Random Variable
Probability Distribution

Peter Büchel

HSLU W

SA: W03

Reminder: Function

- Functions are everywhere in mathematics and therefore in statistics
- Recapitulate properties of functions which are most important for us
- First: "official" definition:

> In mathematics, a function from a set X to a set Y assigns to each element of X exactly one element of Y.
>
> The set X is called the *domain* of the function and the set Y is called the *codomain* of the function.

- This definition is quite abstract and general: Give an example

Example: Grades

- Let P be number of points you can score in an exam, e.g.:

$$P = \{0, 1, 2, \ldots 20\}$$

- Usually, these will be translated in grades G, e.g., in Switzerland:

$$G = \{1, 1.5, \ldots 6\}$$

- Some notation: Let p denote points scored and g corresponding grade

- A *function*, denoted by f, assigns to each p corresponding g:

$$f(p) = g$$

- Variable p: *Independent* variable
- Variable g: *Dependent* variable
- Grade g *depends* on number of points p
- For example:

$$f(0) = 1 \qquad \text{oder} \qquad f(20) = 6 \qquad \text{oder} \qquad f(12) = 4.5$$

- Note: Did not specify so far *how* this assignment works

- For all points in P: *Exactly one* corresponding grade exists in G

- *Not* possible:

$$f(10) = 4 \quad \textit{and} \quad f(10) = 3.5$$

- Lecturer gave for same number of points different grades

- One student would fail, but not the other one

- Someone of you would quite rightly complain

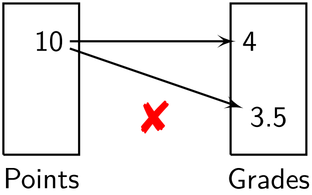
- However, it *is* possible:

$$f(8) = 3.5 \qquad \text{and} \qquad f(10) = 3.5$$

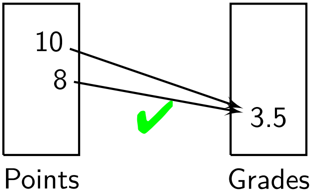
- For different number of points: Same grades are possible

Graphically

- Not possible:



- Possible:



Remarks

- In school and books: Functions are often denoted by $y = f(x)$ or something similar
- Generally: This notation is avoided in this module
- Use notation which is close to problem as in Example above
- Further notions about functions will be introduced as we go along

Random Variable: Example

- Very important notion in statistics: *Random variable*
- Pack of playing cards (Switzerland): 36 different cards, with 4 suits and card values $6, 7, 8, 9, 10$, Jack, Queen, King and Ace each
- Draw three cards in a row, put them back into deck after each draw
- Do this twice and get following results:
    1. 6, Queen, King
    2. 8, Jack, Ace
- Question: Which result is “better”?
- Hard to compare in this form

- A solution: Assign *numbers* to individual playing cards:
  - 6, 7, 8, 9 have value 0
  - 10 has value 10
  - Jack has value 2
  - Queen has value 3
  - King has value 4
  - Ace has value 11
- Now draws are comparable:
  1. 6, Queen, King : $0 + 3 + 4 = 7$
  2. 8, Jack, Ace : $0 + 2 + 11 = 13$
- Second draw with these assigned values better than first

Random Variable

- Example above: Situation occurs frequently in statistics
- Random experiment with sample space Ω
- Numbers are assigned to all elementary events of Ω
- Every elementary event ω has a value

$$X(\omega) = x$$

- X: *Function*, which assigns number x to each elementary event ω
- This function: *Random variable*

Example

- Drawing cards from a deck of playing cards
- Each card a number is assigned:

$$\begin{aligned}
\omega = \text{Ace} \quad &\mapsto \quad X(\omega) = 11 \\
\omega = \text{King} \quad &\mapsto \quad X(\omega) = 4 \\
\vdots \quad & \quad\quad \vdots \\
\omega = \text{Six} \quad &\mapsto \quad X(\omega) = 0
\end{aligned}$$

- With numbers $X(\omega)$, e.g. "average" of drawn cards can be calculated
- Average of "6, Queen, King" equal to $\frac{7}{3}$
- For elementary events "6", "Queen" and "King" without numbers, notion "average" makes no sense at all

