Biostats VHM 801 Course
Fall 2019, Atlantic Veterinary College, PEI
Henrik Stryhn

Index of 3-L

Practical Information

Today's lecture:

- probably the most abstract in course/books (hardly avoidable when dealing with probability),
- topics:
  - randomness — what is probability?[1]
  - probability models and rules,[2]
  - discrete and continuous probability distributions; working with random variables,[3]
- extra topics (outside course syllabus):
  - Bayes theorem and its use for prob. calculations,
  - integration formulae for continuous distributions,
  - formulae for means & variances of random variables.

Other news:

- final exam on December 9 (Monday) — objections?
- do we want to schedule lab review sessions? and if yes, when? (Tuesdays 1-2pm or Thursdays 1-2pm),
- new course feature: Wiki (on Moodle) on course topics:
  - intention is to stimulate *reflection* (and questions),
  - participation is voluntary, but without your input this will be going nowhere. . .

[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

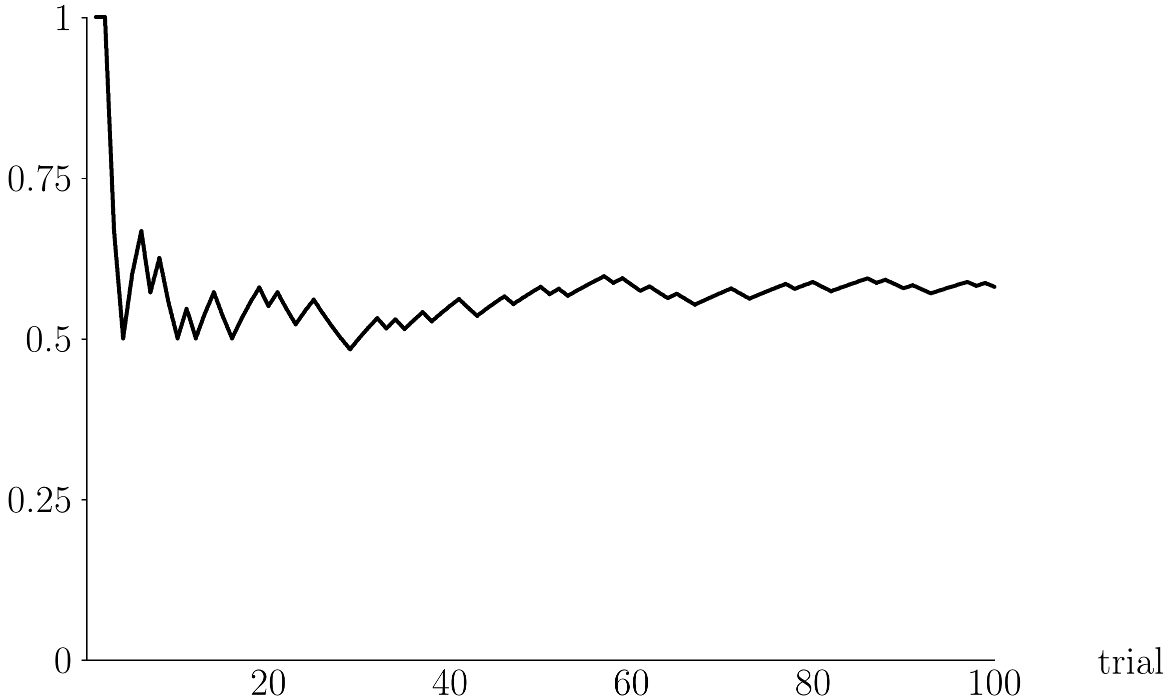


Figure: proportion "point down" plotted against number of throws.

The proportion (relative frequency) of "point down" events seems to stabilize around 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", 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" ("frequentist") approach:

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

Bayesian approach may be more appropriate for complex events with no frequency interpretation, or when different sources of information are available (more later in course).

[4] Subjective probabilities exist also for events with no frequency interpretation, e.g. "there's at least one typo in the notes", or "Canada will have a new PM".

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 like $\{0, 1, 2, \ldots\}$,[5]
  - continuous, e.g. intervals: $(0, 1)$, $(0, \infty)$ or $(-\infty, \infty)$.

Event A:

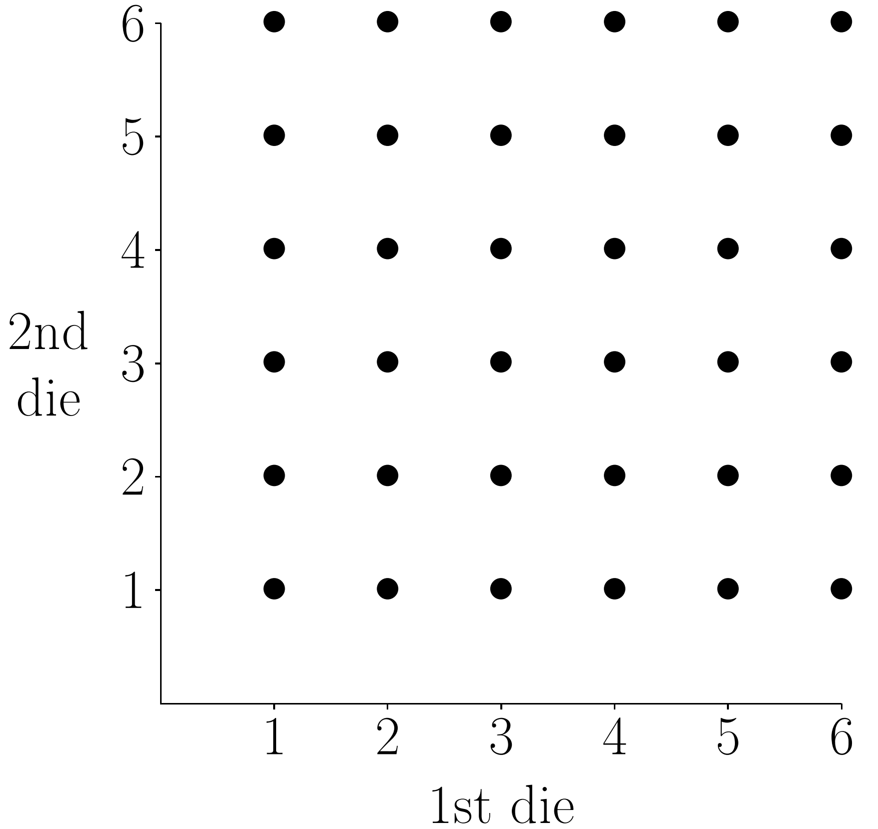
- = subset of the sample space (with some mathematical conditions we won't worry about),
- some standard sets/events:
  - $A = S$: the full set (entire sample space),
  - $A = \{x\}$: a single value x,
  - A or B (often written $A \cup B$): union of A and B,
  - A and B (often written $A \cap B$): intersection 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.

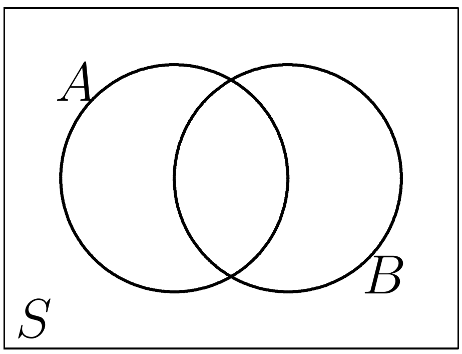
Example: Throwing 2 Dice

Sample space for throwing a pair of dice:



$S = \{(1,1),(1,2),\ldots,(6,5),(6,6)\}$ (36 elements)

Venn diagram:



Example:

$A =$ (1st die shows 5)

$B =$ (sum of dice ≤ 8)

A and $B =$?

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[6]: for two disjoint events A and B (that is, 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.

[6] 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 (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 all 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 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("not forested") = $P(A^c) = 1 - P(A) = 1 - 0.35 = 0.65$,

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

(c) P("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) larger than 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,

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

Definition is only meaningful for events A with $\mathrm{P}(A) > 0$.

Two important implications:

- general multiplication rule (very useful for calculus...): $\mathrm{P}(A \text{ and } B) = \mathrm{P}(A)\mathrm{P}(B|A) = \mathrm{P}(B)\mathrm{P}(A|B)$, [7]
- more intuitive definition of independence: Two events A and B (with $\mathrm{P}(A) > 0$ and $\mathrm{P}(B) > 0$) are *independent* if

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

[7] dividing by $\mathrm{P}(B)$ gives the famous *Bayes' formula* (or rule, or theorem): $\mathrm{P}(A|B) = \mathrm{P}(B|A)\mathrm{P}(A)/\mathrm{P}(B)$, see next page.

Bayes' Formula / Theorem

For events A and B with $P(A), P(B) > 0$ it holds that

$$P(A|B) = P(B|A) \cdot P(A)/P(B),$$

or, for disjoint events $A_1, \ldots, A_n$ with $P(A_i) > 0$ and $\sum_i P(A_i) = 1$:

$$P(A_i|B) = [P(B|A_i) \cdot P(A_i)]/[P(B|A_1) \cdot P(A_1) + \ldots + P(B|A_n) \cdot P(A_n)].$$

Loosely stated, Bayes' formula is for situations where you know some conditional probabilities but want them "the other way round", e.g.

- consider testing for presence of disease ($D+$) in humans/animals,
- the test (T) is imperfect but we know (have estimates for)

$$\begin{aligned} P(T+|D+) &= \text{sensitivity (Se)} = 0.50\,, \\ P(T-|D-) &= \text{specificity (Sp)} = 0.97\,, \end{aligned}$$

- if a subject tests positive ($T+$), we want to know the probability of disease $P(D+|T+)$ (positive predictive value),
- we need to know the prevalence/disease probability prior to testing ($P(D+)$; note that $P(D+) + P(D-) = 1$); then,

$$P(D+|T+) = \frac{P(T+|D+)\,P(D+)}{P(T+|D+)\,P(D+) + P(T+|D-)\,P(D-)}$$

- scenario 1: $P(D+) = 0.003$ (rare disease/unexposed subject),

$$P(D+|T+) = \frac{0.5 \cdot 0.003}{0.5 \cdot 0.003 + 0.03 \cdot 0.997} = 0.048$$

  — an almost 16-fold increase in probability, but very far from certainty and from the Se ... (*BMJ* **327**, 741-744),

- scenario 2: $P(D+) = 0.4$ (common disease/exposed subject),

$$P(D+|T+) = \frac{0.5 \cdot 0.4}{0.5 \cdot 0.4 + 0.03 \cdot 0.6} = 0.917$$

  — likely to be diseased.

