Statistical Methods for NLP

Statistical Inference

Joakim Nivre

Uppsala University
Department of Linguistics and Philology
joakim.nivre@lingfil.uu.se
A stochastic variable (or random variable) $X$ is a function from a sample space $\Omega$ (the domain of $X$) to a value space $\Omega_X$ (the range of $X$).

Examples:

1. The part-of-speech of an arbitrary word from a corpus is a stochastic variable $X$ with $\Omega = \{w|w$ is a corpus token$\}$ and $\Omega_X = \{\text{noun, verb, adjective, \ldots}\}$.

2. The sum of two dice is a stochastic variable $Y$ with $\Omega = \{(x, y)|1 \leq x, y \leq 6\}$ and $\Omega_Y = \{z|2 \leq z \leq 12\}$. 
Types of Variables

- If $\Omega_X$ is a subset of the set of real numbers, then $X$ is said to be numerical; otherwise it is categorical.
- If $\Omega_X$ is finite or countably infinite, then $X$ is said to be discrete.
- Examples:
  1. The part-of-speech $X$ of an arbitrary word from a corpus is a discrete, categorical variable, since $\Omega_X$ is finite and not numerical.
  2. The sum $Y$ of two dice is a discrete numerical variable, since $\Omega_Y$ is finite and numerical.
Frequency Functions

The probability $P(X = x)$ of variable $X$ assuming value $x$ is given by the frequency function $f_X$:

$$f_X(x) = P(X = x)$$

For discrete variables, this is equivalent to summing the probability of all outcomes in $\Omega$ that are mapped to $x$ by $X$:

$$f_X(x) = P(\{u \in \Omega | X(u) = x\}) = \sum_{u : X(u) = x} P(u)$$

Example:

The probability of sampling a noun from a corpus is the sum of the probabilities of sampling each noun.
Expectation

Let $X$ be a discrete numerical variable with value space $\Omega_X$. The expectation of $X$, $E[X]$, is defined as follows:

$$E[X] = \sum_{x \in \Omega_X} x \cdot f_X(x)$$

Example: The expectation of the sum $Y$ of two dice:

$$E[Y] = \sum_{y=2}^{12} y \cdot f_Y(y) = \frac{252}{36} = 7$$
Variance

Let \( X \) be a discrete stochastic variable with expectation \( \mu \). The variance of \( X \), \( \text{Var}[X] \), is defined as follows:

\[
\text{Var}[X] = E[(X - \mu)^2] = \sum_{x \in \Omega_X} (x - \mu)^2 \cdot f_X(x)
\]

Example: The variance of the sum \( Y \) of two dice:

\[
\text{Var}[Y] = \sum_{y=2}^{12} (y - 7)^2 \cdot f_Y(y) = \frac{210}{36} \approx 5.83
\]

If \( X \) is a variable with variance \( \sigma^2 \), then \( \sigma = \sqrt{\sigma^2} \) is the standard deviation of \( X \).
Entropy

Let $X$ be a discrete stochastic variable. The entropy of $X$, $H[X]$, is defined as follows:

$$H[X] = E[- \log_2 f_X] = - \sum_{x \in \Omega_X} f_X(x) \log_2 f_X(x)$$

Example: The entropy of the sum $Y$ of two dice:

$$H[Y] = - \sum_{y=2}^{12} f_Y(y) \log_2 f_Y(y) \approx 3.27$$
More on Entropy

- The entropy of a variable $X$ can be interpreted as the expected amount of information (measured in bits) when learning the value of $X$:

\[ I_X(x) = - \log_2 f_X(x) \]

- Given a finite value space $\Omega_X$ of size $n$, entropy is maximized if $f_X(x) = \frac{1}{n}$ for all $x \in \Omega_X$.

- Example: Entropy of the outcome $Z$ of an 11-sided die (2–12):

\[ H[Z] = - \sum_{z=2}^{12} \frac{1}{11} \log_2 \frac{1}{11} \approx 3.46 \]
Joint and Conditional Probability

Let $X$ and $Y$ be stochastic variables with sample spaces $\Omega_1$ and $\Omega_2$ and value spaces $\Omega_X$ and $\Omega_Y$, respectively.

