Statistical Experiments and Significance Testing

Trevor Barnes

9/1/2020

Design of experiments is a cornerstone of the practice of statistics, with applications in virtually all areas of research. The goal is to design an experiment in order to confirm or reject a hypothesis. Data scientists are faced with the need to conduct continual experiments, particularly regarding user interface and product marketing.

Whenever you see references to statistical significance, t-tests, or p-values, it is typically in the context of 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.

Table 1: **KEY TERMS FOR A/B TESTING**

Term	Definition
Treatment	Something (drug, price, web headline) to which a subject is exposed.
Treatment Group	A group of subjects exposed to a specific treatment.
Control Group	A group of subjects exposed to no (or standard) treatment.
Randomization	The process of randomly assigning subjects to treatments.
Subjects	The items (web visitors, patients, etc.) that are exposed to treatments.
Test Statistic	The metric used to measure the effect of the treatment.

A/B tests are common in web design and marketing, since results are so readily measured. 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.

Why Have a Control Group?

Without a control group, there is no assurance that "other things are equal" and that any difference is really due to the treatment (or to chance). When you have a control group, it is subject to the same conditions (except for the treatment of interest) as the treatment group. If you simply make a comparison to "baseline" or prior experience, other factors, besides the treatment, might differ.

Key Ideas
Subjects are assigned to two (or more) groups that are treated exactly alike, except that the treatment under study differs from one to another. Ideally, subjects are assigned randomly to the groups.

Hypothesis Testing

Hypothesis tests, also called ***significance tests***, are ubiquitous in the traditional statistical analysis of published research. Their purpose is to help you learn whether random chance might be responsible for an observed effect.

Term	Definition
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. 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

This involves a baseline assumption that the treatments are equivalent, and any difference between the groups is due to chance. This baseline assumption is termed the ***null hypothesis***. Our hope is then that we can, in fact, prove the null hypothesis wrong, and show that the outcomes for groups A and B are more different than what chance might produce.

One way to do this is via a resampling permutation procedure, in which we shuffle together the results from groups A and B and then repeatedly deal out the data in groups of similar sizes, then observe how often we get a difference as extreme as the observed difference.

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 B," alternative = "B > A"
- Null = "B is not X% greater than A," alternative = "B is X% greater than A"

One-way, Two-way Hypothesis Test

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 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.

Key Ideas
A ***null hypothesis*** is a logical construct embodying the notion that nothing special has happened, and any effect you observe is due to random chance. The ***hypothesis test*** assumes that the null hypothesis is true, creates a "null model" (a probability model), and tests whether the effect you observe is a reasonable outcome of that model.

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; it was discussed in the previous chapter (see "The Bootstrap"). Permutation tests are used to test hypotheses, typically involving two or more groups, and we discuss those in this section.

Term	Definition	Synonym
Permutatition Test	The procedure of combining two or more samples together, and randomly (or exhaustively) reallocating the observations to resamples.	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

