Statistical Testing in NLP (II)

CS 690N, Spring 2018
Advanced Natural Language Processing
http://people.cs.umass.edu/~brenocon/anlp2018/

Brendan O’Connor
College of Information and Computer Sciences
University of Massachusetts Amherst
Statistical variability in NLP

• How to trust experiment results, given many sources of variability?
  • How was the text data sampled?
  • How were the annotations sampled?
    • How variably do the human annotators behave?
  • How variable are the computational algorithms?
• Today: Variability due to small sample size
Text data variability

- Mathematically, the easiest case to analyze: What if we resampled the tokens/sentences/documents from a similar population as our current data sample?
- Assume units are sampled i.i.d.; then apply your favorite statistical significance/confidence interval testing technique
  - T-tests, binomial tests, ...
  - Bootstrapping
  - Paired tests
- For
  - 1. Null hypothesis testing
  - 2. Confidence intervals
Null hypothesis test

- Must define a null hypothesis you wish to ~disprove
- p-value = Probability of a result as least as extreme, if the null hypothesis was active
- Example: paired testing of classifiers with exact binomial test (R: binom.test)

\[
P_{\text{Binom}}(k; N, \theta) = \binom{N}{k} \theta^k (1 - \theta)^{N-k}
\]
Statistical tests

• Closed-form tests
  • t-tests, exact binomial test, chi-square tests....
• Bootstrapping

• All methods can give both p-values and confidence intervals
Bootstrapping

• Bootstrapped CI methods
  • Percentile
  • Standard error-based normal approx, etc.
• Theoretical guarantees (under various regularity conditions... for a slightly different CI method...):

\[ \mathbb{P}(\theta \in C_n) = 1 - \alpha - O\left(\frac{1}{\sqrt{n}}\right) \]

• How many samples? 10,000-100,000 (governs monte carlo error; can always make nearly 0)
• Paired bootstrap
• Bootstrapped p-values
• (stopped here 2/27)
• Paired bootstrap test
  • (Subtle, debatable bug?)
• Stat. sig results may not transfer domains
• Researcher effects? Or is paired testing working correctly?
This makes the bootstrap very fast in practice. For our metric, we compute that the bootstrap samples reflect the underlying population distribution quite well. Another major benefit of the bootstrap is that any evaluation metric can be used to compute confidence intervals. This is especially useful for metrics like ROUGE, which can be used for summarization, translation, and dependency parsing. In Figure 2, we plot the ROUGE gain against the p-value for all pairs of the 58 participating systems at TAC 2008. We can see that there is a strong positive correlation between ROUGE gain and confidence, with most points falling on or near the diagonal line. This suggests that the bootstrap estimation noise between different runs is correlated. As a result, we can use the bootstrap to quickly and easily compute confidence intervals for many pairs of systems.

In our first experiment, we use the outputs of the 31 systems participating in the 2008 English summarization track of TAC. For this task, test instances correspond to sentences from a single document collection. The test set consists of 48 sentences from the TAC 2008 English summarization test set, which was released publicly available. We obtained the system outputs from all systems that have been evaluated on machine translation (Callison-Burch et al., 2010). Ideally, for a given task and test set we could obtain results for many pairs of systems. For each pair so that the gain, x, is

x = \frac{\text{metric gain}}{\text{true significance level}}

Our first goal is to explore the relationship between metric gain, x, and p-value. As mentioned, a major benefit of the bootstrap is that any evaluation metric can be used to compute confidence intervals. This is especially useful for metrics like ROUGE, which can be used for summarization, translation, and dependency parsing. In Figure 2, we plot the ROUGE gain against the p-value for all pairs of the 58 participating systems at TAC 2008. We can see that there is a strong positive correlation between ROUGE gain and confidence, with most points falling on or near the diagonal line. This suggests that the bootstrap estimation noise between different runs is correlated. As a result, we can use the bootstrap to quickly and easily compute confidence intervals for many pairs of systems.

In our first experiment, we use the outputs of the 31 systems participating in the 2008 English summarization track of TAC. For this task, test instances correspond to sentences from a single document collection. The test set consists of 48 sentences from the TAC 2008 English summarization test set, which was released publicly available. We obtained the system outputs from all systems that have been evaluated on machine translation (Callison-Burch et al., 2010). Ideally, for a given task and test set we could obtain results for many pairs of systems. For each pair so that the gain, x, is

x = \frac{\text{metric gain}}{\text{true significance level}}

Our first goal is to explore the relationship between metric gain, x, and p-value. As mentioned, a major benefit of the bootstrap is that any evaluation metric can be used to compute confidence intervals. This is especially useful for metrics like ROUGE, which can be used for summarization, translation, and dependency parsing. In Figure 2, we plot the ROUGE gain against the p-value for all pairs of the 58 participating systems at TAC 2008. We can see that there is a strong positive correlation between ROUGE gain and confidence, with most points falling on or near the diagonal line. This suggests that the bootstrap estimation noise between different runs is correlated. As a result, we can use the bootstrap to quickly and easily compute confidence intervals for many pairs of systems.

