Statistical Hypothesis Testing, Model Comparison

Milan Straka

January 04, 2021

Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Statistical Hypothesis Testing

Variation of a famous saying states, that there are various kinds of truth:

- truth,
- half-truth,
- lie,
- disgusting lie,
- and statistics.

Statistical Hypothesis Testing

Assume we have a hypothesis testable using observed data of random variables.

There are two slightly differing views on statistical hypothesis testing:

1. In the first one, we assume we have a **null hypothesis** H_0, and we are interested in whether we can **reject it** using the observed data.

   The result is **statistically significant**, if it is very unlikely that the observed data have occurred given the null hypothesis.

   The **significance level** of a test is the threshold of this unlikeliness.

2. In the second view, we have two hypotheses, a null hypothesis H_0 and an **alternative hypothesis** H_1, and we want to distinguish among them.

   We consider only two outcomes of the test:
   - either we "reject" the null hypothesis, if the data is very unlikely to have occurred given the null hypothesis; or
   - we cannot reject the null hypothesis.

   Note that we never "prove" the alternative hypothesis.

Type I and Type II Errors

Consider the *courtroom trial* example, which is similar to a criminal trial, where the defendant is considered not guilty until their guilt is proven.

In this setting, H_0 is “not guilty” and H_1 is “guilty”.

	H_0 is true **Truly not guilty**	**H_1 is true** **Truly guilty**
Not proven guilty Not rejecting H_0	Correct decision True negative	Wrong decision False negative **Type II Error**
Proven guilty Rejecting H_0	Wrong decision False positive **Type I Error**	Correct decision True positive

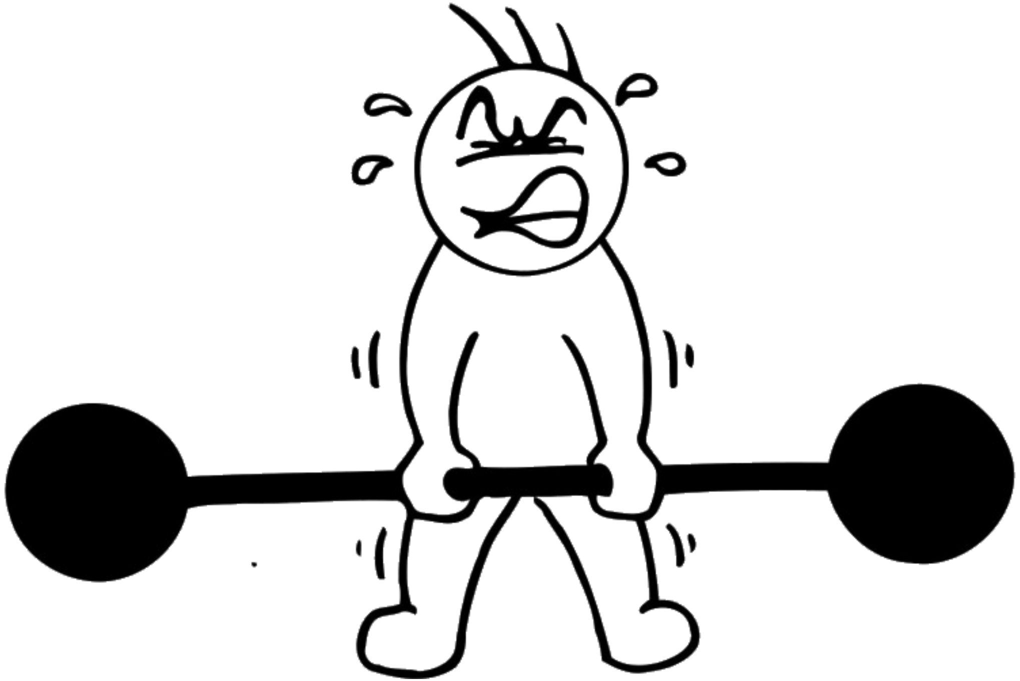
Our goal is to limit the Type 1 errors – the test **significance level** is the type 1 error rate.

Match Analogy

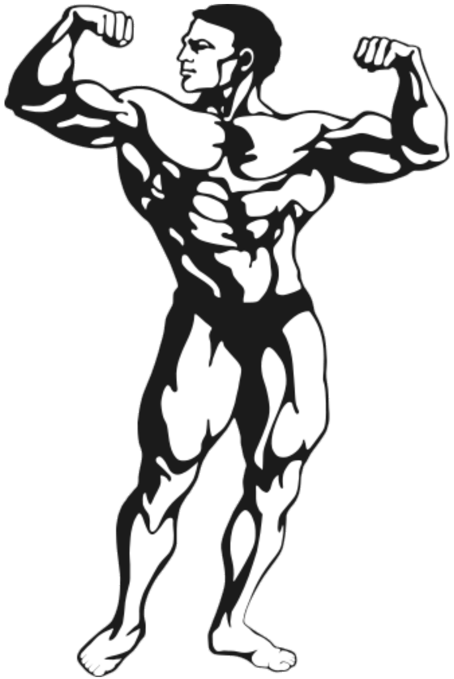
I like the following analogy – if you have a theory and want to convince others that it holds, you devise an *opponent* for it and let them wrestle.

If your theory wins, it may be an indication that it really holds.

However, you must choose an appropriate opponent.



http://clipart-library.com/clipart/qTB7drzT5.htm



http://clipart-library.com/clipart/n892248.htm

Statistical Hypothesis Testing

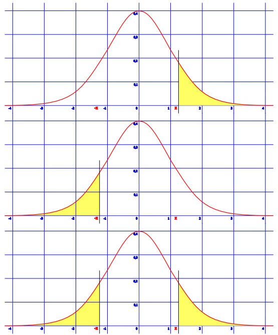
The crucial part of a statistical test is the **test statistic**. It is some summary of the observed data, very often a single value (like mean), which can be used to distinguish the null and the alternative hypothesis.

