Introduction to statistics
Session 1

Grzegorz Chrupała

Saarland University

October 9, 2012

Key concepts

- Axioms of probability

Key concepts

- Axioms of probability
- Chain rule and Bayes theorem

Key concepts

- Axioms of probability
- Chain rule and Bayes theorem
- Random variables

Key concepts

- Axioms of probability
- Chain rule and Bayes theorem
- Random variables
- Entropy

Key concepts

- Axioms of probability
- Chain rule and Bayes theorem
- Random variables
- Entropy
- Hypothesis testing

Key concepts

- Axioms of probability
- Chain rule and Bayes theorem
- Random variables
- Entropy
- Hypothesis testing
- Binomial and normal distributions

Key concepts

- Axioms of probability
- Chain rule and Bayes theorem
- Random variables
- Entropy
- Hypothesis testing
- Binomial and normal distributions
- Linear and logistic regression

Key concepts

- Axioms of probability
- Chain rule and Bayes theorem
- Random variables
- Entropy
- Hypothesis testing
- Binomial and normal distributions
- Linear and logistic regression

If you are familiar with these, you don't need this course!

Course structure

- **Oct 8 - Oct 9**
  Basic concepts of Probability and Information theory with Grzegorz Chrupała
  `gchrupala@lsv.uni-saarland.de`

Course structure

- **Oct 8 - Oct 9**
  Basic concepts of Probability and Information theory with Grzegorz Chrupała
  `gchrupala@lsv.uni-saarland.de`
- **Oct 10 - Oct 12**
  Statistics for experimental science with Francesca Delogu

Course structure

- **Oct 8 - Oct 9**
  Basic concepts of Probability and Information theory with Grzegorz Chrupała
  `gchrupala@lsv.uni-saarland.de`
- **Oct 10 - Oct 12**
  Statistics for experimental science with Francesca Delogu
- **Oct 15 - Oct 17**
  Statistics for NLP – reading group

Course structure

- **Oct 8 - Oct 9**
  Basic concepts of Probability and Information theory with Grzegorz Chrupała
  `gchrupala@lsv.uni-saarland.de`
- **Oct 10 - Oct 12**
  Statistics for experimental science with Francesca Delogu
- **Oct 15 - Oct 17**
  Statistics for NLP – reading group
- **Oct 18 - 19**
  Linear models: Grzegorz Chrupała

Textbook and topics

- Foundations of Statistical NLP, Manning and Schütze
  - For each of the first three sessions next week everybody reads a section of the book.
  - A group of students will present the material (45-60 min).
  - Follow up with exercises and discussion.

Suggested topics

- Collocations (5)
- Lexical acquisition (8)
- Clustering (14)

Other topic possible (talk to me!)

- Organize yourselves into three groups and agree on topics by **tomorrow**
- Within each group, coordinate and decide who presents which subsection

Today: Basic concepts in probability theory

- Probability notation $P(X|Y)$
  - What does this expression mean?
  - How can we manipulate it?
  - How can we estimate its value in practice?

Three aspects of statistics

- **Descriptive.** Mean or median grade at a university. Distribution of heights among a population of country

Three aspects of statistics

- **Descriptive.** Mean or median grade at a university. Distribution of heights among a population of country
- **Confirmatory.** Are the results statistically significant?

Three aspects of statistics

- **Descriptive.** Mean or median grade at a university. Distribution of heights among a population of country
- **Confirmatory.** Are the results statistically significant?
- **Predictive.** Learn from past data to predict future events

Three aspects of statistics

- **Descriptive.** Mean or median grade at a university. Distribution of heights among a population of country
- **Confirmatory.** Are the results statistically significant?
- **Predictive.** Learn from past data to predict future events

Another dimension: Frequentist vs Bayesian (philosophical underpinnings)

Experiments and Sample Spaces

- Consider an experiment or process

Experiments and Sample Spaces

- Consider an experiment or process
- Set of possible basic outcomes: sample space Ω
  - Coin toss:

Experiments and Sample Spaces

- Consider an experiment or process
- Set of possible basic outcomes: sample space Ω
  - Coin toss: $\Omega = \{H, T\}$.
  - Dice:

Experiments and Sample Spaces

- Consider an experiment or process
- Set of possible basic outcomes: sample space Ω
  - Coin toss: $\Omega = \{H, T\}$.
  - Dice: $\Omega = \{1, 2, 3, 4, 5, 6\}$
  - Yes/no poll, correct/incorrect:

