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

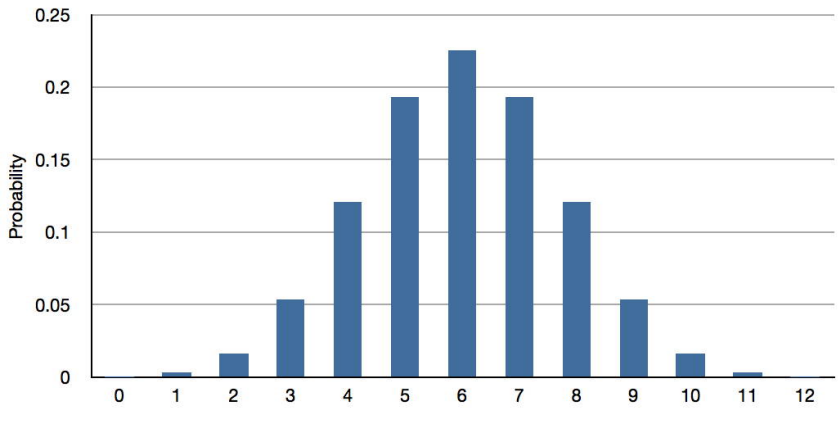
Sign Test (non-parametric, paired)

- 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:

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



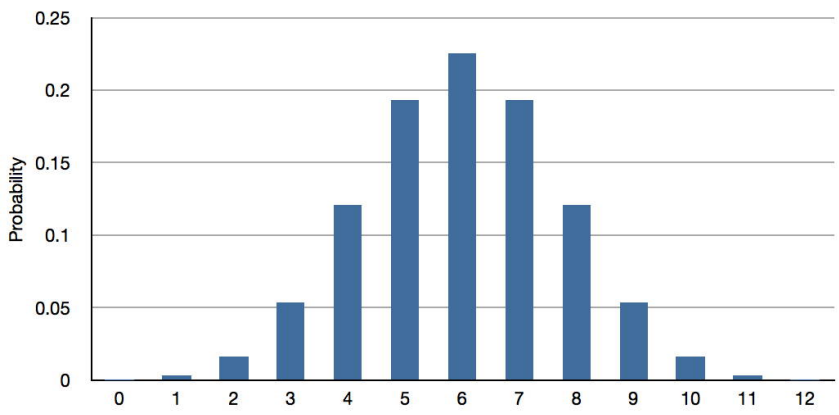
Binomial Distribution $B(N, q)$

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

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

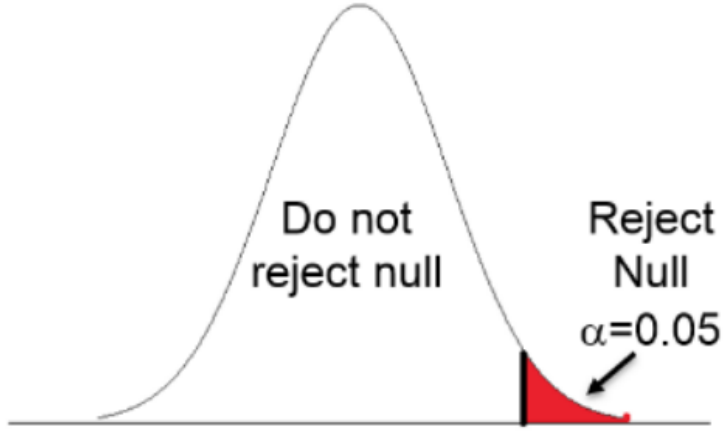
- At most k negative events:

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



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 a less than 5% chance that the effect is due to chance.



Two-Tailed vs. One-Tailed Tests

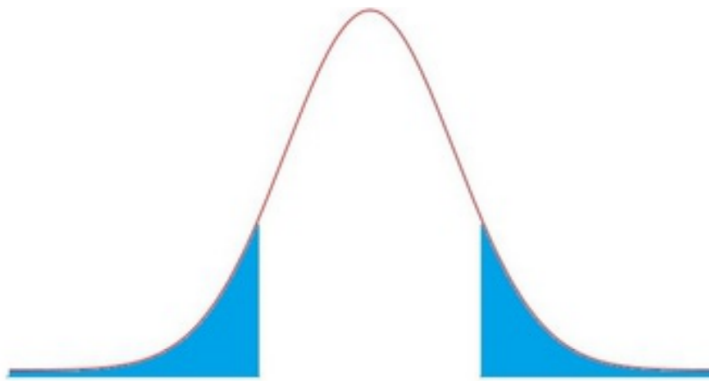
So far, we've been testing difference in a specific direction:

- probability that **at most** 753 out of 2,000 such binary events are negative [One-tailed test]

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

- probability of observing an event that is at least as extreme as 753 out of 2000 **[Two-tailed test]**
- 0.05 of the probability mass gets split up between 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 and System 2 perform equally well.
- We'll be using the two-tailed test.



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 it doesn't exist (Type 1 error).
  - α is the probability that this happens.
  - 1- α is called the specificity of a test.
- That a test declares no difference when it does exist (Type 2 error).
  - β is the probability that this happens.
  - 1-β is called the power of a test.

What to do if you suspect a problem

Power issues (Type 2 errors):

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

Specificity issues (Type 1 errors):

- This should never happen (problem of scientific ethics)
- Develop an intuition when numbers look “too good to be true"
- You probably used the wrong test (which has built-in assumptions that don’t apply)
- Or you applied the test incorrectly.

