Multivariate Verfahren

1. Some probability theory

Hannah Schulz-Kümpel

Department of Statistics, LMU Munich

Summer semester 2024

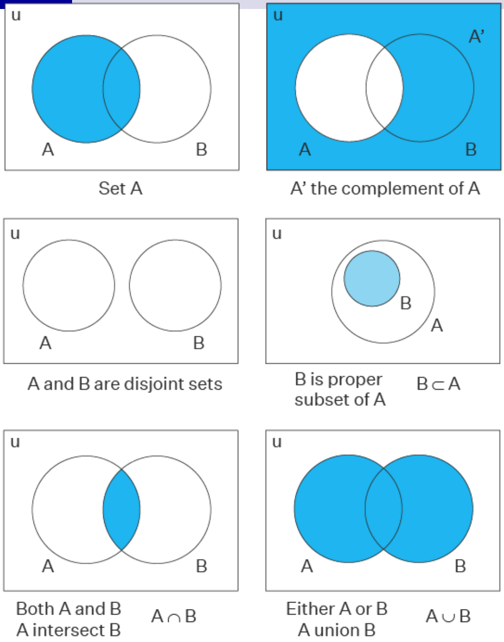
We will start at the very beginning:
The realm of probability theory!

1. Let's get philosophical

2. Probability spaces and operations

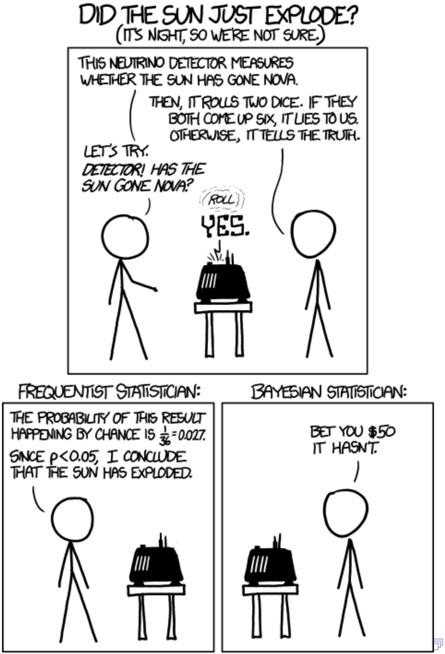
3. Random Variables and univariate distributions

Quick set theory reminder:



QUESTION:

What is your understanding of the term "probability"?


DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)
THIS NEUTRINO DETECTOR MEASURES WHETHER THE SUN HAS GONE NOVA.
THEN, IT ROLLS TWO DICE. IF THEY BOTH COME UP SIX, IT LIES TO US. OTHERWISE, IT TELLS THE TRUTH.
LET'S TRY.
DETECTOR! HAS THE SUN GONE NOVA?
ROLL
YES.
FREQUENTIST STATISTICIAN:
THE PROBABILITY OF THIS RESULT HAPPENING BY CHANCE IS 1/36 = 0.027. SINCE p<0.05, I CONCLUDE THAT THE SUN HAS EXPLODED.
BAYESIAN STATISTICIAN:
BET YOU $50 IT HASN'T.

Mathematics is here to help!

- So is there no "true" definition of probability?!
- Actually, there are two equivalent ways of formalizing the concept of probability:
  - Cox's theorem
  - The axioms of Kolmogorov (probability axioms)
    $\rightarrow$ *what we will focus on, since much more popular.*

Kolmogorov axioms - heuristic version I

- The axiomatic foundations of modern probability theory were laid only as recently as 1933!
- Specifically, they were published in the book *Foundations of the Theory of Probability* by Andrey Kolmogorov.

Kolmogorov axioms - heuristic version II

Heuristically, for an event space $\mathcal{S}$, i.e. the set of all possible events, the axioms state the following:

Axiom 1: For any event E, the probability of E is greater or equal to zero.

Axiom 2: The probability of the union of all events equals 1.

Axiom 3: For a countable sequence of mutually exclusive events $E_1, E_2, E_3, \ldots$ the probability of any of these events occurring is equal to the sum of each of the events occurring.

