BUSA 603

Module 4: Introducing A/B Testing, Experimental Designs, and Analyzing Test Results

Brian Weikel

brian.weikel@franklin.edu

Business Analytics
Franklin University

Spring 2024

Outline[1]

1. A/B Testing
2. Experimental Design
3. Basic Principles of Statistical Testing
4. Analyzing Results
5. Appendix

[1] These lecture notes map to chapters 8, 9 and 13 of Davis (2022).

A/B Testing

A/B Testing

A/B testing is a method for testing the effectiveness of a marketing effort, such as an advertising campaign, via a controlled experiment that tests two or more conditions before exposure to the broader marketplace.

- A/B testing is based on the Scientific Method, a principle used in many ways across many science fields.
- A/B testing is also sometimes called **split testing** or **bucket testing** because audiences members are split, or audience members are placed in different buckets, to see response to different marketing inputs. These terms are generally used synonymously.
- A/B testing is also sometimes called **A/B/N testing**, signifying that more than two inputs can be tested at the same time (N is used to signify any number).

Scientific Method

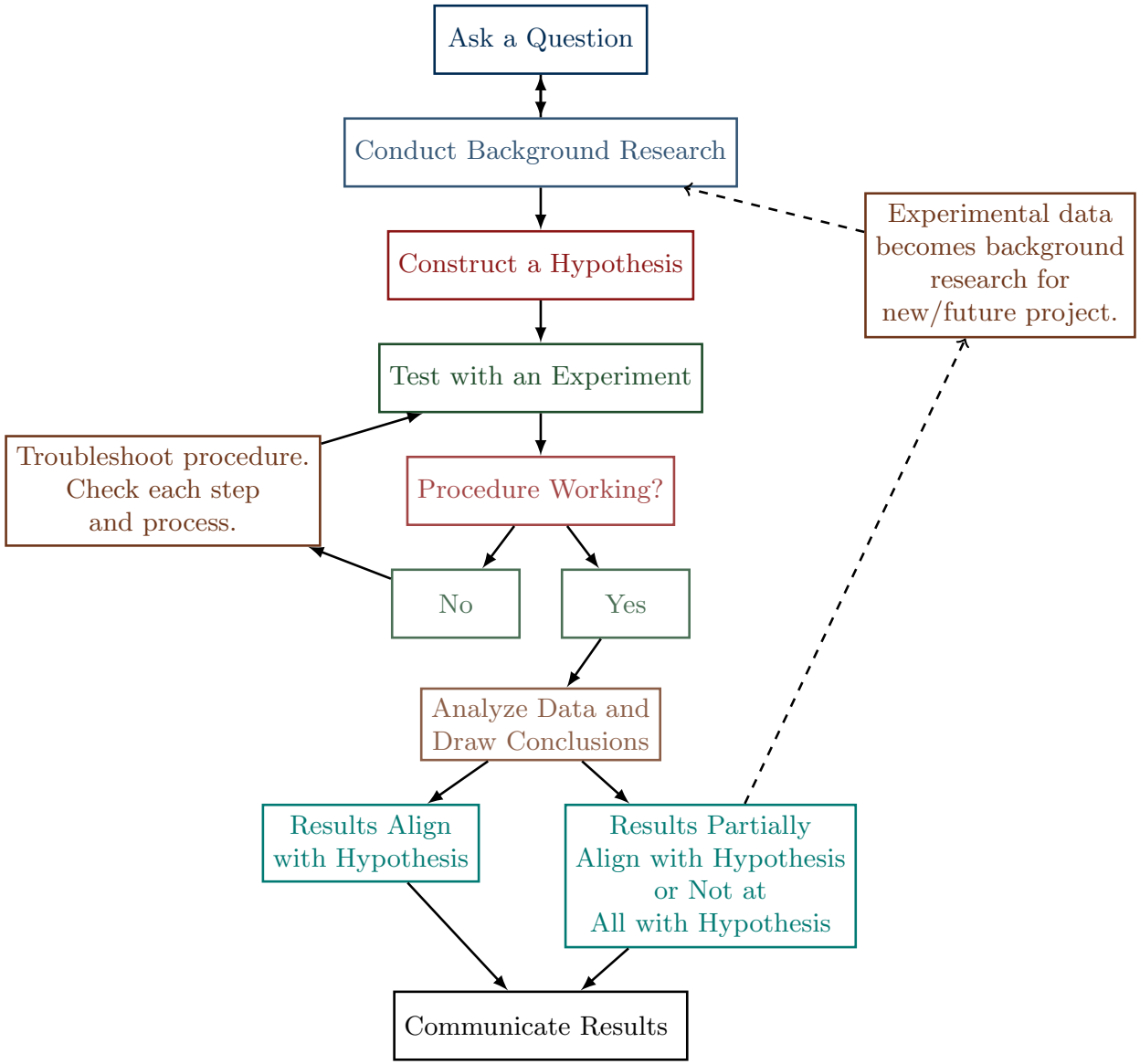


Figure 1: The Scientific Method

Conducting an A/B Test is Quantitative Research

Quantitative research involves structured data collection methods that provide results that can be converted to numbers and analyzed through statistical procedures. Statistical testing, such as A/B testing, is a quantitative research method.

Qualitative research involves unstructured data collection methods where results are subjectively interpreted.

- Qualitative research is typically used for initial exploratory research. However, it can also be used after a descriptive study to explore deeper into the minds of consumers or whoever the research participants may be.
- Since qualitative research involves probing via open-ended questions, the results become subjective. That makes generalizing the findings to a larger population or other consumers more difficult.
- The textbook refers to qualitative research approaches as *intuitive*.

Benefits and Costs of Different Research Approaches

Feature	Qualitative[2]	Quantitative
Research Type	Exploratory, Clinical, Phenomenological	Descriptive, Scientific, Causal
Sample Size	Small	Large
Question Types	Unstructured	Structured
Type of Analysis	Subjective	Objective, statistical
Generalizability	Limited	High
Costs (Typically)	Lower	More expensive
Time Frame (Typically)	Shorter	Longer

Table 1: Comparison of Qualitative and Quantitative Research

[2]For a deeper understanding of qualitative research approaches see Calder (1977).

A Few Marketing Research Approaches

Method	Approach	Knowledge Type	Rationale
Quantitative	Descriptive	Everyday	To find numerical patterns related to everyday concepts. As an example, consumption breakdowns by age.[3]
	Scientific	Scientific	To use numerical measurement to test scientific theories and constructs.
	Causal	Causal	To use numerical measurement to test causal hypotheses using data from randomized control trials or observational data studies.

[3] In the field of data science, this is called exploratory data analysis, or EDA. A classic reference is Tukey (1977). A rather recent survey of EDA tools is Ghosh et al. (2018).

A Few Marketing Research Approaches Continued

Method	Approach	Knowledge Type	Rationale
Qualitative	Exploratory	Prescientific	To generate scientific constructs and to validate them against everyday experience.
	Clinical	Quasi-scientific	To use second-degree scientific constructs without numerical measurement. For example, the use and examination of clinical judgments.
	Phenomenological	Everyday	To understand the everyday experience of the consumer.

Table 2: Summary of a Research Approaches

A/B Testing Stages

The goal of A/B testing in marketing is to identify differences in marketing outcomes after the random assignment of people to marketing input A or marketing input B.

It is called A/B testing because oftentimes two conditions are tested and these conditions are called condition A and condition B.

A/B testing occurs in two stages.

- **Exploration stage**
- **Exploitation stage**

Exploration Stage

In the exploration stage, the marketing analytics professional chooses what portion of the potential audience to explore.

