IN4080 – 2021 FALL

NATURAL LANGUAGE PROCESSING

Jan Tore Lønning

Probabilities

Lecture 3, 2 Sept.

Today – Probability theory

- Probability
- Random variable

The benefits of statistics in NLP:

1. Part of the (learned) model:
   - What is the most probable meaning of this occurrence of *bass*?
   - What is the most probable parse of this sentence?
   - What is the best (most probable) translation of a certain Norwegian sentence into English?

Tagged text and tagging


```
[('They', 'PRP'), ('saw', 'VBD'), ('a', 'DT'), ('saw', 'NN'), ('.', '.')]
[('They', 'PRP'), ('like', 'VBP'), ('to', 'TO'), ('saw', 'VB'), ('.', '.')]
[('They', 'PRP'), ('saw', 'VBD'), ('a', 'DT'), ('log', 'NN')]
```

- In tagged text each token is assigned a "part of speech" (POS) tag
- A tagger is a program which automatically ascribes tags to words in text
  - We will return to how they work
- From the context we are (most often) able to determine the tag.
  - But some sentences are genuinely ambiguous and hence so are the tags.

The benefits of statistics in NLP:

2. In constructing models from examples (”learning”):

- What is the best (=most likely) model given these examples?
  - Given a set of tagged English sentences.
    - Try to construct a tagger from these.
    - Between several different candidate taggers, which one is best?
  - Given a set of texts translated between French and English
    - Try to construct a translations system from these
    - Which system is best

The benefits of statistics in NLP:

3. In evaluation:

- We have two parsers and test them on 1000 sentences. One gets 86% correct and the other gets 88% correct. Can we conclude that one is better than the other
- If parser one gets 86% correct on the 1000 sentences drawn from a much larger corpus. How well will it perform on the corpus as a whole?

Components of statistics

1. Probability theory
   - Mathematical theory of chance/random phenomena
2. Descriptive statistics
   - Describing and systematizing data
3. Inferential statistics
   - Making inferences on the basis of (1) and (2), e.g.
     - (Estimation:) ”The average height is between 179cm and 181cm with 95% confidence”
     - (Hypothesis testing:) ”This pill cures that illness, with 99% confidence”

Probability theory

Basic concepts

- Random experiment (or trial) (no: forsøk)
  - Observing an event with unknown outcome
- Outcomes (utfallene)
  - The possible results of the experiment
- Sample space (utfallsrommet)
  - The set of all possible outcomes

Examples

	Experiment	Sample space, Ω
1	Flipping a coin	{H, T}
2	Rolling a dice	{1,2,3,4,5,6}
3	Flipping a coin three times	{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}
4	Will it rain tomorrow?	{Yes, No}

Examples

	Experiment	Sample space, Ω
1	Flipping a coin	{H, T}
2	Rolling a dice	{1,2,3,4,5,6}
3	Flipping a coin three times	{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}
4	Will it rain tomorrow?	{Yes, No}
5	A word occurrence in ''Tom Sawyer''	{u | u is an English word}
6	Throwing a dice until you get 6	{1,2,3,4, …}
7	The maximum temperature at Blindern for a day	{t | t is a real}

Event

- An event (begivenhet/hendelse) is a set of elementary outcomes

	Experiment	Event	Formally
2	Rolling a dice	Getting 5 or 6	{5,6}
3	Flipping a coin three times	Getting at least two heads	{HHH, HHT, HTH, THH}

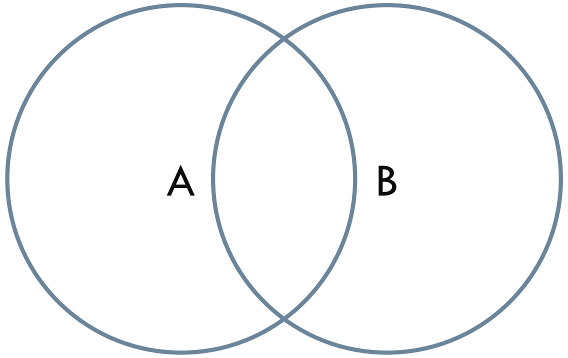
Event

- An event (begivenhet) is a set of elementary outcomes

	Experiment	Event	Formally
2	Rolling a dice	Getting 5 or 6	{5,6}
3	Flipping a coin three times	Getting at least two heads	{HHH, HHT, HTH, THH}
5	A word occurrence in "Tom Sawyer"	The word is a noun	{u | u is an English noun}
6	Throwing a dice until you get 6	An odd number of throws	{1,3,5, …}
7	The maximum temperature at Blindern	Between 20 and 22	$\{t \mid 20 \leq t \leq 22\}$

