

4: Significance Testing

Machine Learning and Real-world Data

Simone Teufel

Computer Laboratory
University of Cambridge

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.
- really = “significantly”

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?
- We need to show that the smoothed system is **significantly** better.

Statistical Significance Testing

- Let's say we observe that System 1 returns a higher overall accuracy than System 2 in our experiment, and now we want to show that System 1 is significantly better.
- Null Hypothesis: two result sets come from the same distribution
 - System 1 is (really) equally good as System 2.
- First, choose a **significance level** (α), 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 significantly better than System 2.” \equiv
“The difference between System 1 and System 2 is statistically significant at $\alpha = 0.01$.”
- Any statements about differences without mentioning significance are strictly speaking **meaningless** if all they are based on is a difference in raw accuracy alone (without a stat test).
- Also note: it is never a measured value (such as accuracy) that is significant; it is always **differences between values** that are significant.

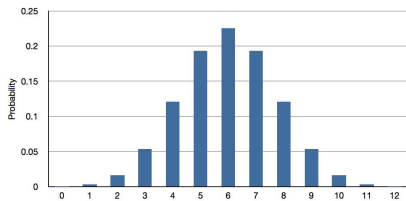
Sign Test (non-parametric, paired)

- 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 :

$$P_q(X = k|N) = \binom{N}{k} q^k (1 - q)^{N-k}$$



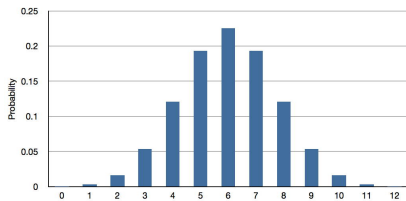
Binomial Distribution $B(N, q)$

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

$$P_q(X = k|N) = \binom{N}{k} q^k (1 - q)^{N-k}$$

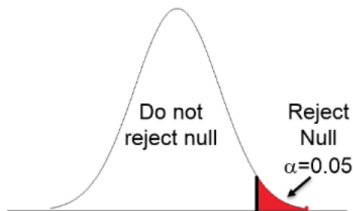
- At most k negative events:

$$P_q(X \leq k|N) = \sum_{i=0}^k \binom{N}{i} q^i (1 - q)^{N-i}$$



Binary Event Model and Statistical Tests

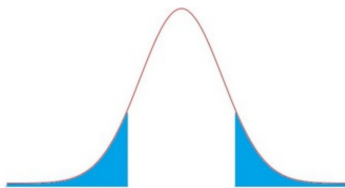
- If the probability of observing the event we saw under the Null Hypothesis is very small (smaller than our pre-selected significance level α , e.g., 0.05), we can safely reject the Null hypothesis.
- The $P(X \leq k)$ we just calculated directly gives us the probability we are interested in.
- If $P(X \leq k) \leq 0.05$, this means there is less than a 5% 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 0.05 of the probability exists in the two tails.
- Due to symmetry of $B(N,0.5)$: if $2P(X \leq k) \leq 0.05$, then there is less than a 5% chance that System 1 does not actually beat System 2.
- You have now measured significance wrt. the chance of an event which is as extreme as the one you observed, or even more extreme
- We'll be using the two-tailed test for this practical.



Specificity and Power of a Test

When we perform significance testing, there are two things we don't want to happen:

- That a test declares a difference when it doesn't exist (Type 1 error).
 - $1 - \alpha$ is the **specificity** of a test.
 - If you suspect that this is happening, change your test!
- That a test declares no difference when it does exist (Type 2 error).
 - β is the probability that this happens. $1 - \beta$ is called the **power** of a test.

What to do if you suspect a problem

Power issues (Type 2 errors):

- Use more data
- Change your system so there is a stronger effect
- Use a more powerful test, for instance permutation test rather than sign test.
- quite common
- Hopefully, your p will decrease and finally reach below α .

Specificity issues (Type 1 errors):

- This could be a problem of how you apply the test

Claims supported by Significance Testing

- Significance tests cannot show that two distributions are the same, they can only potentially ever show a difference.
- As a result, if you pass the test and are able to reject the Null hypothesis, you can report “better”.
- If you fail the test, you have an inconclusive result.
- You are unable to reject the Null hypothesis, but the Null hypothesis is not proven.
- You failed the test because there was too little data **or** because there was no effect.
- If your system performs below your competitor’s system, but you cannot prove a difference with a test, this is not proof that your system is equally good as your competitor’s.

Treatment of Ties

- When comparing two systems in classification tasks, it is common for a large number of ties to occur.
- Disregarding ties will lead to unjustified rejection of Null hypothesis (Type 1 error).
- 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 I

- 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 II

- Create more (potentially better) systems to use the significance test on.
- Improve the simple lexicon-based classifier by weighting terms with stronger sentiment more.
 - You can 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.

Starred Tick — Parameter tuning for NB Smoothing

- Formula for smoothing with a constant ω :

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

- We used add-one smoothing in Task 2 ($\omega = 1$).
- Using the training corpus, we can optimise the smoothing parameter ω .

Literature

- Siegel and Castellan (1988). *Non-parametric statistics for the behavioral sciences*, McGraw-Hill, 2nd. Edition.
 - Chapter 2: The use of statistical tests in research
 - Sign test: p. 80–87