4: Significance Testing

Machine Learning and Real-world Data

Simone Teufel and Ann Copestake

Computer Laboratory
University of Cambridge

Lent 2017

Last session: Zipf's Law and Heaps' Law

- Heaps' Law means that we will systematically always encounter unknown words in new texts.
- Zipf's Law means that the number of low frequency words is large –"long tail"
- Smoothing works by
  - lowering the MLE estimate for seen events
  - redistributing this probability to unseen events (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 the other system.
- Maybe you were just lucky and *all* documents in the test set are in your 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 (p), e.g., $p = 0.01$.
- We then try to reject the null hypothesis with at least probability $1 - p$ (99% in this case)
- That means showing that the observed result is very unlikely to have occurred by chance.

Reporting significance

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

  "The difference between System A and System B is significant at $p \leq 0.01$."

- Any other statements based on raw accuracy differences alone are strictly speaking meaningless.

Sign Test (nonparametric, paired)

- The sign test uses a binary event model.
- Here, n events (corresponding to n documents).
- Events have binary outcomes:
  - Positive: System 1 beats System 2 on this document (*PLUS* times).
  - Negative: System 2 beats System 1 on this document (*MINUS* times).
  - (Tie: System 1 and System 2 do equally well on this document; *NULL* times)
- Call the probability of a positive outcome q (here $q = 0.5$)
- Binary distribution allows us to calculate the probability that, say, at least 1247 out of 2000 such binary events are positive.
- Which equals the probability that at most 753 out of 2000 are negative.

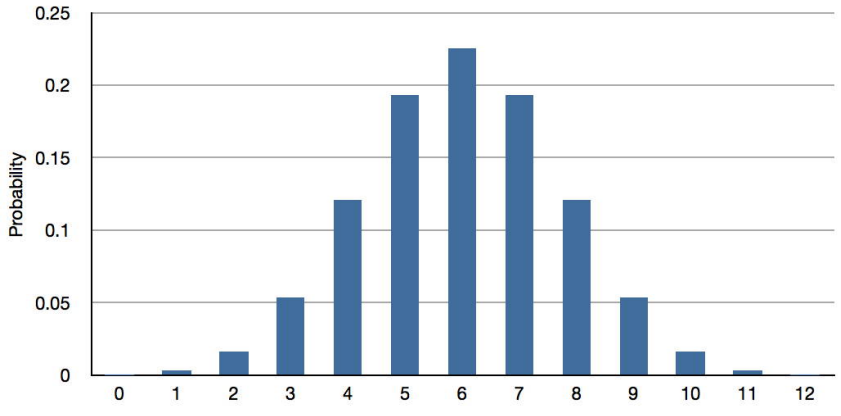
Binomial Distribution $B(N, q)$

- 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:

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

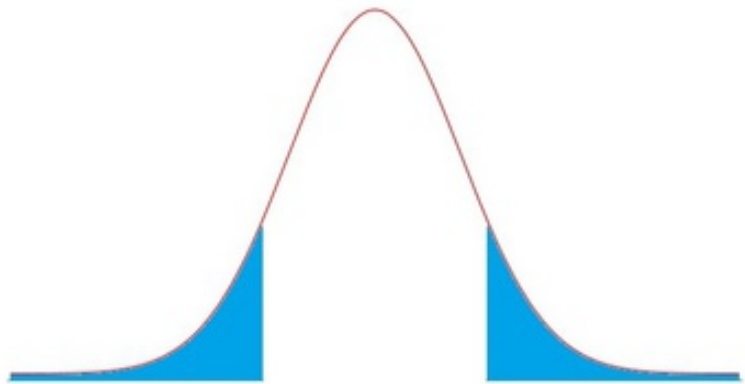


Binary Event Model and Statistical Tests

- If the probability of observing our events under Null Hypothesis is very small (smaller than our pre-selected significance level p, e.g. 1%), we can safely reject the Null hypothesis.
- The $P(X \leq k)$ we just calculated directly gives us the significance level p we are after.
- This means there is only a 1% chance that System 1 and System 2 were not different.
- Well, almost...

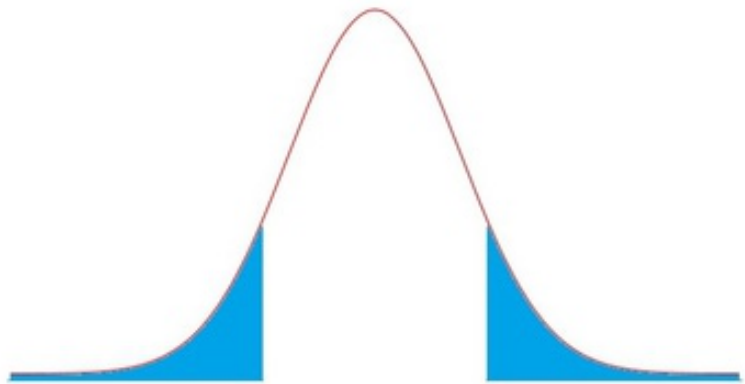
Two-Tailed vs. One-Tailed Tests

- So far we received $P(X \leq k)$ as answer to the question:
  - What is the probability of getting at most 753 negative out of 2000 trials? [One-tailed test]
- But maybe the question should be
  - What is the probability of getting a result that is as extreme as the one I observed (or even more extreme)? [Two-tailed test]
- The answer to this question is $2P(X \leq k)$ (because $B(n, 0.5)$ is symmetric).



Use the two-tailed test

- Why is it safer to use the second question?
  - Because the first question makes the assumption that our chosen system is better.
- Therefore – always use the two-tailed test.



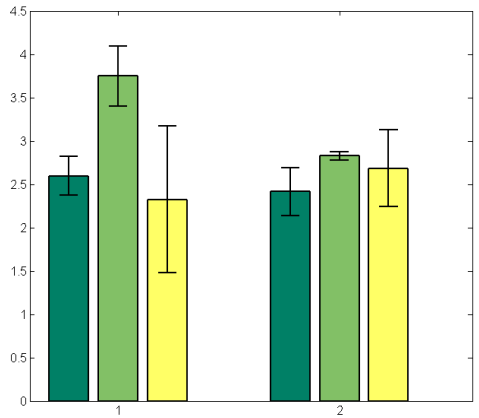
Treatment of Ties

- When comparing two systems in classification tasks, it is common for a large number of ties ("Null" events) to occur.
- Simply disregarding the ties is not an option.
- 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):

$$k = MINUS + \lceil \frac{NULL}{2} \rceil$$

Error bars

- Error bars are another way of communicating statistical significance.
- Error bars show the range of values that might also have been observed under our experimental conditions (instead of the really observed ones), with a given probability.
- 95% error bars are common.
- We can read off the error bars of two systems whether they are significantly different.



Your first task today

- Implement the above-introduced test for statistical significance, so that you can compare two systems.

Your second task today

- Create more (potentially better) systems to use the significance test on.
- Modify the lexicon-based simple 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 setting.
  - Use the training corpus to set your parameters, then test on the 200 documents as before.

Parameter setting – NB Smoothing

- Formula for smoothing with a constant ω:

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

- 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). *Nonparametric statistics for the behavioral sciences*, McGraw-Hill, 2nd. Edition.
  - Chapter 2: The use of statistical tests in research
  - Sign test: p. 80-87