# T-test & t distribution

You will need to have the file femaleMiceWeights.csv in your working directory.

1) The CLT is a result from probability theory. Much of probability theory was originally inspired by gambling. This theory is still used in practice by casinos. For example, they can estimate how many people need to play slots for there to be a 99.9999% probability of earning enough money to cover expenses. Let's try a simple example related to gambling.
   Suppose we are interested in the proportion of times we see a 6 when rolling n=100 die. This is a random variable which we can simulate with `x=sample(1:6, n, replace=TRUE)` and the proportion we are interested in can be expressed as an average: `mean(x==6)`. Because the die rolls are independent, the CLT applies.
   We want to roll n dice 10,000 times and keep these proportions. This random variable (proportion of 6s) has mean p=1/6 and variance p*(1-p)/n. So according to CLT z = (mean(x==6) - p) / sqrt(p*(1-p)/n) should be normal with mean 0 and SD 1. Set the seed to 1, then use replicate to perform the simulation, and report what proportion of times z was larger than 2 in absolute value (CLT says it should be about 0.05).
2) For the last simulation you can make a qqplot to confirm the normal approximation. Now, the CLT is an asympototic result, meaning it is closer and closer to being a perfect approximation as the sample size increases. In practice, however, we need to decide if it is appropriate for actual sample sizes. Is 10 enough? 15? 30?
   In the example used in exercise 1, the original data is binary (either 6 or not). In this case, the success probability also affects the appropriateness of the CLT. With very low probabilities, we need larger sample sizes for the CLT to "kick in".
   Run the simulation from exercise 1, but for different values of p and n. For which of the following is the normal approximation best?
   a) p=0.5 and n=5
   b) p=0.5 and n=30
   c) p=0.01 and n=30
   d) p=0.01 and n=100

3) As we have already seen, the CLT also applies to averages of quantitative data. A major difference with binary data, for which we know the variance is p(1−p), is that with quantitative data we need to estimate the population standard deviation.

   In several previous exercises we have illustrated statistical concepts with the unrealistic situation of having access to the entire population. In practice, we do not have access to entire populations. Instead, we obtain one random sample and need to reach conclusions analyzing that data. dat is an example of a typical simple dataset representing just one sample. We have 12 measurements for each of two populations:

```
X <- filter(dat, Diet=="chow") %>% select(Bodyweight) %>% unlist
Y <- filter(dat, Diet=="hf") %>% select(Bodyweight) %>% unlist
```

   We think of X as a random sample from the population of all mice in the control diet and Y as a random sample from the population of all mice in the high fat diet.

   Define the parameter $\mu_x$ as the average of the control population. We estimate this parameter with the sample average $\bar{X}$. What is the sample average?

4) We don't know $\mu_x$ , but want to use $\bar{X}$ to understand $\mu_x$. Which of the following uses CLT to understand how well $\bar{X}$ approximates $\mu_x$ ?

   a) $\bar{X}$ follows a normal distribution with mean 0 and standard deviation 1.

   b) $\mu_x$ follows a normal distribution with mean $\bar{X}$ and standard deviation $\frac{\sigma_x}{\sqrt{12}}$ where $\sigma_x$ is the population standard deviation.

   c) $\bar{X}$ follows a normal distribution with mean $\mu_x$ and standard deviation where $\sigma_x$ is the population standard deviation.

   d) $\bar{X}$ follows a normal distribution with mean $\mu_x$ and standard deviation $\frac{\sigma_x}{\sqrt{12}}$ where $\sigma_x$ is the population standard deviation.

5) The result above tells us the distribution of the following random variable: $Z = \sqrt{12}\frac{\bar{X} - \mu_x}{\sigma_X}$ . What does the CLT tell us is the mean of Z (you don't need code)?

6) what is your estimate of population standard deviation $\sigma_x$

7) Use the CLT to approximate the probability that our estimate $\bar{X}$ is off by more than 2 grams from $\mu_x$.

8) Now we introduce the concept of a null hypothesis. We don't know $\mu_x$ nor $\mu_y$. We want to quantify what the data says about the possibility that the diet has no effect $\mu_x = \mu_y$.

If we use CLT, then we approximate the distribution of $\overline{X}$ as normal with mean $\mu_x$ and standard deviation $\sigma_x$, and the distribution of $\overline{Y}$ as normal with mean $\mu_y$ and standard deviation $\sigma_y$. This implies that the difference $\overline{Y}$ - $\overline{X}$ has mean 0.

We described the standard deviation of this statistic (the standard error) is $SE(\overline{X} - \overline{Y}) = \sqrt{\frac{\sigma_y^2}{12} + \frac{\sigma_x^2}{12}}$ and that we estimate the population standard deviations $\sigma_x$ and $\sigma_y$ with the sample estimates. What is the estimate of $SE(\overline{X} - \overline{Y}) = \sqrt{\frac{\sigma_y^2}{12} + \frac{\sigma_x^2}{12}}$?

9) So now we can compute $\overline{Y} - \overline{X}$ as well as an estimate of this standard error and construct a t-statistic. What number is this t-statistic?

10) If we apply the CLT, what is the distribution of this t-statistic?

11) Now we are ready to compute a p-value using the CLT. What is the probability of observing a quantity as large as what we computed in 9, when the null distribution is true?

12) CLT provides an approximation for cases in which the sample size is large. In practice, we can't check the assumption because we only get to see 1 outcome (which you computed above). As a result, if this approximation is off, so is our p-value. As described earlier, there is another approach that does not require a large sample size, but rather that the distribution of the population is approximately normal. We don't get to see this distribution so it is again an assumption, although we can look at the distribution of the sample with qqnorm(X) and qqnorm(Y). If we are willing to assume this, then it follows that the t-statistic follows t-distribution. What is the p-value under the t-distribution approximation? Hint: use the t.test function.

13) With the CLT distribution, we obtained a p-value smaller than 0.05 and with the t-distribution, one that is larger. They can't both be right. What best describes the difference?

   a) A sample size of 12 is not large enough, so we have to use the t-distribution approximation.