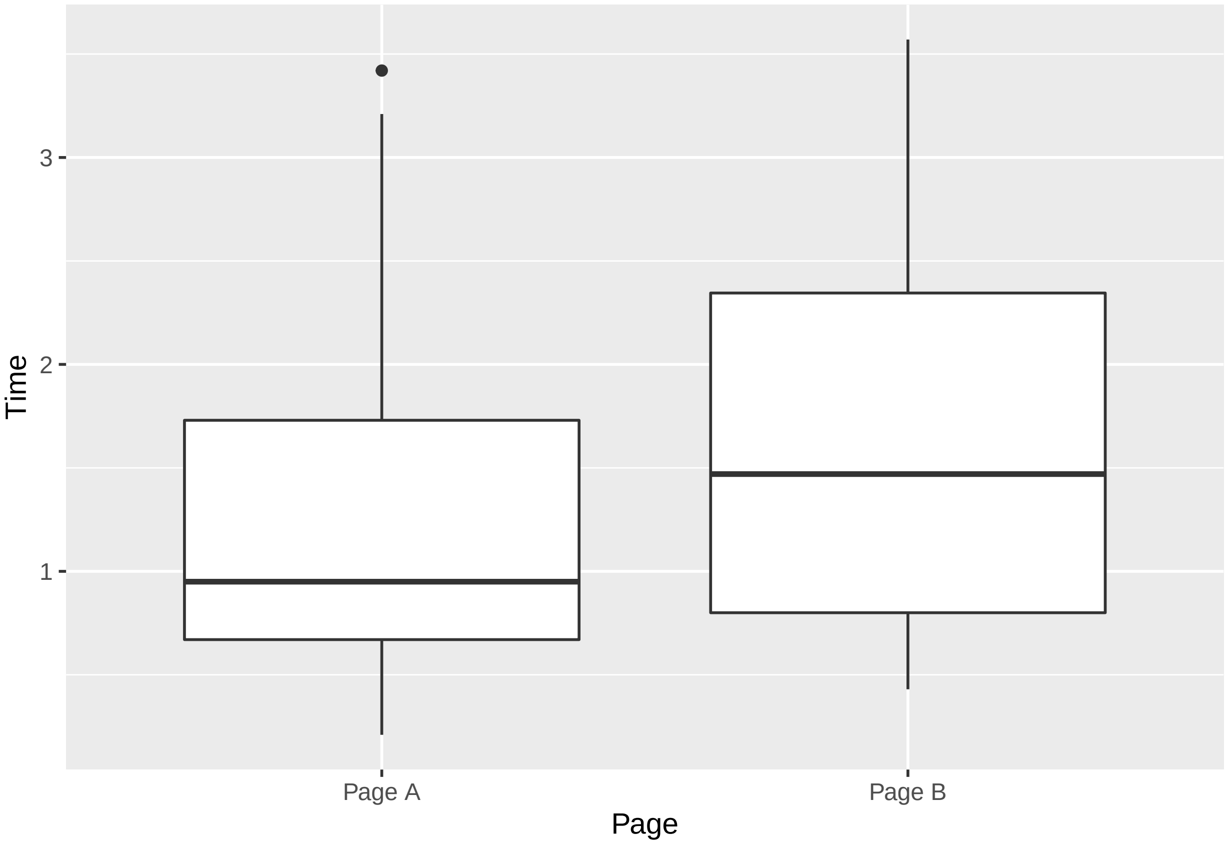
In a ***permutation*** procedure, two or more samples are involved, typically the groups in an A/B or other hypothesis test. ***Permute*** means to change the order of a set of values. The first step in a ***permutation test*** of a hypothesis is to combine the results from groups A and B (and, if used, 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 dataset.
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) are sample of the same size as group B.
4. Do the same for groups C, D, and soon.
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.

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***.

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:

```r
## Plot the session times
ggplot(session_times, aes(x=Page, y =Time)) +
  geom_boxplot()
```



```
## Get the mean session time for page A
mean_a <- mean(session_times[session_times['Page']=='Page A', 'Time'])
## Get the mean session time for page B
mean_b <- mean(session_times[session_times['Page']=='Page B', 'Time'])

## Take the difference and convert to seconds
(mean_b - mean_a) * 60
```

```
## [1] 21.4
```