Example

- Roll a blue and red die together
- Sample space Ω (see earlier): Numbers on dice:

$$\Omega = \{11, 12, \dots 16, 21, 22, \dots, 26, \dots 66\}$$

- Different random variables can be defined on Ω
- Let X be random variable for sum of numbers:
  - Hence:

$$X(16) = 7 \qquad \text{or} \qquad X(31) = 4 \qquad \text{or} \qquad X(13) = 4$$

  - Values that random variable can take is called *range*:

$$\mathbb{W}_X = \{2, 3, 4 \dots, 11, 12\}$$

- Let Y be number on red die:
  - Then:

$$Y(16) = 6 \qquad \text{or} \qquad Y(31) = 1 \qquad \text{or} \qquad Y(13) = 3$$

  - Range:

$$\mathbb{W}_Y = \{1, 2, \ldots, 6\}$$

- Let Z be equal 0 for all elementary events:
  - Then:

$$Z(16) = 0 \qquad \text{or} \qquad Z(31) = 0 \qquad \text{or} \qquad Z(13) = 0$$

  - Range:

$$\mathbb{W}_Z = \{0\}$$

- Last example is a completely legitimate random variable
- How useful this is, is another question

Example

- Randomly select a person
- Sample space Ω: All people on this planet
- Many random variables are conceivable:
  - X: Random variable that assigns income to each person
  - Y: Random variable that assigns height to each person
  - Z: Random variable that assigns age to each person
- Following variables are *not* random variables:
  - Variable V assigns gender to each person
  - Variable W assigns corresponding nationality to each person
- "Result" of a random variable *has to be* a *number*

Definition: Random Variable

- Definition:

Random Variable

A *Random variable* X is a *function*:

$$\begin{array}{rccc} X: & \Omega & \to & W_X \subset \mathbb{R} \\ & \omega & \mapsto & X(\omega) \end{array}$$

Remarks

- $\mathbb{R}$: Real numbers (points on number line, all decimal fractions)
- Statistics X (or $Y, Z, \ldots$): Notation for *random variables* (*functions*)
- Functions in mathematics often in form:

$$y = f(x)$$

- Following distinction is very important:

> - *Random variable* are denoted by *capital* letters X (or Y, Z)
> - Corresponding *lowercase* letters x (or y, z) represents *specific value* that random variable can take
> - Event where random variable X takes value x:
>
> $$X = x$$

- Random variable: *Not* function X is random, but only argument ω:
  - Depending on outcome of random experiment ω, different value $x = X(\omega)$
  - Once ω is chosen, $X(\omega)$ is fixed, *not* random
- x is also called a *realisation* of random variable X

Examples

- Playing cards: Realisation $X = 11$ corresponds to drawing an Ace
- Dice: Realisation $X = 8$ corresponds to eye sum 8

Probability Distribution of Random Variable

- Already seen: Calculation probability $P(E)$ of an event E
- Analogously: Probability of a general realisation x of a random variable X

Example: Playing Cards

- Random variable X: Value of a drawn playing card
- Value of drawn card is 4
- Realisation is $X = 4$
- Corresponding probability:

$$P(X = 4)$$

- Realisation $X = 4$: Corresponds to drawing a King
- Looking for:

$$P(X = 4) = P(\{\omega \mid \omega = \text{ King}\}) = \frac{4}{36} = \frac{1}{9}$$

- Probability that a King is drawn

In General

- Definition:

> Values of random variable X (possible realisations of X) occur with certain probabilities.
>
> Probability that X takes value x:
>
> $$P(X = x) = P(\{\omega \mid \ X(\omega) = x\}) = \sum_{\omega; X(\omega)=x} P(\omega)$$

- Playing cards: For $x = 4$, ω are all possible Kings, whose respective probabilities are added up

Probability Distribution

- Before: Probability of *one* realisation determined
- Now: Calculate probabilities of *all* realisations
- Definition:

> **Probability distribution**
>
> *Probability distribution*: Associated probability is determined for *all* realisations of random variable

- For finite sample space: Probability distribution is a table

Example

- Random variable X: Value of card drawn
- Know already:
$$P(X = 4) = \frac{1}{9}$$
- Probability $P(X = 0)$ with Laplace probability:
$$P(X = 0) = \frac{16}{36} = \frac{4}{9}$$
- Probability $P(X = 2)$: Drawing a Jack:
$$P(X = 2) = \frac{4}{36} = \frac{1}{9}$$

