Statistical Methods – midterm exam solutions – November 21, 2022

Question 1

a) The statement is wrong because convenience samples are based on easily available responses which typically leads to a biased picture (e.g. if family and friends are asked).
Correction: "A convenience sample is typically biased and thus not representative for the population one is interested in."

b) When the sum of both probabilities, $P(A)$ and $P(B)$ is taken, their intersection is counted twice. Thus, it has to be subtracted once:
Correction: "$P(A \cup B) = P(A) + P(B) - P(A \cap B)$."

c) In right-skewed samples, there are typically some very large (not small) "outliers", or, in any case, there are typically not a lot of data points much smaller than the center of the data distribution, compared to those data points larger than the center.
Correction: For right-skewed samples, there are typically many observations much bigger than the sample median, rather than observations which are much smaller than the sample median.

d) Independence and disjointness often exclude each other.
Correction: Two independent events are typically not disjoint and two disjoint events are typically not independent. / Two independent events are not necessarily disjoint and two disjoint events are not necessarily independent.

Question 2

a) $\Omega = \{A, B, \ldots, Z\}^5, \quad P(\omega) = 1/26^5$ for all $\omega \in \Omega$.
(Note: Ω has $26^5 = 11,881,376$ elements.)

b) We are interested in the conditional probability $P(A|B)$, where $B = \{GREAT, GRTEA\}$ and A is a simple event containing either of GREAT or GRTAE.
Since A is contained in B, we have $P(A|B) = (1/26^5)/(2/26^5) = 1/2$.

c)

$$\begin{aligned} E(X) &= \sum_x P(X = x) \cdot x = 0 \cdot 1,000 + 0.01 \cdot 500 + 0.05 \cdot 200 + 0.1 \cdot 50 + 0.2 \cdot 10 + 0.4 \cdot 1 + 0.24 \cdot 0 \\ &= 5 + 10 + 5 + 2 + 0.4 = 22.4. \end{aligned}$$

Question 3

a) Let $A = \{\text{test positive}\}$, $B = \{\text{infected}\}$.
We know: $P(A|B) = 0.95$, $P(\bar{A}|\bar{B}) = 0.98$, and $P(B) = 0.01$.

We are looking for:
$P(A) = P(A|B) \cdot P(B) + P(A|\bar{B}) \cdot P(\bar{B})$ (law of total probability)
$= 0.95 \cdot 0.01 + (1 - 0.98) \cdot 0.99 \approx 0.0293 \approx 0.029$.

b) We are looking for:
$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A|B) \cdot P(B) + P(A|\bar{B}) \cdot P(\bar{B})}$ (Bayes' theorem)
$= \frac{0.95 \cdot 0.01}{0.0293}$ (shortcut using result from a) in combination with the law of total probability for $P(A)$ in the denominator.)
$\approx 32.42\%$. (32.76% if the rounded outcome 0.029 from a) is used.)

c) Let $A_i = \{\text{Test } i \text{ is positive.}\}$, $i = 1, 2$.
We are looking for:
$P(A_1 \cap \bar{A}_2|B) + P(\bar{A}_1 \cap A_2|B)$
$= 2 \cdot P(A_1 \cap \bar{A}_2|B)$
$= 2 \cdot P(A_1|B) \cdot P(\bar{A}_2|B)$ (independence of tests)
$= 2 \cdot 0.95 \cdot 0.05 = 0.095$.

Question 4

a) The law of large numbers applies because we're repeating the same experiment over and over.
Thus, we approximate an average score of $\frac{1}{100}\sum_{i=1}^{100} X_i \approx E(X_1) = 1 \cdot P(X_1 = 6) + 0 \cdot P(X_1 \neq 6) = 1/6$.

b) We are looking for: $P(\overline{X}_{100} < 0.1) \approx P\Big(Z < \frac{0.1-1/6}{\sqrt{(\frac{1}{6}\cdot\frac{5}{6})/100}}\Big) \approx P(Z < -1.789) \approx 0.0367$. (Table 2)
Here, Z is $N(0,1)$-distributed and the first approximation is due to the Central Limit Theorem.

(If they use the "wrong" result 0.2 from a), the resulting z-score in b) will be about -2.683 and the resulting probability 0.0037.)

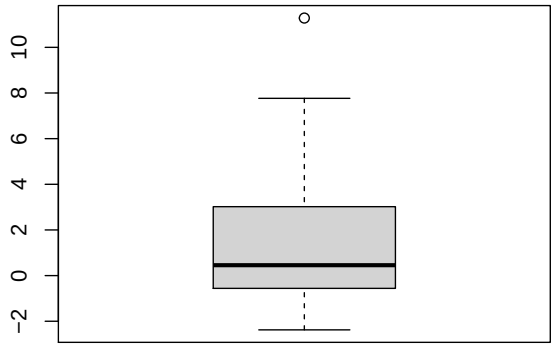
c) $P(Z > 1.24) = 1 - P(Z \leq 1.24)$
$1 - 0.8925 = 0.1075$.

Question 5

a) **Between 1 and 2:**
Similarities: Both are symmetric.
Differences: 2 has heavier left and right tails.

Between 3 and 4:
Similarities: none
Differences: 3 has a heavier left tail
but a lighter right tail than 4.
Thus, 3 is symmetrical whereas 4 is asymmetrical.

b) Something that ressembles the correct boxplot:



c)
A QQ-plot plots the quantiles of one distribution against the quantiles of another. The distributions could be data distributions or theoretical distributions (on either axis).

With a QQ-plot you could check whether two distributions could be modelled to come from the same location-scale family. Each of both distributions could be a data distribution or a theoretical distribution.

If the points of a QQ-plot roughly follow a straight line, it seems appropriate that the distributions are the same location-scale family.

In the given QQ-plot, the tails of the data distribution seem heavier (left tail) or much heavier (right tail) than the tails of the normal distribution.

Thus, it doesn't make sense to model the data normally.