Statistical Hypothesis Testing, Model Comparison

Jindřich Libovický (reusing materials by Milan Straka)

December 15, 2025

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

Today's Lecture Objectives

After this lecture you should be able to

- Explain foundations of statistical hypothesis testing.
- Reason about multiple comparisons problem.
- Use Bootstrap Resampling and Permutation Tests to compare machine learning models.

Statistical Hypothesis Testing

Statistical Hypothesis Testing

Assume we have a hypothesis testable using observed outcomes 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.

   In simple cases when H_0 is just a negation of H_1, rejecting H_0 amounts to accepting H_1.

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.

Statistical Hypothesis Testing

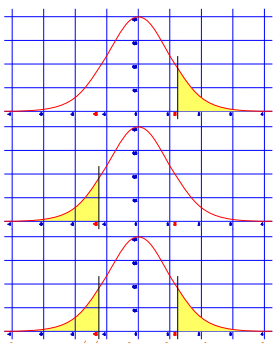
The crucial part of a statistical test is the **test statistic**. It is some summary of the observed data, very often a single value (like mean), which can be used to distinguish the null and the alternative hypothesis.

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



https://upload.wikimedia.org/wikipedia/commons/9/96/DisNormal06.svg

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/lower than/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 a one-sample test on the mean of the differences.

Parametric Test Statistics Distributions

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*.

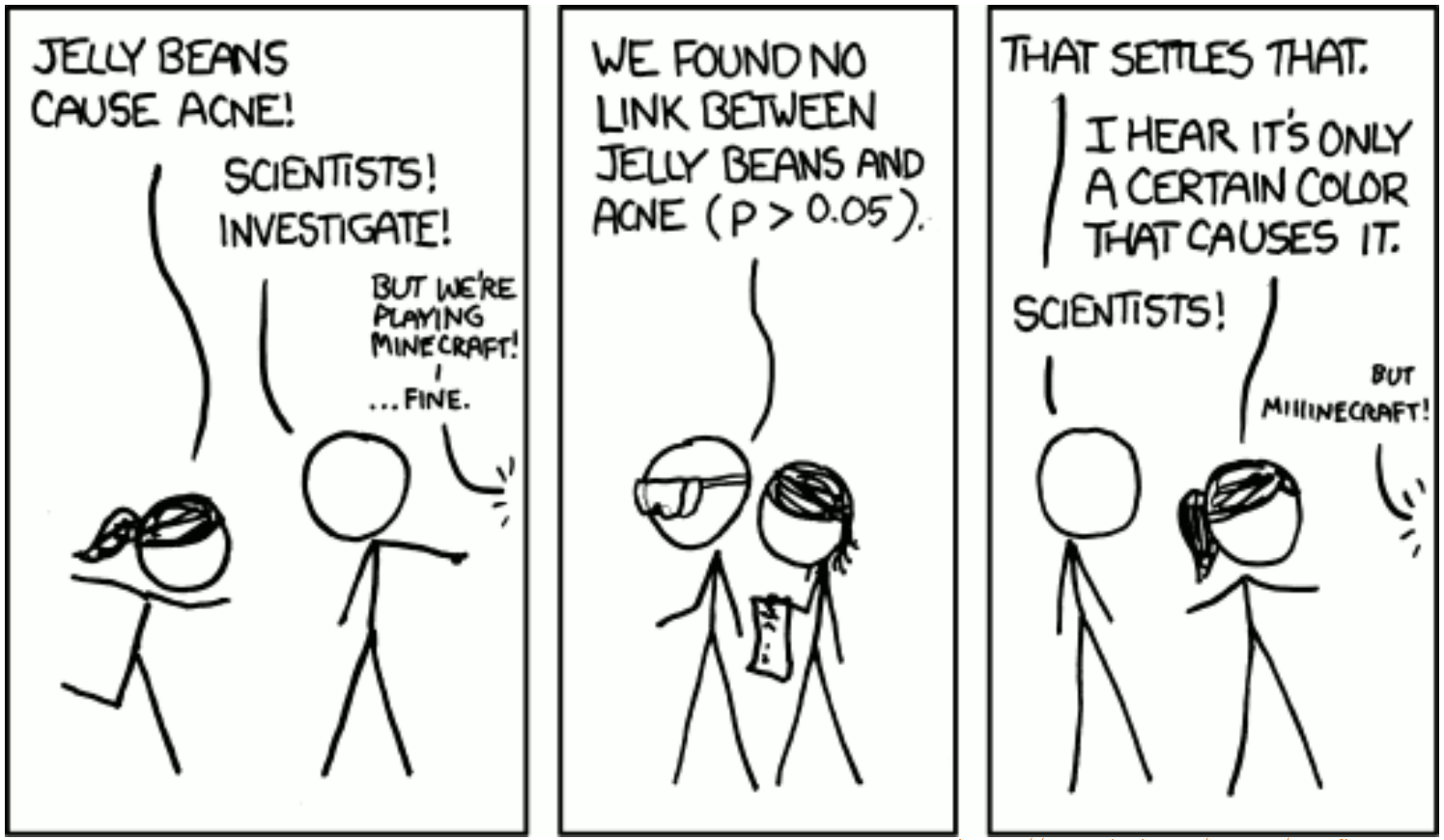
  Therefore, the *t*-test is used when comparing a mean of samples with unknown variance to a given value, or to a mean of samples from another distribution with the same sample size and variance.
- **Chi-squared test** utilizes a test statistic with a chi-squared distribution, which is a distribution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

