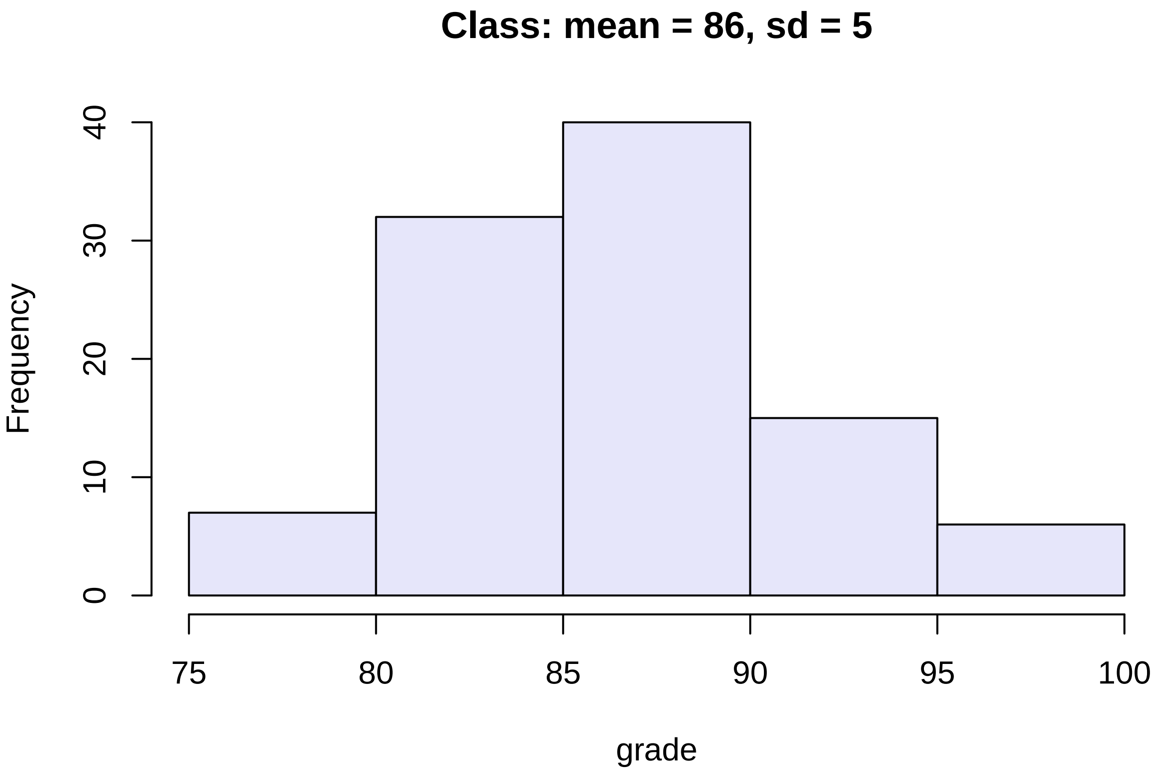
Problem Set 2: Notes

Applied Data Science 2

Semester 1 2022/23

Looking at student grades

You can create a classroom sample by drawing from a normal distribution with the required characteristics. Here is what this distribution may look like (although because you are drawing random numbers, yours will look a bit different)



In order to count students with grades lower or higher than two numbers, make use of the "or" operator: In R, you can write `A | B` for "A or B"

Remember also that Booleans are equivalent to using the number 1 for TRUE and 0 for FALSE, so in order to count the number of true entries in a list of Booleans, you can just take the sum.

```
grade_81_91 <- sum(class<81 | class>91)
grade_81_91
```

```
## [1] 25
```

```
grade_76_96 <- sum(class < 76 | class > 96)
grade_76_96
```

```
## [1] 5
```

As we will see later in this class, normal distributions have an interesting property: We expect around 68% of data points to be within one standard deviation of the mean, and around 95% to be within two standard deviations. How close to these numbers is your particular sample?

What problems may there be with creating a predicted grade distribution in this way? Here are two things you may have considered:

- Limits: You may or may not have paid attention to the upper and lower limits of your virtual classroom. If you sample from a random distribution with mean 86 and standard deviation 5, it may happen (not very often, but it may happen) that you draw a number larger than 100. You may also (albeit very, very rarely) draw numbers smaller than 0. There should be a strategy in place to disregard such numbers, because they are impossible in the context of a class mark.
- Nature of the distribution: We gave you a mean and standard deviation, and you can easily recreate a normal distribution with the same mean and standard deviation. But the distribution that the original mean and standard deviation came from may not have been normal. For instance, it could be that there was one student who performed very badly, but that everyone else performed at or slightly better than the mean. For instance, look at the following:

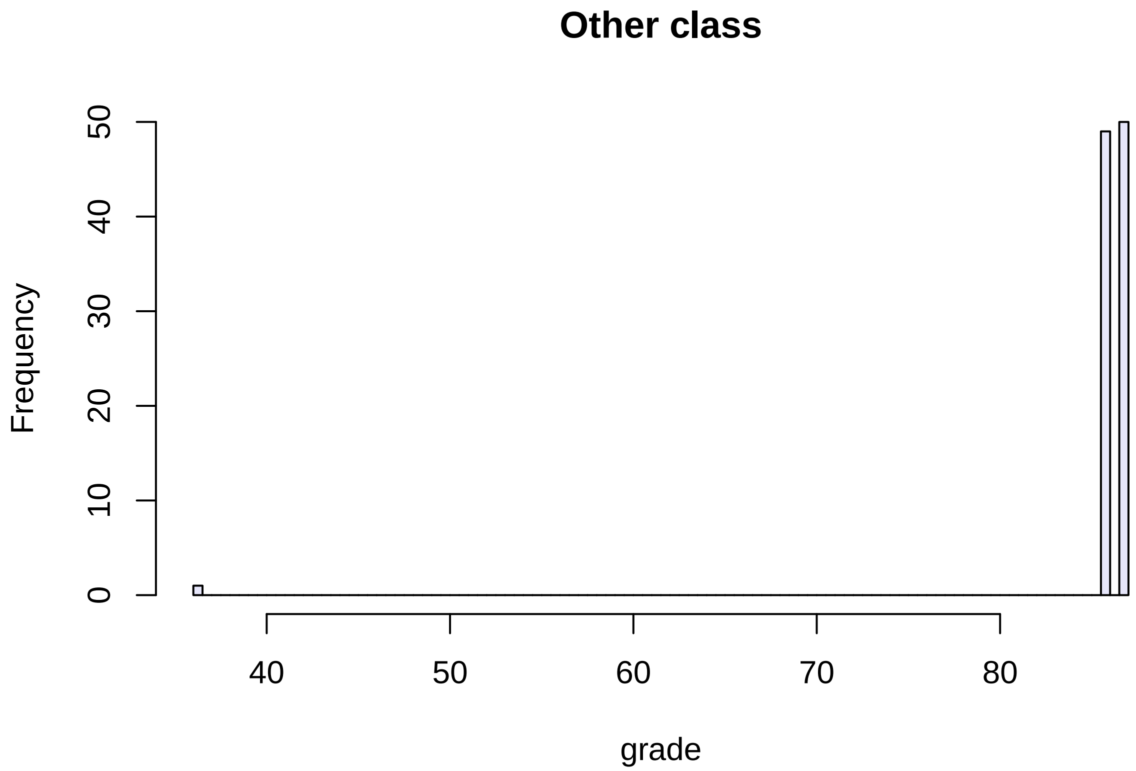
```
other_class = c(36, rep(86,49), rep(87,50))
mean(other_class)
```

```
## [1] 86
```

```
round(sd(other_class))
```

```
## [1] 5
```

```
hist(other_class, xlab="grade", breaks=100, col="lavender", main="Other class")
```



Getting good grades

Scenario 1: Random correct answers

Let's look at the first scenario first: The instructors have randomly chosen A, B, C, or D to be the right answer for each question. There are two ways to look at this: You could set up a simulation in R, or you could do the maths. You only needed to do one of those. But here we will answer the question in three different ways, so you can have a quick look at what you missed!

Simulation-based approach

If we want to run a quick simulation, we can generate a list of correct answers by drawing 20 random numbers between 1 and 4 (1 standing for A, 2 for B etc.)

```
correct <- sample(1:4,20,replace=TRUE)
```

Similarly, we can generate a list of a student's guesses:

```
guesses <- sample(1:4,20,replace=TRUE)
```

Now, all we need to ask in order to get the score is, how often are they the same?

```
score <- sum(correct==guesses)
score
```

```
## [1] 5
```

And is that score equal to or bigger than 10?

```
score >= 10
```

```
## [1] FALSE
```

Now, all you need to do is do this many times *(How many?)*

Here is our result from doing it one million times (this took several seconds to run)

```
## [1] 0.015
```

What if a student who just selects “A” every time? This is easier, because we don’t have to create a list of “guessed” answers, we just need to count how many times the number 1 appears in the list of correct answers. Here is the result (again, from one million simulations)

```
## [1] 0.015
```

Did you get similar results? Are they the same or not? This may be tricky to tell just from the simulation.

