

## Index of 3-L

Page	Title
1	Practical information
2	Replication of simple experiments
3	Probability and randomness
4	Probability models I
5	Example: throwing 2 dice
6	Probability models II
7	General addition rules for probabilities
8	Exercise 4.21 & Birthday problem
9	Conditional probability
10	Bayes' formula/theorem
11	Example for Bayes' formula
12	Continuous theoretical distributions
13	Exercise 1.106
14	Discrete probability distributions
15	Changing the units: linear transformation
16	Parameters and random variables
17	Rules for means and variances of random variables
18	Summary notes / Overview of distributions

## PRACTICAL INFORMATION

### Today's lecture:

- probably the most abstract in course/books (hardly avoidable when dealing with probability) — the **topics**:
  - \* **randomness** — what is probability?<sup>1</sup>
  - \* **probability models** and rules,<sup>2</sup>
  - \* **discrete and continuous** probability distributions, and working with **random variables**,<sup>3</sup>
- **extra topics** (outside course syllabus):
  - \* Bayes theorem and its use for probability calculations,
  - \* integration formulas for calculus with continuous distributions,
  - \* formulas for means and variances of random variables.

### Other news:

- new virtual access to Minitab (as you saw in my attempted demo on Monday), see documentation at Moodle site,
- a short **video review** of the quiz for Session 2 is planned (→ Moodle).

---

<sup>1</sup> PSLS 4e: Chapter 9; IPS7e: Section 4.1; S: Chapter 4.

<sup>2</sup> PSLS 4e: Chapters 9+10; IPS7e: Sections 4.2-4.5; S: Chapter 4.

<sup>3</sup> PSLS 4e: Chapter 9; IPS7e: Sections 4.3-4.4; S: Chapters 5+6.

## REPLICATION OF SIMPLE EXPERIMENTS

Experience has shown **certain regularities** when an experiment is **repeated many times** (independently of each other):

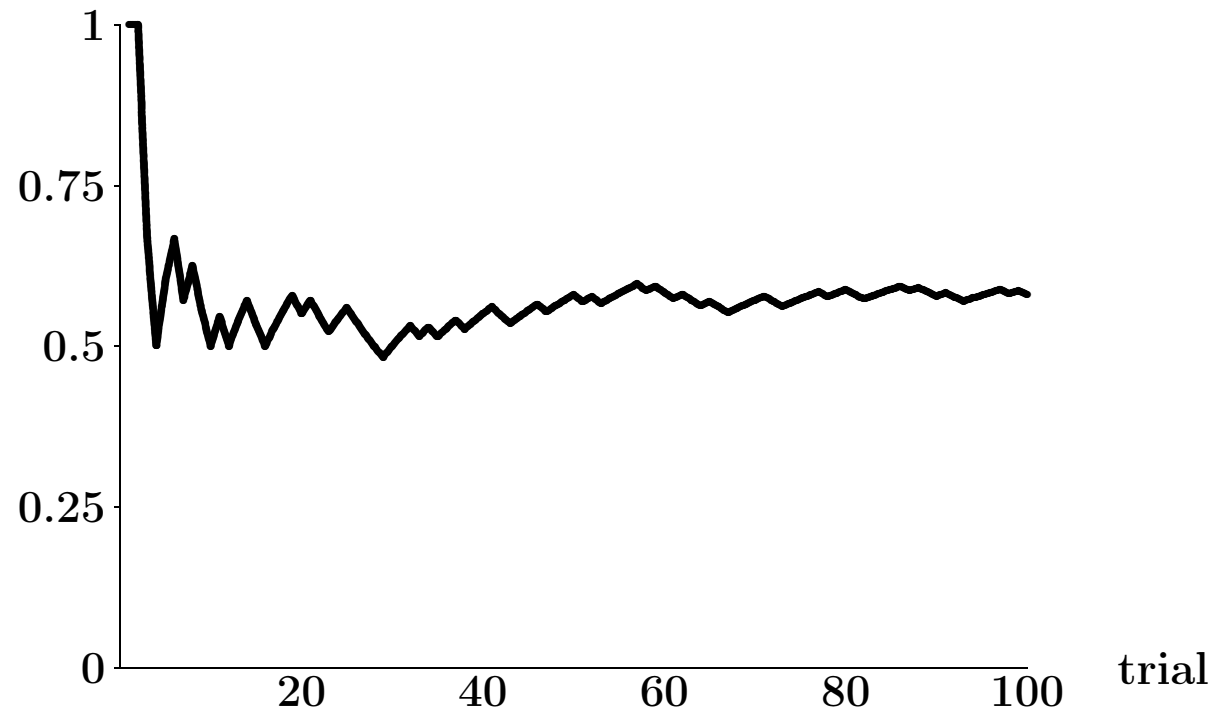
**100 thumbtack throws**

(1 = point down, 0 = point up):