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

Multiple Comparisons Problem

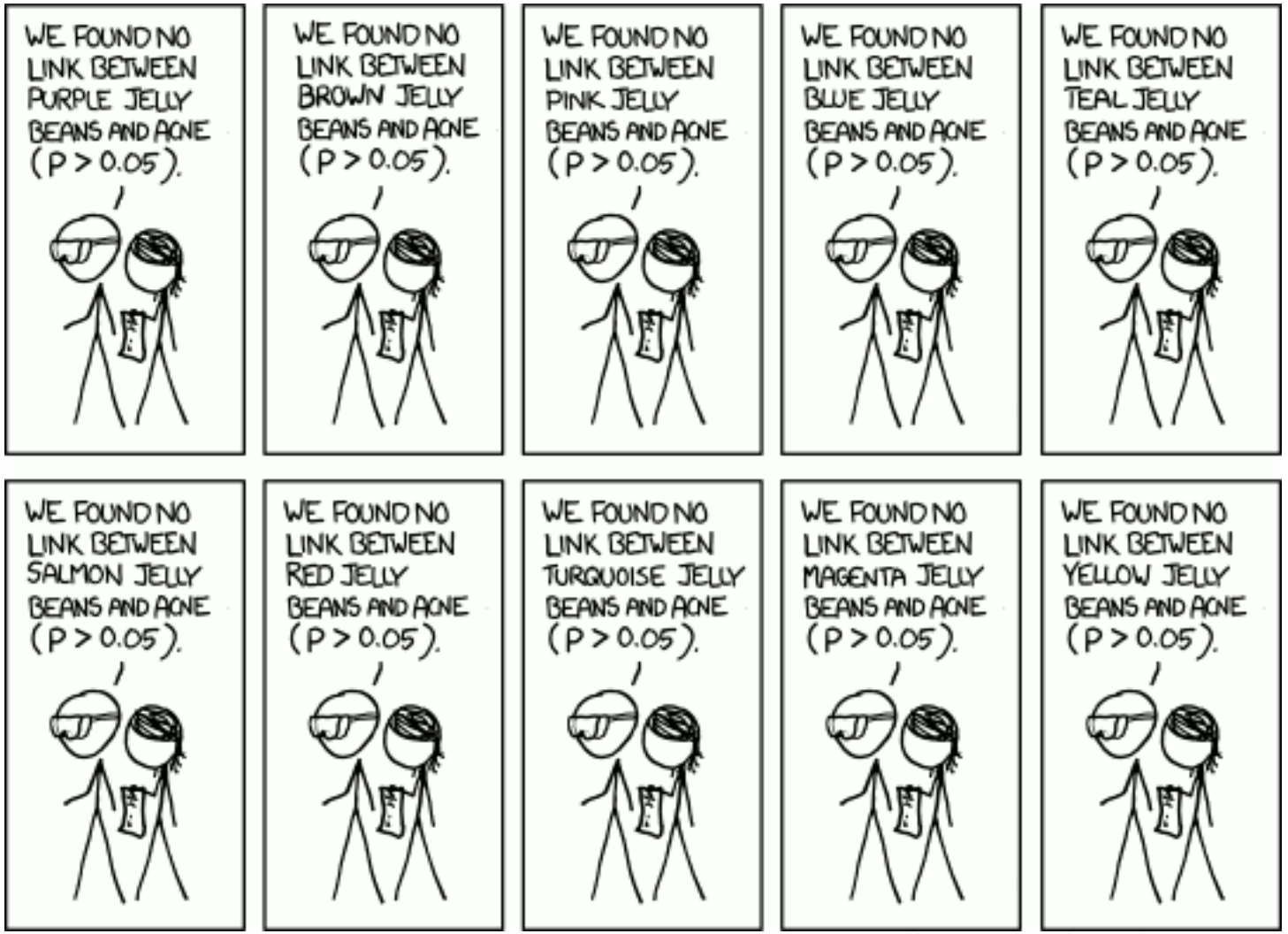
Multiple Comparisons Problem