In our first experiment, we use the outputs of the 31 systems participating in the 2008 English summarization track of TAC. For this task, test instances correspond to sentences from a single document collection. The test set consists of 48 sentences from the TAC 2008 English summarization test set, which was released publicly available. We obtained the system outputs from all systems that have been evaluated on machine translation (Callison-Burch et al., 2010). Ideally, for a given task and test set we could obtain results for many pairs of systems. For each pair so that the gain, x, is

x = \frac{\text{metric gain}}{\text{true significance level}}

Our first goal is to explore the relationship between metric gain, x, and p-value. As mentioned, a major benefit of the bootstrap is that any evaluation metric can be used to compute confidence intervals. This is especially useful for metrics like ROUGE, which can be used for summarization, translation, and dependency parsing. In Figure 2, we plot the ROUGE gain against the p-value for all pairs of the 58 participating systems at TAC 2008. We can see that there is a strong positive correlation between ROUGE gain and confidence, with most points falling on or near the diagonal line. This suggests that the bootstrap estimation noise between different runs is correlated. As a result, we can use the bootstrap to quickly and easily compute confidence intervals for many pairs of systems.

In our first experiment, we use the outputs of the 31 systems participating in the 2008 English summarization track of TAC. For this task, test instances correspond to sentences from a single document collection. The test set consists of 48 sentences from the TAC 2008 English summarization test set, which was released publicly available. We obtained the system outputs from all systems that have been evaluated on machine translation (Callison-Burch et al., 2010). Ideally, for a given task and test set we could obtain results for many pairs of systems. For each pair so that the gain, x, is

x = \frac{\text{metric gain}}{\text{true significance level}}

Our first goal is to explore the relationship between metric gain, x, and p-value. As mentioned, a major benefit of the bootstrap is that any evaluation metric can be used to compute confidence intervals. This is especially useful for metrics like ROUGE, which can be used for summarization, translation, and dependency parsing. In Figure 2, we plot the ROUGE gain against the p-value for all pairs of the 58 participating systems at TAC 2008. We can see that there is a strong positive correlation between ROUGE gain and confidence, with most points falling on or near the diagonal line. This suggests that the bootstrap estimation noise between different runs is correlated. As a result, we can use the bootstrap to quickly and easily compute confidence intervals for many pairs of systems.

In our first experiment, we use the outputs of the 31 systems participating in the 2008 English summarization track of TAC. For this task, test instances correspond to sentences from a single document collection. The test set consists of 48 sentences from the TAC 2008 English summarization test set, which was released publicly available. We obtained the system outputs from all systems that have been evaluated on machine translation (Callison-Burch et al., 2010). Ideally, for a given task and test set we could obtain results for many pairs of systems. For each pair so that the gain, x, is

x = \frac{\text{metric gain}}{\text{true significance level}}

Our first goal is to explore the relationship between metric gain, x, and p-value. As mentioned, a major benefit of the bootstrap is that any evaluation metric can be used to compute confidence intervals. This is especially useful for metrics like ROUGE, which can be used for summarization, translation, and dependency parsing. In Figure 2, we plot the ROUGE gain against the p-value for all pairs of the 58 participating systems at TAC 2008. We can see that there is a strong positive correlation between ROUGE gain and confidence, with most points falling on or near the diagonal line. This suggests that the bootstrap estimation noise between different runs is correlated. As a result, we can use the bootstrap to quickly and easily compute confidence intervals for many pairs of systems.

In our first experiment, we use the outputs of the 31 systems participating in the 2008 English summarization track of TAC. For this task, test instances correspond to sentences from a single document collection. The test set consists of 48 sentences from the TAC 2008 English summarization test set, which was released publicly available. We obtained the system outputs from all systems that have been evaluated on machine translation (Callison-Burch et al., 2010). Ideally, for a given task and test set we could obtain results for many pairs of systems. For each pair so that the gain, x, is

x = \frac{\text{metric gain}}{\text{true significance level}}

Our first goal is to explore the relationship between metric gain, x, and p-value. As mentioned, a major benefit of the bootstrap is that any evaluation metric can be used to compute confidence intervals. This is especially useful for metrics like ROUGE, which can be used for summarization, translation, and dependency parsing. In Figure 2, we plot the ROUGE gain against the p-value for all pairs of the 58 participating systems at TAC 2008. We can see that there is a strong positive correlation between ROUGE gain and confidence, with most points falling on or near the diagonal line. This suggests that the bootstrap estimation noise between different runs is correlated. As a result, we can use the bootstrap to quickly and easily compute confidence intervals for many pairs of systems.
4.2 Empirical Calibration across Domains

