MATH 338

EXAM 4

THURSDAY, AUGUST 3, 2017

Your name: _____

Your scores (to be filled in by Dr. Wynne):

Problem 1: ____/23.5

Problem 2: ____/11.5

Problem 3: ____/9

Total: ____/44

You have 60 minutes to complete this exam. This exam is closed book and closed notes with the exception of your formula sheet.

For full credit, show all work except for final numerical calculations (which can be done using a scientific calculator).

1. In a 1936 paper, the great statistician (and geneticist!) Ronald Fisher reexamined Gregor Mendel's famous pea plant experiments. In a hybrid cross, according to Mendel's Law of Independent Assortment, there is a 3/4 chance of an offspring having the dominant phenotype and a 1/4 chance of having the recessive phenotype (a phenotype is an observed characteristic determined by genetics).

A) [1.5 pts] If a sample of 7324 seeds is observed from a hybrid cross, what is the exact distribution of the number of seeds with a dominant phenotype (shape/type of distribution, and parameters)?

B(7324, 0.75)

0.5 points each for Binomial and parameters n = 7324 and p = 0.75

B) [2.5 pts] Find the mean and standard deviation of the proportion of seeds with a dominant phenotype, when a sample of 7324 seeds is observed.

0.5 pts mean = p = 0.75

1 pt sd = sqrt(0.75*0.25/7324) = 0.005

1 pt work in the realm of proportions rather than counts, i.e., a total of 1.5 points for mean = np = 5493 and sd = sqrt(7324*0.75*0.25) = 37.06

C) [3 pts] To make sure that a pea plant was a hybrid, Mendel investigated a sample of 10 offspring of a self-fertilization. If any of them had a recessive phenotype, Mendel classified the plant as a hybrid. If the Law of Independent Assortment holds, what is the probability of misclassifying a hybrid as not a hybrid?

1 pt Let X be the number of recessive offspring; then X ~ B(10, 0.25)

1 pt P(correct classification) = P(X $\geq$ 1), so P(misclassification) = P(X = 0)

1 pt P(X = 0) = (10 choose 0) * (0.25)^0 * (0.75)^(10-0) = (0.75)^10 = 0.056

D) [6 pts] Out of 7324 seeds from a hybrid cross, Mendel observed 5474 round (dominant) seeds and 1850 wrinkled (recessive) seeds. Construct and interpret a 95% confidence interval for the true proportion of round seeds in the population of pea plants.

1 pt use a large-sample z confidence interval: CI = $\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

1 pt sample proportion = 5474/7324 = 0.747

1 pt standard error = sqrt((5474/7324)*(1850/7324)/7324) = 0.005

1 pt z*=1.96 for a 95% CI, so the margin of error is 1.96*0.005 = 0.01

1 pt The CI is then 0.747 +/- 0.01 = (0.737, 0.757)

1 pt Correct interpretation: We are 95% confident that between 73.7% and 75.7% of pea plants will have round seeds (or an equivalent correct interpretation)

E) [2 pts] Based on your work in part (D), do you agree with Fisher's statement that "the deviation from the expected 3:1 [ratio] is less than its standard error of random sampling" (the standard error of $\hat{p}$)? Show an appropriate comparison to justify your answer.

Yes, we should agree with this statement, since $SE_{\hat{p}} = 0.005$

but our observed sample proportion is only 0.75 – 0.747 = 0.003 away from the theoretical proportion

F) [2 pts] If we had no idea about the true population proportion of round seeds, how many seeds would we need to sample for our confidence interval in part (D) to be guaranteed to have a margin of error no greater than 3 percentage points?

1 pt set up the equation for conservative margin of error, either $n = \frac{1}{4}\left(\frac{z^*}{E}\right)^2$ or $n = \left(\frac{z^*}{E}\right)^2 (p^*)(1-p^*)$

1 pt solve equation to get $n = \frac{1}{4}\left(\frac{1.96}{0.03}\right)^2 = 1067.11$ or 1068 seeds

G) [6.5 pts] In a dihybrid cross, we expect that the offspring will show a 9:3:3:1 ratio of phenotypes (expected proportions are 9/16, 3/16, 3/16, 1/16). In Mendel's experiments, he observed 315 round, yellow seeds; 108 round, green seeds; 101 wrinkled, yellow seeds; and 32 wrinkled, green seeds. Perform an appropriate hypothesis test to assess whether this generation fits the expected ratio.

1.5 pts We perform a chi-square goodness of fit test with H_0: population proportions are (9/16, 3/16, 3/16, 1/16) vs. H_a: population proportions follow some other distribution, and significance level is defined

1.5 pts expected counts are 312.75, 104.25, 104.25, 34.75, so …

1.5 pt chi-square test statistic is (315 – 312.75)^2/312.75 + (108 – 104.25)^2/104.25 + (101 – 104.25)^2/104.25 + (32 – 34.75)^2/34.75 = 0.47

1 pt we have 4 categories, so we have 3 degrees of freedom and the critical value is either 7.81 or 11.34 depending on significance level

1 pt Fail to reject the null hypothesis, so we conclude that the assumed population distribution is reasonable. In fact, the population distribution fits a little bit too well (the p-value is 0.925) – and it did in every experiment Mendel did. Fisher concluded that Mendel might have been duped by a gardening assistant who knew what Mendel expected the outcome of the experiments to be.