Contents

Formalizing probability I

- Of course, to derive the probability calculus and more complex results (like the CLT) which most of applied statistics is built on, we need a formal version of these axioms.
- Luckily, set- and measure- theory have us covered!
- We only need two definitions to get started:

Formalizing probability II

Definition (σ-Algebra)

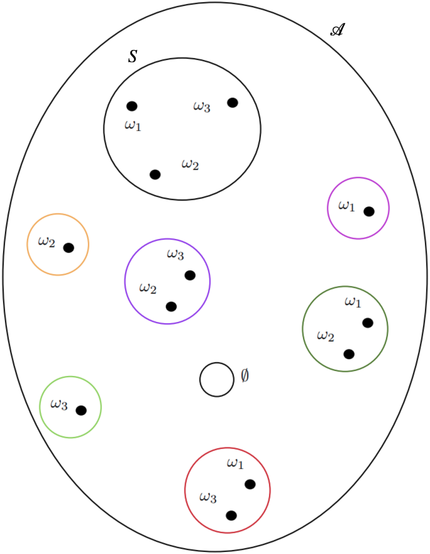
Given a set S, a collection $\mathcal{A}$ of subsets of S is called *σ-algebra over S*, if it satisfies the following properties:

1. $S \in \mathcal{A}$
2. $A \in \mathcal{A} \quad \Rightarrow \quad A^c \in \mathcal{A}$ ($\mathcal{A}$ is closed under complementation)
3. For sets $A_1, A_2, A_3, ... \in \mathcal{A} \Rightarrow \bigcup_{i \in \mathbb{N}} A_i \in \mathcal{A}$ ($\mathcal{A}$ is closed under countable unions)

- For countable sets S, the largest possible σ-algebra is the **power set**, i.e. the set containing all subsets of S, including the empty set and S itself. The power set of S is often denoted by $\mathcal{P}(S)$ or 2^S.

Formalizing probability III

An example:



$\hat{=}\mathcal{P}(S)$

Formalizing probability IV

Definition (Measure)

Consider a σ-algebra $\mathcal{A}$ over a set S. A function $\mu : \mathcal{A} \longrightarrow [0, \infty]$ that meets the following requirements

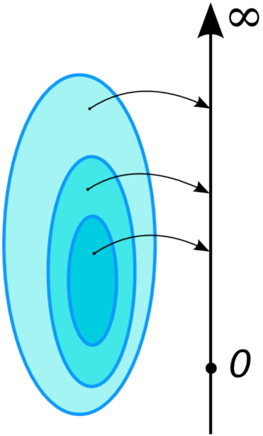
1. $\mu(\emptyset) = 0$
2. $\forall A \in \mathcal{A} : \quad \mu(A) \geq 0$
3. For pairwise disjoint sets $A_1, A_2, A_3, ... \in \mathcal{A} \quad \Rightarrow \quad \mu\Big(\bigcup_{i\in\mathbb{N}} A_i\Big) = \sum_{i\in\mathbb{N}} \mu(A_i).$

is called *measure*.

- **Example:** Cardinality. *We can easily check that the function that maps any set to the number of its elements fulfills the above definition of measure on σ-algebra $\mathcal{P}(S)$ for any finite set S.*

Formalizing probability V

- So measures are mathematical objects that quantify some definition of set-size:



By Oleg Alexandrov - Own work based on: Measure illustration.png, Public Domain, https://commons.wikimedia.org/w/index.php?curid=32489121

Formalizing probability VI

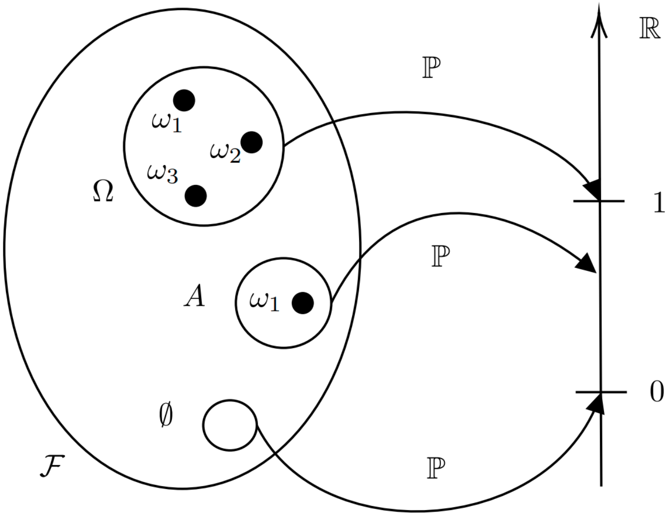
- Having defined the concepts of *σ-algebra* and *measure*, we can formalize the Kolmogorov axioms by
    - representing events as sets and
    - defining probability as a measure.

