Data Analysis for Business Analytics

CIS621

Week 6 - Statistical Significance

Lecture Goals

1. Understand the concept of statistical significance.
2. Understand the benefits of using statistical significance.
3. Understand the pitfalls of statistical significance and how they relate to broader issues in the research community.
4. Understand the practical application of statistical significance.

This lecture will revisit and expand key ideas from last week.

Statistical Significance

- Statistical significance answers the following question: How likely would we have gotten our outcomes by chance?
- It is nothing more or less than the above.
- Statistical significance does not equate to *importance*. A relationship can be statistically significant but not interesting or important in the “real world”.
- Statistical significance only measures if a difference in outcomes is likely “real”, not how large the effect is.

Statistical Significance and the Normal Distribution

- Classically, statistical significance assumes that the outcomes we are comparing fall on a normal distribution.
- That is, the actual mean outcome we reported is *one of many possibilities* drawn from a normal distribution.
- By constructing multiple normal distributions, we can see how often they would overlap if the experiment were re-run under the same conditions.
- We can construct our normal distribution from our actual data and from the empirical rule (otherwise known as the 68–95–99.7 rule).
- Likewise, we can have faith that the outcomes follow a normal distribution thanks to the central limit theorem (As we take more samples, our distributions of means approaches a normal distribution, regardless of distribution of the population).

Further Statistical Significance Assumptions

- Statistical significance assumes that compared groups are independent.
- Statistical significance - and more broadly statistical analysis - assumes that the data in each groups are representative of the broader population (often achieved, though not always perfectly, through randomization).
- Further, statistical significance assumes the compared grouped have equal spread.

Comparison Groups Example

- Invalid groups to compare:
  - The treatment group who received coupons was selected from people who volunteered their email address.
  - The control group who did not receive coupons was comprised of everyone else - people who did not volunteer their email address.
- Valid groups to compare:
  - The treatment group who received coupons was comprised of a random 10,000 people who volunteered their email.
  - The control group who did not receive coupons was comprised of a mutually exclusive random 10,000 people who volunteered their email.

Null Hypothesis vs. Alternative Hypothesis

- Every time we run a statistical test, we have a null hypothesis and an alternative hypothesis.
- Null hypothesis: The status quo (e.g., the outcomes are not different).
- Alternative hypothesis: The results are different (either they are different or one is expressly larger).
- If the difference in outcomes is statistically significant, we “reject the null and accept the alternative”.
- If the difference in outcomes is not statistically significant, we “fail to reject the null hypothesis”.
- Notice our default is the status quo.

P-Values

- How do we determine if outcomes are different in terms of statistical significance? The p-value, of course!
- The p-value is the probability of experiencing an outcome *at least as extreme as the result actually observed* when the null hypothesis is actually correct (language influenced by this Wikipedia article).
- In other words, the p-value is the probability you report a difference in outcomes when no true difference exists.

P-Values Example

- Stripped-Down Example Scenario:
    - People who ate a certain cereal had a 0.5% chance of having a heart attack.
    - People who did not eat that cereal had a 0.6% chance of having a heart attack.
    - We run a test and find the results have a statistically significant difference.
    - Our p-value from the test is 0.04.
    - We can state that “there is a 4% probability that reporting that the two groups have different outcomes is incorrect and that the two outcomes are actually not different”.

Common P-Value Cutoffs

- How confident do we want to be in reporting the “right” outcome?
- In other words, what tolerance do we have for reporting there is a statistically significant difference when there is *not* a statistically significant difference?
- Common p-value cutoffs are 0.10, 0.05, and 0.01. Of these three, the most common is 0.05.
    - “There is a 5% chance we report a statistically significant difference when one actually does not exist”.
- Academically, the p-value cutoff is referred to as alpha.

Type I and Type II Errors

- A Type I error is a false positive.
  - We say a difference exists when one does not.
- A Type II error is a false negative.
  - We say no difference exists when one does.

How should you set alpha? Based on the costliness of each type of error!

