Sample Questions

Question 4 [25 marks]

This question is based on real data.
In order to analyse the impact of English language proficiency on students' academic performance in the QMUL-BUPT Joint Programme, we randomly collected average marks in English modules* and final weighted average marks for 100 graduated students from 4 cohorts. The sample data are saved in a pandas DataFrame `english_finalmarks_sample_df`, with the average English marks* stored in the column '`English Average Mark`' and the final weighted average marks stored in the column '`Weighted average`'.
**The average mark in English modules or average English mark for each student is the average mark of 3 English modules each student took during their 4-year studies.*

a) Please choose a suitable type of plot and fill in the blanks in the following code snippet to visualise the relation between the 100 students' average English marks and their final weighted average marks when they graduated.

[4 marks]

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

fig, ax = plt.subplots(figsize=(4, 3))
sns._____(data=_____, x="_____", y="_____", ax=ax)
plt.show()
```

b) Please fill in the blanks in the following code snippet to compute the Pearson correlation coefficient between the 100 students' average English marks and their final weighted average marks when they graduated.

[3 marks]

```
import numpy as np
import pandas as pd

english_finalmarks_sample_df['_____'].corr(english_finalmarks_sample_df[_____'],
                                           method='_____')
```

c) Suppose we want to make an inference about the correlation between students' average English marks and their final weighted average mark**s** for all Joint Programme students based on the sample data. Please perform a one-tailed t-test with a significance level of 5% to determine the statistical significance of the positive Pearson correlation coefficient based on the sample data of 100 students. Suppose all assumptions for the t-test have been met.
You can refer to the formulas and the t-table in the **Appendix** at the end of this paper.

[12 marks]

d) Briefly describe how to generate the null distribution if we choose to perform a permutation test to test for the significance of the **positive** Pearson correlation coefficient based on the data stored in `english_finalmarks_sample_df`.

[6 marks]

Answer

a)

`` import numpy as np import pandas as pd import matplotlib.pyplot as plt import seaborn as sns fig, ax = plt.subplots(figsize=(4, 3)) sns.scatterplot(data=english_finalmarks_sample_df, x="English Average Mark", y="Weighted average", ax=ax) plt.show() ``` *(1 mark for each blank, case sensitive)*	[4 marks]
TOTAL	**[4 marks]**

b)

`` import numpy as np import pandas as pd english_finalmarks_sample_df['English Average Mark'].corr(english_finalmarks_sample_df['Weighted average'], method='pearson') ``` *(1 mark for each blank, case sensitive)*	[3 marks]
TOTAL	**[3 marks]**

c)

Set up hypothesis: $H_0: \rho \leq 0$ *(2 marks)* $H_1: \rho > 0$ *(2 marks)*	[4 marks]
sample size n=100 *(1 mark)* t-value: $t = r\sqrt{\frac{n-2}{1-r^2}}$ *(1 mark)* $= 0.68\sqrt{\frac{100-2}{1-0.68^2}} = 9.18$ *(1 mark)*	[3 marks]
The hypothesis test is a one-tailed (right-tailed) t-test. So, by looking up the t-table with $\alpha = 0.05$ and df = 98 *(1 mark)* we can get the critical value $t_{0.05,98} = 1.66$ *(1 mark)*	[2 marks]
Because $\|t\| = 9.18 > t_{0.05,98} = 1.66$ *(1 mark)*, we reject the null in favour of the alternative. *(1 mark)* The correlation between students' average English marks and their final weighted average marks in the Joint Programme is statistically significant. *(1 mark)*	[3 marks]

TOTAL	**[12 marks]**

d)

We can 1. explore all possible pairings between the 100 students' English average marks and their final weighted average marks by permuting values in column '`English Average Mark`' and the values in column '`Weighted average`'. *(2 marks)* OR: explore all possible pairings between the 100 students' English average marks and their final weighted average marks by permuting values either in column '`English Average Mark`' or the values in column '`Weighted average`'. *(2 marks)* 2. Calculate Pearson correlation coefficient for each possible pairing between the 100 students' English average marks and their final weighted average marks. *(2 marks)* 3. The Pearson correlation coefficients for all possible pairings form the null distribution. *(2 marks)*	[6 marks]
TOTAL	**[6 marks]**

Appendix

t-Test for Pearson/Spearman correlation coefficients

t-value:

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

where n is the sample size, and r is the correlation coefficient.

t-Table

Table entry for p and C is the point t^* with probability p lying above it and probability C lying between $-t^*$ and t^*.

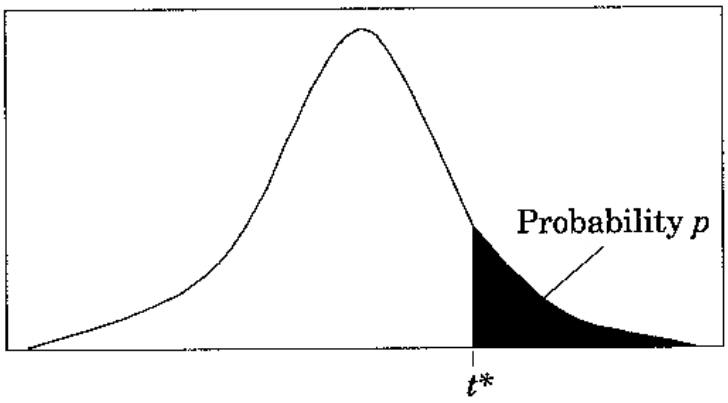

Table B t distribution critical values

	Tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	.765	.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	.741	.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	.727	.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	.718	.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	.711	.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	.706	.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	.703	.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	.700	.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	.697	.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	.695	.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	.694	.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	.692	.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	.691	.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	.690	.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	.689	.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	.688	.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	.688	.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	.687	.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	.686	.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	.686	.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	.685	.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	.685	.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	.684	.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	.684	.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	.684	.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	.683	.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	.683	.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	.683	.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	.681	.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	.679	.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	.679	.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	.678	.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	.677	.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	.675	.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
∞	.674	.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level C											