Berner Fachhochschule
Haute école spécialisée bernoise
Bern University of Applied Sciences

Applied Statistics

Discrete Random Variables

Release FS24

Learning objectives

- Know the definition of a random variable
- Model a measurement correctly with a discrete random variable
- Know how to specify the distribution of a discrete random variable:
  - cdf, pmf
  - expected value, variance
- Recognize situations that must be modeled by a discrete uniform, Bernoulli, binomial or Poisson distribution

Random variables

A **random variable** is a variable that takes numerical values which depend on the outcome of a random experiment. It is an assignment of an event to a real number.

A more mathematical definition:

Definition (Random variable)

A **random variable** X is a function mapping a sample space Ω to $\mathbb{R}$ (or a subset), i.e. $X : \Omega \to \mathbb{R}$.

A random variable X induces a probability measure on $\mathbb{R}$.

Random variables: examples

1. **Family size**: choose a family at random from a population, let X be the number of children. Possible values: $X \in \mathbb{N}_0 = \{0, 1, 2, 3, \ldots\}$. If e.g. 23% of the families have 2 children, then $\mathrm{P}(X = 2) = 0.23$; if 72% of the families have *at most* 2 children, then $\mathrm{P}(X \in \{0, 1, 2\}) = \mathrm{P}(X \leq 2) = 0.72$.
2. **Length of fish**: measure the length of a fish randomly sampled from a population, let X be the length in cm. If e.g. 64% of the fish have a length between 11.5 cm and 16.2 cm, then $\mathrm{P}(X \in [11.5, 16.2]) = \mathrm{P}(11.5 \leq X \leq 16.2) = 0.64$.

What's the probability that the fish has *exactly* a length of 14 cm?

Random variables: notation

- Capital letter, e.g. X: random variable; lower case letter, e.g. x: realized value.
- $\{X = x\}$: elementary event that random variable X takes value x (with induced probability measure).

In words:

- Capital letter: description of an experiment (e.g., "measurement of the length of a fish")
- Lower case letter: outcome of the experiment (e.g., 13.5 cm)

2 types of random variables

- **Discrete random variable**: random variable with finite (or countable) image (i.e. set of possible values): $X : \Omega \to \{x_1, x_2, \dots\}$
- **Continuous random variable**: random variable whose image contains an interval, or $\mathbb{R}$.

The statement $\mathrm{P}(X = x)$ *only makes sense for a discrete random variable* X*.*

Describing the distribution

Next goal: describe probability "distribution" of a discrete random variable X including its characteristics.

Several quantities of interest:

- cumulative distribution function (cdf)
- probability mass function (pmf)
- expected value $\mathrm{E}(X)$
- variance $\mathrm{Var}(X)$
- ...

Cumulative distribution function

Definition (Cumulative distribution function)

The **cumulative distribution function** (cdf) of a random variable X is defined as $F_X(x) := \mathrm{P}(X \leq x)$.

Properties of a cdf F_X:

- F_X is monotonically increasing
- $\lim_{x \to -\infty} F_X(x) = 0$, $\lim_{x \to \infty} F_X(x) = 1$
- $\mathrm{P}(a < X \leq b) = F_X(b) - F_X(a)$

In fact this definition also holds for continuous random variables.

Quantile function

Definition (Quantile function)

Let X be a random variable with distribution function F_X and let $\alpha \in (0, 1)$. The α-**quantile** of X fulfills

$$\mathrm{P}(X \leq q) \geq \alpha \quad \text{and} \quad \mathrm{P}(X \geq q) \geq 1 - \alpha.$$

$\hookrightarrow$ There is an inverse relation between the quantiles and the values of the cdf.

Discrete random variables

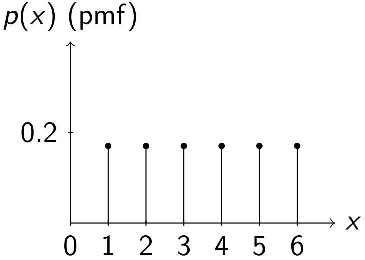
- **Discrete random variable**: random variable with finite (or countable) image (i.e. set of possible values): $X : \Omega \to \{x_1, x_2, \dots\}$
- Characterized by **probability mass function** (pmf) $p(x_k) := \mathrm{P}(X = x_k)$ with the following properties:
  - For each set $A \subset \{x_1, x_2, \dots\}$, we have

