

Statistical Machine Learning for Data Science- BAD702

**Prepared By,
Dr. Anitha DB
Associate Professor & Head
Department of CSE-Data Science
ATME College of Engineering, Mysuru**

[GitHub - gedeck/practical-statistics-for-data-scientists: Code repository for O'Reilly book](https://github.com/gedeck/practical-statistics-for-data-scientists)

Module-1

Exploratory Data Analysis: Estimates of locations and variability, Exploring data distributions, Exploring binary and categorical data, Exploring two or more variables.

Textbook: Chapter 1

Module-2

Data and Sampling Distributions: Random sampling and bias, selection bias, sampling distribution of statistic, bootstrap, confidence intervals, data distributions: normal, long tailed, student's-t, binomial, Chi-square, F distribution, Poisson and related distributions.

Textbook: Chapter 2

Module-3

Statistical Experiments and Significance Testing: A/B testing, hypothesis testing, resampling, statistical significance & p-values, t-tests, multiple testing, degrees of freedom.

Textbook: Chapter 3

Module-4

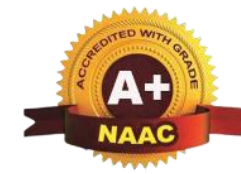
Multi-Arm Bandit algorithm, power and sample size, factor variables in regression, interpreting the regression equation, Regression diagnostics, Polynomial and Spline Regression.

Textbook: Chapter 3 & 4

Module-5

Discriminant Analysis: Covariance Matrix, Fisher's Linear discriminant, Generalized Linear Models, Interpreting the coefficients and odd ratios, Strategies for Imbalanced Data.

Textbook: Chapter 5



Topics

1. **A/B testing:** Why Have a Control Group?, Why Just A/B? Why Not C, D...?
2. **Hypothesis testing:** The Null Hypothesis, Alternative Hypothesis, One-Way, Two-Way Hypothesis Test
3. **Resampling:** Permutation Test, Example: Web Stickiness, Exhaustive and Bootstrap Permutation Test, Permutation Tests: The Bottom Line for Data Science
4. **Statistical significance & p-values:** P-Value, Alpha, Type 1 and Type 2 Errors, Data Science and P-Values
5. **T-tests:** Test statistic, t-statistic, t-distribution
6. **Multiple testing:** Type 1 error, False discovery rate, Adjustment of p-values, Overfitting
7. **Degrees of freedom.**

Textbook: Peter Bruce, Andrew Bruce and Peter Gadeck, “Practical Statistics for Data Scientists”, 2nd edition, O’Reilly Publications, 2020. Chapter 3

[GitHub - gedeck/practical-statistics-for-data-scientists: Code repository for O'Reilly book](https://github.com/gedeck/practical-statistics-for-data-scientists)

Design of experiments is a cornerstone of the practice of statistics. The goal is to design an experiment in order to confirm or reject a hypothesis. This chapter reviews traditional experimental design and discusses some common challenges in data science. It also covers some concepts in statistical inference and explains their meaning and relevance (or lack of relevance) to data science.

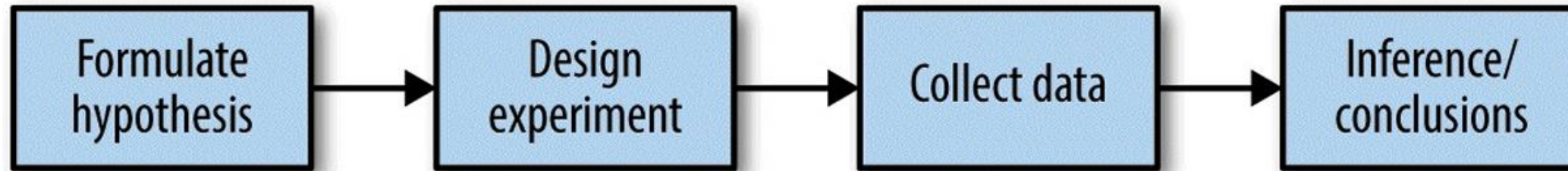


Figure 3-1. The classical statistical inference pipeline

This process starts with a hypothesis (“drug A is better than the existing standard drug,” “price A is more profitable than the existing price B”). An experiment (it might be an A/B test) is designed to test the hypothesis — designed in such a way that, hopefully, will deliver conclusive results. The data is collected and analyzed, and then a conclusion is drawn. The term inference reflects the intention to apply the experiment results, which involve a limited set of data, to a larger process or population

A/B testing

An A/B test is an experiment with two groups to establish which of two treatments, products, procedures, or the like is superior. Often one of the two treatments is the standard existing treatment, or no treatment. If a standard (or no) treatment is used, it is called the *control*. A typical hypothesis is that treatment is better than *control*.

KEY TERMS FOR A/B TESTING

1. **Treatment** :Something (drug, price, web headline) to which a subject is exposed.
2. **Treatment group** : A group of subjects exposed to a specific treatment.
3. **Control group** : A group of subjects exposed to no (or standard) treatment.
4. **Randomization** : The process of randomly assigning subjects to treatments.
5. **Subjects** : The items (web visitors, patients, etc.) that are exposed to treatments.
6. **Test statistic**: The metric used to measure the effect of the treatment.

