Past-paper walk-through

Describing the Distribution of a Quantitative Variable ▪ 1.6

eXercise

Questions in this set

- 2024 Q6
- 2023 Q1
- 2021 Q1
- 2019 Q1
- 2016 Q1

Reading the mark scheme

- **B1** a mark that stands on its own – the point is either made or not.
- **M1** a *method* mark: the right approach, even if the number is wrong.
- **A1** an *answer* mark, and it usually needs its M mark first.
- **C1** a working mark on the way to the answer.
- // separates answers the examiner accepts equally.
- **ecf** = error carried forward: a later part is marked on your own earlier answer.
- **owtte** = "or words to that effect" – the wording need not match.

Question 6

2024 ▪ Q6

Begin your response to **QUESTION 6** on this page.

SECTION II, Part B

Suggested Time—25 minutes

1 Question

Directions: Show all your work. Indicate clearly the methods you use, because you will be scored on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.

6. A company sells a certain type of whistle. The price of the whistle varies from store to store. Julio, a statistician at the company, wants to estimate the mean price, in dollars ($), of this type of whistle at all stores that sell the whistle.

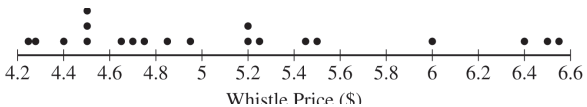
   (a) (i) Identify the appropriate inference procedure for Julio to use.

Question 6

(ii) Describe the parameter for the inference procedure you identified in part (a-i) in context.

Julio called the managers of 20 randomly selected stores that sell the whistle and recorded the price of the whistle at each store. Following is a dotplot of Julio's data.

Price of the Whistle at 20 Stores



Question 6

Question 6

Continue your response to **QUESTION 6** on this page.

The summary statistics for Julio's data are shown in the following table.

Question 6

(b) Julio wants to examine some characteristics of the distribution of the sample of whistle prices.

(i) Describe the shape of the distribution of the sample of whistle prices. Justify your response using appropriate values from the summary statistics table.

Question 6

(ii) Using the $1.5 \times \text{IQR}$ rule, determine whether there are any outliers in the sample of whistle prices. Justify your response.

Question 6

GO ON TO THE NEXT PAGE.

Continue your response to **QUESTION 6** on this page.

It can often be difficult to determine whether the distribution of sample data is skewed by looking at a graph of the data and the summary statistics, particularly when the sample size is small. Thus, statisticians sometimes measure how skewed a data set is. One such measure is Pearson's coefficient of skewness, which is calculated using the following formula.

$$\text{Pearson's Coefficient of Skewness} = \frac{3(\overline{x} - m)}{s}$$

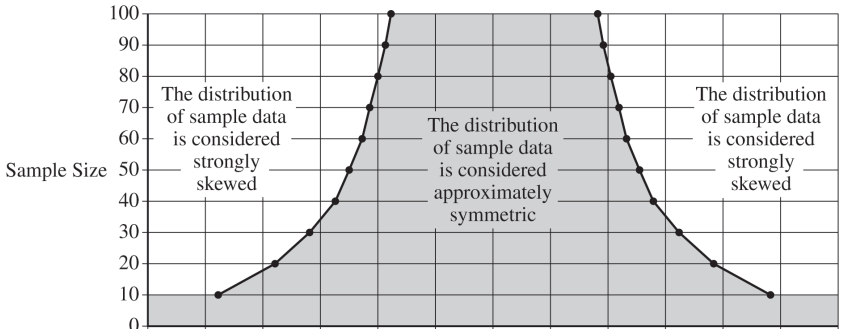
In the formula, $\overline{x}$ is the sample mean, m is the sample median, and s is the sample standard deviation.

(c) (i) Calculate Pearson's coefficient of skewness for Julio's sample of 20 whistle prices. Show your work.

Question 6

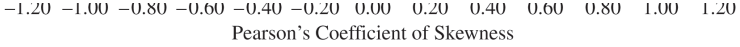
The following graph shows conclusions that can be made about the shape of the distribution of sample data based on Pearson's coefficient of skewness and sample size.

Conclusion from Pearson's Coefficient of Skewness



Question 6

Question 6



(ii) Indicate the value of the Pearson's coefficient of skewness you calculated in part (c-i) for the appropriate sample size by marking it with an "X" on the preceding graph.

Question 6

Julio's inference procedure in part (a-i) needs one of the following requirements to be satisfied to verify the normality condition.

- The sample size is greater than or equal to 30.
- If the sample size is less than 30, the distribution of the sample data is not strongly skewed and does not have outliers.

(ii) Using your response to (d-i) and the preceding requirements, is the normality condition satisfied for Julio's data? Explain your response.