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



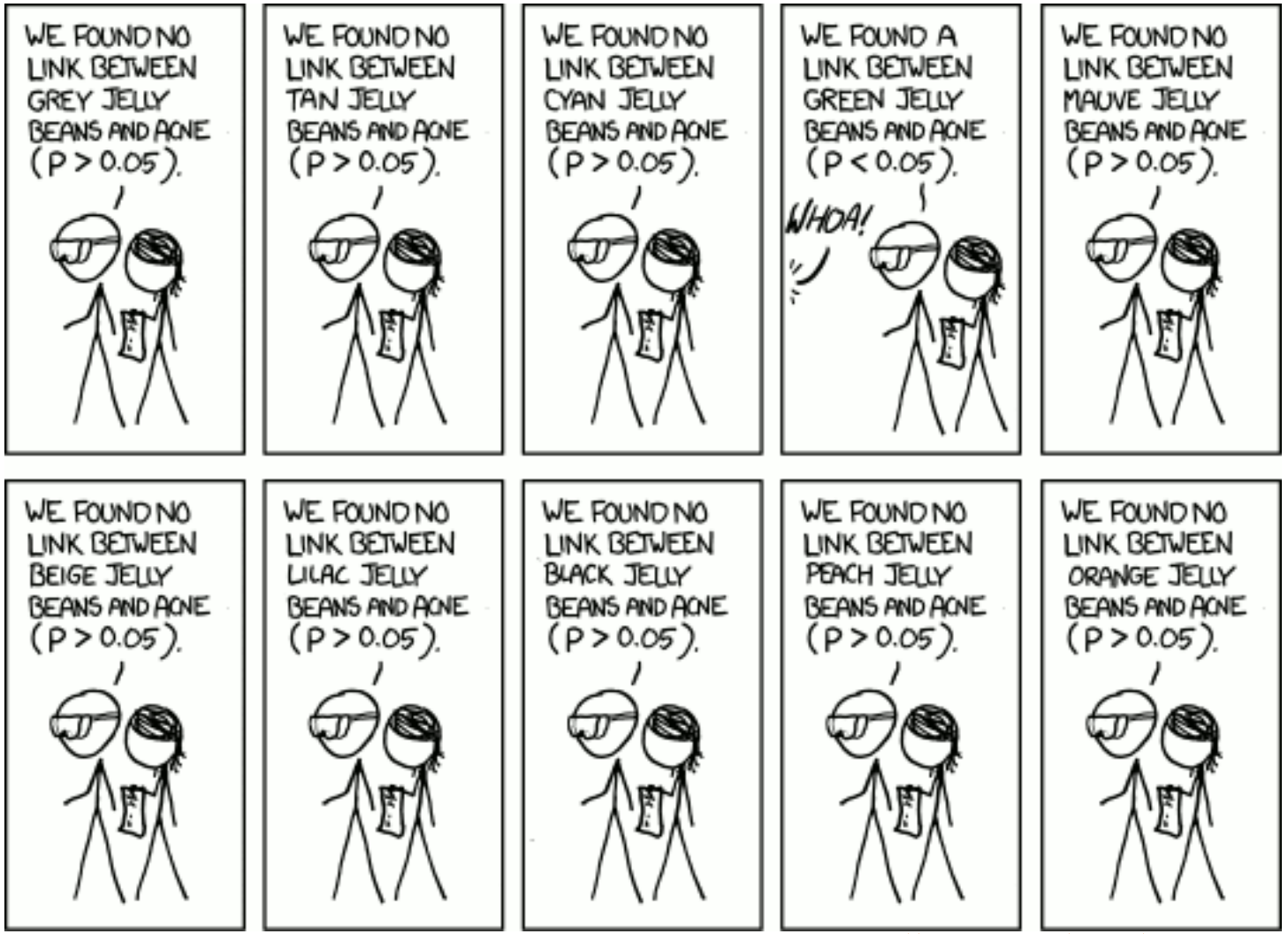
https://imgs.xkcd.com/comics/significant.png