Then A/B testing software randomly selects people to assign to the exploration stage.

The software then again uses randomization to assign a proportion of those chosen for the exploration stage to be exposed only to version A, and one less that proportion to version B.[4]

- The proportion is usually, but not always, a half.
- Those who see version A never know a version B exists, and vice versa.
- Assigning people to see different conditions is now much easier in the digital world than it is in the physical world.

[4] At times you may hear one version called "test" and the other version called "control", where "test" refers to a new treatment, method, structure, ... and "control" is the existing treatment, method, structure, ...

Exploitation Stage

Upon completion of the exploration stage, software determines which version was more successful.

The marketing analytics professional tells the software algorithm what marketing outcome is important in order to test whether version A or B is more successful (e.g., click-through rates, site duration, purchase conversion . . .).

- The winning version is exposed to the large remainder of the potential audience that was not part of the exploration stage.
- This exposure frequently happens automatically, but it is possible to set it up so that the marketing analytics professional is required to complete several steps before enacting the winning version.

New Approaches for A/B Testing

1. **Bandit testing** is very similar to A/B testing but differs in the exploitation stage.
   - As with A/B testing, bandit test's exploration stage compares multiple options such as versions A and B.
   - Unlike A/B testing, bandit testing has an adaptive exploitation stage. Specifically, the software determines which version was more successful and exposes both versions to some of the remaining audience, proportionate to version successfulness.[5]
2. Many software systems allow marketers to analyze the success of options during the exploration stage of A/B testing across different segments by applying the results intelligently across segments in the exploitation stage.

[5] In the exploitation stage of A/B testing, the successful campaign would go to all of the remaining people on the list, but bandit testing would send both the successful and unsuccessful campaigns to another smaller portion in a series of exploration stages, increasingly using the more successful campaign.

Potential Challenges with A/B Testing

Potential challenges with A/B testing include:

1. One of the largest challenges with A/B testing is making decisions from a small sample of customers in the exploration stage or stopping an A/B test before a sufficient number has been tested. To address this challenge, we will rely on statistical **power**. We introduce statistical power in a later section.
2. A **false positive occurs** when a condition seems to be the winner in the exploration stage, but it does not actually perform as well in the exploitation stage. The chances of a false positive are much larger when too many dimensions are tested at once, especially without ample sample size. We expand on false positives in a later section.

Potential Challenges with A/B Testing Continued

3. Running multiple tests at the same time with the same audience can create overlap and problem interactions. This happens when a version in one A/B test affects the perceptions of a version in another A/B test.
4. Distinguishing statistical significance from business significance.
5. In order to rule out seasonality, A/B tests should be conducted for lengthy periods.
6. Hypotheses need to be be rigorously defined and vetted.
7. While an A/B test may optimize a marketing mix variable with its current design, it may be a **local maximum**. That is, there may exist designs that would yield a **global maximum**.[6]

[6] For a more thorough review of A/B testing in the ad tech space, see Kohavi et al. (2009), Kohavi and Longbotham (2011), and Kohavi and Longbotham (2017).

Tools for Running A/B Tests

There are many online tools available for A/B testing. The textbook recommends the following, which are not listed in any particular order.

1. Optimizely
2. VWO
3. Convert Experiences
4. SiteSpect
5. AB Tasty
6. Firebase A/B Testing
7. Qubit
8. Adobe Target

See the textbook for a few other marketing tools with A/B testing and crowdsourcing capabilities.[7]

[7] **Crowdsourcing** is the practice of gathering information by requesting the services of a large number of people, typically through online connection. In general, crowdsourcing is not recommended for rigorous statistical testing since many potential biases may occur, such as **voluntary response bias**, **self-interest bias**, and **social-desirability bias**, to name few.

Experimental Design

Experiments

It is wise to take time and effort to organize an experiment properly to ensure that the right type of data, and enough of it, is available to answer the questions of interest as clearly and efficiently as possible. This process is called **experimental design**.[8]

1. An **experiment** deliberately imposes a treatment (or condition) on a group of objects or subjects in the interest of observing the response.
    - An **observational data study** involves collecting and analyzing data without changing existing conditions.[9]

[8] The National Institute of Standards and Technology has an informative introductory exposition on experimental designs.

[9] Three types of bias can arise in observational data: (i) confounding bias, (ii) selection bias (i.e, improper selection of participants through stratifying, adjusting or selecting), and (iii) measurement bias (i.e., imprecise or biased measurement of variables in analysis).

Treatments and Factors

- Because the validity of an experiment is directly affected by its construction and execution, attention to experimental design is extremely important.

2. In experiments, a **treatment** is something that researchers administer to experimental units. For example, a customer arriving on a landing page layout could be "treated" (i.e., served) one of three different pages.
3. A **factor** of an experiment is a controlled independent variable; a variable whose **levels** are set by the experimenter. A factor is a general type or category of treatments.
   - Different treatments constitute different levels of a factor.
   - Example: Using a website's source of origin for a user, 3 different group's users are subjected to distinct landing pages. The users are the units. There is 1 factor, 'landing page type.' The factor's levels are the 3 different landing pages.

Experimental Bias

Suppose a website designer wishes to evaluate a new site design. Based on the availability of multiple web servers and after unit, integration, regression, and end-user acceptance testing, she persuades the firm's technology organization to conduct an experiment by using the new design between 12:01 AM and 12:00 PM, while using the existing design between 12:01 PM and 12:00 AM.

The problem with this experiment is that the designer has neglected to control the differences in customers. For example, the marketing department knows that those who arrive between 12:01 PM and 12:00 AM tend to have a larger rate of employment, larger disposable income, and larger conversion rates. This is an example of **experimental bias**.

A Benefit of Randomization

Because it is generally extremely difficult for experimenters to eliminate bias using only their expert judgment, the use of **randomization** in experiments is common practice.

- In a randomized experimental design, objects or individuals are randomly assigned to an experimental group.
- Using randomization is the most reliable method of creating homogeneous treatment groups, without involving any potential biases or judgments.
- One caveat is that randomization is not perfect. By chance, it is possible that one group happens to have younger people on average than the comparison group.

Causal Inference

Cause-and-effect approaches are extremely important because decision-makers are usually interested in knowing whether a marketing action might cause a desired marketing outcome, not just whether they happen to occur at the same time. More succinctly, they are interested in causation and not just correlation.

Causal inference is the formal name for cause-and-effect interpretations of research methods like experiments.

- Rubin (1974) and Holland (1986) are seminal papers in the modern development of causal inference in the statistical literature.
- An excellent introduction to causal inference from an econometric perspective is Angrist and Pischke (2009).
- On the use of Bayesian Networks to explain causality, see Pearl (2009) and Pearl and Mackenzie (2018).

Requirements for Causal Inference

Let X be a factor, such as a marketing input. Let Y be the outcome, such as a purchase conversion. For X to cause Y:

1. X must occur before Y chronologically.
2. There must be evidence that X is associated with Y.
3. Other causal factors are ruled out (e.g., "controlled for").

The assumption that other potential causal factors have been eliminated is almost always the most challenging of the three to realize.

Sleeping with Shoes On

It is important to emphasize that correlation is not the same as causation. In general when two variables are correlated, we cannot conclude that changing the value of one variable will cause a change in the value of the other.

Consider the following example:

- Sleeping with one's shoes on is strongly correlated with waking with a headache.
- Sleeping with one's shoes on must cause headaches.
- Thus one should take off one's shoes before sleeping.

There is an error in the causal conclusion. What is a likely better explanation of the correlation?

