Statistical Techniques for CS/BIT

Final Test – Solutions

1. **a.** The 5-number summary is: $x_{(1)} = 75.1$, $Q_1 = \frac{x_{(5)}+x_{(6)}}{2} = 83.9$, $m = \frac{x_{(10)}+x_{(11)}}{2} = 88.5$, $Q_3 = \frac{x_{(15)}+x_{(16)}}{2} = 92.2$, $= x_{(20)} = 102.2$.
The 1.5×IQR interval is $(Q_1 - 1.5 \times \text{IQR}, Q_3 + 1.5 \times \text{IQR}) = (71.45, 104.65)$. There are no observations outside this range.

b. The skewness and kurtosis are very close to the reference values (0 and 0); definitely less than 2 standard deviations away. We can also see that the sample mean and median are very similar, suggesting symmetry in the distribution. Normality seems to be a reasonable assumption.

c. For *any* of the usual α values (1%, 5%, 10%), the statistic $W = 0.985$ is not included in the rejection region. Conclusion: there is no evidence suggesting the data is non-normal.

2. **a.** The data is numeric (from question 1. we know it is likely normal), we have one sample, and the sample size $n = 20$ is relatively small. The population mean and variance are unknown. The correct tests to apply are:

(1) A 1-sample t-test on the difference.

(2) A sign test on the median (note that from question 1. we know that the sample mean and sample median are very similar).

b. We apply the t-test in 8 steps.

1. The speeds $X_1, \ldots, X_{20}$ are independent and follow a normal distribution with unknown mean μ and unknown variance σ^2.
2. We test $H_0 : \mu \geq 90$ versus $H_1 : \mu < 90$ at $\alpha = 5\%$.
3. The test statistic is $T = \dfrac{\overline{X} - \mu_0}{S/\sqrt{n}} = \dfrac{\overline{X} - 90}{S/\sqrt{20}}$.
4. Under the null hypothesis we have $T \sim t_{19}$.
5. The observed value is $t = \dfrac{88.36 - 90}{6.85/\sqrt{20}} = -1.07$.
6. Since this is a left-sided test (cf. the hypotheses), we reject H_0 if $T \leq c$. From the t_{19} table we get $c = -1.729$.
7. The observed value $t = -1.07$ is not in the rejection region, so we fail to reject H_0.
8. At a 5% level of significance the sample does not provide enough evidence that the expected upload speed is less than 90 Mbit/s.

c. For the sign test we have:

(1) Assumptions: the observations are a random sample; the population mean and median are the same.

(2) If X denotes the number of observations smaller than 90, then $X \sim \text{Binom}(20, p)$. We observed $x = 12$. The p-value is $\mathbf{P}(X \geq 12 \,|\, p = 0.5) = 1 - \mathbf{P}(X \leq 11 \,|\, p = 0.5) = 1 - 0.748 = 0.252$.

(3) From such a large p-value we conclude: there is not enough evidence that the expected upload speed is less than 90 Mbit/s.

3. The statements **b**, **d**, **e**, and **g** are correct.

4. **a.** The 95%-CI(p) $= (\hat{p} - c\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + c\sqrt{\frac{\hat{p}(1-\hat{p})}{n}})$
with: ϕ (c)=1-$\frac{1}{2}\alpha$, $\alpha = 0.975$, so c=1.96, n=100 (large enough) and $\hat{p} = \frac{50}{100}$=0.5.
95%-CI(p) $\approx (0.5 - 0.098, 0.5 + 0.098) = (0.402, 0.598)$

b. Using $\hat{p} = 0.6$ and $c = 1.96$ like in part **a**, the condition on the width is as follows:
$2 \times c\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq 0.04$
$\Rightarrow n \geq (\frac{2c}{0.04})^2 \times \hat{p}(1-\hat{p})$

$\Rightarrow n \geq (\frac{2\times 1.96}{0.04})^2 \times (0.6 \times 0.4)$
$\Rightarrow n \geq 98^2 \times (0.6 \times 0.4) \Rightarrow n \geq 2304.96 \Rightarrow n = 2305$