A/B testing

A/B tests are common in web design and marketing, since results are so readily measured. Some examples of A/B testing include:

- Testing two soil treatments to determine which produces better seed germination
- Testing two therapies to determine which suppresses cancer more effectively
- Testing two prices to determine which yields more net profit
- Testing two web headlines to determine which produces more clicks (Figure 3-2)
- Testing two web ads to determine which generates more conversions

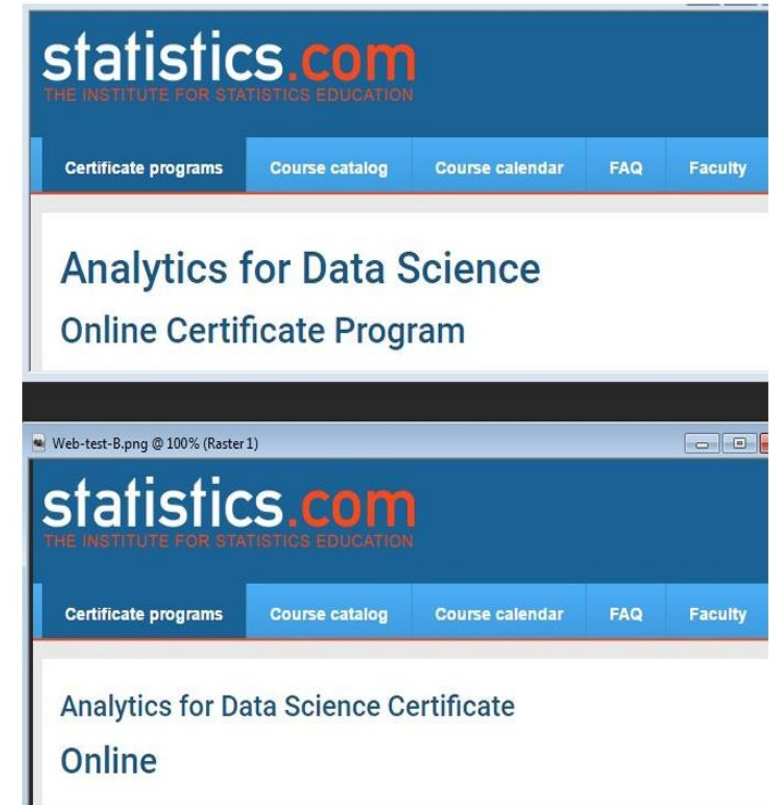


Figure 3-2. Marketers continually test one web presentation against another

A/B testing

A proper A/B test has subjects that can be assigned to one treatment or another. The subject might be a person, a plant seed, a web visitor; the key is that the subject is exposed to the treatment. Ideally, subjects are *randomized* (assigned randomly) to treatments. In this way, you know that any difference between the treatment groups is due to one of two things:

- The effect of the different treatments
- Luck of the draw in which subjects are assigned to which treatments (i.e., the random assignment may have resulted in the naturally better-performing subjects being concentrated in A or B)

You also need to pay attention to the test statistic or metric you use to compare group A to group B. Perhaps the most common metric in data science is a binary variable: click or no-click, buy or don't buy, fraud or no fraud, and so on. Those results would be summed up in a 2×2 table. Table 3-1 is a 2×2 table for an actual price test

Table 3-1. 2×2 table for
ecommerce experiment
results

Outcome	Price A	Price B
Conversion	200	182
No conversion	23,539	22,406

If the metric is a continuous variable (purchase amount, profit, etc.), or a count (e.g., days in hospital, pages visited) the result might be displayed differently. If one were interested not in conversion, but in revenue per page view, the results of the price test in Table 3-1 might look like this in typical default software output:

Revenue/page-view with price A: mean = 3.87, SD = 51.10

Revenue/page-view with price B: mean = 4.11, SD = 62.98

“SD” refers to the standard deviation of the values within each group

Why Have a Control Group?

Why not just apply the treatment to one group and compare results to past experience?

- **Problem with no control group:**

Without a control group, we can't be sure that any observed changes are truly due to the treatment. Other variables may have changed as well.

- **Value of a control group:**

A control group ensures that **all conditions are the same** for both groups *except* for the treatment. This allows for a valid comparison and helps isolate the treatment effect.

- **Comparing to past data is risky:**

If you compare to past experiences instead of using a control group, unknown or uncontrollable factors might influence the outcome.

A control group helps eliminate confounding variables, making your experimental results more reliable and valid.

Why Have a Control Group?

The use of A/B testing in data science is typically in a web context.

Treatments might be

- The design of a web page,
- The price of a product,
- The wording of a headline, or some other item.

Typically the **subject** in the experiment is the web visitor,

The **outcomes** we are interested in measuring are

- Clicks,
- Purchases,
- Visit duration,
- Number of pages visited,
- Whether a particular page is visited.

In a standard A/B experiment, we need to decide on one metric ahead of time.

Multiple behavior metrics might be collected and be of interest, but if the experiment is expected to lead to a decision between treatment A and treatment B, a single metric, or test statistic, needs to be established beforehand.

Selecting a test statistic after the experiment is conducted opens the door to researcher bias.

Why Just A/B? Why Not C, D...?