Question 6

GO ON TO THE NEXT PAGE.

Question 6

STOP

END OF EXAM

Question 6

Question 6

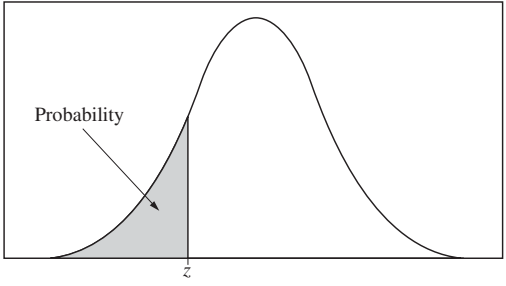


Table entry for z is the probability lying below z.

Question 6

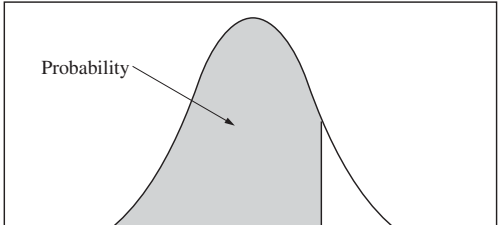
Table A Standard normal probabilities

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−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

Question 6

−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

Table entry for z is the



Question 6

probability lying below z.

Table A (*Continued*)

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

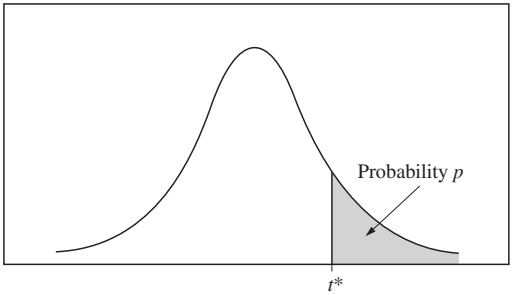
Question 6

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

Question 6

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

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



Question 6

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

Question 6

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

Question 6

Table entry for p is the point (χ^2) with probability p lying above it.

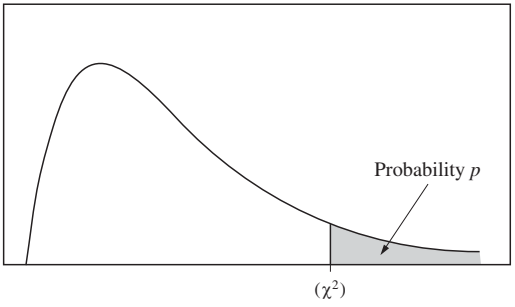


Table C χ^2 **critical values**

Question 6

	Tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.32	1.64	2.07	2.71	3.84	5.02	5.41	6.63	7.88	9.14	10.83	12.12
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82	15.20
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27	17.73
4	5.39	5.99	6.74	7.78	9.49	11.14	11.67	13.28	14.86	16.42	18.47	20.00
5	6.63	7.29	8.12	9.24	11.07	12.83	13.39	15.09	16.75	18.39	20.51	22.11
6	7.84	8.56	9.45	10.64	12.59	14.45	15.03	16.81	18.55	20.25	22.46	24.10
7	9.04	9.80	10.75	12.02	14.07	16.01	16.62	18.48	20.28	22.04	24.32	26.02
8	10.22	11.03	12.03	13.36	15.51	17.53	18.17	20.09	21.95	23.77	26.12	27.87
9	11.39	12.24	13.29	14.68	16.92	19.02	19.68	21.67	23.59	25.46	27.88	29.67
10	12.55	13.44	14.53	15.99	18.31	20.48	21.16	23.21	25.19	27.11	29.59	31.42
11	13.70	14.63	15.77	17.28	19.68	21.92	22.62	24.72	26.76	28.73	31.26	33.14
12	14.85	15.81	16.99	18.55	21.03	23.34	24.05	26.22	28.30	30.32	32.91	34.82
13	15.98	16.98	18.20	19.81	22.36	24.74	25.47	27.69	29.82	31.88	34.53	36.48
14	17.12	18.15	19.41	21.06	23.68	26.12	26.87	29.14	31.32	33.43	36.12	38.11
15	18.25	19.31	20.60	22.31	25.00	27.49	28.26	30.58	32.80	34.95	37.70	39.72
16	19.37	20.47	21.79	23.54	26.30	28.85	29.63	32.00	34.27	36.46	39.25	41.31
17	20.49	21.61	22.98	24.77	27.59	30.19	31.00	33.41	35.72	37.95	40.79	42.88
18	21.60	22.76	24.16	25.99	28.87	31.53	32.35	34.81	37.16	39.42	42.31	44.43
19	22.72	23.90	25.33	27.20	30.14	32.85	33.69	36.19	38.58	40.88	43.82	45.97
20	23.83	25.04	26.50	28.41	31.41	34.17	35.02	37.57	40.00	42.34	45.31	47.50
21	24.93	26.17	27.66	29.62	32.67	35.48	36.34	38.93	41.40	43.78	46.80	49.01
22	26.04	27.30	28.82	30.81	33.92	36.78	37.66	40.29	42.80	45.20	48.27	50.51
23	27.14	28.43	29.98	32.01	35.17	38.08	38.97	41.64	44.18	46.62	49.73	52.00
24	28.24	29.55	31.13	33.20	36.42	39.36	40.27	42.98	45.56	48.03	51.18	53.48

