TECH3 V25 Trial exam solutions

Problem 1:

a) A probability has to be non-negative and be in the interval [0,1].

b) The Central Limit Theorem states that if n becomes large, the sample mean $\overline{X}_n$, is asymptotically normally distributed.

c) Correlation does not imply causation. Therefore, X being correlated with Y, does **not** imply that X causes Y.

d) It is true that if an estimator is consistent, its variance approaches zero when the sample size becomes large. It is not enough though - the estimator must also be unbiased.

e) This is just wrong. The significance level should depend on the situation, and the consequence of making Type I or Type II errors. The "standard" of always using 5% is flawed.

f) The code is a 10-fold cross validation.

Problem 2:

a) Let X denote the gain from a bet. It will then have the distribution $P(X = 1) = 18/37$ and $P(X = -1) = 19/37$. We further have:

b) The expected value:

$$E(X) = \sum_x xP(X = x) = 1 \cdot P(X = 1) + (-1) \cdot P(X = -1) = 1 \cdot \frac{18}{37} + (-1)\frac{19}{37} = -\frac{1}{37} = -0.027.$$

ii) The variance can be found by first finding

$$E(X^2) = \sum_x x^2 P(X = x) = 1^2 \cdot P(X = 1) + (-1)^2 \cdot P(X = -1) = 1^2 \cdot \frac{18}{37} + (-1)^2 \frac{19}{37} = \frac{37}{37} = 1,$$

and then using the short-cut formula:

$$\text{Var}(X) = E(X^2) - E(X)^2 = 1 - (-1/37)^2 = 0.9993.$$

iii) The probability of a positive gain is the probability that the gain equals: $P(X > 0) = P(X = 1) = 18/37 = 0.4865$

b)

According to the CLT, the mean of 50 rounds of gambling, has

i) The same expected value as one game of gambling; $\frac{-1}{37}$.

ii) Variance equal to $\text{Var}(X)/n = 0.9993/50 = 0.0200$

iii)

$$P(\bar{X}_{50} > 0) = P\left(\frac{\bar{X}_{50} - \mu}{\sigma/\sqrt{n}} > \frac{0 - \mu}{\sigma/\sqrt{n}}\right) = P(Z > \frac{0 - (-0.027)}{\sqrt{0.02}})$$
$$= P(Z > 0.19) = 1 - P(Z \leq 0.19) = 0.4247$$

```
from scipy import stats
print(1-stats.norm.cdf(0.19))
```

```
## 0.4246545652652045
```

c) To solve (a) using simulation: I would simulate 10,000 times bets by drawing a random variable that is 1 with probability 18/37 and -1 with probability 19/37. This would give me 10,000 realizations of the outcome of the bet. Let us denote them

$$x_1, \ldots, x_{10\,000}.$$

i) To estimate the expectation would be to calculate the mean of the 10,000 simulated bets:

$$E(X) \approx \overline{x} = \frac{1}{10\,000} \sum_{i=1}^{10\,000} x_i.$$

ii) The variance can be found by calculating the empirical variance of the 10,000 bets:

$$\mathrm{Var}(X) \approx \frac{1}{9999} \sum_{i=1}^{10\,000} (x_i - \overline{x})^2.$$

iii) To find the probability of positive gain, I would find the relative frequency of positive gains:

$$P(X > 0) \approx \frac{1}{10\,000} \sum_{i=1}^{10\,000} I(x_i > 0),$$

where $I(x_i > 0)$ is an indicator function being 1 if $x_i > 0$ and zero otherwise.

Python code for this (not part of the question):

```
import numpy as np
x = np.random.choice([1, -1], size=10000, p=[18/37, 19/37])
print("Expected value:", np.mean(x))
```

```
## Expected value: -0.037
```

```
print("Variance:", np.var(x))
```

```
## Variance: 0.9986310000000002
```

```
print("Probability of positive gain: ", np.mean(x>0))
```

```
## Probability of positive gain:  0.4815
```

Problem 3:

a) $H_0 : \mu = 20\,000$ and $H_A : \mu > 20\,000$.

b) In the formula, $\mu_0 = 20\,000$, so if the sample mean is larger than 20,000, T_1 will be positive, and this would then favor the alternative hypothesis. Therefore, we will reject the null hypothesis for large values of T_1.

c) $t_1 = \frac{21\,253 - 20\,000}{2200/\sqrt{45}} = 3.82$. According to the text, T_1 is student-t distributed with $n - 1$ degrees of freedom. It is therefore the second alternative Python code that is relevant. This gives the probability $P(T_1 < t_1)$, but the p-value is

$$\text{p-value} = P(T_1 > t_1) = 1 - P(T_1 \leq t_1) = 1 - 0.9999334435960688 = 6.655 \cdot 10^{-5}.$$