Operations on events

- Union: A∪B
- Intersection (snitt): A∩B
- Complement



- Venn diagram
- http://www.google.com/doodles/john-venns-180th-birthday

Probability measure, sannsynlighetsmål

- A probability measure P is a function from events to the interval [0,1] such that:

1. $P(\Omega) = 1$
2. $P(A) \geq 0$
3. If $A \cap B = \emptyset$ then $P(A \cup B) = P(A) + P(B)$
    - And if A1, A2, A3, … are disjoint, then $P\left(\bigcup_{j=1}^{\infty} A_j\right) = \sum_{j=1}^{\infty} P(A_j)$

Examples

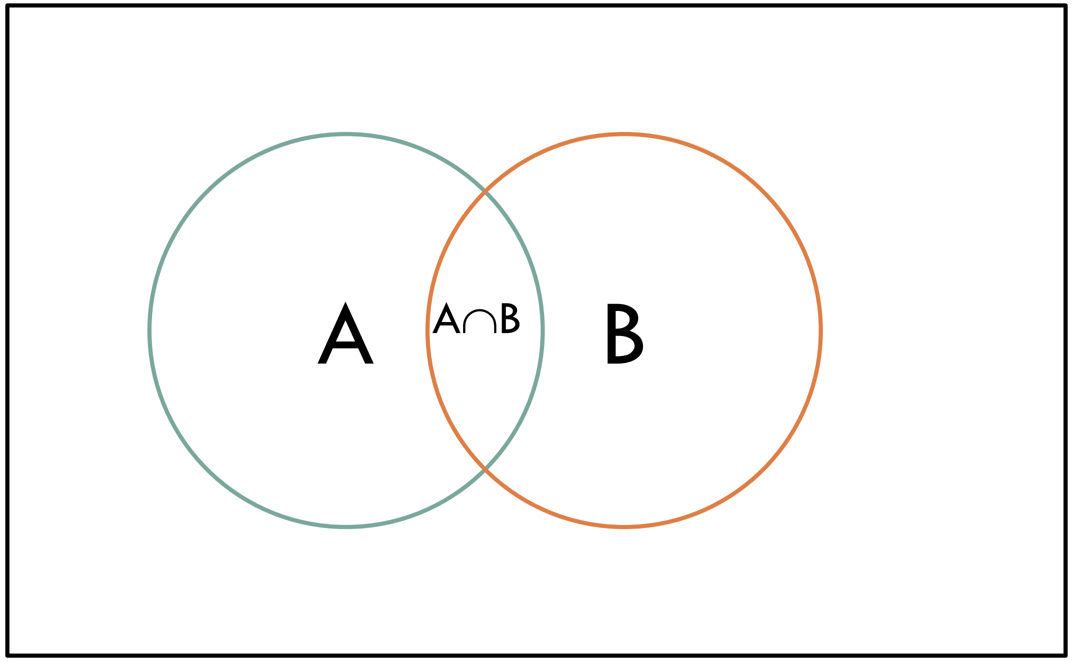
	Experiment	Event	Probability
2	Rolling a fair dice	Getting 5 or 6	P({5,6})=2/6=1/3
3	Flipping a fair coin three times	Getting at least two heads	P({HHH, HHT, HTH, THH}) = 4/8

Examples

	Experiment	Event	Probability
2	Rolling a dice	Getting 5 or 6	P({5,6})=2/6=1/3
3	Flipping a coin three times	Getting at least two heads	P({HHH, HHT, HTH, THH}) = 4/8
5	A word in TS	It is a noun	P({u | u is a noun})= 0.43?
6	Throwing a dice until you get 6	An odd number of throws	P({1,3,5, …})=?
7	The maximum temperature at Blindern at a given day	Between 20 and 22	$P(\{t \mid 20 \leq t \leq 22\})=0.05$

Some observations

- $P(\emptyset) = 0$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$



Some observations

- $P(\emptyset) = 0$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- If Ω is finite or more generally countable, then

$$P(A) = \sum_{a \in A} P(\{a\})$$

- In general, $P(\{a\})$ does not have to be the same for all $a \in A$
  - For some of our examples, like fair coin or fair dice, they are: $P(\{a\}) = 1/n$, where $\#(\Omega) = n$
  - But not if the coin/dice is unfair
  - E.g. $P(\{n\})$, the probability of using *n* throws to get the first 6 is not uniform
  - If A is infinite, $P(\{a\})$ can't be uniform

