Stat 111 Spring 2025 Week 7: Tests and CI's for Count Data

1. **Exact One-Sample Binomial Calculations**
Suppose we observe $X \sim \text{Binom}(n, \theta)$ and wish to make inference about the probability θ. For example, if you test $n = 400$ components and observe $x = 50$ are defective.

   a) Show how to compute the 2-sided P-values reported by binom.test (e.g., for testing $H_o : \theta = 0.1$ vs. $H_a : \theta \neq 0.1$). Explain how, with a discrete test statistic, the 2-sided P-value is the sum of the probabilities for test statistic values that have null probabilities less than or equal to the null probability of the observed statistic value.

   b) Show that the confidence interval generated by binom.test in R with $\alpha = 0.05$ (the default) returns the intersection of two 1-sided 97.5% CI's, defined by inverting 1-sided level 0.025 tests. Show that the probability of getting a Binomial value as large or larger than the observed value is 0.025 if θ is at the lower interval bound, and the probability of getting a Binomial count as small or smaller than the observed value is 0.025 if θ is the upper interval bound.

   c) Use the fact that Uniform order statistics follow a Beta distribution to show the following connection between the Beta and Binomial CDF's.

   ```
   pbeta(p, k, n-k+1) = 1 - pbinom(k-1, n, p)    # for k=1,2,...,n
   ```

   Explain how this equality is used to generate the confidence interval bounds in part b.

   d) Show that, due to the discreteness of the distribution, the actual coverage probabilities for these intervals are greater than 0.95, increasingly so as $|\theta - 0.5|$ increases, but not monotonically in θ. Make graphs showing how the coverage probabilities and mean interval widths change with θ and n.

2. **One-Sample Poisson Calculations** A classic example of Poisson data are the counts of deaths in the Prussian Calvary due to horse-kicks to the head. Data for 200 Corps-Years appear in the table below:

Deaths	Count	Proportion
0	109	0.545
1	65	0.325
2	22	0.110
3	3	0.015
4	1	0.005

   a) Overall there were 122 such deaths in the 10 Army Corps over a 20 year period, for an estimated rate of $\hat{\theta} = 0.61$ deaths per corp-year. For the 200 corps-years, compute the expected counts with 0, 1, 2, and 3 or more deaths. Carry out a Chi-square goodness of fit test to show that the Poisson probabilities seem reasonable (Rice 8.2, 13.3).

   b) For $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Pois}(\theta)$, find the mle and information and construct a large sample approximate 95% CI for θ.

   c) The large-sample approximation considers $Y = \sum X_i \dot{\sim} N(n\theta, n\theta)$, so

$$0.95 \approx P(|Y - n\theta| < 2.0\sqrt{n\theta})$$

   Find the analog to the Binomial plus four CI by solving for the range of $n\theta$ values to make the inequality true. Run simulations to test the coverage probabilities.

d) Use the relationship between the Gamma and Poisson distributions to show the following equality involving their cumulative distribution functions:

```
pgamma(t, k, lambda) = pgamma(lambda, k, t ) = 1 - ppois(k-1, lambda*t)
```

Construct an exact 95% CI for θ.

3. **Binomial Plus-Four Intervals**

a) One year in Stat 11 there were 40 upperclass students (juniors and seniors) and 80 first and second year students. We could model this as the value of a Binomial$(120, \theta)$ random variable, where θ is the overall proportion of Swarthmore students who take Stat 11 as a junior or senior. Use the Normal approximation with the continuity correction to find an approximate p-value for a test of $H_o : \theta = 0.5$ vs. $H_a : \theta \neq 0.5$. Compare to the exact Binomial p-value.

b) For a confidence interval using the Normal approximation, the procedure is not as straightforward, because no null value θ_o is assumed. Show that an upper bound on the Normal approximation to the margin of error for 95% confidence is $1/\sqrt{n}$. The "large-sample" CI sets the margin of error to be $z^*\sqrt{\hat{\theta}(1-\hat{\theta})/n}$, which substitutes the estimated sample proportion $\hat{\theta}$ for the true θ in the standard deviation formula. Compute these CI's for the data in part a.

c) Explain why the test in part a does not show a duality with the CI's in part b. Inverting the test in part a is tricky, but possible. Show that a conservative (and easier to remember) approximate solution is to add 2 successes and 2 failures and construct the large sample CI as though these were actual observations. Demonstrate this Plus-Four confidence interval for a Binomial proportion for the given data. Note how the plus-4 method resembles a Bayesian estimate using a Beta$(2, 2)$ prior distribution.

d) Make a graph as in presentation 1d showing the coverage probabilities of the conservative, large sample and plus four 95% confidence intervals as a function of θ. Show how the plus four method is a nice compromise between the ultra-conservative conservative method and the overly ambitious large-sample method.

