

# Statistical Natural Language Processing

## A refresher on probability theory

Çağrı Çöltekin

University of Tübingen  
Seminar für Sprachwissenschaft

Summer Semester 2018

## Why probability theory?

*But it must be recognized that the notion 'probability of a sentence' is an entirely useless one, under any known interpretation of this term. — Chomsky (1968)*

Short answer: practice proved otherwise.

### Slightly long answer

- Many linguistic phenomena are better explained as tendencies, rather than fixed rules
- Probability theory captures many characteristics of (human) cognition, language is not an exception

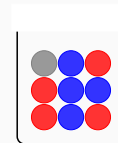
## What is probability?

- Probability is a measure of (un)certainty
- We quantify the probability of an event with a number between 0 and 1
  - 0 the event is impossible
  - 0.5 the event is as likely to happen as it is not
  - 1 the event is certain
- The set of all possible *outcomes* of a trial is called *sample space* ( $\Omega$ )
- An *event* ( $E$ ) is a set of outcomes

Axioms of probability state that

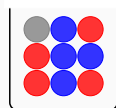
- $P(E) \in \mathbb{R}$ ,  $P(E) \geq 0$
- $P(\Omega) = 1$
- For *disjoint* events  $E_1$  and  $E_2$ ,  $P(E_1 \cup E_2) = P(E_1) + P(E_2)$

## What you should already know



- $P(\{\bullet\}) = 4/9$
- $P(\{\bullet\}) = 4/9$
- $P(\{\bullet\}) = 1/9$
- $P(\{\bullet, \bullet\}) = 8/9$
- $P(\{\bullet, \bullet, \bullet\}) = 1$
- $P(\{\bullet\bullet\}) = 16/81$
- $P(\{\bullet\bullet\}) = 16/81$
- $P(\{\bullet\bullet\}) = 4/81$
- $P(\{\bullet\bullet\}) = 1/81$
- $P(\{\bullet\bullet, \bullet\bullet\}) = 20/81$

## Where do probabilities come from



Axioms of probability do not specify how to assign probabilities to events.

Two major (rival) ways of assigning probabilities to events are

- Frequentist (objective) probabilities: probability of an event is its relative frequency (in the limit)
- Bayesian (subjective) probabilities: probabilities are degrees of belief

## Random variables

- A random variable is a variable whose value is subject to uncertainties
- A random variable is always a number
- Think of a random variable as mapping between the outcomes of a trial to (a vector of) real numbers (a real valued function on the sample space)
- Example outcomes of uncertain experiments
  - height or weight of a person
  - length of a word randomly chosen from a corpus
  - whether an email is spam or not
  - the first word of a book, or first word uttered by a baby

Note: not all of these are numbers

## Random variables

mapping outcomes to real numbers

- Continuous
  - frequency of a sound signal: 100.5, 220.3, 4321.3 ...
- Discrete
  - Number of words in a sentence: 2, 5, 10, ...
  - Whether a review is negative or positive:

Outcome	Negative	Positive
Value	0	1

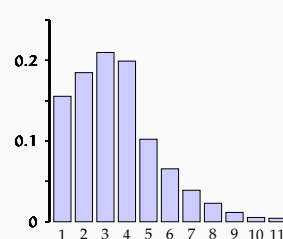
– The POS tag of a word:

Outcome	Noun	Verb	Adj	Adv	...
Value	1	2	3	4	...
...or	1 0 0 0	0 1 0 0	0 0 1 0	0 0 0 1	...

## Probability mass function

Example: probabilities for sentence length in words

- Probability mass function (PMF)* of a *discrete* random variable ( $X$ ) maps every possible ( $x$ ) value to its probability ( $P(X = x)$ ).



x	$P(X = x)$
1	0.155
2	0.185
3	0.210
4	0.194
5	0.102
6	0.066
7	0.039
8	0.023
9	0.012
10	0.005
11	0.004

## Populations, distributions, samples

- In many applications, we use probability theory to make inferences about a possibly infinite *population*
- A probability distribution is a way to characterize a population
- Our inferences are often based on *samples*

A sample from the distribution in the previous slide:

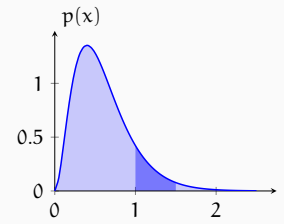
[1, 2, 2, 3, 3, 3, 4, 4, 5, 7, 11]