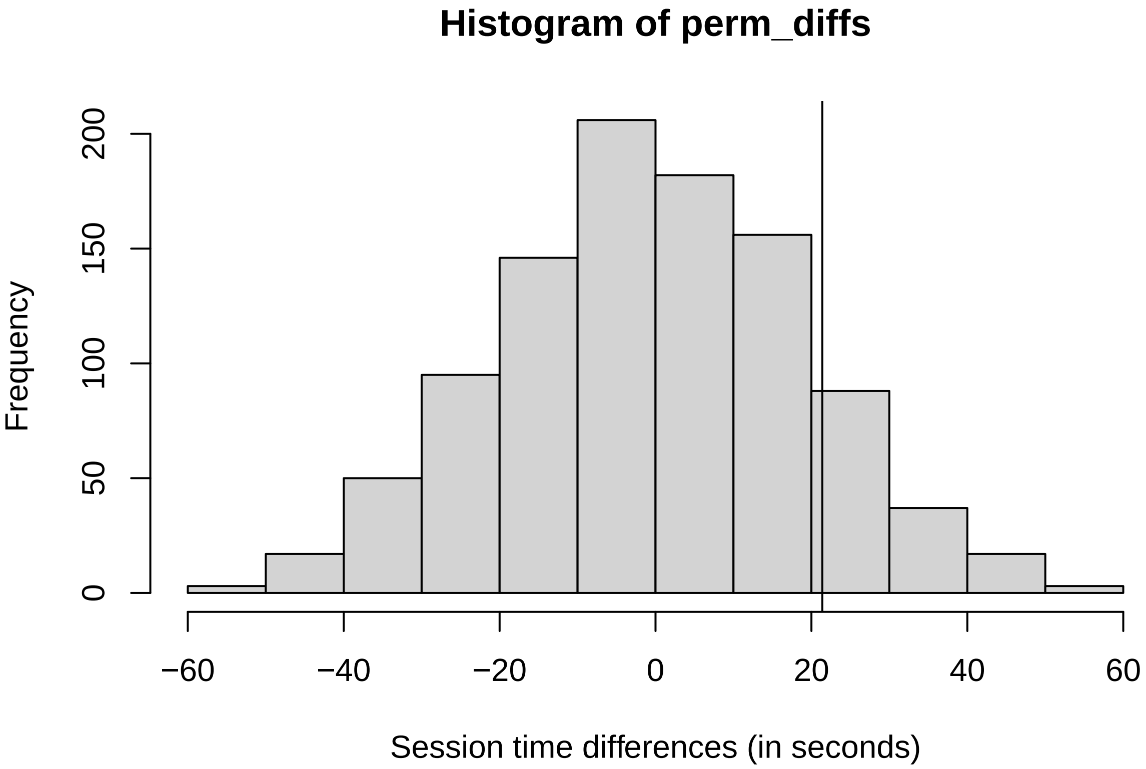
To apply a permutation test, we need a function to randomly assign the 36 session times to a group of 21 (page A) and a group of 15 (page B):

```
perm_fun <- function(x, n1, n2) {
  ## Get the number of samples
  n <- n1 +n2
  ## Sample the number for index b
  idx_b <- sample(1:n, n1)
  ## Get the sample difference
  idx_a <- setdiff(1:n, idx_b)
  ## get the differences between the means
  mean_diff <- (mean(x[idx_b]) - mean(x[idx_a])) * 60

  return(mean_diff)
}
```

This function works by sampling without replacement n2 indices and assigning them to the B group; the remaining n1 indices are assigned to group A. The difference between the two means is returned. Calling this function R = 1,000 times and specifying n2 = 15 and n1 = 21 leads to a distribution of differences in the session times that can be plotted as a histogram.

```
## initiate an array that has 0's
perm_diffs <- rep(0, 1000)
## iterate through the perm_fun 1000 times returning the mean difference
for(i in 1:1000)
  perm_diffs[i] = perm_fun(session_times[,'Time'], 21, 15)
## create a histogram
hist(perm_diffs, xlab='Session time differences (in seconds)')
abline(v = (mean_b - mean_a) * 60)
```



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.

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.

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***.