Sleeping with Shoes On Continued

Solution: There is a 3rd factor that is related to sleeping with one's shoes on and waking with a headache. Sleeping with one's shoes on is not as comfortable as sleeping sans shoes. Thus one must not have all of his/her cognitive abilities prior to falling asleep. For example, being inebriated. Intoxication is a **confounder**.

A confounder is a variable that is related to both the treatment and the outcome. When a confounder is present, it is difficult to determine whether outcome differences are due to it or to the treatment.[10]

- Inebriation is related to both sleeping with one's shoes on and waking with a headache.
- The fact that sleeping with one's shoes on is correlated with waking with a headache does not mean changing one variable will cause the other variable to change.

[10] Confounders may be observed or unobserved.

Sleeping with Shoes On Continued

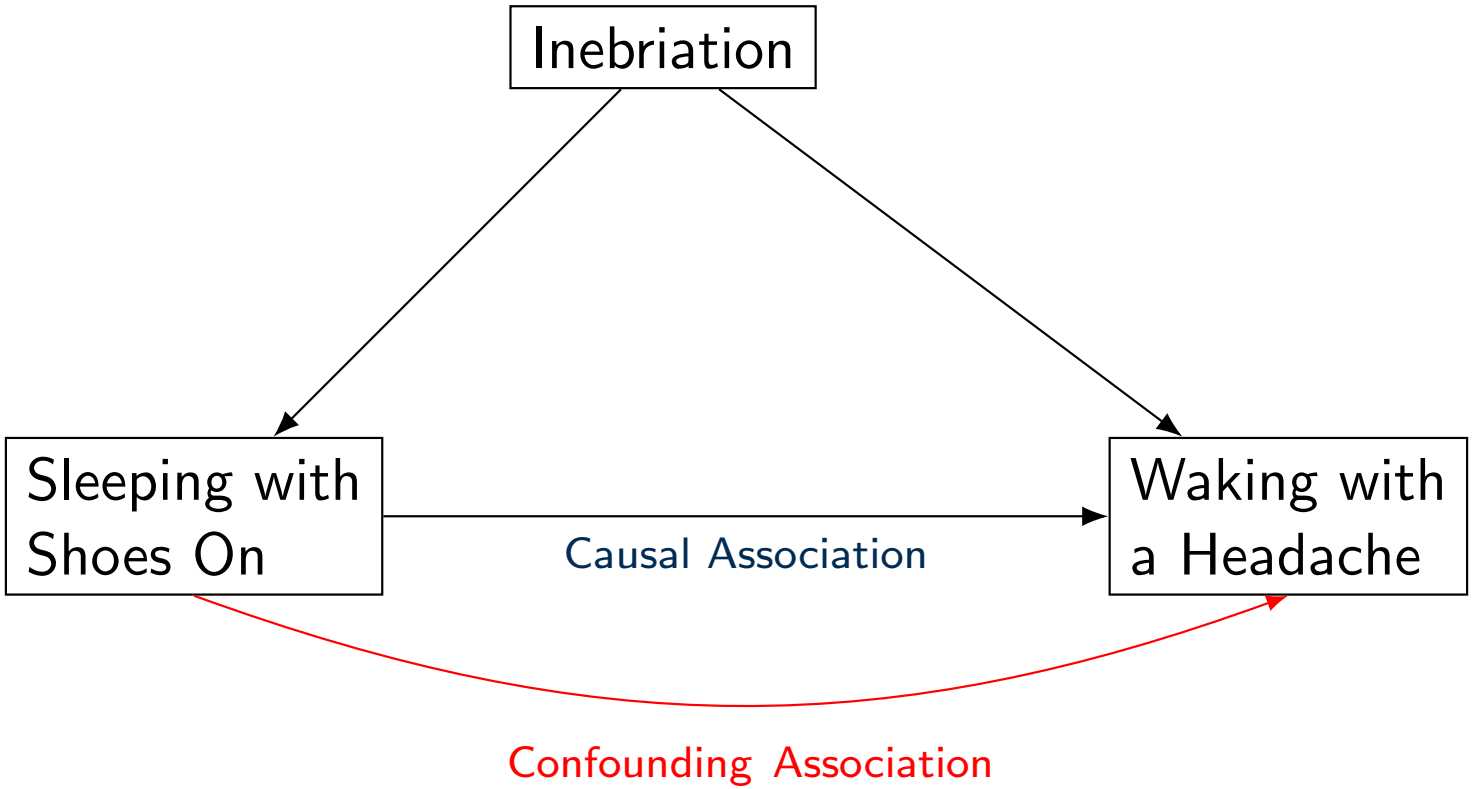


Figure 2: Causal Diagram with A Confounder

Total association (e.g., correlation) is a mixture of causal and confounding association.

By design, experiments fulfill the three requirements of causal inference.

1. The first requirement that X occurs before Y is realized by first manipulating X and then measuring Y. For example, X is a display banner ad with different banner ad colors, sizes, and messages (i.e., different treatments), and Y is the click-through rate.
2. Conducting marketing analytics via statistical methods ensures the second requirement is realized.
3. The third requirement that other causal factors be ruled out or controlled for is typically accomplished through random assignment of treatments. Thus another benefit of randomization!

Simple and Interactive Designs

1. **Simple Designs**
   - The most basic experimental design has a single factor.
   - Examples of a marketing factor include a campaign's message, a campaign's creative (e.g., colors on a display ad), time period of campaign, promotional discounts, ...[11].
2. **Interactive Designs**
   - Pertains to the use of two or more factors.
   - Interactive designs are powerful because each factor can be tested across levels of another factor. For example, consider a women's category landing page where "above the fold" shows one of two items, where each item could be one of three colors. For these two factors, there are $2 \times 3 = 6$ treatment combinations, or **interactions**.

[11] A particular item's regular price and clearance price are almost always determined by the finance department of the company.

After-Only Designs

For an **after-only experimental design** measurements of the outcome variable are *only taken after* the experimental units have been administered a factor treatment.

- In A/B testing, the estimate of the effect is the outcome of one treatment minus the outcome of the other treatment.
  - This is a common approach in advertising research where a sample of target customers are interviewed, questioned, surveyed . . . following exposure to an advertisement and their recall of the product, brand, or sales features is measured.
  - The ad could be one appearing on national broadcast TV, syndicated radio, magazines, newspapers, ad-serving website, or some other media. The amount of information recalled by the sample is taken as an indication of ad effectiveness.

Before-After Designs

- The chief problem with after-only designs in advertising is that they do not afford any control over extraneous factors that could have influenced the post-exposure measurements. If the time between the treatment and outcome measurement is rather short, this problem is typically not too severe.

A **before-after experimental design** requires the researcher to measure the outcome variable *before and after* experimental units have been administered a factor treatment.

- The estimate of the effect is the before-after difference between one treatment and another. In economics, this is referred to as a **difference-in-difference design**.
- After-only designs have the advantage of no pre-test bias compared with the before-after design.

Within-Subject Designs

A **within-subjects design**, or within-units design, is an experiment in which the same experimental unit is administered more than one treatment, each with a distinct outcome measure.

- For example, it is possible to expose the same user to different website category landing pages layouts over time to determine which results in (say) a larger purchase conversion probability.
- The main benefit of this design is that the sample size for analyzing results is increased since each participant is his or her own comparison group.
- However, this design introduces bias, such as seeing one layout and having it influence the way the user views the next layout.

Laboratory Experiments

A/B testing is just one type of experiment.

Non-A/B-test experiments include **laboratory experiments**, which are typically run under highly controlled conditions offline.

In contrast to A/B testing, a laboratory experiment aims to reduce all other possible influences on an outcome in an effort to isolate the role of a marketing effort on a marketing outcome.

