Spring 2024, MATH 408, Final Exam

Wednesday, May 1; 11am–1pm

Instructor — S. Lototsky (KAP 248D; x0–2389; `lototsky@usc.edu`)

Instructions:

- No notes, no books, or other printed materials (including printouts from the web), no collaboration with anybody (or anything, like AI).
- You should have access to a calculator or some other computing device, and to the normal, t, χ^2, and F distribution tables. Instead of the tables, you are welcome to use the statistical functions available on your computing device.
- Answer all questions, show your work, and clearly indicate your answers; upload the solutions to GradeScope.
- **Each problem is worth 10 points.**

A summary of main distributions

Name	Notation	pdf/pmf	Mean	Variance
Binomial	$\mathcal{B}(n,p)$	$\binom{n}{k}p^k(1-p)^{n-k}$	np	$np(1-p)$
Poisson	$\mathcal{P}(\lambda)$	$e^{-\lambda}\lambda^k/k!$	λ	λ
Beta	$\mathrm{Beta}(a,b)$	$\dfrac{x^{a-1}(1-x)^{b-1}}{B(a,b)}$	$\dfrac{a}{a+b}$	$\dfrac{ab}{(a+b)^2(a+b+1)}$
Gamma	$\mathrm{Gamma}(a,\theta)$	$\dfrac{\theta^a x^{a-1}e^{-\theta x}}{\Gamma(a)}$	$\dfrac{a}{\theta}$	$\dfrac{a}{\theta^2}$
Normal	$\mathcal{N}(\mu,\sigma^2)$	$(2\pi\sigma^2)^{-1}\exp\big(-(x-\mu)^2/(2\sigma^2)\big)$	μ	σ^2

Problem 1. Given the set of numbers 30, 55, 60, 70, 65, and assuming that this is an independent random sample from a normal population, construct a 95% confidence interval for the mean. Show your work by filling in the corresponding numerical values:

- sample mean $\bar{X}_n = 56.000$
- $s_n = 15.572$
- the quantile of the t distribution you use: `with` $n=5$ `and` $\alpha/2 = 0.025$`, get` $t_{4,0.025} = 2.776$,
- the final answer: $\bar{X}_n \pm \frac{s_n}{\sqrt{n}}\, t_{4,0.025} \approx 56 \pm 19 = [37, 75]$

Problem 2. Let $X_1, \ldots, X_n$ be an independent random sample such that the pdf of each X_k is

$$f(x;\theta) = \frac{1}{2\theta^3}\, x^2 e^{-x/\theta},\ x>0,\quad \theta>0.$$

Construct the maximum likelihood estimator $\hat{\theta}$ of θ. Make sure to verify that you indeed maximized the likelihood function.

Solution/Answer: the likelihood function is $2^{-n}\theta^{-3n}\left(\prod_{k=1}^{n} X_k^2\right)\cdot\exp\left(-n\bar{X}_n/\theta\right)$, with $\bar{X}_n$ denoting the sample mean. Then the (equivalent) function to maximize is $\ell(\theta) = n(-3\ln\theta - \bar{X}_n\theta^{-1})$. The equation $\ell'(\theta) = 0$ gives $-3\theta^{-1} + \bar{X}_n\theta^{-2}$ or $\hat{\theta} = \bar{X}_n/3$, which, indeed, is the global max of the function: $\ell'(\theta) > 0$ if $\theta < \hat{\theta}$ and $\ell'(\theta) < 0$ if $\theta > \hat{\theta}$; note that ℓ has an inflection point at $\theta = 2\hat{\theta}$.

Problem 3. Suppose that two independent random samples from two populations X and Y resulted in the following numerical values for the sample mean and standard deviation:

$$\bar{X}_n = 11.2,\ s_{n,X} = 8.5,\ \bar{Y}_n = 10.5,\ s_{n,Y} = 7.0.$$

Assume that $n = 550$. Can you claim that the population mean of X is significantly bigger than the population mean of Y? Justify your conclusion by computing the corresponding P value.

Solution/Answer: using large sample approximation (more precisely, combining the CLT with LLN and the Slutsky theorem), $\bar{X}_n$ is approximately normal with mean μ_X (the population mean of X) and variance $\sigma_X^2/n \approx s_{n,X}^2/n$; $\bar{Y}_n$ is approximately normal with mean μ_Y (the population mean of Y) and variance $\sigma_Y^2/n \approx s_{n,Y}^2/n$.
Then $\sqrt{n}(\bar{X}_n - \bar{Y}_n)(s_{n,X}^2 + s_{n,Y}^2)^{-1/2}$ is approximately standard normal and the corresponding one-sided test statistic is $\phi = \sqrt{n}(\bar{X}_n - \bar{Y}_n)(s_{n,X}^2 + s_{n,Y}^2)^{-1/2}$. The observed value is $\phi^* = 1.5$ corresponding to the P value $\mathbb{P}(Z > 1.5) = 0.068 > 0.05$. Therefore, you cannot claim that the population mean of X is significantly bigger than the population mean of Y.

Problem 4. Let $X_1, \ldots, X_n$ be an independent random sample from the distribution with pdf

$$f(x;\theta) = \frac{1}{2\theta^3}\, x^2 e^{-x/\theta},\quad x > 0,\ \theta > 0.$$

Construct the most powerful test of $H_0 : \theta = 2$ against $H_1 : \theta = 5$ at the level $\alpha = 0.05$.

