# Solutions of assignment 2
# Statistics for EMCL

*Tamas Biro*

## Task 1

Original test scores are:

15, 13, 6, 7, 8, 16, 15, 17, 8, 6, 10, 13, 8, 14, 5, 15

In the following table I give the test scores ranked sorted from least to greatest. At the same time, to save space, I also calculate the deviation of each score ( $x_i - \bar{x}$ , knowing that the mean is $\bar{x} = 11$ , as we shall soon see it), as well as the square of these deviation values.

| Test score (ranked) | Deviation from mean | Square of deviation |
|---|---|---|
| 5 | -6 | 36 |
| 6 | -5 | 25 |
| 6 | -5 | 25 |
| 7 | -4 | 16 |
| 8 | -3 | 9 |
| 8 | -3 | 9 |
| 8 | -3 | 9 |
| 10 | -1 | 1 |
| 13 | 2 | 4 |
| 13 | 2 | 4 |
| 14 | 3 | 9 |
| 15 | 4 | 16 |
| 15 | 4 | 16 |
| 15 | 4 | 16 |
| 16 | 5 | 25 |
| 17 | 6 | 36 |

Thick lines show the "position" of 1st quantile, median and 3rd quantile. As there are 16 pieces of data, these three values should be determined such that four cases have a value lower than Q1, four cases have a value between Q1 and the median, four cases have a value between the median and Q3, and finally four cases have a value greater than (or equal to) Q3.

A histogram is a graph that shows which value was observed in how many of the cases. It plots the frequency of each (possible/observed) values of the variable. On the x-axis you show values 5-17 (at least), and each bar is as high as the number of cases having exactly that values (0, 1, 2 or 3 in our case.)

You can observe that the histogram = the distribution of the data is approximately symmetric, not really skewed and has two peaks, that is, two modes. In fact, it is the combination of two “hills”. It is very different from a Normal distribution, but each hill can be seen as a rough approximation to some Normal distribution. (Actually, the number of data is very small to be able to tell whether these hills are Normal-like, or whether they follow a different but still “bell-like” distribution.) It does not make sense to speak of a distribution that is “skewed both to the left and to the right”, as some of you have mentioned.

Two **modes**: 8 and 15 (both having frequency 3), that is, we have a **bimodal distribution**. Most probably, dyslaxic children are those who produce the left “heap” amd the normal children are those who produce the right “heap”.

Using the above table, here are the statistics asked for:

Min: 5 (the lowest value)

Max: 17 (the greatest value)

Median: $\frac{10+13}{2}=11.5$

1st Q: $\frac{7+8}{2}=7.5$

3rd Q: $\frac{15+15}{2}=15$

IQR: 7.5 (= *Q3-Q1*)

Mean: 11 (sum of the first column, divided by *n=16*)

Var n-1: 17.07 (the sum of the third column = 256, divided by *n-1=15*)
(value from the calculator: 17.06666667)

Var n: 16 (the sum of the third column = 256, divided by *n=16*)

Std dev. n-1: 4.13 (root square of var n-1)

Std. Dev. n: 4 (root square of var n)
(value from the calculator: 4.131182235954578010857883093 1012)

No need to display all digits computed by your calculator! Show one or two digits after what is relevant. Now, numbers vary tipically in their first digits left to the point, so give at most two decimal digits. **Do not forget to round up, if the first digit you don't show is 5 or greater!** So the variance n-1 is 17.07, and not 17.06!

# Task 2

**Question 1:**
Population: members of club.
Sample: the twenty people being questioned.
NB: 20 is the size of the sample, not the sample.
Variable: years of membership.
NB: 0, 1, 2, etc., 8, 9 are the possible values of the variable.
Parameter: mean of the variable (average of years of membership) in club.
Statistic: mean of the variable (average of years of membership) in the sample.

**Question 2:**

To create a histogram, you first count which value occurs how often in your sample:
value 0: 2 times,
value 1: 3 times,
etc.
value 9: 6 times.

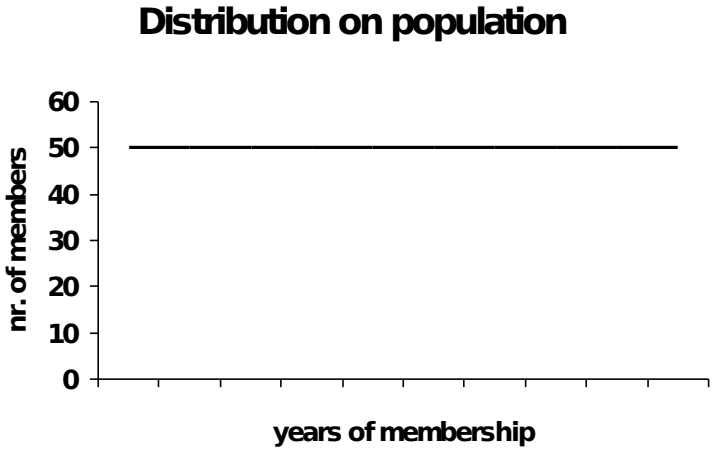
Then you draw the *histogram*, and not – what many of you have done – the cumulative proportion!

Sum of the first 20 digits in row 105 is (by pure coincidence) 105:
$(2 \times 0) + (3 \times 1) + \ldots (6 \times 9) = 105$

Mean: 105/20 = 5.25, and this is the statistic we are measuring on the sample.

