4: Significance Testing
Machine Learning and Real-world Data

Simone Teufel

Computer Laboratory
University of Cambridge

Lent 2020
Last session: Zipf’s Law and Heaps’ Law

- **Zipf’s Law**: small number of very high-frequency words; large number of low-frequency words (“long tail”).
- **Heaps’ Law**: as more text is gathered, there will be diminishing returns in terms of discovery of new word types in the tail.
  - We will systematically always encounter new unseen words in new texts.

- **Smoothing works by**
  - lowering the MLE estimate for seen types
  - redistributing this probability to unseen types (e.g. for words in long tail we might encounter during our experiment).
Observed system improvement

- This produced a better system.
- Or at least, you observed higher accuracies.
- Today: we use a statistical test to gather evidence that one system is really better than another system.
Variation in the data

- Documents are different (writing style, length, type of words used, ...)
- Some documents will make it easier for your system to score well, some will make it easier for some other system.
- Maybe you were just lucky and all documents in the test set are in the smoothed system’s favour?
  - This could be the case if you don’t have enough data.
  - This could be the case if the difference in accuracy is small.
- Maybe both systems perform equally well in reality?
Statistical Significance Testing

- Null Hypothesis: two result sets come from the same distribution
  - System 1 is (really) equally good as System 2.

- First, choose a significance level ($\alpha$), e.g., $\alpha = 0.01$ or 0.05.

- We then try to reject the null hypothesis with confidence $1 - \alpha$ (99% or 95% in this case)

- Rejecting the null hypothesis means showing that the observed result is unlikely to have occurred by chance.
Reporting significance

- If we successfully pass the significance test, and only then, we can report:

  “System 1 is different from System 2.” \iff
  “The difference between System 1 and System 2 is statistically significant at $\alpha = 0.01$.”

- Any other such statements are strictly speaking meaningless if all they are based on is a difference in raw accuracy alone (without a stat test).
Sign Test (non-parametric, paired)

- The sign test uses a **binary event model**.
- Here, events correspond to documents.
- Events have binary outcomes:
  - **Positive**: System 1 beats System 2 on this document.
  - **Negative**: System 2 beats System 1 on this document.
  - *(Tie): System 1 and System 2 do equally well on this document / have identical results – more on this later).*

- Binary distribution allows us to calculate the probability that, say, (at least) 1,247 out of 2,000 such binary events are positive.

- Which is identical to the probability that (at most) 753 out of 2,000 are negative.
Binomial Distribution \( B(N, q) \)

- Call the probability of a negative outcome \( q \) (here \( q=0.5 \))
- Probability of observing \( X = k \) negative events out of \( N \):

\[
P_q(X = k|N) = \binom{N}{k} q^k (1 - q)^{N-k}
\]
Binomial Distribution $B(N, q)$

- Call the probability of a negative outcome $q$ (here $q=0.5$)
- Probability of observing $X = k$ negative events out of $N$:
  \[
P_q(X = k|N) = \binom{N}{k} q^k (1 - q)^{N-k}
\]
- At most $k$ negative events:
  \[
P_q(X \leq k|N) = \sum_{i=0}^{k} \binom{N}{i} q^i (1 - q)^{N-i}
\]
Binary Event Model and Statistical Tests

- If the probability of observing our events under the Null Hypothesis is very small (smaller than our pre-selected significance level $\alpha$, e.g., 0.01), we can safely reject the Null hypothesis.
- The $P(X \leq k)$ we just calculated directly gives us the probability we are interested in.
- If $P(X \leq k) \leq 0.01$, this means there is less than a 1% chance that the effect is due to chance.
Two-Tailed vs. One-Tailed Tests

A more conservative, rigorous test would be a non-directional one (though some debate on this!)

- **Testing for statistically significant difference regardless of direction: a two-tailed test**
- We are now interested in the value of $k$ at which 0.01 of the probability exists in the two tails.

- $B(N, 0.5)$ is symmetric so we are now interested in $2P(X \leq k)$
- For the two-tailed test, if $2P(X \leq k) \leq 0.01$, then there is less than a 1% chance that System 1 does not actually beat System 2.
- We’ll be using the two-tailed test for this practical.
Treatment of Ties

- When comparing two systems in classification tasks, it is common for a large number of ties to occur.
- Disregarding ties will tend to affect a study’s statistical power.
- Here, we will treat ties by adding 0.5 events to the positive and 0.5 events to the negative side (and round up at the end).
Today’s Tasks

■ Implement the above-introduced test for statistical significance, so that you can compare two systems.
■ Implementation details on moodle (including helper code as before)
Today’s Tasks

■ Create more (potentially better) systems to use the significance test on.
■ Modify the simple lexicon-based classifier by weighting terms with stronger sentiment more.
■ The pretester will accept a system where strong indicators have weight 2.
  ■ You can also empirically find out the optimal weight.
  ■ We call this process parameter tuning.
  ■ Use the training corpus to set your parameters, then test on the 200 documents as before.
  ■ We should really use a validation corpus, but I haven’t given you one yet... More on this in Session 5.
Formula for smoothing with a constant $\omega$:

$$\hat{P}(w_i|c) = \frac{\text{count}(w_i, c) + \omega}{\left(\sum_{w \in V} \text{count}(w, c)\right) + \omega|V|}$$

We used add-one smoothing in Task 2 ($\omega = 1$).

Using the training corpus, we can optimise the smoothing parameter $\omega$. 

- Chapter 2: The use of statistical tests in research
- Sign test: p. 80–87