A/B tests are popular in the marketing and ecommerce worlds, but are far from the only type of statistical experiment. Additional treatments can be included. Subjects might have repeated measurements taken. Pharmaceutical trials where subjects are scarce, expensive, and acquired over time are sometimes designed with multiple opportunities to stop the experiment and reach a conclusion.

Traditional statistical experimental designs focus on answering a static question about the efficacy of specified treatments. Data scientists are less interested in the question:

Is the difference between price A and price B statistically significant?

than in the question:

Which, out of multiple possible prices, is best?

For this, a relatively new type of experimental design used is : the **multi-arm bandit** algorithm

Hypothesis Tests

Hypothesis tests, also known as **significance tests**, are fundamental tools in traditional statistical analysis. They help researchers determine whether the results they observe in their data are likely due to **random chance** or if they reflect a **real underlying effect**.

KEY TERMS

Null hypothesis : The hypothesis that chance is to blame.

Alternative hypothesis : Counterpoint to the null (what you hope to prove).

One-way test: Hypothesis test that counts chance results only in one direction.

Two-way test: Hypothesis test that counts chance results in two directions.

An A/B test is typically constructed with a hypothesis in mind. For example, the hypothesis might be that price B produces higher profit.

Why do we need a hypothesis? Why not just look at the outcome of the experiment and go with whichever treatment does better?

The answer lies in the tendency of the human mind to underestimate the scope of natural random behavior.

- One manifestation of this is the failure to anticipate extreme events, or so-called “black swans” .
- Another manifestation is the tendency to misinterpret random events as having patterns of some significance.

Statistical hypothesis testing was invented as a way to protect researchers from being fooled by random chance.

In a properly designed A/B test, you collect data on treatments A and B in such a way that any observed difference between A and B must be due to either:

- Random chance in assignment of subjects
- A true difference between A and B

A statistical hypothesis test is further analysis of an A/B test, or any randomized experiment, to assess whether random chance is a reasonable explanation for the observed difference between groups A and B

The Null Hypothesis

The **null hypothesis** is a starting point that assumes: **There is no real difference** between the two treatments (e.g., Price A and Price B). Any observed difference in outcomes is due to **random variation**. This is the **default assumption**- It protects us from jumping to conclusions based on random noise.

Steps for testing the Null Hypothesis using Resampling Permutation

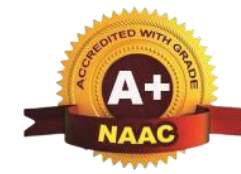
1. **Combine** all data from both groups A and B.
2. **Shuffle** the data randomly, breaking any link between group label and outcome (this simulates the null hypothesis being true—no difference between groups).
3. **Split** the shuffled data back into two groups of the same sizes as the original A and B.
4. **Calculate** the difference in outcome (e.g., average profit, click-through rate) between the new shuffled groups.
5. **Repeat** steps 2–4 many times (e.g., 10,000 times).
6. **Compare**: How often is the difference between these randomized groups as extreme as the difference observed in the actual A/B test?
7. If your actual result is **more extreme than what random chance produces 95% of the time**, you have grounds to **reject the null hypothesis** and conclude the treatments likely differ.

Alternative Hypothesis

Hypothesis tests by their nature involve not just a null hypothesis, but also an offsetting alternative hypothesis. Here are some examples:

- **Null** = “no difference between the means of group A and group B,” **alternative** = “A is different from B” (could be bigger or smaller)
- **Null** = “ $A \leq B$,” **alternative** = “ $B > A$ ”
- **Null** = “B is not X% greater than A,” **alternative** = “B is X% greater than A”

Taken together, the null and alternative hypotheses must account for all possibilities. The nature of the null hypothesis determines the structure of the hypothesis test



One-Way, Two-Way Hypothesis Test

Often, in an A/B test, you are testing a new option (say B), against an established default option (A) and the presumption is that you will stick with the default option unless the new option proves itself definitively better. In such a case, you want a hypothesis test to protect you from being fooled by chance in the direction favoring B. You don't care about being fooled by chance in the other direction, because you would be sticking with A unless B proves definitively better. So you want a directional alternative hypothesis (B is better than A). In such a case, you use a one-way (or one-tail) hypothesis test. This means that extreme chance results in only one direction count toward the p-value.

If you want a hypothesis test to protect you from being fooled by chance in either direction, the alternative hypothesis is bidirectional (A is different from B; could be bigger or smaller). In such a case, you use a two-way (or two-tail) hypothesis. This means that extreme chance results in either direction count toward the p value.

A one-tail hypothesis test often fits the nature of A/B decision making, in which a decision is required and one option is typically assigned "default" status unless the other proves better. Software, however, including R, typically provides a two tail test in its default output, and many statisticians opt for the more conservative two-tail test just to avoid argument.

Resampling

Resampling in statistics means to repeatedly sample values from observed data, with a general goal of assessing random variability in a statistic. It can also be used to assess and improve the accuracy of some machine-learning models (e.g., the predictions from decision tree models built on multiple bootstrapped data sets can be averaged in a process known as bagging). There are **two main types** of resampling procedures: the **bootstrap and permutation** tests. The **bootstrap** is used to assess the reliability of an estimate. **Permutation** tests are used to test hypotheses, typically involving two or more groups.

