Statistical Experiments and Significance Testing

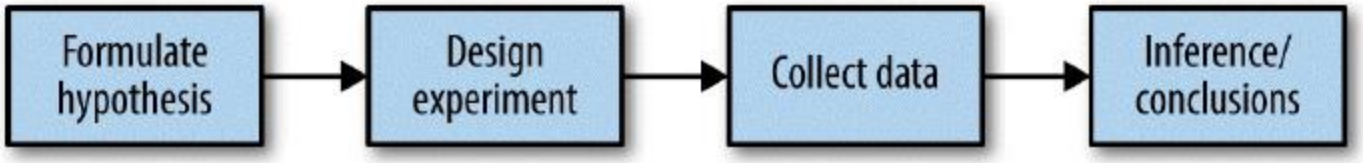


Figure 3-1. The classical statistical inference pipeline

- 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
- 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.

*This presentation is based on the book 'Practical Statistics for Data Scientist' by Peter Bruce & Andrew Bruce

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.**

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.

Hypothesis Tests

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

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.

The Null Hypothesis

- "Given the human tendency to react to unusual but random behavior and interpret it as something meaningful and real, in our experiments we will require proof that the difference between groups is more extreme than what chance might reasonably produce."
- 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.

- **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"

- 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.

- 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.

Resampling

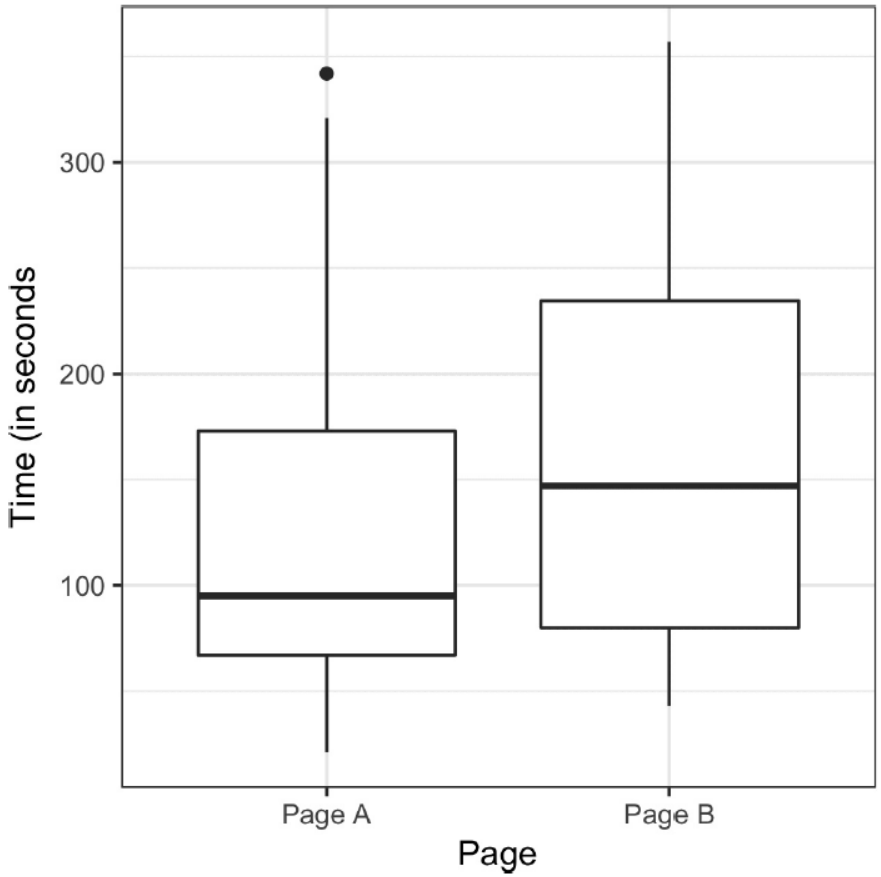
- Resampling in statistics means to repeatedly sample values from observed data, with a general goal of assessing random variability in a statistic.
- There are two main types of resampling procedures: the bootstrap and ***permutation tests***.
- The bootstrap is used to assess the reliability of an estimate; it was discussed in the previous chapter
- Permutation tests are used to test hypotheses, typically involving two or more groups, and we discuss those in this section.