4. **Binomial and Chi-square tests**

For 2-sample Binomial problems it is difficult to escape using a Normal approximation of some sort if you want a non-simulation (so repeatable) answer.

a) Describe the "large-sample procedures for testing $H_o : \theta_1 = \theta_2$ vs. $H_1 : \theta_1 \neq \theta_2$, and for confidence intervals for $\theta_1 - \theta_2$, and guidelines for when we can be reasonably confident in the Normal approximation. Show how the standard error used for the 2-sample test is different from the standard error used for the confidence interval, and explain why this means the test results and interval results may not agree for some values of θ_1 and θ_2 (as with the 1-sample approximate test and CI).

b) I ask about coffee consumption in my class surveys. There were 30 coffee drinkers among the 80 first and second year students, and 25 coffee drinkers among the 40 juniors and seniors. Carry out the large-sample approximate test of $H_o : p_1 = p_2$ vs $H_a : p_1 \neq p_2$ and find the large sample 95% CI for $p_2 - p_1$.

c) Show how the 2-sided approximate Z-test of equal Binomial probabilities is exactly equivalent to the Chi-square test of independence, with test statistic Z^2. Define the Pearson Chi-square test of independence and say when the Chi-square approximation will be reasonably accurate. Explain why the Chi-square test hypotheses are equivalent to the 2-sample Binomial test hypotheses.

Describe the test statistic as a measure of the distance between the observed table and a table with the same marginals that shows perfect independence for rows and columns (with possibly fractional expected counts). Explain why the test rejects for large values of the statistic only (and you do not double the p-value). Say what an unusually small values of the Chi-square statistic would indicate. Also show how sample size factor into the Chi-square stat (e.g., what happens if all the counts are doubled?).

d) The two-sample plus-4 method adds one success and one failure to each sample and uses the large sample method as though these four added outcomes were part of the data. The plus-4 two-sample test uses the same adjustment and standard error, so the test fails to reject when the CI contains 0, and does reject when the CI excludes 0. Demonstrate for the coffee data. Run a small simulation for true parameters like the estimated values, and see how close the type-1 error rate is to α.

5. **Poisson and Chi-Square**

The form of the Pearson Chi-square statistic follows naturally from a Poisson representation.

a) Recall the relationship between pairs of independent Poisson counts and Binomial variables. Show how, with the Poisson representation, the Pearson Chi-square statistic is the sum of squared standardized Poisson variables.

b) Suppose for a 2×2 table the four counts are independent Poisson(θ_{ij}) variables, for rows $i = 1, 2$ and columns $j = 1, 2$. Note how conditioning on the row totals results in Binomial variables with probabilities $\phi_1 = \frac{\theta_{11}}{\theta_{11}+\theta_{12}}$ and $\phi_2 = \frac{\theta_{21}}{\theta_{21}+\theta_{22}}$.

c) Jeffreys' non-informative prior for a Poisson rate θ is $p(\theta) \propto \theta^{-1/2} I(\theta > 0)$. Assuming independent priors for the four Poisson rates, show that the joint posterior density for the θ_{ij}'s is that of four independent Gamma($X_{ij} + 1/2, 1$) random variables, and for ϕ_1 and ϕ_2 it is that of two independent Beta($X_{i1} + 1/2, X_{i2} + 1/2$) variables. Note how this agrees with assuming independent Beta($1/2, 1/2$) prior densities for ϕ_1 and ϕ_2. Find a Bayes posterior 95% interval for $\phi_1 - \phi_2$ for the coffee data and compare to the large-sample CI.

d) The Chi-square test generalizes from 2 samples to many samples, and from Binomial to Multinomial. For example, there are four class years and coffee consumption could have more than two categories, for example, 0 cups, 1-4 cups and 5 or more cups per week. The counts are

	0 cups	1-4 cups	5 or more	Total
Fr	21	11	5	45
So	21	9	5	35
Jr	10	7	8	25
Sr	5	4	6	15
	65	31	24	120

Carry out a Chi-square test on these data and explain your conclusions. In particular, note which cells contribute the most to the Chi-square statistic.