Question 6

25	29.34	30.68	32.28	34.38	37.65	40.65	41.57	44.31	46.93	49.44	52.62	54.95
26	30.43	31.79	33.43	35.56	38.89	41.92	42.86	45.64	48.29	50.83	54.05	56.41
27	31.53	32.91	34.57	36.74	40.11	43.19	44.14	46.96	49.64	52.22	55.48	57.86
28	32.62	34.03	35.71	37.92	41.34	44.46	45.42	48.28	50.99	53.59	56.89	59.30
29	33.71	35.14	36.85	39.09	42.56	45.72	46.69	49.59	52.34	54.97	58.30	60.73
30	34.80	36.25	37.99	40.26	43.77	46.98	47.96	50.89	53.67	56.33	59.70	62.16
40	45.62	47.27	49.24	51.81	55.76	59.34	60.44	63.69	66.77	69.70	73.40	76.09
50	56.33	58.16	60.35	63.17	67.50	71.42	72.61	76.15	79.49	82.66	86.66	89.56
60	66.98	68.97	71.34	74.40	79.08	83.30	84.58	88.38	91.95	95.34	99.61	102.7
80	88.13	90.41	93.11	96.58	101.9	106.6	108.1	112.3	116.3	120.1	124.8	128.3
100	109.1	111.7	114.7	118.5	124.3	129.6	131.1	135.8	140.2	144.3	149.4	153.2

Solution 6

(a)(i)

Mark scheme

- A one-sample t-interval for a population mean. Full credit requires identifying the procedure BY NAME as a one-sample t-interval (e.g. "one-sample t-interval"); a response that refers to a "test" instead of an interval does not satisfy this.

Solution 6

(a)(ii)

Mark scheme

- The parameter is μ = the true mean price, in dollars, of this type of whistle at all stores that sell the whistle. Full credit needs the parameter identified as a MEAN (singular "a mean" is sufficient) AND sufficient context: the population (all stores that sell the whistle) and the variable of interest (the price of this type of whistle) – at minimum, "price" and "whistle" must appear, and if the response refers to the sample mean instead of the population mean it must be clearly defined as the population mean to earn credit.

Solution 6

(b)(i)

Mark scheme

- The distribution of the sample of whistle prices appears slightly skewed to the right, because the mean (5.12) is slightly higher than the median (4.885). Full credit needs: (1) indicates the distribution is skewed right (or approximately symmetric, if justified consistently), and (2) justification using the summary statistics – e.g. the mean is higher than the median, the ratio of mean to median is greater than 1, or comparing the max-to-median gap with the median-to-min gap. A response that indicates skewed LEFT does not satisfy either component.

Solution 6

(b)(ii)

Mark scheme

- Using the $1.5 \times IQR$ rule: $IQR = Q_3 - Q_1 = 5.475 - 4.51 = 0.965$. Lower boundary $= Q_1 - 1.5(IQR) = 4.51 - 1.5(0.965) = 3.0625$; because the minimum (4.25) is greater than 3.0625, there are no outliers to the left. Upper boundary $= Q_3 + 1.5(IQR) = 5.475 + 1.5(0.965) = 6.9225$; because the maximum (6.58) is less than 6.9225, there are no outliers to the right. There are NO outliers in this sample of whistle prices. Full credit needs both boundary values computed with work shown AND the correct no-outliers conclusion stated based on the calculated lower and upper criteria.

Solution 6

(c)(i)

Mark scheme

- Pearson's coefficient of skewness $= \dfrac{3(\bar{x} - m)}{s} = \dfrac{3(5.12 - 4.885)}{0.743} \approx 0.949.$
  Full credit needs the correct value with work shown, substituting $\bar{x} = 5.12$ (mean), $m = 4.885$ (median), and $s = 0.743$ (standard deviation) into the given formula.

Solution 6

(c)(ii)

Mark scheme

- Mark an "X" on the provided graph at the point where sample size $= 20$ (Julio's n) and Pearson's coefficient of skewness ≈ 0.949 (from part (c-i)) – i.e. at approximately $(0.949, 20)$, which lies on the line for sample size 20, between about 0.90 and 1.00 on the horizontal axis, inside the right-hand shaded "strongly skewed" region near the boundary curve. Full credit needs the mark placed at (or consistent with) that coordinate.