1. The joint probability of $X$ and $Y$ is given by their joint probability function $f_{(X,Y)}$:

$$f_{(X,Y)}(x,y) = P(X = x, Y = y) = P(\{(u,v) \in \Omega_1 \times \Omega_2 | X(u) = x, Y(v) = y\})$$

2. The conditional probability of $X$ given $Y$ is given by the conditional probability function $f_{X|Y}$:

$$f_{X|Y}(x|y) = P(X = x | Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}$$
Independence

- Stochastic variables $X$ and $Y$ (defined on the same underlying sample space) are independent if and only if the following holds for all $x \in \Omega_X$ and $y \in \Omega_Y$:

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

- Corollary: If $X$ and $Y$ are independent variables then the following conditions hold (for all $x \in \Omega_X$ and $y \in \Omega_Y$):

1. $P(X = x | Y = y) = P(X = x)$
2. $P(Y = y | X = x) = P(Y = y)$
Part-of-Speech Bigrams 1

Let \((X_1, X_2)\) be the parts-of-speech of an arbitrary bigram and let the following probabilities be given:

1. \(P(X_1 = \text{noun}) = P(X_2 = \text{noun}) = 0.2\)
2. \(P(X_1 = \text{adj}) = P(X_2 = \text{adj}) = 0.05\)
3. \(P(X_1 = \text{det}|X_2 = \text{noun}) = 0.3\)
4. \(P(X_1 = \text{det}|X_2 = \text{adj}) = 0.6\)
5. \(P(X_1 = \text{det}|X_2 \not\in \{\text{noun, adj}\}) = 0\)

Question: What is \(P(X_2 = \text{noun}|X_1 = \text{det})\)?
Part-of-Speech Bigrams 2

- Using Bayes’ law:

\[
P(X_2 = \text{noun}) \cdot P(X_1 = \text{det}|X_2 = \text{noun}) \frac{P(X_1 = \text{det})}{P(X_1 = \text{det})}
\]

- Using the law of total probability:

\[
P(X_2 = \text{noun}) \cdot P(X_1 = \text{det}|X_2 = \text{noun}) \frac{P(X_1 = \text{d}|X_2 = \text{n}) \cdot P(X_2 = \text{n}) + P(X_1 = \text{d}|X_2 = \text{a}) \cdot P(X_2 = \text{a})}{P(X_1 = \text{d}|X_2 = \text{n}) \cdot P(X_2 = \text{n}) + P(X_1 = \text{d}|X_2 = \text{a}) \cdot P(X_2 = \text{a})}
\]

- Putting in the numbers:

\[
\frac{0.2 \cdot 0.3}{0.3 \cdot 0.2 + 0.6 \cdot 0.05} = 0.67
\]
Consider:

1. $P(X_1 = \text{det}) = 0.09$
2. $P(X_2 = \text{noun}) = 0.2$
3. $P(X_1 = d, X_2 = n) = P(X_1 = d) \cdot P(X_2 = n|X_1 = d) = 0.06$
4. $P(X_1 = \text{det}) \cdot P(X_2 = \text{noun}) = 0.2 \cdot 0.09 = 0.018$
5. $0.018 \neq 0.06$

Conclusion: $X_1$ and $X_2$ are not independent variables.
Statistical Inference

- Statistical inference is the science of making predictions or inferences from finite sets of observations (samples) to (potentially infinite) sets of new observations (populations).
- Two main kinds of statistical inference:
  1. Estimation: Use samples and sample variables to predict population variables.
  2. Hypothesis testing: Use samples and sample variables to test hypotheses about population variables.
- Note: In statistical modeling, we often talk about models instead of populations.
Sampling

Let $X$ be a stochastic variable.

1. A vector $(X_1, \ldots, X_n)$ of independent variables $X_i$ with the same distribution as $X$ is said to be a random sample of $X$.
2. A value vector $(x_1, \ldots, x_n)$ such that $X_1 = x_1, \ldots, X_n = x_n$ in a particular experiment is called a statistical material.

Example:

- Consider a corpus $C$ consisting of words $(w_1, \ldots, w_n)$.
- Can we regard $C$ as a statistical material resulting from a sample $(W_1, \ldots, W_n)$ of the word variable $W$?
- Why (not)?
Sample Variables

- Given a random sample of a variable $X$, we can define new stochastic variables that are functions of the sample, called sample variables:

  1. The sample mean: $\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$

  2. The sample variance: $s_n^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$

- These variables are called sample variables to distinguish them from the expectation $\mu$ and (true) variance $\sigma^2$ of $X$, which are called population variables or model parameters.
Estimation

Two kinds of estimation:

1. Point estimation: Use sample variable \( f(X_1, \ldots, X_n) \) to estimate parameter \( \phi \).
2. Interval estimation: Use sample variables \( f_1(X_1, \ldots, X_n) \) and \( f_2(X_1, \ldots, X_n) \) to construct an interval such that \( P(f_1(X_1, \ldots, X_n) < \phi < f_2(X_1, \ldots, X_n)) = p \), where \( p \) is the confidence level adopted.
Maximum Likelihood Estimation (MLE)

Given a statistical material $x_1, \ldots, x_n$ and a set of parameters $\theta$, the likelihood function $L$ is:

$$L(x_1, \ldots, x_n, \theta) = \prod_{i=1}^{n} P_\theta(x_i)$$

where $P_\theta(x_i)$ is the probability that the variable $X_i$ assumes the value $x_i$ given a set of values for the parameters in $\theta$.