11001	10100	10110	01110
10011	00001	11010	11011
10011	10111	01011	11010
01001	00111	10011	11011
00111	10100	10011	11010



**Figure:** proportion “point down” plotted against number of throws.



The proportion (relative frequency) of “point down” events seems to stabilize at 60%.

**Simulated data** (e.g., the Probability applet) show a similar behaviour.

## PROBABILITY AND RANDOMNESS

### Randomness / Variability:

- our inability to predict events, even in the presence of “all” information,
- arises from “nature”, the unknown factors affecting events,
- variability is generally large in biological sciences.

**What does it mean** ... that the probability of, say, “point down” is 60%?

This fundamental question statisticians do not agree about. . .

(a) according to the “**classical**” definition of probability:

in the long run, 60% of events are “point down”,

(b) according to the **Bayesian definition** of subjective probabilities:

I believe (myself) that the event occurs with probability 60% in one throw.<sup>4</sup>

“**Classical**” (often “frequentist”) approach:

- basis for this book and course,
- originates from experimental design and is (in my opinion) well suited for this field.

A **Bayesian approach** (not in course) may be more appropriate for complex events with no frequency interpretation, or when different sources of information are available.

<sup>4</sup> Subjective probabilities exist also for events with no frequency interpretation, e.g. “there’s at least one typo in the notes”, or “there will be a fall election”.

## PROBABILITY MODELS I

### Sample Space $S$ :

- = set of **the possible outcomes** of the “experiment” under study,
- recommended to use “natural” and simple values,
- **types of sample spaces** for a single variable:
  - \* **finite**, e.g.,  $\{0, 1\}$ ,  $\{\text{“up”}, \text{“down”}\}$ ,  $\{1, \dots, 6\}$ ,
  - \* **countable**, e.g. a subset of the integer numbers such as  $\{0, 1, 2, \dots\}$ ,<sup>5</sup>
  - \* **continuous**, e.g. intervals:  $(0, 1)$ ,  $(0, \infty)$  or  $(-\infty, \infty)$ .

### Event $A$ — a subset of the sample space,<sup>6</sup>

- some **standard sets/events**:
  - \*  $A = S$ : the full set (entire sample space),
  - \*  $A = \{x\}$ : a single value  $x$ ,
  - \*  $A$  or  $B$  (often written  $A \cup B$ ): union of  $A$  and  $B$ ,
  - \*  $A$  and  $B$  (often written  $A \cap B$ ): intersection of  $A$  and  $B$ .

As probability models are based on **set theory**, we can use **Venn diagrams** (next slide) for visualization.

---

<sup>5</sup> Probability models on such sample spaces are discussed (only) in PSLS and S.

<sup>6</sup> The eligible subsets must meet some mathematical condition we won't worry about.

