Sampling and hypothesis tests

Part 6: Resampling, permutation test, p-value and types of errors

By: Noureddin Sadawi, PhD

University of London

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.’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

Key terms for resampling

- ‘**Permutation test:** The procedure of combining two or more samples together and randomly (or exhaustively) reallocating the observations to resamples.
  - Synonyms: Randomization test, random permutation test, exact test.
- **Resampling:** Drawing additional samples (“resamples”) from an observed data set.

  With or without replacement: In sampling, whether or not an item is returned to the sample before the next draw.’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

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,…).
- 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.’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

Permutation test procedure

1. ‘Combine the results from the different groups into a single data set.
2. Shuffle the combined data and then randomly draw (without replacement) a resample of the same size as group A (clearly it will contain some data from the other groups).
3. From the remaining data, randomly draw (without replacement) a resample of the same size as group B.
4. Do the same for groups C, D, and so on. You have now collected one set of resamples that mirror the sizes of the original samples.
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.’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

Permutation test

- ‘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*.’

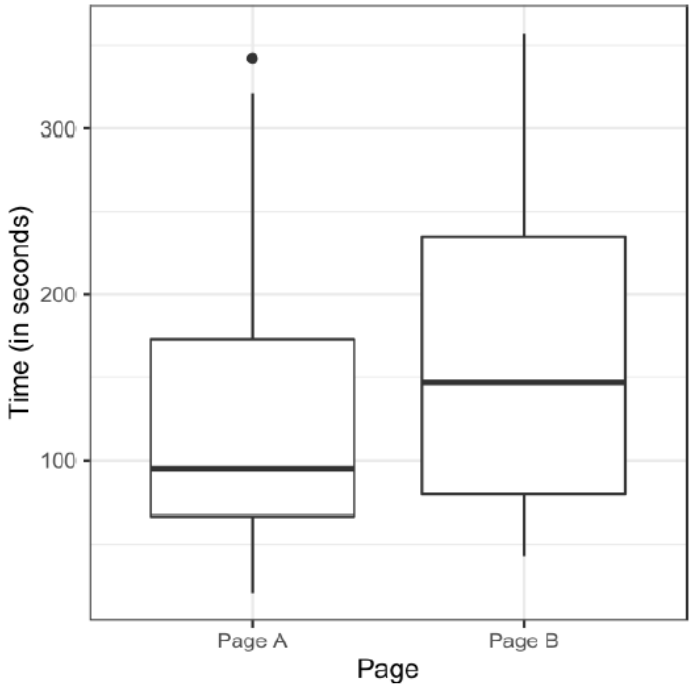
(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

Key ideas

- ‘In a permutation test, multiple samples are combined and 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.’

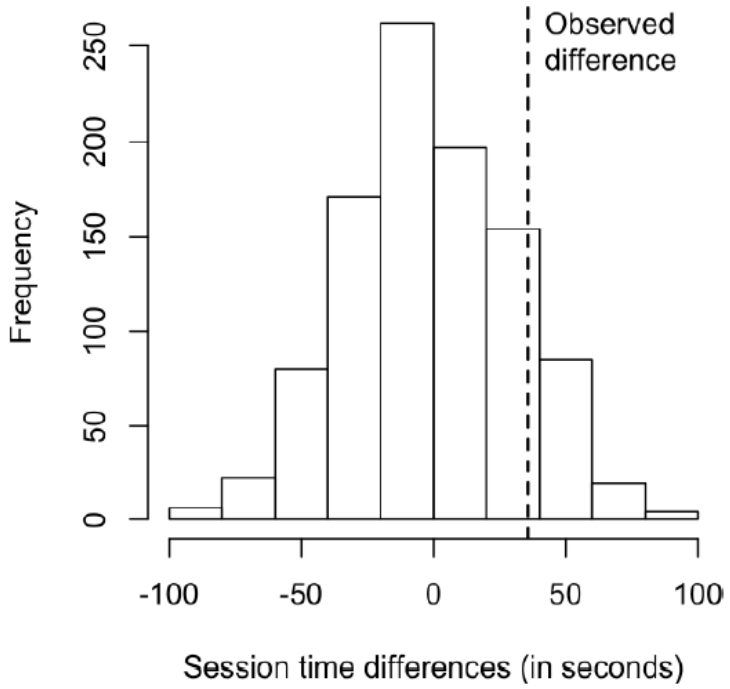
(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

Example



Session times for web pages A and B (21 visits to page A and 15 to page B)

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).



Frequency distribution for session time differences between pages A and B; the vertical line shows the observed difference (well within chance)

Statistical significance and p-value

- ‘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.’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

Key terms for statistical significance and p-value

- ‘**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).’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

p-value

- Consider the web-page visits example.
- ‘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.’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

How to interpret p-value

- Sometimes people interpret the p-value as:

  "*The probability that the result is due to chance.*"
- A better interpretation is:

  "*The probability that, given a chance model, results as extreme as the observed results could occur.*"
- A threshold value of 5% (or 0.05) for the p-value is often used to determine if the results are significant.
- Remember, the question 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?"

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

Type 1 and type 2 errors

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

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

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).