Experiments and Sample Spaces

- Consider an experiment or process
- Set of possible basic outcomes: sample space Ω
  - Coin toss: $\Omega = \{H, T\}$.
  - Dice: $\Omega = \{1, 2, 3, 4, 5, 6\}$
  - Yes/no poll, correct/incorrect: $\Omega = \{0, 1\}$
  - Number of traffic accidents in an area per year:

Experiments and Sample Spaces

- Consider an experiment or process
- Set of possible basic outcomes: sample space Ω
  - Coin toss: $\Omega = \{H, T\}$.
  - Dice: $\Omega = \{1, 2, 3, 4, 5, 6\}$
  - Yes/no poll, correct/incorrect: $\Omega = \{0, 1\}$
  - Number of traffic accidents in an area per year: $\Omega = \mathbb{N}$
  - Misspelling of a word.

Experiments and Sample Spaces

- Consider an experiment or process
- Set of possible basic outcomes: sample space Ω
  - Coin toss: $\Omega = \{H, T\}$.
  - Dice: $\Omega = \{1, 2, 3, 4, 5, 6\}$
  - Yes/no poll, correct/incorrect: $\Omega = \{0, 1\}$
  - Number of traffic accidents in an area per year: $\Omega = \mathbb{N}$
  - Misspelling of a word. $\Omega = Z*$ where Z is an alphabet, and $Z*$ the set of strings over this alphabet
  - Guess a missing word:

Experiments and Sample Spaces

- Consider an experiment or process
- Set of possible basic outcomes: sample space Ω
  - Coin toss: $\Omega = \{H, T\}$.
  - Dice: $\Omega = \{1, 2, 3, 4, 5, 6\}$
  - Yes/no poll, correct/incorrect: $\Omega = \{0, 1\}$
  - Number of traffic accidents in an area per year: $\Omega = \mathbb{N}$
  - Misspelling of a word. $\Omega = Z*$ where Z is an alphabet, and $Z*$ the set of strings over this alphabet
  - Guess a missing word: $|\Omega| =$ vocabulary size

Events

- An event A is a set of basic outcomes. Event A takes place if the outcome of the experiment $\in A$
- $A \subseteq \Omega$, and any $A \in 2^{\Omega}$ (all subsets of Ω).
  - Ω

Events

- An event A is a set of basic outcomes. Event A takes place if the outcome of the experiment $\in A$
- $A \subseteq \Omega$, and any $A \in 2^{\Omega}$ (all subsets of Ω).
  - Ω is the certain event
  - $\emptyset$

Events

- An event A is a set of basic outcomes. Event A takes place if the outcome of the experiment $\in A$
- $A \subseteq \Omega$, and any $A \in 2^{\Omega}$ (all subsets of Ω).
  - Ω is the certain event
  - $\emptyset$ is the impossible event
- Experiment, toss 3 coins

Events

- An event A is a set of basic outcomes. Event A takes place if the outcome of the experiment $\in A$
- $A \subseteq \Omega$, and any $A \in 2^{\Omega}$ (all subsets of Ω).
  - Ω is the certain event
  - $\emptyset$ is the impossible event
- Experiment, toss 3 coins
  - $\Omega =$

Events

- An event A is a set of basic outcomes. Event A takes place if the outcome of the experiment $\in A$
- $A \subseteq \Omega$, and any $A \in 2^{\Omega}$ (all subsets of Ω).
  - Ω is the certain event
  - $\emptyset$ is the impossible event
- Experiment, toss 3 coins
  - $\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$
  - Event A: there were exactly two tails.

Events

- An event A is a set of basic outcomes. Event A takes place if the outcome of the experiment $\in A$
- $A \subseteq \Omega$, and any $A \in 2^{\Omega}$ (all subsets of Ω).
  - Ω is the certain event
  - $\emptyset$ is the impossible event
- Experiment, toss 3 coins
  - $\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$
  - Event A: there were exactly two tails.
    - $A = \{HTT, THT, TTH\}$
  - Event B: there were three heads.

