Significance
State-of-the-art: Statistical significance testing is mostly ignored in NLP and ML in general. [Marie et al., 2021, Ulmer et al., 2022]

Goal: Start reproducibility analysis by significance testing, w/ and w/o incorporation of variability in meta-parameters and data.

Method:
- Train LMEM on performance scores of baseline and SOTA models, obtained w/ or w/o meta-parameter variation during training.
- Apply GLRT to system effect parameter of LMEM.
- Analyze significance w/ and w/o meta-parameter variation and conditional on data properties.
For given dataset of \( N \) input-output pairs \( \{(x^n, y^n)\}_{n=1}^N \), general form of an LMEM is

\[
Y = X\beta + Zb + \epsilon.
\]

- \( Y \) are \( N \) stacked response variables,
- \( X \) and \( Z \) known design matrices,
- \( \beta \) fixed effects,
- \( b \) random effects,
- \( \epsilon \) residual errors,
- where \( b \sim \mathcal{N}(0, \psi_\theta) \), \( \epsilon \sim \mathcal{N}(0, \Lambda_\theta) \).
Significance Testing with LMEMs

GLRTs based on LMEMS

- **Response variables Y for LMEM training**: Performance evaluation scores of meta-parameter variants of baseline and SOTA systems.

- **GLRT**: Train LMEM with fixed effect $\beta_c$ accounting for competing systems on performance scores of baseline and SOTA systems, and compare their likelihood ratio.

- **Pairing of systems on the sentence level**: Incorporation of random sentence effect $b_s$ allows incorporation of meta-parameter variations and reduces residual variance.
The nested models setup [Pinheiro and Bates, 2000]

- **Restricted null hypothesis model** not distinguishing between systems:

  \[ m_0 : Y = \beta + b_s + \epsilon_{res}, \]

  where \( \beta \) is fixed effect for common mean for both systems, \( b_s \) is random effect for sentence-specific deviation with variance \( \sigma^2_s \), and residual error \( \epsilon_{res} \) with variance \( \sigma^2_{res} \).

- **General model with different means** for baseline and SOTA:

  \[ m_1 : Y = \beta + \beta_c \cdot I_c + b_s + \epsilon_{res}, \]

  where indicator function \( I_c \) activates fixed effect \( \beta_c \) for deviation of competing SOTA model from the baseline mean \( \beta \) when data point was obtained by a SOTA evaluation.
GLRTs in the nested models setup

- Restricted model $m_0$ is special case ("nested") of more general model $m_1$ since it restricts factor $\beta_c$ to zero.

- Let $\ell_0$ be likelihood of restricted model $m_0$, $\ell_1$ be likelihood of more general model $m_1$, intuition of GLRT is to reject the null hypothesis if the test statistic of likelihood ratio

$$\lambda = \frac{\ell_0}{\ell_1}$$

yields values close to zero.
Analyzing significance conditional on data properties

- Extend models $m_0$ and $m_1$ by a **fixed effect** $\beta_d$ modeling a test **data property** $d$ like segment length, readability, or word rarity.
- Add **interaction effect** $\beta_{c:d}$ to assess expected system performance for different levels of $d$.
- Perform GLRT comparing

$$m'_1 : Y = \beta + \beta_d \cdot d + (\beta_c + \beta_{c:d} \cdot d) \cdot \mathbb{I}_c + b_s + \epsilon_{res}$$

to null hypothesis model

$$m'_0 : Y = \beta + \beta_d \cdot d + b_s + \epsilon_{res}.$$
Fine-Tuning Neural Machine Translation (NMT) from human feedback [Kreutzer et al., 2020]

- Baseline: NMT system pre-trained on large out-of-domain data.
- SOTA: Fine-tuning on in-domain data annotated with human error markings or error corrections.
- Response variables for LMEM training: TER scores on test data.

[Snover et al., 2006]

- Data properties: Sentence lengths, binned into short (< 15 words), typical (15 – 55 words), very long (> 55 words).
Meta-parameter grid of attention-based RNN for interactive NMT.

<table>
<thead>
<tr>
<th>Meta-parameter</th>
<th>Grid values</th>
</tr>
</thead>
<tbody>
<tr>
<td>learning_rate</td>
<td>0.0001 0.0003 0.0005 0.003</td>
</tr>
<tr>
<td>random_seed</td>
<td>42 43 44</td>
</tr>
<tr>
<td>encoder_dropout</td>
<td>0 0.2 0.4 0.6</td>
</tr>
<tr>
<td>decoder_dropout</td>
<td>0 0.2 0.4 0.6</td>
</tr>
<tr>
<td>decoder_dropout_hidden</td>
<td>0 0.2 0.4 0.6</td>
</tr>
</tbody>
</table>

[Kreutzer et al., 2020]
TER scores for fine-tuning on human error markings or human post-edits compared to baseline, evaluated on test sentences of 3 length classes.

- SOTA systems trained under three different random seeds, thus one replication for each of three random seeds in LMEM input data.
<table>
<thead>
<tr>
<th></th>
<th>p-value</th>
<th>effect size</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline - marking</td>
<td>0.000332</td>
<td>1.24</td>
</tr>
<tr>
<td>baseline - post-edit</td>
<td>0.0000000358</td>
<td>1.28</td>
</tr>
<tr>
<td>marking - post-edit</td>
<td>0.0252</td>
<td>0.589</td>
</tr>
</tbody>
</table>

- *p*-values and effect sizes (standardized mean difference) for comparison of fine-tuning on human error markings or human post-edits to baseline on very long test sentences.
  - *p*-values < 0.05, medium to very large effect sizes
Extended meta-parameter configuration space by grid search over $4 \times 4 \times 4 \times 4 \times 3 = 768$ trained models for each of the fine-tuning runs.
Significance Testing under Meta-Parameter Variation

- **Meta-parameters:**
  - initial learning rate (`learning_rate`),
  - probability of zeroing out connections during training of encoder (`enc_dropout`), decoder (`dec_dropout`), and hidden decoder layers (`dec_dropout_h`),
  - seed of random number generator (`random_seed`).