[[ IMAGE DISCARDED DUE TO `/tikz/external/mode=list and make' ]]

## Probability density function (PDF)

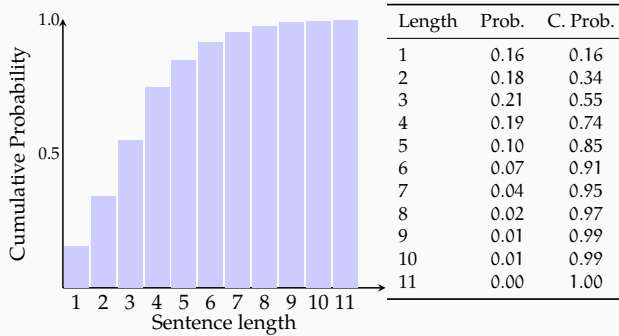
- Continuous variables have *probability density functions*
- $p(x)$  is not a probability (note the notation: we use lowercase  $p$  for PDF)
- Area under  $p(x)$  sums to 1
- $P(X = x) = 0$
- Non zero probabilities are possible for ranges:

$$P(a \leq x \leq b) = \int_a^b p(x) dx$$



## Cumulative distribution function

- $F_X(x) = P(X \leq x)$



## Expected value

- Expected value (mean) of a random variable  $X$  is,

$$E[X] = \mu = \sum_{i=1}^n P(x_i) x_i = P(x_1) x_1 + P(x_2) x_2 + \dots + P(x_n) x_n$$

- More generally, expected value of a function of  $X$  is

$$E[f(X)] = \sum_x P(x) f(x)$$

- Expected value is a measure of central tendency
- Note: it is not the 'most likely' value
- Expected value is linear

$$E[aX + bY] = aE[X] + bE[Y]$$

## Variance and standard deviation

- **Variance** of a random variable  $X$  is,

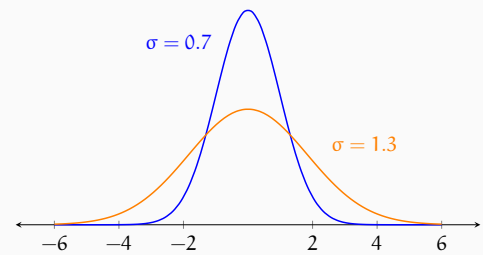
$$\text{Var}(X) = \sigma^2 = \sum_{i=1}^n P(x_i) (x_i - \mu)^2 = E[X^2] - (E[X])^2$$

- It is a measure of spread, divergence from the central tendency
- The square root of variance is called **standard deviation**

$$\sigma = \sqrt{\left( \sum_{i=1}^n P(x_i) x_i^2 \right) - \mu^2}$$

- Standard deviation is in the same units as the values of the random variable
- Variance is not linear:  $\sigma_{X+Y}^2 \neq \sigma_X^2 + \sigma_Y^2$  (neither the  $\sigma$ )

## Example: two distributions with different variances



## Short divergence: Chebyshev's inequality

For any probability distribution, and  $k > 1$ ,

$$P(|x - \mu| > k\sigma) \leq \frac{1}{k^2}$$

Distance from $\mu$	$2\sigma$	$3\sigma$	$5\sigma$	$10\sigma$	$100\sigma$
Probability	0.25	0.11	0.04	0.01	0.0001

This also shows why standardizing values of random variables,

$$z = \frac{x - \mu}{\sigma}$$

makes sense (the normalized quantity is often called the **z-score**).

## Median and mode of a random variable

**Median** is the mid-point of a distribution. Median of a random variable is defined as the number  $m$  that satisfies

$$P(X \leq m) \geq \frac{1}{2} \text{ and } P(X \geq m) \geq \frac{1}{2}$$

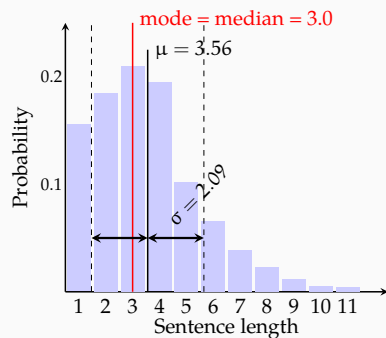
- Median of 1, 4, 5, 8, 10 is 5
- Median of 1, 4, 5, 7, 8, 10 is 6