Field and Natural Experiments

Field experimentation means running experiments in naturally occurring environments rather than the laboratory. A naturally occurring environment may be a city street, a store, a bazaar, or a website.

- **Field experiments**, though defined as occurring in the naturally occurring world, are not to be confused with natural experiments.
- **Natural experiments** are not randomized experiments. In natural experiments, nature makes the treatment assignments.

An experiment's ability to provide cause-and-effect information elevates it as the gold standard for addressing marketing attribution issues.[12]

[12] Recall that **attribution** is defined as assigning credit to event that lead to marketing conversions.

Basic Principles of Statistical Testing

Adding Rigor to Testing

When a sample is drawn from a population the data can be used to make **inferential** statements about population characteristics, such as **point estimates** (e.g., means and proportions) and **confidence intervals**.

Alternatively the sample may be used to assess the validity of a conjecture, or **hypothesis**, that one may have formed about the population.

1. A consumer packaged goods manufacturer of salad dressing claims that on average the contents of a particular bottle size is at least 13 fluid ounces.

2. A niche apparel retailer's distribution center only accepts large deliveries if no more than 1% of items are defective.
3. A website designer wants to know if a local automobile dealer's website appeals equally to men and women.
4. An auto insurer believes that owners of a sports utility vehicle (SUV) are more likely than owners of a pickup truck to file an accident claim per mile driven.

As you may surmise based on our discussion of A/B testing, a random sample of the population, with randomized treatment assignments, can be used to determine the validity of each conjecture.

A General Framework

- Suppose that some hypothesis has been formed about the **parameter** of interest (e.g., population mean, population proportions, ...) and this hypothesis will be believed unless sufficient contrary evidence is produced. This hypothesis is called the **maintained hypothesis** or **null hypothesis**, typically denoted as H_0.
- If the null hypothesis is not true, then some alternative must be true, and in carrying out a hypothesis test, one formulates an **alternative hypothesis**, typically denoted as H_1, against which the null hypothesis is tested.
- Hypotheses may specify a single value or a range of values for the population parameter of interest. These hypotheses are **simple** and **composite** hypotheses, respectively. Alternative hypotheses may be **one-sided** or **two-sided**, or **one-tailed** or **two-tailed**, respectively.

Returning to the Examples of this Section

1. μ is the population mean of salad dressing weights.

$$H_0\colon\ \mu \geq 13$$

$$H_1\colon\ \mu < 13$$

A composite null hypothesis and a composite one-sided (left-tailed) alternative hypothesis.

2. p is the population proportion of defectives.

$$H_0\colon\ p \leq 0.01$$

$$H_1\colon\ p > 0.01$$

A composite null hypothesis and a composite one-sided (right-tailed) alternative hypothesis.

Continuing with the Examples

3. $p_1 - p_2$ is the difference between two population proportions

$$H_0\colon\ p_1 - p_2 = 0$$

$$H_1\colon\ p_1 - p_2 \neq 0$$

A simple null hypothesis and a composite two-tailed alternative hypothesis.

4. $\mu_1 - \mu_2$ is the difference between the population mean for SUV accident claims per mile driven and the population mean for pickup truck accident claims per mile driven.

$$H_0\colon\ \mu_1 - \mu_2 = 0$$

$$H_1\colon\ \mu_1 - \mu_2 > 0$$

A simple null hypothesis and a composite one-sided (right-tailed) alternative hypothesis.

Decision Rule

- After specifying the hypotheses and collecting sample information, a decision about the null hypothesis must be made.
- We may **accept** the null hypothesis or **reject** it in favor of the alternative. At times you will hear, "we do not reject the null hypothesis," instead of, "we accept the null hypothesis."
- To reach a conclusion, a **decision rule** is formulated based on the sample information.
- Due to the randomness of the sample, we cannot know with certainty whether the null hypothesis is true or false. Thus for any decision rule adopted, there is some chance of reaching an erroneous conclusion about the parameter of interest.

Type I and Type II Errors

- If the null hypothesis is rejected even though it is true, then a **Type I error** is made.
- If the null hypothesis is accepted even though it is false, then a **Type II error** is made.

The table below summarizes the errors that can be made.

Null Hypothesis Decision	**State of Nature**	
	True Null Hypothesis	False Null Hypothesis
Accept	Correct decision probability $= 1 - \alpha$	Type II error probability $= \beta$
Reject	Type I error probability $= \alpha$, where α is called the **significance level**	Correct decision probability $= 1 - \beta$, where $1 - \beta$ is called **power**

Table 3: Decision-Making Errors

Choosing the Significance Level

- We would like to have Type I and Type II error rates as small as possible.
- However, there is clearly a trade-off between the two: once a sample has been taken, any adjustment to the decision rule that makes it less likely to reject a true hypothesis will render it more likely to accept the hypothesis when it is false.
- For a given sample size, one fixes at some desired level the probability of a Type I error, that is the significance level. For a given decision rule, the Type II error rate is then given.
- Increasing the sample size decreases the Type II error rate for a given significance level. This is important when designing statistical experiments.

Determining Statistically Significant Results

To summarize, hypothesis testing incorporates the following to determine whether the results are *statistically significant*.[13]

- **Effect size**: The larger the effect size, the less likely it is to be random error.
- **Sample size**: Larger sample sizes allow hypothesis tests to detect smaller effects.
- **Variability**: When sample data has greater variability, random sampling error is more likely to produce considerable differences between the experimental groups even when there is no real effect.

For a given hypotheses, an expected effect size, and variability, sample sizes may be determined to obtain a desired statistical power. This is referred to as **power analysis**.

[13] Across many disciplines there has a growing chorus to abstain from saying, "statistically significant results," or the like. For additional information see Wasserstein et al. (2019).

Power Analysis Curve and Sample Size Determination

Illustrative power analysis curve for an A/B digital test where the test group receives a campaign ad, the control group receives a public service announcement ad (e.g., donate to the American Red Cross), and the outcome is average value (e.g., U.S. dollar) sales per user.

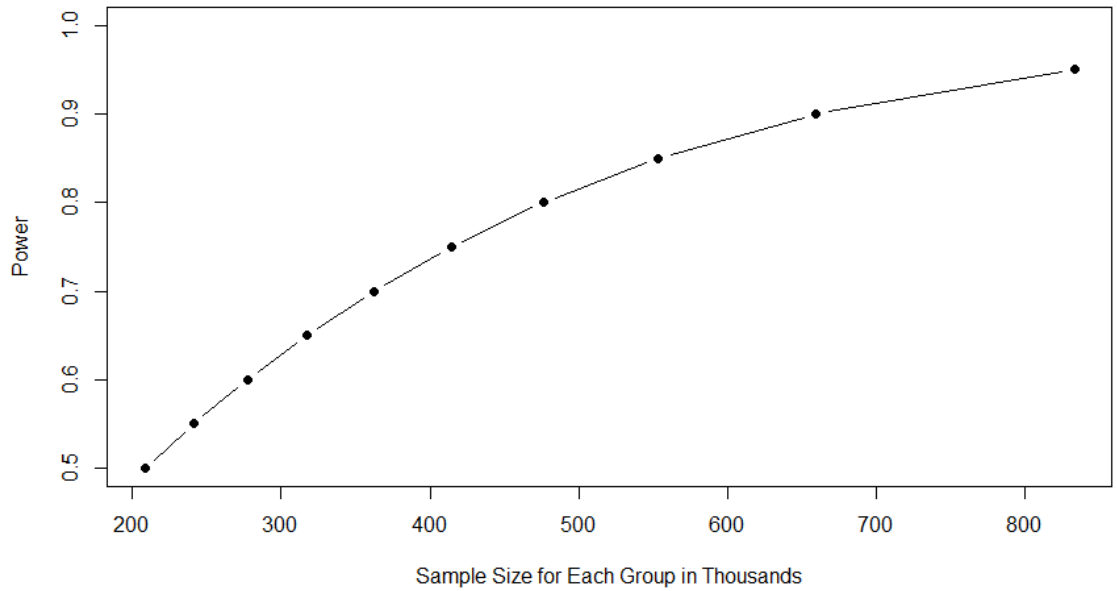


Figure 3: Power Analysis Curve for a Two Sample T-Test[14]

[14] Effect Size: 0.8¢. Pooled Standard Deviation: \$1.57. Significance Level: 0.10. Two-sided alternative hypothesis.

