Statistics I — Class Test 1 — September 19, 2024

Name: **Roll number:**

Write your answers on the question paper. You may use separate sheets for calculations, but submit only the filled in question paper. Write your name and roll number on both pages.

The following table gives binned frequency counts of height data for a subset of the NHANES data.

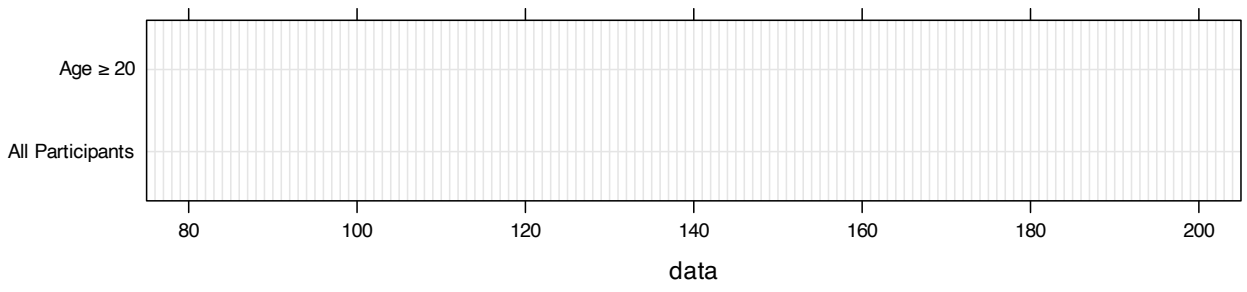
left-endpoint	right-endpoint	midpoint	All Participants	Age $\geq$ 20
75	85	80	6	0
85	95	90	61	0
95	105	100	78	0
105	115	110	119	0
115	125	120	86	0
125	135	130	94	1
135	145	140	152	41
145	155	150	744	577
155	165	160	2123	1810
165	175	170	1219	1074
175	185	180	165	155
185	195	190	0	0
195	205	200	0	0

Suppose we pretend that all data points in a bin are equal to the midpoint of that bin. Then,

1. Compute the five-number summary for both datasets and enter them in the following table. [4]

Dataset	Min	Q1	Q2	Q3	Max
All Participants					
Age $\geq$ 20					

2. Use these values to draw approximate box and whisker plots for the above datasets. [2]



Name: **Roll number:**

Let $X_1, X_2, \ldots, X_n$ be a random sample from some probability distribution P. Assume that the underlying distribution is *continuous*. Although we have not yet formally defined continuous distributions, the only relevant fact that we need to know for this example is that:

- If a is the population median of P, then $P(X_1 < a) = P(X_1 > a) = \frac{1}{2}$ (and as a consequence, $P(X_1 = a) = 0$).
- Similarly, if b is the pth quantile, then $P(X_1 < b) = p$ and $P(X_1 > b) = 1 - p$, for any $0 < p < 1$.

For any $m \in \mathbb{R}$, let $Z(m)$ be the number of observations that are strictly less than m. That is,

$$Z(m) = \sum_{i=1}^{n} \mathbf{1}\{X_i < m\}$$

where $\mathbf{1}(\cdot)$ is the indicator function.

3. If m is the median of P, what is the distribution of $Z(m)$? [2]

4. If m is the 0.8th quantile of P, what is the distribution of $Z(m)$? [2]

5. [Bonus question] Suppose in a given dataset with $n = 10$, we observe that $Z(3) = 4, Z(6) = 5$, and $Z(9) = 8$. Using this information, we want to decide whether 3, 6, and 9 are plausible values of the unknown population median. Can you think of a strategy? Briefly outline your proposal, without going into details.

Good luck!