- $p$-values for all pairwise differences are above 0.05 across different classes of sentence length.
  - **Significance of result difference lost!**
  - Investigate reasons ➔ reliability analysis!
Advantages of Model-Based Significance Testing with LMEMs

- One-stop approach to test statistical significance of performance differences between machine learning models:
  - Variance in evaluation scores due **meta-parameter variation is incorporated naturally** into training data for LMEM.
  - **No matching of evaluation metrics to significance tests required** [Dror et al., 2020] since test statistics is not based on evaluation metrics, but on MLE parameters of LMEM [van der Vaart, 1998].
  - Further key advantage is analysis of **significance of result difference conditional on data properties**.
  - Power of significance test is **intimately related to reliability** of model under analysis - next chapter!
Goal:

- Applicability to arbitrary and arbitrarily complex evaluation metrics (e.g., non-linear combinations of counts like F-score [Manning et al., 2008], BLEU [Papineni et al., 2002], ROUGE [Lin and Hovy, 2003]).
- No restriction to "mean of samples" metrics which is requirement in parametric tests (t-test, Z-test).
- More powerful than nonparametric tests (e.g. sign test).
Examples

- **Bootstrap resampling:** [Efron and Tibshirani, 1993] Sample itself is a representative “proxy” for the population, sampling distribution of test statistic is estimated by repeatedly sampling (with replacement) from the sample itself.

- **Permutation test:** [Fisher, 1935] Principle of stratified shuffling [Noreen, 1989] allows generation of null-hypothesis conditions by shuffling (sampling without replacement) outputs between two systems at strata that partition the data.
Given test set outputs \((A_0, B_0) = (a_i, b_i)_{i=1}^{N}\), where \(a_i\) is the output of system \(A\), and \(b_i\) is the output of system \(B\), on test instance \(i\).

Compute score difference \(\Delta S_0 = S(A_0) - S(B_0)\) on test data.

For \(k = 1, \ldots, K\):
- Generate bootstrap dataset \(S_k = (A_k, B_k)\) by sampling \(N\) examples from \((a_i, b_i)_{i=1}^{N}\) with replacement.
- Compute score difference \(\Delta S_k = S(A_k) - S(B_k)\) on bootstrap data.

Compute \(\overline{\Delta S_k} = \frac{1}{K} \sum_{k=1}^{K} \Delta S_k\).

Set \(c = 0\).

For \(k = 1, \ldots, K\):
- If \(|\Delta S_k - \overline{\Delta S_k}| \geq |\Delta S_0|\)
  - \(c++\)
- \(p = c/K\).

Reject null hypothesis if \(p\) is less than or equal to rejection level \(\alpha\).
Given test set outputs \((A_0, B_0) = (a_i, b_i)_{i=1}^{N}\), where the first element in the ordered pair \((a_i, b_i)\) is the output of system \(A\), and the second element is the output of system \(B\), on test instance \(i\).

Compute score difference \(\Delta S_0 = S(A_0) - S(B_0)\) on test data.

Set \(c = 0\).

For \(r = 1, \ldots, R\):

Compute shuffled outputs \((A_r, B_r)\) where for each \(i = 1, \ldots, N\):

\[
\text{swap}(a_i, b_i) = \begin{cases} 
(a_i, b_i) & \text{with probability 0.5,} \\
(b_i, a_i) & \text{with probability 0.5.}
\end{cases}
\]

Compute score difference \(\Delta S_r = S(A_r) - S(B_r)\) on shuffled data.

If \(|\Delta S_r| \geq |\Delta S_0|\)

\(c++\)

\(p = c/R\).

Reject null hypothesis if \(p\) is less than or equal to rejection level \(\alpha\).
Bootstrap test makes more Type I errors (i.e., rejecting $H_0$ when it is true) and more Type II errors (i.e., not rejecting $H_0$ when it is false) than the permutation test if bootstrap consistency is not given (i.e., if data from which is resampled are not representative of population). [Canty et al., 2006, Riezler and Maxwell, 2005, Berg-Kirkpatrick et al., 2012]

Designed for comparing a pair of selected systems on a single test set, no easy incorporation of variability in meta-parameters or data!
Problems with Permutation

- Permutation test has great power (i.e., high probability of rejecting $H_0$ when it is false) for large samples [Hoeffding, 1952].
- Stratified shuffling principle needs to be applicable, which is not always the case.
- Designed for comparing a pair of selected systems on a single test set, no easy incorporation of variability in meta-parameters or data!
Significance testing across multiple meta-parameter and data settings

- Bootstrap and permutation tests are designed for comparing a pair of selected systems on a single test set - extensions apply this principle to sampling w/ and w/o replacement from system outputs under meta-parameter variations, but ignore variation of data properties. [Clark et al., 2011, Sellam et al., 2021, Bouthillier et al., 2021].

- Statistical significance test based on the stochastic order/dominance of performance score distributions allow incorporation of meta-parameter variation, but still ignore variation of data properties. [Dror et al., 2019, Ulmer et al., 2022]