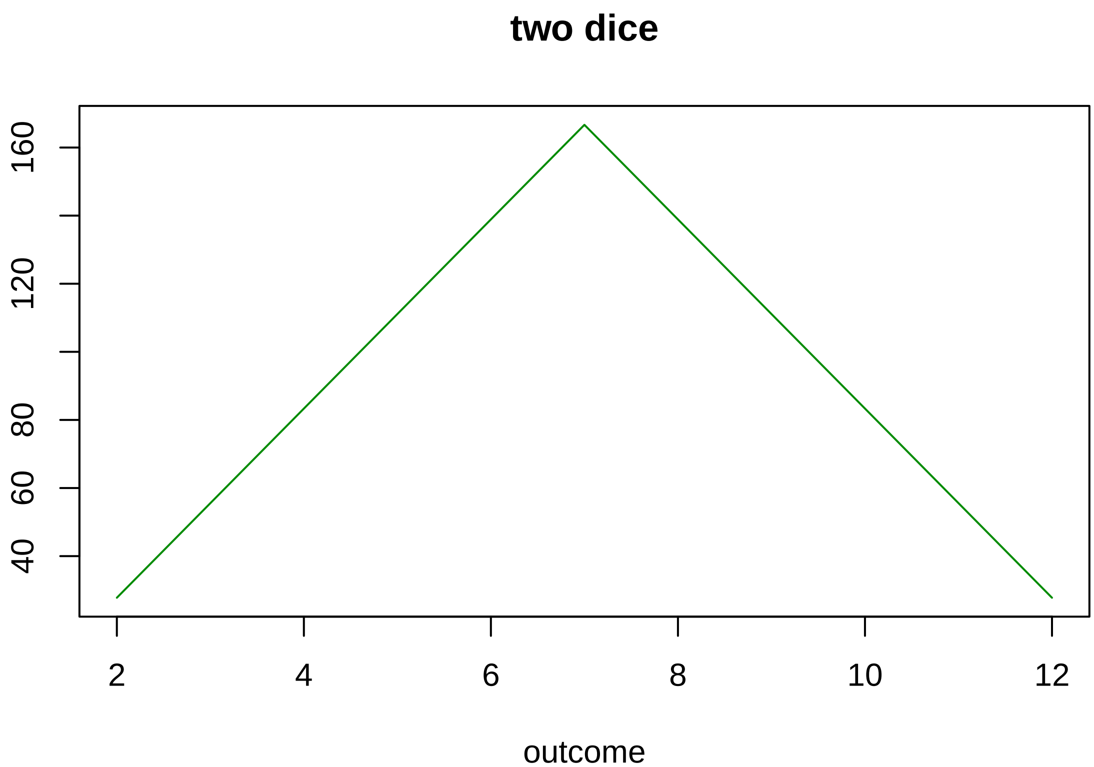
Problem Set 5: Notes

ADS2

Semester 1 2022/23

Rolling Dice

When rolling two dice, we get a distribution that does not look "normal", but rather like a pyramid, going straight up to the mean/median/mode at 7 and then back down again.



To see why, we can just do the maths. If we distinguish between the first and second dice, then there are $6 \times 6 = 36$ possible outcomes. Let's look at the probability of individual outcomes.

The probability of throwing a 2 is $\frac{1}{36}$, because there is exactly one way to do this: both dice must be 1s. Similarly, the probability of throwing a 12 is also $\frac{1}{36}$.

In order to throw a 3, we need one 1 and one 2 - there are exactly two ways to do this (1 on the first dice, 2 on the second, or the other way around), so the probability is $\frac{2}{36}$ - and this is of course the same for the probability of an 11.

You can go through all the numbers and convince yourself that for i where $2 \leq i \leq 7$, the probability $P(i)$ of throwing i with two dice is:

$$P(i) = P(14 - i) = \frac{i-1}{36}$$

We are not going to go through them all here, just a quick "sanity check" to show we are right: How do you get a 7? Basically, there are 6 ways. Why? Because the first dice can land on anything, and you will be able to make a 7 if the second lands just right.

This gives exactly the probability distribution we saw.

How do we see this is "not normal"? This is actually hard (if not impossible) to see. If you look at the histogram, it does have some features of the normal distribution: The distribution is symmetric and the mean, median, and mode are the same. Whether or not it is "bell shaped" can be hard to tell.

