Challenge 1: What is the difference between expected value and mean?

Mean

The mean is the average of the values (outcomes) of a process, or the central tendency of a collection of numbers.

Expected Value

The expected value is the average value of a random variable over a large number of experiments.

There are three types of expected outcomes:

1. Neutral: neither wins nor losses are expected in the long run.
   a. *Ex: You flip the fair coin. Every time you get heads, you lose $1, and every time you get tails, you gain $1.*
2. Positive: the player is expected to gain from playing this game.
   a. *Ex: You flip the fair coin. Every time you get heads, you lose $1, and every time you get tails, you gain $2.*
3. Negative: the player is expected to lose from playing this game.
   a. *Ex: You flip the fair coin. Every time you get heads, you lose $1, and every time you get tails, you gain $1. Additionally, there is a $0.01 fee for every flip regardless of the outcome.*

Difference

The difference between the two in an example would be that the mean is the average value of the outcome (heads or tails), and the expected value is the corresponding return from a multitude of outcomes ($1,5).

The expected value is thus useful in decision making and can be used to measure quantifiable potential returns.

Challenge 2: What is the "problem" in science with p-values?

The problem of ignoring differences between samples because the differences are negligible; as well as the fact that statistical analysis cannot prove nor disprove a theory.

Some definitions before we start:

- The **P value**, or calculated probability, is the probability of finding the observed, or more extreme, results when the null hypothesis (H 0) of a study question is true – the definition of 'extreme' depends on how the hypothesis is being tested.
  - In other words, the likelihood of encountering outliers.
  - The key value for values is $P < 0.05$ (making it statistically significant) and $P > 0.05$ (making it statistically non-significant).

Reliance on thresholds of statistical significance (p-values) can mislead us. We should exert special caution when comparing studies with similar outcomes but different statistical significance. Moreover, a statistical non-significance should not be interpreted as "no difference" or "no effect", since these are wildly different concepts (which according to an analysis of "791 articles across 5 journals found that around half mistakenly assume non-significance means no effect).

The P-value should hence not be used to decide whether a result refutes or supports a scientific hypothesis.

In addition, another two problems are faced when dealing with the improper use of "statistical significance" parameters. Firstly, once a conclusion is made the results are biased in the direction (positive or negative) of the conclusion. Secondly, the necessity of crossing the renown P<0.05 threshold can lead to biased research methods than can yield more attractive results whilst ignoring the veracity of the issue.

In order to avoid obtaining misleading results, the point estimate should be interpreted whilst acknowledging its uncertainty, and without declaring it makes "no difference".

A better practice would include suggesting reasons for the results, but whilst discussing a range of other potential explanations. Factors such as background evidence, study design, data quality and understanding of underlying mechanisms are often more important than statistical measures such as P values or intervals. Alternatively, the statistician could even consider adapting the p-value threshold to the given circumstances of the statistical analysis.

I've included a concise quote on what should be the practice undertaken regarding statistical conclusions, for future reference:

"What will retiring statistical significance look like? We hope that methods sections and data tabulation will be more detailed and nuanced. Authors will emphasize their estimates and the uncertainty in them — for example, by explicitly discussing the lower and upper limits of their intervals. They will not rely on significance tests. When P values are reported, they will be given with sensible precision (for example, P = 0.021 or P = 0.13) — without adornments such as stars or letters to denote statistical significance and not as binary inequalities (P < 0.05 or P > 0.05). Decisions to interpret or to publish results will not be based on statistical thresholds. People will spend less time with statistical software, and more time thinking."

In conclusion, inferences should be scientific, and that goes far beyond the merely statistical. Therefore, statisticians should refrain from confusing correlation with causality and making final conclusions on the effects of the factors they study.

Challenge 3: Applying testing to a specific case: A/B testing.

I decided to apply A/B testing to a specific case of Netflix.

Instead of focusing on the A/B test that they did comparing different images, I'm going to attempt to implement a new visualization mode: a short video teaser/gif of the trailer once the user hovers over the movie title for over 2 seconds.

I'm thus going to use A/B testing to compare the original single image format with the video format, and then compare several different videos to see if they would perform better.

In order to do so, I will follow the following steps:

For starters, I will use a basic 6 group division where 1/6 of the users in my sample will be shown just one image (the one that was already tested for best performance) and the other 5/6 one video from a selection of 5 choices . Since I0m dividing my sample in 6 equal parts, the sample size needs to be significantly big in order for the results to be reliable.

This method will yield the best results since I'm interested in whether a video can outperform our best image (hence why only the best imagine is shown), and at the same time, since I have no way of knowing which would be the best performing video, I can offer a selection and properly compare the best performing video's conversion rates with the image's conversion's rates.

If the best performing video yields better results that the best image, then swapping to video should be considered on a big scale. If it does not then the old model would remain active.

In the case of the former, a multitude of new A/B test would need to be performed to determine what video performs best for each title, just like they did in the case of the images.

Moreover, I would also consider testing different ways of executing the videos (after more or less time of hovering over a title, of longer or shorter duration, with sound or without, etc).