Solution 6

(d)(i)

Mark scheme

- The distribution of the sample of whistle prices should be considered STRONGLY skewed: for a sample size of 20 and a skewness coefficient of 0.949 (from part (c)), that point falls in the "strongly skewed" region of the graph in part (c). Full credit needs the conclusion (strongly skewed) stated, using or referring to the graph or Pearson's coefficient of skewness from part (c) to make the determination. A response that indicates the distribution is strongly skewed to the LEFT does not satisfy this.

Solution 6

(d)(ii) Mark scheme

- No, the normality condition is NOT satisfied for Julio's data. Julio's sample size is only 20, which is less than 30, and (from part (d-i)) Pearson's coefficient of skewness indicates the distribution of the sample data is strongly skewed – so neither of the two allowed criteria (sample size ≥ 30, OR sample size < 30 with the distribution not strongly skewed and no outliers) is met. Full credit needs: (1) states the normality condition is not met (or a statement consistent with part (d-i)), (2) explains that the sample size (20) is less than 30, (3) explains that the distribution of the sample data is strongly skewed, consistent with part (d-i). (A response that says the sample is strongly skewed but claims the normality condition IS satisfied should be scored as not meeting this.)

Question 1

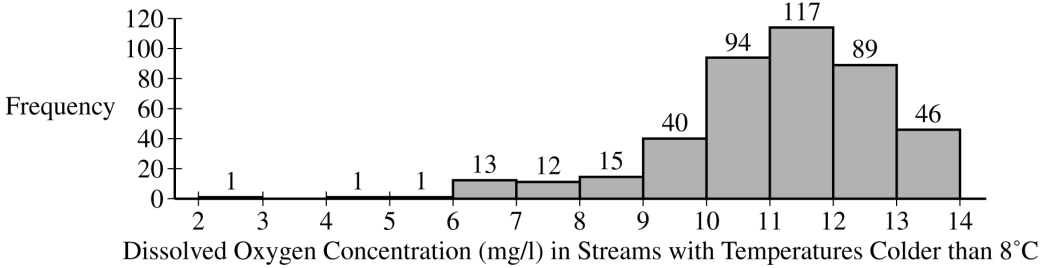
2023 ▪ Q1

As part of a study on the chemistry of Alaskan streams, researchers took water samples from many streams with temperatures colder than $8°$C and from many streams with temperatures warmer than $8°$C. For each sample, the researchers measured the dissolved oxygen concentration, in milligrams per liter (mg/l).
[Histogram titled "Dissolved Oxygen Concentration (mg/l) in Streams with Temperatures Colder than $8°$C"; x-axis "Dissolved Oxygen Concentration (mg/l)" from 2 to 14 (bin width 1); y-axis "Frequency" from 0 to 120. Bar heights (frequency) by bin: [2,3): 1, [3,4): 0, [4,5): 1, [5,6): 1, [6,7): 13, [7,8): 12, [8,9): 15, [9,10): 40, [10,11): 94, [11,12): 117, [12,13): 89, [13,14): 46.]
(a) The researchers constructed the histogram shown for the dissolved oxygen concentration in streams from the sample with water temperatures <u>colder</u>

Question 1

than 8°C. Based on the histogram, describe the distribution of dissolved oxygen concentration in streams with water temperatures <u>colder</u> than 8°C.

Question 1


Frequency
120
100
80
60
40
20
0
1
1
1
13
12
15
40
94
117
89
46
2
3
4
5
6
7
8
9
10
11
12
13
14
Dissolved Oxygen Concentration (mg/l) in Streams with Temperatures Colder than 8°C

Question 1

2023 ▪ Q1

As part of a study on the chemistry of Alaskan streams, researchers took water samples from many streams with temperatures colder than $8°C$ and from many streams with temperatures warmer than $8°C$. For each sample, the researchers measured the dissolved oxygen concentration, in milligrams per liter (mg/l).

[Histogram titled "Dissolved Oxygen Concentration (mg/l) in Streams with Temperatures Colder than $8°C$"; x-axis "Dissolved Oxygen Concentration (mg/l)" from 2 to 14 (bin width 1); y-axis "Frequency" from 0 to 120. Bar heights (frequency) by bin: [2,3): 1, [3,4): 0, [4,5): 1, [5,6): 1, [6,7): 13, [7,8): 12, [8,9): 15, [9,10): 40, [10,11): 94, [11,12): 117, [12,13): 89, [13,14): 46.]