There is another thing we can check for normal distributions - remember the 68-95-99.7 rule! We rolled 2 dice 1000 times and got the following for the mean and standard deviation:

```
## [1] "Mean: 7.05"
```

```
## [1] "Standard deviation: 2.45"
```

What percentage of data points are within 1 standard deviation of that mean?

```
minus1sd = mean(dicerolls)-sd(dicerolls)
plus1sd = mean(dicerolls)+sd(dicerolls)
sum(dicerolls>minus1sd & dicerolls<plus1sd)/10
```

```
## [1] 65.7
```

Depending on the actual sample, this is sometimes not very far away from 68 at all (sometimes a bit further away).

What about 2 standard deviations?

```
## [1] 94.2
```

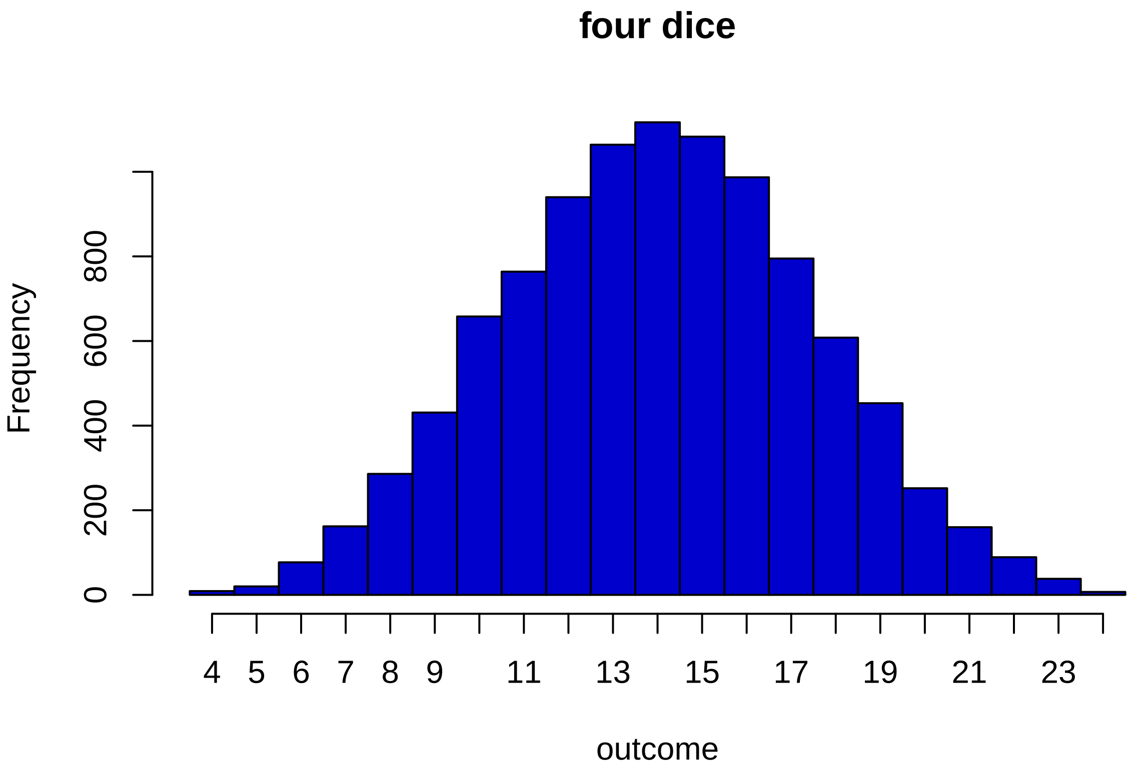
Again, not very far from 95!

