AP Statistics
Unit 08 – Day 01 Notes
Chi Square: Goodness of Fit Test

Name Key
Period______

We use this type of Inference test when we want to examine the distribution of a single categorical variable in a population!

One-way table: a table used to display the distribution of a single categorical variable (ex: M&M color)

Goodness of Fit: a test that tests the null hypothesis that a categorical variable has a specified distribution in the population of interest (ex: 20% yellow, 25% red, 15% blue, etc.)

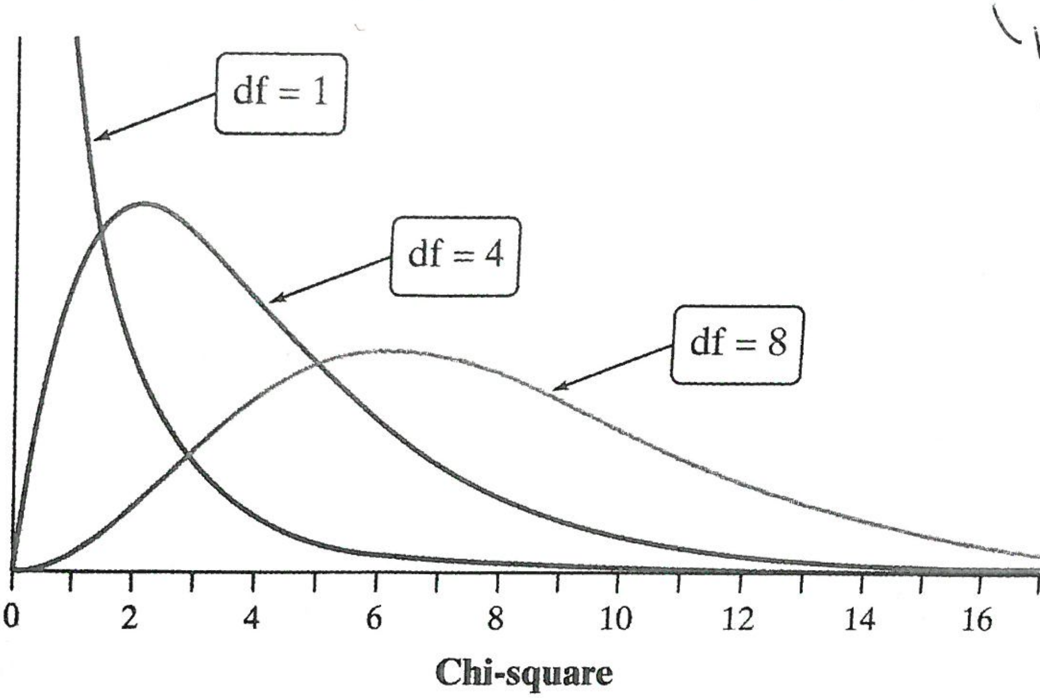
Observed Value: the actual # of individuals in the sample that fall in each cell of the one-way table (ex: 6 yellow M&Ms in a bag of 24)

Expected Value: the expected # of individuals in the sample that would fall in each cell of the one-way table if H_0 were true (ex: $p_0 \cdot n$ for each color of M&M yellow: $0.2 \cdot 24 = 4.8$ yellow M&Ms expected)

The Chi-Square Distribution and P-Values

- The sampling distribution of the chi-square statistic is **NOT** a **normal distribution**
- It is a **right-skewed** distribution that allows only **nonnegative values** (χ^2 = can't have a negative number when you square)
- When the expected counts are all at least 5, the sampling distribution of χ^2 statistic is modeled well by a **chi-square distribution** (you must list them all when checking conditions) Large counts condition
- Chi-square distributions are a family of density curves (kind of like t-distributions!)
- Chi-square distribution with degrees of freedom = the number of categories – 1 (not sample size)
- A particular chi-square distribution is specified by giving its degrees of freedom (see picture below)
- When df > 2, the mode (peak) of the chi-square density curve is df – 2

know this for MC questions!



STATE:

H_o = the stated distribution of the categorical variable in the population of interest is correct.

these get put into context

H_A = the stated distribution of the categorical variable in the population of interest is incorrect.

$\alpha =$ the default is still 0.05

*you can write these using #s but it takes forever so words are easier :)

PLAN:

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

10%: The sample must be less than 10% of the population if we are sampling without replacement

Large Counts: All counts must be 5 or more. YOU MUST WRITE THEM ALL OUT. ← this sucks! too bad :(

State Test: Because our conditions are met, we will perform a **chi-square test for goodness of fit**.

DO:

Chi-square statistic

$$\boxed{\chi^2 = \sum_{all\ cells} \frac{(observed - expected)^2}{expected}}$$

- work shown for the above formula (at least 3 terms +...)
- degrees of freedom df =
- $\chi^2 =$
- P-value =

You have to show all of the stuff even if you do the test on your calculator. No need to put in what you actually enter into your calculator this time (and thank goodness because that would basically require writing out the lists you entered).

CONCLUDE:

- Large values of χ^2 (the test statistic) are evidence against the H_o and in favor of the H_A.
- The P-value is the area to the right of the χ^2 under the chi-square distribution with degrees of freedom (df = number of categories – 1). you can find this using χ^2cdf(,) in the distributions menu (#8).