KEY TERMS

- **Permutation test** : The procedure of combining two or more samples together, and randomly (or exhaustively) re-locating the observations to resamples. Synonyms → Randomization test, random permutation test, exact test.
- **With or without replacement**: In sampling, whether or not an item is returned to the sample before the next draw

Permutation Test

In a permutation procedure, two or more samples are involved, typically the groups in an A/B or other hypothesis test. Permute means to change the order of a set of values. The first step in a permutation test of a hypothesis is to combine the results from groups A and B (and, if used, C, D, ...) together. This is the logical embodiment of the null hypothesis that the treatments to which the groups were exposed do not differ. We then test that hypothesis by randomly drawing groups from this combined set, and seeing how much they differ from one another.

The permutation procedure is as follows:

1. Combine the results from the different groups in a single data set.
2. Shuffle the combined data, then randomly draw (without replacing) a resample of the same size as group A.
3. From the remaining data, randomly draw (without replacing) a resample of the same size as group B.
4. Do the same for groups C, D, and so on.
5. Whatever statistic or estimate was calculated for the original samples (e.g., difference in group proportions), calculate it now for the resamples, and record; this constitutes one permutation iteration.
6. Repeat the previous steps R times to yield a permutation distribution of the test statistic.

Now go back to the observed difference between groups and compare it to the set of permuted differences.

- If the observed difference lies well within the set of permuted differences, then we have not proven anything — the observed difference is within the range of what chance might produce.
- However, if the observed difference lies outside most of the permutation distribution, then we conclude that chance is not responsible. In technical terms, the difference is statistically significant.

Example: Web Stickiness

A company selling a relatively high-value service wants to test which of two web presentations does a better selling job. Due to the high value of the service being sold, sales are infrequent and the sales cycle is lengthy; it would take too long to accumulate enough sales to know which presentation is superior. So the company decides to measure the results with a proxy variable, using the detailed interior page that describes the service.

One potential proxy variable for our company is the number of **clicks** on the detailed landing page. A better one is how long people spend on the page. It is reasonable to think that a web presentation (page) that holds people's attention longer will lead to more sales. Hence, our metric is average session time, comparing page A to page B. Due to the fact that this is an interior, special-purpose page, it does not receive a huge number of visitors. Also note that Google Analytics, which is how we measure session time, cannot measure session time for the last session a person visits. Instead of deleting that session from the data, though, GA records it as a zero, so the data requires additional processing to remove those sessions. The result is a total of 36 sessions for the two different presentations, 21 for page A and 15 for page B. Using ggplot, we can visually compare the session times using side-by-side boxplots

```
ggplot(session_times, aes(x=Page, y=Time)) + geom_boxplot()
```

Example: Web Stickiness

The boxplot, shown in Figure 3-3, indicates that page B leads to longer sessions than page A. The means for each group can be computed as follows:

```
mean_a <- mean(session_times[session_times['Page']=='Page A', 'Time'])
mean_b <- mean(session_times[session_times['Page']=='Page B', 'Time'])
mean_b - mean_a
[1] 21.4
```

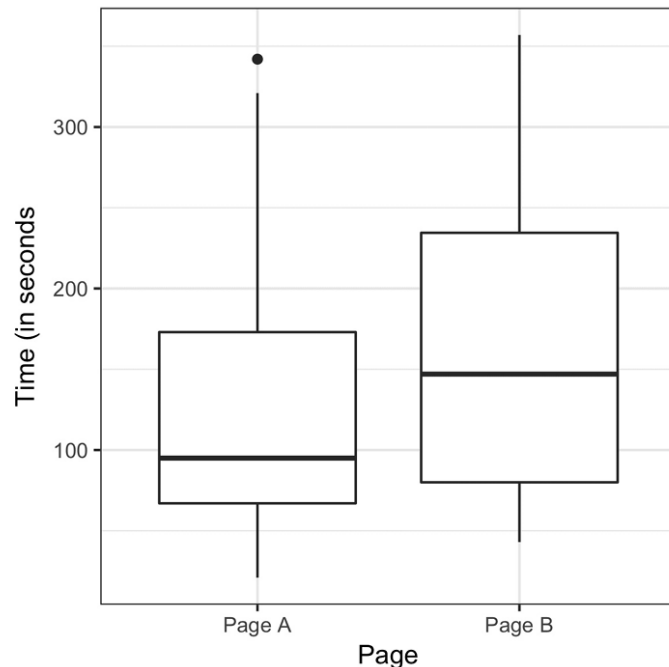


Figure 3-3. Session times for web pages A and B

Page B has session times greater, on average, by 21.4 seconds versus page A. The question is whether this difference is within the range of what random chance might produce, or, alternatively, is statistically significant. One way to answer this is to apply a permutation test — combine all the session times together, then repeatedly shuffle and divide them into groups of 21 (recall that $n = 21$ for page A) and 15 ($n = 15$ for B). To apply a permutation test, we need a function to randomly assign the 36 session times to a group of 21 (page A) and a group of 15 (page B):

```
perm_fun <- function(x, n1, n2)
{
  n <- n1 + n2
  idx_b <- sample(1:n, n1)
  idx_a <- setdiff(1:n, idx_b)
  mean_diff <- mean(x[idx_b]) - mean(x[idx_a])
  return(mean_diff)
}
```

Example: Web Stickiness

This function works by sampling without replacement n_2 indices and assigning them to the B group; the remaining n_1 indices are assigned to group A. The difference between the two means is returned. Calling this function $R = 1,000$ times and specifying $n_2 = 15$ and $n_1 = 21$ leads to a distribution of differences in the session times that can be plotted as a histogram.