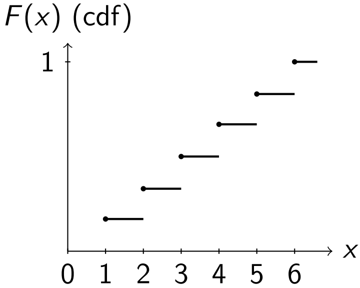
$$\mathrm{P}(X \in A) = \sum_{k : x_k \in A} \mathrm{P}(X = x_k)$$

  - Normalization: $\sum_k \mathrm{P}(X = x_k) = 1$
  - Connection to CDF: $F_X(x) = \mathrm{P}(X \leq x) = \sum_{k : x_k \leq x} \mathrm{P}(X = x_k)$

Example: fair die

A die can take values in $\{1, 2, \ldots, 6\}$; if it is fair, it takes all values with the same probability. Its probability mass function and cumulative distribution function look as follows:





Expected value

What do we expect on average?

- Fair die:

x_k	1	2	3	4	5	6
$\mathrm{P}(X = x_k)$	1/6	1/6	1/6	1/6	1/6	1/6

  On average we expect the mean number of spots, i.e. 3.5.

- Non-fair die:

x_k	1	2	3	4	5	6
$\mathrm{P}(X = x_k)$	1/10	1/10	1/10	1/10	1/10	1/2

  On average we expect the *weighted* mean number of spots, i.e. $0.1 \cdot (1 + 2 + 3 + 4 + 5) + 0.5 \cdot 6 = 4.5$.

Expected value

Definition (Expected value)

The **expected value** of a discrete random variable X is defined as

$$\mathsf{E}(X) := \sum_{k} x_k \cdot \mathrm{P}(X = x_k) \ .$$

Interpretation: The expected value $\mathsf{E}(X)$ corresponds to the weighted mean of all possible values of the random variable X. The weights are determined by the probability mass function.

Variance

How strong does X vary around $\mathsf{E}(X)$?

Definition (Variance)

The **variance** of a discrete random variable X is defined as

$$\mathsf{Var}(X) := \sum_k (x_k - \mathsf{E}(X))^2 \cdot \mathrm{P}(X = x_k) \ .$$

Example fair die: Let X be the result of a fair die. X has expected value $\mathsf{E}(X) = 3.5$ and variance $\mathsf{Var}(X) = 2.917$.

Transformations of random variables

Let X and Y be (continuous or discrete) random variables, and a and b two real numbers.

- $\mathrm{E}(aX + b) = a\,\mathrm{E}(X) + b$
- $\mathrm{Var}(aX + b) = a^2\,\mathrm{Var}(X)$

Furthermore,

- $\mathrm{E}(X + Y) = \mathrm{E}(X) + \mathrm{E}(Y)$
- $\mathrm{Var}(X + Y) \neq \mathrm{Var}(X) + \mathrm{Var}(Y)$
  (*only true, if X and Y are* **independent**)

Discrete probability distributions

We will now consider three discrete probability distributions widely used:

- (Discrete uniform distribution $\hookrightarrow$ dice)
- Bernoulli distribution
- Binomial distribution
- Poisson distribution

Bernoulli distribution

The Bernoulli distribution is the simplest non-trivial discrete probability distribution.

Definition (Bernoulli distribution)

A discrete random variable X that can only take the values 0 and 1 is said to have **Bernoulli distribution**. The distribution is specified by the probability $\pi := \mathrm{P}(X = 1)$.
We write $X \sim \text{Bernoulli}(\pi)$.

Binomial distribution

- Distribution of the sum of independent, identically distributed (iid) Bernoulli random variables
- Distribution of the number of "successes" of n independent trials with individual success probability π

Definition (Binomial distribution)

