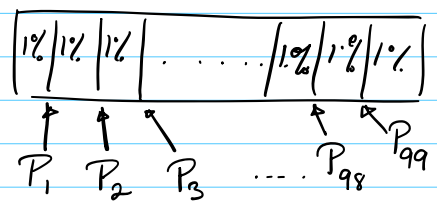
STAT-1301 ; Jan. 23, '24
Lecture 5

Percentiles and Rank Percentiles:



P_k is the k-th Percentile : $k \times 100\%$ of the Sorted data is less than P_k.

P_k = value of the $\frac{k \times n}{100}$ th term in a ranked data set.

Here, k = number of the percentile;

n = Sample Size.

Ex. Find the 42nd Percentile of the following data. Note: the data has already been Sorted in increasing order.

P_{42}

11,669 13,435 14,413 18,103 18,215 **21,088**

26,343 29,920 33,956 40,197 42,082

40,769

1. Sort data in increasing order. (Already done).

2. $L = \frac{k \times n}{100} = \frac{42 \times 12}{100} = 5.04$ is the location of the 42^{nd} percentile (P_{42}).

Text algo.: Since $\frac{k \times n}{100} = 5.04$ is not an integer, round up to the next nearest integer.

i.e. $5.04 \rightarrow 6$

i.e. The 42^{nd} percentile (P_{42}) is the 6^{th} obs'n in the Sorted data.

$$P_{42} = 21,088.$$

Interpretation: 42% of the data is less than 21,088.

Remark: P_{25} = 1^{st} Quartile = Q_1

$P_{50} \Leftrightarrow$ median $\Leftrightarrow Q_2$

$P_{75} \Leftrightarrow Q_3 \Leftrightarrow 3^{rd}$ Quartile.

Percentile Rank of a Value:

$\Leftrightarrow$ $\dfrac{\#\text{ of obs'ns} < x}{n} \times 100\%$ where

n = Sample Size.

Ex. Refer to the previous example data. Find the percentile rank of 29,920?

Sol'n:

1. Sort data in increasing order. Done!

2. 7 obs'ns are less than 29,920.

$\therefore \dfrac{7}{12} \times 100\% = 58.33\%$ is the percentile rank of 29,920.

MCQ Interpretation: Approx. 58% of the data is less than 29,920.

§ 3.6 Box-and-Whisker Plot

- used to understand the Shape of the

distribution (ie. symmetric vs. skewed, etc.)

- Can visualize Outliers (extreme observations).

Problem 3.100: Golf scores of 17 men and 15 women. We will construct boxplots of their scores.

Men: 87 68 92 79 83 67 71 92
112 75 77 102 79 78 85 75 72

1. Sort data in increasing order. Find Q_1, Q_2, Q_3 and IQR.

$Q_1 = 73.5$, $Q_2 = 79$, $Q_3 = 89.5$ and

$IQR = Q_3 - Q_1 = 16$.

2. lower inner fence (LIF): $= Q_1 - 1.5 \times IQR$

$= 73.5 - 1.5 \times 16$

$= 49.5$

Upper inner fence (UIF) := $Q_3 + 1.5 \times IQR$

$= 89.5 + 1.5 \times 16$

$= 113.5$

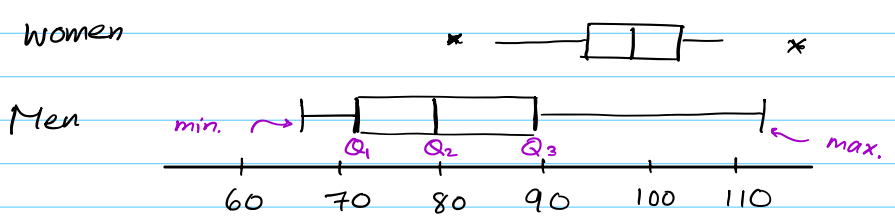
3. Consider the interval (LIF, UIF)

$= (49.5, 113.5)$ from Step 2.

No observations outside of this interval.

No outliers for the men's scores.

4. Draw a box based on Q_1, Q_2, Q_3.



Here are the women's golf scores.

101 100 87 95 98 81 117 107 103

97 90 100 99 94 94.

1. Sort data in increasing order and find Q_1, Q_2, Q_3.

$Q_1 = 94, \quad Q_2 = 98, \quad Q_3 = 101; \quad IQR = 7$

2. $LIF = Q_1 - 1.5 \times IQR = 83.5$

$UIF = Q_3 + 1.5 \times IQR = 111.5$

Since $81 < LIF$ and $117 > UIF$, 81 and 117 are outliers.

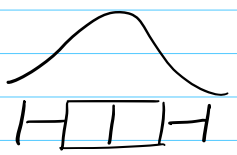
3. How far do the whiskers go?

$(LIF, UIF) = (83.5, 111.5)$.

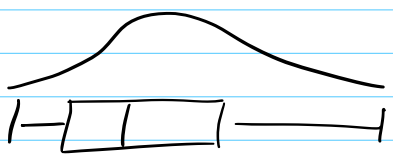
i) Extend from Q_1 to 87, the smallest number in $(83.5, 111.5)$.

ii) Extend from Q_3 to 107 , the largest obs'n in (83.5, 111.5).

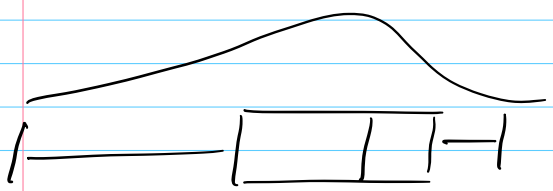
Boxplot Shapes:



=> Symmetric



=> Skewed to the right



=> Skewed to the left

Rule of thumb for outliers:

Let x be an observation.

If $x < LIF$ or $x > UIF$, then x is said to be a (mild) outlier.

Here, $LIF = Q_1 - 1.5 \times IQR$

$$UIF = Q_3 + 1.5 \times IQR.$$

(Know this rule of thumb).

Boxplot construction is not examined.

Back to golf data ....

Data Analysis:

1. On average men are stronger players than the women. (Compared medians).

2. The men's scores have a larger Spread than the women's scores.

3. There are two outliers in the women's scores.

etc.

Ch. 4 Probability

Introduced motivation for studying Probability here. - Aspirin study.