**Mode** is the value that occurs most often in the data.

- Modes appear as peaks in probability mass (or density) functions
- Mode of 1, 4, 4, 8, 10 is 4
- Modes of 1, 4, 4, 8, 9, 9 are 4 and 9

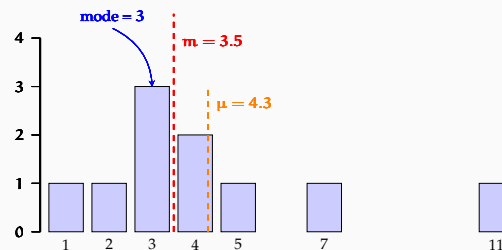
## Mode, median, mean, standard deviation

Visualization on sentence length example

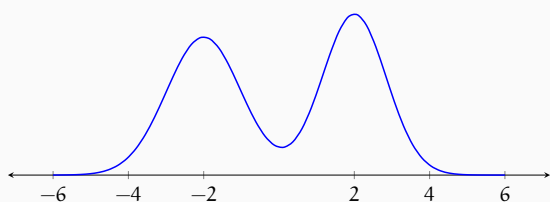


## Mode, median, mean

sensitivity to extreme values



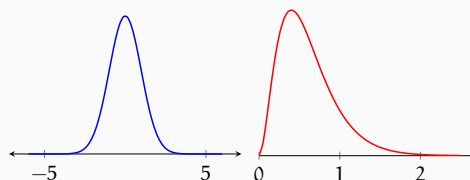
## Multimodal distributions



- A distribution is multimodal if it has multiple modes
- Multimodal distributions often indicate confounding variables

## Skew

- Another important property of a probability distribution is its *skew*
- **symmetric** distributions have no skew
- **positively skewed** distributions have a long *tail* on the right
- **negatively skewed** distributions have a long left tail

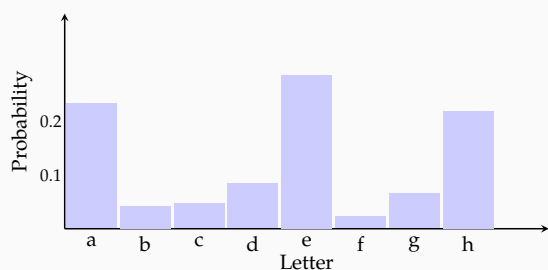


## Another example

A probability distribution over letters

- We have a hypothetical language with 8 letters with the following probabilities

Let.	a	b	c	d	e	f	g	h
Prob.	0.23	0.04	0.05	0.08	0.29	0.02	0.07	0.22



## Probability distributions

- Some random variables (approximately) follow a distribution that can be parametrized with a number of parameters
- For example, Gaussian (or normal) distribution is conventionally parametrized by its mean ( $\mu$ ) and variance ( $\sigma^2$ )
- Common notation we use for indicating that a variable  $X$  follows a particular distribution is

$$X \sim \text{Normal}(\mu, \sigma^2) \quad \text{or} \quad X \sim \mathcal{N}(\mu, \sigma^2).$$

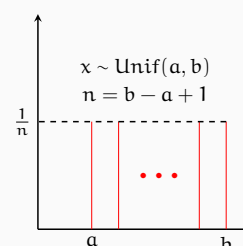
- For the rest of this lecture, we will revise some of the important probability distributions

## Probability distributions (cont)

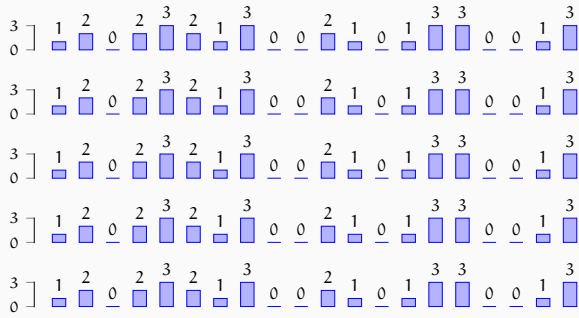
- A probability distribution is called *univariate* if it was defined on real numbers,
- *multivariate* probability distributions are defined on vectors
- Probability distributions are abstract mathematical objects (functions that map events/outcomes to probabilities)
- In real life, we often deal with *samples*
- A probability distribution is generative device: it can generate samples
- Finding most likely probability distribution from a sample is called *inference* (next week)