2. In 2017, a Polish group sent 360 academic journals the resume of a fake researcher (whose name translates as "Dr. Fraud"), who asked to become an editor. Some of these journals were on a controversial list of "Predatory" journals that will publish anything for a fee, and some weren't.

A) [2 pts] Fill in the blanks in the two-way table below. The column variable is "What did the journal do to the request?" and the row variable is, "Did the journal appear on the Predatory journals list?".

	Accepted	Denied	No Response	Total
Yes	41	20	69	130
No	7	88	135	230
Total	48	108	204	360

0.5 pts per answer

B) [1 pt] If a journal is selected at random from the 360, what is the probability it did not respond?

204/360 = 0.567 or 56.7%

C) [2 pts] Using the table, estimate the probability that a Predatory journal would accept the request.

P(accept | predatory) = 41/130 = 0.315 or 31.5%

D) [6.5 pts] Is there a relationship between a journal's "Predatory" status and the journal's decision regarding the fake researcher? Perform an appropriate hypothesis test (the table below may help you organize) and state your conclusion with respect to the real-world question asked.

	Accepted	Denied	No Response	Total
Yes	17.33	39	73.67	130
No	30.67	69	130.33	230
Total	48	108	204	360

1.5 pts We perform a chi-square test of independence with H_0: predatory journal and request decision are independent, H_a: they are not independent, and define the significance level

1.5 pts the expected counts are as given in the above table …

1.5 pts …so the chi-square test statistic is (41-17.33)^2/17.33 + … + (135 – 130.33)^2/130.33 = 65.55

1 pt we have (3-1)*(2-1) = 2 degrees of freedom, so the critical value is either 5.99 or 9.21

1 pt in either case, we reject H_0 and conclude that there is a relationship between the journal's being a Predatory journal and the journal's decision regarding the fake researcher.

3. A certain biomarker test will detect breast cancer in 95% of post-menopausal women who have it, but it will also detect breast cancer in 1% of women who don't have it.

A) [1 pt] What is the sensitivity of this test?

0.95 or 95%

B) [1 pt] What is the specificity of this test?

1 – 0.01 = 0.99 or 99%

C) [5 pts] As of 2011, the accepted prevalence (base rate) of breast cancer in post-menopausal women was 0.4% (1 in 2,500 have breast cancer). Calculate the positive predictive value of the test.

1 pt use Bayes' rule or a tree diagram to calculate the positive predictive value

3 pts plug in correctly (1 pt per set of tree diagram branches if using tree diagram):

$$P(cancer\,|test+) = \frac{P(test+|cancer)\,P(cancer)}{P(test+|cancer)P(cancer) + P(test+|no\ cancer)P(no\ cancer)}$$

$$P(cancer|test+) = \frac{(0.95)(0.004)}{(0.95)(0.004) + (0.01)(0.996)} = 0.276$$

1 pt The positive predictive value of the test is about 0.276 or 27.6%

D) [2 pts] A group of researchers evaluated this biomarker test on a sample of post-menopausal women judged to be at high risk for breast cancer, and (correctly) obtained a positive predictive value of over 99%. Why might there be a discrepancy between your answer in part (C) and the researchers' results?

While sensitivity and specificity are not dependent on the base rate of disease, positive predictive value does depend on the base rate of disease. Since these women are judged to be at high risk, the researchers likely assumed that the probability they had breast cancer was much greater than 0.4%. This would have greatly increased the positive predictive value of the test.

Extra Space. The tables below show a number of critical values *z* for the standard normal variable $Z \sim N(0, 1)$ and the corresponding cumulative proportions, corresponding to P(Z ≤ z).

z-score	Cumulative Proportion
-3.00	0.0013
-2.50	0.0062
-2.00	0.0228
-1.65	0.0495
-1.28	0.1003
-1.00	0.1587
-0.67	0.2514

z-score	Cumulative Proportion
0.67	0.7486
1.00	0.8413
1.28	0.8997
1.65	0.9505
2.00	0.9772
2.50	0.9938
3.00	0.9987

Refer to the following table for z* critical values for confidence intervals:

	C = 0.90 (90%)	C = 0.95 (95%)	C = 0.98 (98%)	C = 0.99 (99%)	C = 0.999 (99.9%)
z* values	1.645	1.960	2.326	2.576	3.291

For a two-sided hypothesis test, use the column corresponding to C = 1 – α

For a one-sided hypothesis test, use the column corresponding to C = 1 - 2α

Refer to the following table for χ^2 critical values:

Degrees of freedom	α = 0.05	α = 0.01
1	3.84	6.63
2	5.99	9.21
3	7.81	11.34
4	9.49	13.28
5	11.07	15.09
6	12.59	16.81

The rest of this space to be used for extra work: