ADS2 Practical 8: Simulation-based comparison of two means

Facilitators: Gedi Lukšys, Chaochen Wang

Mon, November 4, 2019, 14-16

Learning objectives:

- Understand the logic behind hypothesis tests using simulations
- Learn how to compare mean of a sample from a distribution to a value
- Learn how to compare means of two samples from two distributions
- Appreciate the advantages of simulation-based approach

The task:

Imagine that a country runs an exam system using normative scoring, where each student first gets a raw score based on performance in the exam, and later student's normative score is computed as corresponding to the percentile of all students in the country, hence it can be effectively approximated using a uniform distribution between 0 and 100 (for the sake of simplicity let's assume all scores are not rounded).

Question 1

Let's assume that our class has 26 students whose score distribution follows the same one as in the whole country. We want to know the probability that the mean of their normative scores is lower than 40. How can we find this out using simulation?

For uniform distribution you can use function *runif* (use *?* to figure out its parameters)
Once you get your result (probability of the mean being below 40) try to explain it.
Is it sensitive to number of repetitions in your simulation?

We generate many samples of 26 from the described uniform distribution and calculate how many means are below 40 – that corresponds to the probability in question, e.g.

```
> task1 = replicate(10000, mean(runif(26, 0, 100)))
> mean(task1 < 40)
[1] 0.0395
```

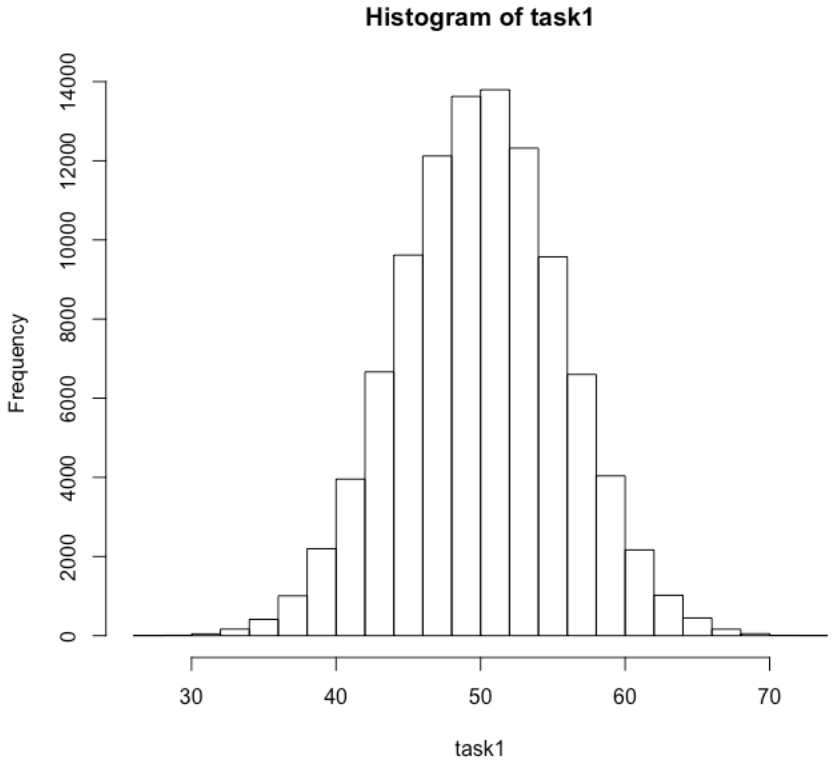
The P value being less than 0.05 suggests that there is a statistically significant difference between 40 and the mean of normative scores. However, it's important to note that this is only the case if we use a one-tailed test. For a two-tailed test with the same significance level we would want $P < 0.025$.

If we change the number of repetitions the resulting probability varies either a lot (in case of few repetitions) or little (in case of many repetitions), e.g.

```
> task1 = replicate(100,mean(runif(26,0,100)))
> mean(task1 < 40)
[1] 0.01
> task1 = replicate(100,mean(runif(26,0,100)))
> mean(task1 < 40)
```

```
[1] 0.03
> task1 = replicate(100,mean(runif(26,0,100)))
> mean(task1 < 40)
[1] 0.06
> task1 = replicate(100000,mean(runif(26,0,100)))
> mean(task1 < 40)
[1] 0.03942
> task1 = replicate(100000,mean(runif(26,0,100)))
> mean(task1 < 40)
[1] 0.03851
> task1 = replicate(100000,mean(runif(26,0,100)))
> mean(task1 < 40)
[1] 0.03769
```

You can plot the distribution of means.
Does it look similar to the uniform distribution? If not, why not?

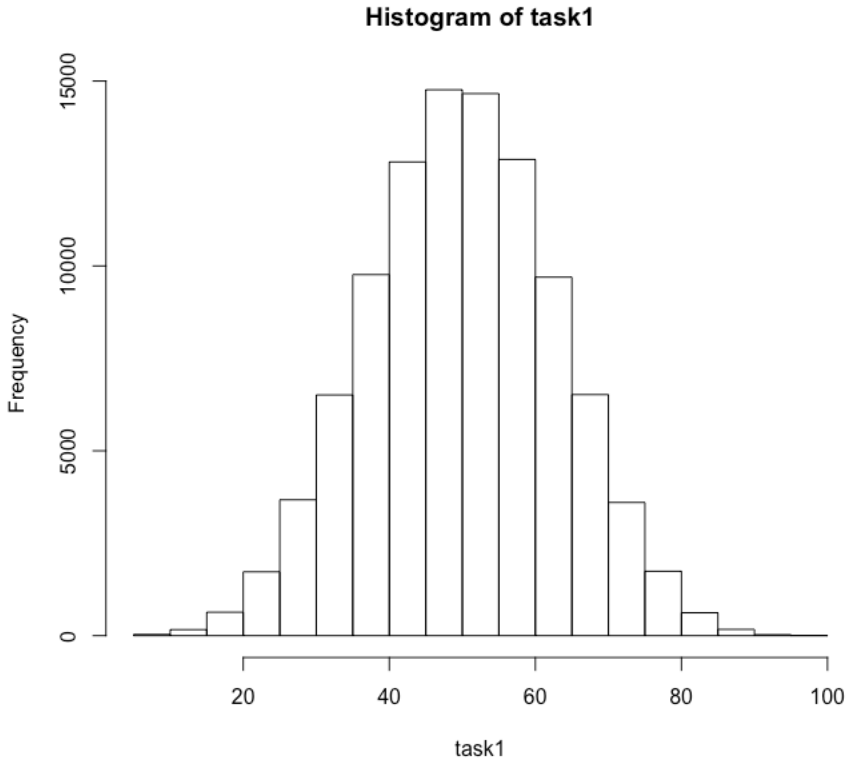


No, it looks similar to the normal distribution because of the Central Limit Theorem.

How would it look if the class only had 5 students?

```
> task1 = replicate(100000,mean(runif(5,0,100)))
> mean(task1 < 40)
[1] 0.2261
> hist(task1)
```

The histogram below also looks similar to the normal distribution; however, it is much wider because standard error of the mean is much higher due to smaller sample size.
The P-value (of mean being below 40) is also considerably higher and well above 0.05.



Question 2

Now imagine that one class of 26 students had a careless administrator who didn't notice the 5th exam question and only printed the first four. Let's **assume** that as a result of this, the normative scores of students from this class followed a uniform distribution between 0 and 80 (with the mean of 40). We now want to know what is the probability that this class did better than another class of 26 that didn't have such bad luck. How can we do this using simulation?

As in question 1, we generate many samples of 26 from the uniform distribution from range [0, 80] and compare their means with other samples randomly generated from the uniform distribution from range [0, 100].

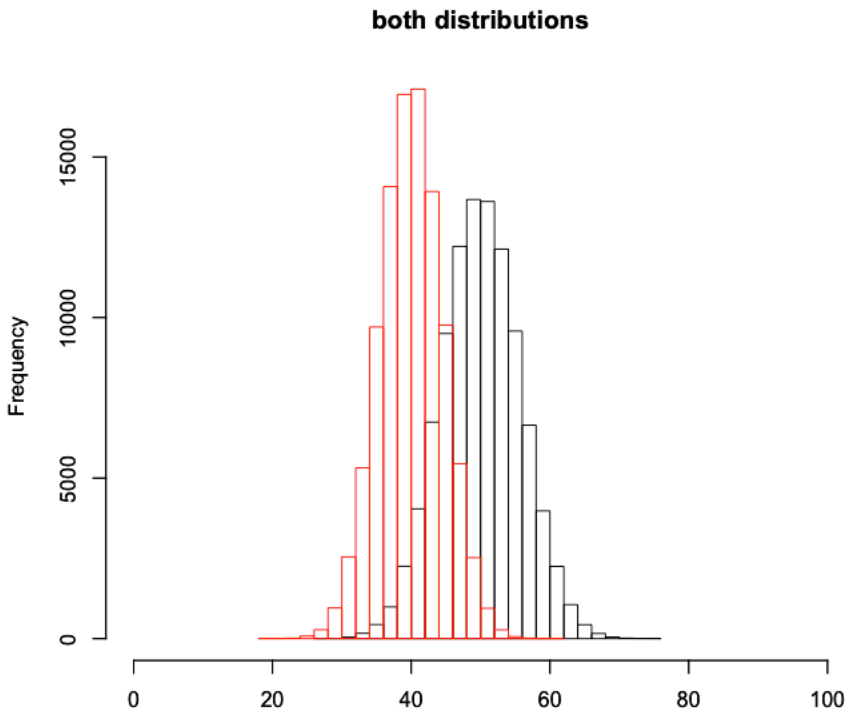
```
> task2unlucky = replicate(100000, mean(runif(26,0,80)))
> task2control = replicate(100000, mean(runif(26,0,100)))
> mean(task2unlucky > task2control)
[1] 0.08399
```

Is the result the same as in the previous question? If not, why not? You can also plot both distributions of means (ideally on one figure).

No, because instead of comparing means of samples from one distribution to a number we now compare means of samples from two distributions, hence the variability of the 2nd distribution also matters. In practical terms it means that the mean of the "unlucky" sample may be compared with 45 as likely as it may be compared with 55 with the resulting P value considerably higher due to non-linearity.

For plotting you can use the following code:

```
> hist(task2control, xlim=c(0,100), ylim=c(0,17000), border="black", xlab =
"", main = "both distributions")
> par(new=TRUE)
> hist(task2unlucky, xlim=c(0,100), ylim=c(0,17000), border="red", xlab =
"", main = "both distributions")
```



Question 3

As this exam had a clear scoring system and students could take a copy of their papers, a student Leonie found that she got a 64. Based on statistics from previous years of students taking a similar exam, she has found out that their raw scores follow a normal distribution with a mean of 50 and standard deviation of 10 (for which you can use function *rnorm*).

What would be the expected normative score Leonie likely got? Why?

It would be a cumulative distribution function of a Gaussian with mean = 50 and s.d. = 10:

```
> pnorm(64,50,10)
[1] 0.9192433
```

We need to multiply by 100, as normative scores are between 0 and 100.
Hence her likely normative score is approximately 92.

Leonie and her three friends got the following raw scores: 64, 63, 62, 59. Her friend Sheldon is a very bright student from another class. He and his friends got the following raw marks: 70, 63, 61, 56. As a result, Sheldon is boastful that their average is higher. However, having a good understanding of data science Leonie thinks that her team will have the last laugh once the normative scores are out. Is she right? Why or why not? Use simulations and plots to support your argument.

We first need to calculate normative scores of both teams and then compare their averages.
For Leonie's team it's

```
> mean(pnorm(c(64,63,62,59),50,10)*100)
[1] 88.08283
```

And for Sheldon's team it's

```
> mean(pnorm(c(70,63,61,56),50,10)*100)
[1] 86.76326
```

As you see, Leonie is indeed right and the reason the order of averages changes is non-linearity of the cumulative distribution function. Basically, the normative score of the lowest scoring student in Sheldon's team is much lower compared to the lowest scoring student of Leonie's team than Sheldon's normative score is higher compared to Leonie's.
The same result can be computed using simulations – see a similar example in Question 5.

Question 4

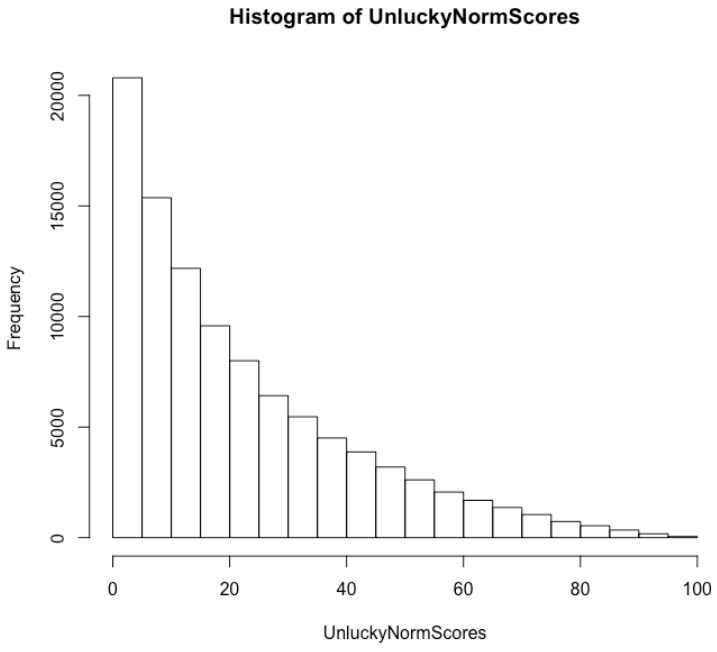
Now remember the unlucky class for which one exam question was not printed. What is wrong with the assumption in bold in question 2? Let's find and plot a distribution of normative (percentile-based) scores for this class if 1 out of 5 answers are missing, hence their raw scores are on average 20% lower but follow the same distribution (normal with mean = 40 and standard deviation = 8).

The assumption says that normative scores are lower by 20%, whereas that should be the case for raw scores, not normative scores. In order to obtain the distribution of normative scores for the unlucky class, we need to generate a distribution of raw scores for the unlucky class using *rnorm(100000, 40, 8)* and compute the corresponding normative scores based on cumulative distribution function values of the normal class distribution (as in Q3):

```
> UnluckyNormScores = pnorm(rnorm(100000,40,8), 50, 10)*100
```

How does this distribution look? Why is it such? What is its mean value?

It is an exponential distribution with mean = 21.8:



The intuitive reason why it is exponential is again because of non-linearity of conversion from raw to normative scores. The mathematical reason is much harder to derive. The raw score of 40 (which is average for the unlucky class) corresponds to a normative score of 15.9 (use *pnorm(40, 50, 10)* to check), so basically half of students in the unlucky class have their normative scores under 15.9, whereas only a very small number of students have scores above 50 (either raw or normative).

What is the probability that this class got a higher mean normative score than another class of 26 that didn't have bad luck of missing one problem?

We again use the same approach as in Q2, only now we sample from *UnluckyNormScores*:

```
> task4unlucky = replicate(100000, mean(sample(UnluckyNormScores, 26)))
> task4control = replicate(100000, mean(runif(26, 0, 100)))
> mean(task4unlucky > task4control)
[1] 4e-05
```

As the resulting P value is very small, it's important to run simulations sufficiently many times (100000 is an absolute minimum…)

Does the school's principal have a valid reason to worry that the unlucky class would be the worst in the country, assuming it has about 10000 classes of the same size?

As the probability of the unlucky class having a higher mean of normative scores than a normal class is less than 1/10000, the principal's fears that this would not occur with 10000 comparisons are indeed justified.

Use simulations to answer these questions.

Now assume the number of students per class actually varies, following a uniform distribution between 5 and 40 (with 10000 classes in total). Would that alleviate the principal's worries of having a class with the worst normative score average in the country? Why or why not?

Although the underlying distributions are the same, now the distribution of means changes for normal classes, as variation in class size for normal classes increases the variability of their means (again, as the increase of n=5 compared to n=26 is much more than decrease of n=40 compared to n=26). As a result, the probability of a class with a smaller mean than that of the unlucky class increases:

```
> task4control = replicate(100000, mean(runif(round(runif(1,5,40)),0,100)))
> mean(task4unlucky > task4control)
[1] 0.00117
```

As now it is well above 1/10000, it is no longer likely that the unlucky class would be the worst.

Question 5 (homework)

In reality it is difficult to expect that student **raw scores** from different schools and classes will follow the same distribution even if the vast majority of classes didn't have bad luck of missing out one question in five. We expect that due to differences in teaching quality and

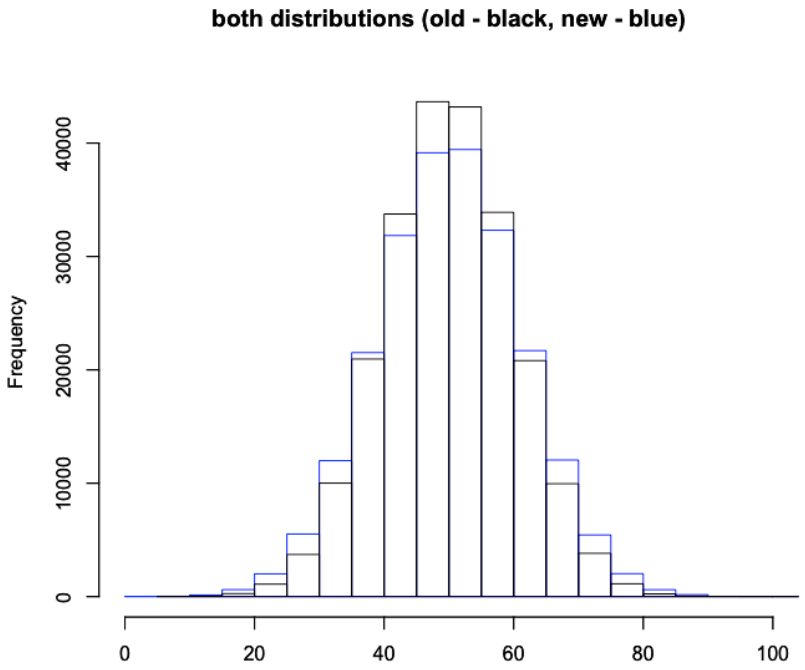
learning environment (some schools are better, others worse) **the means of each class** follow a normal distribution with mean = 50 and standard deviation = 5 whereas standard deviation is stable across the country (and is equal to 10, except for the unlucky class, for which it's 8). Class sizes also vary as in *question 4*, following a uniform distribution between 5 and 40, with 10000 classes in total.

With these changes generate your new student **raw score** distribution in the country and based on this recalculate a distribution of **normative scores** for the unlucky class (*as you did in question 4*). How do both distributions look and what are their means?

Unlike in previous questions, now we no longer can use *pnorm*. Instead, we first need to generate a distribution of raw scores in the country, then calculate the distribution of all normative scores for the unlucky class. The distribution of normative scores for the whole country remains uniform with mean = 50, but due to uneven class size and systematic differences between them, to obtain class means we can no longer sample randomly from the overall distribution. Instead, we should mark the raw scores for each class, so that after conversion into normative scores we could compute their means. As we need to have data both for each class and overall, we save the raw scores both as a vector and as a list.

```
> allrawvector = c()
> allrawlist = c()
> for (i in 1:10000) {
+ sampleclass = rnorm(round(runif(1,5,40)),rnorm(1,50,5),10)
+ allrawvector = c(allrawvector, sampleclass)
+ allrawlist = c(allrawlist, list(sampleclass)) }
```