Analyzing Results

A CPG Digital Advertising Campaign

Suppose a consumer packaged goods (CPG) manufacturer ran an ad campaign announcing a new product line extension for a particular brand.[15] The campaign was solely executed on a particular digital platform.

Suppose the company that owns the digital platform has many user-level attributes, such as gender, age, location, . . . , browsers, operating systems, . . . , written content created by the user, content provided by the user via camera feature, content a user viewed or interacted with, . . . ,

The aforementioned attributes are used to select an audience, which is a set of users who may be exposed to an ad of the focal campaign.

[15] While this case study is similar to the one presented in Module 1, there are differences.

Intent-to-Treat

A test group user's exposure is a function of the user logging into the platform, the digital platform's targeting algorithms, and the bidding process.

The audience chosen for the campaign summarized must have had a frequent shopper card (FSC) and must have had at least $75 in monthly total FSC gross value sales for 10 of the 12 months before the campaign commenced.[16] The FSC requirement permits test and control group users to be linked to sales outcomes such that sales ad effectiveness can be determined.[17]

[16] Recall the FSC historical purchase requirement is referred to as a *static*. Using a static ensures that a user is using his/her/their FSC cards with some cross-time regularity.

[17] In the A/B testing lexicon of the platform, the test group is 'A' and the control group is 'B'.

Ad Campaign Objectives

The campaign ran for 6 weeks, where in addition to increasing awareness and consideration of the line extension, increasing brand conversion rates, average brand purchase occasions, and average value (i.e., U.S. dollar) sales. The sales outcome post campaign period was 2 weeks. The design was after-only.

Since the treatment, a brand ad, cannot always be guaranteed to be delivered and the targeting process occurs after a user is assigned to either the test and control group, an intent-to-treat (ITT) testing approach was used to determine the sales effectiveness of the campaign. Fifty percent of the audience was randomly assigned to the test group and fifty percent to the control group.[18]

[18] For additional information on the ITT approach, see Gordon et al. (2019).

Campaign Exposure Contingency Table

Elements of the deliverables provided by the platform after the campaign has completed are a set of contingency tables.

Age	Gender	Exposure		Total
		No	Yes	
18-29	Unknown/Other	9,678	1,136	10,814
	Male	44,431	19,847	64,278
	Female	72,625	24,639	97,264
30-39	Unknown/Other	14,532	1,809	16,341
	Male	65,942	30,618	96,560
	Female	107,748	38,106	145,854
40-49	Unknown/Other	14,821	1,939	16,760
	Male	67,277	32,504	99,781
	Female	108,571	40,911	149,482
50-59	Unknown/Other	9,362	1,195	10,557
	Male	42,671	20,038	62,709
	Female	68,881	24,968	93,849
60+	Unknown/Other	45,71	545	5,116
	Male	21,043	9,470	30,513
	Female	34,400	11,668	46,068
Unknown	Unknown/Other	2,994	423	3,417
	Male	13,607	6,568	20,175
	Female	22,170	8,292	30,462
	Total	725,324	274,676	1,000,000

Table 4: Campaign Exposure-Demographic Table

Campaign Test Variant Contingency Table

Age	Gender	Test Variant		Total
		Test	Control	
18-29	Unknown/Other	5,458	5,356	10,814
	Male	32,060	32,218	64,278
	Female	48,853	48,411	97,264
30-39	Unknown/Other	8,114	8,227	16,341
	Male	48,258	48,302	96,560
	Female	73,008	72,846	145,854
40-49	Unknown/Other	8,303	8,457	16,760
	Male	49,920	49,861	99,781
	Female	74,799	74,683	149,482
50-59	Unknown/Other	5,296	5,261	10,557
	Male	31,527	31,182	62,709
	Female	46,903	46,946	93,849
60+	Unknown/Other	2,602	2,514	5,116
	Male	15,338	15,175	30,513
	Female	22,978	23,090	46,068
Unknown	Unknown/Other	1,774	1,643	3,417
	Male	10,043	10,132	20,175
	Female	15,248	15,214	30,462
	Total	500,482	499,518	1,000,000

Table 5: Campaign Test Variant-Demographic Table

Table 5 suggests that the random assignment of a user to a test or control group is working as planned.

Campaign Test Variant-Exposure Table

Exposure	Test Variant		Total
	Test	Control	
No	225,806	499,518	725,324
Yes	274,676	0	274,676
Total	500,482	499,518	1,000,000

Table 6: Campaign Test Variant-Exposure Table

As may be expected, no user in the control group was exposed to a campaign ad.

55.3% of test group users were exposed to campaign ad.

In practice, a set of statistical tests would be conducted to determine if the test and control groups were balanced, say by age and gender.

Campaign Sales Effectiveness

Table 7 contains large sample test results for conversion rate, average purchase occasions, and average value sales. Obviously the campaign was not sales outcome effective.[19]

Outcome	Measure	Test	Control
Conversion Rate	Point Estimate	0.0163	0.0160
	Alternative Hypothesis	Two-Sided	
	p-value	0.3639	
Average Purchase Occasions	Point Estimate	0.0171	0.0168
	Alternative Hypothesis	Two-Sided	
	p-value	0.2878	
Average Value Sales	Point Estimate	\$0.1922	\$0.1885
	Alternative Hypothesis	Two-Sided	
	p-value	0.2741	

Table 7: Campaign Sales Effectiveness Using Test/Control Groups

[19] If the sample was small and purchase occasions and value sales are normally distributed, and given the unknown variances, one would use a *t*-test in lieu of the large sample test. A small sample test of proportions could use bootstrap tests (Efron and Tibshirani (1993) and Shao and Tu (1996)).

Be Wary of Exposed/Unexposed Group Test Results

If we only knew users who were exposed and not exposed, that is we do not know which test variant a user was assigned, we would have an observational data study. Thus one may wish to conduct statistical inference with exposed and unexposed groups.[20]

Outcome	Measure	Exposed	Unexposed
Conversion Rate	Point Estimate	0.0170	0.0158
	Alternative Hypothesis	Two-Sided	
	p-value	< 0.0001	
Average Purchase Occasions	Point Estimate	0.0179	0.0168
	Alternative Hypothesis	Two-Sided	
	p-value	< 0.0001	
Average Value Sales	Point Estimate	\$0.2038	\$0.1854
	Alternative Hypothesis	Two-Sided	
	p-value	< 0.0001	

Table 8: Campaign Sales Effectiveness Using Exposure Groups

[20] Note it is not appropriate to conduct such an exercise, either from a statistical or causal inference point of view, without proper statistical adjustments. Potential approaches are mentioned later in this section.

Revisiting Causal Inference

Notice that the exposed outcome estimates of Table 8 are larger than the control outcome estimates of Table 7.