## Uniform distribution (discrete)

- A uniform distribution assigns equal probabilities to all values in range  $[a, b]$ , where  $a$  and  $b$  are the parameters of the distribution
- Probabilities of the values outside range is 0
- $\mu = \frac{1}{b-a+1}$
- $\sigma^2 = \frac{(b-a+1)^2-1}{12}$
- There is also an analogous continuous uniform distribution



## Samples from a uniform distribution



## Bernoulli distribution

Bernoulli distribution characterizes simple random experiments with two outcomes

- Coin flip: heads or tails
- Spam detection: spam or not
- Predicting gender: female or male

We denote (arbitrarily) one of the possible values with 1 (often called a success), the other with 0 (often called a failure)

$$P(X = 1) = p$$

$$P(X = 0) = 1 - p$$

$$P(X = k) = p^k(1 - p)^{1-k}$$

$$\mu_X = p$$

$$\sigma_X^2 = p(1 - p)$$

## Binomial distribution

Binomial distribution is a generalization of Bernoulli distribution to  $n$  trials, the value of the random variable is the number of 'successes' in the experiment

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

$$\mu_X = np$$

$$\sigma_X^2 = np(1 - p)$$

Remember that  $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ .

## Categorical distribution

- Extension of Bernoulli to  $k$  mutually exclusive outcomes
- For any  $k$ -way event, distribution is parametrized by  $k$  parameters  $p_1, \dots, p_k$  ( $k - 1$  independent parameters) where

$$\sum_{i=1}^k p_i = 1$$

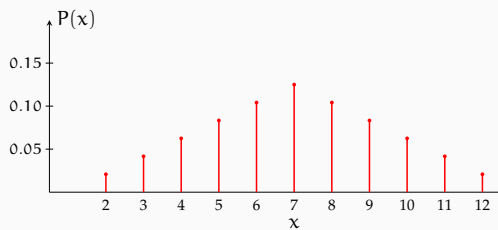
$$E[x_i] = p_i$$

$$\text{Var}(x_i) = p_i(1 - p_i)$$

- Similar to Bernoulli-binomial generalization, *multinomial* distribution is the generalization of categorical distribution to  $n$  trials

## Categorical distribution example

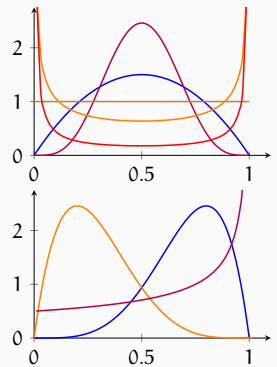
sum of the outcomes from roll of two fair dice



## Beta distribution

- Beta distribution is defined in range  $[0, 1]$
- It is characterized by two parameters  $\alpha$  and  $\beta$

$$p(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}}$$



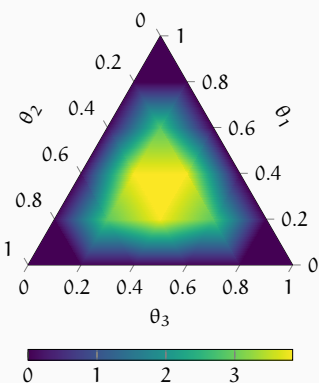
## Beta distribution

where do we use it

- A common use is the random variables whose values are probabilities
- Particularly important in Bayesian methods as a conjugate prior of Bernoulli and Binomial distributions
- *Dirichlet distribution* generalizes Beta to  $k$ -dimensional vectors whose components are in range  $(0, 1)$  and  $\|x\|_1 = 1$ .
- Dirichlet distribution is also used often in NLP, e.g., *latent Dirichlet allocation* is a well know method for topic modeling

