Entropy and decisions

CSC401/2511 – Natural Language Computing – Spring 2019
Lecture 3, Frank Rudzicz and Chloé Pou-Prom
University of Toronto
This lecture

- Information theory and entropy.
- Decisions.
  - Classification.
  - Significance.

Can we quantify the statistical structure in a model of communication?
Can we quantify the meaningful difference between statistical models?
Information

• Imagine Darth Vader is about to say either “yes” or “no” with equal probability.
  • You don’t know what he’ll say.

• You have a certain amount of uncertainty – a lack of information.
Imagine you then **observe** Darth Vader saying “no”. Your uncertainty is **gone**; you’ve **received information**.

**How much** information do you **receive** about event $E$ when you observe it?

$$I(E) = \log_2 \frac{1}{P(E)}$$

For the units of measurement

$$I(no) = \log_2 \frac{1}{P(no)} = \log_2 \frac{1}{1/2} = 1 \text{ bit}$$
Information

• Imagine Darth Vader is about to roll a fair die.
• You have more uncertainty about an event because there are more possibilities.
• You receive more information when you observe it.

\[ I(5) = \log_2 \frac{1}{P(5)} \]
\[ = \log_2 \frac{1}{1/6} \approx 2.59 \text{ bits} \]
Information is additive

• From \( k \) independent, equally likely events \( E \),

\[
I(E^k) = \log_2 \frac{1}{P(E^k)} = \log_2 \frac{1}{P(E)^k} \quad I(k \text{ binary decisions}) = \log_2 \frac{1}{(1/2)^k} = k \text{ bits}
\]

• For a unigram model, with each of 50K words \( w \) equally likely,

\[
I(w) = \log_2 \frac{1}{1/50000} \approx 15.61 \text{ bits}
\]

and for a sequence of 1K words in that model,

\[
I(w^k) = \log_2 \frac{1}{(1/50000)^{1000}} \approx ???
\]
Information with unequal events

- An information source $S$ emits symbols without memory from a vocabulary $\{w_1, w_2, \ldots, w_n\}$. Each symbol has its own probability $\{p_1, p_2, \ldots, p_n\}$

- What is the average amount of information we get in observing the output of source $S$?

- You still have 6 events that are possible – but you’re fairly sure it will be ‘No’.
Entropy

- **Entropy**: n. the *average* amount of information we get in observing the output of source $S$.

$$H(S) = \sum_i p_i I(w_i) = \sum_i p_i \log_2 \frac{1}{p_i}$$

Note that this is *very* similar to how we define the expected value (i.e., ‘average’) of something:

$$E[X] = \sum_{x \in X} p(x) x$$
Entropy – examples

\[ H(S) = \sum_i p_i \log_2 \frac{1}{p_i} \]

\[ = 0.7 \log_2 (1/0.7) + 0.1 \log_2 (1/0.1) + \ldots \]

\[ = 1.542 \text{ bits} \]

There is **less** average uncertainty when the probabilities are ‘skewed’.

\[ H(S) = \sum_i p_i \log_2 \frac{1}{p_i} = 6 \left( \frac{1}{6} \log_2 \frac{1}{1/6} \right) \]

\[ = 2.585 \text{ bits} \]
Entropy characterizes the distribution

- ‘Flatter’ distributions have a higher entropy because the choices are more equivalent, on average.
- So which of these distributions has a lower entropy?
Low entropy makes decisions easier

- When predicting the next word, e.g., we’d like a distribution with lower entropy.
- Low entropy ≡ less uncertainty
Bounds on entropy

- **Maximum**: uniform distribution $S_1$. Given $M$ choices,
  
  $$H(S_1) = \sum_i p_i \log_2 \frac{1}{p_i} = \sum_i \frac{1}{M} \log_2 \frac{1}{1/M} = \log_2 M$$

- **Minimum**: only one choice, $H(S_2) = p_i \log_2 \frac{1}{p_i} = 1 \log_2 1 = 0$
Coding symbols efficiently

• If we want to transmit Vader’s words efficiently, we can encode them so that more probable words require fewer bits.
• On average, fewer bits will need to be transmitted.