Definition (Probability measure)

Consider a σ-algebra $\mathcal{F}$ over a set Ω. A measure $\mathrm{P} : \mathcal{F} \longrightarrow [0, \infty]$ with $\mathrm{P}(\Omega) = 1$ is called a **probability measure** on $\mathcal{F}$.

- Note that by the definition of measure, the following has to hold for any probability measure: $\forall A \in \mathcal{F} : \ \mathrm{P}(A) \in [0, 1]$. This is why probability measures are often directly defined via $\mathrm{P} : \mathcal{F} \longrightarrow [0, 1]$.

Visualizing probability measures



Source: https://maurocamaraescudero.netlify.app/post/visualizing-measure-theory-for-markov-chains/

Probability spaces

Definition (Probability space)

A probability space $(\Omega, \mathcal{F}, \mathrm{P})$ consists of a nonempty set Ω, a σ-algebra $\mathcal{F}$ over Ω and a probability measure P on $\mathcal{F}$.

Now, by the definition of σ-algebra and probability measure the Kolmogorov axioms automatically hold and can be formally expressed as follows:

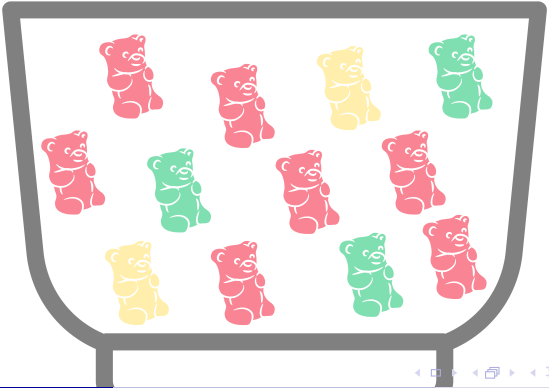
Axiom 1: $\mathrm{P}(A) \geq 0 \quad \forall A \in \mathcal{F}$.

Axiom 2: $\mathrm{P}(\Omega) = 1$.

Axiom 3: For pairwise disjoint sets $A_1, A_2, A_3, \ldots \in \mathcal{A}$
$\mathrm{P}\left(\bigcup_{i\in\mathbb{N}} A_i\right) = \sum_{i\in\mathbb{N}} \mathrm{P}(A_i)$.

Example: Gummy bears

- Consider a bowl with 2 yellow, 3 green, and 7 red gummy bears from which we want to randomly pick one.



Example: Gummy bears

- Here, we have a probability space consisting of
  - $\Omega = \{\{red\}, \{green\}, \{yellow\}\}$
  - $\mathcal{F} = \Big\{\emptyset, \{red\}, \{green\}, \{yellow\}, \{\{red\}, \{green\}\},$
    $\{\{red\}, \{yellow\}\}, \{\{yellow\}, \{green\}\}, \Omega\Big\}$ $\rightarrow$ Why?
  - $\mathrm{P} : \mathcal{F} \longrightarrow [0,1], \quad \mathrm{P}(A) \mapsto \begin{cases} \frac{7}{12}, & \text{if } A = \{red\}, \\ \frac{1}{4}, & \text{if } A = \{green\}, \\ \frac{1}{6}, & \text{if } A = \{yellow\}, \\ 0, & \text{otherwise.} \end{cases}$

Basic probability operations

- From the thus far established theory, we already automatically get some fundamental rules of probability, such as, for a probability space $\big(\Omega, \mathcal{F}, \mathrm{P}\big)$ and $A, B \in \mathcal{F}$:
  - $\mathrm{P}(A) = 1 - \mathrm{P}(A^c)$, because $1 = \mathrm{P}(\Omega) = \mathrm{P}(A \cup A^c) = \mathrm{P}(A) + \mathrm{P}(A^c)$.
  - $\mathrm{P}(\emptyset) = 0$, because $\Omega^c = \emptyset$.
  - $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$, with $\mathrm{P}(A \cap B) = 0$ for mutually exclusive events A and B, obviously.
- But we are still missing something, right?
  YES - **the concept of dependence**!