Multiple Comparisons Problem



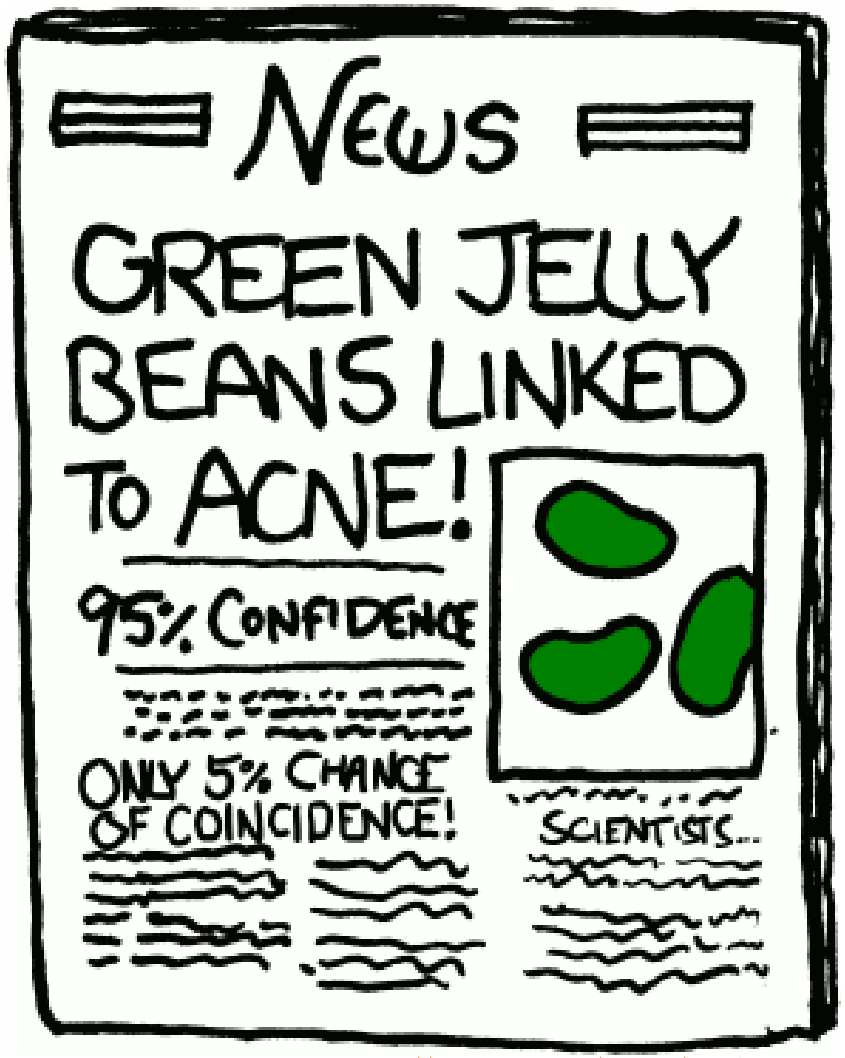
https://imgs.xkcd.com/comics/significant.png

Multiple Comparisons Problem



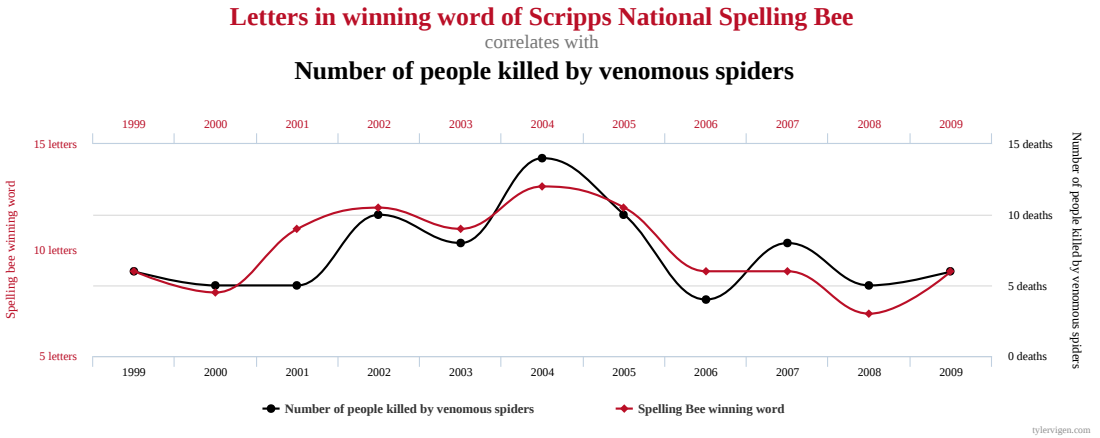
https://imgs.xkcd.com/comics/significant.png