<table>
<thead>
<tr>
<th>Word (sorted)</th>
<th>Linear Code</th>
<th>Huffman Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>000</td>
<td>0</td>
</tr>
<tr>
<td>Yes</td>
<td>001</td>
<td>11</td>
</tr>
<tr>
<td>Destiny</td>
<td>010</td>
<td>101</td>
</tr>
<tr>
<td>Darkside</td>
<td>011</td>
<td>1001</td>
</tr>
<tr>
<td>Maybe</td>
<td>100</td>
<td>10000</td>
</tr>
<tr>
<td>Sure</td>
<td>101</td>
<td>10001</td>
</tr>
</tbody>
</table>

Yes (0.1)  No (0.7)  Maybe (0.04)  Sure (0.03)  Darkside (0.06)  Destiny (0.07)
Coding symbols efficiently

• Another way of looking at this is through the (binary) Huffman tree (r-ary trees are often flatter, all else being equal):

<table>
<thead>
<tr>
<th>Word (sorted)</th>
<th>Linear Code</th>
<th>Huffman Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>000</td>
<td>0</td>
</tr>
<tr>
<td>Yes</td>
<td>001</td>
<td>11</td>
</tr>
<tr>
<td>Destiny</td>
<td>010</td>
<td>101</td>
</tr>
<tr>
<td>Darkside</td>
<td>011</td>
<td>1001</td>
</tr>
<tr>
<td>Maybe</td>
<td>100</td>
<td>10000</td>
</tr>
<tr>
<td>Sure</td>
<td>101</td>
<td>10001</td>
</tr>
</tbody>
</table>
Alternative notions of entropy

• Entropy is equivalently:
  • The average amount of information provided by symbols in a vocabulary,
  • The average amount of uncertainty you have before observing a symbol from a vocabulary,
  • The average amount of ‘surprise’ you receive when observing a symbol,
  • The number of bits needed to communicate that alphabet
    • Aside: Shannon showed that you cannot have a coding scheme that can communicate the vocabulary more efficiently than $H(S)$
Entropy of several variables

- Joint entropy
- Conditional entropy
- Mutual information
Entropy of several variables

- Consider the vocabulary of a meteorologist describing temperature and wetness.
  - Temperature = \{hot, mild, cold\}
  - Wetness = \{dry, wet\}

\begin{align*}
P(W = \text{dry}) &= 0.6, \\
P(W = \text{wet}) &= 0.4
\end{align*}

\begin{align*}
P(T = \text{hot}) &= 0.3, \\
P(T = \text{mild}) &= 0.5, \\
P(T = \text{cold}) &= 0.2
\end{align*}

\begin{align*}
H(W) &= 0.6 \log_2 \frac{1}{0.6} + 0.4 \log_2 \frac{1}{0.4} = 0.970951 \text{ bits} \\
H(T) &= 0.3 \log_2 \frac{1}{0.3} + 0.5 \log_2 \frac{1}{0.5} + 0.2 \log_2 \frac{1}{0.2} = 1.48548 \text{ bits}
\end{align*}

But $W$ and $T$ are not independent, $P(W, T) \neq P(W)P(T)$

Example from Roni Rosenfeld
Joint entropy

• **Joint Entropy:** *n.* the **average** amount of information needed to specify **multiple** variables **simultaneously**.

\[ H(X, Y) = \sum_x \sum_y p(x, y) \log_2 \frac{1}{p(x, y)} \]

• **Hint:** this is *very* similar to univariate entropy – we just replace univariate probabilities with joint probabilities and sum over everything.
Entropy of several variables

• Consider joint probability, \( P(W, T) \)

<table>
<thead>
<tr>
<th></th>
<th>cold</th>
<th>mild</th>
<th>hot</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>dry</td>
<td>0.1</td>
<td>0.4</td>
<td>0.1</td>
<td>0.6</td>
</tr>
<tr>
<td>wet</td>
<td>0.2</td>
<td>0.1</td>
<td>0.1</td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td>0.3</td>
<td>0.5</td>
<td>0.2</td>
<td>1.0</td>
</tr>
</tbody>
</table>