(In)dependence

Definition

Again, consider a probability space $\big(\Omega, \mathcal{F}, \mathrm{P}\,\big)$.

- Two events $A, B \in \mathcal{F}$ are called **independent**, if

$$\mathrm{P}(A \cap B) = \mathrm{P}(A)\,\mathrm{P}(B)\,.$$

- For $B \in \mathcal{F}$, the **conditional probability given** B for any $A \in \mathcal{F}$ is defined by

$$\mathrm{P}(A|B) := \begin{cases} \frac{\mathrm{P}(A\cap B)}{\mathrm{P}(B)}, & \text{if } \mathrm{P}(B) > 0, \\ 0, & \text{otherwise.} \end{cases}$$

Bayes' formula

- Note that, since $A \cap B = B \cap A$, it follows that

$$\mathrm{P}(A \cap B) = \mathrm{P}(A|B)\,\mathrm{P}(B) = \mathrm{P}(B|A)\,\mathrm{P}(A) = \mathrm{P}(B \cap A)\,.$$

- From the equality in the middle, we immediately get **Bayes' formula**

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

for any $B \in \mathcal{F}$ with $\mathrm{P}(B) \neq 0$.

Frequentist vs. Bayesian approach

Probability (mathematics)

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

Everyone uses Bayes' formula when the prior $P(H)$ is known.

Bayesian path

Frequentist path

Statistics (art)

$$P_{\text{Posterior}}(H|D) = \frac{P(D|H)P_{\text{prior}}(H)}{P(D)}$$

Bayesians require a prior, so they develop one from the best information they have.

Likelihood $L(H;D) = P(D|H)$

Without a known prior frequentists draw inferences from just the likelihood function.

<u>source:</u> Philippe Rigollet. 18.650 Statistics for Applications. Fall 2016. Massachusetts Institute of Technology: MIT OpenCourseWare, https://ocw.mit.edu.

Contents

Random variables (formal definition)

- You are probably already at least vaguely aware that random variables are functions, but usually ignore this fact in practice.
- Let's take another look at the definition of random variables, given the theoretical background we have just established.

Definition (Random Variables)

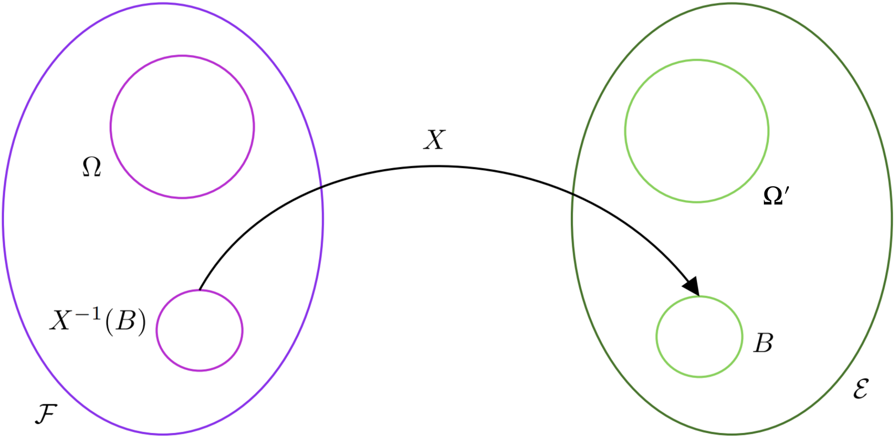
Consider a probability space $(\Omega, \mathcal{F}, \mathrm{P})$ and a measurable space $(\Omega', \mathcal{E})$, i.e. Ω' is a nonempty set and $\mathcal{E}$ a σ-algebra over Ω'.
A **random variable** with values in $(\Omega', \mathcal{E})$ is any measurable function

$$X : \Omega \longrightarrow \Omega', \quad \omega \mapsto X(\omega)\,,$$

i.e. any function $X : (\Omega, \mathcal{F}) \longrightarrow (\Omega', \mathcal{E})$ with

$$\forall E \in \mathcal{E} : \quad X^{-1}(E) := \{\omega \in \Omega | X(\omega) \in E\} \in \mathcal{F}\,.$$

Visualizing random variables



Source: https://maurocamaraescudero.netlify.app/post/visualizing-measure-theory-for-markov-chains/