Group Discussion

- Depending on the situation, we may weight false positives and false negatives differently.
- In what situations would we want to minimize false positives?
- In what situations would we want to minimize false negatives?

Two-Sample T-Test

- The two-sample t-test assess if two samples means differ from one another.
- Example: Is the mean sale price of houses in Kansas City different from the mean sale price of houses in St. Louis?
- Remember, this entire comparison is built on the idea that outcomes fall on a normal distribution.
- If the mean in Kansas City is $200K and the mean in St. Louis is $210, those values are not fixed in stone. Under a different sample, we *might* see Kansas City with a mean of $210K and St. Louis with a mean of $200K! Statistical significance, more or less, compares how likely the two normal distributions are to overlap.

One-Sample T-Test

- The one-sample t-test assess if a calculated mean is different from a hypothetical value.
- Example: Is the mean sale price of this sample of houses different from 90K?
- The logic is very similar to the two-sample t-test. More or less, instead of comparing two actual sample distributions we are comparing a sample distribution against a hypothesized one.

Technical Note: Z-Test vs. T-Test

- There are two many tests to determine if means are different in a statistically significant way:
  - The z-test is used when we know the population standard deviation *and* when the sample size is large (>= 30). It assumes data follows a normal distribution.
  - The t-test is used when the population standard deviation is not known or when we have a small size (< 30). It assumes a t-distribution, which is closely related to a normal distribution, but it has more kurtosis with small sample sizes.
- Since most of the time we do not know the population standard deviation we use the t-test.
  - “Population” means we have collected all possible data; we have do no sampling.
- Likewise, once we get above 30 samples, the t-distribution becomes basically identical to the normal distribution, so much so we can say they are the same.

T-Test vs. Z-Test Citations and Reading Materials

https://www.statology.org/normal-distribution-vs-t-distribution/

https://www.wallstreetmojo.com/z-test-vs-t-test/#h-differences-between-z-test-and-t-test

Application: A/B Testing

- A classic place we see statistical significance is in A/B testing.
- We randomly assign website users to see one of two versions of a webpage: one with a red button and one with a black button.
- We want to determine if the red button encourages more form submits than the black one.

Notes on A/B Testing in the Real World

- First, we have to assign participants randomly. (We *can* do non-random assignment, but we have to *control for* the non-random factors in our analysis).
- Second, for a true A/B test, we only change one element (the button is red in one version and black in another…nothing else is different).
- The foregoing makes it easier to determine what is driving a difference in behavior.
- If we change multiple items (e.g. a button color and a tagline), this is known as multivariate testing, and the driving factor is less clear. To tease out the driving factor, we have to let a statistical model assign attribution. This can work - but there is no guarantee the model will assign the proper attribution.

An Extension: Multi-Armed Bandit

- In the real world, an A/B test can be costly. If we assign individuals to a clearly inferior option, core business metrics will likely suffer.
- An alternative is a multi-armed bandit (named after slot machines).
- It will balance exploration with exploitation.
  - Exploration: trying “new things”
  - Exploitation: doing the “thing” we know works
- The classic algorithm to balance the explore-exploit tradeoff is called epsilon greedy.
  - To start off, the assignment will be 50/50. If a certain option underperforms, we will start decreasing the number of instances that get assigned to it, though it will still get a certain percentage under the explore construct.

Statistical significance in action with https://abtestguide.com/calc/

One- and Two-Sample T-Tests in Action with Python

Bonferroni Correction

- We set our p-value cutoff to be 0.05. We run 1,000 t-tests. What is going to happen?
- We are going to have ~50 false positives! This is known as the multiple-comparisons problem.
- To counteract this, we can use the Bonferroni Correction.
- This process sets the significance cut-off at the initial significance cutoff (called alpha) / number of tests (called n).
- For example, with 20 tests and α = 0.05, you would then reject a null hypothesis if the p-value is less than 0.0025.
- A more aggressive approach is the Benjamini–Hochberg procedure.

