Students performances

Data set analysis: StudentsPerformance.csv

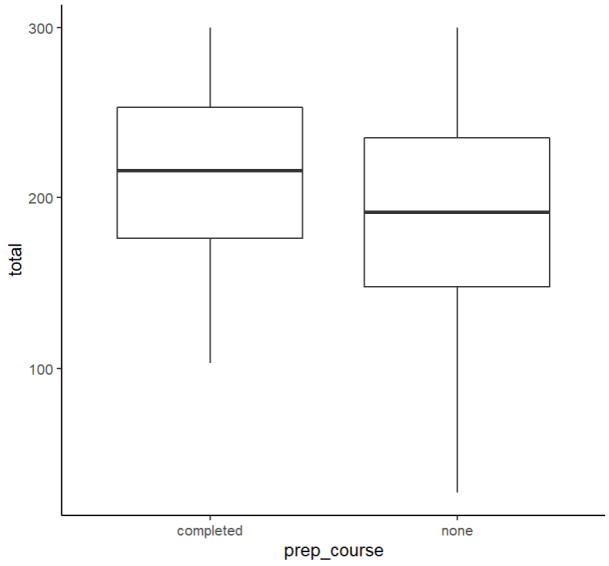
This data set provides pieces of information on 1000 American students including marks in tests of basic skills like math, reading and writing. Here it is a snapshot of the rearranged tibble:

gender	ethnicity	parents_ed	lunch	prep_course	math	reading	writing
female	group B	bachelor's	standard	none	72	72	74
female	group C	some college	standard	completed	69	90	88
female	group B	master's	standard	none	90	95	93
male	group A	associate's	free/reduced	none	47	57	44
male	group C	some college	standard	none	76	78	75

where the categorical variable *lunch* stands for the fee paid for lunch at school and the binary variable *prep_course* displays whether the student attended the preparatory course for the test or not.
However, in order to have a better understanding of the situation, it is interesting to add four additional variables: the mean score in language skills, the overall mean score, the total score and a boolean variable showing whether the test was passed or not with a threshold of 60 in the overall mean score.
Now that we have everything we need, we can start do some analysis.

1 The importance of preparation

The first correlation we might be interested to address is the one between the attendance to the preparatory course and the total grade.



We can infer from the boxplot that there is some discrepancy, but let's be more precise and carry out a proper p-value hypothesis testing on it.

- **STEP 0**: set a null hypothesis H_0.
  Suppose there is no difference at all between the overall grade of the two groups.

- **STEP 1**: compute a test statistic t_0 of the data.
  Since our null hypothesis is that the preparatory course has no influence on the overall grade, it seems

reasonable to set the test statistic equal to the difference of of the total scores mean for those who completed the course and for those who did not attend it.

$$t_0 = mean(total[\text{completed}]) - mean(total[\text{none}]) = 218 - 195.1 = 22.9$$

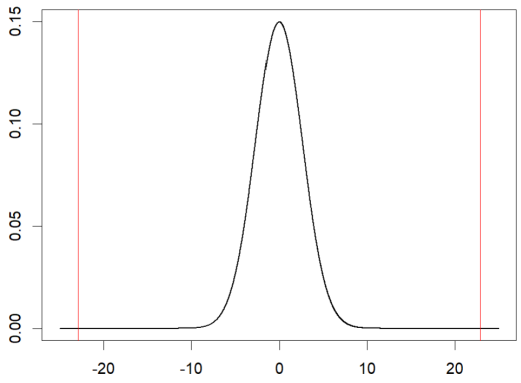
- **STEP 2**: assess the distribution T of t_0 under H_0.
  It is evident that the total average results of both those who completed the course and those who did not attend it are normally distributed around the mean score with variance equal to the variance of the group.

$$t[\text{completed}] \sim \mathcal{N}(218, 1530)$$
$$t[\text{none}] \sim \mathcal{N}(195.1, 1811)$$