(b) The researchers computed the summary statistics shown in the table for the dissolved oxygen concentration in streams from the sample with water temperatures <u>warmer</u> than $8°C$. Use the summary statistics to construct a box plot for

Question 1

the dissolved oxygen concentration in streams with water temperatures <u>warmer</u> than 8°C. Do not indicate outliers.

Min	Q1	Median	Q3	Max	Mean	Std. Dev.
2.10	4.39	5.43	6.12	13.45	5.54	1.64

[Blank grid provided for the box plot; x-axis "Dissolved Oxygen Concentration (mg/l) in Streams with Temperatures Warmer than 8°C" numbered 2 to 14.]

Question 1

2023 ▪ Q1

As part of a study on the chemistry of Alaskan streams, researchers took water samples from many streams with temperatures colder than $8°C$ and from many streams with temperatures warmer than $8°C$. For each sample, the researchers measured the dissolved oxygen concentration, in milligrams per liter (mg/l).

[Histogram titled "Dissolved Oxygen Concentration (mg/l) in Streams with Temperatures Colder than $8°C$"; x-axis "Dissolved Oxygen Concentration (mg/l)" from 2 to 14 (bin width 1); y-axis "Frequency" from 0 to 120. Bar heights (frequency) by bin: [2,3): 1, [3,4): 0, [4,5): 1, [5,6): 1, [6,7): 13, [7,8): 12, [8,9): 15, [9,10): 40, [10,11): 94, [11,12): 117, [12,13): 89, [13,14): 46.]

(c) The researchers believe that streams with higher dissolved oxygen concentration are generally healthier for wildlife. Which streams are generally healthier for wildlife, those with water temperature <u>colder</u> than $8°C$ or those with water

Question 1

temperature <u>warmer</u> than 8°C? Using characteristics of the distribution of dissolved oxygen concentration for temperatures <u>colder</u> than 8°C and characteristics of the distribution of dissolved oxygen concentration for temperatures <u>warmer</u> than 8°C, justify your answer.

Solution 1

(a) Mark scheme

- The distribution of dissolved oxygen concentration (mg/l) in Alaskan streams colder than 8°C is unimodal and skewed left, with median between 11 and 12 mg/l. Q1 falls in the 10-11 mg/l bin and Q3 in the 12-13 mg/l bin, so IQR $\approx$ 2 mg/l (acceptable IQR range 1.5-2.5; acceptable range-of-data 10-12 mg/l; acceptable SD 1.7-1.8 mg/l). There are no apparent high outliers, but the 2-3, 4-5, and 5-6 mg/l bins are all more than 1.5$\times$IQR below Q1, so there are several potential low outliers (a comment on a gap between 3 and 4 mg/l is also acceptable for this component). Full credit requires component 1 (context: dissolved oxygen concentration/amount of oxygen/mg per liter) AND at least 3 of: (2) shape - skewed left (describing it as normal/approximately normal fails this

Solution 1

component), (3) center - a numeric value for median/mean between 11 and 12 mg/l (definitive language like 'the median is 11.5' does NOT satisfy this component; it must be phrased approximately, e.g. 'approximately 11.5'), (4) spread - any correctly used measure of variation (IQR, range, or SD) within the bounds above, (5) unusual features - potential outliers or a gap (stating a *different* unusual feature, or definitive language like 'there is an outlier', does not satisfy this component). Partial credit: component 1 + only 2 of components 2-5, OR at least 3 of components 2-5 without component 1.

Solution 1

(b) Mark scheme

- Construct a boxplot (not any other graph type) for dissolved oxygen concentration in streams warmer than 8°C using: Min = 2.10, Q1 = 4.39, Median = 5.43, Q3 = 6.12, Max ≈ 13 mg/l (do not flag outliers - draw the whiskers all the way to the min and max). The box spans from Q1 to Q3 with the median line inside; whiskers extend from the box ends to the minimum and maximum values. Full credit (4 or 5 of 5 components): (1) box begins at Q1 (between the 4 and 4.5 tick marks), (2) box ends at Q3 (between 6 and 6.5), (3) median located within the box (between 5 and 5.5), (4) line extended to the minimum value (between 2 and 2.5), (5) line extended to the maximum value (between 13 and 13.5). Partial

Solution 1

credit: only 3 of the 5 components. A response that draws a graph other than a boxplot (e.g. a histogram) should be scored incorrect regardless of accuracy.