Multiple Comparisons Problem



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

It is problematic if 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

Family-Wise Error Rate

Family-Wise Error Rate

There are several ways to handle the multiple comparisons 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.

$$\mathrm{FWER} = P\left(\bigcup_i \left(p_i \leq \alpha\right)\right).$$

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 $P\left(\bigcup_i A_i\right) \leq \sum_i P(A_i)$, we get that

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

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

Model Comparison

Model Comparison

The goal of model comparison is to test whether some model delivers better performance 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 an 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

Bootstrap Resampling

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

Input: Test set $\{(\boldsymbol{x}_1, t_1), \ldots, (\boldsymbol{x}_N, t_N)\}$, model predictions $\{y(\boldsymbol{x}_1), \ldots, y(\boldsymbol{x}_N)\}$, 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

An analogous approach is sometimes used to perform model comparison – a procedure sometimes called the **paired bootstrap test**.

Even if a two-sample test could be used, 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. Therefore, we perform a **paired** 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 than 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. Finally, we compute the proportion of bootstrap samples where the performance difference is less than or equal to zero.

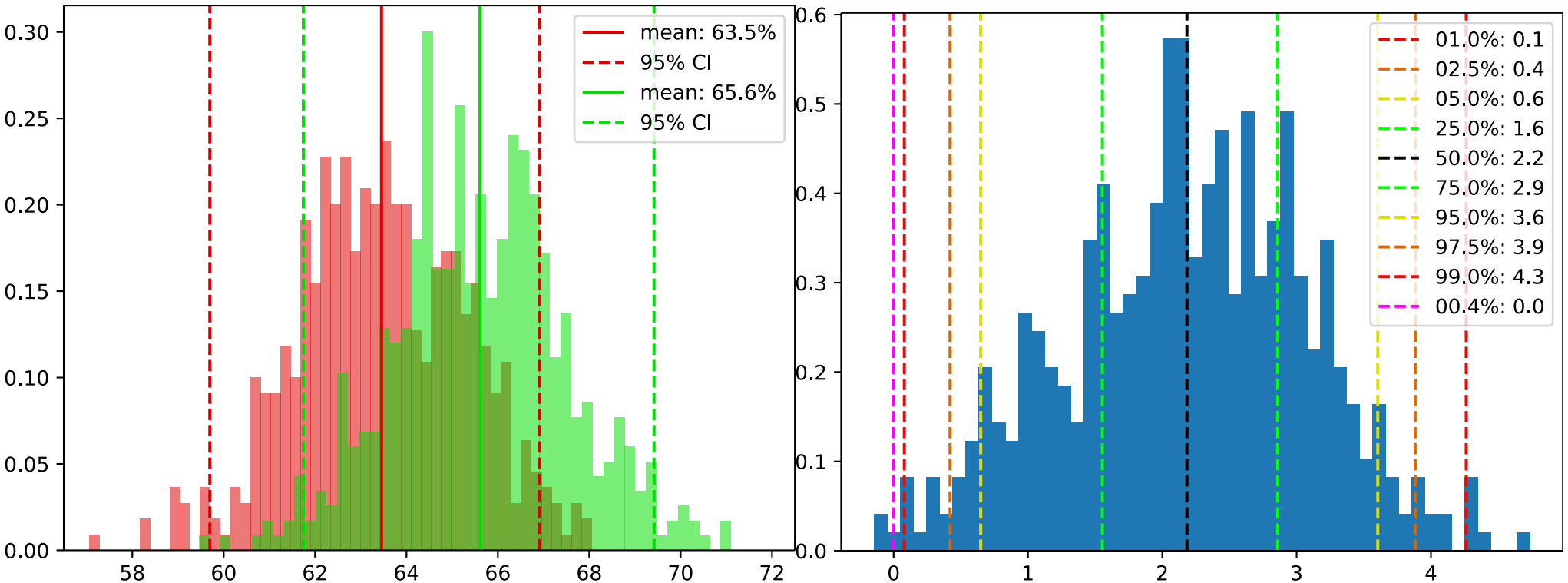
Paired Bootstrap Test Algorithm

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 $\{z(\boldsymbol{x}_1), \ldots, z(\boldsymbol{x}_N)\}$, metric E, number of resamplings R.
Output: Estimated probability of the model y performing worse or equal to z (beware that such a quantity is not a p-value).