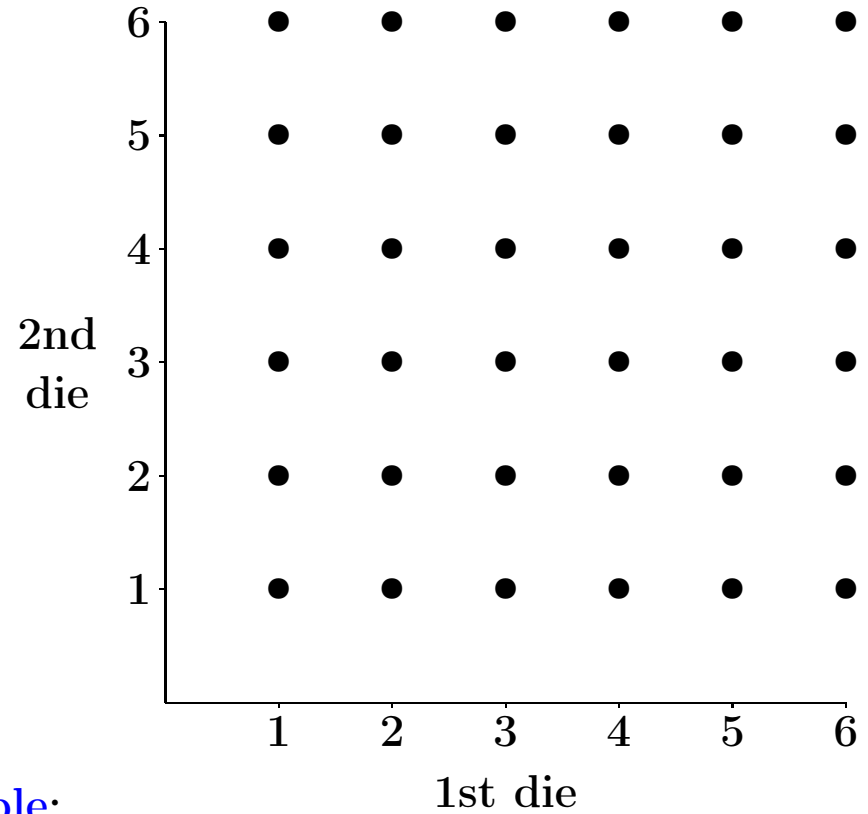
EXAMPLE: THROWING 2 DICE

**Sample space**

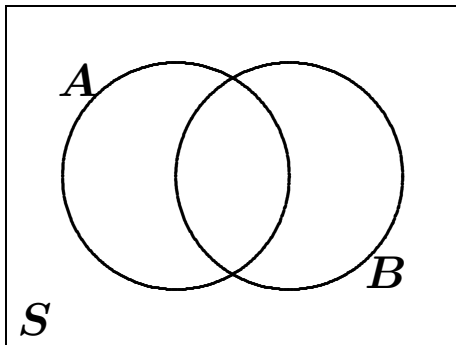
for throwing a pair of dice:

$$S = \{(1, 1), (1, 2), (1, 3), \dots \\ \dots, (6, 4), (6, 5), (6, 6)\}$$

(36 elements)



**Venn diagram:**



**Example:**

$$A = \text{(1st die shows 5)}$$

$$B = \text{(sum of dice } \leq 8)$$

$A$  and  $B = ?$

$A$  or  $B = ?$

## PROBABILITY MODELS II

A **probability distribution**  $P$  (or  $\Pr$ , my favorite notation) gives a value  $P(A)$  to any subset  $A$  (is mathematically, a function  $A \mapsto P(A)$ ) such that

- $0 \leq P(A) \leq 1$ , and  $P(S) = 1$ ,
- **addition rule**<sup>7</sup>: for two disjoint events  $A$  and  $B$  (i.e., their intersection ( $A$  and  $B$ ) is empty), it holds that

$$P(A \text{ or } B) = P(A) + P(B).$$

**Further rules and definitions:**

- **complement rule**: for any event  $A$ , the complement  $A^c$  is the event that  $A$  does not occur, and

$$P(A^c) = 1 - P(A)$$

- **definition**: two events  $A$  and  $B$  are said to be **independent**, if (the **multiplication rule** holds)

$$P(A \text{ and } B) = P(A)P(B).$$

Independence intuitively corresponds to unrelated events.

**Conversely**, the multiplication rule **only holds** for independent events.

---

<sup>7</sup> Mathematically, the addition rule must hold not only for two sets, but for “countably” many sets to properly define a probability distribution.

## GENERAL ADDITION RULES FOR PROBABILITIES

- **3-event addition rule:** if  $A$ ,  $B$  and  $C$  are pairwise disjoint events (i.e., all two-set intersections are empty), then it holds that

$$P(A \text{ or } B \text{ or } C) = P(A) + P(B) + P(C),$$

- **many event addition rule:** for any finite number of events  $A_1, \dots, A_n$  such that all pairs  $A_i$  and  $A_j$  are disjoint (their intersection is empty), it holds that

$$P(A_1 \text{ or } A_2 \text{ or } \dots \text{ or } A_n) = P(A_1) + P(A_2) + \dots + P(A_n).$$

- **non-disjoint events:** for any events  $A$  and  $B$ , it holds that

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B).$$

Intuitively, in the last formula we compensate for “counting” the event ( $A$  and  $B$ ) twice (both in  $A$  and in  $B$ ) by subtracting its probability.