Solution 1

(c) Mark scheme

- Streams with water temperature colder than 8°C are generally healthier for wildlife (assuming the researchers' belief that higher dissolved oxygen means healthier streams). The colder-stream distribution has a higher center - its median (between 11 and 12 mg/l) is larger than the warmer streams' median (5.43 mg/l). The two distributions differ in shape: colder streams are skewed left while warmer streams are skewed right. Both distributions have similar spread - their IQRs are comparable ($\approx$2 mg/l for colder streams vs. $\approx$1.73 mg/l for warmer streams). Full credit requires component 1 (states colder streams are healthier for wildlife) AND at least 2 of: (2) directly compares the centers of the two distributions, (3) indicates warmer streams are skewed right and colder streams are skewed left, (4)

Solution 1

directly compares the spreads (comparing at least one of range, IQR, or SD between the two distributions). Partial credit: component 1 + only 1 of components 2-4. A response based on an incorrect part (a) or (b) may still earn credit here if it is consistent with those earlier answers. Comparing outliers between the two distributions is not necessary for this part.

Question 1

2021 ▪ Q1

The length of stay in a hospital after receiving a particular treatment is of interest to the patient, the hospital, and insurance providers. Of particular interest are unusually short or long lengths of stay. A random sample of 50 patients who received the treatment was selected, and the length of stay, in number of days, was recorded for each patient. The results are summarized in the following table and are shown in the dotplot.

Length of stay (days)	**5**	**6**	**7**	**8**	**9**	**12**	**21**
Number of patients	4	13	14	11	6	1	1

Question 1

[Dotplot titled “Length of Stay (days)”; x-axis labeled Length of Stay (days), tick marks at 5, 7, 9, 11, 13, 15, 17, 19, 21. Stacked dots above each value show the counts from the table: 4 dots at 5, 13 dots at 6, 14 dots at 7, 11 dots at 8, 6 dots at 9, 1 dot at 12, 1 dot at 21 (no dots shown at 6, 10, 11, or 13-20 other than the isolated points at 12 and 21).]

(a) Determine the five-number summary of the distribution of length of stay.

(b)(i) Consider two rules for identifying outliers, method A and method B. Let method A represent the $1.5 \times$ IQR rule, and let method B represent the 2 standard deviations rule.

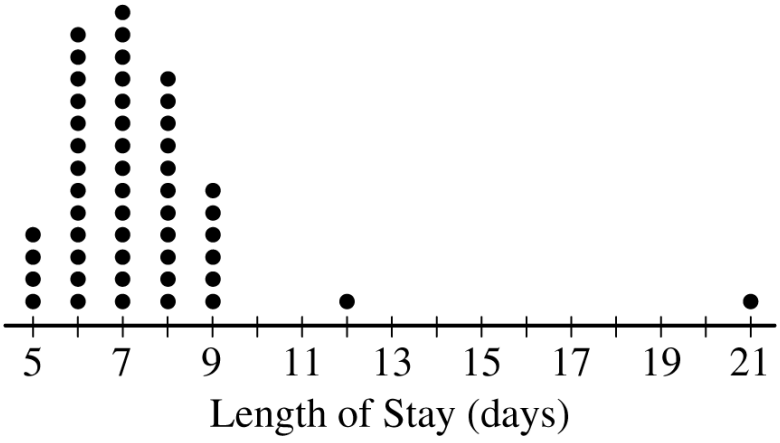
Using method A, determine any data points that are potential outliers in the distribution of length of stay. Justify your answer.

Question 1

(b)(ii) The mean length of stay for the sample is 7.42 days with a standard deviation of 2.37 days. Using method B, determine any data points that are potential outliers in the distribution of length of stay. Justify your answer.

(c) Explain why method A might identify more data points as potential outliers than method B for a distribution that is strongly skewed to the right.

Question 1



Solution 1

(a)

Mark scheme

- Five-number summary of length of stay (days): Minimum $= 5$, Lower quartile $Q_1 = 6$, Median $= 7$, Upper quartile $Q_3 = 8$, Maximum $= 21$. Full credit requires all five values, each correctly labeled (min, Q_1, median, Q_3, max); 3-4 correct labeled values earns partial credit.

Solution 1

(b)(i)

Mark scheme

- Method A ($1.5\times IQR rule$) :IQR $= Q_3 - Q_1 = 8 - 6 = 2$. Lower fence $= Q_1 - 1.5(IQR) = 6 - 1.5(2) = 3$; Upper fence $= Q_3 + 1.5(IQR) = 8 + 1.5(2) = 11$. Any value below 3 or above 11 is a potential outlier. No values are below 3; the patients who stayed 12 days and 21 days are above 11, so those two data points (12 and 21 days) are the potential outliers. Credit requires both fence values shown/calculated AND the two correct outliers identified.

Solution 1

(b)(ii)

Mark scheme

- Method B (2 standard deviations rule): mean = 7.42 days, SD = 2.37 days. Mean $\pm 2 \times SD = 7.42 \pm 2(2.37) = 7.42 \pm 4.74$, giving the interval $(2.68, 12.16)$. Any value outside this interval is a potential outlier. Only the patient who stayed 21 days falls outside ($21 > 12.16$), so 21 days is the only outlier under method B. Credit requires the interval calculation shown AND the one correct outlier (21 days) identified.

