Nate Warner STAT 300 Summer 2024

PSET 1 - Due: Sunday, June 23

1. A small survey was conducted in which each respondent was asked how many times, in the previous two-week period, they had eaten at a fast food restaurant. The data appear below

0, 2, 1, 5, 2, 2, 3, 4, 1, 2, 7, 1, 3, 4, 1, 0, 1, 4, 2, 1, 3, 3, 2, 1, 9, 1

(a) Construct a frequency histogram. The histogram should be neat, accurate, and well labeled.

(b) How would you describe the shape the distribution?

(c) Find the proportion of the respondents described by each of the following.

(i) Ate at a fast food restaurant at least four times

(ii) Ate at a fast food restaurant fewer than two times

Given the above data set, we can first find the frequencies of each data point. We can then construct a frequency distribution to get a better look at the data

Number of days	Frequency
0	2
1	8
2	6
3	4
4	3
5	1
6	0
7	1
8	0
9	1

Table 1: *Number of times $n = 26$ people ate a fast food restaurant in the previous two weeks organized as a frequency distribution*

From this, we create the histogram

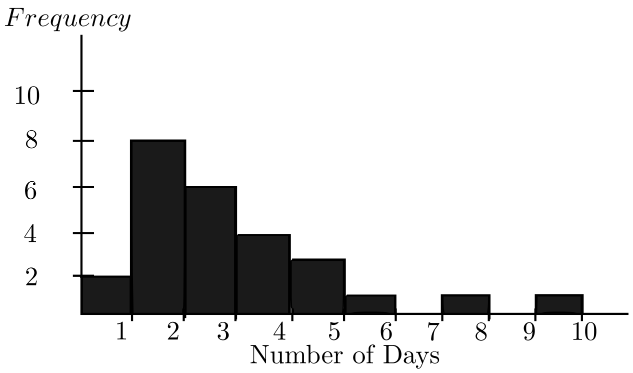

Figure 1: *Number of times $n = 26$ people ate a fast food restaurant in the previous two weeks organized as a histogram for discrete data*

b.) From the histogram, we can see that the shape of the distribution is skewed positively (right)

c.) Furthermore, we can find the relative frequency (proportion) of respondents who ate at a fast food restaurant at least four times by dividing the summation of the frequencies $x \in \{4, 5, 6, ..., 9\}$ by $n = 26$ (the total number of respondents). We find

$$\frac{3+1+0+1+0+1}{26} = 0.2308 \approx 23\%.$$

We preform a similar calculation to find the number of respondents who ate at a fast food restaurant fewer than two times, in this case we get

$$\frac{2+8}{26} = 0.3846 \approx 38\%.$$

2. The blood glucose levels (in milligrams per deciliter) for 25 patients at a medical facility appear below.

63 65 66 68 69 71 73 74 75 75 76 76 77
79 79 81 81 81 83 84 86 87 90 91 95

(a) Construct a relative frequency histogram. Use nine class intervals starting with $60 \leqslant \text{Glucose} < 64$.

(b) How would you describe the shape of the distribution?

(c) What proportion of the patients had a blood glucose level described by each of the following?

(i) At most 71

(ii) Between 72 and 91 (inclusive of the endpoints)

We again begin by organizing the data via a frequency distribution

Blood glucose levels (mg/dL)	Frequecy	Relative frequency
60-64	1	0.04
64-68	2	0.08
68-72	3	0.12
72-76	4	0.16
76-80	5	0.2
80-84	4	0.16
84-88	3	0.12
88-92	2	0.08
92-96	1	0.04

Table 2: *Blood glucose levels for $n = 25$ patiens at a medical facility organized as a frequency distribution*

Using the relative frequencies found above, we construct a simple continuous data, equal class width histogram