ANOVA: Comparing More than Two Means

- Similar to the T-Test, but used to compare the means among three or more groups.

Wilcoxon Rank Sum Test: Non-Parametric T-Test

- What are the assumptions of the t-test?
    - Data is independent
    - Groups have equal variance
    - Data is normally distributed
- The Wilcoxon Rank Sum Test drops the assumption of normality.
- When we have small samples, we want to use this test.
- Related Note: We can use a family of power transforms to make non-normal data normal. This is another option, though it still may not work well for small sample sizes. It might be a better idea when we have larger samples that are not quite normal.

Chi-Squared Test: Dealing with Categorical Dependence

- The Chi-Squared Test determines if one category depends on the other.
- For example, does purchasing a product depend on the channel used to acquire the customer?

Example: Linear Regression

- Estimates the relationship between a continuous target and a set of predictor variables.
- The model coefficients can be interpreted as how the target moves as we increase the predictor variable by one unit, holding all other variables constant. We can determine if a coefficient is statistically significant.

How Much Data Do We Need to Detect Statistical Significance?

- The above is a common question you might get from a business counterpart.
- The answer - we don’t know in advance!
- Detecting statistical significance is based on sample size (also comparatively among groups) *and* the effect size.
- We know that 1,000 samples is better than 100, but what we don’t know beforehand is the effect size - that’s why we are collecting data!
- The greater the difference across groups, the less data we need. The smaller the difference across groups, the more data we need.
  - In our sample, if group 1 has a mean of 90 and group 2 has a mean of 7, we don’t need as much data to determine those are different in terms of statistical significance. Remember our discussion on multi-armed bandits?
- Based on a hypothesized effect difference, we then could estimate the amount of data we need…but only then! (We’ll cover statistical power soon).

A Related Note: Why is More Data Better?

- Experiment 1:
  - Group 1 Conversions / Observations: 10 / 100 (10%)
  - Group 2 Conversions / Observations: 14 / 100 (14%)
  - Two-Sided P-Value: 0.3832
- Experiment 2:
  - Group 1 Conversions / Observations: 10 / 1000 (1%)
  - Group 2 Conversions / Observations: 14 / 1000 (1.4%)
  - Two-Sided P-Value: 0.4113
- In this case, more data does not make us more confident that Group 2 has a *different* conversion rate because the effect size gets lower as we simply add observations.

A Related Note: Why is More Data Better?

- Experiment 1:
  - Group 1 Conversions / Observations: 50 / 500 (10%)
  - Group 2 Conversions / Observations: 60 / 500 (12%)
  - Two-Sided P-Value: 0.3119
- Experiment 2:
  - Group 1 Conversions / Observations: 200 / 2000 (10%)
  - Group 2 Conversions / Observations: 240 / 2000 (12%)
  - Two-Sided P-Value: 0.0431
- In this case, more data makes us more confident that Group 2 has a *different* conversion rate than Group 1. This is because we can better define (i.e. narrow) the expected distributions as we get more data.
- In more technical terms, the variance of the sample mean (known as the standard error) decreases as we add data.

Let’s see this in action using the AB test site!

The Concept of Statistical Power

- Statistical Power: the probability of finding an effect if there is an effect to be found.
- Power Analysis: Estimates the minimum sample size for an experiment, given a significance level, effect size, and statistical power (default is 0.80).
- The higher the power, the lower the chance of producing a *false negative*. Remember, significance deals with *false positives*.

A Related Note: Why is More Data Better?