Solution/Answer: The ratio of likelihoods that we want to be small when rejecting H_0 is, up to a constant, $\exp\left((-0.5+0.2)n\bar{X}_n\right) = \exp(-0.3n\bar{X}_n)$, where $\bar{X}_n$ is the sample mean. As a result, we reject H_0 if $n\bar{X}_n$ is LARGE.
Following the notations in the table, the original distribution is $\mathrm{Gamma}(3, 1/\theta)$. Because, under the null hypothesis $\theta = 2$ the population is $\mathrm{Gamma}(3, 1/2)$, we conclude that, under the null hypothesis, $n\bar{X}_n$ is $\mathrm{Gamma}(3n, 1/2)$, so that the rejection rule at level 0.05 is $n\bar{X}_n \geq \mathrm{Gamma}(3n, 1/2)_{0.05}$, where the quantile on the right corresponds to the ''area to the right of the point''. From problem 2 and the table, we know that the sample mean is MLE for 3θ, that is, the bigger the θ, the bigger the value of $\bar{X}_n$ we expect to measure, confirming that the rejection rule makes sense. Note also that $\mathrm{Gamma}(3n, 1/2) = \chi^2_{6n}$.

Problem 5. For the first-year students at a certain university, the correlation between SAT scores and the amount of money borrowed to pay for the study was -0.36. Assume the joint distribution of the SAT scores and the borrowed money is normal. Predict the percentile rank on the amount of money borrowed for a student whose percentile rank on the SAT was 85%.

Answer: with Φ denoting the standard normal cdf, we use one of the versions of the regression line to compute the predicted ranking as $\Phi(-0.36\Phi^{-1}(0.85)) = \Phi(-0.36\cdot 1.0364) =$

$1 - 0.64546 \approx 35\%$.

Problem 6. Below is part of a two-way ANOVA table. Fill out the rest of the table.

Source	SS	df	MS	F	Prob $> F$
Blocks	80	4	20	2.105	0.118045835...
Treatments	210	5	42	4.42	0.007124324...
Error	190	20	9.5		
Total	480	29			

Answers are in smaller font. Note that, using the F distribution table, you can only conclude that $\mathbb{P}(F_{4,20} > 2.105) > 0.1$ and $\mathbb{P}(F_{5,20} > 4.42) \in (0.005, 0.01)$

Problem 7. To test whether a die is fair, 64 rolls were made, and the corresponding outcomes were as follows:

Face value	Observed frequency
1	8
2	9
3	15
4	15
5	9
6	8

Would you consider the die fair using a χ^2 goodness of fit test? Justify your conclusion.

Solution: for the value φ^* of the test statistic, which is the sum of observed-minus-expected-squared-over-expected, with expected equal to $64/6 = 32/3$, we get

$$\varphi^* = 2((24-32)^2 + (27-32)^2 + (45-32)^2)/(32 \cdot 3) = (64 + 25 + 169)/48 = 5.375$$

and the P-value is

$$\mathbb{P}(\chi^2_5 > \varphi^*) = 0.37184731...$$

Using the table, you can only conclude that the P-value is bigger than 0.1 (and less than 0.9).
Either way, the answer to the question in the problem is YES: based on the P-value, there are no reasons NOT to conclude that the die is fair.

Problem 8. Assume that the following is an independent random sample from population X with a continuous cdf $F_X(x) = F(x)$:

$$14.4 \quad 15.5 \quad 13.3 \quad 12.1 \quad 12.2,$$

and assume that the following is an independent random sample from population Y with cdf $F_Y(x) = F(x+\theta)$:

$$18.8 \quad 15.0 \quad 10.7 \quad 9.4 \quad 10.6.$$

Compute the P-value of the sign test for the null hypothesis $\theta = 0$ against the alternative $\theta > 0$. Note that the alternative means that the random variable X is more likely to be large, that is, $\mathbb{P}(X > Y) > 1/2$.

Solution: with the test statistic $M = \sum_{k=1}^{5} I(X_k > Y_k)$ we get $M^* = 4$ (except for the first pair, all other X samples are bigger than the corresponding Y samples), and therefore
P-value= $\mathbb{P}(\mathcal{B}(5, 1/2) \geq 4) = \mathbb{P}(\mathcal{B}(5, 1/2) = 4) + \mathbb{P}(\mathcal{B}(5, 1/2) = 5) = (5 + 1) \cdot 2^{-5} = 3/16$.

Problem 9. For the two samples in Problem 3, compute the Spearman rank correlation coefficient.

Solution. For the ranks of X, that is, the positions of X_k in the sample arranged in increasing order, we get $4, 5, 3, 1, 2$; the corresponding ranks of Y_k are $5, 4, 3, 1, 2$ and the sum of the squares of the differences of the ranks is 2.
Using the ''no-tie formula'' for r_s, with $n = 5$, we get $r_s = 1-(6\cdot 2)/(5\cdot 24) = 1-0.1 = 0.9$.

Problems 10. In a large class, the number of students absent at a particular lecture can be modeled as a Poisson random variable with mean value λ; the value of λ can be estimated using the Bayesian approach.

Suppose there were three students absent at the first lecture; this suggests $\text{Gamma}(3, 1)$ as a prior distribution for λ. Assuming all the independence you need, compute the posterior distribution of λ if, over the next $n = 20$ lectures, there were a total $N = 71$ absences. Use the resulting posterior mean to compute the (posterior) probability that lecture number 22 will have full attendance.

Solution. Using the idea of conjugate priors (Gamma/Poisson), we conclude that, given the prior $\text{Gamma}(3, 1)$, the posterior is $\text{Gamma}(3+N, n+1)$, with posterior mean $\lambda^* = (3+N)/(n+1) = 74/21 \approx 3.5$. The (posterior) probability of having full attendance at any subsequent lecture is then

$$\mathbb{P}\Big(\mathcal{P}(\lambda^*) = 0\Big) = e^{-\lambda^*} \approx 0.03.$$