We either will reject the null or fail to reject the null and explain in context why. We can use the same sentence frames as before:

Because our p-value _____ is (greater than/less than) our significance level = ___, we (reject/fail to reject) the null. There (is/is not) convincing evidence that (alternative hypothesis in context).

ON THE CALCULATOR:

1. Enter your data into lists:
   a. STAT → Enter
      i. Observed Values in L1
      ii. Expected Values in L2
2. Choose the appropriate test:
   a. STAT → TESTS
   b. Option D: χ^2GOF-Test
      i. df: # categories – 1
      ii. Calculate

AND/OR

1. Calculate the test statistic by hand and then use the calculator to find the p-value:
   a. 2ND VARS
   b. Option 8: χ^2cdf(
      i. lower: your χ^2 value
      ii. upper: 10,000 (a very large number since it is skewed to the right)
      iii. df: # categories – 1

basically ∞

EXAMPLE 1:

Baseball is a remarkable sport, in part because so much data are available. We have the birth dates of every one if the 16,804 players who ever played in a major league game. Since the effect we're suspecting may be due to relatively recent policies (and to keep the same size moderate), we'll consider the birth months of the 1478 major league players born since 1975 to 2006. We can also look up the national demographic statistic to find what percentage of people who were born in each month. Let's test whether the observed distribution of ball players birth months shows just random fluctuations or whether it represents a real deviation from the national pattern.

L_1

Month	Ballplayer Count	National Birth %	Month	Ballplayer Count	National Birth %
1	137	8%	7	102	9%
2	121	7%	8	165	9%
3	116	8%	9	134	9%
4	121	8%	10	115	9%
5	126	8%	11	105	8%
6	114	8%	12	122	9%
			TOTAL	1478	100%

STATE:

H_0: The distribution of birthdays of all MLB players is the same as the gen. pop. (national pattern)

H_A: The distribution of birthdays of all MLB players is not the same as the general population.

α = 0.05

PLAN:

Random: all players between 1975 and 2006 ✓

10% condition: n=1478 14780 < 16804 players ever in MLB ✓

Large Counts: (this part sucks)

L_2 Find the expected values for each category (month)

Month	Expected	Month	Expected	Month	Expected
1	1478 · 0.08 = 118.24	5	118.24	9	133.02
2	103.46	6	118.24	10	133.02
3	118.24	7	133.02	11	118.24
4	118.24	8	133.02	12	133.02

all expected counts are ≥ 5. ✓

Because our conditions are met... we will perform a χ^2 test for goodness of fit.

DO: $\chi^2 = \frac{(137-118.24)^2}{118.24} + \frac{(121-103.46)^2}{103.46} + \frac{(116-118.24)^2}{118.24} + \ldots = 26.48$

df = 12 − 1 = 11

χ^2 = 26.48

(in calculator: Ball player count = L_1, expected = L_2) ← no need to write this :)

P-value = 0.00549

CONCLUSION: Because our p-value = 0.00549 is less than our significance level α = 0.05, we reject the null. There is convincing evidence that the distribution of MLB players' birthdays deviates from the national pattern.

EXAMPLE 2:

We have counts of 256 randomly selected executives in 12 zodiac sign categories. The natural null hypothesis is that birth dates of executives are divided equally among all the zodiac signs. The test statistic looks at how closely the observed data math this idealized situation. Are zodiac signs of CEO's distributed uniformly?

Births		Signs
Observed Values	Expected Values	
23	$256 \cdot 1/12 = 21.\bar{3}$	Aries
20	$21.\bar{3}$	Taurus
18	$21.\bar{3}$	Gemini
23	$21.\bar{3}$	Cancer
20	$21.\bar{3}$	Leo
19	$21.\bar{3}$	Virgo
18	$21.\bar{3}$	Libra
21	$21.\bar{3}$	Scorpio
19	$21.\bar{3}$	Sagittarius
22	$21.\bar{3}$	Capricorn
24	$21.\bar{3}$	Aquarius
29	$21.\bar{3}$	Pisces
TOTAL = 256		

State: H_0: Birthdates of executives are evenly distributed across all zodiac signs.

H_A: Birthdates of executives are not evenly distributed across all zodiac signs.

$\alpha = 0.05$.

Plan: random: 256 randomly selected executives ✓

10% condition: $n = 256$ $2560 <$ all executives ever ✓

Large counts: if birth dates are evenly distributed across all 12 zodiacs in the year, then the expected count for each zodiac is $256 \cdot 1/12 = 21.\bar{3}$ executives ≥ 5. ✓

Because our conditions are met, we will perform a χ^2 test for goodness of fit.

DO: $\chi^2 = \frac{(23 - 21.\bar{3})^2}{21.3} + \frac{(20 - 21.3)^2}{21.3} + \frac{(18 - 21.3)^2}{21.3} + \ldots \; 5.09$

df = 11

p-value = 0.9265

Conclude: Because our p-value = 0.9265 is greater than our significance level $\alpha = 0.05$, we fail to reject the null. There is not convincing evidence that the birthdates of executives are not evenly distributed among all 12 zodiac signs.