Permutation Test

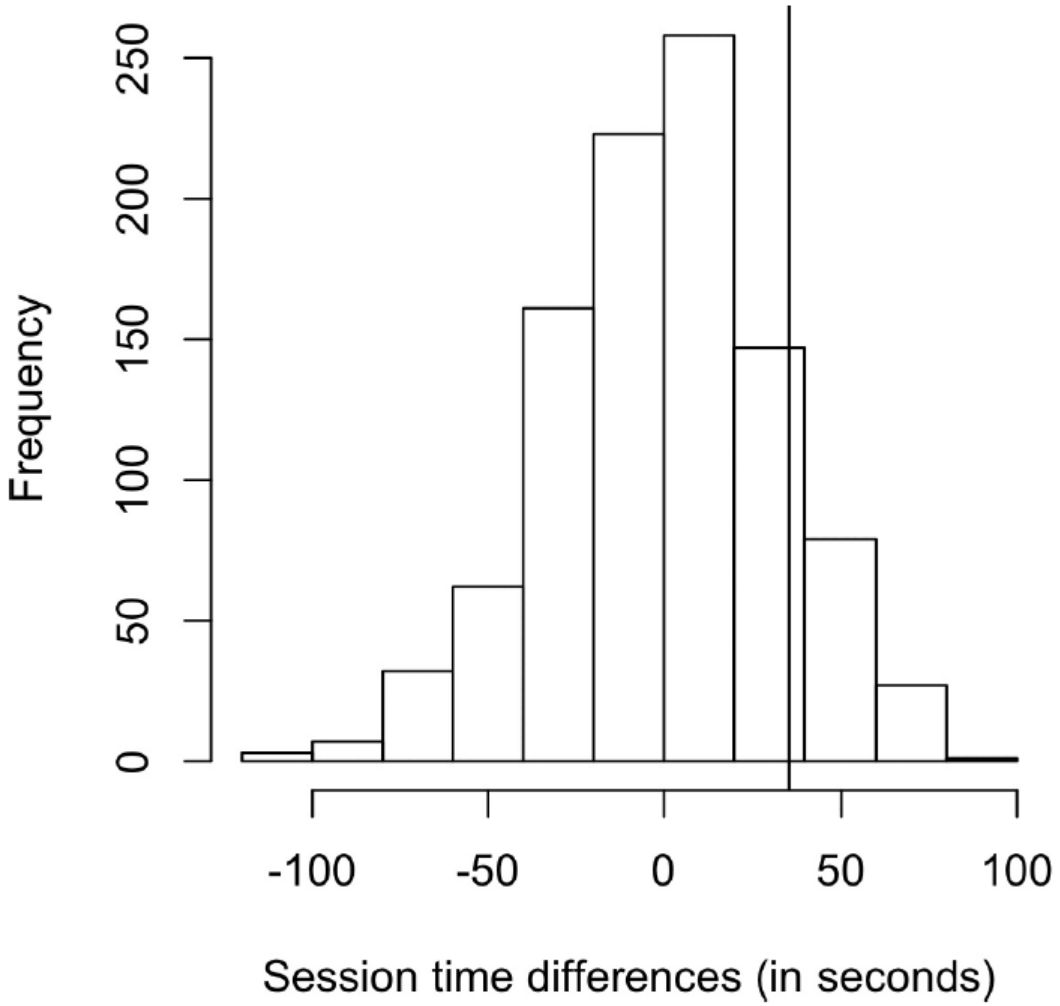
- Combine the results from the different groups in a single data set.
- Shuffle the combined data, then randomly draw (without replacing) a resample of the same size as group A.
- From the remaining data, randomly draw (without replacing) a resample of the same size as group B.
- Whatever statistic or 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.
- 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*.

- 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).



- The histogram 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.



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***.

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

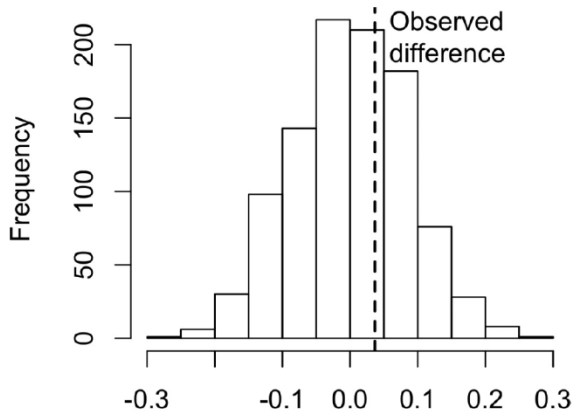
Mistakenly concluding an effect is real (when it is due to chance).

Type 2 error

Mistakenly concluding an effect is due to chance (when it is real).

P-Value

- This is the frequency with which the chance model produces a result more extreme than the observed result.



- 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.
- Smaller the p-value, more statistically significant the difference is.

- Alpha: 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%.

Type 1 and Type 2 Errors

- 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.”

- 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.