• Joint entropy, \( H(W, T) \), computed as a sum over the space of joint events (\( W = w, T = t \))

\[
H(W, T) = 0.1 \log_2 \frac{1}{0.1} + 0.4 \log_2 \frac{1}{0.4} + 0.1 \log_2 \frac{1}{0.1} + 0.2 \log_2 \frac{1}{0.2} + 0.1 \log_2 \frac{1}{0.1} + 0.1 \log_2 \frac{1}{0.1} = 2.32193 \text{ bits}
\]

Notice \( H(W, T) \approx 2.32 < 2.46 \approx H(W) + H(T) \)
Entropy given knowledge

- In our example, joint entropy of two variables together is lower than the sum of their individual entropies
  - \( H(W, T) \approx 2.32 < 2.46 \approx H(W) + H(T) \)

- Why?

- Information is shared among variables
  - There are dependencies, e.g., between temperature and wetness.
  - E.g., if we knew exactly how wet it is, is there less confusion about what the temperature is ... ?
Conditional entropy

- **Conditional entropy**: the **average** amount of information needed to specify one variable given that you know another.

- A.k.a ‘equivocation’

\[ H(Y|X) = \sum_{x \in X} p(x)H(Y|X = x) \]

- **Hint**: this is *very* similar to how we compute expected values in general distributions.
Entropy given knowledge

- Consider **conditional** probability, $P(T|W)$

\[
P(T|W) = \frac{P(W, T)}{P(W)}
\]

<table>
<thead>
<tr>
<th>$P(W,T)$</th>
<th>$T = \text{cold}$</th>
<th>mild</th>
<th>hot</th>
</tr>
</thead>
<tbody>
<tr>
<td>$W = \text{dry}$</td>
<td>0.1</td>
<td>0.4</td>
<td>0.1</td>
</tr>
<tr>
<td>wet</td>
<td>0.2</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>0.3</td>
<td>0.5</td>
<td>0.2</td>
<td>1.0</td>
</tr>
</tbody>
</table>

| $P(T | W)$ | $T = \text{cold}$ | mild | hot |
|----------|------------------|------|-----|
| $W = \text{dry}$ | 0.1/0.6 | 0.4/0.6 | 0.1/0.6 | **1.0** |
| wet | 0.2/0.4 | 0.1/0.4 | 0.1/0.4 | **1.0** |
Entropy given knowledge

• Consider conditional probability, \( P(T|W) \)

| \( P(T|W) \) | \( T = \text{cold} \) | \( \text{mild} \) | \( \text{hot} \) |
|----------------|----------------|---------|--------|
| \( W = \text{dry} \) | 1/6 | 2/3 | 1/6 | 1.0 |
| \( \text{wet} \) | 1/2 | 1/4 | 1/4 | 1.0 |

\[
H(T|W = \text{dry}) = H\left(\left\{\frac{1}{6}, \frac{2}{3}, \frac{1}{6}\right\}\right) = 1.25163 \text{ bits}
\]

\[
H(T|W = \text{wet}) = H\left(\left\{\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right\}\right) = 1.5 \text{ bits}
\]

• Conditional entropy combines these:

\[
H(T|W) = [p(W = \text{dry})H(T|W = \text{dry})] + [p(W = \text{wet})H(T|W = \text{wet})]
\]

\[
= 1.350978 \text{ bits}
\]
Equivocation removes uncertainty

• Remember $H(T) = 1.48548$ bits
• $H(W, T) = 2.32193$ bits
• $H(T|W) = 1.350978$ bits

• How much does $W$ tell us about $T$?
  • $H(T) - H(T|W) = 1.48548 - 1.350978 \approx 0.1345$ bits
  • Well, a little bit!

Entropy (i.e., confusion) about temperature is reduced if we know how wet it is outside.
Perhaps $T$ is more informative?