Usual choices for $(\Omega', \mathcal{E})$ I

- Statisticians almost exclusively deal with **real random variables**, i.e. random variables that take values in $\mathbb{R}$ (or, depending on an authors definition $\mathbb{R}^p$, $p \in \mathbb{N}$) - we too will only consider real random variables from here on out.
- While this course's objective is to cover *multivariate statistics*, we will focus on one dimensional random variables in this lecture (i.e. $X : \Omega \longrightarrow \Omega' \subseteq \mathbb{R}$) and extend to higher dimensions a bit later.
- Fundamentally, we will usually deal with two different "kinds" of random variables:

Usual choices for $(\Omega', \mathcal{E})$ II

- **Discrete random variables** have a countable image $\Omega' \subseteq \mathbb{R}$, such as the natural numbers $\mathbb{N}$.[1]

  The power set $\mathcal{P}(\Omega')$ is usually chosen as the corresponding σ-algebra.

- **Continuous random variables** have image $\Omega' = \mathbb{R}$ and[2] the *Borel σ-algebra* $\mathcal{B}(\mathbb{R})$ is usually chosen as the corresponding σ-algebra.
  - There is some more complex theory behind Borel- sets and σ-algebras, but for the purposes of this lecture you may simply remember the following:
  - $\mathcal{B}(\mathbb{R})$ is the σ-algebra generated by the open sets, i.e., if $\mathcal{O}$ denotes the collection of all open subsets of $\mathbb{R}$, then $\mathcal{B}(\mathbb{R}) = \sigma(\mathcal{O})$.

[1] Technically, there is an alternative construction option - ask about it if you are interested ;)

[2] Having $\Omega' = \mathbb{R}$ is not technically a sufficient condition for a random variable to be continuous, they also need a suitable density - more on that later.

Distributions (formal definition)

- At first glance, this formal definition might seem a little unnecessarily complicated, but this formal set up gives rise to all kinds of relevant properties and results that are constantly used in applied statistics!

- The same goes for the formal definition of distribution:

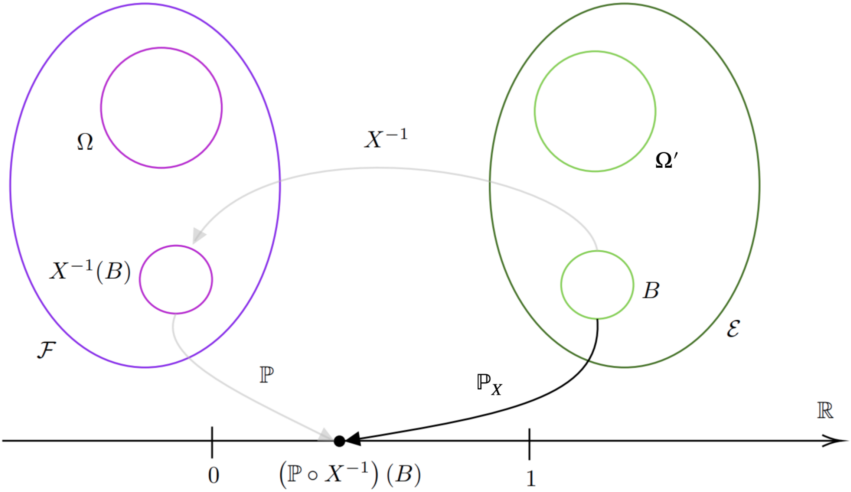
Definition (Distributions)

Given a probability space $(\Omega, \mathcal{F}, \mathrm{P})$ and a random variable X with values in $(\Omega', \mathcal{E})$, we define the **distribution** of X as the probability measure

$$\mathrm{P}_X := \mathrm{P} \circ X^{-1},$$

i.e. a function $P_X : \mathcal{E} \longrightarrow [0, 1]$.

Visualizing formal distributions



Source: https://maurocamaraescudero.netlify.app/post/visualizing-measure-theory-for-markov-chains/

Distributions as we routinely use them

- You are probably already familiar with the **cumulative distribution function (CDF)** $F(x) \equiv \mathrm{P}(X \leq x)$ of a random variable X.

- Given the established formal definition of distribution, we can now understand the formal definition of CDF as, for a probability space $(\Omega, \mathcal{F}, \mathrm{P})$ and random variable X with values in $(\Omega', \mathcal{E})$:

$$F(x) := \mathrm{P}_X([-\infty, x]) = \mathrm{P}\left(\{\omega \in \Omega | X(\omega) \leq x\}\right) \quad \forall x \in \mathbb{R}\,.$$

- The common notation $\mathrm{P}(X \leq x)$ is therefore a simplification of the term $\mathrm{P}\left(\{\omega \in \Omega | X(\omega) \leq x\}\right)$.