- Probability distribution of X as table:

x	0	2	3	4	10	11
$P(X = x)$	$4/9$	$1/9$	$1/9$	$1/9$	$1/9$	$1/9$

- Values for $P(X = 1)$ or $P(X = 178)$ are not listed in table

- Reason: These values cannot occur

- Nevertheless assign probability, i.e. 0:

$$P(X = 1) = 0 \qquad \text{or} \qquad P(X = 178) = 0$$

- Add all values of probability distribution: 1
- A realisation *must* be drawn
- Hence:

$$P(X=0)+P(X=2)+P(X=3)+P(X=4)+P(X=10)+P(X=11)=1$$

Example: Eye Sum of Two Dice

- Probability distribution for random variable X:

x	2	3	4	5	6	7	8	9	10	11	12
$P(X = x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

- Fractions are not reduced: Better to see that sum of values must be 1

In General

Probability Distribution

List of $P(X = x)$ for all *possible* values $x_1, x_2, \ldots, x_n$ is called *discrete probability distribution* of discrete random variable X

Equation

$$P(X = x_1) + P(X = x_2) + \ldots + P(X = x_n) = 1$$

has to be satisfied

With Σ notation:

$$\sum_{\text{All possible } x} P(X = x) = 1$$

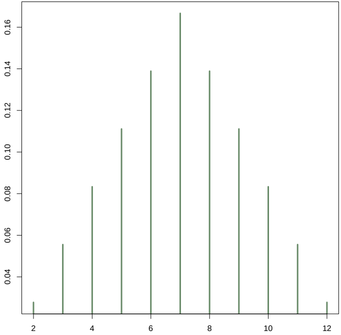
All values of a probability distribution add up to 1

Example: Eye Sum of Two Dice

- Already seen:

x	2	3	4	5	6	7	8	9	10	11	12
$P(X = x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

- Sketch:



Example: Eye Sum of Two Dice

- X: Random variable of eye sum rolled
- What is probability to roll exactly eye sum 6 ?
  - Sought: $P(X = 6)$:

$$P(X = 6) = \frac{5}{36}$$

- What is probability to roll eye sum 6 or 8 ?
  - Sought: $P(X = 6) + P(X = 8)$:

$$P(X = 6) + P(X = 8) = \frac{5}{36} + \frac{5}{36} = \frac{10}{36} = \frac{5}{18}$$

- What is probability to roll at most eye sum 3 ?
  - Sought:
$$P(X \leq 3) = P(X = 2) + P(X = 3)$$
  - In other words:
$$P(X = 2) + P(X = 3) = \frac{1}{36} + \frac{2}{36} = \frac{3}{36} = \frac{1}{12}$$

- What is probability to roll at least eye sum 3?
  - Sought:
$$P(X \geq 3) = P(X = 3) + \ldots + P(X = 12)$$
  - Simpler:
$$P(X \geq 3) = 1 - P(X = 2)$$
  - So:
$$1 - P(X = 2) = 1 - \frac{1}{36} = \frac{35}{36}$$

- What is probability to roll eye sum of 3 to 5 ?
  - Sought:

$$P(3 \leq X \leq 5) = P(X = 3) + P(X = 4) + P(X = 5)$$

  - So:

$$P(3 \leq X \leq 5) = \frac{2}{36} + \frac{3}{36} + \frac{4}{36} = \frac{9}{36} = \frac{1}{4}$$

Key Figures of Distribution

- Any (discrete) distribution: Simplified by 2 key figures:
  - *Expected value* $\mathrm{E}(X)$: Central location of distribution
  - *Standard deviation* $\sigma(X)$: Spread (dispersion) of distribution about $\mathrm{E}(X)$

- Discrete random variable X: Possible values $x_1, x_2, \ldots, x_n$
- Definition:

Expected value and standard deviation

- *Expected value*:

$$\begin{aligned} \mathrm{E}(X) &= x_1 \cdot P(X = x_1) + x_2 \cdot P(X = x_2) + \ldots + x_n \cdot P(X = x_n) \\ &= \sum_{\text{All possible } x} x \cdot P(X = x) \end{aligned}$$

- *Variance* and *standard deviation*:

$$\begin{aligned} \mathrm{Var}(X) &= (x_1 - \mathrm{E}(X))^2 \cdot P(X = x_1) + \ldots + (x_n - \mathrm{E}(X))^2 \cdot P(X = x_n) \\ &= \sum_{\text{All possible } x} (x - \mathrm{E}(X))^2 P(X = x) \end{aligned}$$

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

Example

- Rolling fair die: All 6 possible numbers have equal probability to occur
- Random variable X; Number rolled
- Expected value $\mathrm{E}(X)$:

$$
\begin{aligned}
\mathrm{E}(X) &= x_1 \cdot P(X = x_1) + x_2 \cdot P(X = x_2) + \ldots + x_6 \cdot P(X = x_6) \\
&= 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} \\
&= \frac{1}{6} \cdot (1 + 2 + 3 + 4 + 5 + 6) \\
&= 3.5
\end{aligned}
$$

- Expected value 3.5: Average of eye numbers

- How can this value be interpreted?
- Rolling fair die 100 times: Average mostly *not* exactly 3.5
- But: Average *close* to 3.5
- Approximation should always get better the more the die is rolled
- Roll 100 billion times: Average not *exactly* 3.5, but very close
- Interpretation: For many rolls average is very close to expected value
- See simulation later

- Calculate standard deviation with R:

```
x <- 1 : 6
p <- 1 / 6

E_X <- sum(x * p)

var_X <- sum((x - E_X)^2 * p)
sd_X <- sqrt(var_X)

sd_X

[1] 1.707825
```

- Means: Deviation on “average” 1.7 from 3.5

Example: Playing Cards

- Distribution:

x	0	2	3	4	10	11
$P(X = x)$	$4/9$	$1/9$	$1/9$	$1/9$	$1/9$	$1/9$

- Draw a card from deck
- What is average value of card being drawn?
- Calculate expected value $\mathrm{E}(X)$:

$$\mathrm{E}(X) = 0 \cdot \frac{4}{9} + 2 \cdot \frac{1}{9} + 3 \cdot \frac{1}{9} + 4 \cdot \frac{1}{9} + 10 \cdot \frac{1}{9} + 11 \cdot \frac{1}{9} = 3.33$$

- Values above each other in table are multiplied and then added up

- R:

```
x <- c(0, 2, 3, 4, 10, 11)
p <- 1 / 9 * c(4, 1, 1, 1, 1, 1)

E_X <- sum(x * p)
E_X

[1] 3.333333
```

- Average value to be expected if card is drawn and put back into deck very often
- Many cards with value 0: Expected value rather low

- Variance and standard deviation:

$$
\begin{aligned}
\operatorname{Var}(X) = {} & (0-3.33)^2 \cdot \frac{4}{9} + (2-3.33)^2 \cdot \frac{1}{9} + (3-3.33)^2 \cdot \frac{1}{9} \\
& + (4-3.33)^2 \cdot \frac{1}{9} + (10-3.33)^2 \cdot \frac{1}{9} + (11-3.33)^2 \cdot \frac{1}{9} \\
= {} & 16.67
\end{aligned}
$$

and

$$
\sigma(X) = \sqrt{16.67} = 4.08
$$

- R:

```
var_X <- sum((x - E_X)^2 * p)
sd_X <- sqrt(var_X)

sd_X


[1] 4.082483
```

- "Average" deviation 4.1: Rather large because of values 10, 11

Remarks

- Expected value of a discrete random variable: Weighted mean of all possible values, weighted by their probability of occurring
- Expected value: Often also referred to as μ_X:
  - Index X will often be omitted if random variable is clear
- Probabilities for all values $x_1, x_2, \ldots, x_n$ equal: Expected value equal mean of values
- Variance is square of spread of value of random variable from expected value weighted with respective weight
- Standard deviation has same unit as X: Unit of variance squared:
  - E.g. X in meters (m) measured: $\text{Var}(X)$ in square meter (m^2)
  - $\sigma(X)$ again dimension meter (m)

Difference between Empirical and Theoretical Key Figures

- Already seen: Average:

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- *Empirical* Variance:

$$\text{Var}(x) = \frac{(x_1 - \overline{x}_n)^2 + (x_2 - \overline{x}_n)^2 + \cdots + (x_n - \overline{x}_n)^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2$$

- Empirical standard deviation s_x:

$$s_x = \sqrt{\text{Var}(x)} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2}$$

- How are this definition and definition for mean and standard deviation for random variable X related?