Solution 1

(c) Mark scheme

- For a distribution strongly skewed to the right, the mean is pulled toward the extreme values in the long right tail more than the median/quartiles are, and the ratio of SD to IQR tends to be larger than for a symmetric distribution. This means method B's outlier boundary (mean + $2\times SD$) $shifts further into the right tail than method A's boundary$ ($\mathrm{Q}_3 + 1.5 \times IQR$), so method B's threshold for calling a point an outlier becomes looser (higher). This decreases method B's ability to detect outliers relative to method A, so method A may flag more potential outliers than method B for a strongly right-skewed distribution. Credit requires (1) stating the mean/SD are pulled more toward the

Solution 1

tail than the median/quartiles (or IQR), AND (2) linking that effect to method A detecting relatively more outliers than method B.

Question 1

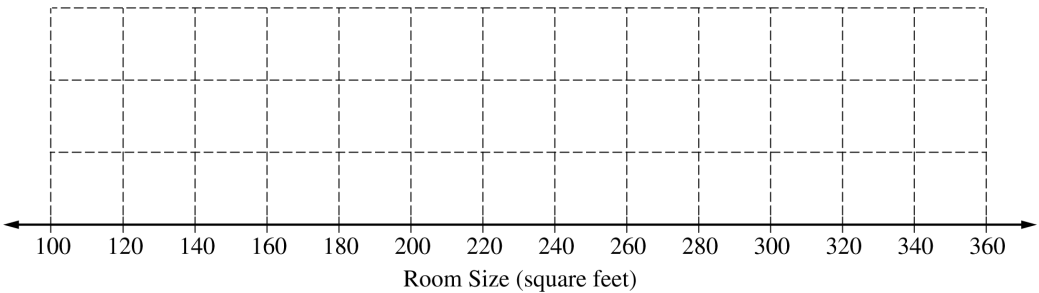
2019 ▪ Q1

The sizes, in square feet, of the 20 rooms in a student residence hall at a certain university are summarized in the following histogram.

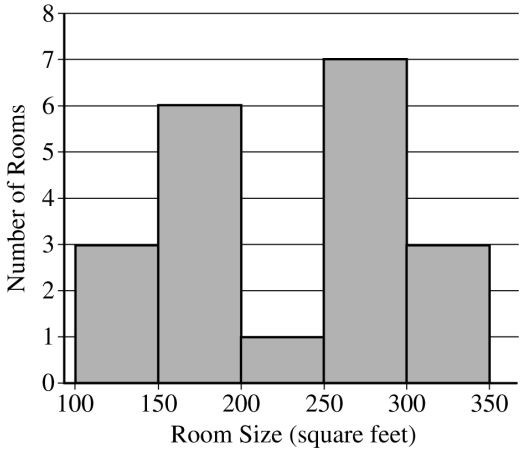
[Histogram titled "Room Size (square feet)"; y-axis "Number of Rooms" from 0 to 8; x-axis "Room Size (square feet)" with tick marks at 100, 150, 200, 250, 300, 350. Bars: 100-150 has height 3; 150-200 has height 6; 200-250 has height 1; 250-300 has height 7; 300-350 has height 3.]

(a) Based on the histogram, write a few sentences describing the distribution of room size in the residence hall.

Question 1


100 120 140 160 180 200 220 240 260 280 300 320 340 360
Room Size (square feet)

Question 1


Number of Rooms
0
1
2
3
4
5
6
7
8
100
150
200
250
300
350
Room Size (square feet)

Question 1

Mean	Standard Deviation	Min	Q1	Median	Q3	Max
231.4	68.12	134	174	253.5	292	315

Question 1

2019 ▪ Q1

The sizes, in square feet, of the 20 rooms in a student residence hall at a certain university are summarized in the following histogram.
[Histogram titled "Room Size (square feet)"; y-axis "Number of Rooms" from 0 to 8; x-axis "Room Size (square feet)" with tick marks at 100, 150, 200, 250, 300, 350. Bars: 100-150 has height 3; 150-200 has height 6; 200-250 has height 1; 250-300 has height 7; 300-350 has height 3.]

(b) Summary statistics for the sizes are given in the following table.

Mean	Standard Deviation	Min	Q1	Median	Q3	Max
231.4	68.12	134	174	253.5	292	315

Question 1

Determine whether there are potential outliers in the data. Then use the following grid to sketch a boxplot of room size.

[Number line grid titled "Room Size (square feet)" with tick marks from 100 to 360 in increments of 20 (100, 120, 140, 160, 180, 200, 220, 240, 260, 280, 300, 320, 340, 360), for sketching a boxplot.]