Examples

Backgammon

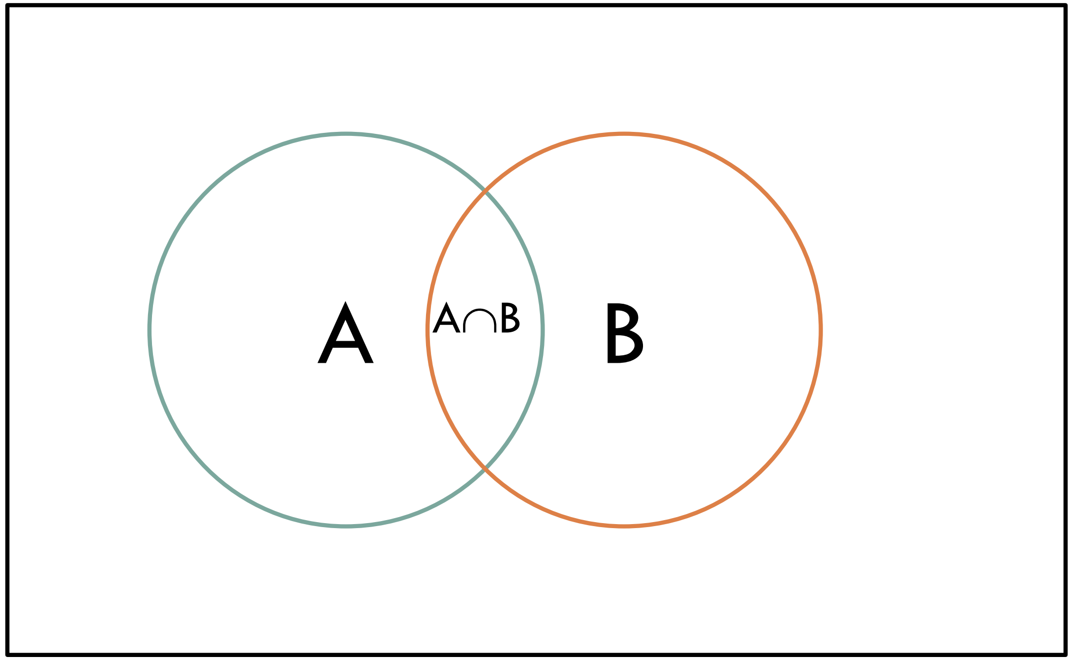
- Throwing two dice
- Both numbers matter, but not which dice is which:
  - e.g., {3,5} = {5,3}
- 21 = 15 + 6 possible outcomes
- P({{3,5}}=2/36
- P({{5,5}}=1/36

The sum of two dice

- 11 possible outcomes:
  - 2, 3, …,12
- P({2}) = P({12})=1/36
- P({8}) = 5/36

Joint probability

- $P(A \cap B)$
  - Both A and B happens



Examples

6-sided fair dice, find the following probabilities

- Two throws: the probability of 2 sixes?
- The probability of getting a six in two throws?
- 5 dices: the probability of getting 5 equal dices?
- 5 dices: the probability of getting 1-2-3-4-5?
- 5 dices: the probability of getting no 6-s?

Counting methods

Given all outcomes equally likely

- P(A) = number of ways A can occur/
  total number of outcomes
- Multiplication principle:
  if one experiment has *m* possible outcomes
  and another has *n* possible outcomes,
  then the two have $m*n$ possible outcomes

Sampling

How many different samples?

- Ordered sequences:
  - Choose k items from a population of n items with replacement: n^k
  - Without replacement:

$$n(n-1)(n-2)\cdots(n-k+1) = \prod_{i=0}^{k-1}(n-i) = \frac{n!}{(n-k)!}$$

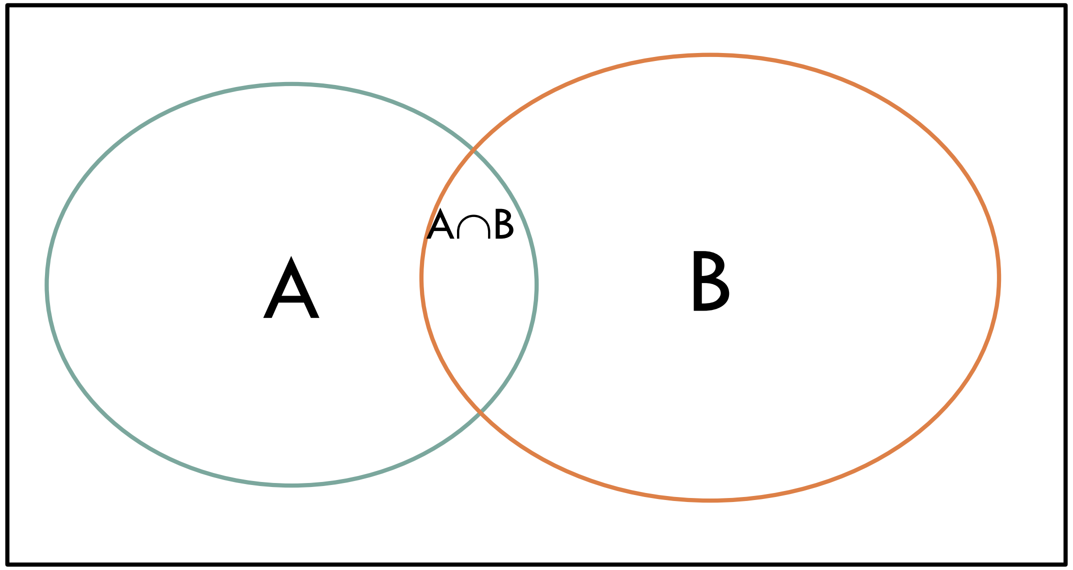
- Unordered sequences
  - Without replacement: $\frac{1}{k!}\left(\frac{n!}{(n-k)!}\right) = \left(\frac{n!}{k!(n-k)!}\right) = \binom{n}{k}$
    - = the number of ordered sequences/
      the number of ordered sequences containing the same k elements

Conditional probability

- **Conditional probability** (betinget sannsynlighet)

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

  - The probability of A happens if B happens



Conditional probability

- Conditional probability (betinget sannsynlighet)

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

  - The probability of A happens if B happens
- Multiplication rule $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$
- A and B are independent iff $P(A \cap B) = P(A)P(B)$

Example

- Throwing two dice
  - A: the first dice is 1
  - B: the second dice is 4
  - $P(A \cap B) = P(\{(1,6)\}) = 1/36 = P(A)P(B)$
  - Hence: A and B are independent
- Also throwing two dice
  - B: the first dice is 1
  - C: the sum of the two is 5
    - $P(C) = 4/36 = 1/9$
    - $P(C \cap B) = P(\{(1,4)\}) = 1/36$
    - $P(C)P(B) = 1/9 * 1/6 = 1/54$
  - Hence: B and C are not independent

Bayes theorem

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

- Jargon:
  - P(A) – prior probability
  - P(A|B) – posterior probability
- Extended form

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

Example: Corona test (May 2020, Norway)

- The test has a good sensitivity (= recall) (cf. Wikipedia):
  - It recognizes 80% of the infected
  - $P(pos|c19) = 0.8$
- It has an even better specificity: 99.9%
  - Specificity: $P(neg|-c19) = 0.999$
  - If you are not ill, there is only 0.1% chance for a positive test
  - $P(pos|-c19) = 0.001$
- What is the chances you are ill if you get a positive test?
- (Numbers from FHI web page May 2020, removed by Aug 2021)

Source

- The numbers were published at:
  - https://www.fhi.no/nettpub/coronavirus/testing-og-oppfolging-av-smittede/testkriterier/
- This is a page which is updated and changed weekly
- But the actual version can be recovered with the Way Back Machine:
  - https://web.archive.org/web/20200619010655/https://www.fhi.no/nettpub/coronavirus/testing-og-oppfolging-av-smittede/testkriterier/

Example: Corona test, contd.

- $P(pos|c19) = 0.8\,,\, P(pos|-c19) = 0.001$
- We also need the prior probability.
  - Before the summer 2020, it was assumed to be something like $P(c19) = \frac{1}{10{,}000}$
  - i.e. 10 in 100,000 or 500 in Norway
- Then $P(c19|pos) =$

$$\frac{P(pos|c19)P(c19)}{P(pos|c19)P(c19) + P(pos|-c19)P(-c19)}$$

$$= \frac{0.8 \times 0.0001}{0.8 \times 0.0001 + 0.001 \times 0.9999} = 0.074$$

What to be learned for the example?

- Most probably you are not ill, even if you get a positive test.
- But it is much more probable that your are ill after a positive test (posterior probability) than before the test (prior probability).
- It doesn't make sense to test large samples (without symptoms) to find out how many are infected.
  - Why we don't test everybody.
- Repeating the test might help.

Exercise

- The number of infected is much larger 1 Sept. 2021, maybe 25,000 in Norway.
- What is the probability you are infected given a positive test today?

