Index of 3-L

Practical Information

Today's lecture:

- probably the most abstract in course/books (hardly avoidable when dealing with probability) — the topics:
  - randomness — what is probability?[1]
  - probability models and rules,[2]
  - discrete and continuous probability distributions, and working with random variables,[3]
- 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:

- Monday's lab session rescheduled to Wednesday 9am-12pm,
- Moodle quiz for Sessions $1-2$ accessible (optional for you to try),
- do we want to schedule lab review sessions? and when? (Tuesdays/Thursdays 1-2pm)

[1] PSLS 4e: Chapter 9; IPS7e: Section 4.1; S: Chapter 4.
[2] PSLS 4e: Chapters 9+10; IPS7e: Sections 4.2-4.5; S: Chapter 4.
[3] PSLS 4e: Chapter 9; IPS7e: Sections 4.3-4.4; S: Chapters 5+6.

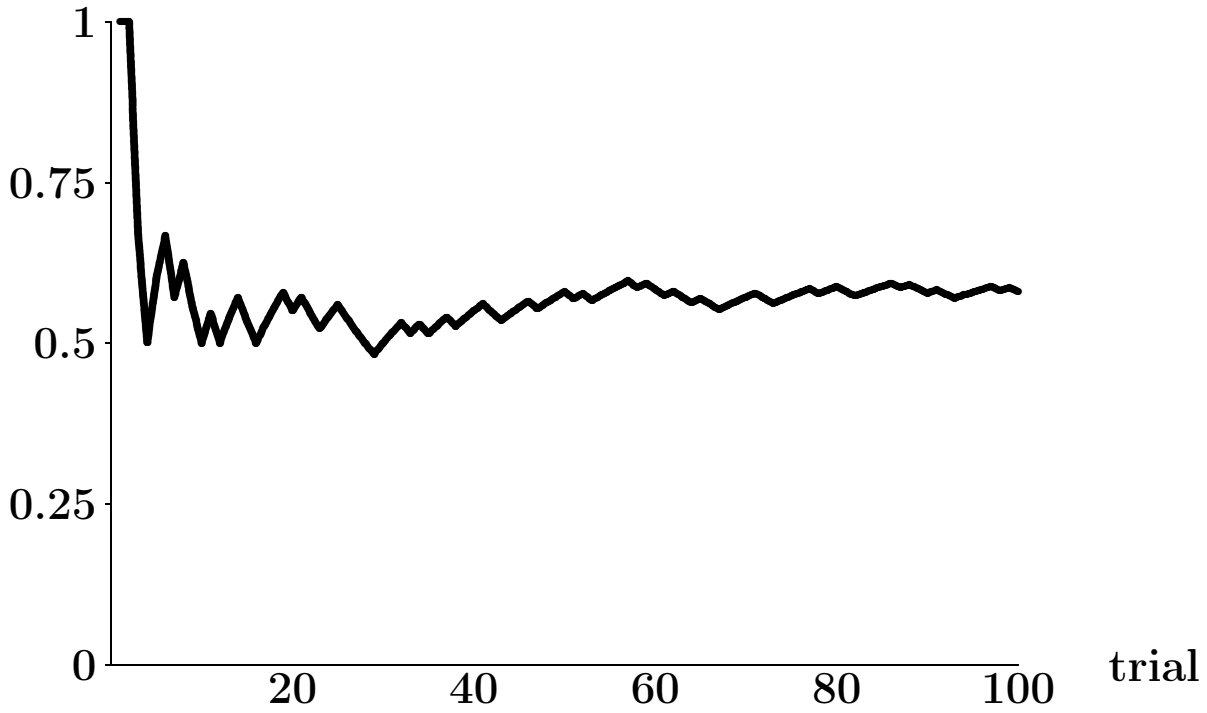
Replication of Simple Experiments

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.[4]

"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.

[4] Subjective probabilities exist also for events with no frequency interpretation, e.g. "There's at least one typo in the notes", or "The next person to set foot on the moon will be from the U.S".

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., $\{0, 1\}$, {"up","down"}, $\{1, \ldots, 6\}$,
  - countable, e.g. a subset of the integer numbers such as $\{0, 1, 2, \ldots\}$,[5]
  - continuous, e.g. intervals: $(0, 1)$, $(0, \infty)$ or $(-\infty, \infty)$.

Event A — a subset of the sample space,[6]

- 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.