- Consider another conditional probability, $P(W|T)$

| $P(W|T)$ | $T = \text{cold}$ | mild | hot |
|----------|-------------------|------|-----|
| $W = \text{dry}$ | 0.1/0.3 | 0.4/0.5 | 0.1/0.2 |
| $W = \text{wet}$ | 0.2/0.3 | 0.1/0.5 | 0.1/0.2 |
| $H(W|T)$ | 1.0 | 1.0 | 1.0 |

- $H(W|T = \text{cold}) = H\left(\frac{1}{3}, \frac{2}{3}\right) = 0.918295$ bits
- $H(W|T = \text{mild}) = H\left(\frac{4}{5}, \frac{1}{5}\right) = 0.721928$ bits
- $H(W|T = \text{hot}) = H\left(\frac{1}{2}, \frac{1}{2}\right) = 1$ bit
- $H(W|T) = 0.8364528$ bits
Equivocation removes uncertainty

- $H(T) = 1.48548$ bits
- $H(W) = 0.970951$ bits
- $H(W, T) = 2.32193$ bits
- $H(T|W) = 1.350978$ bits
- $H(T) - H(T|W) \approx 0.1345$ bits

How much does $T$ tell us about $W$ on average?
- $H(W) - H(W|T) = 0.970951 - 0.8364528 \approx 0.1345$ bits

Interesting ... is that a coincidence?
Mutual information

- **Mutual information**: *n.* the **average** amount of information **shared** between variables.

\[
I(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \\
= \sum_{x,y} p(x, y) \log_2 \frac{p(x,y)}{p(x)p(y)}
\]

- **Hint**: The amount of uncertainty **removed** in variable \(X\) if you know \(Y\).
- **Hint2**: If \(X\) and \(Y\) are **independent**, \(p(x, y) = p(x)p(y)\), then

\[
\log_2 \frac{p(x,y)}{p(x)p(y)} = \log_2 1 = 0 \ \forall x, y \ – \text{there is no mutual information!}
\]
Relations between entropies

\[ H(X, Y) = H(X) + H(Y) - I(X; Y) \]
Preview – the noisy channel

• Messages can get **distorted** when passed through a **noisy** conduit – _how much information is lost/retained?_

  - Signals
  - Symbols
  - Languages

  **Sexual abuse**

  **Locker room talk**

  **Hello, computer**

  **Bonjour, ordinateur**
Relating corpora
Relatedness of two distributions

• How similar are two probability distributions?
• e.g., Distribution $P$ learned from Kylo Ren
  Distribution $Q$ learned from Darth Vader
Relatedness of two distributions

- A Huffman code based on Vader (Q) instead of Kylo (P) will be less *efficient* at coding symbols that Kylo will say.
- What is the **average number of extra bits** required to code symbols from P when using a code based on Q?
Kullback-Leibler divergence

- **KL divergence**: \( n. \) the average log difference between the distributions \( P \) and \( Q \), relative to \( Q \). a.k.a. relative entropy.

  caveat: we assume \( 0 \log 0 = 0 \)
Kullback-Leibler divergence

\[ D_{KL}(P \parallel Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \]

- Why log \( \frac{P(i)}{Q(i)} \)?
  - \( \log \frac{P(i)}{Q(i)} = \log P(i) - \log Q(i) = \log \left( \frac{1}{Q(i)} \right) - \log \left( \frac{1}{P(i)} \right) \)
- If word \( w_i \) is less probable in \( Q \) than \( P \) (i.e., it carries more information), it will be Huffman encoded in more bits, so when we see \( w_i \) from \( P \), we need \( \log \frac{P(i)}{Q(i)} \) more bits.
Kullback-Leibler divergence

- KL divergence:
  - is somewhat like a ‘distance’:
    - $D_{KL}(P||Q) \geq 0 \ \forall P, Q$
    - $D_{KL}(P||Q) = 0$ iff $P$ and $Q$ are identical.
  - is not symmetric, $D_{KL}(P||Q) \neq D_{KL}(Q||P)$

