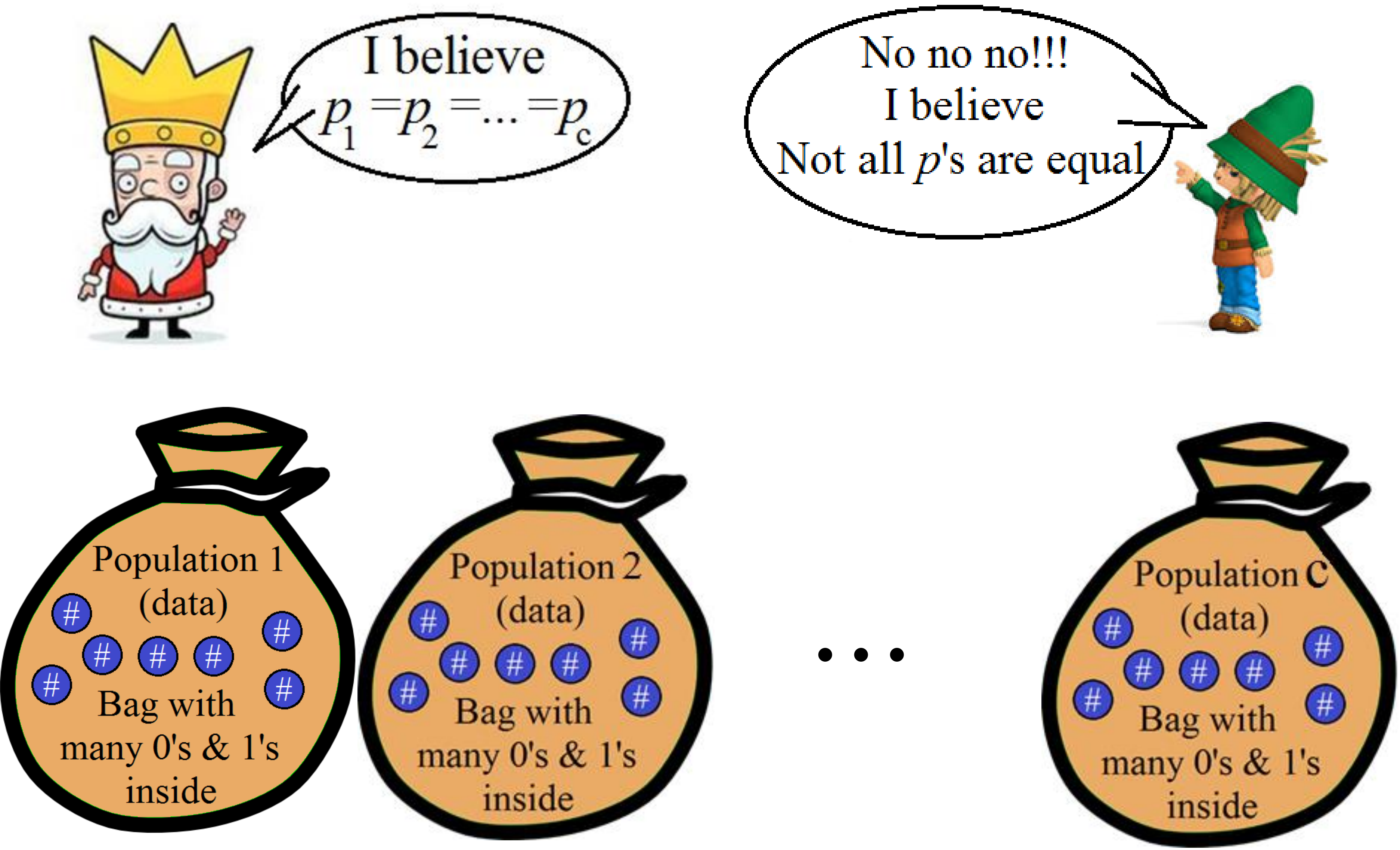
Marital Status

Happiness Level

	Married	Widowed	Divorced/Separated	Never Married
Very Happy	600	63	112	144
Pretty Happy	720	142	355	459
Not Too Happy	93	51	119	127

Sec. 12.2: Tests for Independence & The Homogeneity of Proportions

Idea Behind a Homogeneity of Proportions Test



Idea Behind a Homogeneity of Proportions Test

- You have many different yes or no questions (*c* such questions)
- Each population is assigned one of these questions
  1. Take a sample from each population and ask everyone in the sample its corresponding question
  2. Form a contingency table where each column represents a question and the rows represent yes's and no's (the numbers in the table are the observed frequencies *O*)
  3. Assuming all *p*'s are equal, calculate the *E*'s

If the *O*'s and *E*'s are "close together", believe that ALL *p*'s are equal

If the *O*'s and *E*'s are "far apart", believe that NOT ALL *p*'s are equal

Homogeneity of Proportions Test Formulas & Info

- H_0 will always be

$$H_0\text{:}\ \ p_1 = p_2 = \cdots = p_c$$

- H_1 will always be

$$H_1\text{:}\ \ Not\ all\ p's\ are\ equal$$

or

$$H_1\text{:}\ \ At\ least\ one\ of\ the\ p's\ is\ different\ from\ the\ others$$

Homogeneity of Proportions Test Formulas & Info

Sample size (same as grand total): n

# of rows: r

# of columns: c

Observed frequency: O

Expected frequency: $E = \dfrac{(row\ total)(column\ total)}{grand\ total}$

Homogeneity of Proportions Test Formulas & Info

Probability Distribution: χ^2

Degrees of Freedom: $df = (r-1)(c-1)$

Test Statistic Formula: $\chi^2 = \sum \frac{(O-E)^2}{E}$

(tells you how far apart the observed and expected frequencies are)

Conditions:

1. All expected counts (E's) are greater than or equal to 1
2. No more than 20% of the expected counts (E's) are less than 5

Homogeneity of Proportions Test Formulas & Info

Note:

- All tests for independence are RIGHT TAILED tests

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

Ex 5 (Sec. 12.2, Ex 4, pg. 607): Zocor is a drug manufactured by Merck and Co. that is meant to reduce the level of LDL (bad) cholesterol and increase the level of HDL (good) cholesterol. In clinical trials of the drug, patients were randomly divided into three groups. Group 1 received Zocor, group 2 received a placebo, and group 3 received cholestyramine, a cholesterol-lowering drug currently available. The table below contains the number of patients in each group who did and did not experience abdominal pain as a side effect. Is there evidence to indicate that the proportion of subjects in each group who experienced abdominal pain is different at the $\alpha = 0.01$ significance level?

a) Use the P-value method

b) Use the rejection region method

Ex 5 (Sec. 12.2, Ex 4, pg. 607):

	Group 1 (Zocor)	Group 2 (Placebo)	Group 3 (Cholestyramine)
# of people who had abdominal pain	51	5	16
# of people who did not have abdominal pain	1532	152	163