- $\mathrm{differences} \leftarrow [\,]$
- 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 than or equal to zero

Paired Bootstrap Test Visualization

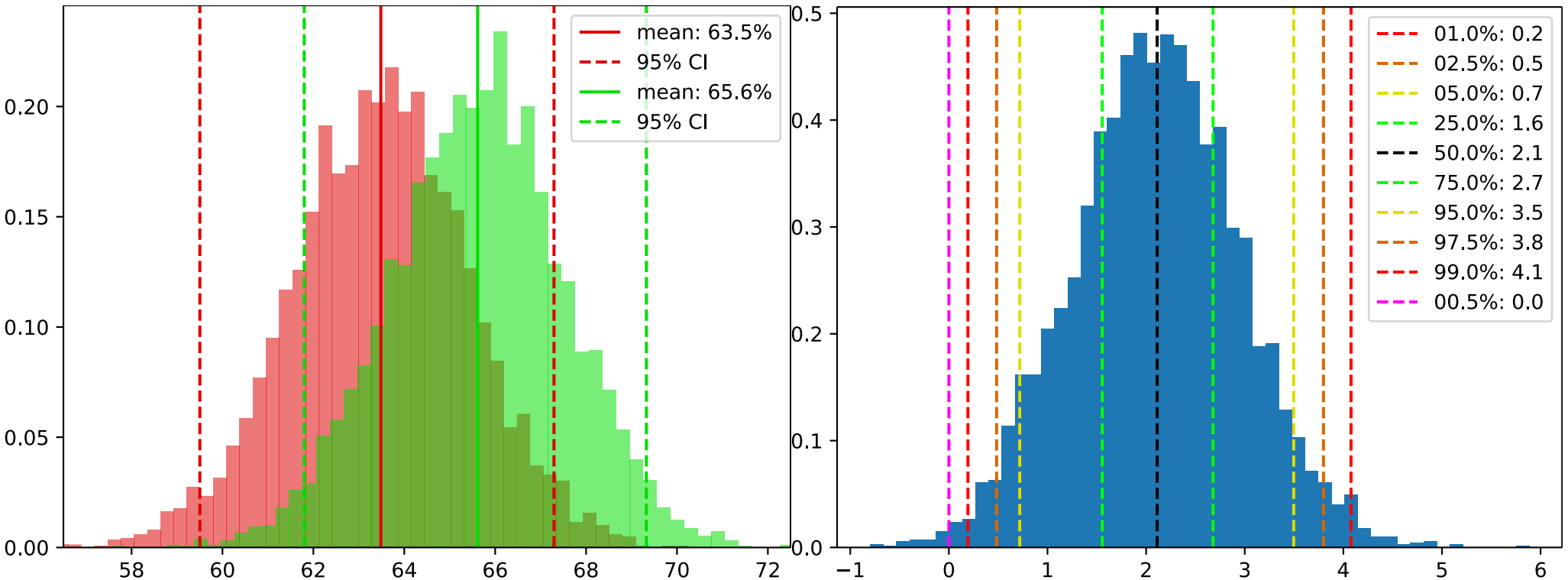
For illustration, consider models for the `isnt_it_ironic` competition utilizing either 3 (red) or 4 (green) in-word character n-grams. On the left, there are distributions of the individual model performances, while on the right there is a distribution of their differences.



The histograms are generated using 50 bins and 500 resamplings.

