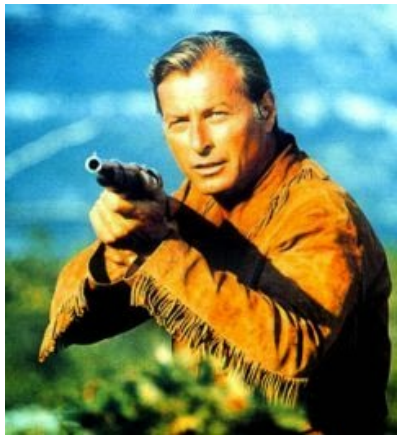
Probability Theory

Patrick van der Smagt

July 2017

slides almost completely by Justin Bayer

Some random events



Why Probability theory?

Different things have different inherent certainties, i.e. *entropy*:

- We cannot really say whether a tennis ball hitting the net on top goes to right or left—unless we have a *very sophisticated* physical model.
- We cannot predict Lotto at all, since the process is random by design.
- The stock market seems to correlate with all kinds of things.
- Old Shatterhand never misses. Or does he?

Why Probability theory?

- If the universe were deterministic, would we need probabilities?
- Yes, because we do not know everything.
- We need probability to express our *subjective uncertainty* about something.

$$\text{Probability} = \text{boolean logic} + \text{uncertainty}$$

"*The probability of an event is simply a fraction whose numerator is the number of favourable cases and whose denominator is the number of all the cases possible.*"

Pierre-Simon Laplace

Probability theory really comes down to counting. But counting can be hard.

Events and Random Variables 1

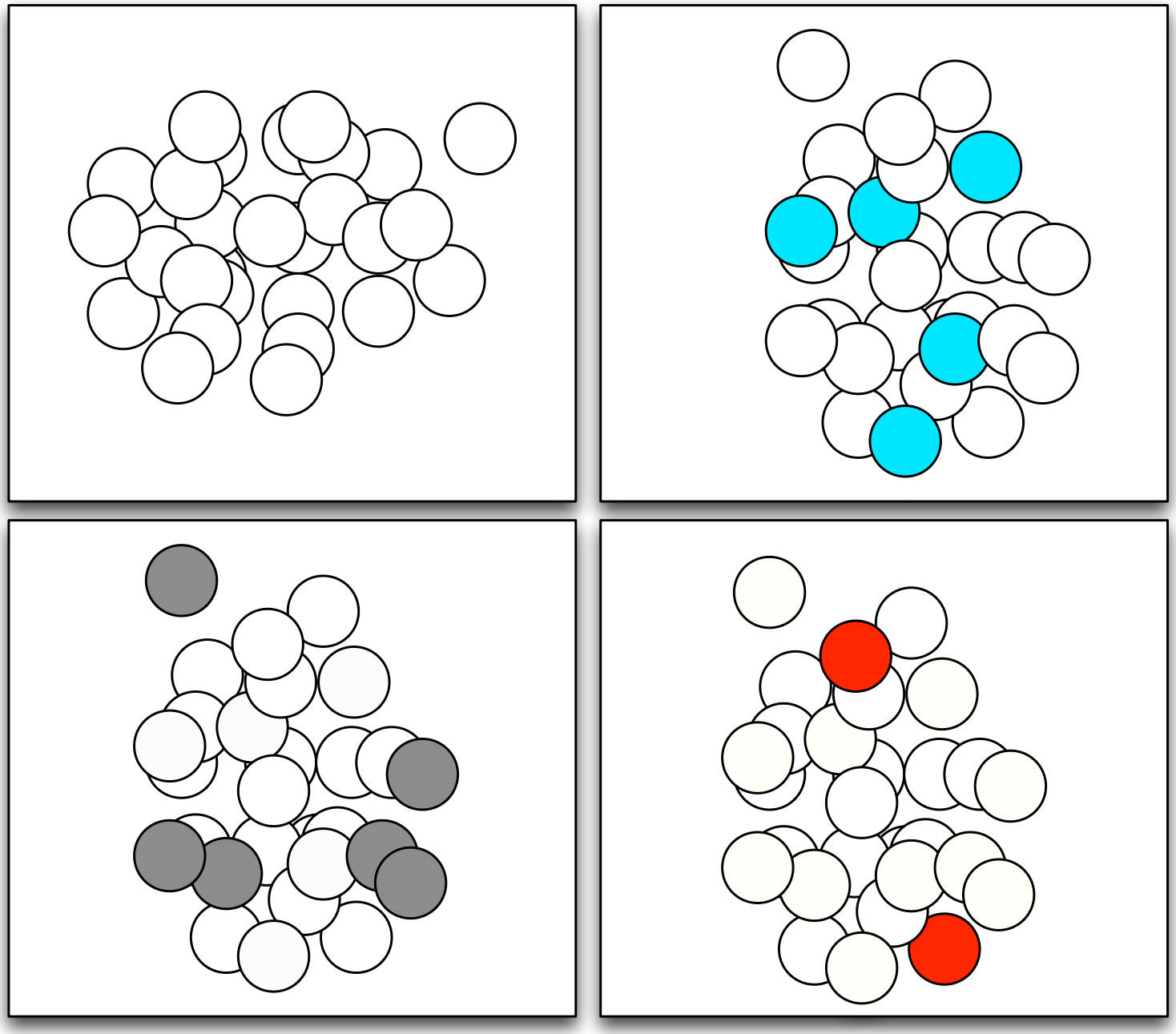
Consider a box with balls. Each ball has a number from $1 \ldots 10$ written on it. We draw a single ball.