The histogram, shown in Figure 3-4 shows that mean difference of random permutations often exceeds the observed difference in session times (the vertical line). This suggests that the observed difference in session time between page A and page B is well within the range of chance variation, thus is not statistically significant.

```
perm_diffs <- rep(0, 1000)
for(i in 1:1000)
  perm_diffs[i] = perm_fun(session_times[, 'Time'], 21, 15)

hist(perm_diffs, xlab='Session time differences (in seconds)')
abline(v = mean_b - mean_a)
```

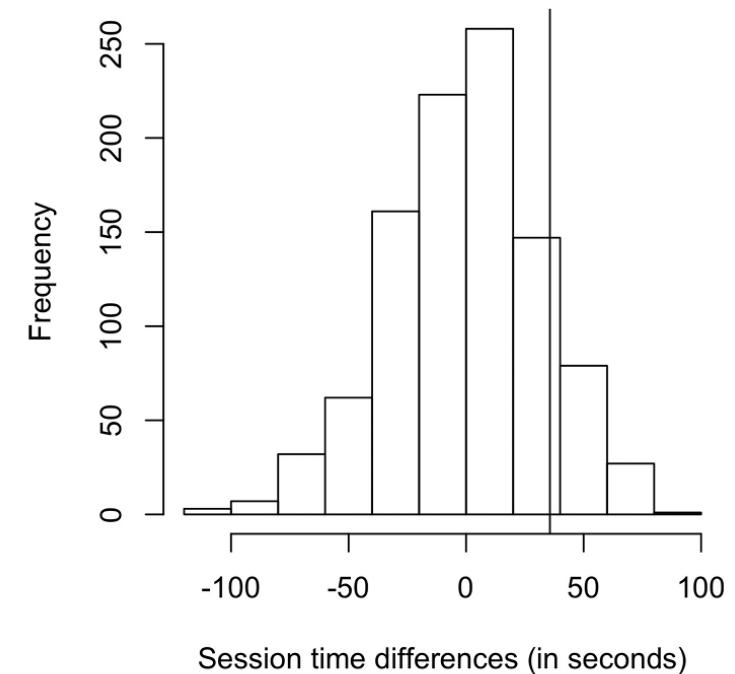


Figure 3-4. Frequency distribution for session time differences between pages A and B

Exhaustive and Bootstrap Permutation Test

In addition to the preceding random shuffling procedure, also called a random permutation test or a randomization test, there are two variants of the permutation test:

- An exhaustive permutation test
- A bootstrap permutation test

In an **exhaustive permutation test**, instead of just randomly shuffling and dividing the data, we actually figure out all the possible ways it could be divided. This is practical only for relatively small sample sizes. With a large number of repeated shuffling's, the random permutation test results approximate those of the exhaustive permutation test, and approach them in the limit. Exhaustive permutation tests are also sometimes called exact tests, due to their statistical property of guaranteeing that the null model will not test as “significant” more than the alpha level of the test .

In a **bootstrap permutation test**, the draws outlined in steps 2 and 3 of the random permutation test are made with replacement instead of **without replacement**. In this way the resampling procedure models not just the random element in the assignment of treatment to subject, but also the random element in the selection of subjects from a population.

Both procedures are encountered in statistics, and the distinction between them is somewhat convoluted and not of consequence in the practice of data science

Permutation Tests: The Bottom Line for Data Science

Permutation tests are useful heuristic procedures for exploring the role of random variation. They are relatively easy to code, interpret and explain, and they offer a useful detour around the formalism and “false determinism” of formula-based statistics.

One virtue of resampling, in contrast to formula approaches, is that it comes much closer to a “one size fits all” approach to inference. Data can be numeric or binary. Sample sizes can be the same or different. Assumptions about normally distributed data are not needed.

KEY IDEAS

- In a permutation test, multiple samples are combined, then shuffled.
- The shuffled values are then divided into resamples, and the statistic of interest is calculated.
- This process is then repeated, and the resampled statistic is tabulated.
- Comparing the observed value of the statistic to the resampled distribution allows you to judge whether an observed difference between samples might occur by chance

Statistical Significance and P-Values

Statistical significance is how statisticians measure whether an experiment (or even a study of existing data) yields a result more extreme than what chance might produce. If the result is beyond the realm of chance variation, it is said to be statistically significant.

KEY TERMS

- **P-value:** Given a chance model that embodies the nul hypothesis, the p-value is the probability of obtaining results as unusual or extreme as the observed results.
- **Alpha :** The probability threshold of “unusualness” that chance results must surpass, for actual outcomes to be deemed statistically significant
- **Type 1 error :** Mistakenly concluding an effect is real (when it is due to chance).
- **Type 2 error:** Mistakenly concluding an effect is due to chance (when it is real)

Consider in Table 3-2 the results of the web test

Price A converts almost 5% better than price B (0.8425% versus 0.8057% — a difference of 0.0368 percentage points), big enough to be meaningful in a high volume business. We have over 45,000 data points here, and it is tempting to consider this as “big data,” not requiring tests of statistical significance (needed mainly to account for sampling variability in small samples). However, the conversion rates are so low (less than 1%) that the actual meaningful values — the conversions — are only in the 100s, and the sample size needed is really determined by these conversions. We can test whether the difference in conversions between prices A and B is within the range of chance variation, using a resampling procedure. By “chance variation,” we mean the random variation produced by a probability model that embodies the null hypothesis that there is no difference between the rates .