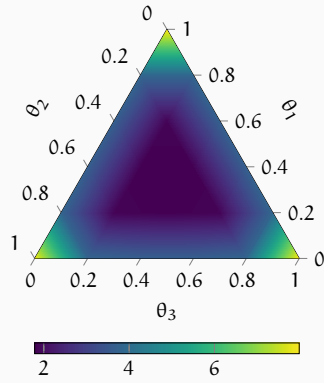
## Example Dirichlet distributions

$\theta = (2, 2, 2)$



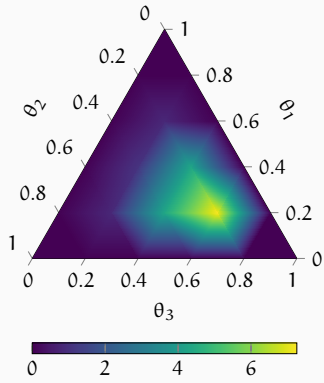
## Example Dirichlet distributions

$$\theta = (0.8, 0.8, 0.8)$$



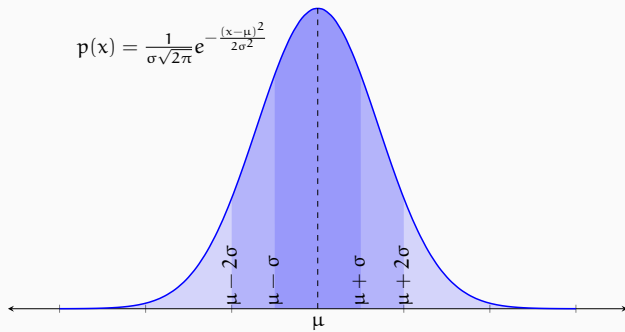
## Example Dirichlet distributions

$$\theta = (2, 2, 4)$$



## Gaussian (normal) distribution

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



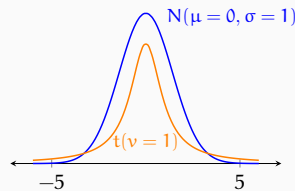
## Short detour: central limit theorem

**Central limit theorem** (CLT) states that the sum of a large number of independent and identically distributed variables (i.i.d.) is normally distributed.

- Expected value (average) of means of samples from any distribution will be distributed normally
- Many (inference) methods in statistics and machine learning works because of this fact

## Student's t-distribution

- T-distribution is another important distribution
- It is similar to normal distribution, but it has heavier tails
- It has one parameter: *degree of freedom* ( $v$ )



## Joint and marginal probability

Two random variables form a *joint probability distribution*.

An example: consider the letter bigrams.

	a	b	c	d	e	f	g	h	
a	0.04	0.02	0.02	0.03	0.05	0.01	0.02	0.06	0.23
b	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.04
c	0.02	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.05
d	0.02	0.00	0.00	0.01	0.02	0.00	0.01	0.02	0.08
e	0.06	0.02	0.01	0.03	0.08	0.01	0.01	0.07	0.29
f	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.02
g	0.01	0.00	0.00	0.01	0.02	0.00	0.01	0.02	0.07
h	0.08	0.00	0.00	0.01	0.10	0.00	0.01	0.02	0.22
	0.23	0.04	0.05	0.08	0.29	0.02	0.07	0.22	

## Expected values of joint distributions

$$\mathbb{E}[f(X, Y)] = \sum_x \sum_y P(x, y) f(x, y)$$

$$\mu_X = \mathbb{E}[X] = \sum_x \sum_y P(x, y) x$$

$$\mu_Y = \mathbb{E}[Y] = \sum_x \sum_y P(x, y) y$$

We can simplify the notation by vector notation, for  $\mu = (\mu_X, \mu_Y)$ ,

$$\mu = \sum_{x \in XY} x P(x)$$

where vector  $x$  ranges over all possible combinations of the values of random variables  $X$  and  $Y$ .

## Variances of joint distributions

$$\sigma_X^2 = \sum_x \sum_y P(x, y) (x - \mu_X)^2$$

$$\sigma_Y^2 = \sum_x \sum_y P(x, y) (y - \mu_Y)^2$$

$$\sigma_{XY} = \sum_x \sum_y P(x, y) (x - \mu_X)(y - \mu_Y)$$

- The last quantity is called *covariance* which indicates whether the two variables vary together or not

Again, using vector/matrix notation we can define the *covariance matrix* ( $\Sigma$ ) as

$$\Sigma = \mathbb{E}[(x - \mu)^2]$$

## Covariance and the covariance matrix

$$\Sigma = \begin{bmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{YX} & \sigma_Y^2 \end{bmatrix}$$

- The main diagonal of the covariance matrix contains the variances of the individual variables
- Non-diagonal entries are the covariances of the corresponding variables
- Covariance matrix is symmetric ( $\sigma_{XY} = \sigma_{YX}$ )
- For a joint distribution of  $k$  variables we have a covariance matrix of size  $k \times k$