Continuous Theoretical Distributions

A continuous distribution for a single variable is given by a density curve[8] — 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 stand. dev. by:

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

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.

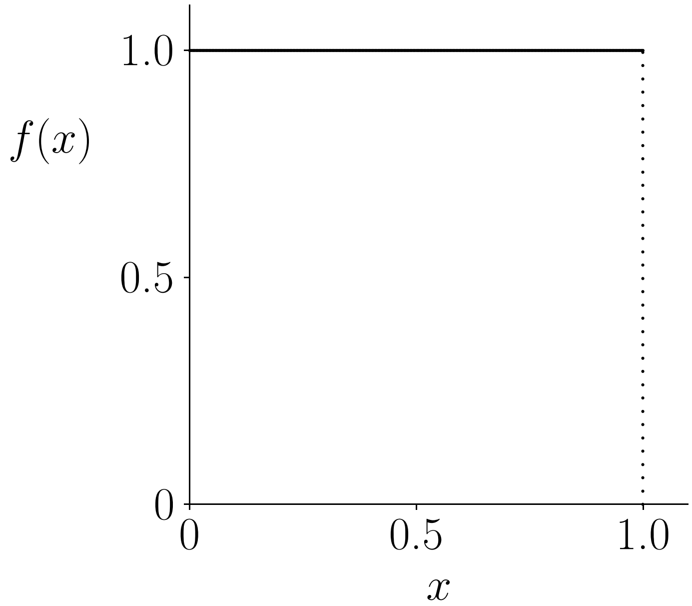
Why theoretical distributions?

- "easier to work with",
- 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).

[8] 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:



(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 define 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 with N points: $p(x) = 1/N$.[9]

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

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

These 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 prob. function

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

That is, all data points have equal weight.

[9] 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, resp.,
- conversion formulae — linear transformations:

  $x_F = 32 + (9/5)x_C$ and $x_C = (-160/9) + (5/9)x_F$,
- question: how to translate measures of center and spread?

Effect of linear transformation on center and spread:

- scaling ($a = 0$) with b — effectively: $x \mapsto bx$,
  center $\mapsto b \cdot$ center; spread $\mapsto b \cdot$ spread,
- translation ($b = 1$) with a — effectively: $x \mapsto a + x$,
  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:

- assume $\bar{x}_F = 95$°F and $s_F = 9$°F for some data,
- measured in °C we would have

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

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); in the next lecture we will meet the normal distribution $\mathrm{N}(\mu, \sigma^2)$ and the binomial distribution $\mathrm{B}(n, p)$,
- unknown, because (i) we don't know them (exactly), (ii) we aim to get as close to their true value as possible,
- denoted by Greek letters to distinguish them from the quantities we calculate from data.

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

- $P(X > 0)$ to denote the probability that an observation from the distribution (of X) is > 0,
- $\mathrm{E}(X)$ or $\mathrm{E}X$ to denote the mean (expectation) of the distribution for 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 determine the distribution of a new variable.

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,
- $\sigma_X = \mathrm{sd}X =$ standard deviation of X.

Rules for one variable — same old rules, new "disguise":
Let X be a random variable, and let $Y = a + bX$, where a, b are numbers. Then it holds that,

- $\mu_Y = \mu_{a+bX} = a + b\mu_X$, 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:
Let X and Y be random variables. Then it holds that,[10]

- $\mu_{X+Y} = \mu_X + \mu_Y$ 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, respectively, are independent),[10]
  $\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2$ or $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}Y$,
  $\sigma_{X-Y}^2 = \sigma_X^2 + \sigma_Y^2$ 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 $\mathrm{Var}(X-Y)$ (and the stand. devs.).

[10] "Addition rules" for means and variances, respectively.

Summary Notes / Overview of Distributions

Key words and concepts:

- probability: sample space, event, probab. distrib.,
- addition rule, independence / multiplication rule,
- continuous, theoretical distribution: density curve,
- discrete, theoretical distribution: probability function,
- parameter of a probability distribution,[11]
- random variable[12]; linear transformation; mean, variance, stand. dev. of distrib. and random var.

Brief 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 $f(x)$	prob. 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	$\bar{x} = \frac{1}{n}\sum x_i$	$\mu = \int xf(x)dx$	$\mu = \sum xp(x)$
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

[11] Parameters are often denoted by a Greek letter, e.g. μ.
[12] Random variables are usually denoted by capital letters, e.g. X.