  Assume the properties of test are the same.

The nature of probabilities

What are probabilities?

- Example throwing a dice:

1. Classical view:
   - The six outcomes are equally likely
2. Frequenist:
   - If you throw the dice many, many, many times, the number of 6s approach 16.6666…%
3. Bayesian: subjective beliefs

Random variables

Random variable

- A variable X in statistics is a property (feature) of an outcome of an experiment.
  - Formally it is a function from a sample space (utfallsrom) Ω to a value space Ω_X.
- When the value space Ω_X is numerical (roughly a subset of R^n), it is called a random variable
- There are two kinds:
  - Discrete random variables
  - Continuous random variables
- A third type of variable: categorical variable, when Ω_X is non-numerical

Examples

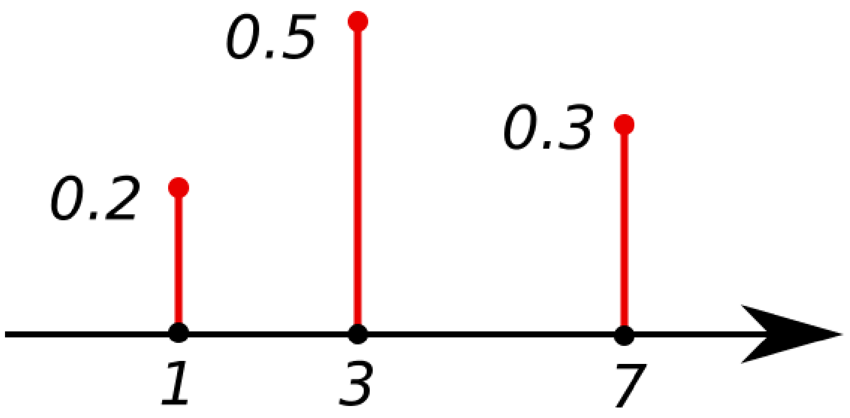
1. Throwing two dice,
   - $\Omega=\{(1,1), (1,2),\ldots(1,6),(2,1),\ldots(6,6)\}$
   1. The number of 6s is a random variable X, $\Omega_X=\{0, 1, 2\}$
   2. The number of 5 or 6s is a random variable Y, $\Omega_Y= \Omega_X$
   3. The sum of the two dice, Z, $\Omega_Z=\{2, 3, \ldots, 12\}$
2. A random person:
   1. X, the height of the person $\Omega_X=[0, 3]$ (meters)
   2. Y, whether age ≥ 18 or not, $\Omega_Y=\{0, 1\}$ (1 for ≥ 18)

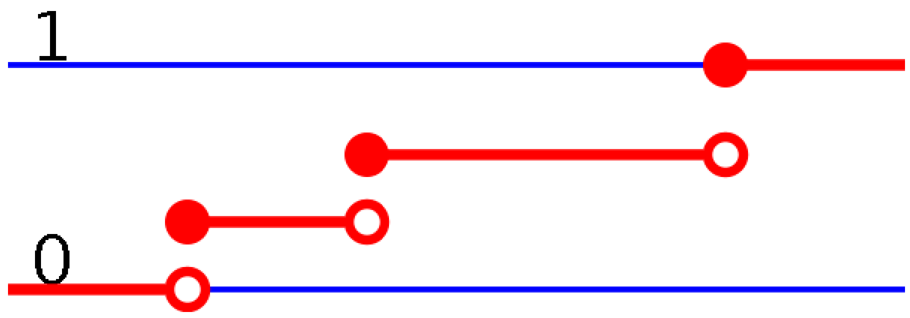
- Ex 2.1 is continuous, the other are discrete