Question 1

2019 ▪ Q1

The sizes, in square feet, of the 20 rooms in a student residence hall at a certain university are summarized in the following histogram.

[Histogram titled "Room Size (square feet)"; y-axis "Number of Rooms" from 0 to 8; x-axis "Room Size (square feet)" with tick marks at 100, 150, 200, 250, 300, 350. Bars: 100-150 has height 3; 150-200 has height 6; 200-250 has height 1; 250-300 has height 7; 300-350 has height 3.]

(c) What characteristic of the shape of the distribution of room size is apparent from the histogram but not from the boxplot?

Solution 1

(a) Mark scheme

- The distribution of room sizes is bimodal and roughly symmetric, with two clusters: about 100-200 square feet and about 250-350 square feet. The center of the distribution is between 200 and 300 square feet (e.g., median ≈ 253.5, mean ≈ 231.4). The spread can be stated as: the range is a value between 150 and 250 square feet, OR the IQR is a value between 50 and 150 square feet, OR all room sizes are between 100 and 350 square feet. The response must include context (room size in square feet). Full credit requires all four components: (1) bimodal shape / two peaks / two clusters (NOT unimodal or normal), (2) center between 200 and 300 sq ft, (3) a valid spread statement, (4) context.

Solution 1

(b) Mark scheme

- Outlier check: IQR = Q3 - Q1 = 292 - 174 = 118 square feet. Fences: $Q_1 - 1.5(IQR) = 174 - 1.5(118) = -3$ and $Q_3 + 1.5(IQR) = 292 + 1.5(118) = 469$. There are no potential outliers because the minimum (134 sq ft) does not fall below -3, and the maximum (315 sq ft) does not exceed 469. Boxplot: sketch a standard box-and-whisker (no mean shown) using the five-number summary min=134, Q1=174, median=253.5, Q3=292, max=315 – on the grid the minimum should fall between 120-140, Q1 between 160-180, the median between 240-260, Q3 between 280-300, and the maximum between 300-320 (square feet).

Solution 1

(c)

Mark scheme

- The histogram clearly shows the bimodal nature of the distribution of room sizes (two distinct peaks/clusters), but this bimodal shape is not apparent from the boxplot – a boxplot only displays the five-number summary and cannot reveal multiple modes. (A response based only on symmetry/skewness, rather than the bimodal/unimodal shape, does not earn credit here.)

Question 1

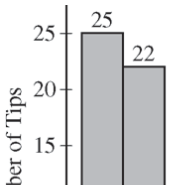
2016 ▪ Q1

STATISTICS
SECTION II
Part A
Questions 1-5
Spend about 65 minutes on this part of the exam.
Percent of Section II score—75

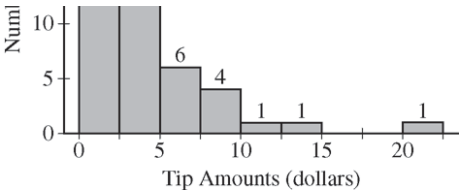
Directions: Show all your work. Indicate clearly the methods you use, because you will be scored on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.

Question 1

1. Robin works as a server in a small restaurant, where she can earn a tip (extra money) from each customer she serves. The histogram below shows the distribution of her 60 tip amounts for one day of work.



Question 1



(a) Write a few sentences to describe the distribution of tip amounts for the day shown.

(b) One of the tip amounts was $8. If the $8 tip had been $18, what effect would the increase have had on the following statistics? Justify your answers.

The mean:

The median:

Solution 1

(a) Mark scheme

- The distribution of Robin's tip amounts is skewed to the right, with a gap between the largest tip (in the $20-22.50$ interval) and the next-largest tip (in the $12.50-15$ interval) – the largest tip appears to be an outlier. The median tip is between $2.50 and 5.00$ (the mean is between $2.62 and 5.13$). Tip amounts vary from a minimum between $0 and 2.50$ up to a maximum between $20.00 and 22.50$; about 78% of tips are between $0 and 5$. Full credit (E) requires reasonable comments on all five components – shape, outlier/gap, center, variability, and context (tip amounts in dollars); partial credit (P) for 3-4 of the five; incorrect (I) for at most 2.

Solution 1

(b)

Mark scheme

- Mean: if the \$8 tip had been \$18, the mean would increase by \$10 divided by 60 tips, i.e. by \$\tfrac{1}{6} \approx \$0.17\$, since the total sum of tips increases by \$10 while the number of tips (n = 60) stays the same. Median: the median would NOT change, because the current median lies between \$2.50 and \$5.00, and both \$8 and \$18 are above that value, so the tip's position among the ordered values doesn't shift the middle observation(s). Full credit requires all four components: states mean increases, correctly justifies why, states median unchanged, correctly justifies why.