Short Answers to Quiz #5 (Optional Quiz), Stat 50, Spring 2025

1. (0.75 pts) Let p denote the actual (unknown) proportion of all cars registered in a state, whose emission levels exceed the acceptable state standards. Consider the following hypotheses about p:

   $$H_0 : p \geq 0.25 \text{ versus } H_1 : p < 0.25.$$

   Based on a random sample of cars in the state, about 20% of the cars tested had emission levels above the state standards, and an upper-bound confidence interval for p at 95% confidence level was computed to be $(0, 0.262]$. Using the evidence from the interval above, can you reasonably reject H_0 and conclude that p is smaller than 0.25? Explain.

   Answer. Even though the sample proportion, $\hat{p} = 0.2$, is below 0.25, the upper-bound confidence interval covers proportions as high as 0.262. So, we don't have compelling evidence to reject H_0.

2. The average investment amount in the bio-tech industry in a state is believed to be at around 20 million dollars. We want to test if this amount is accurate.

   (a) (0.5 pts) Construct the relevant null and alternative hypotheses of a two-tailed test about the mean investment amount, μ (in million dollars).

   Answer. $H_0 : \mu = 20$ versus $H_1 : \mu \neq 20$, which is a two-tailed test.

   (b) (0.5 pts) In a random sample of 16 bio-tech investments in the state, the sample average was 18.15 million dollars with a sample standard deviation of 2.40 million dollars. Moreover, the sample didn't include any outlier. Explain why a z-test doesn't apply to this scenario.

   Answer. The assumptions of a z-test are not satisfied. Since the population std. dev. (σ) is unknown and we don't know if the population distribution is roughly bell-shaped, the sample size is too small to even consider an approximate z-test procedure.

   (c) (1 pt) Does a t-test apply to the sample in part (b)? Justify your answers. If you said "yes", then compute also the value of the corresponding t-statistics. You don't need to estimate a P-value in this problem.

   Answer. The assumptions of a t-test are not fully satisfied but we can use an approximate t-test method relying on the sample standard deviation ($s = 2.40$) in the absence of outliers.

3. Data science experts of an engineering company run a large-scale program for a machine learning application to estimate its average run time (in seconds) on a particular computer. In 81 trials of the simulation study, the average run time was 28.0 seconds with a standard deviation of 4.5 seconds.

   (a) (1 pt) The experts wonder if it can be concluded that the average run time of the program is below 30 seconds. To this end, the following left-tailed test is designed:

   $$H_0 : \mu \geq 30 \text{ versus } H_1 : \mu < 30.$$

   Briefly explain whether a z-test or a t-test can be applied to this problem, and then determine the relevant test statistic (z_0 or t_0).

   Answer. The assumptions of a t-test are satisfied (a sufficiently large random sample when population distribution is not guaranteed to be normal and with unknown σ) with $df = 80$. An approximate z-test would also apply since n is very large (more than 50), but the more relevant test statistic is $t_0 = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} = \frac{28-30}{4.5/\sqrt{81}} = -4.0$.

   (b) (1.5 pts) Using a P-value approach, conduct a suitable hypothesis test at 5% level of significance, and state your conclusion in context. Show your work and justify your steps.

   Answer. P-value is the left tail area under a t-curve (with $df = 80$) determined by the cutoff point, $t_0 = -4$. Using a t-table (with positive t-scores), we observe that this would be a very small area, even less than 0.005 (using *tcdf* function of Matlab, we actually get $P = 7.022 \times 10^{-5}$). Similarly, an approximate z-test would give $P = P(Z \leq -4) < 0.0001$ ($P \simeq 3.2 \times 10^{-5}$ using *normcdf* function of Matlab).