- Aside:
  \[
  I(P; Q) = D_{KL}(P(X,Y)||P(X)P(Y))
  \]
Kullback-Leibler divergence

- KL divergence generalizes to **continuous** distributions.
- Below, $D_{KL}(\text{green} || \text{blue}) > D_{KL}(\text{purple} || \text{blue})$
Applications of KL divergence

- Often used towards some other purpose, e.g.,
  - In evaluation to say that purple is a better model than green of the true distribution blue.
  - In machine learning to adjust the parameters of purple to be, e.g., less like green and more like blue.
Entropy as intrinsic LM evaluation

• **Cross-entropy** measures how difficult it is to encode an event drawn from a *true probability* $p$ given a **model** based on a distribution $q$.

• What if we don’t know the *true probability* $p$?
  • We’d have to estimate $p$.
  • We estimate $p$ by estimating the **probability of a test corpus** $C$ using the distribution $q$:

$$P_q(C)$$
Probability of a corpus?

- The probability $P(C)$ of a corpus $C$ requires similar assumptions that allowed us to compute the probability $P(s_i)$ of a sentence $s_i$.

<table>
<thead>
<tr>
<th>Chain rule</th>
<th>Sentence</th>
<th>Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(s_i) = P(w_1) \prod_{t=2}^{n} P(w_t</td>
<td>w_{1:(t-1)})$</td>
<td>$P(C) = P(w_1) \prod_{t=2}^{</td>
</tr>
<tr>
<td>Approx.</td>
<td>$P(s_i) \approx \prod_{t} P(w_t)$</td>
<td>$P(C) \approx \prod_{i} P(s_i)$</td>
</tr>
</tbody>
</table>

- Regardless of the LM used for $P(s_i)$, we can assume complete independence between sentences.
Intrinsic evaluation – Cross-entropy

• **Cross-entropy** of a LM $M$ and a *new* test corpus $C$ with size $\|C\|$ (total number of words), where sentence $s_i \in C$, is *approximated* by:

$$H(C; M) = -\frac{\log_2 P_M(C)}{\|C\|} = -\frac{\sum_i \log_2 P_M(s_i)}{\sum_i \|s_i\|}$$

• **Perplexity** comes from this definition:

$$PP_M(C) = 2^{H(C; M)}$$
Decisions
Deciding what we know

- **Anecdotes** are often useless except as proofs by contradiction.
  - E.g., “I saw Google used as a verb” does not mean that Google is **always** (or even **likely** to be) a verb, just that it is **not always** a noun.

- **Shallow statistics** are often not enough to be truly meaningful.
  - E.g., “My ASR system is 95% accurate on my test data. Yours is only 94.5% accurate, you horrible knuckle-dragging idiot.”
    - What if the test data was **biased** to favor my system?
    - What if we only used a **very small** amount of data?

- Given all this potential ambiguity, we need a **test** to see if our statistics actually **mean** something.
Differences due to sampling

• We saw that KL divergence essentially measures how different two distributions are from each other.

• But what if their difference is due to randomness in sampling?

• How can we tell that a distribution is really different from another?
Hypothesis testing

• Often, we assume a null hypothesis, $H_0$, which states that the two distributions are the same (i.e., come from the same underlying model, population, or phenomenon).

• We reject the null hypothesis if the probability of it being true is too small.
  • This is often our goal – e.g., if my ASR system beats yours by 0.5%, I want to show that this difference is not a random accident.
  • I assume it was an accident, then show how nearly impossible that is.

• As scientists, we have to be very careful to not reject $H_0$ too hastily.
  • How can we ensure our diligence?
Confidence

• We reject $H_0$ if it is too improbable.
  • How do we determine the value of ‘too’?