It is crucial to be able to compute the distribution of the test statistic, which allows the **p-values** to be calculated.

A **p-value** is the probability of obtaining test statistic value at least as extreme as the one actually observed, assuming validity of the null hypothesis. A very small p-value indicates that the observed data are very unlikely under the null hypothesis.

Given a test statistic, we usually perform one of

- a one-sided right-tail test, when the p-value of t is $P(\textit{test statistic} > t|H_0)$;
- a one-sided left-tail test, when the p-value of t is $P(\textit{test statistic} < t|H_0)$;
- a two-sided test, when the p-value of t is twice the minimum of $P(\textit{test statistic} < t|H_0)$ and $P(\textit{test statistic} > t|H_0)$. For a symmetrical centered distribution, $P(\mathrm{abs}(\textit{test statistic}) > \mathrm{abs}(t)|H_0)$ can also be used.



https://upload.wikimedia.org/wi

Statistical Hypothesis Testing

Therefore, the whole procedure consists of the following steps:

1. Formulate the null hypothesis H_0, and optionally the alternative hypothesis H_1.
2. Choose the test statistic.
3. Compute the observed value of the test statistic.
4. Calculate the p-value, which is the probability of a test statistic value being at least as extreme as the observed one, under the null hypothesis H_0.
5. Reject the null hypothesis H_0 (in favor of the alternative hypothesis H_1), if the p-value is less than the chosen significance level α (a standard is to use α at most 5%; common choices include 5%, 1%, 0.5% or 0.1%, but vary a lot in different fields).

Test Statistics

There are several kinds of test statistics:

- **one-sample tests**, where we sample values from one distribution.

  Common one-sample tests usually check for
  - the mean of the distribution to be greater/ than and/or equal to zero;
  - the goodness of fit (that the data comes from a normal or categorical distribution of given parameters).

- **two-sample tests**, where we sample independently from two distributions.

- **paired tests**, in which case we also sample from two distributions, but the samples are paired (i.e., evaluating several models on the same data).

  In paired tests, we usually compute the difference between the paired members and perform one-sample test on the mean of the differences.

Test Statistics

There are many commonly used test statistics, with different requirements and conditions. We only mention several commonly-used ones, but it is by no means a comprehensive treatment.

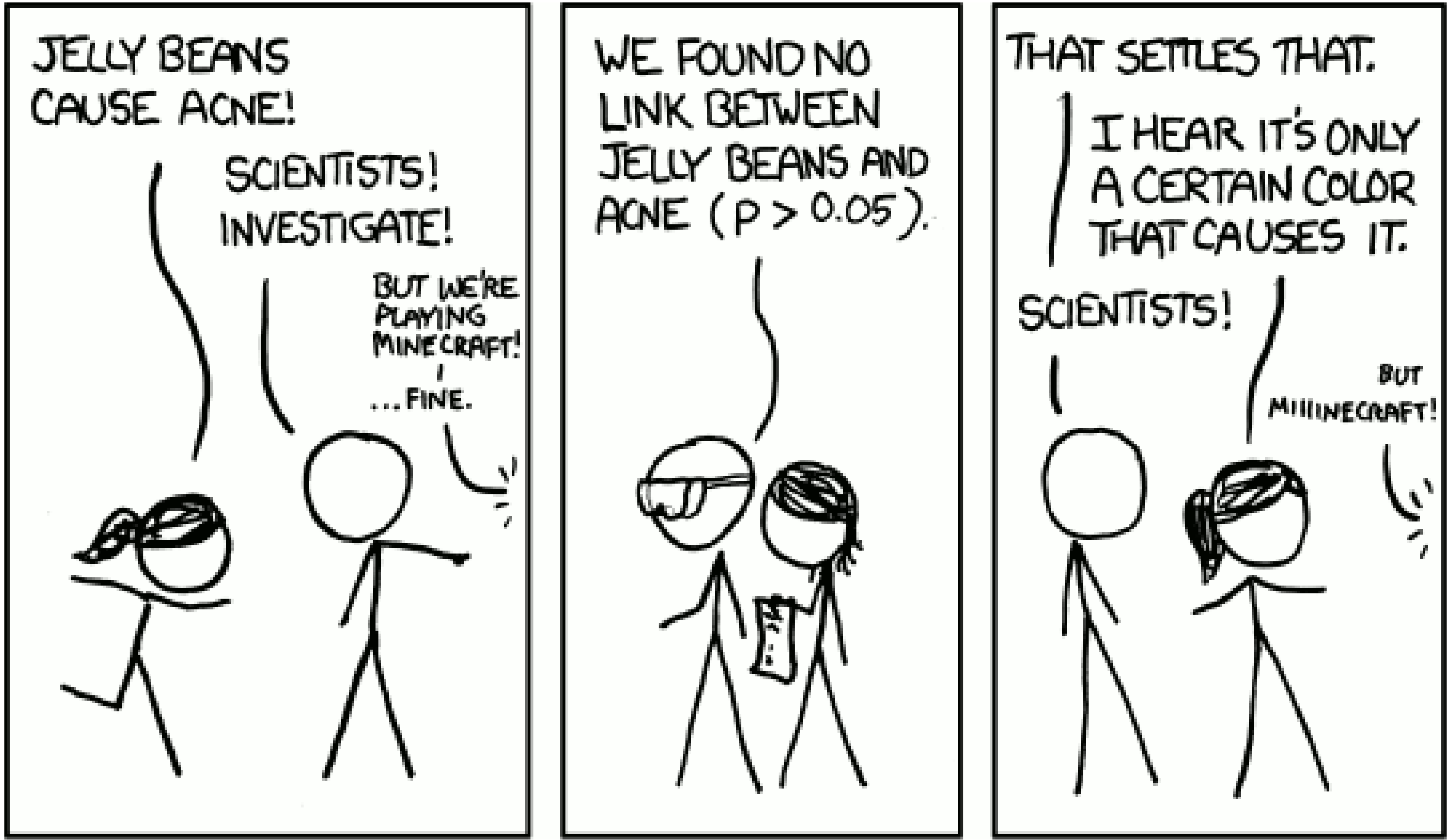
- **Z-Test** is a test, where the test statistic can be approximated by a normal distribution. For example, it can be used when comparing a mean of samples *with known variance* to a given value.
- In **Student's *t*-test** the test statistic follow a Student's *t*-distribution (where Student is the pseudonym used by the real author W. S. Gosset), which is the distribution of a sample mean of normally-distributed population *with unknown variance*.