Maths-ing it out

Let’s do it again, but this time with a bit of maths! We will walk through the steps slowly. Let’s start with the scenario where the student guesses the answer to every question (i.e. chooses A, B, C, or D at random)

1. We are looking for the probability of a student passing the course, i.e. getting at least 10 questions correctly.
2. Let’s look at a single question. If the examiner chose the answer at random and the student did as well, the probability of a match is 0.25
3. If there are 10 questions, the probability of guessing all of them correctly would be 0.2510 (remember what you learned last year about the joint probability of a number of independent events: you just multiply). Also, there is a probability of 0.75^10 of *not* guessing the other 10 questions correctly. So the probability of getting, say, the first 10 questions correct and the last 10 questions wrong is 0.2510 ×0.75^10. In case you are interested, this is

```
(0.25^(10)) * (0.75^(10))
```

```
## [1] 5.370475e-08
```

4. This sounds bad, but there is good news: The student has several opportunities to get 10 questions correct. How many? As many as there are ways of choosing 10 questions out of 20. This number is written as $\binom{20}{10}$ (pronounced “20 choose 10”). In general, $\binom{n}{k}$ is $\frac{n!}{n!(n-k)!}$. (Where the exclamation mark is pronounced “factorial”, and the factorial of an integer is that integer multiplied by all integers that are smaller than it - for instance $4! = 4 \times 3 \times 2 \times 1 = 24$).

Maybe this is familiar to you, but maybe it’s not (or maybe it’s been a while). If you need a short exercise to refamiliarise yourself with this, try to figure out all the ways to choose 2 out of a set of 5, and convince yourself that this number is indeed $\binom{5}{2}$

Back to our student: They have $\frac{20!}{10!(20-10)!}$ chances of selecting the 10 answers where they guess correctly out of the 20 answers in the quiz.

So, the probability of getting any 10 questions right (and the other 10 wrong) is

$$0.25^{10} \times 0.75^{10} \times \binom{20}{10}$$

In R:

```
0.25^10*0.75^10*choose(20,10)
```

```
## [1] 0.009922275
```

(Note that here, we are just using R as a calculator, not a simulation engine. Note also that R conveniently has a choose() function)

5. This still sounds bad, but there is good news: You don’t only pass with 10 correct answers, you also pass with 11, 12, 13, 14, etc. For each of these, the probability of getting that exact score (let’s call it s) is

$$0.25^{s} \times 0.75^{20-s} \times \binom{20}{s}$$

And then we just have to sum over all those to get the passing probability.

```
p_passing = 0
for (s in 10:20) {
  p_s = 0.25^s*0.75^(20-s)*choose(20,s)
  p_passing = p_passing+p_s
}
p_passing
```

```
## [1] 0.01386442
```

6. This took a while! But it’s an exact result. Compare it to the result of the above simulations - what do you think?

7. And what if the student did not choose at random, but chose “A” for all answers? Does this change any of the steps 1-6 above?

An even maths-ier way

There is another way of "seeing" the answer, which is even simpler, if you enjoy a more abstract way of looking at things (but don't worry if you don't!)

Note that we asked what strategy gives a higher passing probability, not what that passing probability is! So, you can actually answer that question without figuring out the probability itself. All you need is a tiny bit of set theory.

- Imagine the set of all possible quizzes with 20 questions and four possible answer choices. This set Q contains every ordered list of numbers between 1 and 4 that has exactly 20 elements.
- A student's guess of all answers for a particular quiz is an element of that set. Let's call it g. You can imagine a subset of Q "around" that guess, which contains all the quizzes that a student could pass using that particular guess g. (I.e. all other elements that are "similar enough" to g: they need to have 10 or more elements in common with g). Let's call that subset G. Now, G will be much smaller than Q - in fact, the probability of passing any quiz with the given guess is just the size of G divided by the size of Q.
- We could compute this probability, but we don't have to! All you need to ask yourself is, will the size of G depend on what particular g you choose? And in particular, if

$$g = (1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1)$$

will it be different from, say

$$g = (1, 4, 2, 2, 3, 2, 2, 4, 2, 4, 1, 3, 3, 3, 1, 3, 1, 4, 3, 1)?$$

Once you have answered this question, you are done.

Scenario 2: Same number of A, B, C, and D

Here is a twist: The instructors have made it so each of the answers "A", "B", "C", and "D" is chosen as the correct answer the same number of times (i.e. 5 times each).

We got a passing probability of ~0.014 if you choose your answers at random and 0 if you select "A" for every answer. You should be able to get to the same result by using one of the three approaches outlined for Scenario 1.

Originally created by MI Stefan in 2019, CC-BY-SA 3.0

Last update by DJ MacGregor in 2022