Discrete random variables

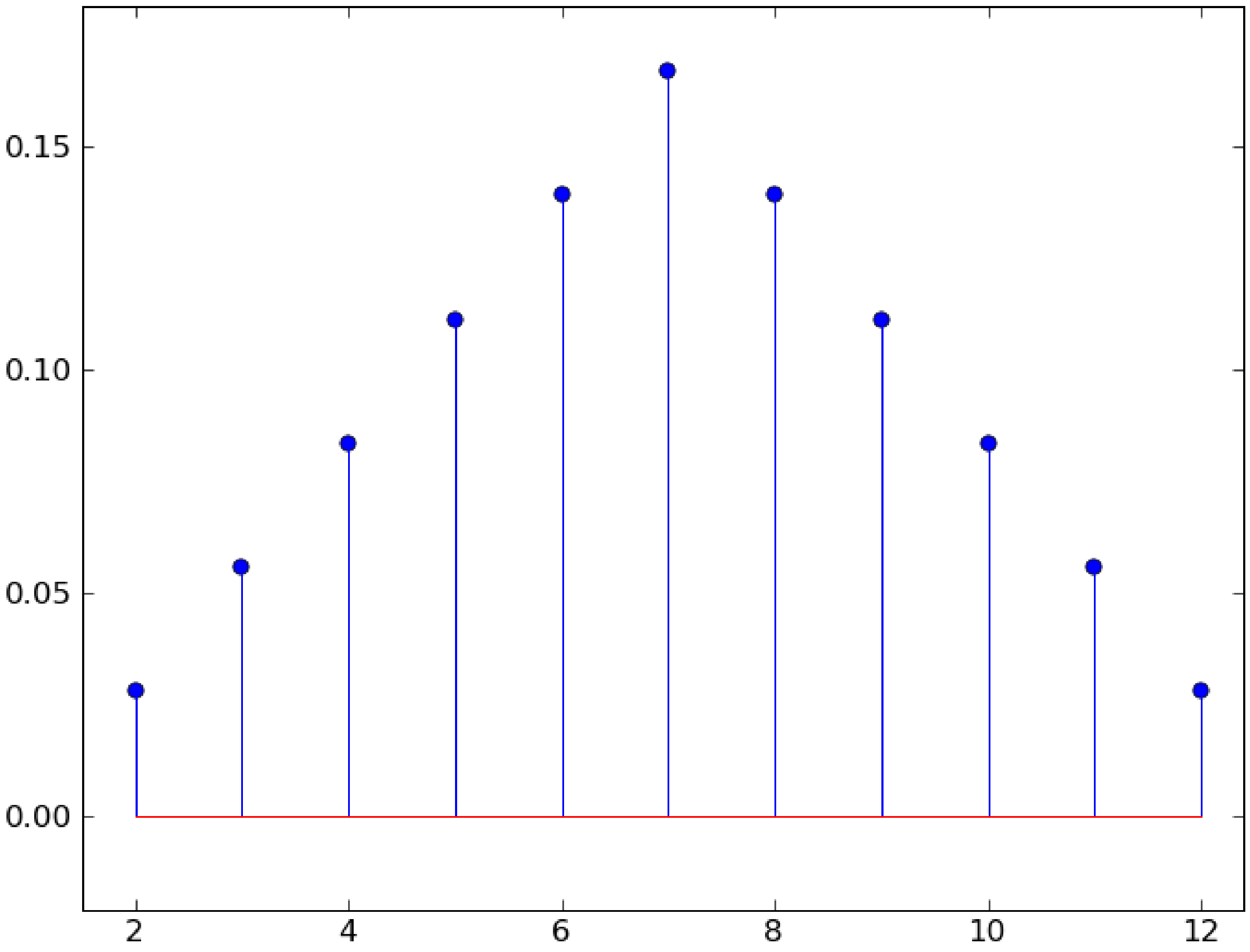
Discrete random variable

- The value space is a finite or a countable infinite set of numbers {x1, x2,…, xn, …}
- The probability mass function, pmf, p, also called frequency function, which to each value yields
  - $p(x_i) = P(X=x_i) = P(\{\omega\in\Omega \mid X(\omega)=x_i\})$
- The cumulative distribution function, cdf,
  - $F(x_i) = P(X \leq x_i) = P(\{\omega\in\Omega \mid X(\omega) \leq x_i\})$