  Therefore, the *t*-test is used when comparing a mean of samples with unknown variance to a given value, or to a mean of samples from another distribution with the same sample size and variance.
- **Chi-squared test** utilizes a test statistic with a chi-squared distribution, which is a distribution of a sum of squares of k independent normally distributed variables.

  The essential Pearson's chi-squared test can be used to evaluate a goodness of fit of k random categorical samples with respect to a given categorical distribution.

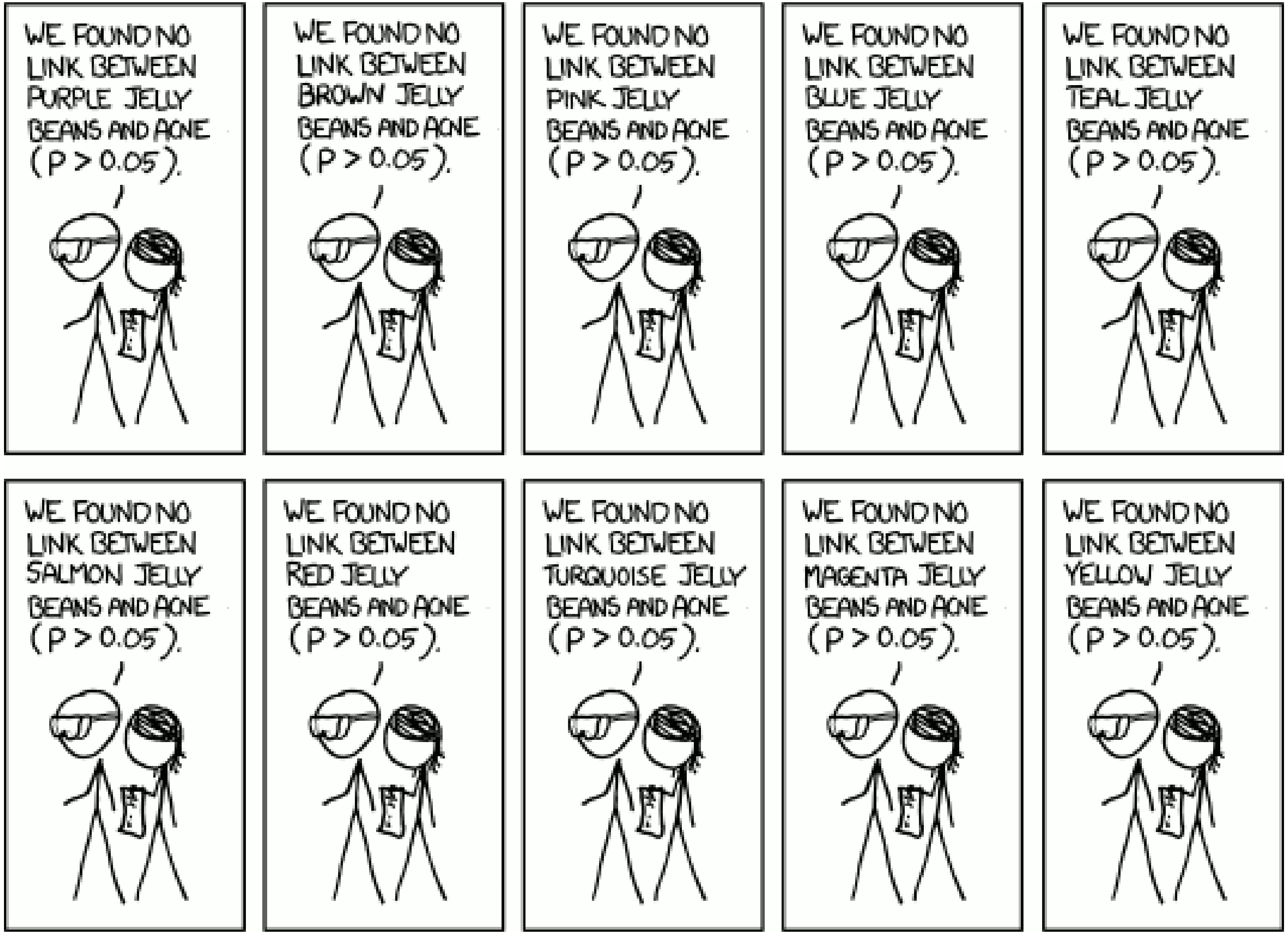
Multiple Comparisons Problem

A **multiple comparisons problem** (or multiple testing problem) arises, if we consider many statistical hypotheses tests using the same observed data.