Each ball corresponds to an *event*. We can design *random variables* based on that event; each random variable is a *function* of the event space.

Some examples. The number on the ball is

1. odd,
2. a prime number,
3. greater than 12,
4. less than 3.

These are all *indicator variables*, i.e. they are either *true* or *false*. That is statisticianlish for *binary*.



Events and Random Variables 2

We can also directly retrieve the number with a random variable:

$$n : \Omega \to \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$$

Ω is the set of events, i.e. the set of all balls.

Probabilities are fractions

Informal:

$$p(\text{“number on ball is odd”}) = \frac{\#\text{odd balls}}{\#\text{all balls}}$$

Formal:

Let $\omega \in \Omega$. Then the probability of the outcome of some indicator random variable $F : \Omega \to \{0, 1\}$ is

$$p(F = 1) = \frac{|\{\omega \in \Omega : F(\omega) = 1\}|}{|\Omega|}$$

More general:

$$p(F = f) = \frac{|\{\omega \in \Omega : F(\omega) = f\}|}{|\Omega|}$$

More than one random variable

Let us consider two draws from the box and call the random variables n_0 and n_1.

Question: Does it make a difference for the second ball if we put the first one back after looking at it?

Answer: Yes, because the distribution in the box is changed.

Question: Under what circumstances would it not make a difference?

Answer: If there were so many balls in the box, that one missing has only a very small effect. I.e., infinitely many.

Joint probability

We can also calculate the probability of two random variables taking certain values simultaneously:

$$p(F_0 = f_0, F_1 = f_1)$$

Example: Probability that a number on a ball is odd and divisible by 3. These random variables can highly influence each other!

Conditional probability

Consider we have already observed that a ball is odd. What is the probability that it is divisible by 3?

$$p(F_0 = f_0 \mid F_1 = f_1)$$

Marginal probability

Consider we have modelled the world in terms of F_0 and F_1. Then only considering one of them is called the *marginal probability*:

$$p(F_1 = f_1)$$

The most important slide of this set

Sum rule

$$p(X) = \sum_{Y} p(X, Y)$$

Product rule

$$p(X, Y) = p(X \mid Y)p(Y)$$

Bayes' theorem

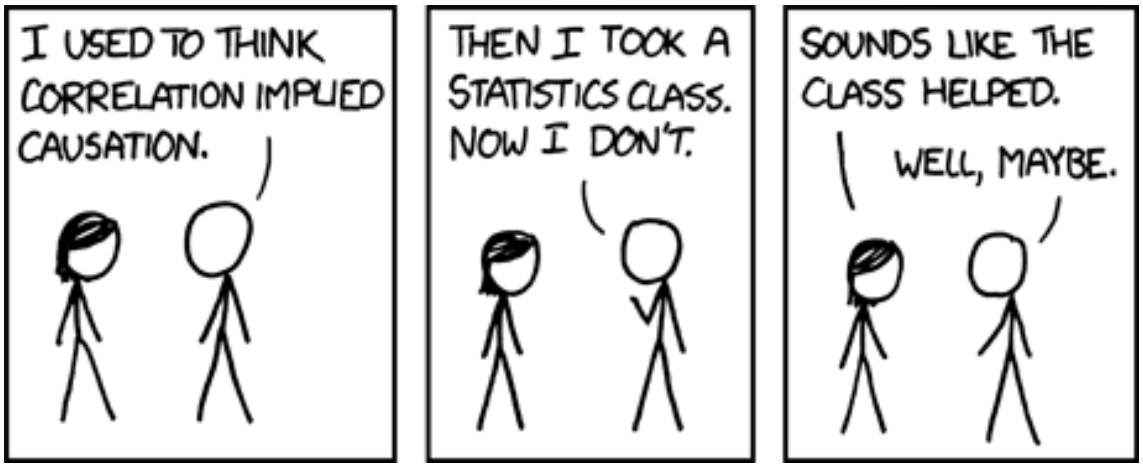
$$p(X \mid Y) = \frac{p(Y \mid X)p(X)}{p(Y)}$$

Independence

Two random variables are *independent* if they do not influence their respective outcomes.

$$
\begin{aligned}
p(X \mid Y) = p(X) &\Rightarrow \\
p(X, Y) = p(X)p(Y) &\Rightarrow \\
p(Y \mid X) = p(Y)
\end{aligned}
$$

But beware!