Maximum likelihood estimation means choosing $\theta$ so that the likelihood function is maximized:

$$\max_\theta L(x_1, \ldots, x_n, \theta)$$
MLE: Example 1

- Given a random sample \((X_1, \ldots, X_n)\) of a numerical variable \(X\), the sample mean \(\overline{X}_n\) is a maximum likelihood estimate of the expectation \(E[X]\).

- The average sentence length \(X\) in a certain type of text can be estimated with the mean sentence length in a representative sample:

\[
\hat{E}[X] = \overline{X}_n
\]
MLE: Example 2

- Given a random sample $(X_1, \ldots, X_n)$ of a categorical variable $X$, the relative frequency of the value $x$, $f_n(x)$, is a maximum likelihood estimate of the probability $P(X = x)$.
- The probability of an arbitrary word from a text being a noun can be estimated with the relative frequency of nouns in a suitable corpus of texts:

$$\hat{P}(\text{noun}) = f_n(\text{noun})$$
The Rationale of MLE

We want to choose the most probable model given the data:

\[
P(\theta|x_1,\ldots,x_n) = \frac{P(x_1,\ldots,x_n|\theta)P(\theta)}{P(x_1,\ldots,x_n)}
\]

\[
\arg\max_{\theta} P(\theta|x_1,\ldots,x_n) = \arg\max_{\theta} P(x_1,\ldots,x_n|\theta)P(\theta)
\]

If we assume a uniform distribution for \(P(\theta)\), then

\[
\arg\max_{\theta} P(\theta|x_1,\ldots,x_n) = \arg\max_{\theta} P(x_1,\ldots,x_n|\theta)
\]

The status of \(P(\theta)\) is controversial in statistical theory (Bayesians vs. Frequentists)
MLE and Smoothing

- MLE is usually a good solution to the estimation problem if the statistical material is large enough.
- For language data, MLE is often suboptimal because of sparse data and requires smoothing (or regularization).
- Example:
  - Additive smoothing:
    \[
    \hat{P}_{\text{add}}(X = x) = \frac{f_n(x) + m}{n + m \cdot |\Omega_x|}
    \]
    where \( m \) is a constant (usually \( m \leq 1 \)).
  - Note: \( \hat{P}_{\text{add}}(X = x) \neq 0 \).
Interval Estimation

- In general, we can derive a 95% confidence interval for our maximum likelihood estimate $\hat{\mu}$ of a mean as follows:

$$\hat{\mu} \pm 1.96\sqrt{\frac{\sigma^2}{n}}$$

- Examples:
  1. Sentence length: $\hat{E}[X] = \bar{X}_n \pm 1.96\sqrt{\frac{\sigma^2}{n}}$
  2. Noun probability: $\hat{P}(noun) = f_n(noun) \pm 1.96\sqrt{\frac{\sigma^2}{n}}$

- Note:
  - The interval grows with increasing variance $\sigma^2$.
  - The interval shrinks with the sample size $n$. 
More on Interval Estimation

- Where does the number 1.96 come from?

- Assumptions:
  1. Parameter has a normal distribution – okay for large $n$.
  2. True variance $\sigma^2$ is known – usually not the case.
Hypothesis Testing

- When are differences in sample variables $f(x_1, \ldots, x_m)$ and $f(y_1, \ldots, y_n)$ significant?
  - Do they reflect differences in population variables $\xi$ and $\upsilon$?
  - Null hypothesis $H_0$: $\xi = \upsilon$

- Test procedure:
  - Choose a test statistic $f(X, Y)$ whose distribution is known when $H_0$ is true.
  - Calculate the probability $p$ of $f(x, y)$ given $H_0$.
  - If $p < \alpha$, reject $H_0$ at significance level $\alpha$. 
Example: Z-test

- **Data:**
  - Mean sentence length in 50 novels from 1950s: $\bar{X}_1 = 19.3$.
  - Mean sentence length in 50 novels from 2000s: $\bar{X}_2 = 16.4$.
  - $X$ is normally distributed with variance $\sigma^2 = 134.2$.

- **Test statistic:**
  
  $$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma^2}{n_1 + n_2}}} = \frac{19.3 - 16.9}{\sqrt{\frac{134.2}{50 + 50}}} = 2.28$$

- **Probability calculation:**
  - $Z = 2.28$ corresponds to $p = P(Z = 2.28 | H_0) = 0.0226$.
  - Reject $H_0$ at $\alpha = 0.05$ (but not $\alpha = 0.01$).
  - Note: This does not mean that $P(H_1) > 0.95$. 
Tips and Tricks

▶ What to do in real life?
  ▶ When variance is not known – use sample variance and a $t$-test (instead of $Z$-test):

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

▶ When distribution is not normal – use a non-parametric test.

▶ The special case of proportions:
  ▶ If $X$ is binary with $P(X = 1) = p$, then $\text{Var}[X] = p(1 - p)$:

$$\hat{p} = \frac{f_n(1)}{n} \pm \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$