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIINECRAFT!

https://imgs.xkcd.com/comics/significant.png

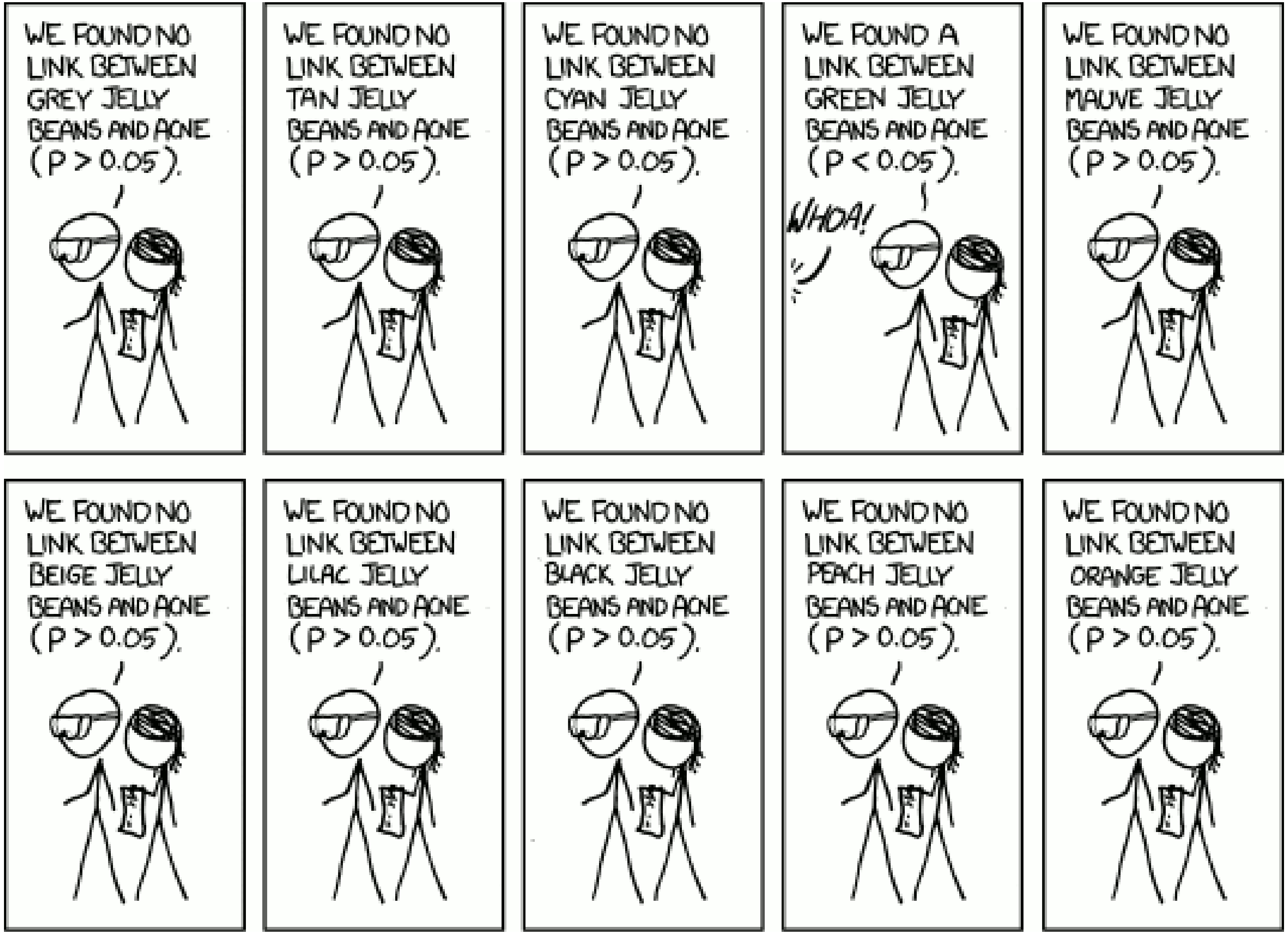
Multiple Comparisons Problem



WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).

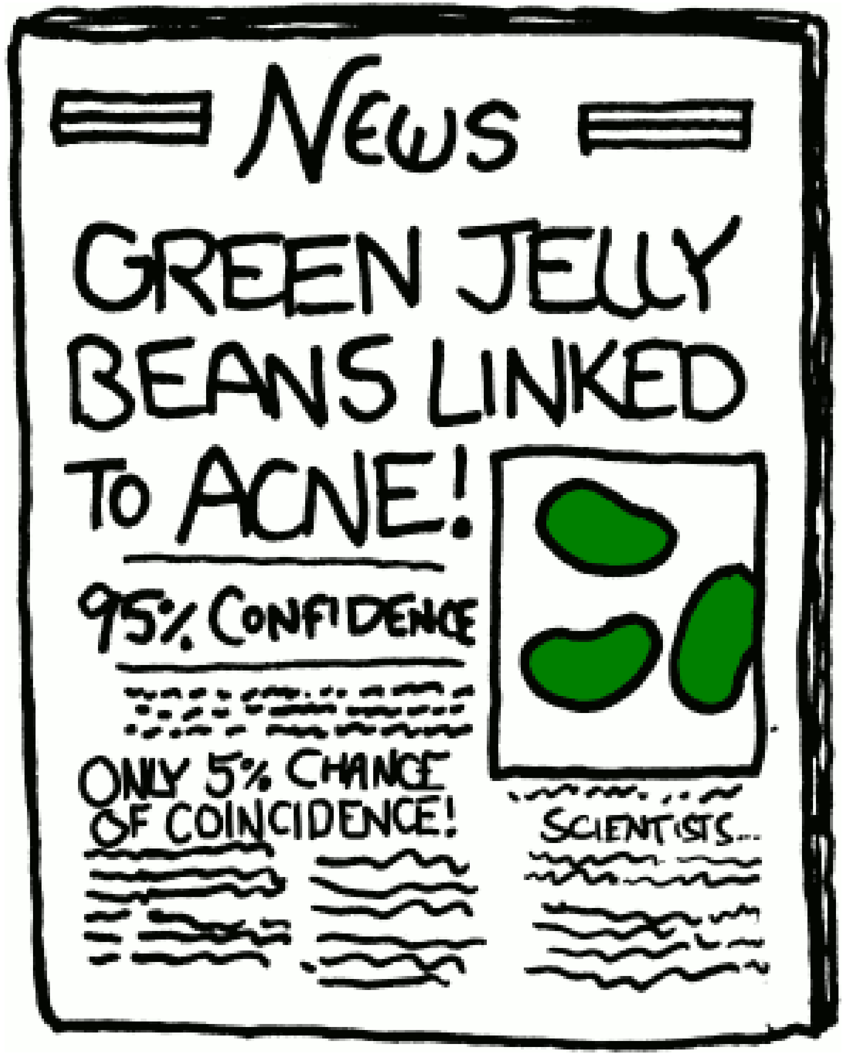
https://imgs.xkcd.com/comics/significant.png