Two broadly divided causal inference approach types:

1. Statistical adjustment to control confounding and arrive at a causal estimate.
   - These approaches rely on the assumption that there is no remaining unmeasured confounding and no measurement error after the application of the methods.
   - Effective statistical adjustment for confounding requires knowing what to measure, and measuring it accurately.
2. Design-based methods such as randomized control trials, and its digital space A/B testing implementation.[21]

[21] The methods do not rely on the supposition that there is no remaining unmeasured confounding and no measurement error.

Observational Data Study Causal Inference Methods

Approaches that rely on statistical adjustment are likely to have similar, or at least related, sources of bias. Design-based approaches are more likely to have different sources of bias.

Observational data study methods used in an effort to realize accurate and precise causal inference include, but are not limited to:

1. Matching Methods
   - Exact Matching
   - Stratification
   - Nearest Neighbor Covariate Matching
   - Propensity Scores

Instrumental Variables, Matching Methods . . .

2. Instrumental Variables
3. Doubly Robust Estimators
   - Target Maximum Likelihood
   - Doubly/Debiased Machine Learning
4. Difference-in-Differences
5. Regression Discontinuity
6. Synthetic Control Method

Taddy et al. (2023) contains introductions to several of the methods mentioned above. Hernán and Robin (2020) is a more rigorous treatment of several of the approaches.

Experimental Design Testing Approaches

To conduct statistical inference for A/B tests, one typically uses large sample tests or t-tests. These tests yield **frequent statistics**. Table 7 tests are large sample frequentist tests.

An alternative to the frequentist approach to testing is **Bayesian statistical inference**. It is an approach to data analysis and parameter estimation based on Bayes' theorem.[22]

Unique for Bayesian statistics is that all observed and unobserved parameters in a statistical model are given a joint probability distribution.

[22] The Appendix contains a brief review of Bayes' theorem.

Prior, Likelihood and Posterior Distribution

1. **Prior distribution**: It is the beliefs held by researchers about the parameters in a statistical model before seeing the data, expressed as probability distributions.
2. **Likelihood function**: The conditional probability distribution of the observed data given the parameters, defined up to a constant. This is called a likelihood because for a given pair of data and parameters it registers how 'likely' is the data.
3. **Posterior distribution**: A way to summarize one's updated knowledge, balancing prior knowledge with observed data, and is used to conduct inferences. More precisely, the likelihood is combined with the prior distribution to form the posterior distribution.

Posterior Distribution Specification

The posterior distribution results from applying Bayes' rule. Let D represent the data and θ the parameters of interest (e.g., means, proportions, regression coefficients ...). An application of Bayes' Rule yields.

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}, \quad \text{where:} \tag{1}$$

- $P(\theta)$ is the prior distribution. This is the strength in our belief of θ without considering the evidence.
- $P(D|\theta)$ is the likelihood function. This is the probability of observing the data as generated by a model with parameter θ.
- $P(D)$ is the evidence. This is the probability of the data as determined by summing (or integrating) across all possible values of θ, weighted by how strongly we believe in the particular values of θ.
- $P(\theta|D)$ is the posterior distribution. This is the refined belief of θ once the evidence has been taken into account.

Bayesian Inference is Becoming More Popular

For every frequentist approach to a particular statistical problem, there is usually at least one Bayesian approach. For example, there are many Bayesian one-sample and two-sample t-test variants.[23]

Bayesian inference can be computationally demanding. As computation has become more efficient and cost-effective, Bayesian inference popularity and use has experienced increased industry use.

[23] Examples include those of Gönen et al. (2005), Wang and Liu (2016), Abdelrazeq et al. (2020), Gronau et al. (2020), and Al-Labadi et al. (2022).

Analysis of Variance

Recall that an A/B test is only one type of experiment. Specifically, it is *a field experiment in the digital world*. A/B tests have 1 factor with 2 treatments.

For tests with multiple factors and hence potential interactions, we may wish to compare many pairs of treatments at the same time. One method to conduct such inference is via Analysis of Variance (ANOVA). ANOVA allows for simultaneous comparisons of more than two factors.[24]

- While the main result of an ANOVA is again the p-value, this time the p-value is associated with an F-statistic.
- The F-statistic can handle more factors than the one factor standard normal statistics (i.e., "z statistics") and t-statistics.

[24] As an example consider the previously discussed different women's category landing page where "above-the-fold" shows one of two items, where each item could be one of three colors. For these two factors, there are $2 \times 3 = 6$ interactions.

A Retailer Experimental Design

A mall-based retailer with stores nationwide is considering changing its front-facing windows and store-layout, or "floorset". Before making such changes, the company wants to ensure that the change will have a positive effect on store revenue. In particular, the proposed changes would ideally result in larger gross value sales per store selling foot. Of the more than 1,000 stores, 80 stores were randomly selected to participate in the test.

The company designed the following test:

- Total number of Stores: 80
- Factor 1: Window (W)
  - Treatment W_1: Window display from last year. Referred to as the *champion window*.
  - Treatment W_2: Proposed new window display. Referred to as the *challenger window*.

A Retailer Experimental Design Continued

- 50% of the 80 stores were randomly selected to be assigned W_2. The other 50% of stores were assigned W_1.

- Factor 2: Floorset (FS)
  - Treatment FS_1: Floorset from last year. Referred to as the *champion floorset*.
  - Treatment FS_2: Proposed new floorset. Referred to as the *challenger floorset*.
  - 50% of the 80 stores were randomly assigned FS_2. The other 50% of stores were assigned FS_1.
- This design has 2 factors, each with 2 levels, and thus there is the possibility of an interaction effect.
- The test window is 4 weeks.

Visualizing the Data

We should take a look at the data before conducting statistical inference via an ANOVA.

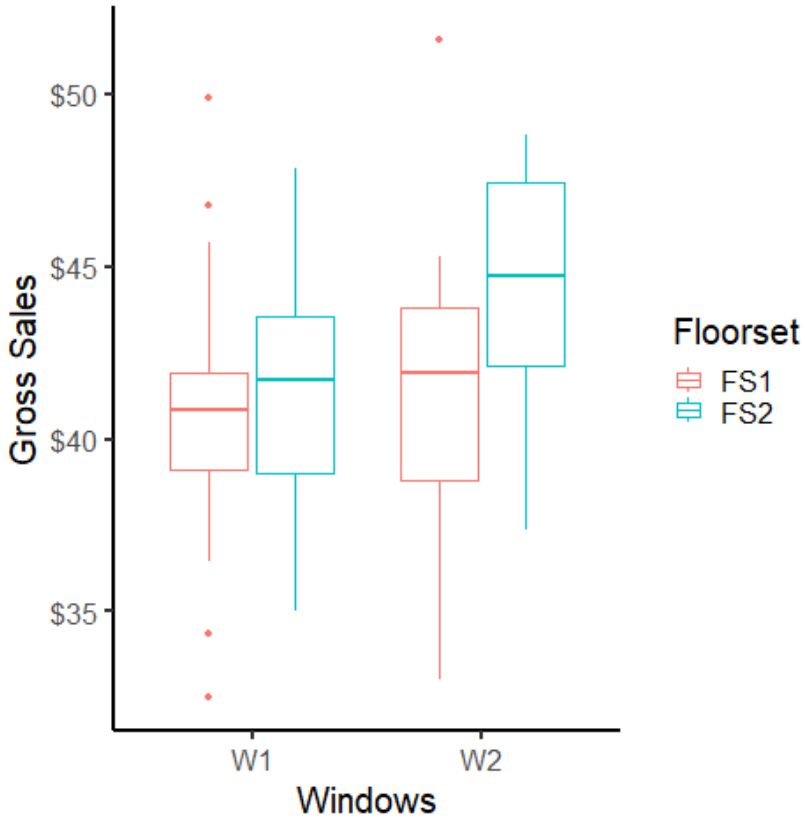


Figure 4: Boxplot of Window and Floorset Treatment Gross Sales per Selling Square Foot Across Stores

ANOVA Specification

For i indexing a store, $j = \{W_1, W_2\}$, and $k = \{FS_1, FS_2\}$, the two-way ANOVA specification for this test is:

$$Y_{ijk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \epsilon_{ijk}, \tag{2}$$

where $\epsilon_{ijk} \sim \mathrm{N}(0, \sigma^2)$ are independent and $\sum_{\forall j} \alpha_j = \sum_{\forall k} \beta_k = \sum_{\forall j, \forall k} (\alpha_j \beta_k) = 0$.