**EXERCISE 4.21 & BIRTHDAY PROBLEM**

**Exercise 4.21**

For a randomly chosen acre of Canadian land,  $A$  and  $B$  are the events that the land is forest and pasture, respectively:  $P(A) = 0.35$ ,  $P(B) = 0.03$ .

- (a)  $P(\text{“not forested”}) = P(A^c) = 1 - P(A) = 1 - 0.35 = 0.65$ ,
- (b)  $P(\text{“either forest or pasture”}) = P(A \text{ or } B) = P(A) + P(B) = 0.38$ ,
- (c)  $P(\text{“other than forest or pasture”}) = P((A \text{ or } B)^c) = 1 - P(A \text{ or } B) = 1 - 0.38 = 0.62$ .

**The birthday problem:** How many unrelated people does it take to make the probability that at least two of them have birthday on the same day (of the year)  $> 50\%$ ?

**Solution of birthday problem:**

Assume  $n$  people present, and assume their birthdays independent of each other (plausible) and equally likely to happen on all days of the year, disregarding 29/2 (not quite correct). We will calculate  $p_n$  = the probability of all birthdays being on different days:

- For two persons ( $n=2$ ):  $p_2 = 364/365$ ,
- For  $n=3$ :  $p_3 = (364/365) \times (363/365)$ ,
- generally:  $p_n = (364/365) \times (363/365) \times \dots \times (365 - n + 1)/365$ .

$n$	2	3	5	10	13	20	21	22
$p_n$	0.997	0.992	0.960	0.859	0.806	0.556	0.524	0.493

## CONDITIONAL PROBABILITY

**Intuitively**, conditional probabilities

- are for situations where you already know that **some event** has occurred,
- give a probability distribution on a sample space consisting of those outcomes **in agreement with** the actually occurred event,
- for example, if the first die was 5, we *know* the probability of the two dice summing up to less than 5 (it is zero!), even without throwing the second.

The **conditional probability of  $B$  given  $A$**  is the probability of the part of  $B$  in agreement with  $A$ , that is,

$$P(B|A) = P(A \text{ and } B)/P(A).^8$$

Two important **implications**:

- general **multiplication rule** (very useful for calculus...):

$$P(A \text{ and } B) = P(A) P(B|A) = P(B) P(A|B),^9$$

- **more intuitive definition of independence**:

Two events  $A$  and  $B$  (with  $P(A) > 0$  and  $P(B) > 0$ ) are **independent** if

$$P(B) = P(B|A) \quad \text{and/or} \quad P(A) = P(A|B).$$

---

<sup>8</sup> The definition is valid (**only meaningful**) for events  $A$  with  $P(A) > 0$ .

<sup>9</sup> Dividing by  $P(B)$  gives the famous **Bayes' formula** (or rule, or theorem):  $P(A|B) = P(B|A)P(A)/P(B)$ , (next page).

## BAYES' FORMULA / THEOREM

**Formula/Theorem:** For events  $A$  and  $B$  with  $P(A), P(B) > 0$ , it holds that

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)},$$

or, for a collection of disjoint events  $A_1, \dots, A_n$  with  $P(A_i) > 0$  and  $\sum_i P(A_i) = 1$ ,

$$P(A_i|B) = \frac{P(B|A_i) \cdot P(A_i)}{P(B|A_1) \cdot P(A_1) + \dots + P(B|A_n) \cdot P(A_n)}.$$

### Comments:

- (technical) other versions exist of the formula for density and probability functions,
- **loosely stated**, Bayes' formula is for situations where you know some conditional probabilities but want them “the other way round”, the most prominent example being **interpretations of diagnostic tests**:
  - \* consider testing for presence of disease ( $D+$ ) in humans/animals,
  - \* a test result ( $T$ ) comes back, and the researcher/doctor/patient wants to know what it says about the subject's disease status:
    - if the test is **perfect**, the test result gives the answer,
    - tests are generally not perfect — they can give both **false positive** and **false negative** results; we want to know the **conditional probability**  $P(D+ | T+)$ .