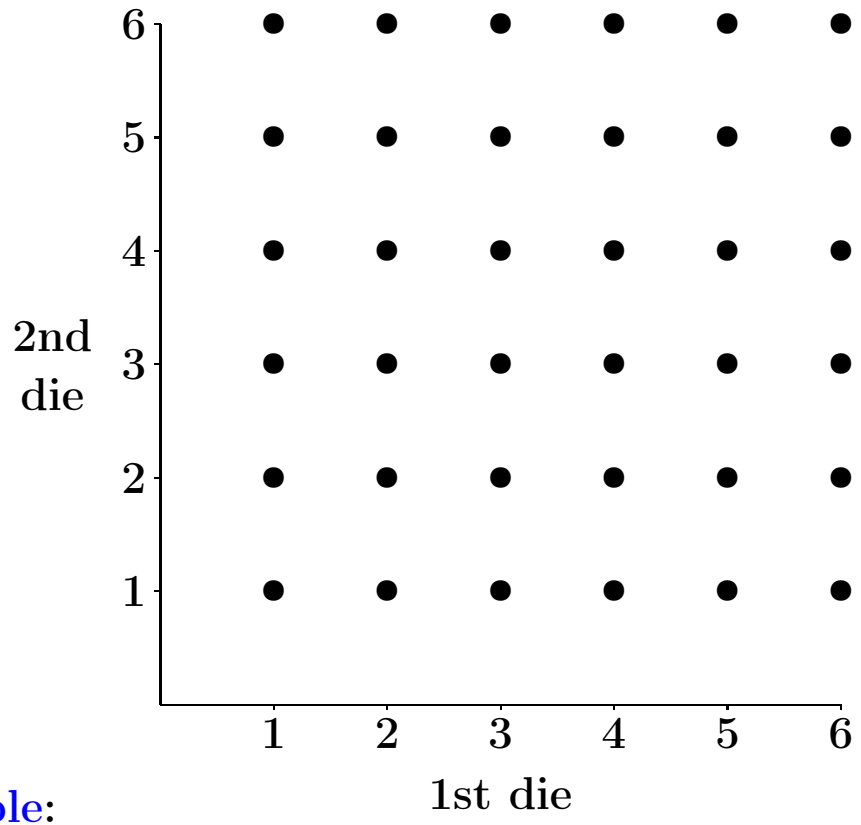
[5] Probability models on such sample spaces are discussed (only) in PSLS and S.
[6] The eligible subsets must meet some mathematical condition we won't worry about.

Example: Throwing 2 Dice

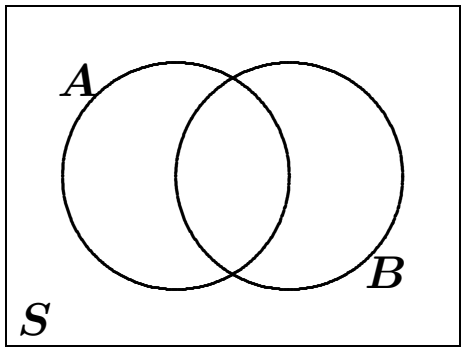
Sample space
for throwing a pair of dice:

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

(36 elements)



Venn diagram:



Example:

A = (1st die shows 5)

B = (sum of dice ≤ 8)

A and B = ?

A or B = ?

Probability Models II

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

- $0 \leq \mathrm{P}(A) \leq 1$, and $\mathrm{P}(S) = 1$,
- addition rule[7]: for two disjoint events A and B (i.e., their intersection (A and B) is empty), it holds that

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

Further rules and definitions:

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

$$\mathrm{P}(A^c) = 1 - \mathrm{P}(A)$$

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

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

  Independence intuitively corresponds to unrelated events.
  Conversely, the multiplication rule only holds for independent events.

[7] 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

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

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

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

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

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

  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: $\mathrm{P}(A) = 0.35$, $\mathrm{P}(B) = 0.03$.

(a) P("not forested") $= \mathrm{P}(A^c) = 1 - \mathrm{P}(A) = 1 - 0.35 = 0.65$,

(b) P("either forest or pasture") $= \mathrm{P}(A \text{ or } B) = \mathrm{P}(A) + \mathrm{P}(B) = 0.38$,

(c) P("other than forest or pasture") $= \mathrm{P}((A \text{ or } B)^c) = 1 - \mathrm{P}(A \text{ or } B) = 1 - 0.38 = 0.62$.

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: $p_n = (364/365) \times (363/365) \times \ldots \times (365-n+1)/365$.

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 ocurred,
- 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,

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

Two important implications:

- general multiplication rule (very useful for calculus...):
  $$\mathrm{P}(A \text{ and } B) = \mathrm{P}(A)\,\mathrm{P}(B|A) = \mathrm{P}(B)\,\mathrm{P}(A|B),\ ^{9}$$
- more intuitive definition of independence:
  Two events A and B (with $\mathrm{P}(A) > 0$ and $\mathrm{P}(B) > 0$) are independent if
  $$\mathrm{P}(B) = \mathrm{P}(B|A) \quad \text{and/or} \quad \mathrm{P}(A) = \mathrm{P}(A|B).$$

[8] The definition is valid (only meaningful) for events A with $\mathrm{P}(A) > 0$.
[9] Dividing by $\mathrm{P}(B)$ gives the famous Bayes' formula (or rule, or theorem): $\mathrm{P}(A|B) = \mathrm{P}(B|A)\mathrm{P}(A)/\mathrm{P}(B)$, (next page).