5. In this question, the givens are n=50, $\overline{x} = 4.05$ hrs, σ=0.2 hrs. From previous tests, the life of these batteries follows a normal distribution and that the standard deviation is known

a. We test for μ while the standard deviation σ (and the variance σ^2) is known. Therefore we do a *normal* test (and not a t-test).
The null and alternative hypothesis are as follows: $H_0 : \mu \leq 4$ versus $H_1 : \mu > 4$ at $\alpha = 5\%$
OR
$H_0 : \mu = 4$ versus $H_1 : \mu > 4$ at $\alpha = 5\%$. The test statistic is $\overline{X}$

b. Power is the probability of avoiding a Type II error OR as equivalent, the power of the test is the probability to reject $H_0 : \mu = 4$ if in reality $\mu = 4.15$.
We first need to determine the rejection region for the test. The critical value c is given by

$$P(\overline{X} \geq c \,|\, \mu = 4) = 5\%.$$

This means

$$\frac{c-4}{0.2/\sqrt{50}} = 1.645, \quad \text{which means} \quad c = 4.047.$$

Power is $P(\overline{X} \geq c \mid \mu = 4.15) = 1 - P(\overline{X} \leq 4.047 \mid \mu = 4.15)$,

$$\phi\left(\frac{4.047 - 4.15}{0.2/\sqrt{50}}\right) = \phi(-3.64) = 0.0001.$$

Therefore the power is 0.9999, which is almost 100%.

6. Note that we have two distinct groups: early eaters and late eaters. Each individual is either in one group or the other. In particular, Subject #1 of Group #1 is different from Subject #1 of Group #2. Therefore, we have two **independent** samples, and we want to estimate the difference in the population means.
(The differences were reported but are irrelevant for the problem!)

a. In order to choose the test we analyze the situation:

- The variables are numeric.
- The population means and variances are unknown.
- From the Q-Q plots we see no strong deviation from the line, so we may assume the populations are normally distributed.

- We have two independent samples.
- The sample variances are very similar. Moreover, the p-value of 0.814 in the F-test confirms that it is safe to assume equal variances.

From this information we deduce that the correct test to apply is a *2 independent samples t-test on the difference of the means*. The assumptions of the test are: independent random samples, normality, and equal variances; which are all satisfied.

b. The confidence interval $(0.106, 7.577)$ excludes the value 0. We can state with 95% confidence that the difference in weight loss is greater than zero. Equivalently, the weight loss for early eaters is significantly different from that for late eaters at a significance level of 5%.

However, this **does not** prove the author's claim as stated in $(*)$. The statement "The earlier one eats during the day, the larger the weight loss will be", implies that even within one group (eg. early eaters) the weight loss is even greater for those who eat extra early. But our test only compares the two groups, and has no way of further analyzing the relationship between weight loss and timing of meals.

c. The statement is incorrect. The symbol μ represents the true difference between the population means. This is a fixed constant, even if unknown to us. Either μ is inside the interval or it is not. We don't know, but it is not a matter of probability because μ is not a random variable.

7. This problem is clearly about a test on a population proportion.

a. The binomial test is the standard choice here. When the sample size is large enough, we may approximate the binomial distribution by a normal one (and thus perform a Z-test). But the sample size is not particularly large, and we can perfectly work with the exact probabilities of the binomial distribution without having to approximate anything. A t-test would never apply in this situation.

b. For a binomial test the only assumption is that we have a **random sample** of observations; meaning they are independent and come from the same population. In this experiment all the samples come from the same population. However we **cannot** claim that the samples are independent! The psychologist polled 21 children from 10 households. This means that there are multiple instances where siblings were polled. Their responses are not independent: it is likely that they will provide the same response because they have the same parents.

c. The psychologist's claim is invalid, because it was derived from a test in which the assumptions were not met.