## Correlation

Correlation is a normalized version of covariance

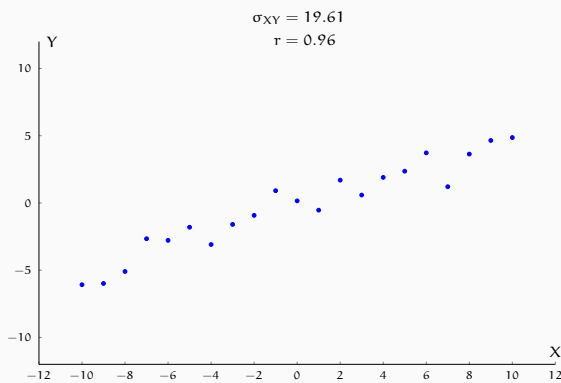
$$r = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

Correlation coefficient ( $r$ ) takes values between  $-1$  and  $1$

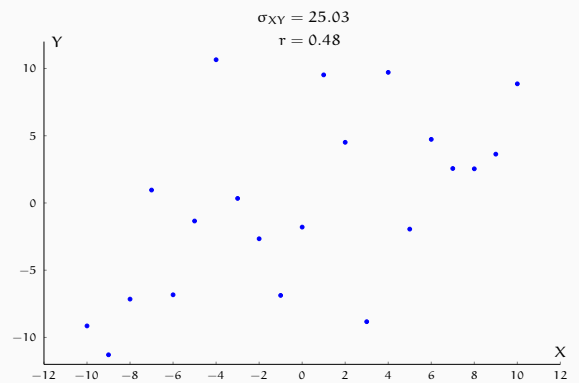
- $1$  Perfect positive correlation.
- $(0, 1)$  positive correlation:  $x$  increases as  $y$  increases.
- $0$  No correlation, variables are independent.
- $(-1, 0)$  negative correlation:  $x$  decreases as  $y$  increases.
- $-1$  Perfect negative correlation.

Note: like covariance, correlation is a symmetric measure.

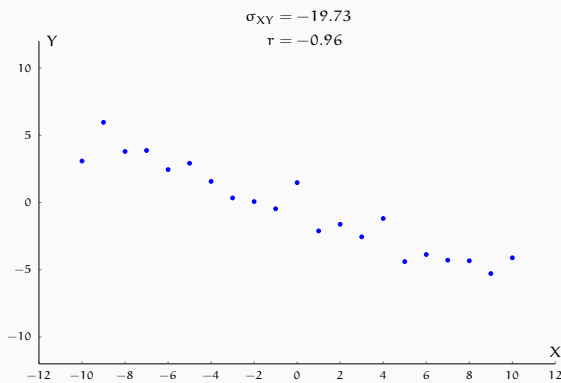
## Correlation: visualization (1)



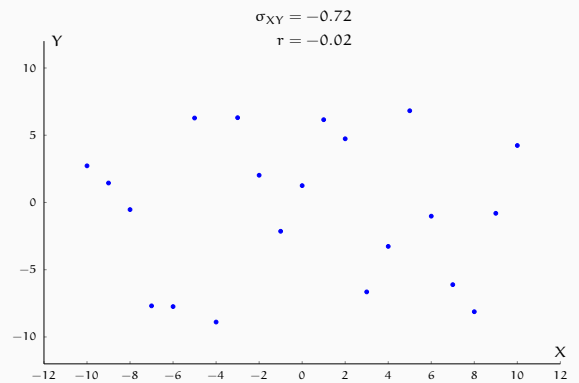
## Correlation: visualization (2)



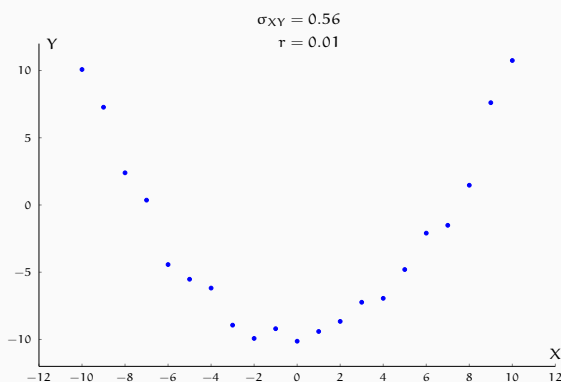
## Correlation: visualization (3)



## Correlation: visualization (4)



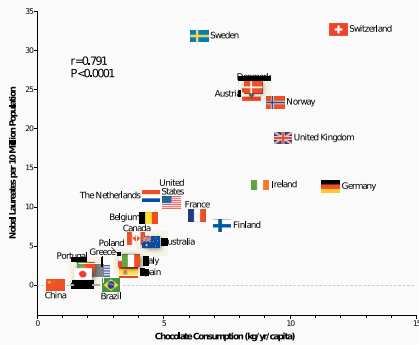
## Correlation: visualization (5)



## Correlation and independence

- Statistical (in)dependence is an important concept (in ML)
- The covariance (or correlation) of independent random variables is 0
- The reverse is not true: 0 correlation does not imply independence
- Correlation measures a linear dependence (relationship) between two variables, non-linear dependence may not be measured by covariance

## Short divergence: correlation and causation



From Messerli (2012).

