Midterm Exam Statistical Methods for CS (X_401020) and AI (XB_0080)

Vrije Universiteit Amsterdam, Faculty of Science

18.45 – 20.30h, November 21, 2022

- In general: **always motivate your answers.**
- Write your answers in English.
- You are allowed to use only a simple, non-graphical, non-programmable calculator.
- The exam is closed-book. Some formulas and tables are available on the last 2 pages of this exam.
- The total number of points available is 40.5 and your midterm grade is the number between 1 and 10 determined as Grade $= 1 + \frac{\text{points}}{4.5}$.
- The points are awarded for each question and its parts as follows:

Question	1	2	3	4	5
Part a)	2	2	3	2	3
Part b)	2	1.5	3.5	4	2.5
Part c)	2	3.5	2.5	2	3
Part d)	2	–	–	–	–
Total	8	7	9	8	8.5

1. Explain why the following statements are wrong, and reformulate them in a way, such that the statement about *convenience sampling, probabilities of unions of events, right-skewed samples, and independence* is correct.

   a) A convenience sample is generally representative for the whole population one is interested in.

   b) The probability of the union of the events A and B, i.e. $A \cup B$, is equal to the sum of the probabilities $P(A) + P(B)$.

   c) For right-skewed samples, there are typically many observations much smaller than the sample median, rather than observations which are much greater than the sample median.

   d) Two independent events are always disjoint but two disjoint events are not necessarily independent.

2. A game similar to the game *Wordle* asks the player to guess a 5-letter word. In each round of the game, it is displayed how many and which of the just-typed letters are at the correct position (highlighted in dark grey: A) and which letters appear in the word but are at a wrong position (highlighted in light grey: A).

   In the following, we assume for simplicity that the 26 'regular' letters A-Z are used and that also non-existent 5-letter words (like 'AAAAA') can be solution words. All letters (at each position) are assumed equally likely.

   a) Give the sample space Ω and the probability measure P for the first round of the game.
   Hint: because there are too many possibilities, think of an economically feasible way to summarize the sample space.

b) Suppose that, after 4 rounds of the game, the following intermediate result has been established:
G R A T E
(The first two letters, GR are highlighted in dark grey and the last three letters, ATE, are highlighted in light gray.)
Find the conditional probability that the word will be correctly guessed in the next round, given the available information.
Remember: also non-existent words are allowed.
Also, it is expected of you that all available information is used, e.g. that the first two letters have already been found.

The game uses a scoring system which gives more points if the solution word is guessed in early rounds of the game. The following table summarizes the scores depending on the round of the game and also the (rounded) probabilities (determined from vast records) that the solution word was guessed in a certain round (up to 6). If the word has not been guessed after round 6, the game declares that the word has not been guessed at all and the game ends.

Round k of the game	1	2	3	4	5	6	not at all
Score for guessing the word in round k	1,000	500	200	50	10	1	0
Probability of guessing the word in round k	0	0.01	0.05	0.1	0.2	0.4	0.24

Let the random variable X model the score of the game.

c) Compute the expectation $E(X)$ of X.

Note: for c), state the formula for the expected value that you are using.

---

3. Suppose that currently 1% of Dutch citizens is acutely infected with SARS-CoV-2. Furthermore, suppose there is an antigen test which correctly detects an actual infection (result: "positive") with a probability of 95% (sensitivity). If one does actually not have the virus, the test correctly confirms this (result: "negative") with a probability of 98% (specificity).

   Suppose we ask a randomly selected person to apply the antigen test to him-/herself.

   a) Compute the probability that the test result is positive.
   b) If the test result was positive, compute the conditional probability that this person is indeed infected.
   c) Calculate the probability that, for two *independent* tests applied to the same *actually infected* person, one of the tests is negative and the other is positive.

   Note: always name the formulas or theorems that you are using, if they have a name, and repeat the formulas in the way you are using them.

---

4. The computer science student Jake is a board game fan. He would like to calculate or approximate some probabilities and other important parameters that are relevant for his favorite board game. Rolling 6's on a fair 6-sided die plays a major role in this: in particular, each roll of a 6 gives a score of 1, whereas all other die roll outcomes give a score of 0.

   a) After 100 independent rolls of the fair die, indicate and motivate what is approximately the average score obtained per die roll.
   b) After 100 independent rolls of the fair die, calculate the (approximate) probability that the average score is less than 0.1.
   Tip 1: you may use without further verification that the random variable

$$X_i = \text{ score obtained by the } i\text{-th die roll}, \quad i = 1, \ldots, 100,$$

   has a variance of $\sigma^2 = 5/36$.
   Tip 2: the result from a) also plays a role in finding the solution in b). If you were not able to solve a), you may (wrongly) assume for solving b) that the answer in a) is 0.2.

c) Unrelated to the previous questions:
find the probability that a standard normally distributed random variable is bigger than 1.24.

Note: always name the formulas or theorems that you are using, if they have a name, and repeat the formulas in the way you are using them.

---

5. a) The two plots in Figure 1 below display two density functions each. Make qualitative comparisons of densities 1 and 2 and also of densities 3 and 4 by pointing out the most important similarities and differences.