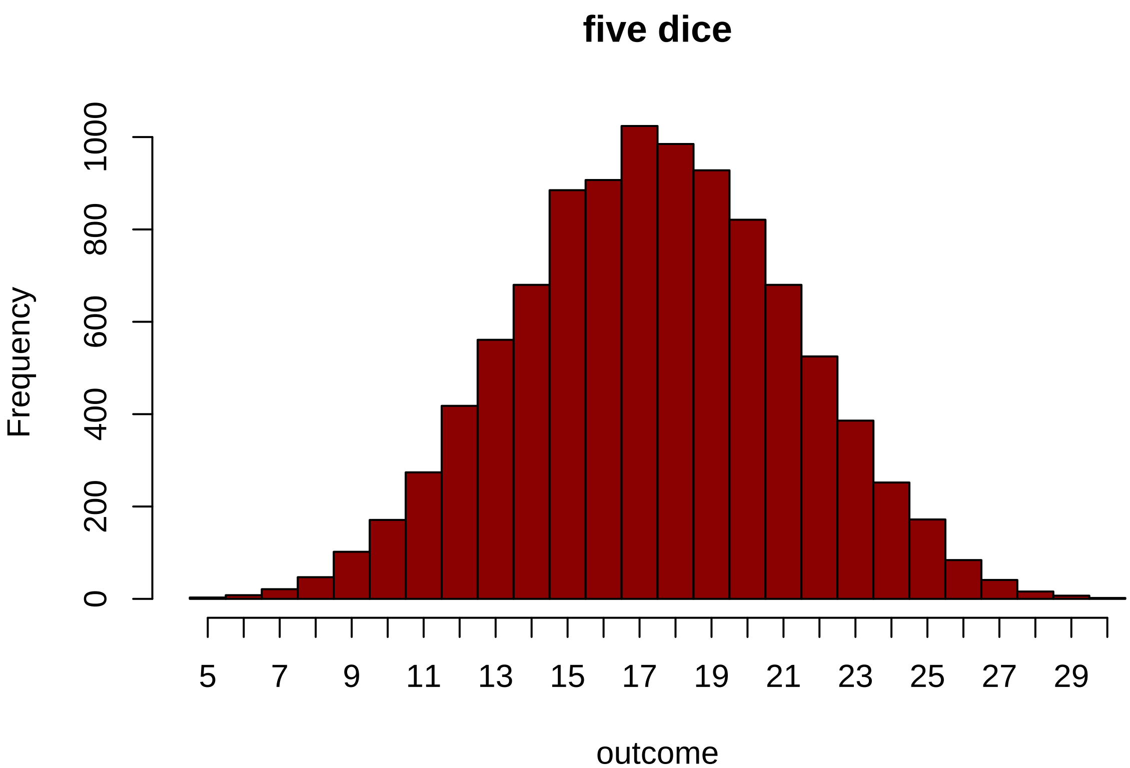
Three standard deviations:

```
## [1] 100
```

So, even on the 68-95-99.7 front, this is not too bad - if we saw a similar sample "in the wild", it would be easy to mistake it for a normally distributed sample. (Part of the reason here is that there are not that many different values that the sample can take)

So, just for fun, let's look at 3, 4, and 5 dice and see what we get. (We rolled each of them ten thousand times, just to get a bigger sample, so it looks nicer)



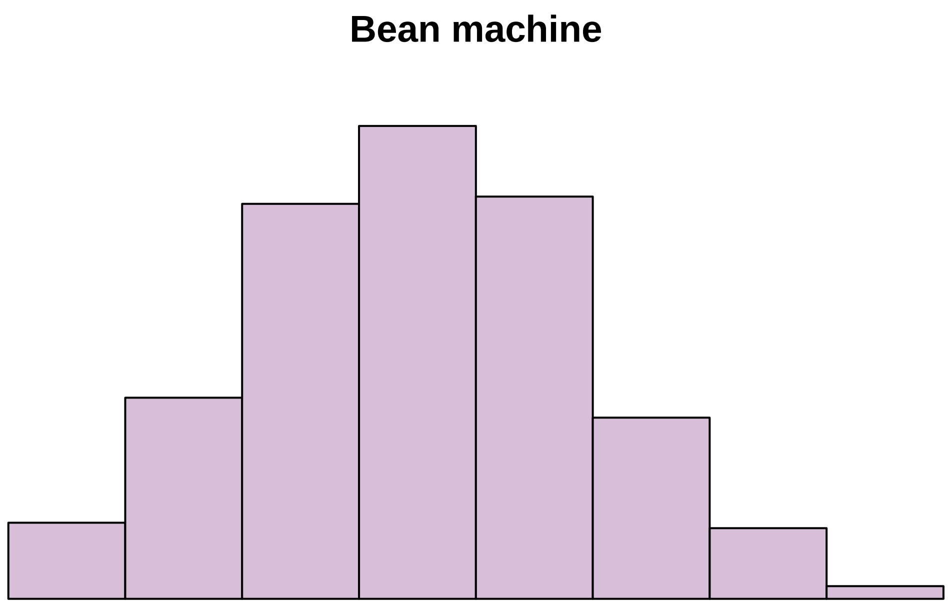


These distributions are definitely starting to look more bell-shaped as we add more dice. If you need more convincing, you can try and check the 68-95-99.7 rule again!

