**1. Choosing the right statistical test**

We'll now look into choosing the right statistical test for analyzing experimental data.

**2. Selecting the right test**

Just as choosing the right book or the right measurement tool for is vital to research, choosing the right statistical test is foundational to any data analysis. Understanding our dataset's features and the hypotheses under examination is vital. It necessitates assessing the data types—categorical or continuous—their distributions, often assumed to be normal by many statistical tests, and the number of variables in the study. It's essential to align the chosen statistical method with the dataset's properties and the study's goals to ensure accurate and dependable outcomes. In this video, we'll explore how to apply t-tests, ANOVA, and Chi-Square tests, focusing on analyzing experimental data.

**3. The dataset: athletic performance**

We'll work with a DataFrame called athletic\_perf containing athletes' performance data, focusing on the effects of different training programs and diets on athletic performance. Key variables are the type of training program, assigned diet, initial fitness level, and the observed performance increase as a percentage.

**4. Independent samples t-test**

An independent samples t-test is used to compare the means of two distinct groups to determine if there is a statistically significant difference between them. This test relies on the assumptions that the response data for both groups are normally distributed and have equal variances, ensuring the validity and reliability of the test results. We'll use an alpha of 0.5 and compare the mean athletic performance improvements between two groups undergoing High-Intensity Interval Training (HIIT) and Endurance training by assigning their performance increases to group1 and group2. Next we call ttest\_ind on group1 and group2 and retrieve the test statistics and p-value. A large p-value here leads us to conclude that there is no significant difference in the mean performance increase between the HIIT and Endurance groups.

**5. One-way ANOVA**

A one-way ANOVA test is employed to determine if there are statistically significant differences among the means of more than two groups. The one-way corresponds to ANOVA with a single independent variable, and it assumes that the variances among the groups are equal. For our example, we gather the athletic performance increase data for each training program type into a list of groups using a list comprehension. The f\_oneway function from scipy.stats is then used to conduct the ANOVA test across these groups by unpacking the groups list using an asterisk. The relatively high P-value implies that, based on the provided data, we cannot confidently assert that different training programs lead to different mean increases in athletic performance.

**6. Chi-square test of association**

The Chi-square test of association is a statistical method used to assess whether there is a significant association between two categorical variables. Unlike many other statistical tests, the chi-square test does not require assumptions about the distribution of the data. To prepare for the test, we start by creating a contingency table using crosstab from pandas, which cross-tabulates athletes by their Training\_Program and Diet\_Type.

**7. Chi-square test of association**

The chi2\_contingency function from scipy.stats is then employed to conduct the chi-square test on the contingency table. The large P-value suggests that any observed association between training programs and diet types is not statistically significant.

**Post-hoc analysis following ANOVA**

After conducting ANOVA, we often need to understand specific differences between groups. This is where post-hoc analysis comes in, providing detailed insights into pairwise comparisons.

**2. When to use post-hoc tests**

Post-hoc tests are pivotal when ANOVA reveals significant differences among groups. They allow us to pinpoint which specific pairs of groups differ, allowing us to peek behind the curtain to explore the inner workings of pairwise differences.

**3. Key post-hoc methods**

There are two common post-hoc methods: Tukey's HSD, named after statistician John Tukey, which is known for its robustness in multiple comparisons. There's also the Bonferroni correction, named after mathematician Carlo Bonferroni, which adjusts p-values to control for Type I errors. For broader comparisons, use Tukey's HSD; Bonferroni is better for reducing false positives in more focused tests.

1. 1 https://www.amphilsoc.org/item-detail/photograph-john-wilder-tukey
2. 2 https://en.wikipedia.org/wiki/Carlo\_Emilio\_Bonferroni

**4. The dataset: marketing ad campaigns**

We'll work with a dataset of marketing campaigns, examining the Click\_Through\_Rate for different Ad campaigns to identify differences and which strategy is most effective.

**5. Data organization with pivot tables**

Pivot tables in pandas can be extremely helpful for organizing data, especially before conducting post-hoc analysis. It provides a clear comparison of the mean Click\_Through\_Rates for each campaign type.