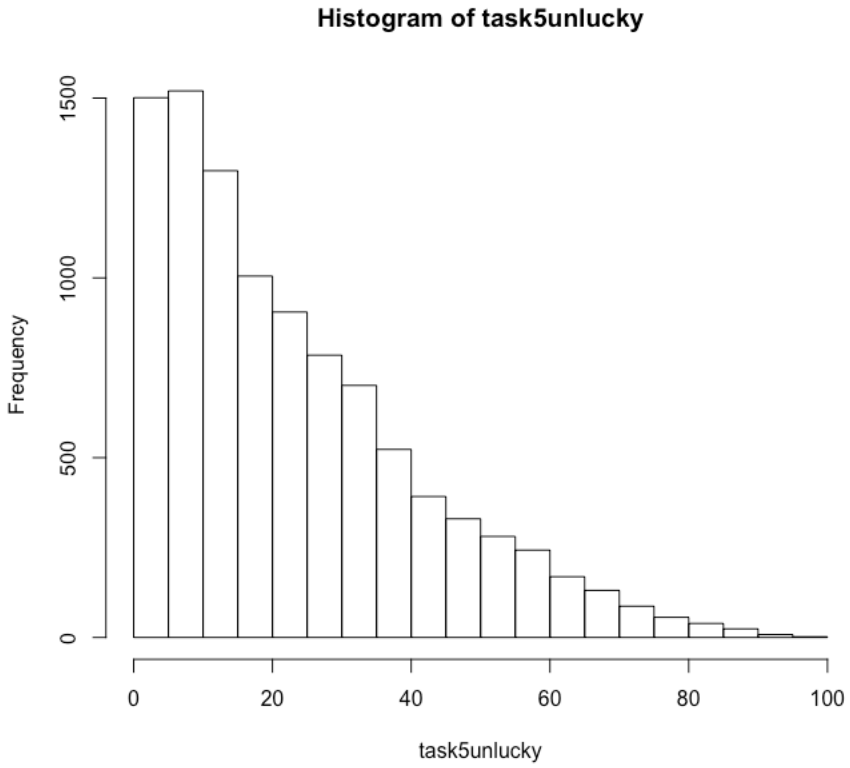
We can plot the new distribution of raw scores and compare it to the one before (normal with mean = 50 and s.d. = 10):



It is evident that they look similar; however, the new distribution is a bit wider than the old one as means of its classes vary no longer only randomly but also at the point of generation.

Now we generate normative score distribution for the unlucky class

```
> task5unlucky = c()
> for (i in 1:10000) {
+ task5unlucky = c(task5unlucky, mean(rnorm(1,40,8) > allrawvector)*100) }
> mean(task5unlucky)
[1] 23.03395
> hist(task5unlucky)
```



It is still similar to the exponential distribution (although no longer exactly it); however, the mean is larger than before, as the distribution of raw scores is now wider than before.

We also compute normative scores for each class, so that we can later compare their means

```
> allnormlist = c()
> for (i in 1:10000) {
+ classnorm = c()
+ for (j in 1:(length(allrawlist[[i]])))
+   classnorm = c(classnorm, mean(allrawlist[[i]][j] > allrawvector)*100)
+ allnormlist = c(allnormlist, list(classnorm)) }
```

This code can take a while to complete.

What is the probability that the unlucky class did better than another class in the country that didn't have such bad luck?

We simply take 10000 samples of size 26 from the normative score distribution of the unlucky class and compare their means with means of normative scores of 10000 normal classes computed above:

```
> count = 0
> for (i in 1:10000) {
```

```
+ if (mean(sample(task5unlucky,26)) > mean(allnormlist[[i]]))
+ count = count + 1 }
> count
[1] 288
```

As you can see this probability is now only 0.0288, which if using two-tailed test wouldn't even be reaching significance at 0.05 level.

How does the average of Leonie's team's normative scores now compare to that of Sheldon's team (as in *question 3*)?

```
> Leonie1 = mean(64 > allrawvector) * 100
> Leonie2 = mean(63 > allrawvector) * 100
> Leonie3 = mean(62 > allrawvector) * 100
> Leonie4 = mean(59 > allrawvector) * 100
> mean(c(Leonie1, Leonie2, Leonie3, Leonie4))
[1] 85.50875
> Sheldon1 = mean(70 > allrawvector) * 100
> Sheldon2 = mean(63 > allrawvector) * 100
> Sheldon3 = mean(61 > allrawvector) * 100
> Sheldon4 = mean(56 > allrawvector) * 100
> mean(c(Sheldon1, Sheldon2, Sheldon3, Sheldon4))
[1] 84.55328
```

Now due to a wider distribution of raw scores Sheldon's advantage over Leonie increases. However, Leonie's team still wins.

Use simulations to obtain these values and inferences (this time *pnorm* may not be of much help, as the overall distribution of raw scores in the country no longer follows the normal distribution with mean = 50 and s.d. = 10 – only that of each class does, albeit with different means that are themselves normally distributed)