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

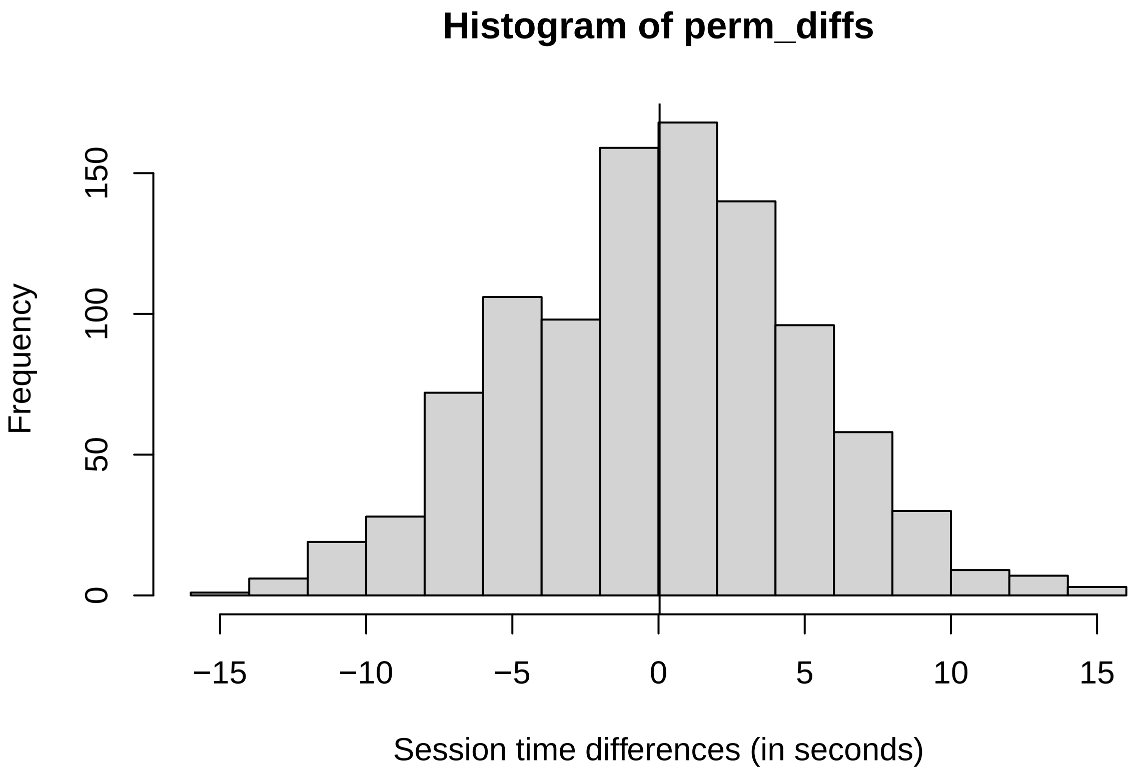
Statistical significance is how statisticians measure whether an experiment (or even a study of existing data) yields a result more extreme than what chance might produce. If the result is beyond the realm of chance variation, it is said to be statistically significant.

Term	Definition
P-value	Given a chance model that embodies the null 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).

Reusing the function perm_fun defined earlier, we can create a histogram of randomly permuted differences in conversion rate:

```r
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)
```



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

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] 0.511
```

The p-value is 0.308, which means that we would expect to achieve a result as extreme as this, or more extreme, 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%.

Value of the p-value

Considerable controversy has surrounded the use of the p-value in recent years. One psychology journal has gone so far as to "ban" the use of p-values in submitted papers on the grounds that publication decisions based solely on the p-value were resulting in the publication of poor research.

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 night 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.

Term	Definition
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 null 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.

These formulas are not shown here because all statistical software, as well as R and Python, include commands that embody the formula. In R, the function is `t.test`:

```r
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 0.1959674
## sample estimates:
## mean in group Page A mean in group Page B
##             1.263333             1.620000
```

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. In the formula world, many variations present themselves, and they can be bewildering. Statisticians need to navigate that world and learn its map, but data scientists do not—they are typically not in the business of sweating the details of hypothesis tests and confidence intervals the way a researcher preparing a paper for presentation might.

Multiple Testing

As we've mentioned previously, there is a saying in statistics: "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.

Term	Definition
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. As previously discussed, this is called a ***Type 1 error***. You 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 20 1 correctly test nonsignificant is 0.95 × 0.95 × 0.95 … or 0.95 = 0.36. The probability that at least one predictor will (falsely) test significant is the flip side of this probability, or 1 − (***probability that all will be nonsignificant***) = 0.64.

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

In the documentation and settings to many statistical tests, you will see reference to "degrees of freedom." The concept is applied to statistics calculated from sample data, and refers to the number of values free to vary. For example, if you know the mean for a sample of 10 values, and you also know 9 of the values, you also know the 10th value. Only 9 are free to vary.

Term	Definition
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.

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).

ANOVA

Suppose that, instead of an A/B test, we had a comparison of multiple groups, say A-B-C-D, each with numeric data. The statistical procedure that tests for a statistically significant difference among the groups is called ***analysis of variance***, or ***ANOVA***.

Term	Definition
Pairwise Comparison	A hypothesis test (e.g., of means) between two groups among multiple groups.
Omnibus Test	A single hypothesis test of the overall variance among multiple group means.
Decomposition of Variance	Separation of components. contributing to an individual value (e.g., from the overall average, from a treatment mean, and from a residual error).
F-statistic	A standardized statistic that measures the extent to which differences among group means exceeds what might be expected in a chance model.
SS	"Sum of squares," referring to deviations from some average value.

The more such ***pairwise*** comparisons we make, the greater the potential for being fooled by random chance. Instead of worrying about all the different comparisons between individual pages we could possibly make, we can do a single overall ***omnibus*** test that addresses the question, "Could all the pages have the same underlying stickiness, and the differences among them be due to the random way in which a common set of session times got allocated among the four pages?"

The procedure used to test this is ANOVA. The basis for it can be seen in the following resampling procedure (specified here for the A-B-C-D test of web page stickiness): 1. Combine all the data together in a single box 2. Shuffle and draw out four resamples of five values each 3. Record the mean of each of the four groups 4. Record the variance among the four group means 5. Repeat steps 2–4 many times (say 1,000)

What proportion of the time did the resampled variance exceed the observed variance? This is the p-value.

This type of permutation test is a bit more involved than the type used in "Permutation Test". Fortunately, the aovp function in the lmPerm package computes a permutation test for this case:

```r
library(lmPerm)
summary(aovp(Time ~ Page, data=four_sessions))
```

```
## [1] "Settings:  unique SS "
## Component 1 :
##             Df R Sum Sq R Mean Sq Iter Pr(Prob)
## Page1        3    831.4    277.13 4035  0.08203 .
## Residuals   16   1618.4    101.15
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

The p-value, given by Pr(Prob), is 0.09278. The column Iter lists the number of iterations taken in the permutation test. The other columns correspond to a traditional ANOVA table and are described next.

F-Statistic

Just like the t-test can be used instead of a permutation test for comparing the mean of two groups, there is a statistical test for ANOVA based on the ***F-statistic***. The F-statistic is based on the ratio of the variance across group means (i.e., the treatment effect) to the variance due to residual error. The higher this ratio, the more statistically significant the result. If the data follows a normal distribution, then statistical theory dictates that the statistic should have a certain distribution. Based on this, it is possible to compute a p-value.