  Also, since we have 1000 observation, the data set is considerable large enough to apply the central limit theorem: therefore, T is nothing more than the difference of two gaussian random variables with mean equal to the difference of the means, that is 0 under H_0, and variance equal to the sum of the variances, each one divided by the size of the group.

$$\begin{aligned} T &\sim \mathcal{N}(mean, \frac{1530}{358}) - \mathcal{N}(mean, \frac{1811}{642}) \\ &\sim \mathcal{N}(0, 7.09) \end{aligned}$$

- **STEP 3**: computation of the p-value.
  Now that we have produced our model, we shall compute the probability, if H_0 was true, to obtain absolute values of t_0 equal or bigger than the one we have registered.



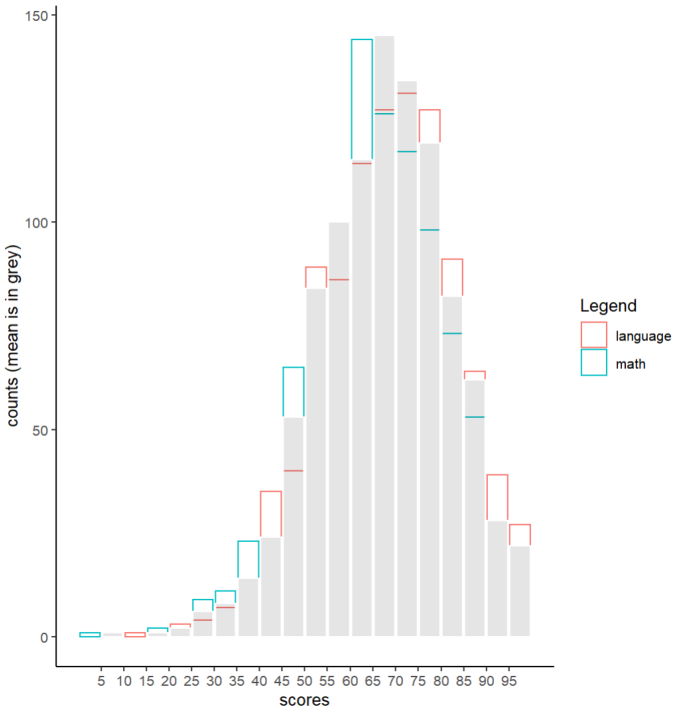
  The corresponding two-sided p-value is:

$$2 \cdot pnorm(22.9, 0, \sqrt{7.09}, lower.tail = FALSE) = 7.95e-18$$

  and leads us to reject the null hypothesis H_0 in favor of a new hypothesis H_1: students who attended the preparatory course do better than the others.
  Still, we have to be aware of possible confounders or other aspects that could contribute in making such results arise: for example, students that are good at studying tend to be more committed and to take seriously anything that has to do with school, so they are more likely to attend the course, even though they wouldn't need it.

2 Skills differentiation

Another interesting figure in the data set is the variation in the results for different parts of the test, especially when considering the language score and the math score.

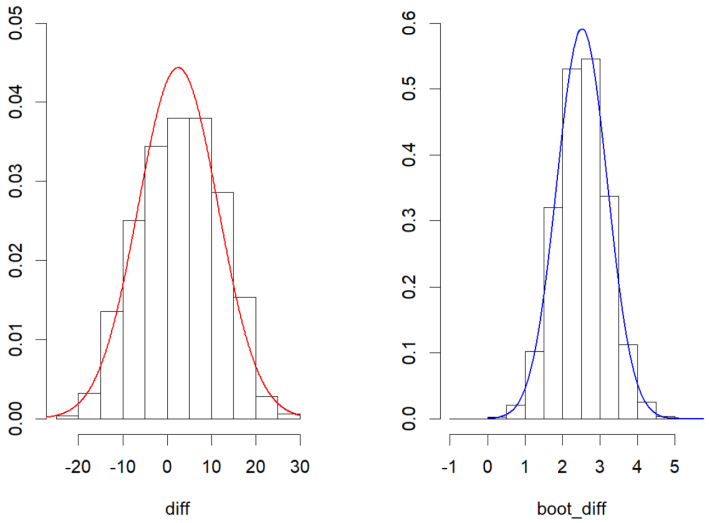


Again, a rough representation of the data helps us inferring something, but it is always better to proceed with hypothesis testing: in this case for example, it might be insightful to use a bootstrap simulation.