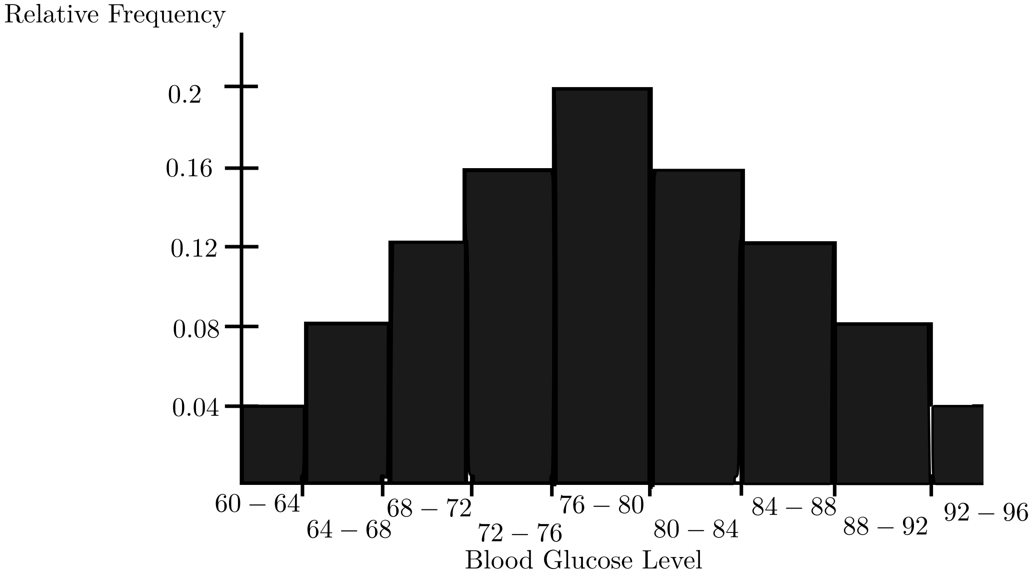

Figure 2: *Blood glucose levels for $n = 25$ patiens at a medical facility organized as a continous data histogram with equal class widths*

Note:-

If a data point falls on the upper boundary of a class, it is instead placed in the class immediately to the right. For example 68 would be placed in the class 68-72 as opposed to 64-68

b.) The shape of this distribution is approximately normal (symmetric)

c.) We find the proportion of patients with a glucose level of at most 71 by summing the relative frequencies from all classes in the range 60-71. This gives

$$0.04 + 0.08 + 0.12 = 0.24 \approx 24\%.$$

We do a similar calculator to find the proportion of patients with a glucose level from 72 and 91 inclusive

$$0.16 + 0.2 + 0.16 + 0.12 + 0.08 = 0.72 \approx 72\%.$$

Note:-

The class 68-72 is not included, any patient with a blood glucose level of 72 would be instead placed in the class 72-76

3. Refer to Problem 1.

   (a) Calculate the sample mean.

   (b) Calculate the sample median.

   (c) How do the answers to (a) and (b) compare?

      (i) Why should we have been able to anticipate this?

a.) The sample mean is given by

$$\begin{aligned}\bar{x} &= \frac{1}{n}\sum x_i \\ &= \frac{1}{25}(63+65+66+68+69+71+73+74 \\ &\qquad +75+75+76+76+77+79+79+81 \\ &\qquad +81+81+83+84+86+87+90+91+95) \\ &= 77.84.\end{aligned}$$

Remark. b.) The sample median is given by

$$\tilde{x} = \begin{cases} \left(\frac{n+1}{2}\right)^{th} \text{ ordered element} & \text{if } n = 2k+1,\ k \in \mathbb{Z} \\ \left(ave\left(\frac{n}{2}, \frac{n+2}{2}\right)\right)^{th} \text{ ordered element} & \text{if } n = 2k\, k \in \mathbb{Z} \end{cases}.$$

In our case, since n is odd we use case one and the median is given by

$$\begin{aligned}\tilde{x} &= \left(\frac{26}{2}\right) = 13^{th} \text{ ordered element} \\ &= 77.\end{aligned}$$

c.) In this case, the sample mean $\bar{x}$ and the sample median $\tilde{x}$ are (approximately) the same. This can be anticipated when the distribution of data is symmetric (normal).