- Experiment 1:
  - Group 1 Conversions / Observations: 50 / 500 (10%)
  - Group 2 Conversions / Observations: 60 / 500 (12%)
  - Two-Sided P-Value: 0.3119
  - Statistical Power: 44.34%
- Experiment 2:
  - Group 1 Conversions / Observations: 200 / 2,000 (10%)
  - Group 2 Conversions / Observations: 240 / 2,000 (12%)
  - Two-Sided P-Value: 0.0431
  - Statistical Power: 82.72%
- Experiment 3:
  - Group 1 Conversions / Observations: 1,000 / 10,000 (10%)
  - Group 2 Conversions / Observations: 1,200 / 10,000 (12%)
  - Two-Sided P-Value: 0.0000
  - Statistical Power: 100%

Power and Significance

- Low power (50%) and low significance (alpha of 0.05)
    - There’s a good chance we might say there is no effect when there is an effect. However, if we say there is an effect, we are confident in that declaration. (This is often where we operate in the real world!)
- Low power (50%) and high significance (alpha of 0.10)
    - There’s a good chance we might say there is no effect when there is an effect. However, if we do detect an effect, we might be picking up on sheer luck.
- High power (90%) and low significance (alpha of 0.05)
    - There’s a good chance we will detect an effect if it exists. If we detect an effect, we are confident in that declaration. (This is where we want to operate!)
- High power (90%) and high significance (alpha of 0.10)
    - There’s a good chance we will detect an effect if it exists. But even if we detect an effect, we might be picking up on sheer luck - we have cast our net too wide!

Ideally, we want high power and high significance thresholds. We cannot always control power due to data collection limitations, but we can control our significance threshold. Recall our alpha can be guided by the relative costs of false positives and false negatives!

If we are very interested in detecting and effect and OK with some risk of false positives, when we have low power, we might stomach a higher alpha.

Likewise, if we have high power, we want consider pairing with a low alpha to reduce the chance of picking up on “noise”.

More Benefits to More Data

- As we take larger samples, we increase the odds of each sample being more representative.
- For example:
  - Our known distribution of age is that 50% are above 50.
  - We take two samples of 10 people each.
  - Sample 1 ends up with 80% of people over the age of 50.
  - Sample 2 ends up with 20% of people over the age of 50.
  - These samples are skewed and no longer representative. Therefore, the sample means we calculate will not represent the population - this is not good!
  - When taking small samples, such large skews are possible. However, as we increase the same size, the less probable they become. Bigger samples give us more confidence that our data is representative and that the means will fall on the “right” normal distribution.

Benefits of Statistical Significance

- Statistical significance takes some subjectivity out of our evaluation of whether or not groups have different means.
- Inherently, it views the world as a distribution rather than a single point, which is more realistic.
- It is based on logical and rigorous statistical theory.

Issues with Statistical Significance

- The classic alpha of 0.05 is somewhat arbitrary…and means we will have 5% false positives over the long run!
- Choosing an alpha may encourage p-hacking - that is, doing just enough data massaging to get statistical significance. (Do some Googling on this topic).
- The multi-comparisons problem if not addressed!
- Parametric tests make assumptions about how the world works.
- It does not communicate if results are *meaningful*, only that groups are *different*.
- Can lull us into “physics envy” - that is, that processes follow a defined pattern and adhere to laws. The real world (social sciences) are much more messy.

Statistical significance is a useful tool, but it is not the end-all be-all. The over-reliance on statistical significance has resulted in many issues in the research industry.

Industry Changes Around Statistical Significance

- Some research journals now de-emphasize p-values!

“Why Most Published Research is Wrong”

- “Why Most Published Research Findings Are False” - 2005 research paper by John Ioannidis.
- Research findings in a scientific field are less likely to be true:
  - the smaller the studies conducted.
  - the smaller the effect sizes.
  - the greater the number and the lesser the selection of tested relationships.
  - the greater the flexibility in designs, definitions, outcomes, and analytical modes.
  - the greater the financial and other interests and prejudices.
  - the hotter the scientific field (with more scientific teams involved).

Replication Issues

- This is a widespread issue but has been especially prevalent with psychology research.
- In 2015, a *Science* journal article reported that only roughly ⅓ of results from a set of psychology studies could be replicated. Part of this has to do with issues with statistical significance!

Study: Sensitivity of Statistical Significance

- In other words, how easy is it to p-hack?