## Conditional probability

In our letter bigram example, given that we know that the first letter is **e**, what is the probability of second letter being **d**?

	a	b	c	d	e	f	g	h	
a	0.04	0.02	0.02	0.03	0.05	0.01	0.02	0.06	0.23
b	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.04
c	0.02	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.05
d	0.02	0.00	0.00	0.01	0.02	0.00	0.01	0.02	0.08
e	0.06	0.02	0.01	0.03	0.08	0.01	0.01	0.07	0.29
f	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.02
g	0.01	0.00	0.00	0.01	0.02	0.00	0.01	0.02	0.07
h	0.08	0.00	0.00	0.01	0.10	0.00	0.01	0.02	0.22

$$P(L_1 = e, L_2 = d) = 0.025940365$$

$$P(L_1 = e) = 0.28605090$$

$$P(L_2 = d|L_1 = e) = \frac{P(L_1 = e, L_2 = d)}{P(L_1 = e)}$$

## Conditional probability (2)

In terms of probability mass (or density) functions,

$$P(X|Y) = \frac{P(X, Y)}{P(Y)}$$

If two variables are **independent**, knowing the outcome of one does not affect the probability of the other variable:

$$P(X|Y) = P(X) \quad P(X, Y) = P(X)P(Y)$$

More notes on notation/interpretation:

$P(X = x, Y = y)$  Probability that  $X = x$  and  $Y = y$  at the same time (joint probability)

$P(Y = y)$  Probability of  $Y = y$ , for any value of  $X$  ( $\sum_{x \in X} P(X = x, Y = y)$ ) (marginal probability)

$P(X = x|Y = y)$  Knowing that we  $Y = y$ ,  $P(X = x)$  (conditional probability)

## Bayes' rule

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

- This is a direct result of rules of probability
- It is often useful as it 'inverts' the conditional probabilities
- The term  $P(X)$ , is called **prior**
- The term  $P(Y|X)$ , is called **likelihood**
- The term  $P(X|Y)$ , is called **posterior**

## Example application of Bayes' rule

We use a test  $t$  to determine whether a patient has condition/illness  $c$

- If a patient has  $c$  test is positive 99% of the time:  
 $P(t|c) = 0.99$
- What is the probability that a patient has  $c$  given  $t$ ?
- ...or more correctly, can you calculate this probability?
- We need to know two more quantities. Let's assume  
 $P(c) = 0.00001$  and  $P(t|\neg c) = 0.01$

$$P(c|t) = \frac{P(t|c)P(c)}{P(t)} = \frac{P(t|c)P(c)}{P(t|c)P(c) + P(t|\neg c)P(\neg c)} = 0.001$$

## Chain rule

We rewrite the relation between the joint and the conditional probability as

$$P(X, Y) = P(X|Y)P(Y)$$

We can also write the same quantity as,

$$P(X, Y) = P(Y|X)P(X)$$

For more than two variables, one can write

$$P(X, Y, Z) = P(Z|X, Y)P(Y|X)P(X) = P(X|Y, Z)P(Y|Z)P(Z) = \dots$$

In general, for any number of random variables, we can write

$$P(X_1, X_2, \dots, X_n) = P(X_1|X_2, \dots, X_n)P(X_2, \dots, X_n)$$

## Conditional independence

If two random variables are conditionally independent:

$$P(X, Y|Z) = P(X|Z)P(Y|Z)$$

This is often used for simplifying the statistical models. For example in spam filtering with Naive Bayes classifier, we are interested in

$$P(w_1, w_2, w_3|\text{spam}) = P(w_1|w_2, w_3, \text{spam})P(w_2|w_3, \text{spam})P(w_3|\text{spam})$$

with the assumption that occurrences of words are independent of each other given we know the email is spam or not,

$$P(w_1, w_2, w_3|\text{spam}) = P(w_1|\text{spam})P(w_2|\text{spam})P(w_3|\text{spam})$$