4. A new type of outdoor paint was tested by painting six homes in the same geographic area. The number of months the paint lasted before fading was recorded and yielded the values below.

$$10, 60, 50, 30, 40, 20$$

(a) Calculate the sample range.

(b) Calculate the sample mean.

(c) Calculate the sample variance using the definition formula (i.e. by using the squared deviations).

(d) Calculate the sample variance using the short-cut formula.

(e) Calculate the sample standard deviation.

(f) Give the units of measurement for (b); for (c); and for (e).

a.) The sample range is computed by finding the difference between the largest and smallest data point. Thus,

$$\text{Range} = 60 - 10 = 50.$$

b.) The sample mean is given by

$$\begin{aligned}\bar{x} &= \frac{1}{n}\sum x_i = \frac{1}{6}(10 + 60 + 50 + 30 + 40 + 20) \\ &= 35.\end{aligned}$$

c.) The sample variance is given by

$$\begin{aligned}s^2 &= \frac{1}{n-1}\sum(x_i - \bar{x})^2 \\ &= \frac{1}{5}\left((10-35)^2 + (60-35)^2 + (50-35)^2 + (30-35)^2 + (40-35)^2 + (20-35)^2\right) \\ &= 350.\end{aligned}$$

d.) The alternative variance formula is derived by simplifying the numerator to

$$\begin{aligned}S_{xx} &= \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} \\ &= \frac{n\sum x_i^2 - \left(\sum x_i\right)^2}{n}.\end{aligned}$$

This allows us to express variance as

$$\begin{aligned}s^2 &= \frac{1}{n-1}S_{xx} \\ &= \frac{n\sum x_i^2 - \left(\sum x_i\right)^2}{n^2 - n}.\end{aligned}$$

Using this, we again find

$$\begin{aligned}s^2 &= \frac{6(10^2 + 60^2 + 50^2 + 40^2 + 30^2 + 20^2) - (10 + 60 + 50 + 30 + 40 + 20)^2}{6^2 - 6} \\ &= 350.\end{aligned}$$

e.) Using the variance found above, we can then compute the std dev

$$s = \sqrt{s^2} = \sqrt{350} = 18.7083.$$

f.) The units are months.

5. Refer to Problem 4.

   (a) Add 10 to each data value in Problem 4.

      (i) Re-calculate the sample mean.

      (ii) How does the answer in (i) compare to that in 4(b)? Be specific.

      (iii) Re-calculate the sample variance. Note – using the definition formula might be more illuminating.

      (iv) How does the answer in (iii) compare to that in 4(c)? Be specific.

   (b) Divide each data value in Problem 4 by 10 (i.e. multiply each data value by 1/10).

      (i) Re-calculate the sample mean.

      (ii) How does the answer in (i) compare to that in 4(b)? Be specific.

      (iii) Re-calculate the sample variance. Note – using the definition formula might be more illuminating.

      (iv) How does the answer in (iii) compare to that in 4(c)? Be specific.

a.i) Adding 10 to each data point gives the set

20 70 60 40 50 30

Using this new set, we calculate the sample mean

$$\begin{aligned}\bar{x} &= \frac{1}{6}(20+70+60+40+50+30)\\ &= 45.\end{aligned}$$

a.ii) As expected, the new mean is exactly 10 more than that found in 4.b. The reason for this can be easily shown

$$\begin{aligned}\bar{x} &= \frac{1}{n}\sum_{i=1}^{n}(x_i+10)\\ &= \frac{1}{n}\sum_{i=1}^{n}x_i + \frac{1}{n}\sum_{i=1}^{n}10\\ &= \frac{1}{n}\sum_{i=1}^{n}x_i + \frac{1}{n}10n\\ &= 10+\frac{1}{n}\sum_{i=1}^{n}x_i.\end{aligned}$$

Thus, finding the mean after adding 10 to each data point is exactly the same as adding 10 to the original mean.

a.iii) Using the new data set and the new sample mean, we can compute the new sample variance.