• Significance level $\alpha$ ($0 \leq \alpha \leq 1$) is the maximum probability that two distributions are identical allowing us to disregard $H_0$.
  • In practice, $\alpha \leq 0.05$. Usually, it’s much lower.
  • Confidence level is $\gamma = 1 - \alpha$
  • E.g., a confidence level of 95% ($\alpha = 0.05$) implies that we expect that our decision is correct 95% of the time, regardless of the test data.
Confidence

• We will briefly see three types of statistical tests that can tell us how confident we can be in a claim:

1. A $t$-test, which usually tests whether the means of two models are the same. There are many types, but most assume Gaussian distributions.

2. An analysis of variance (ANOVA), which generalizes the $t$-test to more than two groups.

3. The $\chi^2$ test, which evaluates categorical (discrete) outputs.
1. The *t*-test

- The **t-test** is a method to compute if distributions are significantly different from one another.

- It is based on the mean ($\bar{x}$) and variance ($\sigma$) of $N$ samples.
- It compares $\bar{x}$ and $\sigma$ to $H_0$ which states that the samples are drawn from a distribution with a **mean** $\mu$.

- If

  $$t = \frac{\bar{x} - \mu}{\sqrt{\frac{\sigma^2}{N}}}$$

  (the “t-statistic”) is large enough, we can reject $H_0$.

There are actually several types of *t*-tests for different situations...
Example of the $t$-test: tails

- Imagine the average tweet length of a McGill ‘student’ is $\mu = 158$ chars.
- We sample $N = 200$ UofT students and find that our average tweet is $\bar{x} = 169$ chars (with $\sigma^2 = 2600$).
- Are UofT tweets significantly longer than much worse McGill tweets?

- We use a ‘one-tailed’ test because we want to see if UofT tweet lengths are significantly higher.
  - If we just wanted to see if UofT tweets were significantly different, we’d use a two-tailed test.
Example of the $t$-test: freedom

- Imagine the average tweet length of a McGill ‘student’ is $\mu = 158$ chars.
- We sample $N = 200$ UofT students and find that our average tweet is $\bar{x} = 169$ chars (with $\sigma^2 = 2600$).
- Are UofT tweets significantly longer than much worse McGill tweets?

- **Degrees of freedom (d.f.):** *n.pl.* In this $t$-test, this is the sum of the number of observations in each group, minus 2 (because there are two groups).

- In our example, we have $N_{UofT} = 200$ for DCS students, but because we don’t sample at McGill, $N_{McGill} \approx \infty$, so $d.f. = \infty$.
  - (this example is adapted from Manning & Schütze)
Example of the $t$-test

- Imagine the average tweet length of a McGill ‘student’ is $\mu = 158$ chars.
- We sample $N = 200$ UofT students and find that our average tweet is $\bar{x} = 169$ chars (with $\sigma^2 = 2600$).
- Are UofT tweets significantly longer than much worse McGill tweets?

- So $t = \frac{\bar{x} - \mu}{\sqrt{\frac{\sigma^2}{N}}} = \frac{169 - 158}{\sqrt{\frac{2600}{200}}} \approx 3.05$

- In a $t$-test table, we look up the minimum value of $t$ necessary to reject $H_0$ at $\alpha = 0.005$ (we want to be quite confident) for a 1-tailed test...
Example of the $t$-test

- So $t = \frac{\bar{x} - \mu}{\sqrt{\frac{\sigma^2}{N}}} = \frac{169 - 158}{\sqrt{\frac{2600}{200}}} \approx 3.05$

- In a $t$-test table, we look up the minimum value of $t$ necessary to reject $H_0$ at $\alpha = 0.005$, and find 2.576.
  - Since $3.05 > 2.576$, we can reject $H_0$ at the 99.5% level of confidence ($\gamma = 1 - \alpha = 0.995$); UofT students are significantly more verbose.
Example of the $t$-test

• Some things to observe about the $t$-test table:
  • We need more evidence, $t$, if we want to be more confident (left-right dimension).
  • We need more evidence, $t$, if we have fewer measurements (top-down dimension).
  • A common criticism of the $t$-test is that picking $\alpha$ is ad-hoc. There are ways to correct for the selection of $\alpha$.