Events

- An event A is a set of basic outcomes. Event A takes place if the outcome of the experiment $\in A$
- $A \subseteq \Omega$, and any $A \in 2^\Omega$ (all subsets of Ω).
  - Ω is the certain event
  - $\emptyset$ is the impossible event
- Experiment, toss 3 coins
  - $\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$
  - Event A: there were exactly two tails.
    - $A = \{HTT, THT, TTH\}$
  - Event B: there were three heads.
    - $B = \{HHH\}$

Events and probability

- $P(\text{Germany wins the game}|\text{no rain}) = 0.9$

Events and probability

- $P(\text{Germany wins the game}|\text{no rain}) = 0.9$
- Past performance. Germany won 90% of games with no rain

Events and probability

- $P(\text{Germany wins the game}|\text{no rain}) = 0.9$
- Past performance. Germany won 90% of games with no rain
- Hypothetical performance. If they played the game in many parallel universes

Events and probability

- $P(\text{Germany wins the game}|\text{no rain}) = 0.9$
- Past performance. Germany won 90% of games with no rain
- Hypothetical performance. If they played the game in many parallel universes
- Subjective strength of belief. Would bet up to 90 cents for a chance to win 1 euro.

Events and probability

- $P(\text{Germany wins the game}|\text{no rain}) = 0.9$
- Past performance. Germany won 90% of games with no rain
- Hypothetical performance. If they played the game in many parallel universes
- Subjective strength of belief. Would bet up to 90 cents for a chance to win 1 euro.
- Output of some computable formula

Probability notation

$$\boxed{P(\underbrace{\text{Germany wins the game}}_{\text{Event A}} \mid \underbrace{\text{no rain}}_{\text{Event B}})}$$

- Given that event B happens, how likely is event A?
- *Germany wins the game* is a **predicate** which selects the outcomes that are members of event A

Frequentist probability

- For series i
  - Repeat experiment many times
  - Record how many times event A occured: $\text{count}_i(A)$
- The ratios $\frac{\text{count}_i(A)}{T_i}$, where T_i is the number of experiments in series i, are close to some unknown but constant value
- We can call this constant $P(A)$

Estimating probabilities

- The constant $P(A)$ is unknown, but we can estimate it:
  - From a single series i: $P(A) = \frac{\text{count}_i A}{T_i}$ (the common case)
  - Or take the weighted average of all series i

Example

- Toss three coins.
  - $\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$
- A: there were exactly three tails
  - $A = \{HTT, THT, TTH\}$
- Run 1000 times
- Got one of HTT, THT, TTH 386 times out of 1000
- $\hat{P}(A) = 0.386$
- Run several times: 373, 399, 355, 372, 406, 359
- $\hat{P}(A) = 0.379$
- **If each outcome in Ω is equally likely** $P(A) = 3/8 = 0.375$

P as a function of sets of outcomes

$$P(\text{Germany wins}|\text{no rain}) = \frac{P(\text{Germany wins, no rain})}{P(\text{no rain})}$$

P as a function of sets of outcomes

$P(A\|B)$	$=$	P(	A , B	) / P(	B	)
notation			conjunction		predicate	

Axioms of probability

- $P(\emptyset) =$

Axioms of probability

- $P(\emptyset) = 0$
- $P(\Omega) =$

Axioms of probability

- $P(\emptyset) = 0$
- $P(\Omega) = 1$
- $P(A) \leq P(B)$ for any $A \subseteq B$

Axioms of probability

- $P(\emptyset) = 0$
- $P(\Omega) = 1$
- $P(A) \leq P(B)$ for any $A \subseteq B$
- $P(A) + P(B) = P(A \cup B)$ provided

Axioms of probability

- $P(\emptyset) = 0$
- $P(\Omega) = 1$
- $P(A) \leq P(B)$ for any $A \subseteq B$
- $P(A) + P(B) = P(A \cup B)$ provided $A \cap B = \emptyset$

Joint and conditional probability

- Joint probability and the meaning of commas
  - $P(A, B) = P(A \cap B)$
  - $P(\text{Germany wins, no rain}) = P(\text{Germany wins} \wedge \text{no rain})$

Joint and conditional probability

- Joint probability and the meaning of commas
  - $P(A, B) = P(A \cap B)$
  - $P(\text{Germany wins, no rain}) = P(\text{Germany wins} \wedge \text{no rain})$
- $P(A|B) = P(A, B)/P(B)$
  - Estimate from counts

$$\begin{aligned} P(A|B) &= \frac{P(A,B)}{P(B)} && (1) \\ &= \frac{\text{count}(A \cap B)/T}{\text{count}(B)/T} && (2) \\ &= \frac{\text{count}(A \cap B)}{\text{count}(B)} && (3) \end{aligned}$$