The p-value is very low, which gives strong evidence against the null hypothesis.

d) $H_0 : \mu_1 = \mu_2 + 5000$ vs $H_A : \mu_1 > \mu_2 + 5000$.

e) For T_2, d_0 is 5000. So if H_0 is true, we expect $\bar{X}_1$ to be near $\bar{X}_2 + 5000$, and T_2 should be close near zero. If H_A is true, we expect $\bar{X}_1$ to be larger than $\bar{X}_2 + 5000$, and thus the numerator of T_1 is positive. Since the denominator of T_2 is always positive, this means that T_2 should be positive if H_A is true. We therefore reject the null hypothesis if T_2 is large.

f) The test statistic is

$$t_2 = \frac{\bar{x}_1 - \bar{x}_2 - d_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} = \frac{21\,253 - 15990 - 5000}{\sqrt{2200^2/45 + 1500^2/45}} = 0.663.$$

We find, from the hint, that

$$\text{p-value} = P(T_2 > t_2) = 1 - P(T_2 \leq t_2) = 1 - 0.745 = 0.255.$$

g) The price of the premimum lightbulb is 50% more, but the premium light bulb only lasts about 33% longer than the economy one ($21\,253/15\,990 = 1.33$). If the decision variable is money invested per hour of light, the economy light bulb would be preferred. If you hate changing the light bulb or take into environmental aspects to your decision, it may be worth the additional cost to prolong the life expectancy of the light bulb.

Problem 4:

a)

$$\log P_{\text{wind}} = \log(\frac{1}{2}\rho A v^3) = -\log 2 + \log \rho A + 3 \log v.$$

The relationship between logarithmic wind power potential and logarithmic wind speed is linear.

b) The fitted model for data point i, can be written as

$$\log P_i = \beta_0 + \beta_1 \log v_i + \epsilon = 7.11 + 2.86 \cdot \log v_i + \epsilon_i.$$

c) The variance explained by the model is $R^2 = 98.3\%$ - which is a **very high** R^2.

d) The intercept is the estimated logarithmic wind power at a wind speed of 1 m/s (when $\log v = 0$). This means, at 1 m/s, we expect logarithmic power generation to be 7.11.

The slope parameter associated with wind speed is 2.86. This means that an increase in logarithmic wind speed of 1 unit, is associated with an increase in logaritmic wind power of 2.86. According to the theoretical expression for the relationship between wind power and wind speed in (a), this number should be close to 3, which is the case here. We are not able to extract the full theoretical power avaiable in the wind, but not that far from it. Note, however, that 3 is not included in the 95% confidence interval for β_1 (in the notation used in (b)), so a hypothesis test would reject the null hypothesis that $\beta_1 = 3$ against a two sided alternative ($\beta_1 \neq 3$).

e) The assumptions are:

√ Linearity

√ Independent errors

× Constant variance

√ Normality

√ No multicolinearity (not relevant here - only one predictor)

The linearity is fulfilled based on the log-log plot in the exercise text. The first plot in Appendix B indicate no systematic pattern between fitted values and residuals (indication of independent errors), but the spread

is increasing with the fitted value, indicating violation of constant variance assumption - we do not have homoskedastic residuals. The qq-normality plot shows some deviations from normality in the tails - heavier tails than for the normal distribution. In my opinion, these are not major deviations, and I would have no problems assuming normality here. The histogram looks symmetric around zero.

f) i) At $v = 8\text{m/s}$, we get a predicted value for the logaritmic power of:

$$\widehat{\log p} = 7.11 + 2.86 \cdot \log 8 = 13.06$$

corresponding to $\widehat{p} = e^{\widehat{\log p}} = e^{13.06} = 469\,770W = 470\text{kW}$.

ii) Correspondingly, for $v = 17\text{m/s}$:

$$\widehat{\log p} = 7.11 + 2.86 \cdot \log 17 = 15.21.$$

$$\widehat{p} = e^{\widehat{\log p}} = e^{15.21} = 4\,032\,915\,\text{W} = 4.03\,\text{MW}.$$

The data used to train the model includes wind speeds from roughly 3 to 12 m/s. While the first prediction, with 8 m/s, is well within this interval, the second, 17 m/s, is outside the range of wind speeds in the data. This prediction is therefore an extrapolation, and there more uncertain. I would trust the first prediction, but perhaps not the second one. For instance, we know that wind turbines reach a maximum power they are able to produce and, at some point, must be shut off if the wind speed is too high. Since we do not have data for these wind speeds, we cannot tell whether this is the case here.

g) The prediction interval is always wider than a confidence interval, because it also takes into account the uncertainty of a new observation, not just the uncertainty in the mean.