Assignment 11, Introduction to Statistics

Oleg Sivokon

<2015-03-27 Fri>

Contents

1 Problems

1.1 Problem 1

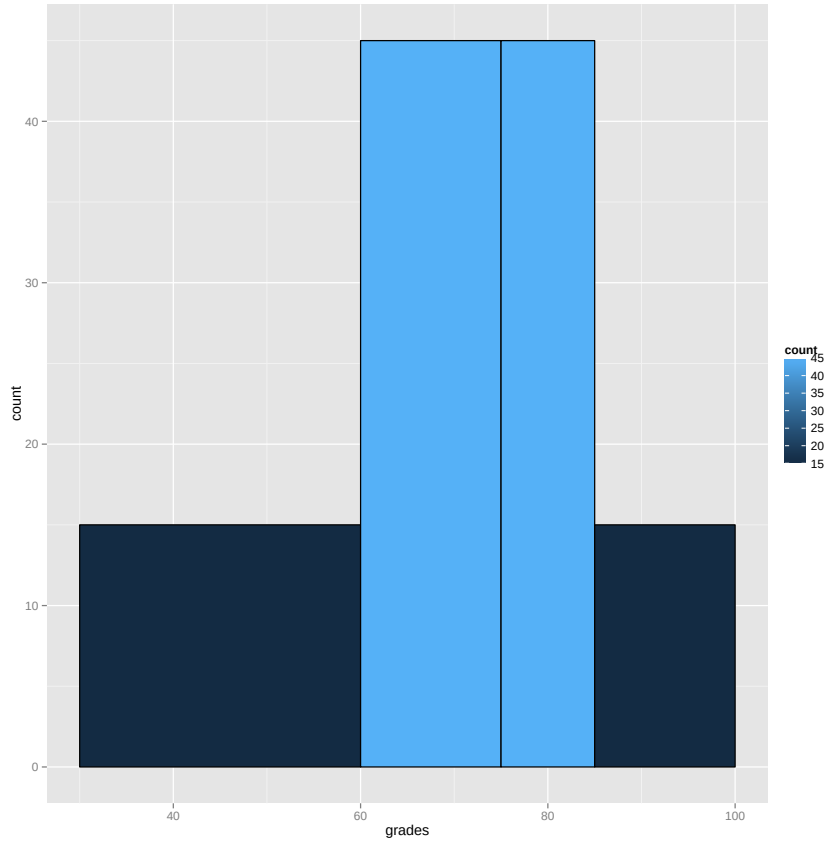
Given the grades in an engineering faculty were as follows:

lower	higher	graded
30	60	15
60	75	45
75	85	45
85	100	15

1. Present the data using a histogram.
2. Calculate mode, median, algebraic average and variance.
3. Calculate the number of students who earned at least 82 points.
4. Given the following data for the preceding year for 80 students where the average grade was 70 and variance was 200, find the average and the variance for two years combined.

1.1.1 Answer 1

```
library(ggplot2)
tbl$avg <- (tbl$lower + tbl$higher) / 2
ggplot(tbl) +
    geom_histogram(
        aes(x = avg, weight = graded, fill = ..count..),
        breaks = unique(append(tbl$lower, tbl$higher)),
        position = "identity", colour = "black") +
            xlab("grades")
```



1.1.2 Answer 2

As easy to see from the diagram, there are two **modes**: one is in the (60, 75] range, an another is in the (75, 85] range.

The **median** is the value at the 60'th studen, which is easy to see from

the table as being the 75 points.

$$
\begin{aligned}
Md &= \frac{\frac{n}{2} - F(x_{m-1})}{f(x_m)} * (L_1 - L_0) + L_0 \\
Md &= \frac{\frac{120}{2} - 60}{45} * (85 - 75) + 75 \\
Md &= \frac{60 - 60}{45} * (85 - 75) + 75 \\
Md &= 75
\end{aligned}
$$

The **average** is given by the formula:

$$
\begin{aligned}
\frac{15 * \frac{30+60}{2} + 45 * \frac{60+75}{2} + 45 * \frac{75+85}{2} + 15 * \frac{85+100}{2}}{15 + 45 + 45 + 15} &= \\
\frac{15 * 90 + 45 * 135 + 45 * 160 + 15 * 185}{2 * 120} &= \\
\frac{17400}{240} &= \\
72.5
\end{aligned}
$$

And the **variance**:

$$
\begin{aligned}
S^2 &= \frac{\sum_1^n (x - \overline{x})^2 * f(x)}{n} \\
S^2 &= \frac{(45 - 72.5)^2 * 15 + (67.5 - 72.5)^2 * 45 + (80 - 72.5)^2 * 45 + (92.5 - 72.5)^2 * 15}{120} \\
S^2 &= \frac{756.25 * 15 + 25 * 45 + 56.25 * 45 + 400 * 15}{120} \\
S^2 &= \frac{21000}{120} \\
S^2 &= 175
\end{aligned}
$$

1.1.3 Answer 3

Using a slightly altered formula for the median, we can calculate the 82'nd percentile. It is easy to see that the 82'nd percentile falls in the third group, viz. [75, 85) interval. Assuming values are uniformly distribute in this

interval, $\frac{7}{10}$ of these are below 82 and $\frac{3}{10}$ are above. In other words, we need to take 0.3 of the 45 students in this category, i.e. 13.5, together with 15 students who earned more points this gives roughly 19 students.

1.1.4 Answer 4

New average is just the weighted average of both averages: $\frac{70*80+72.5*120}{80+120} = 71.5$.