**6. Performing ANOVA**

We start with ANOVA to assess if there's a significant difference in these Click\_Through\_Rates among the campaigns. This sets the stage for further analysis if significant differences are found. First, we specify the different campaign types. Then we create the groups using a list comprehension to extract the Click\_Through\_Rate for each Ad\_Campaign. Next, we perform the ANOVA across the three campaign types, unpacking the groups using an asterisk, to compare their mean click-through rates. The very small P-value here indicates significant differences in these means.

**7. Tukey's HSD test**

If ANOVA indicates significant differences, Tukey's HSD test helps us understand exactly which campaigns differ. The pairwise\_tukeyhsd function from statsmodels.stats takes arguments for the continuous response variable, Click\_Through\_Rate in this case, the categorical variable with more than two groups, Ad\_Campaign, and alpha. To interpret the results of this table, we focus on the meandiff, p-adj (adjusted P-value), and reject columns. For the first row, Loyalty Reward versus New Arrival, the mean difference is 0.2211, with a p-value less than 0.05, indicating that the Loyalty Reward group has a significantly higher mean than the New Arrival group. For Loyalty Reward versus Seasonal Discount, on row 2, the mean difference is -0.2738. With a p-value less than 0.05, it suggests that the Loyalty Reward group has a significantly lower mean than the Seasonal Discount group. Lastly, for New Arrival versus Seasonal Discount, the mean difference is -0.4949, with a p-value less than 0.05, indicating that the New Arrival group has a significantly lower mean than the Seasonal Discount group.

**8. Bonferroni correction set-up**

The Bonferroni correction is a stringent method to adjust p-values when conducting multiple pairwise comparisons, effectively reducing the chances of a Type I error. A little more data preparation is required before applying the Bonferroni correction. We begin by creating an empty P-values list. Then, we lay out a list of tuples containing the pairwise comparisons that we will iterate over. Next, we iterate over the tuples in comparisons, using the tuple elements to extract the Click\_Through\_Rate for both groups. We run ttest\_ind on the click through rates in a pairwise fashion, and append the p-values to our list.

**9. Performing Bonferroni correction**

Now we apply the Bonferroni correction using the multipletests function. The resulting p-values for the three comparisons are all extremely small. This again provides evidence that each of the three groups have significant click through rate differences.

**P-values, alpha, and errors**

In this video, we'll deepen our understanding of p-values, alpha levels, and experimental errors. This will prepare us for the next video, where we'll tackle a key concept in experimental design called power analysis!

**2. P-values and alpha**

P-values and alpha can be viewed as a game. Think of conducting a scientific experiment where we are trying to determine whether a certain strategy (our hypothesis) leads to winning (or a significant result) more often than just by chance. P-values help us understand the likelihood of observing our data if the null hypothesis was true. That is they serve as the scoreboard of the game. Setting an alpha level, often 0.05, allows us to determine the threshold at which we consider our results statistically significant, akin to setting the rules of a game before playing. Alpha is like establishing a rule for what counts as a "remarkable" win in this game. If your P-value is below this alpha level, it's as if we've achieved a high score or a remarkable performance in the game, leading us to conclude that our strategy (the alternative hypothesis) might indeed be effective, and it's not just the luck of the draw.

**3. The dataset: crop yields**

We'll work with a dataset of crop yields from different fields, where each field was treated with either organic or synthetic fertilizer. Our goal is to analyze this data to determine if there's a significant difference in crop yields between the two fertilizer types.

**4. Visualizing the data**

It's helpful to visualize the crop yields for each fertilizer type. By plotting the kernel density estimates (kde), we get a sense of how the two fertilizers might differ in terms of their effect on crop yields and whether there's an overlap between their effects. It appears that Organic tends to produce a higher yield than Synthetic with some overlap.

**5. Conducting an independent samples t-test**

We set our alpha to the standard five-percent level. To compare the effectiveness of organic versus synthetic fertilizers, we perform a t-test on the crop yields from the two groups. The p-value is smaller than alpha suggesting that fertilizer type has a statistically significant impact on crop yield.