Study: Sensitivity to Random Assignment

- In other words, how effective is random assignment at, well, accomplishing randomization?

Don’t Believe Every Study!

- Just because a study says something is true, don’t simply believe it. Think about issues with statistical significance, statistical power, sample sizes, and bias of all kinds.
- Science is an iterative process that builds over time.
- Distilling “science” to the results of a single study is antithetical to the scientific process. By nature, science is skeptical and requires strong evidence to move past the null hypothesis.
- This viewpoint is emphatically pro-science. Science *is* rigor and skepticism.

How to Practically Use Statistical Significance

- Use it as a tool in your larger toolbox. It is a useful guide - not an end all be all.
- Use your knowledge to question results that are statistically significant. How likely should you accept these results? At best, given the classic alpha of 0.05, there’s still a 5% chance the results are incorrect. Given other factors we’ve discussed, this figure in actually will likely be much higher.

Statistical Significance when Dealing with Populations

- The idea of statistical significance is built around samples. That is, we took a sample of data from Group X and a sample of data from Group Y. Are the sample means different?
- What about situations where we have data for an *entire population* and didn't sample? For example, you can take the mean price of every house in Kansas City and every house in St. Louis.
- You could directly compare the means and say something like, "the mean price of houses in Kansas City is greater than the mean price of houses in St. Louis". There's nothing wrong with this.
- However, it is often useful to still *treat these data as samples* that lie on a distribution and that could be compared using a test for statistical significance.
- Why? A single point never describes the world as well as a distribution. In effect, we have taken a snapshot of one universe. Under even slightly different circumstances, the prices of some homes would not be the same. Treating data as a sample hedges for uncertainty and randomness.

Statistical Significance in the Age of Prediction

- When we care about prediction, the concept of statistical significance largely goes out the door. We can about predictive power, not necessarily about explanatory power. At the end of the day, we care about if our model can predict well, not if we can perfectly tease out relationships.
- Most modern machine learning models have no concept of statistical significance (e.g., tree-based models, neutral nets). This is partly because the idea of statistical significance makes less sense with non-linear relationships.
- scikit-learn, the main Python ML library, doesn't even report statistical significance for linear regression models.
- SHAP - the most popular way to provide local and global model explanations - does not have statistical significance.
- We do care about understanding what features drive predictions (and therefore connect to our target) but in a more nuanced way (e.g., these two features are predictive together when they in between these ranges).

Confidence Intervals

- A confidence interval displays a range of values rather than a single point. Typical confidence levels are 90%, 95%, and 99%.
- 95% confidence level: If we ran this experiment 100 times, the outcome would fall in this range 95 times.
- A 95% confidence interval does not state that we are 95% confident in the outcome!
- This is a natural extension of statistical significance as it indicates that our results are not fixed to a single point but rather fall on a distribution.

A Slide Aside: Prediction Intervals

- If we are making predictions using a regression model, we could produce prediction intervals. These are different from confidence intervals since we are projecting the future.
- The Python library MAPIE provides a model-agnostic way to produce prediction intervals.
    - For example, with 90% confidence we predict the value with be between 5 and 10, with a mean predicted value of 7.
- We should always use a prediction interval in conjunction with our mean prediction!

A Related Slight Aside: Classification Model Calibration

- When building classification models, we should likewise show our level of uncertainty.
- Let’s say we are building a model to predict churn vs. no churn.
- Rather than returning the predicted label (churn or no churn), we should return the probability of a person being a churn. Our takeaways and actions can be drastically different if the predicted probability is 51% vs. 98%.
- Most classification models - except logistic regression - have idiosyncrasies in how they produce probabilities (ex: oddly, 40% predicted probability does not map to a 40% real-world probability).
    - For instance, random forest models tend to push probabilities toward the middle of the distribution and are less likely to predict probabilities at the extremes (e.g. 5% or 95%).
- Therefore, we have to calibrate them to produce better probabilities. We can accomplish this by optimizing for log loss and using scikit-learn’s CalibratedClassifierCV.