Midterm

Short Answer

[5 pts] 1. Researcher Joe wants to analyze internet traffic flowing in and out of computer labs on the ISU campus. He talk to the technical support person in Snedecor Hall, who agrees to set up a logger on 20 randomly selected machines in Snedecor labs. After a week, Joe collects the logs and prepares a report, including many summary statistics, about internet traffic in ISU computer labs. Identify the *population* in this study.

Solution: The actual population is computers in Snedecor Hall labs. Joe might like to make conclusions about the population of computer labs or computers in labs on campus, but he has not sampled randomly from such computers or labs on campus.

[5 pts] 2. You have a multi-agent system that models forest fires and fighting them. You run it n times and record whether the fire is contained within a pre-specified amount of time, or not. Y_i indicates whether trial $i \in \{1, 2, \ldots, n\}$ is contained. What is an exact sampling distribution for statistic $X_n = Y_1 + \cdots + Y_n$ (provide name and parameter(s) of the distribution)?

Solution:

$$X_n \sim \text{Binomial}(n, p)$$

where p is the probability of containment.

[5 pts] 3. Continuing the above question, what is an approximate distribution for X_n, especially close to the truth as n increases (provide name and parameter(s) of the distribution)?

Solution:

$$X_n \sim N(np, np(1-p))$$

[5 pts] 4. Find the 55% sample percentile in the following (ordered) sample:

$$0.0106, 0.1564, 0.2088, 0.2452, 0.2497, 0.3786, 0.4288, 0.7888, 0.8430, 0.9462$$

Solution: 0.3786

[5 pts] 5. Consider $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, 1)$. You know that the sample mean $\bar{X}$ is an estimator of μ. The sample median $\phi_{0.5}$ also estimates μ. It can be shown that the sample mean has $\text{Var}(\phi_{0.5}) = \frac{\pi}{2N}$. Which estimator is more efficient?

Solution: The sample mean is more efficient, because it has $\text{Var}(\bar{X}) = \frac{1}{n} < \frac{\pi}{2n}$.

Worked Problems

[10 pts] 6. (a) Find a 90% confidence interval for population mean μ given population variance $\sigma^2 = 9$ and data that yields $\bar{X} = 3.4$.

Solution:

$$\left(3.4 - \frac{3\phi_{0.95}}{\sqrt{n}}, 3.4 + \frac{3\phi_{0.95}}{\sqrt{n}}\right)$$

(b) Suppose (L, U) is the confidence interval you found in part a. What is wrong with the following expression?

$$P(L < \mu < U)$$

Solution: There is nothing random among the three numbers μ, L, and U, so it makes no sense to write the probability of event $\{L < \mu < U\}$. It is true or it isn't. The probability that it is true for still-to-be calculated L and U, where they are random because they depend on as-yet unseen data X, is 90%, assuming the data come from a normal distribution with the specified variance..

[15 pts] 7. Suppose you observe $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Gamma}(k, \theta)$, where the probability density function of this parameterization of the Gamma is given as

$$f(x) = \frac{x^{k-1} e^{-x/\theta}}{\theta^k \Gamma(k)}, \quad x > 0.$$

Write down the likelihood function $L(k, \theta)$. Use the log likelihood to find the maximum likelihood estimate $\hat{\theta}$ in terms of the other parameter k. (There is no closed form solution for $\hat{k}$, so I don't recommend you try anything more than that!)

Solution:

$$L(\theta, k) = \prod_{i=1}^{n} f(X_i) = \prod_{i=1}^{n} \frac{X_i^{k-1} e^{-X_i/\theta}}{\theta^k \Gamma(k)}$$

$$
\begin{aligned}
&= \frac{e^{-\sum_{i=1}^{n} X_i/\theta} \prod_{i=1}^{n} \left(X_i^{k-1}\right)}{\theta^{nk} \left[\Gamma(k)\right]^n} \\
l(\theta, k) &= -\frac{1}{\theta} \sum_{i=1}^{n} X_i - nk \ln \theta - C(k) \qquad C(k) \text{ is a constant with respect to } \theta \\
\frac{dl}{d\theta} &= \frac{n}{\theta^2} \bar{X} - \frac{nk}{\theta} \qquad \text{set to 0 to find } \hat{\theta} \\
\hat{\theta} &= \frac{\bar{X}}{k}
\end{aligned}
$$

[15 pts] 8. Students were asked to *intuitively* (without calculation) provide the probability of two events A and B. The instructor collecting the data suspected that students who were paying attention in class would intuitively understand that event A was more probable than event B. The data are shown in the table on the left below. X_A is the probability (reported as percentage) for event A and X_B for event B, reported for each of 16 students.

Subject	X_A	X_B
1	78	78
2	24	24
3	64	62
4	45	48
5	64	68
6	52	56
7	30	25
8	50	44
9	64	56
10	50	40
11	78	68
12	22	36
13	84	68
14	40	20
15	90	58
16	72	32

Sample size	Critical Value
6	0
7	2
8	3
9	5
10	8
11	10
12	13
13	17
14	21
15	25
16	29

Hesitant to assume normality of the data and doubting whether the CLT would help for $n = 16$, the instructor performed a non-parametric test. Please repeat the instructor's calculations up until the point of calculating a p-value. The second table provides the critical values for a two-tailed test with type I error set to $\alpha = 0.05$ (the result is significant, if you want to verify your work).