## EXAMPLE FOR BAYES' FORMULA

How good are doctors in communicating test results?:<sup>10</sup>

- a number of physicians were asked to communicate a positive test result for an individual being tested for colorectal cancer by a particular test, with the properties

$$\begin{aligned}P(T+ | D+) &= \text{sensitivity (Se)} = (1 - \text{false negative rate}) = 0.5, \\P(T- | D-) &= \text{specificity (Sp)} = (1 - \text{false positive rate}) = 0.97,\end{aligned}$$

- prevalence/disease probability prior to testing (in the population):  $P(D+) = 0.003$ ,
- now Bayes' formula gives:

$$\begin{aligned}P(D+ | T+) &= \frac{P(T+ | D+) P(D+)}{P(T+ | D+) P(D+) + P(T+ | D-) P(D-)} \\&= [0.5 \cdot 0.003] / [0.5 \cdot 0.003 + 0.03 \cdot 0.997] = 0.048\end{aligned}$$

— an almost 16-fold increase in probability, but very far from certainty and from the Se (the physicians' answers were all over the place, and many of them close to 0.5),

- if instead  $P(D+) = 0.4$  ( $\sim$  common disease/exposed subject), we would get:

$$P(D+ | T+) = [0.5 \cdot 0.4] / [0.5 \cdot 0.4 + 0.03 \cdot 0.6] = 0.917$$

— likely to be diseased.

<sup>10</sup> Based on Gigerenzer and Edwards (2003), *BMJ* 327, 741–744; see VHM 801 media page.

## CONTINUOUS THEORETICAL DISTRIBUTIONS

A **continuous distribution** for a single variable is given by a **density curve**<sup>11</sup> — a function  $f(x)$  such that

$$f(x) \geq 0 \text{ everywhere, and } \int_S f(x)dx = 1.$$

**Probabilities** in the distribution are calculated as **areas** under the curve, e.g. for intervals  $(-\infty, a)$  and  $(a, b)$ :

$$P(-\infty, a) = \int_{-\infty}^a f(x)dx, \quad P(a, b) = \int_a^b f(x)dx.$$

We also define the distribution's **mean**, **variance** and **standard deviation** by:

- **mean**  $\mu = \int_S x f(x)dx$ ,
- **variance**  $\sigma^2 = \int_S (x - \mu)^2 f(x)dx$ , and **standard deviation**  $\sigma = \sqrt{\sigma^2}$ .

**Comments:**

- $f(x) \sim$  the likelihood of observations around  $x$  (counter-intuitively, the probability of  $x$  is zero),
- histogram and density curve matched by having area 1.

**Why theoretical distributions?** (besides being “easier to work with”)

- lead to a **separation of systematic and random parts** of our data:
  - \* distribution  $\sim$  systematic features (repeatable in a similar situation, and therefore of primary interest),
  - \* variation (or variance) in the distribution  $\sim$  random features of our data (non-repeatable).

---

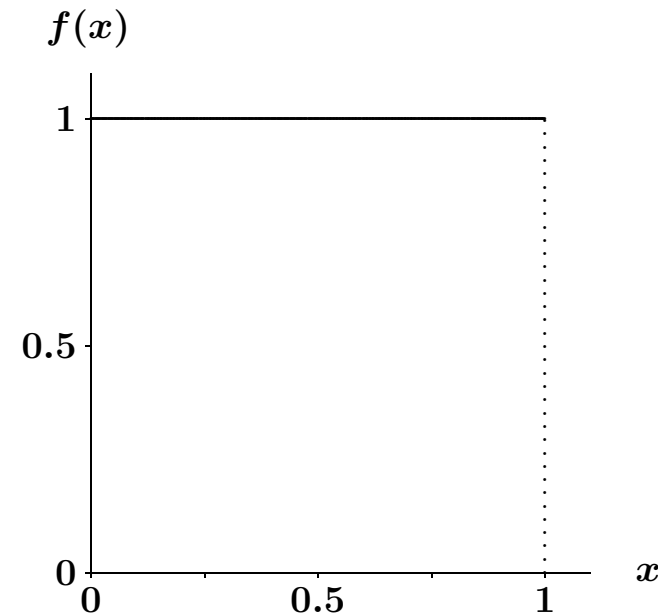
<sup>11</sup> Also termed: density function, probability density, or just density.

**EXERCISE 1.106**

The uniform distribution on the unit interval  $(0,1)$  has the density curve  $f(x)$  as shown:

Answers:

- (a) a square with side  $a$  has area  $a^2$ ;  
here,  $a = 1$ ,
- (b) 0.25 (computed as  $(1 - 0.75) \cdot 1$ ),
- (c) 0.5 (computed as  $(0.75 - 0.25) \cdot 1$ ).