Using the estimated coefficients of the model, we may test treatment **contrasts**.

A contrast is essentially a difference in regression coefficients.

Interpreting the ANOVA Table

The table below is the ANOVA table for the retailer's test design.

```
Analysis of Variance Table
Response: gross_sales
                Df  Sum Sq Mean Sq F value  Pr(>F)
window           1   61.49  61.493  4.0433 0.04789 *
floorset         1   59.49  59.491  3.9117 0.05158 .
window:floorset  1   25.55  25.554  1.6803 0.19881
Residuals       76 1155.84  15.208
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
                0.1 ' ' 1
```

1. The upper part of the table tells us we are looking at an ANOVA table where the outcome (i.e., response) was `gross_sales`, which is truly gross value sales.

Interpreting the ANOVA Table Continued

2. The 3 lines that begin with `window` value are for testing the following null hypotheses:
   - $H_{0,W}$: Gross sales per selling square foot of the challenger window is equal to that of the champion window.
   - $H_{0,FS}$: Gross sales per selling square foot of the challenger floorset is equal to that of the champion floorset.
   - $H_{0,W\times FS}$: There is no gross sales per selling square foot interaction effect for the challenger window and challenge floorset.
3. The alternative for each null hypothesis is two-sided.

Interpreting the ANOVA Table Continued

4. The `F value` is the test statistic for each variable (i.e., factor or interaction of factors). These provide a measure of how large and consistent the effects associated with each variable are. Each `F value` has a pair of degrees of freedom associated with it: one belonging to the variable itself, the other due to the error (residual). Together, the `F value` and its degrees of freedom determines the p-value, which is `Pr(>F)` in the table.

5. The p-value gives the probability that the differences between the set of means for each variable in the model, or a more extreme difference, could have arisen through sampling variation under the null hypothesis of no difference.

6. The ANOVA table tells us nothing about the direction of the effects. Via the least squares fit of the model, we have the following estimates: $\hat{\mu} = 40.80$, $\hat{\alpha}_{W_2} = 0.62$, $\hat{\beta}_{FS_2} = 0.59$ and the interaction coefficient for FS_2 and W_2 is $\hat{\gamma}_{W_2,FS_2} = 2.26$.

Average Gross Value Sales Effectiveness

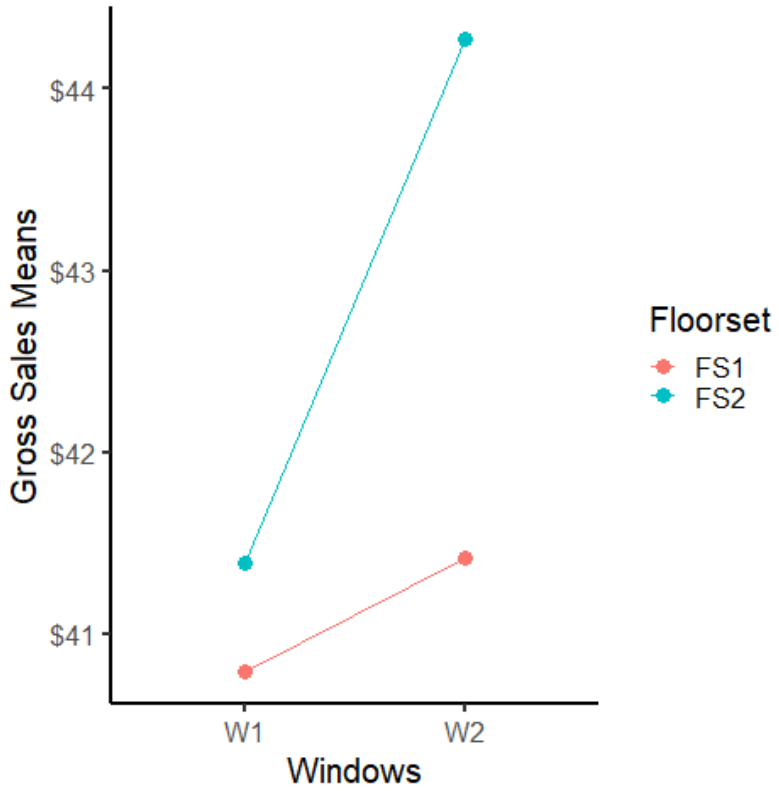


Figure 5: Average Gross Value Sales Per Selling Square Foot For Each Factor's Treatment

Based on these results, do you think the business should implement the challenger floorset, the challenger window, both, or none? What are the reasons for your answer?

Appendix

The General Multiplication Rule for Probabilities

For two events A and B, the **General Multiplication Rule for Probabilities** is $P(A \cap B) = P(A)P(B|A)$.

Two events are **independent** if the occurrence of one does not affect the probability that the other event occurs. If two events are not independent, we say they are **dependent**.

The **Multiplication Rule for Independent Events** says if A and B are independent events, then

$$P(A \cap B) = P(A)P(B). \tag{3}$$

This rule can be extended to the case where there are more than two independent events. If $A, B, C, \ldots$ are independent events, then

$$P(A \cap B \cap C \cap \ldots) = P(A)P(B)P(C)\ldots. \tag{4}$$

Adjusting Probability Statements

Bayes' theorem provides a way of revising conditional probabilities by using available information. It also provides a procedure for determining how probability statements should be adjusted, given additional information.

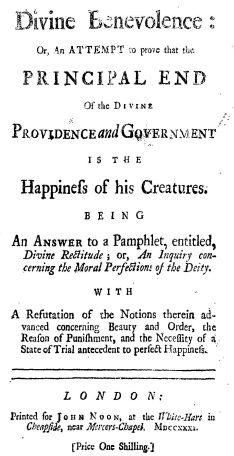
Reverend Thomas Bayes (1702–1761) developed Bayes' theorem, originally published in 1763 after his death and again in 1958. Because games of chance — and, hence, probability — were considered to be works of the devil, the results were not widely publicized.

Since World War II a major area of statistics and a major area of management decision theory have developed based on the original works of Thomas Bayes.

Reverend Thomas Bayes

Bayes' first publication was a theological work, *Divine Benevolence*, where he was trying to address the motivating source of God's actions.[25]

Reverend Thomas Bayes



Divine Benevolence :

Or, An ATTEMPT to prove that the

PRINCIPAL END

Of the DIVINE

PROVIDENCE *and* GOVERNMENT

IS THE

Happineſs of his Creatures.

BEING

An ANSWER to a Pamphlet, entitled, *Divine Rectitude*; or, *An Inquiry concerning the Moral Perfections of the Deity.*

WITH

A Refutation of the Notions therein advanced concerning Beauty and Order, the Reaſon of Punishment, and the Neceſſity of a State of Trial antecedent to perfect Happineſs.