Table 3-2. 2×2 table for ecommerce experiment results

Outcome	Price A	Price B
Conversion	200	182
No conversion	23539	22406

The following permutation procedure asks “if the two prices share the same conversion rate, could chance variation produce a difference as big as 5%?”

1. Create an urn with all sample results: this represents the supposed shared conversion rate of 382 ones and 45,945 zeros = $0.008246 = 0.8246\%$.
2. Shuffle and draw out a resample of size 23,739 (same n as price A), and record how many 1s.
3. Record the number of 1s in the remaining 22,588 (same n as price B).
4. Record the difference in proportion 1s.
5. Repeat steps 2–4.
6. How often was the difference ≥ 0.0368 ? in conversion rate:

Reusing the function `perm_fun` defined in “Example: Web Stickiness”, we can create a histogram of randomly permuted differences.

See the histogram of 1,000 resampled results in Figure 3-5: as it happens, in this case the observed difference of 0.0368% is well within the range of chance variation

```
obs_pct_diff <- 100*(200/23739 - 182/22588)
conversion <- c(rep(0, 45945), rep(1, 382))
perm_diffs <- rep(0, 1000)
for(i in 1:1000)
  perm_diffs[i] = 100*perm_fun(conversion, 23739, 22588 )
hist(perm_diffs, xlab='Session time differences (in seconds)')
abline(v = obs_pct_diff)
```

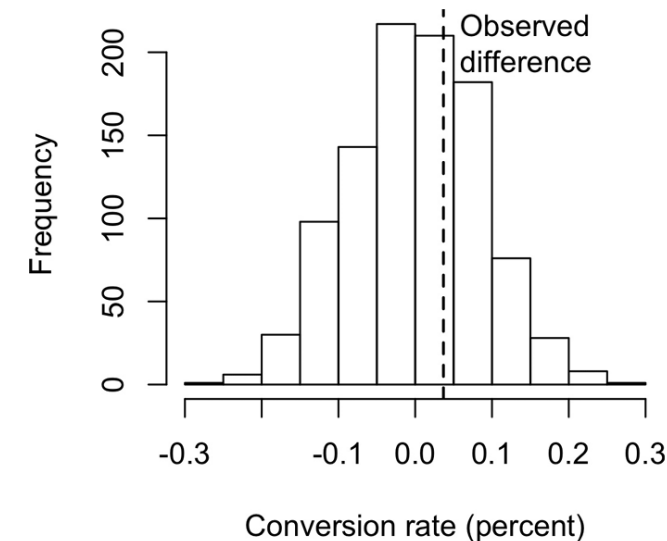


Figure 3-5. Frequency distribution for the difference in conversion rates between pages A and B

P-Value

Simply looking at the graph is not a very precise way to measure statistical significance, so of more interest is the p-value. This is the frequency with which the chance model produces a result more extreme than the observed result. We can estimate a p-value from our permutation test by taking the proportion of times that the permutation test produces a difference equal to or greater than the observed difference

```
mean(perm_diffs > obs_pct_diff)
[1] 0.308
```

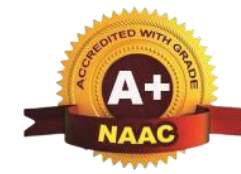
The p-value is 0.308, which means that we would expect to achieve the same result by random chance over 30% of the time. In this case, we didn't need to use a permutation test to get a p-value. Since we have a binomial distribution, we can approximate the p-value using the normal distribution. In R code, we do this using the function `prop.test`

```
> prop.test(x=c(200,182), n=c(23739,22588), alternative="greater")

2-sample test for equality of proportions with continuity correction

data:  c(200, 182) out of c(23739, 22588)
X-squared = 0.14893, df = 1, p-value = 0.3498
alternative hypothesis: greater
95 percent confidence interval:
 -0.001057439  1.000000000
sample estimates:
 prop 1      prop 2 
0.008424955 0.008057376
```

The argument `x` is the number of successes for each group and the argument `n` is the number of trials. The normal approximation yields a p-value of 0.3498, which is close to the p-value obtained from the permutation test



Alpha

Statisticians frown on the practice of leaving it to the researcher's discretion to determine whether a result is “too unusual” to happen by chance. Rather, a threshold is specified in advance, as in “more extreme than 5% of the chance (null hypothesis) results”; this threshold is known as alpha. Typical alpha levels are 5% and 1%. Any chosen level is an arbitrary decision — there is nothing about the process that will guarantee correct decisions $x\%$ of the time. This is because the probability question being answered is not “what is the probability that this happened by chance?” but rather “given a chance model, what is the probability of a result this extreme?” We then deduce backward about the appropriateness of the chance model, but that judgment does not carry a probability. This point has been the subject of much confusion

Value of the p-value

The American Statistical Association (ASA) statement stressed six principles for researchers and journal editors:

1. P-values can indicate how incompatible the data are with a specified statistical model.
2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
4. Proper inference requires full reporting and transparency.
5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis

Type 1 and Type 2 Errors

In assessing statistical significance, two types of error are possible:

- Type 1 error, in which you mistakenly conclude an effect is real, when it is really just due to chance
- Type 2 error, in which you mistakenly conclude that an effect is not real (i.e., due to chance), when it really is real

Actually, a Type 2 error is not so much an error as a judgment that the sample size is too small to detect the effect. When a p-value falls short of statistical significance (e.g., it exceeds 5%), what we are really saying is “effect not proven.” It could be that a larger sample would yield a smaller p-value.