In R, we can compute an ***ANOVA table*** using the `aov` function:

```r
summary(aov(Time ~ Page, data=four_sessions))
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Page         3  831.4   277.1    2.74 0.0776 .
## Residuals   16 1618.4   101.2
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Two-Way ANOVA

The A-B-C-D test just described is a “one-way” ANOVA, in which we have one factor (group) that is varying. We could have a second factor involved—say, “weekend versus weekday”—with data collected on each combination (group A weekend, group A weekday, group B weekend, etc.). This would be a “two-way ANOVA,” and we would handle it in similar fashion to the one-way ANOVA by identifying the “interaction effect.” After identifying the grand average effect, and the treatment effect, we then separate the weekend and the weekday observations for each group, and find the difference between the averages for those subsets and the treatment average.

You can see that ANOVA, then two-way ANOVA, are the first steps on the road toward a full statistical model, such as regression and logistic regression, in which multiple factors and their effects can be modeled.

Key Ideas
ANOVA is a statistical procedure for analyzing the results of an experiment with multiple groups.
It is the extension of similar procedures for the A/B test, used to assess whether the overall variation among groups is within the range of chance variation.
A useful outcome of an ANOVA is the identification of variance components associated with group treatments, interaction effects, and errors.

Chi-Square Test

The chi-square test is used with count data to test how well it fits some expected distribution. The most common use of the ***chi-square*** statistic in statistical practice is with $r * c$ contingency tables, to assess whether the null hypothesis of independence among variables is reasonable.

Term	Definition
Chi-Square Statistic	A measure of the extent to which some observed data departs from expectation.
Expectation or Expected	How we would expect the data to turn out under some assumption, typically the null hypothesis.
d.f.	Degrees of freedom.

The Pearson residual is defined as:

$R = \frac{Observed - Expected}{\sqrt{Expected}}$

R measures the extent to which the actual counts differ from these expected counts.

The chi-squared statistic is defined as the sum of the squared Pearson residuals:

$$\chi = \sum_{i}^{r} \sum_{j}^{c} R^2$$

where r and c are the number of rows and columns, respectively. The chi-squared statistic for this example is 1.666. Is that more than could reasonably occur in a chance model? We can test with this resampling algorithm: 1. Constitute a box with 34 ones (clicks) and 2,966 zeros (noclicks). 2. Shuffle,take three separate samples of 1,000, and count the clicks in each. 3. Find the squared differences between the shuffled counts and the expected counts, and sum them. 4. Repeat steps 2 and 3, say, 1,000 times. 5. How often does the resampled sum of squared deviations exceed the observed? That's the p-value.

The function `chisq.test` can be used to compute a resampled chi-square statistic. For the click data, the chi-square test is:

```
chisq.test(clicks, simulate.p.value=TRUE)
```

```
## Error in chisq.test(clicks, simulate.p.value = TRUE): all entries of 'x' must be nonnegative and fin
```

Chi-Squared Test: Statistical Theory

Asymptotic statistical theory shows that the distribution of the chi-squared statistic can be approximated by a ***chi-square distribution***. The appropriate standard chi-square distribution is determined by the ***degrees of freedom***. For a contingency table, the degrees of freedom are related to the number of rows (r) and columns (s) as follows:

$DegreesOfFreedom = (r - 1) * (c - 1)$

The chi-square distribution is typically skewed, with a long tail to the right for the distribution with 1, 2, 5, and 10 degrees of freedom. The further out on the chi-square distribution the observed statistic is, the lower the p-value.

The function chisq.test can be used to compute the p-value using the chi- squared distribution as a reference:

```r
chisq.test(clicks, simulate.p.value=FALSE)
```

```
## Error in chisq.test(clicks, simulate.p.value = FALSE): all entries of 'x' must be nonnegative and fi
```

The p-value is a little less than the resampling p-value: this is because the chi- square distribution is only an approximation of the actual distribution of the statistic.

Fisher's Exact Test

The chi-square distribution is a good approximation of the shuffled resampling test just described, except when counts are extremely low (single digits, especially five or fewer). In such cases, the resampling procedure will yield more accurate p-values. In fact, most statistical software has a procedure to actually enumerate all the possible rearrangements (permutations) that can occur, tabulate their frequencies, and determine exactly how extreme the observed result is. This is called Fisher's exact test after the great statistician R. A. Fisher. R code for Fisher's exact test is simple in its basic form:

```r
fisher.test(clicks)
```

```
## Error in fisher.test(clicks): all entries of 'x' must be nonnegative and finite
```

The p-value is very close to the p-value of 0.4853 obtained using the resampling method.

Relevance for Data Science

Most standard uses of the chi-square test, or Fisher's exact test, are not terribly relevant for data science. In most experiments, whether A-B or A-B-C…, the goal is not simply to establish statistical significance, but rather to arive at the best treatment. For this purpose, multi-armed bandits (see "Multi-Arm Bandit Algorithm") offer a more complete solution.

Chi-square tests are used widely in research by investigators in search of the elusive statistically significant p-value that will allow publication. Chi-square tests, or similar resampling simulations, are used in data science applications more as a filter to determine whether an effect or feature is worthy of further consideration than as a formal test of significance.

Key Ideas
A common procedure in statistics is to test whether observed data counts are consistent with an assumption of independence (e.g., propensity to buy a particular item is independent of gender).
The chi-square distribution is the reference distribution (which embodies the assumption of independence) to which the observed calculated chi- square statistic must be compared.

Multi-Arm Bandit Algorithm

Multi-arm bandits offer an approach to testing, especially web testing, that allows explicit optimization and more rapid decision making than the traditional statistical approach to designing experiments.

Term	Definition
Multi-Arm Bandit	An imaginary slot machine with multiple arms for the customer to choose from, each with different payoffs, here taken to be an analogy for a multitreatment experiment.
Arm	A treatment in an experiment (e.g., “headline A in a web test”).
Win	The experimental analog of a win at the slot machine (e.g., “customer clicks on the link”).

Bandit algorithms, which are very popular in web testing, allow you to test multiple treatments at once and reach conclusions faster than traditional statistical designs. They take their name from slot machines used in gambling, also termed one-armed bandits (since they are configured in such a way that they extract money from the gambler in a steady flow). If you imagine a slot machine with more than one arm, each arm paying out at a different rate, you would have a multi-armed bandit, which is the full name for this algorithm.

Your goal is to win as much money as possible, and more specifically, to identify and settle on the winning arm sooner rather than later. The challenge is that you don’t know at what rate the arms pay out—you only know the results of pulling the arm. Suppose each “win” is for the same amount, no matter which arm. What differs is the probability of a win.

A more sophisticated algorithm uses “Thompson’s sampling.” This procedure “samples” (pulls a bandit arm) at each stage to maximize the probability of choosing the best arm. Of course you don’t know which is the best arm—that’s the whole problem!—but as you observe the payoff with each successive draw, you gain more information. Thompson’s sampling uses a Bayesian approach: some prior distribution of rewards is assumed initially, using what is called a ***beta distribution*** (this is a common mechanism for specifying prior information in a Bayesian problem). As information accumulates from each draw, this information can be updated, allowing the selection of the next draw to be better optimized as far as choosing the right arm.

Bandit algorithms can efficiently handle 3+ treatments and move toward optimal selection of the “best.” For traditional statistical testing procedures, the complexity of decision making for 3+ treatments far outstrips that of the traditional A/B test, and the advantage of bandit algorithms is much greater.

Key Ideas
Traditional A/B tests envision a random sampling process, which can lead to excessive exposure to the inferior treatment.
Multi-arm bandits, in contrast, alter the sampling process to incorporate information learned during the experiment and reduce the frequency of the inferior treatment.
They also facilitate efficient treatment of more than two treatments.
There are different algorithms for shifting sampling probability away from the inferior treatment(s) and to the (presumed) superior one.

Power and Sample Size

Term	Defintion
Effect Size	The minimum size of the effect that you hope to be able to detect in a statistical test, such as “a 20% improvement in click rates”.
Power	The probability of detecting a given effect size with a given sample size.

Term	Defintion
Significant Level	The statistical significance level at which the test will be conducted.

Power is the probability of detecting a specified ***effect size*** with specified sample characteristics (size and variability). For example, we might say (hypothetically) that the probability of distinguishing between a .330 hitter and a .200 hitter in 25 at-bats is 0.75. The effect size here is a difference of .130. And “detecting” means that a hypothesis test will reject the null hypothesis of “no difference” and conclude there is a real effect. So the experiment of 25 at-bats (n = 25) for two hitters, with an effect size of 0.130, has (hypothetical) power of 0.75 or 75%.

Sample Size

The most common use of power calculations is to estimate how big a sample you will need.

In summary, for calculating power or required sample size, there are four moving parts: * Sample Size * Effect size you want to detect * Significance level (alpha) at which the test will be conducted * Power

Key Ideas
Finding out how big a sample size you need requires thinking ahead to the statistical test you plan to conduct. You must specify the minimum size of the effect that you want to detect. You must also specify the required probability of detecting that effect size (power). Finally, you must specify the significance level (alpha) at which the test will be conducted.