AP Statistics
Unit 08 – HW #1
Chi Square: Goodness of Fit Test

Name Key
Period

1. Acme Toy Company prints baseball cards. The company claims that 30% of the cards are rookies, 60% veterans, and 10% are All-Stars. The cards are sold in packages of 100. Suppose a randomly selected package of cards has 50 rookies, 45 veterans, and 5 All-Stars. Is this consistent with Acme's claim? Use a 0.05 level of significance.

State: H_0: The stated distribution of Acme Toy Company's baseball cards is correct.

$\alpha = 0.05$ H_A: The stated distribution of Acme Toy Company's baseball cards is not correct.

Plan: Random: randomly selected package ✓

10% condition: (Check for each card type) 500 < all rookie cards/players ✓
450 < all veteran cards/players ✓
500 < all All-Star cards/players

Large Counts: expected counts: are all ≥ 5
30 ≥ 5 ✓ rookies
60 ≥ 5 ✓ veterans
10 ≥ 5 ✓ All-Stars

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

DO: $\chi^2 = \frac{(50-30)^2}{30} + \frac{(45-60)^2}{60} + \frac{(5-10)^2}{10} = 19.583$

df = 2

p-value = 0.00005592

Conclude: Because our p-value 0.00005592 is less than our significance level $\alpha = 0.05$, we reject the null. There is convincing evidence that the stated distribution of Acme Toy Company's baseball cards is incorrect.

or, the dist. of students who register for stats classes w/ diff. name is equal across all classes, vs. not (H_A)

2. A study was performed to determine whether or not the name of a course had an effect on student registrations. A statistics course in a large school district was given 4 different names in a course catalog. Each name corresponded to the exact same statistics course. A random sample of student registrations was recorded and the results are given below:

Course Name	Number of Registrations	Expected:
Statistical Applications	25	117 · 1/4 = 29.25
Statistical Reasoning	22	29.25
Statistical Analysis	30	29.25
The Practice of Statistics	40	29.25
TOTAL	117	117

Do these data suggest the name of the course has an effect on student registrations? Conduct an appropriate statistical test to support your conclusion.

State: H_0: The names of statistics courses do not have an effect on the number of students who register for them.

$\alpha = 0.05$

H_A: The names of statistics courses do have an effect on the number of students who register for them.

Plan: Random: random sample of student registrations. ✓
10% condition: $n = 117$, $1170 <$ all students who take statistics ✓
Large counts: all expected counts are $117 \cdot 1/4 = 29.25 \geq 5$ ✓
Because our conditions are met, we will perform a χ^2 test for goodness of fit.

DO: $\chi^2 = \frac{(25-29.25)^2}{29.25} + \frac{(22-29.25)^2}{29.25} + \frac{(30-29.25)^2}{29.25} + \dots$

df = 3
test statistic = 6.385
p-value = 0.09433

Conclude: Because our p-value 0.09433 is greater than our significance level $\alpha = 0.05$, we fail to reject the null. There is not convincing evidence that the names of statistics courses have an effect on the number of students who register for them.

or: H_0: birds do not prefer particular tree types vs. they do. (H_A)

3. Researchers studied the behavior of birds that were searching for seeds and insects in an Oregon forest. In this forest, 54% of the trees were Douglas firs, 40% were ponderosa pines, and 6% were other types of trees. At a randomly selected time during the day, the researchers observed 156 red-breasted nuthatches: 70 were in Douglas firs, 79 in ponderosa pines, and 7 in other types of trees. Do these data provide convincing evidence that nuthatches prefer particular types of trees when they're searching for seeds and insects?

State: H_0: The distribution of red-breasted nuthatches is equal among all tree types in an Oregon forest.

$\alpha = 0.05$ H_A: The distribution of red-breasted nuthatches is not equal among all tree types in an Oregon forest.

Plan: random: randomly selected time ✓ (aka 156 randomly observed birds)

10% condition: 1560 < all red-breasted nuthatches in an oregon forest. ✓

n = 156

Large counts: $156 \cdot 0.54 = 84.24 \geq 5$ douglas firs ✓

$156 \cdot 0.40 = 62.4 \geq 5$ ponderosa pines ✓

$156 \cdot 0.06 = 9.36 \geq 5$ other trees ✓

Because our conditions are met, we will perform a χ^2 test for Goodness Of Fit.

DO: $$\chi^2 = \frac{(70-84.24)^2}{84.24} + \frac{(79-62.4)^2}{62.4} + \frac{(7-9.36)^2}{9.36} = 7.4182$$

df = 2

p-value = 0.024499

Conclude: Because our p-value 0.0245 is less than our significance level $\alpha = 0.05$, we reject the null. there is convincing evidence that red-breasted nuthatches prefer particular types of trees when they're searching for seeds and insects (their distribution is not even among all tree types).

4. Faked numbers in tax returns, invoices, or expense account claims often display patterns that aren't present in legitimate records. Some patterns are obvious and easily avoided by a cleaver crook. Others are subtler. It is a striking fact that the first digits of numbers in legitimate records often follow a model knows and Benford's law. Call the first digit of a randomly chosen record X for short. Benford's law gives this probability model for X (note that the first digit cannot be 0).

1st 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
expected:	75.25	44	31.25	24.25	19.75	16.75	14.5	12.75	11.5

A forensic accountant who is familiar with Benford's law inspects a random sample of 250 invoices from a company that is accused of committing fraud. The table below displays the sample data.

1st Digit	1	2	3	4	5	6	7	8	9
Probability	61	50	43	34	25	16	7	8	6

a. Are these data inconsistent with Benford's law? Carry out an appropriate test where $\alpha = 0.05$ level to support your answer.

State: H_0: The distribution of first digits at a particuliar company follows Benford's law.

$\alpha = 0.05$ H_A: The distribution of first digits at a particullar company does not follow Benford's Law.

Plan: random: random sample of 250 invoices ✓
10% condition: $n=250$ 2500 < all invoices at this company ✓
Large Counts: all expected counts (shown above) are ≥ 5 ✓
Because our conditions are met, we will perform a χ^2 test for Goodness of Fit.

Do: $\chi^2 = \frac{(61-75.25)^2}{75.25} + \frac{(50-44)^2}{44} + \frac{(43-31.25)^2}{31.25} + \ldots$

test statistic = 40.9073 df = 8 p-value = 0.000002169

Conclude: Because our p-value 2.169×10^{-6} is less than our significance level $\alpha = 0.05$, we reject the null. There is convincing evidence that the distribution of first digits for this company does not follow Benford's Law. FRAUDS!

b. Describe a Type 1 error and a Type II error in this setting, and give a possible consequence of each. Which do you think is more serious?

Type I: The dist. does follow the Law but we think it doesn't so we accused them of fraud wrongly.

Type II: The dist. doesn't follow the Law but we think it does so they get away with fraud.