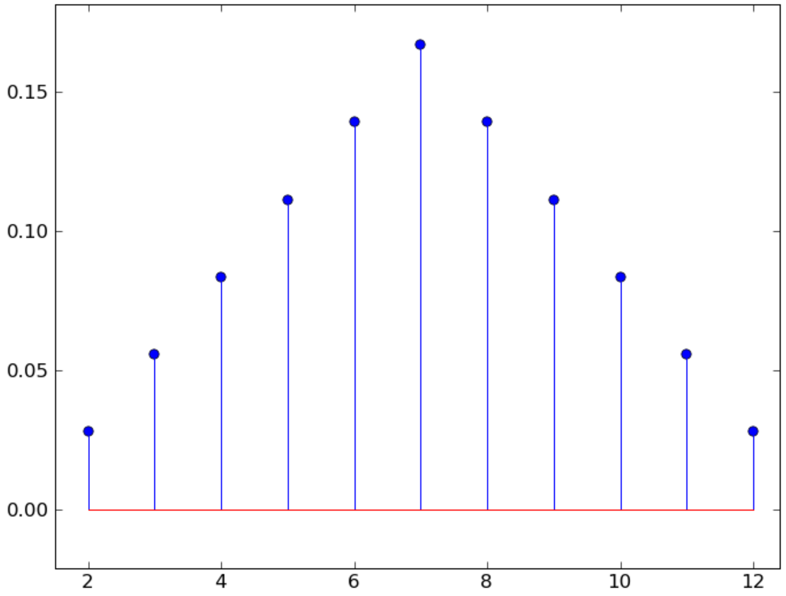


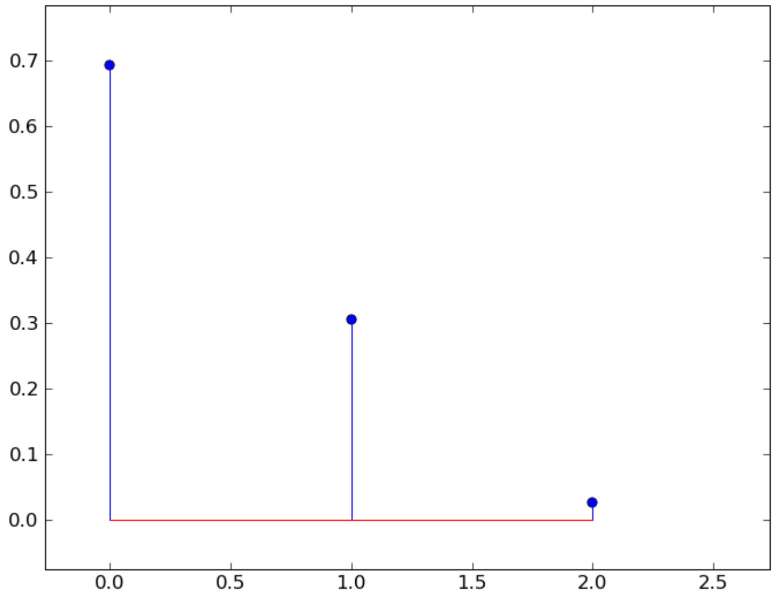
Diagrams: Wikipedia


0.15
0.10
0.05
0.00
2
4
6
8
10
12

Examples

- Throwing two dice,
  - $\Omega=\{(1,1), (1,2),\ldots(1,6),(2,1),\ldots(6,6)\}$
  - (1.3) The sum of the two dice, Z,
    $\Omega_Z=\{2, 3, \ldots, 12\}$
    - $p_Z(2) = P(\{(1,1)\} = 1/36$
    - $p_Z(7) = 6/36$
    - $F_Z(7) = 1+2+\ldots+6=21/36$
  - (1.1) The number of 6s X, $\Omega_X=\{0, 1, 2\}$
    - $p_X(2) = P(\{(6,6)\} = 1/36$
    - $p_X(1) = P(\{(6,x)|x\neq 6\}+ P(\{(x,6)|x\neq 6\}=10/36$
    - $px(0) = 25/36$





Mean – example

- Throwing two dice, what is the mean value of their sum?
- (2+3+4+5+6+7+
  3+4+5+6+7+8+
  4+5+6+7+8+9+
  5+6+7+8+9+10+
  6+7+8+9+10+11+
  7+8+9+10+11+12)/36=
- (2 + 2*3 + 3*4 + 4*5 + 5*6 + 6*7 + 5*8 +…2*11+12)/36=
- (1/36)2 + (2/36)*3 + (3/36)*4 + … + (1/36)*12 =
- p(2)*2 + p(3)*3 + p(4)*4 + … p(12)*12 =
- Σp(x)*x

Mean of a discrete random variable

- The mean (or expectation) (forventningsverdi) of a discrete random variable X:

$$\mu_X = E(X) = \sum_x p(x)x$$

- Useful to remember

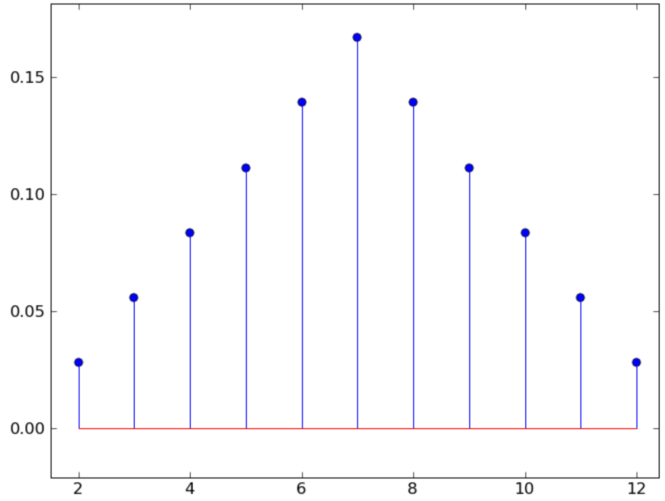
$$\mu_{(X+Y)} = \mu_X + \mu_Y$$

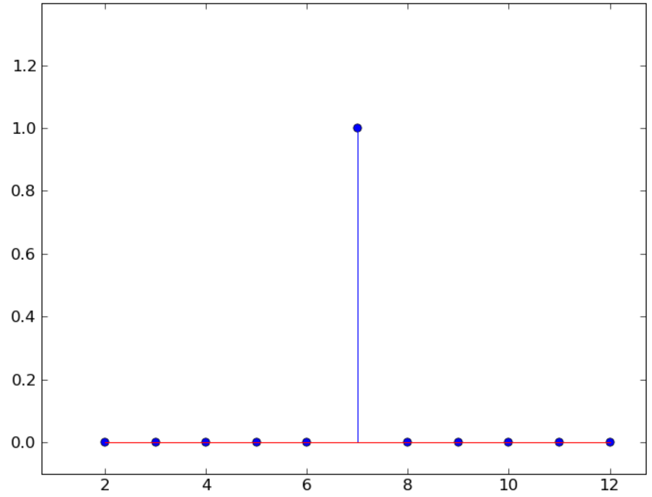
$$\mu_{(a+bX)} = a + b\mu_x$$

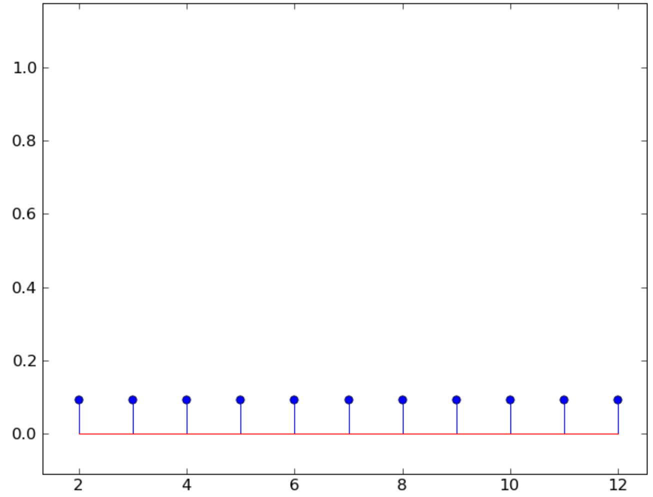
Examples:
One dice: 3.5
Two dice: 7
Ten dice: 35

More than mean

- Mean doesn't say everything
- Examples
  - (1.3) The sum of the two dice, Z, i.e.
    - $p_Z(2) = 1/36, \ldots, p_Z(7) = 6/36$ etc
- (3.2) p_2 given by:
  - $p_2(7)=1$
  - $p_2(x)= 0$ for $x \neq 7$
- (3.3) p_3 given by:
  - $p_3(x)= 1/11$ for $x = 2,3,\ldots,12$
- Have the same mean but are very different







Variance

- The variance of a discrete random variable X

$$Var(X) = \sigma^2 = \sum_x p(x)(x-\mu)^2$$

- The standard deviation of the random variable

$$\sigma = \sqrt{Var(X)}$$

Examples

- Throwing one fair dice
  - $\mu = (1+2+..+6)/6=7/2$
  - $\sigma^2 = ((1-7/2)^2 + (2-7/2)^2 + \ldots (6-7/2)^2)/6 = (25+9+1)/4*3=35/12$
- (Ex 1.3) Throwing two fair dice: 35/6
- (Ex 3.2) p_2, where $p_2(7)=1$ has variance 0
- (Ex 3.3) p_3, the uniform distribution, has variance:
  - $((2-7)^2 + \ldots (12-7)^2)/11 = (25+16+9+4+1+0)*2/11 = 10$

Take home

- Probability space
  - Random experiment (or trial) (no: forsøk)
  - Outcomes (utfallene)
  - Sample space (utfallsrommet)
  - An event (begivenhet/hendelse)
  - Bayes theorem
- Discrete random variable
  - The probability mass function, pmf
  - The cumulative distribution function, cdf
  - The mean (or expectation) (forventningsverdi)
  - The variance of a discrete random variable X
  - The standard deviation of the random variable