Multiple Comparisons Problem


WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).

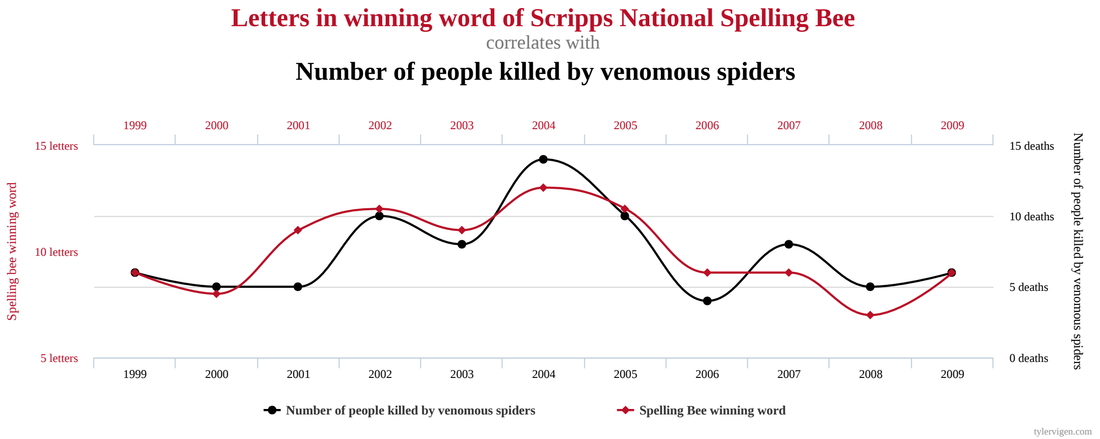
https://imgs.xkcd.com/comics/significant.png

Multiple Comparisons Problem


NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

https://imgs.xkcd.com/comics/significant.png

Usually, the problem is that we perform many statistical tests and only report the ones with statistically significant results.



https://upload.wikimedia.org/wikipedia/commons/0/0c/Spurious_correlations_-_spelling_bee_spiders.svg

Multiple Comparisons Problem: Post-Mortem Salmon Study

Even world-class researchers make mistakes in multiple comparisons problem. Consider the paper

> *Neural Correlates of Interspecies Perspective Taking in the Post-Mortem Atlantic Salmon: An Argument For Proper Multiple Comparisons Correction*

Functional magnetic resonance imaging (fMRI) is a technique for monitoring brain activity via measuring the changes in blood oxygenation. The measurement is performed for every *voxel* in the brain; the authors claim that 130k voxels are common in a single fMRI measurement.

The correlation is usually computed for every voxel, and usually a cluster of some number of neighboring voxels, all of which must pass statistically significant test, is required.

However, with such a large number of voxels, a positive result can be caused by chance without some multiple comparison correction.

Multiple Comparisons Problem: Post-Mortem Salmon Study

The authors perform the following experiment. Citing:

> *One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon measured approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning. It is not known if the salmon was male or female, but given the post-mortem state of the subject this was not thought to be a critical variable.*
>
> ...
>
> *The task administered to the salmon involved completing an open-ended mentalizing task. The salmon was shown a series of photographs depicting human individuals in social situations with a specified emotional valence, either socially inclusive or socially exclusive. The salmon was asked to determine which emotion the individual in the photo must have been experiencing.*