Paired Bootstrap Test Visualization

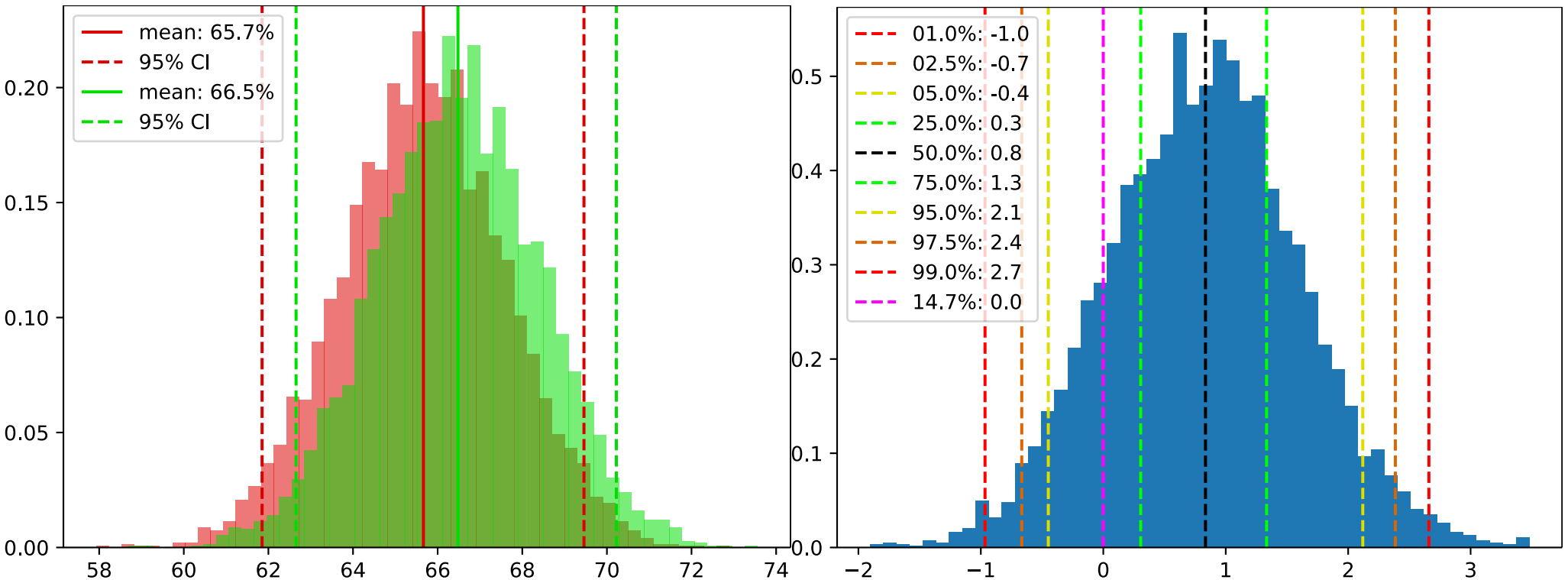
For illustration, consider models for the `isnt_it_ironic` competition utilizing either 3 (red) or 4 (green) in-word character n-grams. On the left, there are distributions of the individual model performances, while on the right there is a distribution of their differences.



The histograms are generated using 50 bins and 5000 resamplings.

Paired Bootstrap Test Visualization

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.

Paired Bootstrap Test Problems

Unfortunately, the value returned by the algorithm is not really a p-value.

The reason is that the distribution of differences was obtained **under the true distribution**.

However, to perform the statistical test, we require the distribution of the test statistic **under the null hypothesis**.

Nevertheless, you can encounter such paired bootstrap tests "in the wild".

Permutation Test

Permutation Test

To obtain a principled p-value for a model comparison, we can turn to a **permutation test**.

The main idea is that

> *If the models are equally good, it does not matter if we utilize predictions from the first or the second one.*

Therefore, if we consider all possible choices of prediction origins, we obtain a distribution of performances under the hypothesis that the models are equally good.

Finally, the p-value is the proportion of permutations where the performance of the model in question is at least as extreme as observed.

Of course, enumerating all assignments is not feasible. Therefore, we sample only some number of random assignments, resulting in a **random** or **Monte Carlo** or **approximate** permutation test.