- Distinguish between them very carefully:

  - Arithmetic mean $\bar{x}$: Calculated from *concrete* data: From observations $x_1, \ldots, x_n$ according to formula above for $\bar{x}_n$
  - Expected value $\mathrm{E}(X)$: *Theoretical value*, which results from *model*, i.e. distribution

- Hopefully: Arithmetic mean $\bar{x}$ approaches theoretical value $\mu_X = \mathrm{E}(x)$ for more and more experiments, provided data follow distribution of X

- If this is not the case: Something is wrong with model (e.g. all sides of die have *not* same probability to be rolled)

Example: Fair Die

- Each side has same probability of being rolled
- Justified assumption because of *symmetry*: All sides equal
- Expected value:
$$\mathrm{E}(X) = 3.5$$
- Rolling ideal, fair die $n = 10$ times
- Although die is fair: Mean very unlikely exactly 3.5
- Ideal fair die: R:

```
set.seed(4)
x <- sample(1:6, size = 10, replace = T)
x

mean(x)

[1] 3.8
```

- Average 3.8: Somewhat off 3.5
- Roll die again 10 times: Usually different result
- Simulation with R
- Simulate 10 times 10 tosses: Calculate corresponding averages

```
set.seed(3)
for (i in 1:10)
{
x <- sample(1:6, size = 10, replace = T)
cat(mean(x)," ")
}

3.3  3.5  4.2  3.4  2.9  3  3.9  2.9  4.1  2.7
```

- Averages from 2.7 to 4.2
- Close to 3.5 but not very accurate

- Roll die $n = 100$:

```
set.seed(2)
x <- sample(1:6, size = 100, replace = T)
x

[1] 5 6 6 1 5 1 4 5 1 2 3 1 3 6 2 3 1 6 1 4 3 6 1 6 5 6 6 3 1 5 5 6 6 2
    2 3 4 3 1 1 5 1 2 4 5 6 5 4 2 5 6 5 2 6 4 4 4 4 1 2 2 6 6 3 5 3 6 5
    5 1 5 6 1 2 1 5 4 1 6 1 5 3 1 2 6 5 3 1 4 1 2 1 4 4 1 4 6 1 5 6


mean(x)

[1] 3.57
```

- Average 3.57: Already relatively close to theoretical value of 3.5

- Do this 10 times:

```
3.39  3.43  3.55  3.5  3.48  3.61  3.46  3.64  3.28  3.41
```

- Averages are between 3.28 and 3.64 ($\approx \pm 0.2$ off expected value)
- Same for rolling die $n = 1000$:

```
3.475  3.49  3.435  3.437  3.407  3.479  3.567  3.474  3.498  3.565
```

- Averages are between 3.407 and 3.567 ($\approx \pm 0.1$ off EV)
- For $n = 1\,000\,000$: Averages rounded to 3 decimal places:

```
498  3.497  3.501  3.496  3.501  3.5  3.497  3.5  3.503  3.499
```

- Averages are between 3.496 and 3.503 ($\approx \pm 0.005$ off EV)
- Average of concrete numbers for ever larger n ever closer to 3.5

- Assumption in example: Fair die
- This is *not* realistic
- *No* real die is completely symmetric, i.e. not all numbers are equally probable
- Can try to construct die that is *very* fair and all rolled numbers occur with a probability of *almost* one sixth
- But to do this *exactly* is not possible

- Same applies for standard deviation:

  - Empirical standard deviation: s_X calculated from *concrete* data: From observations $x_1, \ldots, x_n$ according to formula above s_X
  - Standard deviation σ_X: Theoretical value resulting from model of distribution

- Hopefully: Empirical standard deviation s_X approaches theoretical value σ_X for more and more experiments, if data follow distribution of X

- If this is not the case: Something is wrong with model (e.g. all sides of die have *not* same probability to be rolled)