Extra questions: Determine also (d) the median, (e) the quartiles  $Q_1$  and  $Q_3$ , and (f) the mean.

- (d) 0.5 (by splitting the area 0.5 : 0.5),
- (e)  $Q_1 = 0.25$  and  $Q_3 = 0.75$ ,
- (f) 0.5 (in a symmetrical distribution, the mean and median are the same).

## DISCRETE PROBABILITY DISTRIBUTIONS

A **discrete** probability distribution:

- is a probability distribution on a **discrete sample space**,
- has a **probability function**  $p(x)$  (for  $x$  in  $S$ ), so that for any event  $A$ :

$$P(A) = \sum_{x \text{ in } A} p(x),$$

—  $p(x)$  is interpreted as the probability of the set  $\{x\}$ .

We can create probability distributions by specifying  $p(x)$  such that  $0 \leq p(x) \leq 1$  and  $\sum_x p(x) = 1$ . **Simplest example**: uniform distribution on finite  $S$  ( $N$  elements):  $p(x) = 1/N$ .<sup>12</sup>

For a **discrete, quantitative** (theoretical) distribution with probability function  $p(x)$ , we define the

- **mean**  $\mu = \sum_x x p(x)$ ,
- **variance**  $\sigma^2 = \sum_x (x - \mu)^2 p(x)$ , and **standard deviation**  $\sigma = \sqrt{\sigma^2}$ .

The definitions are very similar to those for observed data and continuous distributions.

An **observed** (or empirical) distribution of observations  $x_1, \dots, x_n$  can be thought of as a **discrete** distribution with probability function ( $\sim$  all data points having **equal weight**):

$$p_e(x) = (\text{no. of } x'_i\text{'s} = x) / n, \quad \text{for } x \text{ in } \{x_1, \dots, x_n\}.$$

<sup>12</sup> In the “Throwing 2 dice” example,  $S = \{(1, 1), \dots, (6, 6)\}$  and  $N = 36$ .

## CHANGING THE UNITS: LINEAR TRANSFORMATION

**Linear transformation:**  $x \mapsto a + bx$ .

**Example:** conversion Fahrenheit/Celsius

- $x_C, x_F$  = temperature measured in °C and °F, respectively,
- conversion formulae — linear transformations:

$$x_F = 32 + (9/5)x_C \quad \text{and} \quad x_C = (-160/9) + (5/9)x_F,$$

- **question:** how to translate measures of center and spread?

**Effect of linear transformation on center and spread:**

- **scaling** ( $a=0$ ) with  $b$  (i.e.,  $x \mapsto bx$ ): **center**  $\mapsto b \cdot$  center; **spread**  $\mapsto b \cdot$  spread,
- **translation** ( $b=1$ ) with  $a$  (i.e.,  $x \mapsto a+x$ ): **center**  $\mapsto a +$  center; **spread** unchanged,
- linear transformation  $a + bx$ : **center**  $\mapsto a + b \cdot$  center; **spread**  $\mapsto b \cdot$  spread,
- formulae apply to **all statistics for center and spread**.

**Example** for temperature scales:

- say we have  $\bar{x}_F = 95^\circ\text{F}$  and  $s_F = 9^\circ\text{F}$  for some data,
- measured in °C we would then have:

$$\bar{x}_C = (-160/9) + (5/9)\bar{x}_F = 35^\circ\text{C}, \quad \text{and} \quad s_C = (5/9)s_F = 5^\circ\text{C}.$$

## PARAMETERS AND RANDOM VARIABLES

**Parameters** = **unknown constants** associated with theoretical distributions, to allow us to adapt them to real data,

- e.g., the **mean** ( $\mu$ ) and the **standard deviation** ( $\sigma$ ),<sup>13</sup>
- **unknown**, because (i) we don't know them (exactly), but (ii) we aim to get as close to their true value as possible,
- denoted by **Greek letters** to distinguish them from statistics calculated from data.

**Random variables** = notation to work with distributions, using **capital, Latin letters**; some typical examples:

- $P(X > 0)$  to denote the prob. that an observation from the distribution (of  $X$ ) is  $> 0$ ,
- $E(X)$  or  $EX$  to denote the mean (expectation) in the distribution of  $X$ ,
- random variables can be manipulated just as data values; we can e.g. compute
  - \*  $X+1$  and  $X-0.5$  (both a **translation**),
  - \*  $2X$  and  $X/100$  (both a **scaling**),
  - \*  $Y = 32 + 1.8X$  and  $Z = (X - \mu)/\sigma$  (both a **linear transformation**, the latter also a **standardization**),

and we can ask about (and maybe determine) the distribution of a new variable.