Random Permutation Test Algorithm

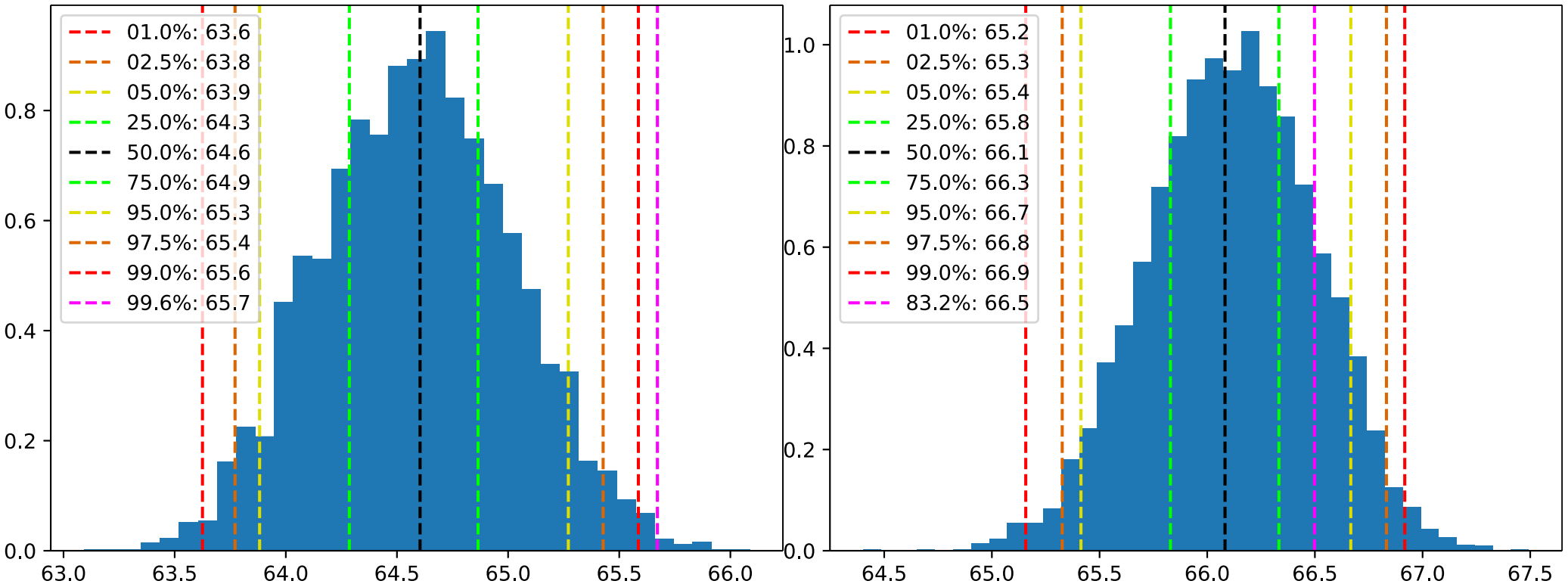
Input: Test set $\{(\boldsymbol{x}_1, t_1), \ldots, (\boldsymbol{x}_N, t_N)\}$, model predictions $\{y(\boldsymbol{x}_1), \ldots, y(\boldsymbol{x}_N)\}$, model predictions $\{z(\boldsymbol{x}_1), \ldots, z(\boldsymbol{x}_N)\}$, metric E, number of resamplings R.

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

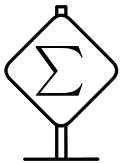
- $\mathrm{performances} \leftarrow [\,]$
- repeat R times:
  - for each test set example, uniformly randomly choose which model to obtain the prediction from
  - measure the performance of the obtained test set prediction using the metric E, and append the score to $\mathrm{performances}$
- return the ratio of the $\mathrm{performances}$ which are greater than or equal to the performance of the model y.

Random Permutation Test Visualization

Again considering the `isnt_it_ironic` models, we compare the 4-vs-3 in-word character n-grams in the left graph and the 5-vs-4 in-word character n-grams in the right graph, using a random permutation test with 5000 resamplings. Note that the resulting p-values are not much different from the probabilities computed by the paired bootstrap test.



Random Permutation Test Strikes Back



Formally, because we did not consider all possible assignments of predictions, we do not obtain the true p-value, but just an approximation of it. In other words, if the algorithm returns β, the probability that the real p-value fulfills

$$p < \beta$$

is only roughly 50%.

Nevertheless, we are usually interested only in deciding whether $p < \alpha$ for a pre-defined α.

In such a case, if $\beta < \alpha$, the probability that $p < \alpha$ does not hold converges to zero as the number of resamplings increases (because of the concentration inequalities, for example Hoeffding's inequality; in other words, the confidence interval of real p-value gets smaller around β as the number of resampling increases). Therefore, it suffices to perform enough resamplings.

For details and a tight bound on the number of resamplings, see the paper

> *Alex Gandy: Sequential Implementation of Monte Carlo Tests With Uniformly Bounded Resampling Risk*
> https://arxiv.org/abs/math/0612488.

Today's Lecture Objectives

After this lecture you should be able to

- Explain foundations of statistical hypothesis testing.
- Reason about multiple comparisons problem.
- Use Bootstrap Resampling and Permutation Tests to compare machine learning models.