**Question 3:**
The distribution is a horizontal line, because there are 50 people at value 0, another 50 people at value 1, etc. and finally 50 people at value 9.

| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 |




In total, there are 10 cohorts of 50 people, that is, 500 people in the Club. The mean of this distribution is:

$$(50 \times 0 + 50 \times 1 + 50 \times 2 + \ldots + 50 \times 9\ ) / (50 \times 10) =$$
$$= (0 + 1 + 2 + \ldots + 9) / 10 = 4.5.$$

For some funny reason, most of you have summed up the values 50, 100, 150, etc. up till 500, and so you got a mean of 5.5, instead of 4.5. But nobody has been a member of the club for 10 years!

**Questions 4 (column 3) and Question 5 (columns 2 and 4):**

| **Nr line** | **First 10 Mean** | **First 20 Mean** | **All 40 Mean** |
|---|---|---|---|
| 101 | 3.8 | 4.10 | 4.175 |
| 102 | 4.6 | 4.35 | 4.425 |
| 103 | 5.0 | 4.80 | 4.650 |
| 104 | 5.2 | 4.60 | 4.625 |
| 105 | 5.0 | 5.25 | 5.250 |
| 106 | 3.8 | 4.35 | 4.525 |
| 107 | 5.8 | 4.65 | 4.650 |
| 108 | 3.4 | 4.30 | 4.925 |
| 109 | 4.2 | 4.20 | 4.800 |
| 110 | 6.3 | 5.70 | 5.225 |

Please note: As the mean of the first ten numbers has always one digit after the point, I don't give more digits. Similarly, I **always** use two digits for the mean of the first 20 numbers, and three decimal digits for the mean of all forty numbers. (Well, the last one is certainly uninteresting, so I should have omitted it.)

I always give the seemingly redundant 0 as the last digit. This shows that, for instance, 5.70 has to be understood as being between 5.6950 and 5.7049. If I gave 5.7, that would mean something between 5.650 and 5.749.

Now, I calculate the mean and the distribution of the ten sample means, for each column (each value of $n$).

**Second column:** *n=10*
mean = 4.71 , standard deviation = 0.929.
**Third column:** *n=20*
mean = 4.63, standard deviation = 0.505.
**Fourth column:** *n=40*
mean = 4.725, standard deviation = 0.337.

So, we are interested in the distribution of the sample means, because the Central Limit Theorem says that if we took all possible samples (not only ten of them), then this distribution would be a Normal distribution, it would have the same mean as the population (4.5 in our case), and a standard deviation that decreases as the sample size increases.

We are *not* interested now in the standard deviation of each sample, and the mean of these std. deviations; only in the standard deviation of the means.

What can we observe? The mean of the ten statistics (ten different sample means) is quite close to the parameter (the population mean: 4.50, see Question 3) – for each sample sizes. It is noteworthy that sample size $n$ = 40 (fourth column) gives a slightly "worse" result – but that just happens if you do random processes. Still, even in this case, the value of the parameter (4.50) is amply within one standard deviation of the mean of the sampling distribution of the statistic being measured on ten samples (4.725 ± 0.337).

The second observation is the standard deviation of the sampling distribution diminishes as sample size grows. Consistently with the Central Limit Theorem, as the sample size grows four time larger (from 10 to 40), the standard deviation of the sampling distribution diminishes to approximately its half (from 0.929 to 0.337). Increasing the sample size to its double diminishes the standard deviation of the sampling distribution approximately to its √2 = 1.41.

The standard deviation of the mean can be calculated thus:

$$sigma^2 = \frac{1}{499}\sum_{i=0}^{9} 50\cdot(x_i - 4.5)^2 = \frac{50\times 2}{499}(4.5^2+3.5^2+2.5^2+1.5^2+0.5^2)$$

Therefore, the standard deviation of the population is *sigma* = 2.88.
You can check the prediction of the **Central Limit Theorem** for the standard deviation of the sampling distribution of the means:
For *n=10, sigma*/√n = 0.912, and we got 0.929.
For *n=20, sigma*/√n = 0.645, and we got 0.505.
For *n=40, sigma*/√n = 0.456, and we got 0.337.

# Task 3

**1. 69.15%**

Explanation: $x = 92$ is exactly at 0.5 standard deviation left to the mean. Therefore, we check $z = -0.5$. The area under the standard Normal curve left to $z = -0.5$ is 0.3085 according to the Standard Normal Table.

As the total area under the curve is 1, you find $1 - 0.3085 = 0.6915$ (which is 69.15%) right of this $z$ value.

**2. 120.5.**

We search for 0.9000 in the Standard Normal Table, that is, the $z$ value left of which 90% of the area can be found. At $z = 1.28$ we find 0.8997 and at $z = 1.29$ we find 0.9015.

So 0.9000 corresponds approximately to $z = 1.282$. That is, 1.282 times standard deviation right of the mean you find the $90^{th}$ percentile. In the case of our distribution (mean=100, standard deviation = 16), this corresponds to $100 + 1.282 \times 16 = 120.512$.

What I missed in almost all assignments of yours, is **interpolation**. You usually looked for the value that is closest to 0.9, which is indeed for *z=1.28*. (Some of you have used *z=1.29* instead, which falls further away from 0.9.) *Interpolation* means that in such a case you guess which value between 1.28 and 1.29 would approximately return 0.9000. The difference between 0.8997 and 0.9015 is 0.0018. Now, 0.9000 is 0.0003 away from 0.8997, that is, 1/6 of "distance" 0.0018. Supposing that the values are approximately linearly distributed between *z=1.28* and *z=1.29*, you would guess $1.28 + 0.01 \times (1/6)$, and this is how I got the guess z = 1.282.