Now that we have a way of generating outputs for thousands of pairs of systems, we can check empirically the practical reliability of significance testing. Recall that the bootstrap p-value \( p(\mathcal{X} > \mathcal{X} | H_0) \) is an approximation to \( p(\mathcal{X} > \mathcal{X} | H_0) \). However, we often really want to determine the probability that the new system is better than the baseline on the underlying test distribution or even the distribution from another domain. There is no reason a priori to expect these numbers to coincide.

In our next experiment, we treat the entire Brown corpus, which consists of 24K sentences, as the true population of English sentences. For each system generated in the way described in Section 3.2.5 we compute F1 on all of Brown. Since we are treating the Brown corpus as the actual population of English sentences, for each pair of parsers we can say that the sign of the F1 difference indicates which is the truly better system. Now, we repeatedly resample small test sets from Brown, each consisting of 1,600 sentences, drawn by sampling sentences with replacement. For each pair of systems, and for each resampled test set, we compute p-value \( p(\mathcal{X} > \mathcal{X} | H_0) \) using the bootstrap. Out of the 4K bootstraps computed in this way, 942 had p-value between 0.04 and 0.06, 869 of which agreed with the sign of the F1 difference we saw on the entire Brown corpus. Thus, 92% of the significance tests with p-value in a tight range around 0.05 correctly identified the better system. This result is encouraging. It suggests that statistical significance computed using the bootstrap is reasonably well calibrated. However, test sets are almost never drawn i.i.d. from the distribution of instances the system will encounter in practical use. Thus, we also wish to compute how calibration degrades as the domain of the test set changes. In another experiment, we look at how significance near p-value \( p = 0.05 \) on section 23 of the WSJ corpus predicts performance on sections 22 and 24 and the Brown corpus. This time, for each pair of generated systems we run a bootstrap on section 23. Out of all these bootstraps, 58 system pairs had p-value between 0.04 and 0.06. Of these, only 83% had the same sign of F1 difference on section 23 as they did on section 22, 71% the had the same sign on section 23 as on section 24, and 48% the same sign on section 23 as on the Brown corpus. This indicates that reliability degrades as we switch the domain. In the extreme, achieving a p-value near 0.05 on section 23 provides no information about performance on the Brown corpus.

If we intend to use our system on out-of-domain data, these results are somewhat discouraging. How low does p-value \( p(\mathcal{X} > \mathcal{X} | H_0) \) have to get before we start getting good information about out-of-domain performance? We try to answer this question for this particular parsing task by running the same domain calibration experiment for several different ranges of p-value. The results are shown in Table 1. From these results, it appears that for constituency parsing, when testing on section 23, a p-value level below 0.00125 is required to reasonably predict performance on the Brown corpus.

It should be considered a good practice to include statistical significance testing results with empirical evaluations. The bootstrap in particular is easy to run and makes relatively few assumptions about the task or evaluation metric. However, we have demonstrated some limitations of statistical significance testing for NLP. In particular, while statistical significance is usually a minimum necessary condition to demonstrate that a performance difference is real, it’s also important to consider the relationship between test set performance and the actual goals of the systems being tested, especially if the system will eventually be used on data from a different domain than the test set used for evaluation.

<table>
<thead>
<tr>
<th>Sec. 23 p-value</th>
<th>% Sys. A &gt; Sys. B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sec. 22</td>
</tr>
<tr>
<td>0.00125 - 0.0025</td>
<td>97%</td>
</tr>
<tr>
<td>0.0025 - 0.005</td>
<td>92%</td>
</tr>
<tr>
<td>0.005 - 0.01</td>
<td>92%</td>
</tr>
<tr>
<td>0.01 - 0.02</td>
<td>88%</td>
</tr>
<tr>
<td>0.02 - 0.04</td>
<td>87%</td>
</tr>
<tr>
<td>0.04 - 0.08</td>
<td>83%</td>
</tr>
</tbody>
</table>

Table 1: **Empirical calibration**: p-value on section 23 of the WSJ corpus vs. fraction of comparisons where system A beats system B on section 22, section 24, and the Brown corpus. Note that system pairs are ordered so that A always outperforms B on section 23.
• Statistical significance != practical significance
• CI width, statistical power, data size
• Many other confounds we don’t have models for, but know can be very significant
  • Researcher bias
  • File-drawer bias
  • Generalization (e.g. across domains)
  • Tuning on test sets
  • Reusing test set over multiple papers