A discrete random variable $X \in \{0, 1, \ldots, n\}$ has **binomial distribution**, if

$$\mathrm{P}(X = x) = \binom{n}{x} \pi^x (1 - \pi)^{n-x} \; .$$

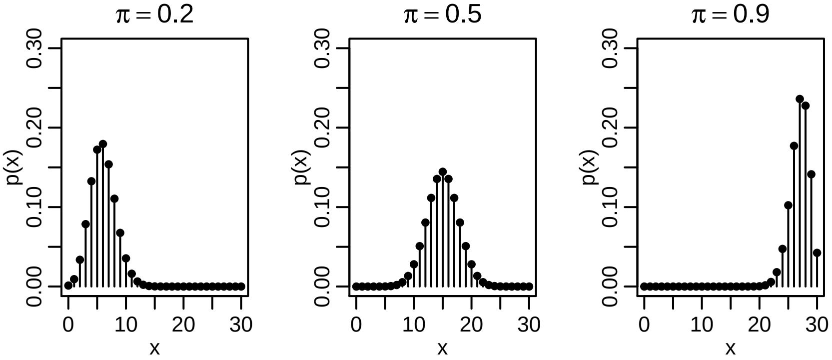
We write $X \sim \mathrm{Bin}(n, \pi)$, $n \in \mathbb{N}$, $\pi \in (0, 1)$.

Binomial distribution

What is the expected value of X?
Expected value: $\mathrm{E}(X) = n\pi$
Variance: $\mathrm{Var}(X) = n\pi(1 - \pi)$

Binomial distribution

Probability mass function of binomial distributions $\text{Bin}(30, \pi)$ for different probabilities π:



What would the corresponding cdf's look like?

Example: test of a new drug

- A new drug is tested on $n = 200$ patients. Subjects with a rare genetic disposition (incidence of $\pi = \frac{1}{1000}$) may have severe side effects.
- What's the probability that one patient in the study has this genetic disposition? Let X be the number of patients with this genetic disposition: $X \sim \text{Bin}(200, 0.001)$

$$\mathrm{P}(X = 1)200 \cdot 0.001 \cdot 0.999^{199} = 0.1639$$

- What's the probability that at least 3 patients have the disposition?

Python functions

In the library `scipy.stats` many useful functions are implemented

```
#---------------------------------------
# import libraries
from scipy.stats import binom, poisson
#---------------------------------------
>>> n, p = 200, 0.001
>>> binom.pmf(1,n,p)
0.16389365955527033
>>> 1-(binom.pmf(0,n,p)+binom.pmf(1,n,p)+binom.pmf(2,n,p))
0.0011337680974732312
>>> 1-binom.cdf(2,n,p)
0.001133768097462684
```

Poisson process

- Events occur *independently* of each other at random times
- Occurrence at a constant rate $\bar{\lambda}$ (expected number of events per time unit)
- Continuous time, not single discrete "trials" are considered
- Applications: Failure of a component of a machine; new customer joining the back of the queue; mutations on a chromosome; accidents on a certain stretch of road; etc.

Poisson distribution

Counting number of events in a given time interval d of a Poisson process, leads to a Poisson distributed random variable with parameter $\lambda = \bar{\lambda} \cdot d$ (expected value).

Definition (Poisson distribution)

A discrete random variable $X \in \mathbb{N}$ has **Poisson distribution** with parameter λ if

$$\mathrm{P}(X = x) = \mathrm{e}^{-\lambda} \frac{\lambda^x}{x!} \ .$$

We write $X \sim \mathrm{Po}(\lambda)$, $\lambda > 0$.

Expected value: $\mathrm{E}(X) = \lambda$
Variance: $\mathrm{Var}(X) = \lambda$

Link to binomial distribution

- Binomial distribution: range of possible values is limited
- Poisson distribution: number of successes in potentially infinitely many trials
- Poisson distribution as approximation of the binomial distribution: $\mathrm{E}(X) = n\pi = \lambda$, a constant, as $n \to \infty$

Example: Complaints

- \# complaints in a ward was 2, 5, 4, 3 in the last 4 months
- \# complaints $\sim$ Po(3.5)
- A new ward manager arrived and the number of complaints was 6 last month.

Are these surprisingly many complaints?

```
>>> poisson.pmf(6,3.5)
0.0770983498752680l
>>> poisson.pmf(5,3.5)
0.13216859978617376
>>> 1-poisson.cdf(5,3.5)
0.14238644690422175
```

The probability htat there are *at least* 6 complaints (given Po(3.5)) is 14% – so my answer would be **no**.

Properties of Poisson distributions

Proposition (Sum of Poisson random variables)

Let $X \sim Po(\lambda_1)$ and $Y \sim Po(\lambda_2)$ be independent. Then, $X + Y \sim Po(\lambda_1 + \lambda_2)$.

Is $\frac{1}{2}(X + Y)$ also Poisson distributed?