Multiple Comparisons Problem: Post-Mortem Salmon Study

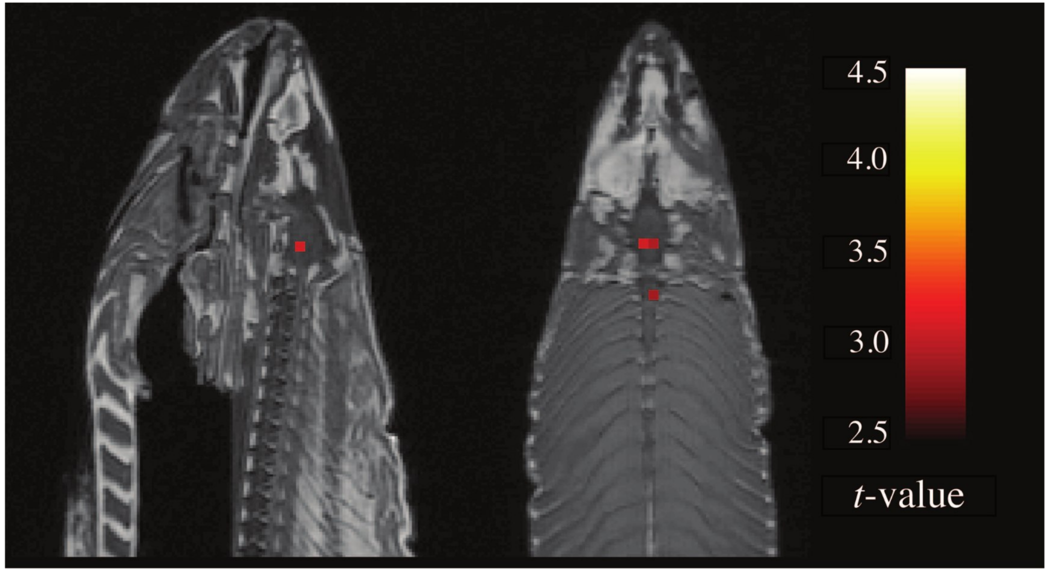


Fig. 1. Sagittal and axial images of significant brain voxels in the task > rest contrast. The parameters for this comparison were $t(131) > 3.15$, $p(\text{uncorrected}) < 0.001$, 3 voxel extent threshold. Two clusters were observed in the salmon central nervous system. One cluster was observed in the medial brain cavity and another was observed in the upper spinal column.

Figure 1 of the paper "Neural Correlates of Interspecies Perspective Taking in the Post-Mortem Atlantic Salmon" by C. Bennet et al.

A t-contrast was used to test for regions with significant BOLD signal change during the presentation of photos as compared to rest. The relatively low extent threshold value was chosen due to the small size of the salmon's brain relative to voxel size. Several active voxels were observed in a cluster located within the salmon's brain cavity (see Fig. 1). The size of this cluster was 81 mm^3 with a cluster-level significance of $p = 0.001$.

Multiple Comparisons Problem: Post-Mortem Salmon Study

The authors claim that:

> *Sadly, while methods for multiple comparisons correction are included in every major neuroimaging software package these techniques are not always invoked in the analysis of functional imaging data. For the year 2008 only 74% of articles in the journal NeuroImage reported results from a general linear model analysis of fMRI data that utilized multiple comparisons correction (193/260 studies). Other journals we examined were Cerebral Cortex (67.5%, 54/80 studies), Social Cognitive and Affective Neuroscience (60%, 15/25 studies), Human Brain Mapping (75.4%, 43/57 studies), and the Journal of Cognitive Neuroscience (61.8%, 42/68 studies). … The issue is not limited to published articles, as proper multiple comparisons correction is somewhat rare during neuroimaging conference presentations. During one poster session at a recent neuroscience conference only 21% of the researchers used multiple comparisons correction in their research (9/42). A further, more insidious problem is that some researchers would apply correction to some contrasts but not to others depending on the results of each comparison.*

Family-Wise Error Rate

There are several ways to handle the multiple comparison problem; one of the easiest (but often overly conservative) is to limit the **family-wise error rate**, which is the probability of at least one type 1 error in the family.

$$\mathrm{FWER} = P\Big(\bigcup_i \big(p_i \leq \alpha\big)\Big).$$

One way of controlling the family-wise error rate is the **Bonferroni correction**, which rejects the null hypothesis of a test in the family of size m when $p_i \leq \frac{\alpha}{m}$.

Assuming such a correction and utilizing the Boole's inequality $P\big(\bigcup_i A_i\big) \leq \sum_i P(A_i)$, we get that

$$\mathrm{FWER} = P\Big(\bigcup_i \Big(p_i \leq \frac{\alpha}{m}\Big)\Big) \leq \sum_i P\Big(p_i \leq \frac{\alpha}{m}\Big) = m \cdot \frac{\alpha}{m} = \alpha.$$

Note that there exist other more powerful methods like Holm-Bonferroni or Šidák correction.

Model Comparison

The goal of model comparison is to test whether some model will deliver better perfomance on unseen data than another one.

However, we usually only have a single fixed-size test set. For the rest of the lecture, we assume the test set instances are independently sampled from the data generating distribution.

Even if comparing the models on the given test set is unbiased, we would like to obtain some significance level of the result.

Therefore, we perform a statistical test with alternative hypothesis that a model y is better than a model z; therefore, the null hypothesis is that the model y is the same or worse than the model z.

However, we only have one sample (the result of a model on the test set). We therefore turn to **bootstrap resampling**.

Bootstrap Resampling

In order to obtain multiple samples of model performance, we exploit the fact that the test set consists of *multiple* examples.

Therefore, we can generate different test sets by bootstrap resampling. Notably, we obtain a same-sized test set by sampling the original test set examples *with replacement*. Naturally, we can easily measure the performance of any given model on such generated test set.

Input: Test set $\{(\boldsymbol{x}_1, t_1), \ldots, (\boldsymbol{x}_N, t_N)\}$, model predictions $\{y(\boldsymbol{x}_1), \ldots, y(\boldsymbol{x}_N)\}$, a metric E, number of resamplings R.

Output: R samples of model performance.

```
• $\mathrm{performances} \leftarrow []$
• repeat $R$ times:
    ○ sample $N$ test set examples with replacements, together with corresponding model predictions
    ○ measure the performance of the sampled data using the metric $E$ and append the result to $\mathrm{performances}$
```

Bootstrap Resampling – Confidence Intervals

When using bootstrap resampling on a single model, we can measure the confidence intervals of model performance.

For a given confidence level (95% is the most common value), the **confidence interval** is an estimate of a value range of some unknown parameter (like a mean performance of some model on unseen data), such that the confidence interval contains the true value of the unknown parameter with the frequency given by the confidence level.

When given the empirical distribution of model performances produced by bootstrap resampling, we can estimate the 95% confidence interval as a range from the 2.5 percentile and 97.5 percentile of the empirical distribution (the so-called *percentile bootstrap*).