<table>
<thead>
<tr>
<th>d.f.</th>
<th>$\alpha$ (one-tail)</th>
<th>0.05</th>
<th>0.025</th>
<th>0.01</th>
<th>0.005</th>
<th>0.001</th>
<th>0.0005</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.314</td>
<td>12.71</td>
<td>31.82</td>
<td>63.66</td>
<td>318.3</td>
<td>636.6</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1.812</td>
<td>2.228</td>
<td>2.764</td>
<td>3.169</td>
<td>4.144</td>
<td>4.587</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>1.725</td>
<td>2.086</td>
<td>2.528</td>
<td>2.845</td>
<td>3.552</td>
<td>3.850</td>
<td></td>
</tr>
<tr>
<td>$\infty$</td>
<td>1.645</td>
<td>1.960</td>
<td>2.326</td>
<td>2.576</td>
<td>3.091</td>
<td>3.291</td>
<td></td>
</tr>
</tbody>
</table>
Another example: collocations

• **Collocation:** *n.* a ‘turn-of-phrase’ or usage where a sequence of words is ‘**perceived**’ to have a meaning ‘**beyond**’ the sum of its parts.

• E.g., ‘*disk drive*’, ‘*video recorder*’, and ‘*soft drink*’ **are** collocations. ‘*cylinder drive*’, ‘*video storer*’, ‘*weak drink*’ **are not** despite some near-synonymy between alternatives.

• Collocations are **not** just highly frequent bigrams, otherwise ‘*of the*’, and ‘*and the*’ would be collocations.

• How can we test if a bigram is a collocation or not?
Hypothesis testing collocations

• For collocations, the null hypothesis \( H_0 \) is that there is no association between two given words beyond pure chance.
  • I.e., the bigram’s actual distribution and pure chance are the same.
  • We compute the probability of those words occurring together if \( H_0 \) were true. If that probability is too low, we reject \( H_0 \).

• E.g., we expect ‘of the’ to occur together, because they’re both likely words to draw randomly
  • We could probably not reject \( H_0 \) in that case.
Example of the \( t \)-test on collocations

- Is ‘new companies’ a collocation?
- In our corpus of 14,307,668 word tokens, \( new \) appears 15,828 times and \( companies \) appears 4,675 times.
- Our null hypothesis, \( H_0 \) is that they are independent, i.e.,

\[
H_0: P(new \ companies) = P(new)P(companies) = \frac{15828}{14307668} \times \frac{4675}{14307668} \approx 3.615 \times 10^{-7}
\]
Example of the $t$-test on collocations

• The Manning & Schütze text claims that if the process of randomly generating bigrams follows a **Bernoulli distribution**.

  • i.e., assigning 1 whenever *new companies* appears and 0 otherwise gives $\bar{x} = p = P(\text{new companies})$

  • For Bernoulli distributions, $\sigma^2 = p(1 - p)$. Manning & Schütze claim that we can assume $\sigma^2 = p(1 - p) \approx p$, since for most bigrams, $p$ is very small.
Example of the $t$-test on collocations

- So, $\mu = 3.615 \times 10^{-7}$ is the expected mean in $H_0$.
- We **actually count** 8 occurrences of *new companies* in our corpus.
  - $\bar{x} = \frac{8}{14307667} \approx 5.591 \times 10^{-7}$
  - $\therefore \sigma^2 \approx p = \bar{x} = 5.591 \times 10^{-7}$
- So $t = \frac{\bar{x} - \mu}{\sqrt{\sigma^2/N}} = \frac{5.591 \times 10^{-7} - 3.615 \times 10^{-7}}{\sqrt{5.591 \times 10^{-7}/14307667}} \approx 0.9999$
- In a **$t$-test table**, we look up the minimum value of $t$ necessary to reject $H_0$ at $\alpha = 0.005$, and find **2.576**.
  - Since **0.9999** < **2.576**, we cannot reject $H_0$ at the 99.5% level of confidence.
  - We **don’t have enough evidence** to think that *new companies* is a collocation (we can’t say that it definitely *isn’t*, though!).
2. Analysis of variance (aside)

- **Analyses of variance (ANOVAs)** (there are several types) can be:
  - A way to generalize *t*-tests to more than two groups.
  - A way to determine which (if any) of several variables are responsible for the variation in an observation (and the interaction between them).

