Spring 2024, MATH 408, Mid-Term Exam 1: `Solutions and more`

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 standard deviation. Show your work by filling in the corresponding numerical values:

- sample mean = 56.000
- $s_n = 15.572$
- the two quantiles of χ^2 you use: with $n = 5$, use χ^2_4, so that the upper quantile is 11.143 (area above/to the right is 0.025), the lower is 0.483 (area above is 0.975)
- the final answer: $[9, 45]$

More: How will your solution change if, instead of the individual numbers, you are given the summary statistics $\sum_{k=1}^n X_k$, $\sum_{k=1}^n X_k^2$? How will you proceed if n is large (e.g. $n = 256$)?

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

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

Construct the maximum likelihood estimator of θ. Make sure to verify that you indeed maximized the likelihood function.

`Solution/Answer: the likelihood function is` $2^{-n}\theta^{3n}\left(\prod_{k=1}^n X_k^2\right)\cdot\exp\left(-n\theta\bar{X}_n\right)$, `with` $\bar{X}_n$ `denoting the sample mean. Then the (equivalent) function to maximize is` $\ell(\theta) = n(3\ln\theta - \theta\bar{X}_n)$. `The equation` $\ell'(\theta) = 0$ `gives` $\theta = 3/\bar{X}_n$, `which is the global max of the function:` $\ell''(\theta) = -3/\theta^2 < 0$. `Thus,` $\widehat{\theta}^{(MLE)} = 3/\bar{X}_n$.

More: think how you would construct a confidence interval for θ given numerical values of n and $\bar{X}_n$. You can try both normal approximation (for large n) and the exact pivot based on the distribution of $n\bar{X}_n$. Then think how to construct a confidence interval for $1/\sqrt{\theta}$ or some other non-linear function of θ.

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 $n = 550$. Is there a (statistically) significant **difference** between the population means of X and 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` $\sigma_X^2/n \approx s_{n,X}^2/n$.

`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 two-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` $\mathbb{P}(|Z| > 1.5) = 0.136$, `meaning that the numbers do not indicate a statistically significant difference between the population means.`

More: assuming the same values for the sample means and standard deviations, think about the following questions. Will the conclusion change if $n = 500$ (or any $n < 550$, for that matter) and why? What is the smallest value of n that would lead to a different conclusion? What if we are only interested in whether the sample mean of X is bigger?

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

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

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, $\exp\Big((-2+5)n\bar{X}_n\Big) = \exp(3n\bar{X}_n)$, where $\bar{X}_n$ is the sample mean. As a result, we reject H_0 if $n\bar{X}_n$ is small.

Call the original distribution $\mathrm{Gamma}(3,\theta)$. Because, under the null hypothesis $\theta = 2$ the population is $\mathrm{Gamma}(3,2)$, we conclude that, under the null hypothesis, $n\bar{X}_n$ is $\mathrm{Gamma}(3n,2)$, so that the rejection rule at level 0.05 is $n\bar{X}_n \le \mathrm{Gamma}(3n,2)_{0.95}$, where the quantile on the right corresponds to the ''area to the right of the point''. From problem 2 we know that the sample mean is MLE for $3/\theta$, that is, the bigger the θ, the smaller the value of $\bar{X}_n$ we expect to measure,
confirming that the rejection rule makes sense.

More: assuming $n = 30$, what is the power of this test? See if you get upper and lower bounds using the χ^2 tables and a more precise value using your favorite statistical software.

Problem 5. For the first-year students at a certain university, the correlation between SAT scores and first-year GPA was 0.36. Assume the distribution of the scores is jointly normal. Predict the percentile rank on the first-year GPA for a student whose percentile rank on the SAT was 85%.

Answer: with Φ denoting the standard normal cdf, we compute the predicted ranking as $\Phi(0.36\Phi^{-1}(0.85)) = \Phi(0.36 * 1.0364) = 0.64546 = 65\%$

More: how will your answer change if, for some strange reasons, the correlation is negative: $\rho = -0.36$? Can you think of a situation where the correlation is naturally negative and propose a corresponding problem?