Treatment of Ties

- Standard textbooks (assuming continuous distributions) often recommend to ignore ties.
- When comparing two systems in classification tasks, it is common for a large number of ties to occur.
- Disregarding ties will lead to unjustifed 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).

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 that doesn’t mean that the Null hypothesis is proven.
- If your system performs **below** your competitor’s system, **and** your significance test fails, the test failure is **not** proof that your system is equally good as your competitor’s.
- You failed the test because there was too little data **or** because there was no effect.

Effect Size and Significance

"System 1 is significantly better than System 2." $\equiv$
"The difference between System 1 and System 2 is statistically significant at $\alpha = 0.01$*."*

- Effect size = difference in measured results between systems
- Significance = binary flag
- Report both, separately but in neighbouring tables
- Any statements about differences without (mentioning) significance are strictly speaking meaningless.
- Also note: the only thing that can ever be significant is a delta, never a single measured value.

Example of what that looks like

	System					
	A	B	C	D	E	F
weighted F	79.1	71.4	74.9	75.2	69.5	70.2
macro-F	69.6	57.7	64.3	63.3	57.8	60.1
BLC	73.0	65.6	65.6	56.7	64.6	63.5
Sub	67.5	57.5	59.0	61.5	55.2	57.1
Super	46.2	19.0	42.4	42.4	26.1	35.7
Abstract	91.9	88.8	90.3	92.5	85.1	84.1

System A: Our best system
System B: System A without property-based indicators
System C: System A without perceptual indicators
System D: System A without WSD
System E: CT21 with WSD
System F: CT21 (original)

System B (71.4)	*				
System C (74.9)					
System D (75.9)					
System E (69.5)	**			*	
System F (70.2)	**				
	System A (79.1)	System B (71.4)	System C (74.9)	System D (75.2)	System E (69.5)

Permutation test results between Results in Weighted F for all systems, * $p < 0.1$, ** $p < 0.05$.

Significance Reporting – the three deadly sins

- Case 1 No significance test performed, statements of “better" and “outperform" and the like are used only based on raw comparisons of numbers

Significance Reporting – the three deadly sins

- Case 1 No significance test performed, statements of “better" and “outperform" and the like are used only based on raw comparisons of numbers $\rightarrow$ **Methodological Unsoundness**

Significance Reporting – the three deadly sins

- Case 1 No significance test performed, statements of “better" and “outperform" and the like are used only based on raw comparisons of numbers $\rightarrow$ **Methodological Unsoundness**
- Case 2 Significance test was performed, but statements of “better" etc are still made for all differences, even the insignificant ones

Significance Reporting – the three deadly sins

- Case 1 No significance test performed, statements of “better" and “outperform" and the like are used only based on raw comparisons of numbers $\rightarrow$ **Methodological Unsoundness**
- Case 2 Significance test was performed, but statements of “better" etc are still made for all differences, even the insignificant ones $\rightarrow$ **Scientific illiteracy**

Significance Reporting – the three deadly sins

- Case 1 No significance test performed, statements of “better" and “outperform" and the like are used only based on raw comparisons of numbers $\rightarrow$ **Methodological Unsoundness**
- Case 2 Significance test was performed, but statements of “better" etc are still made for all differences, even the insignificant ones $\rightarrow$ **Scientific illiteracy**
- No test are performed, statements of “better" etc are made just on basis of raw differences, and the keyword “significant" is still used, often to refer to big effect sizes

Significance Reporting – the three deadly sins

- Case 1 No significance test performed, statements of “better" and “outperform" and the like are used only based on raw comparisons of numbers $\rightarrow$ **Methodological Unsoundness**
- Case 2 Significance test was performed, but statements of “better" etc are still made for all differences, even the insignificant ones $\rightarrow$ **Scientific illiteracy**
- No test are performed, statements of “better" etc are made just on basis of raw differences, and the keyword “significant" is still used, often to refer to big effect sizes $\rightarrow$ **Scientific Fraud**

Significant Digits

- When reporting your results, you will typically get some fraction, not exact numbers
- Where should you round when reporting your means?
- Doesn’t it look “scientific" if I report many post-point digits?
- No, you should only report as many significant digits as are “meaningful".

Significant Digits

- For instance, when deciding whether you should report 0.34 or 0.344, you should only report three digits if the difference between 0.340 and 0.344 is likely to be significant on your dataset.
- This is not easily testable for each case.
- Therefore, err on the safe side
- You don’t want to imply there is significance in numbers when there isn’t (potential fraud case)
- In almost all cases, this means reporting fewer digits
- Sure sign of a rookie paper: 6 post-digital point "significance" digits reported but significance test fails even between systems with adjacent first or second post-digital digits.

Today's Tasks I

- Implement the above-introduced test for statistical significance
- Implementation details on moodle (including helper code as before)
- Make sure you have submitted your judgements on the 4 reviews in session 1

Today's Tasks II

- Use the significance test on pairs of systems
- Create more (potentially better) systems
- 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 call the test corpus "validation corpus" in that case.
  - You will use a validation corpus in the way they were intended in Session 5.
- Record everything in your lab book
- Practice reporting your results in a scientific tone

Trouble-shooting your classifiers

- How many of the predictions of each classifier are positive? How many are negative?
- And how many of these separate predictions are correct?
- Getting accuracies per each class (POS and NEG) can sometimes point out problems.
- Voluntary action for better trouble shooting: calculate these statistics for each classifier you implement.

Starred Tick — Parameter tuning for 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 (ω = 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