Paired Bootstrap Test

To perform the model comparison statistical test, we could use a two-sample test. However, such a test does not consider the fact that some of the inputs might be more difficult than others, and takes into account cases when a weaker model achieves higher performance on a simpler test set than a stronger model on a more difficult test set.

Instead, we perform a paired bootstrap test. Our alternative hypothesis is that the mean of the model performance differences is larger than zero, and the null hypothesis is that it is less or equal to zero. We then repeatedly sample a test set with repetition and compute the difference of the model performances on the sampled test set, obtaining a distribution of differences *under the true distribution*.

However, to perform the statistical test, we require the distribution of the differences *under the null hypothesis*. One way of obtaining this distribution is to assume that the distribution of differences is translation invariant, under which assumption we obtain the wanted distribution as the mean-centered bootstrap distribution. Finally, assuming symmetry, we can estimate the p-value as the ratio of the bootstrapped differences which are less or equal to zero. (See *permutation tests* for a different way of estimating the p-values.)

Paired Bootstrap Test

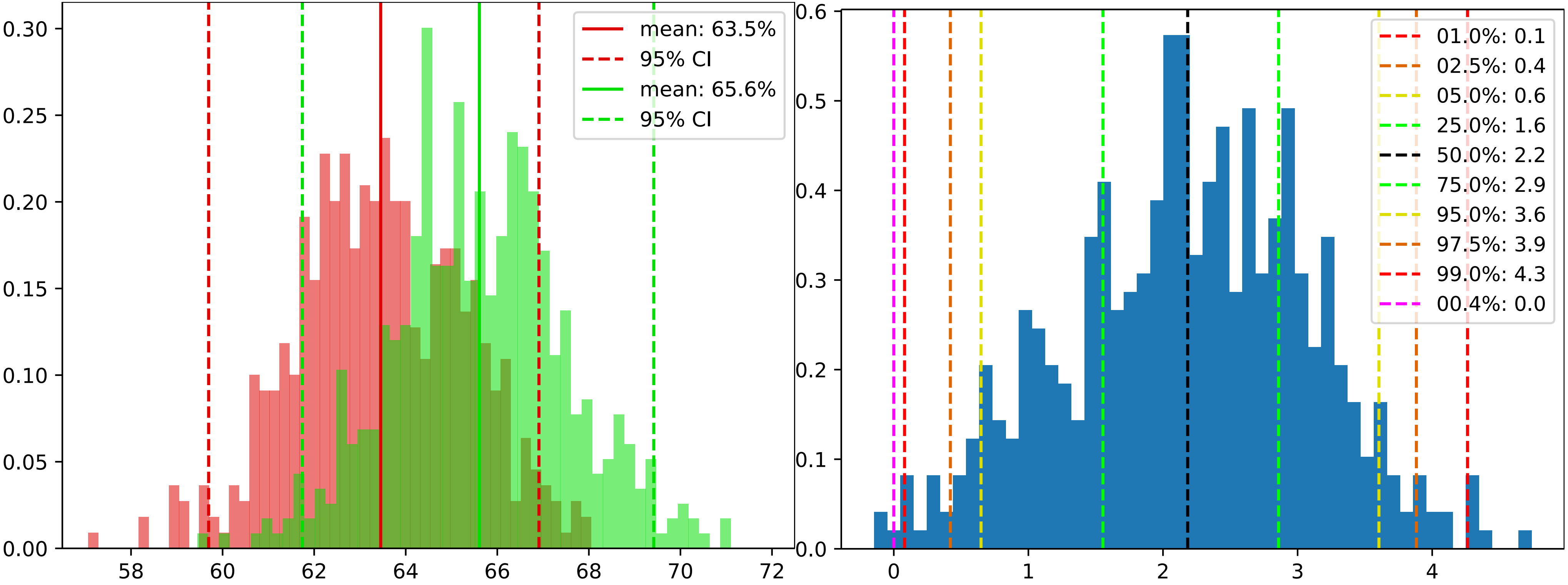
Input: Test set $\{(\boldsymbol{x}_1, t_1), \ldots, (\boldsymbol{x}_N, t_N)\}$, model predictions $\{y(\boldsymbol{x}_1), \ldots, y(\boldsymbol{x}_N)\}$, model predictions $\{z(\boldsymbol{x}_1), \ldots, z(\boldsymbol{x}_N)\}$, a metric E, number of resamplings R.

Output: Estimated p-value assuming that the model y performance is worse or equal to z.

- $\mathrm{differences} \leftarrow []$
- repeat R times:
  - sample N test set examples with replacements, together with the corresponding predictions of the models
  - measure the performances of the models y and z on the sampled data using the metric E and append their difference to $\mathrm{differences}$
- return the ratio of the $\mathrm{differences}$ which are less or equal to zero

