

9.07

Introduction to Statistical Methods

Homework 8

Name: _____

1. Recall the Jurassic Park problem on your first homework assignment. Author Michael Crichton had given the following histogram of the number of dinosaurs of each size:

Image removed due to copyright considerations. Please see histogram charting the height distribution of procompsognathids in: Crichton, Michael. *Jurassic Park*. New York, NY: Ballantine Books, 1993.

We complained that there were two things wrong with this histogram. One is that there seem to be, e.g. 0.7 dinosaurs that have a height of 27cm! There shouldn't be any fractional counts in a histogram. Second of all, we complained that this histogram is too good to be true – it's unlikely with only about 67 dinosaurs that we would get a histogram that looked so much like a normal distribution.

Now that we're talking about χ^2 tests, we have a tool for actually testing whether this histogram is "too good to be true."

My reading of the data on this histogram is that we have the following count-height pairs:

Height (cm)	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41
# dinosaurs	0	1	1	2	4	6	9	10	10	9	6	4	3	1	1	0

Find the parameters of the best fit normal distribution:

(a) What is the mean height of these dinosaurs?

9.07

Introduction to Statistical Methods

Homework 8

Name: _____

(b) What is the standard deviation of the heights? (Use $1/(n-1)$).

Next, we'll use these parameters to generate expected frequencies for each of the bins. Let the first bin include heights < 26.5 cm, the second bin include heights ≥ 26.5 cm and < 27.5 cm, and so on, with the last bin including heights ≥ 40.5 cm.

(c) Find the z-score for each bin boundary (26.5 cm, 27.5 cm, etc.)

(d) For each z-score, find the corresponding area in the z-table. I recommend using a web tool like http://davidmlane.com/hyperstat/z_table.html to get reasonable precision, or the matlab command `normcdf(z)`, which gives you the area under the z-distribution from $-\infty$ to z.

(e) For each bin, find the % of the distribution that we expect to land in that bin.

(f) For each bin, find the expected frequency for that bin, given the number of dinosaurs in the sample.

9.07

Introduction to Statistical Methods

Homework 8

Name: _____

(g) Show the table of observed and expected frequencies, like the one in the examples in class. Compute χ^2 .

(h) Compare χ^2 with the critical value corresponding to 1% of the area under the lower tail. How many degrees of freedom are there? Is this data “too good to be true”?

(i) Do any of the requirements for the χ^2 test make you concerned about the validity of the test you’ve just run?

9.07

Introduction to Statistical Methods

Homework 8

Name: _____

2. A student creates a program to generate uniformly distributed random digits 0, 1, 2, 3, ..., 9. This program generates 250 digits, and we observe the following frequencies:

Digit	0	1	2	3	4	5	6	7	8	9
Observed frequency	17	31	29	18	14	20	35	30	20	36

(a) What are the expected frequencies, for a uniform distribution?

(b) Does the observed distribution differ significantly from the expected distribution?

3. In Mendel's experiments with peas, he observed 315 round & yellow, 108 round & green, 101 wrinkled & yellow, and 32 wrinkled & green. According to his theory, these 4 numbers should be in the proportion 9:3:3:1. Is there any evidence to doubt his theory at the 0.05 level of significance? (Check both whether the data seem to fit the theory, and also whether they seem to be too good to be true.)

9.07

Introduction to Statistical Methods

Homework 8

Name: _____

4. An urn contains a large number of marbles of 4 colors: red, orange, yellow, and green. A sample of 12 marbles drawn at random from the urn yields 4 red, 5 orange, 2 yellow, and 1 green marble. Test the hypothesis that the urn contains equal numbers of the 4 colors of marbles, at the 0.05 level of significance.

5. In the table below are indicated the number of students passed and failed by 3 instructors, each teaching Calculus I. The math department would like the grading by the three instructors to be consistent, so that students with teacher A are not getting a very different distribution of grades than students with teacher B or C.

Are the fraction of students passed and failed by the three instructors equal – i.e. is the grading independent of instructor? Use a two-way χ^2 test to answer this question, at the 0.05 level of significance.

	Instructor A	Instructor B	Instructor C
Passed	52	48	56
Failed	5	13	8

9.07

Introduction to Statistical Methods

Homework 8

Name: _____

6. Do review exercise #3, p. 543.

7. In the general population, the distribution of political party affiliation is 30% Republican, 55% Democrat, and 15% other. Is this distribution also found among the elderly? We sample 100 senior citizens, and find 18 Republicans, 64 Democrats, and 18 other. (a) What are H_0 and H_a ? At the $\alpha=0.05$ level, what do you conclude about the party affiliation in the population of senior citizens?

9.07

Introduction to Statistical Methods

Homework 8

Name: _____

8. The following data reflect the frequency with which people voted in the last election, and whether or not they are satisfied with the officials elected:

	Satisfied	Not satisfied
Voted	48	35
Didn't vote	33	52

You want to test if there is a difference between those who voted and those who did not.

(a) What are the null and alternative hypotheses? (b) With $\alpha=0.05$, what do you conclude about the relationship between the two variables (voting record and satisfaction)?