How is $\mathrm{P}(X \leq x)$ calculated? I

- The general idea for calculating $\mathrm{P}(X \leq x)$ is to calculate it as in interval $\int_{-\infty}^{x} \mathrm{d}\,\mathrm{P}_X$, which is defined separately for continuous and discrete random variables:

Definition

For a discrete random variable X, we have neatly chosen a construction where X has the **countable** image Ω'.
So, given the function $p : \mathbb{R} \longrightarrow [0,1]$, $x \mapsto \mathrm{P}_X(\{x\})$ with support $\operatorname{supp}(p) \equiv \{x \in \mathbb{R} : p(x) \neq 0\} \subset \Omega'$, we have

$$F(x) = \int_{-\infty}^{x} \mathrm{d}\,\mathrm{P}_X = \sum_{a \in [-\infty, x] \cap \operatorname{supp}(p)} p(a)\,.$$

The function p is referred to as **probability (mass) function**.
Note that, by definition, we automatically get $\sum_{x \in \operatorname{supp}(p)} p(x) = 1$.

How is $\mathrm{P}(X \leq x)$ calculated? II

Definition

For a continuous random variable X, we have

$$F(x) = \int_{-\infty}^{x} \mathrm{d}\,\mathrm{P}_X = \int_{-\infty}^{x} f(x)\mathrm{d}\lambda(x)\,,$$

where λ denotes the *Lebesgue measure* and f the **probability density function**, often simply density, defined as the derivative of the CDF.

Formally, we say that a probability measure has a density w.r.t. the Lebesgue measure λ, if the CDF F is absolutely continuous w.r.t. λ and then $f(x) := \frac{\partial F(x)}{\partial x}$.

Note that we now have, by definition of P_X, that any density f must be a function $f : \mathbb{R} \longrightarrow \mathbb{R}$ with $f(x) \geq 0 \;\forall x \in \mathbb{R}$ and $\int_{\mathbb{R}} f(x)\mathrm{d}x \big(\equiv \int_{\mathbb{R}} f(x)\mathrm{d}\lambda(x)\big) = 1$, which is the commonly taught definition of density.

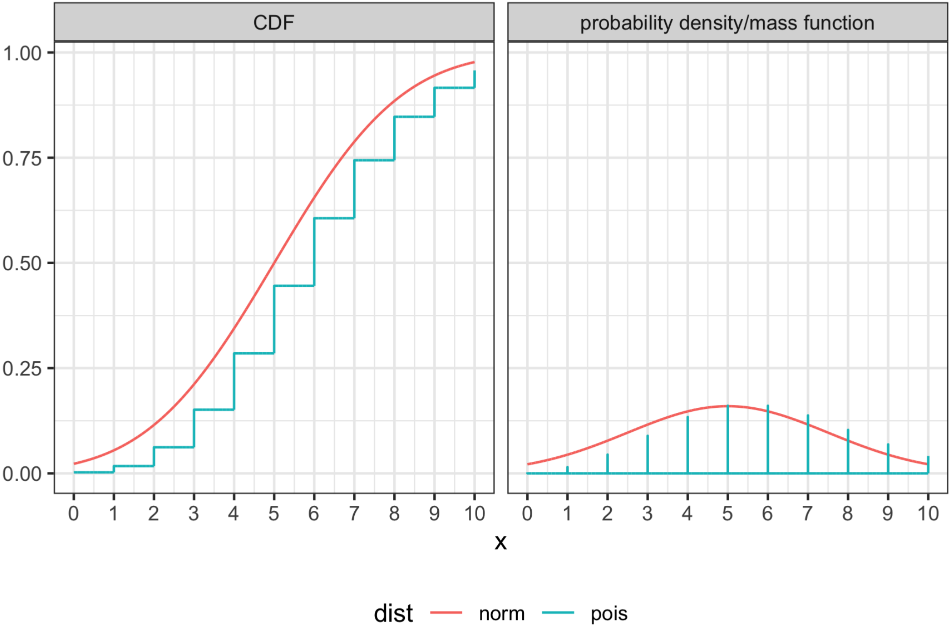
Example: Normal and Poisson distributions

```
library(dplyr)
library(tidyr)
library(ggplot2)

x<-seq(0,10,by=0.001)
df<-data.frame(x=rep(x,2),which=c(rep("probability density/
        mass function",length(x)),rep("CDF",length(x))))
df$pois<-c(dpois(x,6),ppois(x,6))
df$norm<-c(dnorm(x,5,2.5),pnorm(x,5,2.5))

df<-gather(df,dist,value,3:4) %>% as.data.frame()

ggplot(df,aes(x,value, colour = dist))+geom_line()+
  theme_bw()+scale_x_continuous(breaks=0:10)+
  ylab("")+theme(legend.position="bottom")+
  facet_wrap(~which)
```