Bean machines

For the bean machine, each layer can be modelled as a binary choice: 0 means the ball goes left, 1 means it goes right. (You may instead have chosen to sample from {-1, +1}, that's also fine. We are interested in relative end positions, not absolute numbers - this is why we chose not to label the x axis). Since we have 8 layers of pegs, we do this 8 times, and this determines the end position.

```
ends <- c()
for (i in 1:1000) {
  end <- sum(sample(c(0,1),8,replace=T))
  ends <- c(ends, end)
}
hist(ends, main="Bean machine", col="thistle", xlab="", axes=F, ylab="")
```



Again, what we are doing (very fundamentally) is taking 8 random samples from a given distribution (this time, it's a binary distribution) and summing them. And that means, again, we are in Central Limit Theorem territory, and this will look normal if our sample is large enough.

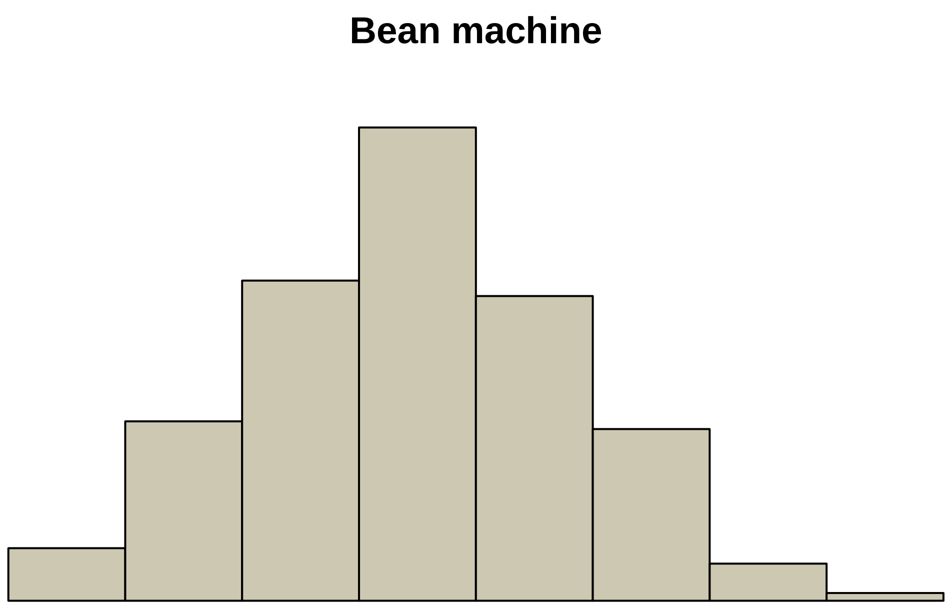
Remember that the Central Limit Theorem holds for distributions in general. So, even if the pegs were asymmetrical, with an 80% probabiltiy of going left and only a 20% probability of going right, the distribution of sums (i.e. end points) would still end up looking normal, as long as we have a large enough sample size (in this case, a large enough number of layers of pegs).

If you want to try this, the thing to keep in mind is that you can tell the "sample" function the probability with which it should sample each choice. (If not specified, it will just give each choice equal probabilities).

```
sample(c(0,1),8,replace=T,p=c(0.8,0.2))
```

```
## [1] 0 0 0 0 0 1 1 0
```

We did it and got something like this:



Class grades

For simplicity, let's just look at multiple-choice quizzes with four answer choices per question, where exactly one choice is correct and where all questions are weighted equally.

Of course, a test score is just a sum of individual question scores, and we can model each question score as a binary random variable (the two possible outcomes being "correct" and "incorrect").

Note that in order to do this, we need not assume that the underlying process is truly "random". If it is (i.e. the student is just guessing), then we would have $P(incorrect) = 0.75$ and $P(correct) = 0.25$. But even if the student is not guessing, we can still model this as a random variable, just (hopefully) with a bigger $P(correct)$.

Let's assume that this $P(correct)$ is the same for all questions and all students. Since we are computing the overall quiz score by summing the number of correct answers, we can again use the Central Limit Theorem - as long as there are enough questions, overall quiz scores will be approximately normally distributed.

But what if $P(correct)$ is not the same for all qusetions and all students?

Then it depends. If questions are just of different levels of difficulty, then you can still model the quiz as a sum of binary random processes and just use the average $P(correct)$. You are only interested in the sum (overall score) anyways. (Basically if a test contains one very easy question and one very hard question, the overall score will be the same as if they were just two medium questions instead.)

If different students have different probabilities of answering a question correctly, then this averaging of $P(correct)$ will not work anymore, and we may indeed end up with a distribution that is not normal.

Last update by DJ MacGregor in 2022