$$\begin{aligned}s_y^2 &= \frac{1}{n-1}\sum_{i=1}^{n} (y_i - \bar{y})^2\\ &= \frac{1}{5}((20-45)^2 + (70-45)^2 + (60-45)^2\\ &+ (40-45)^2 + (50-45)^2 + (30-45)^2)\\ &= 350.\end{aligned}$$

a.iv) Here we see that the variance remains unchanged. The reason for this can be shown mathematically

$$\begin{aligned}s^2 &= \frac{1}{n-1}\sum_{i=1}^{n} ((x_i+10) - (\bar{x}+10))^2\\ &= \frac{1}{n-1}\sum_{i=1}^{n} (x_i + 10 - \bar{x} - 10)^2\\ &= \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2 \ .\end{aligned}$$

Thus, we notice adding a constant to each value in a data set produces no change in the variance.

b) Multiply each value in the original data set by $\frac{1}{10}$ gives the new set

$$1\ 6\ 5\ 3\ 4\ 2$$

b.i) With this, we can calculate the new sample mean

$$\begin{aligned}\bar{x} &= \frac{1}{6}(1+6+5+3+4+2)\\ &= \frac{7}{2} = 3.5.\end{aligned}$$

b.ii) As we see here, this new sample mean is exacty the original sample mean divided by 10. This can again be shown using simple properties of summation. In specific we know $\sum cx_i = c\sum x_i$

$$\begin{aligned}\bar{x} &= \frac{1}{n}\sum_{i=1}^{n} \left(x_i \cdot \frac{1}{10}\right)\\ &= \frac{1}{10n}\sum_{i=1}^{n} x_i\\ &= \frac{\frac{1}{n}\sum\limits_{i=1}^{n} x_i}{10}.\end{aligned}$$

b.ii) Thus, we see that multiply or dividing a data set by a constant factor also multiplys or divides the mean by that same constant factor.

b.iii) Computing the new sample variance gives

$$\begin{aligned}s^2 &= \frac{1}{5}((1-3.5)^2 + (6-3.5)^2\\ &+ (5-3.5)^2 + (3-3.5)^2 + (4-3.5)^2 + (2-3.5)^2)\\ &= 3.5\ .\end{aligned}$$

b.iv) We see that the new sample variance does actually change this time, by a factor of $\frac{1}{10^2}$. Again, we can show why this is the case

$$
\begin{aligned}
s_2^2 &= \frac{1}{n-1}\sum_{i=1}^{n} \left(\frac{x_i}{10} - \frac{\bar{x}}{10}\right)^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n} \left(\frac{1}{10}(x_i - \bar{x})\right)^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n} \left(\frac{1}{10}\right)^2 (x_i - \bar{x})^2 \\
&= \frac{1}{10^2}\left(\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right).
\end{aligned}
$$

6. Suppose that $x_1, x_2, ..., x_n$ is a set of data with $\bar{x} = 40$, $\tilde{x} = 60$, $s_x^2 = 25$ and $s_x = 5$ Suppose that each x_i is transformed to a new data value $y_i = 2x_i - 1$

(a) Compare $\bar{x}$ to $\tilde{x}$. What do the values potentially tell us about the shape of the distribution?

(b) Find each of the following

(i) $\bar{y}$

(ii) $\tilde{y}$

(iii) s_y^2

(iv) s_y

6.a) Since the mean and median are far apart, specifically the median being larger than the mean, we can expect the shape of the distribution to be negatively skewed

Remark. Assume that Y is a linear transformation of X. Then,

$$Y = bX + A$$

- Mean of Y:
$$\text{Mean of } Y = b(\text{Mean of } X) + A$$
- Median of Y:
$$\text{Median of } Y = b(\text{Median of } X) + A$$
- Standard Deviation of Y:
$$\text{sd of } Y = b(\text{sd of } X)$$
- Variance of Y:
$$\text{Variance of } Y = b^2(\text{Variance of } X)$$

b.i) The new mean can be calculated by

$$2(40) - 1 = 79.$$

b.ii) Median:

$$2(60) - 1 = 119.$$

b.iii) Variance:

$$4(25) = 100.$$

b.iv) Standard deviation:

$$2(5) = 10.$$