**6. Exploring experimental errors**

In experimental design, we encounter two main types of errors. Type I errors occur when we incorrectly reject a true null hypothesis, akin to a false alarm. Type II errors happen when we fail to reject a false null hypothesis, similar to a missed detection.

**7. More on alpha**

Alpha, or the significance level, is crucial in hypothesis testing; it indicates the probability of a Type I error—rejecting a true null hypothesis. Common alpha levels include 0.05, 0.01, and 0.10, representing risks of 5%, 1%, and 10%, respectively, for such errors. Selecting an alpha hinges on the study's context and a balance between tolerating a Type I error and the risk of overlooking a true effect, known as a Type II error. The choice should align with the study's goals and the implications of potential errors. Conventionally, 0.05 is the standard for statistical significance across many disciplines. For more rigorous scrutiny, particularly where the cost of a Type I error is high, an alpha of 0.01 is preferred. In preliminary studies, where a higher error tolerance is permissible, an alpha of 0.10 may be utilized, allowing for a broader exploration of potential effects with subsequent validation through more stringent testing.

**Power analysis: sample and effect size**

We now dive into the intricacies of power analysis, focusing on understanding effect size and how it influences sample size.

**2. A primer on effect size**

Effect size quantifies the magnitude of the difference between groups, beyond just noting if the difference is statistically significant. Cohen's d is a commonly used measure, calculated as the difference in means divided by a pooled standard deviation.

**3. The dataset: video game engagement**

A video game company conducted an experiment with sixty participants to understand player engagement across two game genres: Action and Puzzle. They recorded the average number of hours players spent engaged to assess which type tends to captivate players more effectively.

**4. Calculating power overview**

Power analysis revolves around the probability that our test will correctly reject a false null hypothesis. This corresponds to identifying a true effect, avoiding a Type II error. A type II error is denoted as beta, so power is one minus beta and it ranges from zero to one, where one is certainty in our ability to detect a true effect. To calculate power, we first assume an effect size. Here we choose a value of 1, derived from historical data comparing the engagement scores of video game genres. We can also use our sample data to make an estimate of the effect size, but traditionally power analysis is done prior to the data collection. This can also help us determine how big of a sample size we should use in our study. We initialize the power object and call the .solve\_power() method, using a sample size for how many video game players were assessed in either group (30), our assumed effect size, and our alpha of 0.05. This high power tells us the likelihood that our test will detect a significant result, given our effect size and sample size.

**5. Cohen's d formulation**

To calculate Cohen's d as an effect size, we define a function. Its two inputs are numeric data corresponding to the two groups from our sample data. We calculate the difference in the means of the two groups, their sample sizes, and their variances. Next, we determine a pooled standard deviation using this information. Lastly, Cohen's d is the difference in means divided by the pooled standard deviation.

**6. Cohen's d for video game data**

To apply this to the video game data, we first split the data into two groups based on the genre. Then we apply our function to get the effect size. The result here is near the theoretical result of 1 we assumed earlier.

**7. Understanding sample size and power**

Balancing the need for sufficient power with practical constraints on sample size is a fundamental aspect of planning a study, such as comparing engagement times across different video game genres. A larger sample size can enhance an experiment's power, improving the likelihood of detecting a true effect.

1. 1 https://grabngoinfo.com/power-analysis-for-sample-size-using-python/

**8. Sample size calculation in context**

Let's contextualize this within our video game study. Assuming our calculated value for Cohen's d engagement time between game genres, we calculate the sample size needed for each group to achieve 99% power with an alpha of 0.05 and equally-sized groups with a ratio of 1. This calculation is pivotal in ensuring our study is adequately powered to detect meaningful differences in player engagement across genres. Assuming we have an effect size of around 1.2, we would need at least 28 participants in each group to achieve a power of 99%. Recall we collected 30 participants, so we can feel confident about our experiment's power.

**9. Visualizing sample size requirements**

We next build a visualization illustrating the relationship between effect size measured as Cohen's d and required sample size for our video game study, by plotting varying effect sizes against required sample sizes.

**10. Visualizing sample size requirements**

As effect size increases, the required sample size for each group decreases, highlighting the importance of understanding the expected magnitude of differences when planning a study.