Paired Bootstrap Test

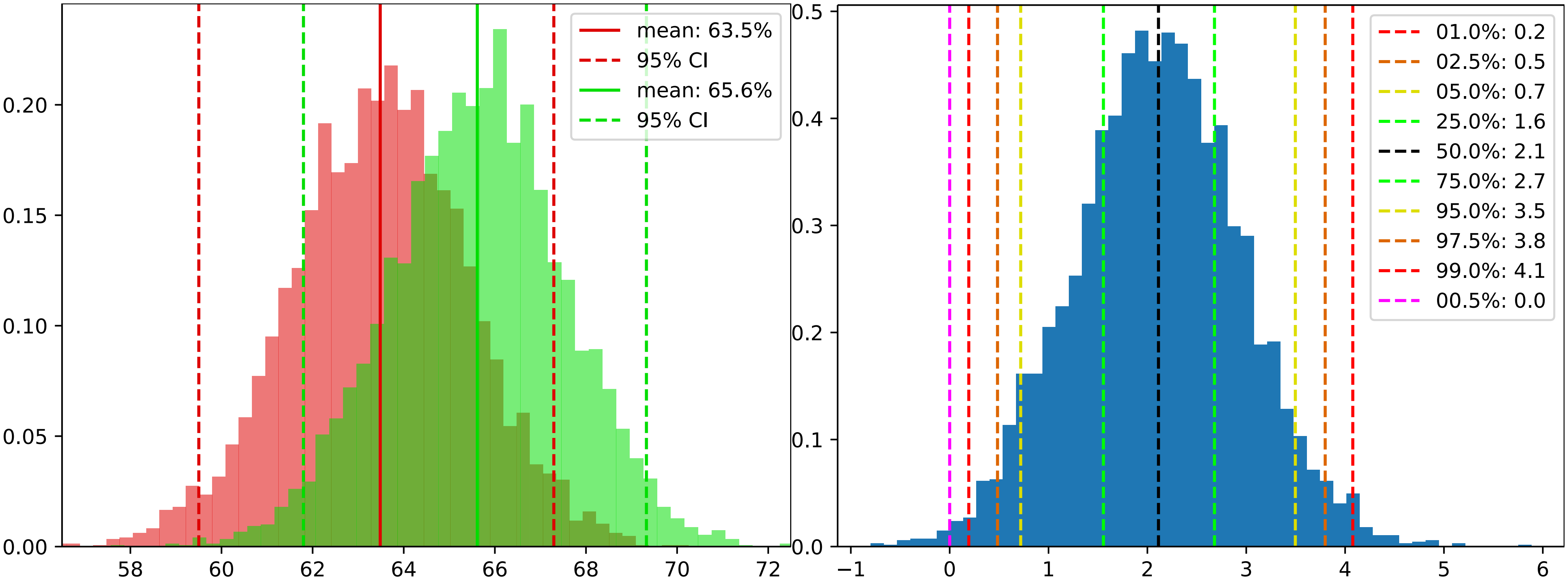
For illustration, consider models for the `isnt_it_ironic` competition utilizing either 3 (red) or 4 (green) in-word character n-grams. On the left, there are distributions of the individual model performances, while on the right there is a distribution of their differences.



The histograms are generated using 50 bins and 500 resamplings.

Paired Bootstrap Test

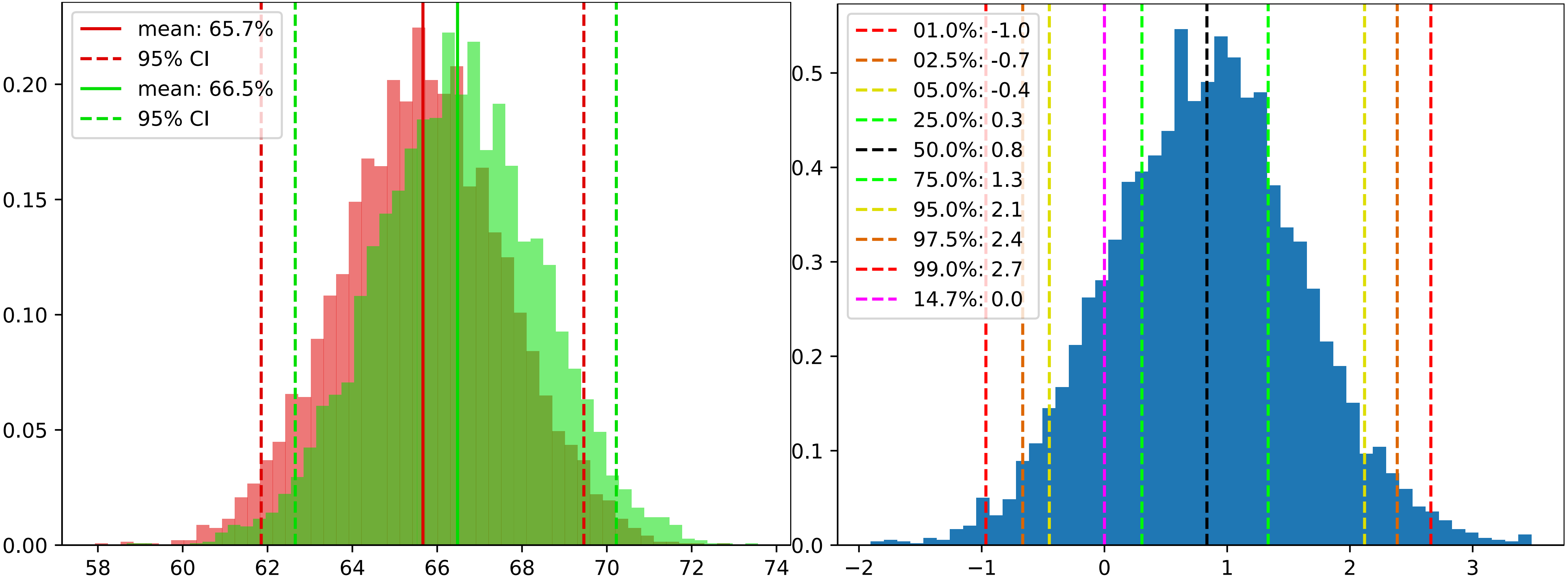
For illustration, consider models for the `isnt_it_ironic` competition utilizing either 3 (red) or 4 (green) in-word character n-grams. On the left, there are distributions of the individual model performances, while on the right there is a distribution of their differences.



The histograms are generated using 50 bins and 5000 resamplings.

Paired Bootstrap Test

For illustration, consider models for the `isnt_it_ironic` competition utilizing either 4 (red) or 5 (green) in-word character n-grams. On the left, there are distributions of the individual model performances, while on the right there is a distribution of their differences.



The histograms are generated using 50 bins and 5000 resamplings.