- E.g., we measure the **accuracy** of an ASR system for different settings of **empirical parameters** $M$ and $Q$ (more on these later in the course...).

<table>
<thead>
<tr>
<th>Accuracy (%)</th>
<th>$M = 2$</th>
<th>$M = 4$</th>
<th>$M = 16$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Q = 2$</td>
<td>53.33</td>
<td>66.67</td>
<td>53.33</td>
</tr>
<tr>
<td></td>
<td>26.67</td>
<td>53.33</td>
<td>40.00</td>
</tr>
<tr>
<td></td>
<td>0.00</td>
<td>40.00</td>
<td>26.67</td>
</tr>
<tr>
<td>$Q = 5$</td>
<td>93.33</td>
<td>26.67</td>
<td>100.00</td>
</tr>
<tr>
<td></td>
<td>66.67</td>
<td>13.33</td>
<td>80.00</td>
</tr>
<tr>
<td></td>
<td>40.00</td>
<td>0.00</td>
<td>60.00</td>
</tr>
</tbody>
</table>

$H_0$: no effect of source variables.

<table>
<thead>
<tr>
<th>Source</th>
<th>d.f.</th>
<th>$p$ value</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$Q$</td>
<td>1</td>
<td>0.179</td>
<td>Accept $H_0$</td>
</tr>
<tr>
<td>$M$</td>
<td>2</td>
<td>0.106</td>
<td>Accept $H_0$</td>
</tr>
<tr>
<td>interaction</td>
<td>2</td>
<td>0.006</td>
<td>Reject $H_0$ at $\alpha = 0.01$</td>
</tr>
</tbody>
</table>

A completely fictional example
3. Pearson’s $\chi^2$ test (details aside)

- The $\chi^2$ test applies to **categorical** data, like the output of a classifier.
- Like the $t$-test, we decide on the degrees of freedom (number of categories minus number of parameters), compute the test-statistic, then look it up in a table.
- The test statistic is:

$$\chi^2 = \sum_{c=1}^{C} \frac{(O_c - E_c)^2}{E_c}$$

where $O_c$ and $E_c$ are the observed and expected number of observations of type $c$, respectively.
3. Pearson's $\chi^2$ test

- For example, is our die from Lecture 2 fair or not?
- Imagine we throw it 60 times. The expected number of appearances of each side is 10.

| $c$ | $O_c$ | $E_c$ | $O_c - E_c$ | $(O_c - E_c)^2$ | $(O_c - E_c)^2 / E_c$
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>10</td>
<td>-5</td>
<td>25</td>
<td>2.5</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>10</td>
<td>-2</td>
<td>4</td>
<td>0.4</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
<td>10</td>
<td>-1</td>
<td>1</td>
<td>0.1</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>10</td>
<td>-2</td>
<td>4</td>
<td>0.4</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>20</td>
<td>10</td>
<td>10</td>
<td>100</td>
<td>10</td>
</tr>
</tbody>
</table>

Sum $(\chi^2)$ = 13.4

- With $df = 6 - 1 = 5$, the critical value is 11.07 $< 13.4$, so we throw away $H_0$: the die is biased.
- We'll see $\chi^2$ again soon...
Reading

• Manning & Schütze: 2.2, 5.3-5.5
Entropy and decisions

- **Information theory** is a vast ocean that provides statistical models of communication at the heart of *cybernetics*.
  - We’ve only taken a first step on the beach.
  - See the ground-breaking work of Shannon & Weaver, e.g.

- So far, we’ve mainly dealt with **random variables** that the world provides – e.g., words tokens, mainly.

- What if we could transform those inputs into new random variables, or **features**, that are directly engineered to be useful to decision tasks...