Stylometry: Federalist Papers and Chi Squared

- Explain what a chi squared test measures and how to compute it.
- Compute a chi squared value using a spreadsheet and using R.
- Compare chi squared values to help identify authorship.

Most Frequent Words

Word length and sentence length are crude measures of literary style.

Word choice is a slightly more sophisticated measure.

Let's assume temporarily that the Disputed papers were written by Hamilton.

If we combine the Hamilton papers and Disputed papers, and look at the top 10 words most frequently used words, these top 10 words should occur in about the same proportions in both the known Hamilton subset and the Disputed subset (if Hamilton really wrote them and his style was consistent).

Consder this table of the top 10 words from both sub-corpuses combined, and their frequency in the two sub-coprpuses separately.

Word	Both Combined	Hamilton	Disputed
the	12938	10598	2340
of	8835	7370	1465
to	5382	4614	768
in	3368	2833	535
and	3323	2730	593
a	3036	2507	529
be	2885	2300	585
that	2058	1717	341
it	1858	1549	309
is	1615	1329	286
TOTAL	45298	37547	7751

How could we see if the proportions are what we expect, or how far off they are?

Chi Squared Test

Election (fake) examples

Suppose that in some election, we have the following breakdown of votes by age group for a sample of 100 voters.

	Over Age 50	Under Age 50
Democrat	35	25
Republican	30	10

Goal: Is the proportion of Democrats vs. Republicans among older voters similar or different from the proportion of Democrats vs. Republicans among younger voters?

What fraction of the total vote is Democrat vs. Republican?

If there were no relationship between age group and voting preferences, based on the overall fraction of Democrats, how many of the 65 older people would you expect to vote Democrat vs. Repbublican?

And how many of the 35 younger people would you expect to vote Democrat vs. Republican?

We can quantify the differences between these expected counts and the actual, observed counts by looking at

$$\chi = \text{sum}\frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

What is the value of χ?

Why do you think we take the square of the difference, and not just the differences themselves?

Why do you think we divide by the expected? Hint: if your observed minus expected is 6, does this surprise you more if your expected is 10 or if your expected is 1000?

Question. Which of the following is correct:

(a) If χ^2 is small, this means that the voting habits of older vs. younger people are very different, and if χ^2 is large, this means that the voting habits are similar, and differences may be attributable to chance.

(b) If χ^2 is large, this means that the voting habits of older vs. younger people are very different, and if χ^2 is small, this means that the voting habits are similar, and differences may be attributable to chance.

In this example, the Chi Squared statistic measures ...

Compute a Chi Squared value for the attend-skip vs. pass-fail example. According to this (fake) example, does skipping class vs. attending affect your chances of passing vs. failing?

	Attended	Skipped
Pass	25	8
Fail	6	15

In this example, the Chi Squared statistic measures ...

Compute a Chi Squared value for the malaria example on the same spreadsheet. Is the distribution of malaria types significantly different in Asia vs. Africa?

	Asia	Africa
Malaria A	31	14
Malaria B	2	5
Malaria c	53	45

Chi Squared Values for Word Frequency

How can we use the chi squared statistic to help determine authorship of the disputed federalist papers?

- First, compare the Hamilton and Disputed papers.
  - Find the 50 most frequently used words in the combined papers from the Hamilton corpus and the Disputed corpus.
  - Count how many times each of those words is used in the Hamilton corpus and how many times it is used in the Disputed corpus.
  - Compute the chi-squared statistic
- Next, compare the Madison and Disputed pspers.
  - Find the 50 most frequently used words in the combined papers from the Hamilton corpus and the Disputed corpus.
  - Count how many times each of those words is used in the Hamilton corpus and how many times it is used in the Disputed corpus.
  - Compute the chi-squared statistic
- Finally, look at the two chi squared values? Which is bigger?
  - If the Hamilton vs. Disputed χ^2 is bigger than the Madison vs. Disputed χ^2, then _____ likely wrote the disputed papers.
  - If the Madison vs. Disputed χ^2 is bigger than the Hamilton vs. Disputed χ^2, then _____ likely wrote the disputed papers.

- Compute a Chi Squared statistic for Hamilton and Disputed for the top 50 words, using the spreadsheet in tinyurl.com/math115unc > Week 15.

- Compute a Chi Squared statistic for Madison and Disputed for the top 50 words.

- Which author has the biggest chi squared value when comparing this author to the disputed corpus?

- What evidence does this give you for who wrote the disputed papers?

Chi Squared values in Python

See the StylometryFederalistPapersPart2 Python Colab document in tinyurl.com/math115unc > Week 14

What improvements could we make to this method of comparing word frequency?