The basic function of significance tests (also called hypothesis tests) is to protect against being fooled by random chance; thus they are typically structured to minimize Type 1 errors

Data Science and P-Values

The work that data scientists do is typically not destined for publication in scientific journals, so the debate over the value of a p-value is somewhat academic. For a data scientist, a p-value is a useful metric in situations where you want to know whether a model result that appears interesting and useful is within the range of normal chance variability. As a decision tool in an experiment, a p value should not be considered controlling, but merely another point of information bearing on a decision. For example, p-values are sometimes used as intermediate inputs in some statistical or machine learning models — a feature might be included in or excluded from a model depending on its p-value

KEY IDEAS

- Significance tests are used to determine whether an observed effect is within the range of chance variation for a null hypothesis model.
- The p-value is the probability that results as extreme as the observed results might occur, given a null hypothesis model.
- The alpha value is the threshold of “unusualness” in a null hypothesis chance model.
- Significance testing has been much more relevant for formal reporting of research than for data science (but has been fading recently, even for the former).

t-Tests

There are numerous types of significance tests, depending on whether the data comprises count data or measured data, how many samples there are, and what's being measured. A very common one is the t-test, named after Student's t-distribution, originally developed by W. S. Gossett to approximate the distribution of a single sample mean.

KEY TERMS

- **Test statistic** : A metric for the difference or effect of interest.
- **t-statistic** : A standardized version of the test statistic.
- **t-distribution**: A reference distribution (in this case derived from the nul hypothesis), to which the observed t-statistic can be compared

All significance tests require that you specify a test statistic to measure the effect you are interested in, and help you determine whether that observed effect lies within the range of normal chance variation.

In a resampling test, the scale of the data does not matter. You create the reference (null hypothesis) distribution from the data itself, and use the test statistic as is.

In the 1920s and 30s, when statistical hypothesis testing was being developed, it was not feasible to randomly shuffle data thousands of times to do a resampling test. Statisticians found that a good approximation to the permutation (shuffled) distribution was the t-test, based on Gossett's t-distribution. It is used for the very common two-sample comparison — A/B test — in which the data is numeric.

But in order for the t-distribution to be used without regard to scale, a standardized form of the test statistic must be used. A classic statistics text would at this stage show various formulas that incorporate Gossett's distribution and demonstrate how to standardize your data to compare it to the standard t-distribution. In R, the function is `t.test`

```
> t.test(Time ~ Page, data=session_times, alternative='less' )
Welch Two Sample t-test

data:  Time by Page
t = -1.0983, df = 27.693, p-value = 0.1408
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
 -Inf 19.59674
sample estimates:
mean in group Page A mean in group Page B
126.3333          162.0000
```

The alternative hypothesis is that the session time mean for page A is less than for page B. This is fairly close to the permutation test p-value of 0.124.

In a resampling mode, we structure the solution to reflect the observed data and the hypothesis to be tested, not worrying about whether the data is numeric or binary, sample sizes are balanced or not, sample variances, or a variety of other factors.

KEY IDEAS

- Before the advent of computers, resampling tests were not practical and statisticians used standard reference distributions.
- A test statistic could then be standardized and compared to the reference distribution.
- One such widely used standardized statistic is the t-statistic.

Multiple Testing

“torture the data long enough, and it will confess.” → This means that if you look at the data through enough different perspectives, and ask enough questions, you can almost invariably find a statistically significant effect.

KEY TERMS

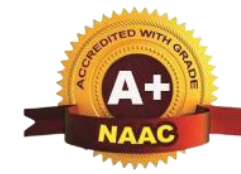
- **Type 1 error** : Mistakenly concluding that an effect is statistically significant.
- **False discovery rate**: Across multiple tests, the rate of making a Type 1 error.
- **Adjustment of p-values** :Accounting for doing multiple tests on the same data.
- **Overfitting**: Fitting the noise

For example, if you have 20 predictor variables and one outcome variable, all randomly generated, the odds are pretty good that at least one predictor will (falsely) turn out to be statistically significant if you do a series of 20 significance tests at the $\alpha = 0.05$ level. This is called a Type 1 error.

We can calculate this probability by first finding the probability that all will correctly test nonsignificant at the 0.05 level. The probability that one will correctly test nonsignificant is 0.95, so the probability that all 20 will correctly test nonsignificant is $0.95 \times 0.95 \times 0.95 \dots$ or $0.95^{20} = 0.36$.¹

The probability that at least one predictor will (falsely) test significant is the flip side of this probability, or $1 - (\text{probability that all will be nonsignificant}) = 0.64$. This issue is related to the problem of overfitting in data mining, or “fitting the model to the noise.”