LONDON:

Printed for JOHN NOON, at the *White-Hart* in *Cheapſide*, near *Mercers-Chapel*. MDCCXXXI.

[Price One Shilling.]

Divine Benevolence

[25] Since no author appears on the title page of *Divine Benevolence*, or anywhere else, it is sometimes considered to be of doubtful authorship. For example, the *National Union Catalog* ascribes authorship to Joshua Bayes. However, Thomas Bayes was probably the author of this work. Richard Price, a friend of Bayes, refers to the book in his own work *A Review of the Principal Questions in Morals*, and says that it was written by Thomas Bayes.

Bayes Theorem

Let A and B be two events with respective probabilities $P(A)$ and $P(B)$. The General Multiplication Rule for Probabilities gives

$$P(A \cap B) = P(B|A)P(A), \tag{5}$$

and also

$$P(A \cap B) = P(A|B)P(B). \tag{6}$$

Since the left-hand sides of Equations (5) and (6) are the same, so must the fight hand sides, so that $P(B|A)P(A) = P(A|B)P(B)$. Dividing through this equation by $P(A)$, assuming it is not zero, gives Bayes theorem: For any two events A and B where $P(A) > 0$,

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}. \tag{7}$$

Subjective Probability Interpretation

The most interesting interpretation of Bayes' theorem is in terms of subjective probabilities.

- Suppose an individual is interested in the event B and forms a subjective view of the probability that B will occur; in this context $P(B)$ is called a **prior** probability.
- If the individual acquires an additional piece of information – namely, that event A has occurred – this may cause a modification of the initial judgement as to the **likelihood** of the occurrence of B.
- Since A is known to have happened, the relevant probability for B is now $P(B|A)$, which is referred to as the **posterior** probability.
- Thus Bayes' theorem can be thought as a mechanism to update a prior probability to a posterior probability when the additional information that event A has occurred becomes available. The mechanism is by a multiplication of $P(B)$; specifically, by $P(A|B)/P(B)$.

Extending Bayes Theorem

Bayes theorem is often expressed in a different but equivalent form. Let $E_1, E_2, \ldots, E_K$ be K mutually exclusive and collectively exhaustive events, and let A be some other event.

For some i, we want to find $P(E_i|A)$. This can be obtained directly by Bayes' theorem by setting B in Equation (7) equal to E.

However, the denominator on the right-hand side of that equation can be expressed in terms of conditional probabilities for A given the E_j and probabilities of each E_j. Thus it follows that

$$P(A) = \sum_{k=1}^{K} P(E_k \cap A). \tag{8}$$

An Alternative Statement of Bayes Theorem

Furthermore using the General Multiplication Rule for Probabilities, $P(E_j \cap A) = P(A|E_j)P(E_j)$, $j \in \{1, 2, \ldots, K\}$, we have

$$P(A) = \sum_{k=1}^{K} P(A|E_k)P(E_k). \tag{9}$$

Finally, the restatement of Bayes' theorem is obtained by substituting E_i for B and the right-hand side of Equation (9) for $P(A)$ in Equation (7). Thus assuming at least one $P(E_j) > 0$, $j \in \{1, 2, \ldots, K\}$, the alternative statement of Bayes' theorem is

$$P(E_i|A) = \frac{P(A|E_i)P(E_i)}{\sum_{k=1}^{K} P(A|E_k)P(E_k)}. \tag{10}$$

Spam Emails

Spamlaws estimates that 45% of emails are spam emails. Many email service providers deploy software to filter spam emails before they reach an inbox. A particular software solution claims that it can detect 99% of spam emails, and the probability for a false positive, a non-spam email detected by the software as spam, is 5%.

If an email is detected as spam, what is the probability that it is truly a non-spam email?

Spam Emails Continued

Solution:

First, the following events are defined.

$$A = \text{event that an email is detected as spam,}$$
$$B = \text{event that an email is spam,}$$
$$B^c = \text{event that an email is not spam.}$$

From above we know, $P(B) = 0.45$, $P(B^c) = 0.55$, $P(A|B) = 0.99$, and $P(A|B^c) = 0.05$. By Bayes' theorem,

$$\begin{aligned} P(B^c|A) &= \frac{P(A|B^c)P(B^c)}{P(A|B)P(B) + P(A|B^c)P(B^c)}, \\ &= \frac{0.05 \times 0.55}{(0.99 \times 0.45) + (0.05 \times 0.55)}, \\ &= \frac{0.028}{0.028 + 0.446}, \\ &= 0.058. \end{aligned}$$

References

Abdelrazeq, I., Al-Labadi, L., and Alzaatreh, A. (2020). On one-sample bayesian tests for the mean. *Statistics*, 54(2):424–440.

Al-Labadi, L., Cheng, Y., Fazeli-Asl, F., Lim, K., and Weng, Y. (2022). A bayesian one-sample test for proportion. *Stats*, 5(4):1242–1253.

Angrist, J. D. and Pischke, J. S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion* . Number 8769 in Economics Books. Princeton University Press, Princeton, NJ.

Calder, B. J. (1977). Focus groups and the nature of qualitative marketing research. *Journal of Marketing Research*, 14(3):353–364.

Davis, B. (2022). *Marketing Analytics*. Edify Publ., ISBN: 978-1-7346888-4-9.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman and Hall, New York, NY.

Ghosh, A., Nashaat, M., Miller, J., Quader, S., and Marston, C. (2018). A comprehensive review of tools for exploratory analysis of tabular industrial datasets. *Visual Informatics*, 2(4):235–253.

Gönen, M., Johnson, W. O., Lu, Y., and Westfall, P. H. (2005). The bayesian two-sample t test. *The American Statistician*, 59(3):252–257.

Gordon, B. R., Zettelmeyer, F., Bhargava, N., and Chapsky, D. (2019). A comparison of approaches to advertising measurement: Evidence from big field experiments at Facebook. *Marketing Science*, 38(2):193–225.

Gronau, Q. F., Ly, A., and Wagenmakers, E.-J. (2020). Informed bayesian t-tests. *The American Statistician*, 74(2):137–143.

Hernán, M. A. and Robin, J. M. (2020). *Causal Inference: What If*. Chapman & Hall/CRC Press, Boca Raton, FL.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960.

Kohavi, R. and Longbotham, R. (2011). Unexpected results in online controlled experiments. *SIGKDD Explor. Newsl.*, 12(2):31–35.

Kohavi, R. and Longbotham, R. (2017). Online controlled experiments and A/B testing. In Sammut, C. and Webb, G. I., editors, *Encyclopedia of Machine Learning and Data Mining*, pages 922–929. Springer US, Boston, MA.

Kohavi, R., Longbotham, R., Sommerfield, D., and Henne, R. M. (2009). Controlled experiments on the web: Survey and practical guide. *Data Mining and Knowledge Discovery*, 18(1):140–181.

Pearl, J. (2009). *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York, NY, 2nd edition.

Pearl, J. and Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. Basic Books, Inc., New York, NY, 1st edition.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701.

Shao, J. and Tu, D. (1996). *The Jackknife and Bootstrap*. Springer, New York, NY.

Taddy, M., Hendrix, L., and Harding, M. C. (2023). *Modern Business Analytics*. McGraw Hill, New York, NY.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading, MA.

Wang, M. and Liu, G. (2016). A simple two-sample bayesian t-test for hypothesis testing. *The American Statistician*, 70(2):195–201.

Wasserstein, R. L., Schirm, A. L., and Lazar, N. A. (2019). Moving to a world beyond "$p < 0.05$". *The American Statistician*, 73(sup1):1–19.