Chain rule

- $P(A|B) = \frac{P(A,B)}{P(B)}$
- Therefore $P(A,B) = P(A|B)P(B)$

Chain rule

- $P(A|B) = \frac{P(A,B)}{P(B)}$
- Therefore $P(A,B) = P(A|B)P(B)$
- Generalization:
  $P(A_1, A_2, \ldots, A_n)$

$$
\begin{aligned}
&= P(A_1|A_2, \ldots, A_n)P(A_2, \ldots, A_n) \\
&= P(A_1|A_2, \ldots, A_n)P(A_2|A_3, \ldots, A_n)P(A_3, \ldots, A_n) \\
&= \prod_{i=1}^{n} P(A_i|A_{i+1}, \ldots, A_n)
\end{aligned}
$$

Independence

- Two events A and B are **independent** if $P(A, B) = P(A)P(B)$

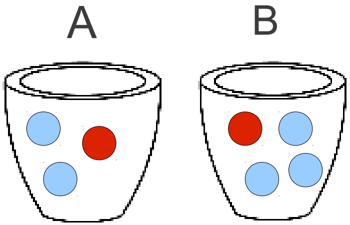
Independence

- Two events A and B are **independent** if $P(A, B) = P(A)P(B)$
- For independent A, B, does $P(A|B) = P(A)$ hold?

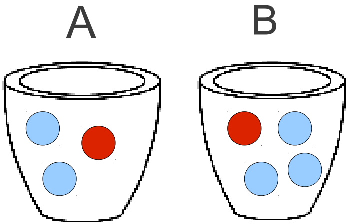
Independence

- Two events A and B are **independent** if $P(A, B) = P(A)P(B)$
- For independent A, B, does $P(A|B) = P(A)$ hold?
- A and B are **conditionally independent** if $P(A, B|C) = P(A|C)P(B|C)$

There are two urns:



There are two urns:

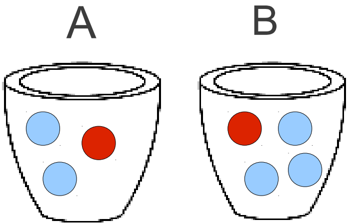


- Suppose we pick an urn uniformly at random and then select a random ball from that urn. What is probability that you pick urn A, and take a blue ball from it?

Marginal probability

- Given $P(A, B_i)$ for disjoint events B_i, find out $P(A)$.
- Use last axiom

$$
\begin{aligned}
P(A) &= P((A \cap B_1) \cup (A \cap B_2) \cup \cdots \cup (A \cap B_n)) \\
&= P(A \cap B_1) + P(A \cap B_2) + \cdots + P(A \cap B_n) \\
&= \sum_{i=1}^{n} P(A \cap B_i)
\end{aligned}
$$



- Given the same ball-picking procedure as before, what is the probability of picking the blue ball?

Bayes rule

- $P(A, B) = P(B, A)$ since $A \cap B = B \cap A$
- $P(B, A) = P(B|A)P(A)$

Bayes rule

- $P(A, B) = P(B, A)$ since $A \cap B = B \cap A$
- $P(B, A) = P(B|A)P(A)$

Therefore

$$\begin{aligned} P(A|B) &= \frac{P(A, B)}{P(B)} \\ &= \frac{P(B, A)}{P(B)} \\ &= \frac{P(B|A)P(A)}{P(B)} \end{aligned}$$

Bayes rule

If we are interested in comparing the probability of events $A_1, A_2, \ldots$ given B, we can ignore $P(B)$ since it's the same for all A_i

$$\begin{aligned}\operatorname*{argmax}_i P(A_i|B) &= \operatorname*{argmax}_i \frac{P(B|A_i)P(A_i)}{P(B)} \\ &= \operatorname*{argmax}_i P(B|A_i)P(A_i)\end{aligned}$$

Bayes rule

If we are interested in comparing the probability of events $A_1, A_2, \ldots$ given B, we can ignore $P(B)$ since it's the same for all A_i

$$\begin{aligned}\operatorname*{argmax}_i P(A_i|B) &= \operatorname*{argmax}_i \frac{P(B|A_i)P(A_i)}{P(B)} \\ &= \operatorname*{argmax}_i P(B|A_i)P(A_i)\end{aligned}$$

- This idea is sometimes expressed as

$$P(A|B) \propto P(B|A)P(A)$$

Example

Suppose we are interested in a test to detect a disease which affects one in 100,000 people on average. A lab has developed a test which works but is not perfect.

- If a person has the disease it will give a positive result with probability 0.97
- if they do not, the test will be positive with probability 0.007.

You took the test, and it gave a positive result. What is the probability that you actually have the disease?

Credits

Some material adapted from:

- Foundations of Statistical NLP
- Intro to NLP slides by Jan Hajic
- How to use probabilities slides by Jason Eisner