The more variables you add, or the more models you run, the greater the probability that something will emerge as “significant” just by chance. In **supervised learning** tasks, a holdout set where models are assessed on data that the model has not seen before mitigates this risk. In **statistical and machine learning** tasks not involving a labeled holdout set, the risk of reaching conclusions based on statistical noise persists.



In statistics, there are some procedures intended to deal with this problem in very specific circumstances. For example, if you are comparing results across multiple treatment groups you might ask multiple questions. So, for treatments A–C, you might ask:

Is A different from B?

Is B different from C?

Is A different from C?

Or, in a clinical trial, you might want to look at results from a therapy at multiple stages. In each case, you are asking multiple questions, and with each question, you are increasing the chance of being fooled by chance. Adjustment procedures in statistics can compensate for this by setting the bar for statistical significance more stringently than it would be set for a single hypothesis test. These adjustment procedures typically involve “dividing up the alpha” according to the number of tests. This results in a smaller alpha (i.e., a more stringent bar for statistical significance) for each test. One such procedure, the Bonferroni adjustment, simply divides the alpha by the number of observations n .

However, the problem of multiple comparisons goes beyond these highly structured cases and is related to the phenomenon of repeated data “dredging” that gives rise to the saying about torturing the data. Put another way, given sufficiently complex data, if you haven’t found something interesting, you simply haven’t looked long and hard enough. More data is available now than ever before, and the number of journal articles published nearly doubled between 2002 and 2010. This gives rise to lots of opportunities to find something interesting in the data, including multiplicity issues such as:

- Checking for multiple pairwise differences across groups
- Looking at multiple subgroup results (“we found no significant treatment effect overall, but we did find an effect for unmarried women younger than 30”)
- Trying lots of statistical models Including lots of variables in models
- Asking a number of different questions (i.e., different possible outcomes)

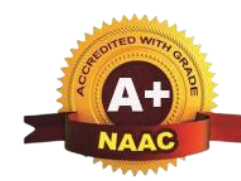
For a variety of reasons, including especially this general issue of “multiplicity,” more research does not necessarily mean better research. For example, the pharmaceutical company Bayer found in 2011 that when it tried to replicate 67 scientific studies, it could fully replicate only 14 of them. Nearly two-thirds could not be replicated at all

In any case, the **adjustment procedures** for highly defined and structured statistical tests are too specific and inflexible to be of general use to data scientists. The bottom line for data scientists on multiplicity is:

- For predictive modeling, the risk of getting an illusory model whose apparent efficacy is largely a product of random chance is mitigated by cross validation (see “Cross-Validation”), and use of a holdout sample.
- For other procedures without a labeled holdout set to check the model, you must rely on:
 - Awareness that the more you query and manipulate the data, the greater the role that chance might play; and
 - Resampling and simulation heuristics to provide random chance benchmarks against which observed results can be compared.

KEY IDEAS

- Multiplicity in a research study or data mining project (multiple comparisons, many variables, many models, etc.) increases the risk of concluding that something is significant just by chance.
- For situations involving multiple statistical comparisons (i.e., multiple tests of significance) there are statistical adjustment procedures.
- In a data mining situation, use of a holdout sample with labeled outcome variables can help avoid misleading results



Degrees of Freedom

KEY TERMS

n or sample size :The number of observations (also called rows or records) in the data.

d.f. :Degrees of freedom

The number of **degrees of freedom** is an input to many statistical tests. For example, degrees of freedom is the name given to the $n - 1$ denominator seen in the calculations for variance and standard deviation. Why does it matter? When you use a sample to estimate the variance for a population, you will end up with an estimate that is slightly biased downward if you use n in the denominator. If you use $n - 1$ in the denominator, the estimate will be free of that bias.

A large share of a traditional statistics course or text is consumed by various standard tests of hypotheses (t-test, F-test, etc.). When sample statistics are standardized for use in traditional statistical formulas, degrees of freedom is part of the standardization calculation to ensure that your standardized data matches the appropriate reference distribution (t-distribution, F-distribution, etc.).

Is it important for data science? Not really, at least in the context of significance testing. For one thing, formal statistical tests are used only sparingly in data science. For another, the data size is usually large enough that it rarely makes a real difference for a data scientist whether, for example, the denominator has n or $n - 1$.

Degrees of Freedom

There is one context, though, in which it is relevant: the use of factored variables in regression (including logistic regression). Regression algorithms choke if exactly redundant predictor variables are present. This most commonly occurs when factoring categorical variables into binary indicators (dummies). Consider day of week. Although there are seven days of the week, there are only six degrees of freedom in specifying day of week. For example, once you know that day of week is not Monday through Saturday, you know it must be Sunday. Inclusion of the Mon–Sat indicators thus means that also including Sunday would cause the regression to fail, due to a multicollinearity error.

KEY IDEAS

- The number of degrees of freedom (d.f.) forms part of the calculation to standardize test statistics so they can be compared to reference distributions (t-distribution, F-distribution, etc.).
- The concept of degrees of freedom lies behind the factoring of categorical variables into $n - 1$ indicator or dummy variables when doing a regression (to avoid multicollinearity).