## Continuous random variables

some reminders

The rules and quantities we discussed above apply to continuous random variables with some differences

- For continuous variables,  $P(X = x) = 0$
- We cannot talk about probability of the variable being equal to a single real number
- But we can define probabilities of ranges
- For all formulas we have seen so far, replace summation with integrals
- Probability of a range:

$$P(a < X < b) = \int_a^b p(x) dx$$

## Multivariate continuous random variables

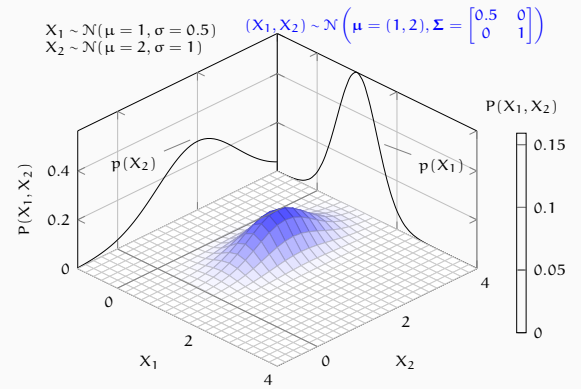
- Joint probability density

$$p(X, Y) = P(X|Y)P(Y) = P(Y|X)P(X)$$

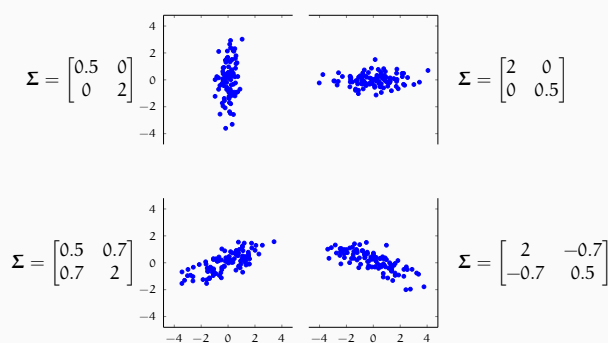
- Marginal probability

$$P(X) = \int_{-\infty}^{\infty} p(x, y) dy$$

## Multivariate Gaussian distribution



## Samples from bi-variate normal distributions



## Summary: some keywords

- Probability, sample space, outcome, event
- Outcome, event, sample space
- Random variables: discrete and continuous
- Probability mass function
- Probability density function
- Cumulative distribution function
- Expected value
- Variance / standard deviation
- Median and mode
- Skewness of a distribution
- Joint and marginal probabilities
- Covariance, correlation
- Conditional probability
- Bayes' rule
- Chain rule
- Some well-known probability distributions:
 

Bernoulli	binomial
categorical	multinomial
beta	Dirichlet
Gaussian	Student's t

## Next

Now Information theory  
 Wed Exercises  
 Fri ML Intro / regression  
 Mon Classification

## References and further reading

- MacKay (2003) covers most of the topics discussed, in a way quite relevant to machine learning. The complete book is available freely online (see the link below)
- See Grinstead and Snell (2012) a more conventional introduction to probability theory. This book is also freely available
- For an influential, but not quite conventional approach, see Jaynes (2007)



Chomsky, Noam (1968). "Quine's empirical assumptions". In: *Synthese* 19.1, pp. 53–68. doi: 10.1007/BF00569049.



Grinstead, Charles Miller and James Laurie Snell (2012). *Introduction to probability*. American Mathematical Society. ISBN: 9780821894149. URL: [http://www.dartmouth.edu/~chance/teaching\\_aids/books\\_articles/probability\\_book/book.html](http://www.dartmouth.edu/~chance/teaching_aids/books_articles/probability_book/book.html)



Jaynes, Edwin T (2007). *Probability Theory: The Logic of Science*. Ed. by G. Larry Bretthorst. Cambridge University Press. ISBN: 978-05-2159-271-0.

## References and further reading (cont.)



MacKay, David J. C. (2003). *Information Theory, Inference and Learning Algorithms*. Cambridge University Press. ISBN: 978-05-2164-298-9. URL: <http://www.inference.phy.cam.ac.uk/itprnn/book.html>



Messerli, Franz H (2012). "Chocolate consumption, cognitive function, and Nobel laureates". In: *The New England journal of medicine* 367.16, pp. 1562–1564.