Solution: The signed rank is appropriate because the data are naturally paired (each individual provides two responses, one each for event A and B). The differences are already conveniently ranked. They are

$$X_A - X_B = 0, 0, 2, -3, -4, -4, 5, 6, 8, 10, 10, -14, 16, 20, 32, 40$$

The ranks are

$$\text{ranks} = 1.5, 1.5, 3, 4, 5.5, 5.5, 6, 7, 8, 9.5, 9.5, 11, 12, 13, 14, 15$$

and the sum of the negative signed ranks will be smaller. The statistic is

$$W = 4 + 5.5 + 5.5 + 11 = 26$$

which is smaller than the critical value 29. Since the sum of the ranks of the smaller collection of negatively signed data is unusually small, we reject the null of equal medians (means if the data are from a normal or symmetric distribution).

[15 pts] 9. You do a quick little survey of computer usage by male and female faculty in your department, producing the following data table.

	Female	Male
PC	0	5
Mac	4	1

Compute the p-value for rejecting the null hypothesis of no association of computer preference and gender.

Solution: The various configurations of the data with the same row and column sums is given below:

$$\left|\begin{matrix} 0 & 5 \\ 4 & 1 \end{matrix}\right| \quad \left|\begin{matrix} 1 & 4 \\ 3 & 2 \end{matrix}\right| \quad \left|\begin{matrix} 2 & 3 \\ 2 & 3 \end{matrix}\right| \quad \left|\begin{matrix} 3 & 2 \\ 1 & 4 \end{matrix}\right| \quad \left|\begin{matrix} 4 & 1 \\ 0 & 5 \end{matrix}\right|$$

The probability of the configurations are

$$\frac{\binom{5}{0}\binom{5}{4}}{\binom{10}{4}} \quad \frac{\binom{5}{1}\binom{5}{3}}{\binom{10}{4}} \quad \frac{\binom{5}{2}\binom{5}{2}}{\binom{10}{4}} \quad \frac{\binom{5}{3}\binom{5}{1}}{\binom{10}{4}} \quad \frac{\binom{5}{4}\binom{5}{0}}{\binom{10}{4}}$$

The first and the last configuration are equally likely, and all others must be more likely, so the p-value is

$$2 \times \frac{\binom{5}{0}\binom{5}{4}}{\binom{10}{4}} = \frac{2 \cdot 5 \cdot 4 \cdot 3 \cdot 2}{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6} = \frac{1}{3 \cdot 2 \cdot 7 \cdot 3} = 0.0079$$

[20 pts] 10. Suppose it has been long known that two algorithms, A and B, differ in efficiency, i.e. mean run-lengths μ_A and μ_B, with $\mu_B < \mu_A$. You've produced algorithm C to compete with algorithm B, and you think it has different efficiency than B (you hope better, but you are not sure). You can't prove the result, but instead run both algorithms A and C on a test set, for which it is well known that $\mu_B - \mu_A = 0.4$. (For proprietary reasons, you cannot run algorithm B.) Your data yields $\bar{X}_A = 1.4$ for $n_A = 50$ and $\bar{X}_C = 1.95$ also for $n_C = 50$. In addition, your estimates of standard deviation are $S_A = 1.42$ and $S_C = 1.03$.

(a) Assuming the variance of algorithms A and C are equal, can you say with confidence $1-\alpha = 0.95$ that your algorithm C better than algorithm B? Use notation $t_{\text{df}}\left(\frac{2-\alpha}{2}\right)$ to refer to quantiles of the t distribution with df degrees of freedom.

Solution: We want to test

$$H_0 : \mu_C - \mu_A = 0.4$$

The pooled sample variance is

$$S_p^2 = \frac{49 \times 1.42^2 + 49 \times 1.03^2}{98} = \frac{1.42^2 + 1.03^2}{2} = 1.54$$

Our statistic

$$T = \frac{\bar{X}_C - \bar{X}_A - 0.4}{S_p/5} \sim t_{48}$$

follows a t distribution with 48 degrees of freedom when the null hypothesis is true. If $|T| > t_{48}(0.975)$, then we reject the null hypothesis and conclude algorithm C is different from B. We could only conclude algorithm C was *better* than B using a one-sided test, but in the statement of the problem we only *hope* C is better than B. That is not good enough to commit to a one-sided test.

(b) What sample size would be required to detect an improvement of 0.2 run time in algorithm C over B with 90% power? (Keep $\alpha = 0.05$ and assume the variance estimate provided by this data is the true common variance.)

Solution:

$$\begin{aligned}
0.90 &= P\left(|T| > t_{48}(0.975) \mid \mu_C - \mu_A = 0.6\right) \\
&\approx P\left(T > t_{48}(0.975) \mid \mu_C - \mu_A = 0.6\right) && T \text{ most likely falls in right tail} \\
&= 1 - P(T \le t_{48}(0.975) \mid \mu_C - \mu_A = 0.6) \\
0.10 &= P(\bar{X}_C - \bar{X}_A \le \sqrt{2/n} S_p t_{48}(0.975) + 0.4 \mid \mu_C - \mu_A = 0.6) \\
&= P(Z \le t_{48}(0.975) - 0.2/(\sqrt{2/n} S_p)) \\
\phi_{0.10} &= t_{48}(0.975) - 0.2\sqrt{n/2}/S_p
\end{aligned}$$

$$n = 2\left[\frac{S_p(\phi_{0.10} - t_{48}(0.975))}{0.2}\right]^2$$

You can accomplish this task using the usual power calculation for $H_0 : \mu_C = \mu_B$ because both null and alternative distributions are merely shifted by a factor of 0.4 and probability calculations are unchanged.