Bayes' Formula / Theorem

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

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

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

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

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 $\mathrm{P}(D+|T+)$.

Example for Bayes' Formula

How good are doctors in communicating test results?:[10]

- 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

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

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

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

  — 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 $\mathrm{P}(D+) = 0.4$ ($\sim$ common disease/exposed subject), we would get:
  $\mathrm{P}(D+\,|T+) = [0.5 \cdot 0.4] \,/\, [0.5 \cdot 0.4 + 0.03 \cdot 0.6] = 0.917$
  — likely to be diseased.

[10] 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[11] — a function $f(x)$ such that

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

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.

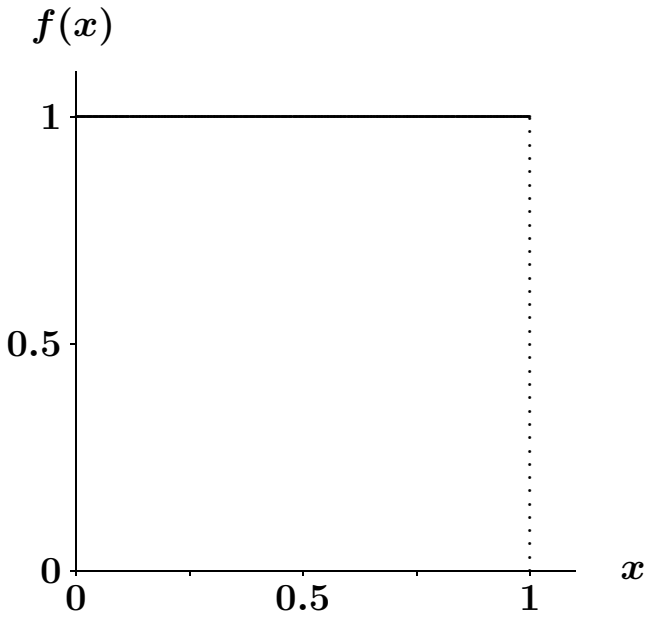
Why theoretical distributions? (besides being "easier to work with")

- 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).

[11] 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:

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

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

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

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 $x_1, \ldots, x_n$ can be thought of as a discrete distribution with probability function ($\sim$ all data points having equal weight):

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

[12] In the "Throwing 2 dice" example, $S = \{(1,1), \ldots, (6,6)\}$ 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 $\bar{x}_F = 95$°F and $s_F = 9$°F for some data,
- measured in °C we would then have:

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

Parameters and Random Variables

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

- e.g., the mean (μ) and the standard deviation (σ),[13]
- 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,
- $\mathrm{E}(X)$ or $\mathrm{E}X$ 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.

[13] In the next lecture, we will meet the normal distribution $\mathrm{N}(\mu, \sigma)$ and the binomial distribution $\mathrm{B}(n, p)$.

Rules for Means and Variances of Random Variables

New notation:

- $\mu_X = \mathrm{E}X =$ mean (expectation) of X,
- $\sigma_X^2 = \mathrm{Var}X =$ variance of X, and $\sigma_X = \mathrm{sd}X =$ 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 $\mathrm{E}Y = \mathrm{E}(a+bX) = a + b\,\mathrm{E}X$,
- $\sigma_Y = \sigma_{a+bX} = |b|\sigma_X$, or $\mathrm{sd}(Y) = \mathrm{sd}(a+bX) = |b|\,\mathrm{sd}X$.

Rules for two variables — new rules, for random variables X and Y it holds that:[14]

- $\mu_{X+Y} = \mu_X + \mu_Y$ or $\mathrm{E}(X+Y) = \mathrm{E}X + \mathrm{E}Y$,
  $\mu_{X-Y} = \mu_X - \mu_Y$ or $\mathrm{E}(X-Y) = \mathrm{E}X - \mathrm{E}Y$,
- if X and Y are independent (i.e., all pairs of events involving X and Y are independent),
  $\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y$ or $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}Y$,
  $\sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y$ or $\mathrm{Var}(X-Y) = \mathrm{Var}(X) + \mathrm{Var}Y$,
- $\mathrm{sd}(X+Y) = \mathrm{sd}(X-Y) = \sqrt{\mathrm{sd}(X)^2 + \mathrm{sd}(Y)^2}$,
- if X and Y are dependent, their correlation enters into $\mathrm{Var}(X+Y)$ and $\mathrm{Var}(X-Y)$ (in addition to the variances/standard deviations).

[14] "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. μ),
- 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 $x_1, \ldots, x_n$	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 $\mathrm{P}(X < x) = 0.5$	
examples	"descriptive stats"	normal & uniform	binomial