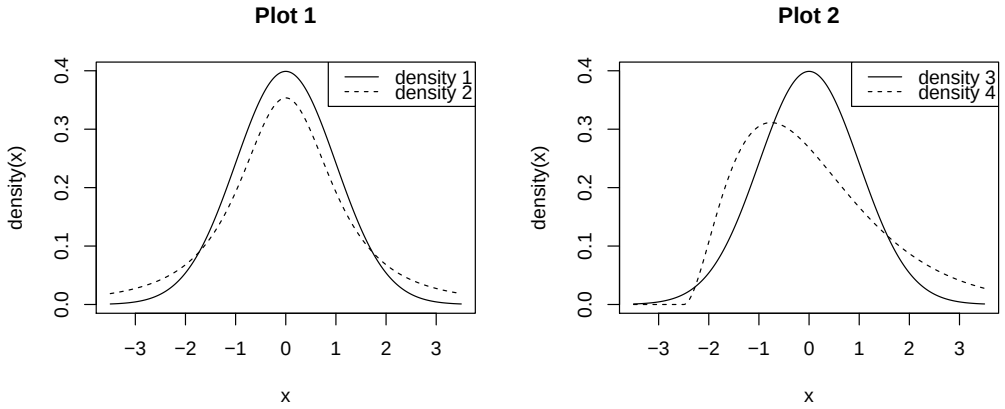

Figure 1: In each plot, two model distributions are compared.

b) Draw a boxplot that represents a random sample (of sufficiently large size) drawn from the distribution illustrated by density4. *Note: it is impossible to draw a completely accurate boxplot; but the most important features should be visible.*

c) What is a QQ-plot: Describe what's on the axes, what you can check with a QQ-plot, and how you can check that. In Figure 2 we see a *normal* QQ-plot. What can you conclude about the tails of the data distribution shown in this plot.

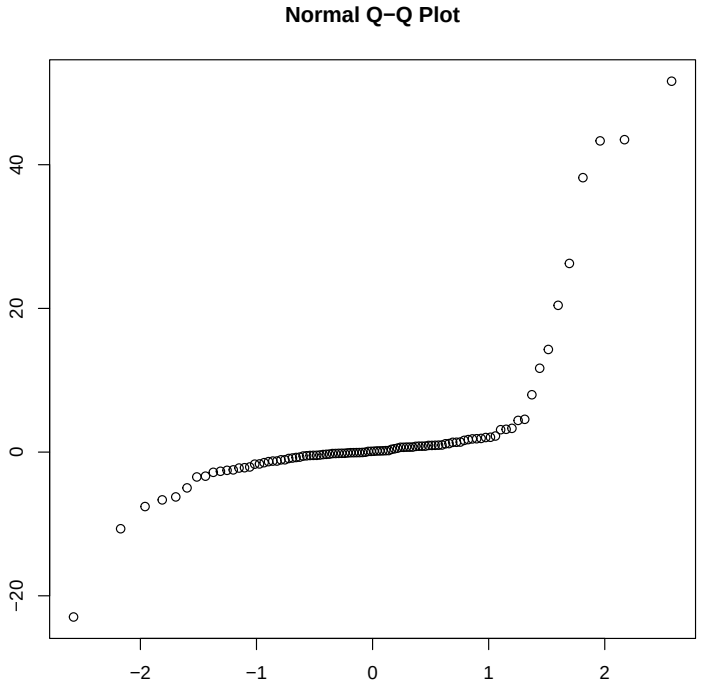

Figure 2: Normal QQ-plot. The labels of the axes are intentionally left blank, because that's question 2c.

Formulas and Tables for Exam Statistical Methods

Probability

We use the following notation:
Ω sample space, P probability measure.
$B, A_1, A_2, \ldots, A_m$ events,
$A_1, A_2, \ldots, A_m$ a partition of Ω with $P(A_i) > 0$ for all $i \in \{1, 2, \ldots, m\}$.

Law of Total Probability:

$$P(B) = \sum_{i=1}^{m} P(B \cap A_i) = \sum_{i=1}^{m} P(B|A_i)P(A_i).$$

Bayes' Theorem:

$$P(A_r|B) = \frac{P(A_r \cap B)}{\sum_{i=1}^{m} P(B|A_i)P(A_i)} = \frac{P(B|A_r)P(A_r)}{\sum_{i=1}^{m} P(B|A_i)P(A_i)}.$$

NEGATIVE *z* Scores

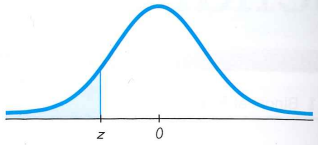

Table 2 Standard Normal (z) Distribution: Cumulative Area from the LEFT

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.50 and lower	.0001									
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051 *	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505 *	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

(continued)

NOTE: For values of z below 3.49, use 0.0001 for the area.
*Use these common values that result from interpolation:

z Score	Area
−1.645	0.0500
−2.575	0.0050

POSITIVE *z* Scores

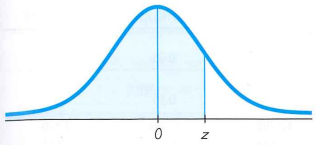

Table 2 *(continued)* Cumulative Area from the LEFT

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495 *	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949 *	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998
3.50 and up	.9999									

NOTE: For values of z above 3.49, use 0.9999 for the area.
*Use these common values that result from interpolation:

z Score	Area
1.645	0.9500
2.575	0.9950

Common Critical Values

Confidence Level	Critical Value
0.90	1.645
0.95	1.96
0.99	2.575