`https://xkcd.com/552/`

Continuous random variables 1

Most interesting data is continuous, i.e. $\mathbf{x} \in \mathbb{R}^D$: sensor readings, images, ...

A hack would be to just quantise those data. The explosion of dimensions prohibits this approach—also there is the so-called *curse of dimensionality*.

- The distance between two points becomes meaningless in high dimensions (they approach roughly the same value),
- Algorithms typically scale super-linear with dimensionality.

There is really no way around using continuous random variables.

Continuous random variables 2

Thankfully, many things from the previous slides still hold—with a few changes.

- Instead of counting events, we will have to integrate the area of an event (i.e. replace $\sum$ with $\int$),
- Consequently, $\int_x p(x)dx = 1$,
- The values of a probability function can now exceed 1,
- That is because $p(X = x)$ is meaningless—it has an area of 0,
- We now have to look at intervals, i.e. $p(a < X \leq b)$.

Expectation of a random variable

The expectation of a random variable is a way to summarise its outcome. It is the value it will take if we average infinitely many outcomes:

$$\mathbb{E}[f] = \sum_x f(x)p(x)$$

If we can draw samples from x, we can approximate it using S independent samples:

$$\mathbb{E}[f] \approx \frac{1}{S} \sum_{s=1}^{S} f(x_s), x_s \sim X.$$

This is special case of Monte Carlo integration.

Variance of a random variables

The expectation does not tell us anything about the spread of a random variable.

We might do a gamble if it will earn us $1'000 in expectation. But if it has a chance of 50% that we will lose $1'000'000 at the same time?

Variance captures this to some extent:

$$\begin{aligned}\mathrm{Var}[f] &= \sum_x \big(f(x) - \mathbb{E}[f]\big)^2 p(x) \\ &= \mathbb{E}\big[(f - \mathbb{E}[f])^2\big].\end{aligned}$$

It is the expected squared distance from the mean.

Covariance of two random variables

Often, we want to know how much two variables *vary together*. This can be quantified by the *covariance*:

$$\mathrm{Cov}[x, y] = \mathbb{E}[xy] - \mathbb{E}[f]\mathbb{E}[y].$$

If we look at a multidimensional **x** with multiple entries, the covariance matrix is defined as:

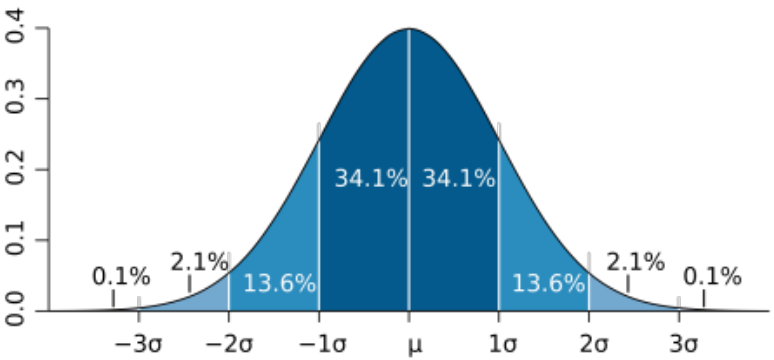
$$\mathrm{Cov}[\mathbf{x}, \mathbf{x}] = \mathbb{E}[\mathbf{x}, \mathbf{x}^T] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{x}]^T.$$

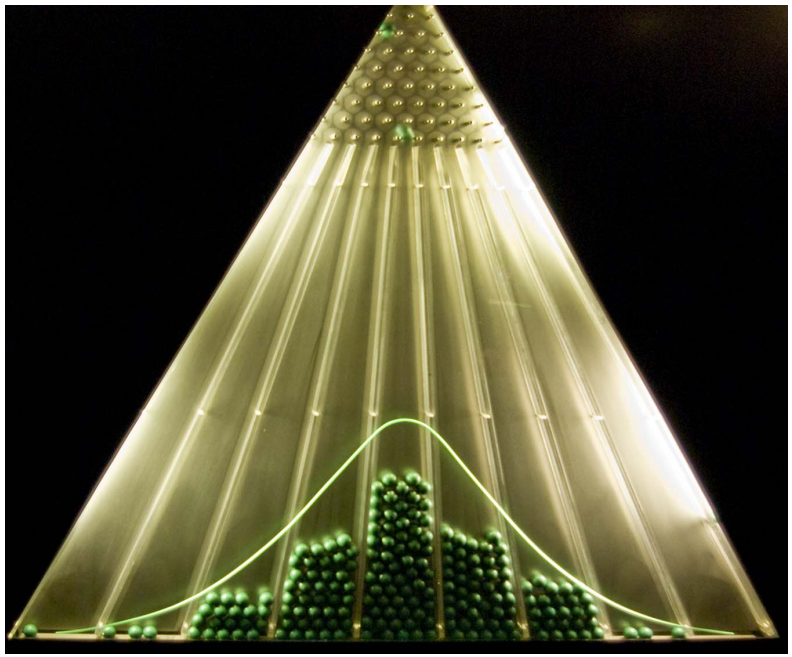
We will get to this again later during linear algebra!