The total variance is calculated using

$$\begin{aligned}
s_1 &= n(S_1 + \overline{x}_1^2) \\
s_1 &= 120 * (175 + 72.5^2) \\
s_1 &= 651750. \\
s_2 &= n(S_2 + \overline{x}_2^2) \\
s_2 &= 80 * (200 + 70^2) \\
s_2 &= 408000.
\end{aligned}$$

$$\begin{aligned}
S^2 &= \frac{s_1 + s_2}{n_1 + n_2} - \overline{x}^2 \\
S^2 &= \frac{651750 + 408000}{200} - 71.5^2 \\
S^2 &= \frac{651750 + 408000}{200} - 5112.25 \\
S^2 &= 186.5
\end{aligned}$$

1.2 Problem 2

Given the number of assignments submitted in 2009 (shown below):

assignments	students
0	11
1	18
2	28
3	22
4	15
5	16

1. Draw a bar chart representing the data.
2. Find mode, median, algebraic average and variance.
3. In addition to the number of assignments submitted, students also received final grades. Let X be the number of assignments submitted, let Y be the grade the student received. Given also that Pearson coefficient is $r = 0.75$.

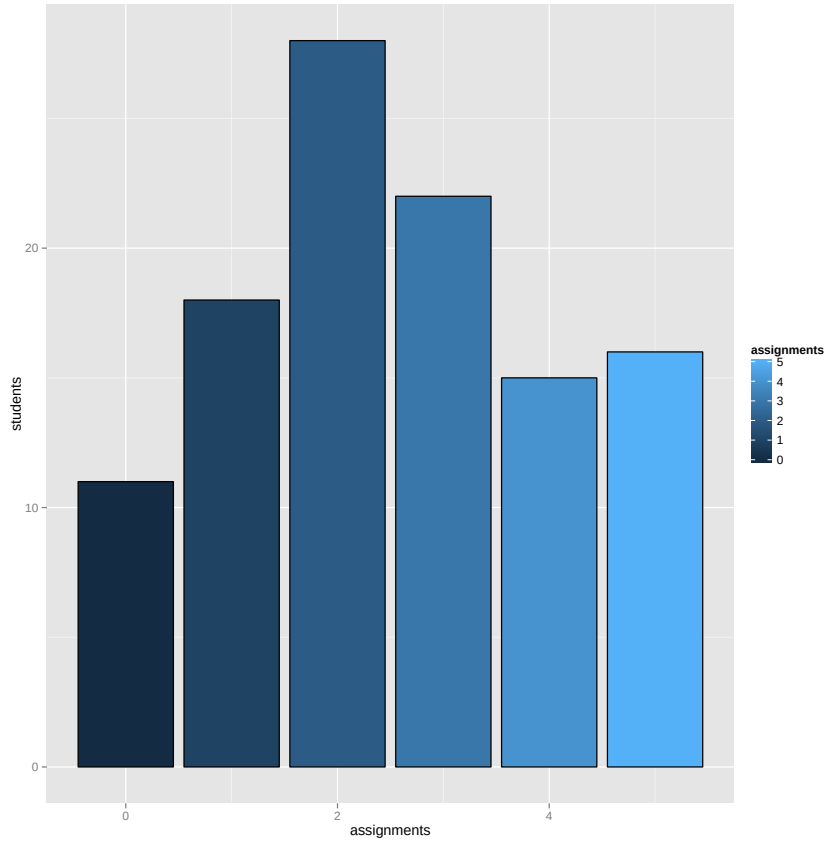
   Prove or disprove:

   (a) $Y = 0.75X + 96.25$.

   (b) The number of assignments submitted negatively correlates with the final grade they received.
4. An investigation found data on 10 more students. 5 of them didn't submit any assignment and 5 of them submitted 5 assignments each. Describe what will happen to each metric calculated in question 2 relying on the previously obtained values.

1.2.1 Answer 5

```
library(ggplot2)
ggplot(data = tbl,
       aes(x = assignments, y = students, fill = assignments)) +
           geom_bar(colour = "black", stat = "identity")
```



1.2.2 Answer 6

Mode for the assignments data is 2, i.e. most of the students submitted only two assignments.

Median is $2 + (3 - 2)/2 = 2.5$ (because there is an even number of bins).

Weighted **average** can be obtained using $\frac{0*11+1*18+2*28+3*22+4*15+5*16}{110} = \frac{280}{110} = 2.45(45)$.

Variance can be obtained using:

$$S^2 = \frac{(0*11)^2 + (1*18)^2 + (2*28)^2 + (3*22)^2 + (4*15)^2 + (5*16)^2}{110} - 2.45^2$$
$$S^2 = \frac{17816}{110} - 6.0025$$
$$S^2 = 161.963636364$$
$$S^2 \simeq 162$$

1.2.3 Answer 7

While it seems appealing, it isn't really possible to have a determination coefficient predict the value of another variable with absolute certainity unless the coefficient is equal to one. Thus $Y = 0.75X + 96.25$ is at best a lucky coincidence.

Positive coefficient means that there exists positive correlation between two variables. In particular, it means that roughly in three fourth of all cases the number of assignments submited perfectly predicted the final grade. So the claim is obviously false.

1.2.4 Answer 8

After adding ten more observations the **mode** will not change as the observations fall into the bins which aren't as dense as the most dense one, i.e. the first bin will contain 11+5=16 students *(fewer than 28 of the densest bin)* and 6+5=11 in the last bin.

The **median** will not change either because we are adding observations to the outermost bins in an equal measure.

The **average** will almost not change because the added observations will "cancel out", however, it will shift very slightly towards the upper end, since we increased the relative weight of the bin of those who submitted the most assignments.

Since **variance** is affected by the mean, it is hard to tell without doing full recalculation whether it will or will not change. However, since we added more borderline observations, which are most likely to be far away from the mean, we'd expect the variance to grow.