CDF
probability density/mass function
1.00
0.75
0.50
0.25
0.00
0 1 2 3 4 5 6 7 8 9 10
0 1 2 3 4 5 6 7 8 9 10
x
dist
norm
pois

Outlook: Probabilistic modelling for regression I

Let's quickly consider how probabilistic thinking comes into play for the most simple of linear regressions. *(This will be discussed in more detail later!)*

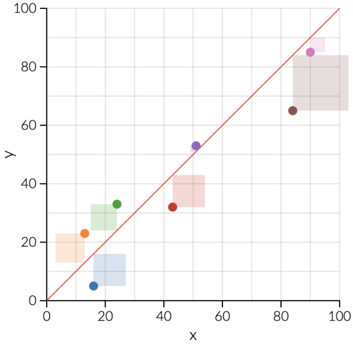
- Setting: we would like to model an outcome variable Y as a **linear** function of some regressor X.
- Probably you have seen

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \, ,$$

  where ε_i is an error term.
- Now, one approach to solving this problem (i.e. finding values for β_0 and β_1) is simply minimizing the error terms with regards to some loss function.

Outlook: Probabilistic modelling for regression II

If we choose squared loss, we get the popular OLS, i.e. minimizing the sum of squares in the following graphic:



(screenshotted from a very cool interactive post on OLS).

Outlook: Probabilistic modelling for regression III

- For the OLS solution, which we will talk more about in the next lecture, no probabilistic modelling is required at all!
- However, our interpretation is technically also limited - how would we phrase predicitions based on this?
  (keywords: causal inference; probabilistic modelling)
- Now, let's consider the following setting:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \, ,$$

with $\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$.

Outlook: Probabilistic modelling for regression IV

- It immediately follows that we consider the y_i to be realizations of a random variable $Y \sim N(\mathbb{E}[Y|X], \sigma^2)$ with

$$\mathbb{E}[Y|X] = \beta_0 + \beta_1 X \,.$$

- Now, if we take a frequentist view of things - *do not worry, this will be discussed more later* - all our information is given by the **Likelihood**

$$\begin{aligned}\mathcal{L}\Big(y; \beta = (\beta_0, \beta_1)^\top\Big) &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_i-(\beta_0+\beta_1 x_i)}{\sigma}\right)^2} \\ &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{\sum_{i=1}^{n} -\frac{1}{2}\left(\frac{y_i-(\beta_0+\beta_1 x_i)}{\sigma}\right)^2}\end{aligned}$$

and we find suitable estimates for β_0 and β_1 by maximizing the Likelihood $\Rightarrow \hat{\beta} = \underset{\beta=(\beta_0\beta_1)^\top\in\mathbb{R}^2}{\operatorname{argmax}} \mathcal{L}(y;\beta) = \underset{\beta=(\beta_0\beta_1)^\top\in\mathbb{R}^2}{\operatorname{argmax}} \log\big(\mathcal{L}(y;\beta)\big)$.

Outlook: Probabilistic modelling for regression V

- This results in (we will look at the general Maximum Likelihood transform for linear regression later)

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$$

- We will later see that the maximum likelihood estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are the same as the OLS ones for linear regression!

Outlook: Probabilistic modelling for regression VI

- But, a cool thing about specifically specifying the model using probabilistic tools is that we can then say

  "*For an observed X-value x_{value}, we predict the expectation of the target variable Y to be equal to $\hat{\beta}_0 + \hat{\beta}_1 x_{value}$*".

- Still, we should never loose sight of all the assumptions that we are making! What are they in our specific example?