Gaussian distribution

Most ~~important~~ convenient distribution for continuous data out there.

- Linear combinations of Gaussian random variables are Gaussian.
- Sums of sufficiently many independent random variables are Gaussian (aka central limit theorem).
- Many important operations can be done in closed form: sampling, evaluation of $p(x)$, conditioning, marginalisation, entropy, ...
- If we know nothing about our data, assuming Gaussianity is the smallest sin we can do.





Gaussian distribution 2

$$\mathcal{N}(x \mid \mu, \sigma^2) = \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}}}_{\text{normalisation}} \underbrace{\exp\left[-\frac{(x-\mu)^2}{\sqrt{2\sigma^2}}\right]}_{\geq 0}$$

The important part is in the exponential—the first part just makes sure that $\int_x \mathcal{N}(x \mid \mu, \sigma^2) dx = 1$.

Maximum Likelihood Estimation 2

If we observe a couple of random numbers $\mathcal{D} = \{x_i\}_{i=1}^N$, what can we say about their distribution?

We can calculate the mean and the variance. But why?

One way to get there is to assume Gaussianity of the random numbers and identify the parameters μ and σ^2 of the distribution using the maximum likelihood principle:

$$\hat{\mu}, \hat{\sigma}^2 = \arg\max_{\mu,\sigma^2} \prod_{i=1}^{N} \mathcal{N}(x_i \mid \mu, \sigma^2)$$

We maximise the likelihood of the data—i.e. we find the parameters, under which the parameters are the most likely.

Question: Why the product?

Answer: Because we—possibly wrongly—assumed that all x_i are independent of each other.

Maximum Likelihood Estimation 3

We will only do the expectation for now.

Logarithm to the rescue. Two important properties:

- Logarithm turns products into sums: $\log(\prod x_i) = \sum(\log x_i)$,
- Logarithm does not change the location of extrema: $\hat{z} = \arg\max_z f(z) \Rightarrow \hat{z} = \arg\max_z \log(f(z))$.

Maximum Likelihood Estimation 4

$$\begin{aligned}\mathcal{L}(\mu, \sigma^2) &= \log \prod_i p(x_i) \\ &= \sum_i \log \mathcal{N}(x_i \mid \mu, \sigma^2)\end{aligned}$$

Let's look at each of those terms separately.

$$\begin{aligned}\log \mathcal{N}(x_i \mid \mu, \sigma^2) &= \log \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[-\frac{(x_i - \mu)^2}{\sqrt{2\sigma^2}} \right] \right] \\ &= \underbrace{\log \frac{1}{\sqrt{2\pi\sigma^2}}}_{\text{const wrt } \mu} + \cancel{\log \exp} \left[-\frac{(x_i - \mu)^2}{\sqrt{2\sigma^2}} \right] \\ &\propto -(x_i - \mu)^2.\end{aligned}$$

Surprise: Fitting a Gaussian is the same as minimising squared error!

Maximum Likelihood Estimation 5

$$\begin{aligned}\mathcal{L}(\mu) &= \log \prod_i p(x_i) \\ &= \sum_i \log \mathcal{N}(x_i|\mu, \sigma^2) \\ &\propto -\sum_i (x_i - \mu)^2.\end{aligned}$$

Use analysis to find the minimum of this function. Necessary condition:

$$\begin{aligned}&\frac{\partial \sum_i -(x_i - \mu)^2}{\partial \mu} = 0 \\ \Rightarrow &-\sum_i \frac{\partial (x_i - \mu)^2}{\partial \mu} = 0 \\ \Rightarrow &-\sum_i \cancel{2} x_i - \cancel{2}\mu = 0 \\ \Rightarrow &\sum_i x_i = N\mu \Rightarrow \mu = \frac{\sum_i x_i}{N}.\end{aligned}$$

Summary

- We have seen quite a lot of complicated math today.
- It is not necessary to be able to get each detail.
- The overall idea is important.