---

<sup>13</sup> In the next lecture, we will meet the normal distribution  $N(\mu, \sigma^2)$  and the binomial distribution  $B(n, p)$ .

**RULES FOR MEANS AND VARIANCES OF RANDOM VARIABLES**

**New notation:**

- $\mu_X = \mathbf{E}X =$  mean (expectation) of  $X$ ,
- $\sigma_X^2 = \mathbf{Var}X =$  variance of  $X$ , and  $\sigma_X = \mathbf{sd}X =$  standard deviation of  $X$ .

**Rules for one variable** — same rules (as 3L–15), new “disguise”:

Let  $X$  be a random variable, and let  $Y = a + bX$ , where  $a, b$  are numbers; then:

- $\mu_Y = \mu_{a+bX} = a + b\mu_X$ , or  $\mathbf{E}Y = \mathbf{E}(a + bX) = a + b\mathbf{E}X$ ,
- $\sigma_Y = \sigma_{a+bX} = |b|\sigma_X$ , or  $\mathbf{sd}(Y) = \mathbf{sd}(a + bX) = |b|\mathbf{sd}X$ .

**Rules for two variables** — new rules, for random variables  $X$  and  $Y$  it holds that:<sup>14</sup>

- $\mu_{X+Y} = \mu_X + \mu_Y$  or  $\mathbf{E}(X + Y) = \mathbf{E}X + \mathbf{E}Y$ ,  
 $\mu_{X-Y} = \mu_X - \mu_Y$  or  $\mathbf{E}(X - Y) = \mathbf{E}X - \mathbf{E}Y$ ,
- if  $X$  and  $Y$  are **independent** (i.e., all pairs of events involving  $X$  and  $Y$  are independent),
 

$\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2$	or	$\mathbf{Var}(X + Y) = \mathbf{Var}(X) + \mathbf{Var}Y$ ,
$\sigma_{X-Y}^2 = \sigma_X^2 + \sigma_Y^2$	or	$\mathbf{Var}(X - Y) = \mathbf{Var}(X) + \mathbf{Var}Y$ ,
- $\mathbf{sd}(X + Y) = \mathbf{sd}(X - Y) = \sqrt{\mathbf{sd}(X)^2 + \mathbf{sd}(Y)^2}$ ,
- if  $X$  and  $Y$  are **dependent**, their **correlation** enters into  $\mathbf{Var}(X+Y)$  and  $\mathbf{Var}(X-Y)$  (in addition to the variances/standard deviations).

---

<sup>14</sup> “Addition rules” for means and variances, respectively.

## SUMMARY NOTES / OVERVIEW OF DISTRIBUTIONS

### Key words and concepts:

- probability: sample space, event, probability distribution,
- addition rule, independence / multiplication rule,
- parameter of a probability distribution (often denoted by a Greek letter, e.g.  $\mu$ ),
- random variable (usually, in this course, denoted by a capital letter, e.g.  $X$ ); linear transformation; mean, variance, stand. dev. of distribution and random variable.

### Summary of concepts

for distributions:

Concept	Distribution		
type values	observed discrete	theoretical continuous	theoretical discrete
given by	actual data $x_1, \dots, x_n$	density curve $f(x)$	probability function $p(x)$
typical value	$x_i$	$X$	$X$
prob. of $\{x\}$	(no. of $x_i$ 's = $x$ )/ $n$	0	$p(x)$
prob. of $A$	(no. of $x_i$ 's in $A$ )/ $n$	$\int_{x \text{ in } A} f(x) dx$	$\sum_{x \text{ in } A} p(x)$
mean	$\bar{x} = \frac{1}{n} \sum x_i$	$\mu = \int x f(x) dx$	$\mu = \sum x p(x)$
variance	$s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$	$\sigma^2 = \int (x - \mu)^2 f(x) dx$	$\sigma^2 = \sum (x - \mu)^2 p(x)$
stand. dev.	$s = \sqrt{s^2}$	$\sigma = \sqrt{\sigma^2}$	$\sigma = \sqrt{\sigma^2}$
median	“mid-observation”	point $x$ where $P(X < x) = 0.5$	
examples	“descriptive stats”	normal & uniform	binomial