- **STEP 0**: set a null hypothesis H_0.
  Suppose that each student have equally developed both language skills and math skills.
- **STEP 1**: compute a test statistic t_0 of the data.
  As in the previous example, the correct thing to do here is to set the test statistic equal to the difference of the means in the two different scores.

$$t_0 = mean(\text{language}) - mean(\text{math}) = 68.61 - 66.1 = 2.51$$

- **STEP 2**: simulate the distribution T of t_0.
  With bootstrapping, we can easily simulate the distribution of the test statistic by ideally repeating the experiment a huge number of times: just extract a new sample from the original one with replacement, compute the average grade in language and math and then store their difference.
  The results are reported below.



We can notice that the set of values for the possible differences in the bootstrap case is concentrated around the mean value 2.52 with variance 0.45.

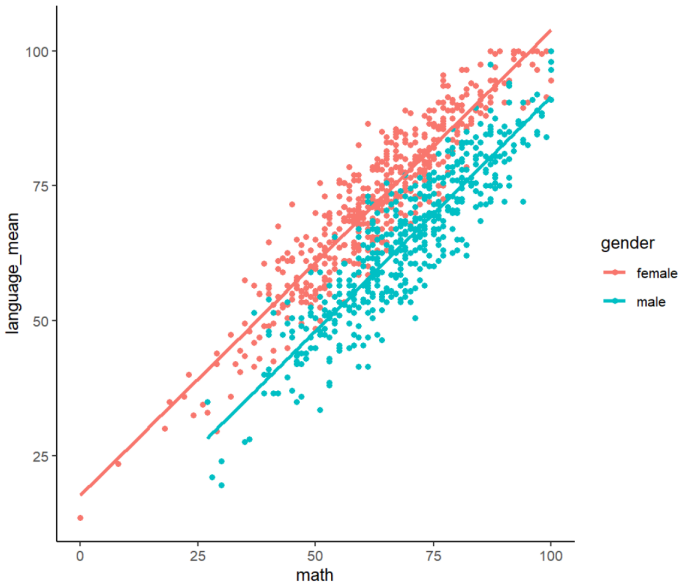
- **STEP 3**: creation of the confidence interval.
  Differently from what we have done before, now that we have found an approximation of our test statistic, we shall calculate the probability that such t_0 takes on values around the one proposed in H_0.
  To do so, we can simply set a confidence interval for the result got with the bootstrap simulation, say a 95% confidence interval since we usually set the p-value for statistical significance to be 5%.

$$qnorm(.025, 2.52, \sqrt{.45}) = 1.2$$
$$qnorm(.975, 2.52, \sqrt{.45} = 3.84$$
$$CI = (1.2, 3.84)$$

  This interval does contain neither 0 nor values around it and clearly leads us to discard the null hypothesis H_0 in favor of the new hypothesis H_1: students have better language skills than math skills.

3 Gender differentiation

We have just seen that there is some discrepancy in the different skills score, so to conclude our analysis we could investigate whether such variation has something to do with the gender of the student or not.



The gap is evident in this scatter plot and the rigorous hypothesis testing that we should apply here is very similar to the one about the correlation between the preparatory course and the overall score, so we can skip that part right here and still state that males are better in math and females in language skills.
However, what would be interesting now is investigating whether such discrepancy lasts even when guys and girls grow up, maybe comparing university enrollment rates.