1 Basic Statistics

Das ist eine reduzierte Version des Kapitels 1 aus **B. Krenn, Chr. Samuelsson: The Linguist's Guide to Statistics. 1997.** *Das Exzerpt deckt die Inhalte der Vorlesungseinheit "Einführung in die Wahrscheinlichkeitstheorie" ab.*

1.1 The Stability of the Relative Frequency

Although an elegant theory in its own right, the main reason that statistics has grown as popular as it has is something that is known as "The stability of the relative frequency". This is the empirical observation that there is some structure also in random processes. If we for example flip a coin a large number of times, we will note that in approximately half the cases the outcome is "heads" and in approximately half the cases it is "tails". If we flip a coin a small number of times, say only once, twice or three times, this is not consistently the case.

The proportion of times a certain outcome occurs is called the *relative frequency* of the outcome. If n_u is the number of times the outcome u occurs in n trials, then $\frac{n_u}{n}$ is the relative frequency of u. The relative frequency is often denoted f_n.

Empirically, there seems to be some number which the relative frequency stabilizes around after a large number of trials. A fundamental assumption in statistics is that such numbers exist. These numbers are called *probabilities*.

1.2 Elementary Probability Theory

Introductory presentations of statistics often disguise probability theory in set theory, and this is no exception.

1.2.1 Sample Space

The *sample space* is a set of *elementary outcomes*. An *event* is a subset of the sample space. Sample spaces are often denoted Ω, and events are often called A, B, C, etc. Let's get this grounded with an example:

> **Example:** For a normal die[1], the six elementary outcomes are *One*, *Two*, *Three*, *Four*, *Five* and *Six*, and thus the sample

[1] "Die" is the singular form of "dice", a sort of mechanical random generators used in games like Craps, Monopoly and Fia-med-knuff.

space Ω is the set $\{One,\ Two,\ Three,\ Four,\ Five,\ Six\}$. The events are various subsets of this set, e.g., "the outcome is One", $\{One\}$; "the outcome is less than four", $\{One,\ Two,\ Three\}$; the outcome is odd, $\{One,\ Three,\ Five\}$; etc. In fact, there are $2^6 = 64$ different subsets of Ω, i.e., there are 64 distinct events in Ω, including the empty set $\emptyset$ and Ω itself, see Section 1.2.7.

Statisticians are a kind of mathematicians, and like to distinguish between for example the outcome One, which is a basic element, and the event $\{One\}$, which is a set consisting of one element. We will try to keep this up for a while.

1.2.2 Probability Measures

A *probability measure* P is a function from events in the sample space Ω, i.e., from the set of subsets of Ω, [2] to the set of real numbers in $[0, 1]$ that has the following properties:

$$\begin{array}{ll} 1) & 0 \ \leq \ P(A) \ \leq \ 1 \quad \text{for each event } A \subseteq \Omega \\ 2) & P(\Omega) \ = \ 1 \\ 3) & A \cap B \ = \ \emptyset \ \Rightarrow \ P(A \cup B) \ = \ P(A) + P(B) \end{array}$$

Intuitively, the total mass of 1 is distributed throughout the set Ω by the function P. This will assign some particular mass to each subset A of Ω. This mass is the probability of event A, denoted $P(A)$. $A \cap B \ = \ \emptyset$ means that A and B are disjoint, i.e., that they have no common element.

Example: For a fair (unbiased) die, where as we recall the sample space is the set $\{One,\ Two,\ Three,\ Four,\ Five,\ Six\}$, the mass 1 is evenly and justly divided among the six different singleton sets. Thus $P(\{One\}) = P(\{Two\}) = P(\{Three\}) = P(\{Four\}) = P(\{Five\}) = P(\{Six\}) = \frac{1}{6}$. If A is the event of the outcome being divisible by two, i.e., the subset $\{Two,\ Four,\ Six\}$, and if B is the event of the outcome being divisible by three, i.e., the subset $\{Three,\ Six\}$, then $P(A) = \frac{1}{2}$ and $P(B) = \frac{1}{3}$.

A loaded die, used by cheaters, would not have the probability mass as evenly distributed, and could for example assign $P(\{Six\})$ a substantially larger value than $\frac{1}{6}$.

[2]This is called the *power set* of Ω and is denoted 2^Ω.

Some immediate corollaries that are worth remembering fall out from the definition of a probability measure:

$$\begin{array}{ll} a) & P(B \setminus A) \ = \ P(B) - P(A \cap B) \\ b) & A \subseteq B \ \ \Rightarrow \ \ P(A) \ \le \ P(B) \\ c) & P(\bar{A}) \ = \ 1 - P(A) \\ d) & P(\emptyset) \ = \ 0 \\ e) & P(A \cup B) \ = \ P(A) + P(B) - P(A \cap B) \end{array}$$

$B \setminus A$ denotes the difference set B minus A, i.e., the set of elements in B that are not members of A. $\bar{A}$ denotes the complement of A, i.e., $\Omega \setminus A$.

Proofs:

$$\begin{array}{ll} a) & B \ = \ (B \setminus A) \cup (A \cap B) \ ; \ (B \setminus A) \cap (A \cap B) \ = \ \emptyset \ \Rightarrow \\ & P(B) \ = \ P((B \setminus A) \cup (A \cap B)) \ = \ P(B \setminus A) + P(A \cap B) \\ b) & A \subseteq B \ \Rightarrow \ A \cap B \ = \ A \\ & 0 \ \le \ P(B \setminus A) \ = \ P(B) - P(A \cap B) \ = \ P(B) - P(A) \\ c) & \bar{A} \ = \ \Omega \setminus A \ ; \ A \cap \Omega \ = \ A \ \Rightarrow \\ & P(\bar{A}) \ = \ P(\Omega \setminus A) \ = \ P(\Omega) - P(A \cap \Omega) \ = \ 1 - P(A) \\ d) & A \ = \ A \cup \emptyset \ ; \ A \cap \emptyset \ = \ \emptyset \ \Rightarrow \\ & P(A) \ = \ P(A \cup \emptyset) \ = \ P(A) + P(\emptyset) \\ e) & A \cup B \ = \ A \cup (B \setminus A) \ ; \ A \cap (B \setminus A) \ = \ \emptyset \ \Rightarrow \\ & P(A \cup B) \ = \ P(A \cup (B \setminus A)) \ = \ P(A) + P(B \setminus A) \ = \\ & = \ P(A) + P(B) - P(A \cap B) \end{array}$$

□

Note how a) follows from 3) and how a) is used to prove b), c) and e). However, d) follows more easily from 3) directly.

1.2.3 Independence

The probability of two events A and B both occurring is the probability of the intersection of the sets A and B, $P(A \cap B)$.

Two events A and B are said to be *independent* iff[3]

$$P(A \cap B) \ = \ P(A) \cdot P(B) \tag{1}$$

[3] "Iff" means if and only if.

Intuitively, this means that the probability of A and B occurring simultaneously can be established directly from the individual probabilities of A and B.

> **Example:** To continue the example of the fair die, with the events A of the outcome being divisible by two and B of the outcome being divisible by three, we note that $P(A \cap B) = P(\{Six\}) = \frac{1}{6}$, and that $P(A) \cdot P(B) = \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6}$. Thus A and B are independent.
>
> Let C be the event of the outcome being divisible by four and note that $P(C) = P(\{Four\}) = \frac{1}{6}$. Intuitively, the property of being divisible by four is related to the property of being divisible by two, and if we calculate on the one hand $P(A \cap C) = P(\{Four\}) = \frac{1}{6}$, and on the other hand $P(A) \cdot P(C) = \frac{1}{2} \cdot \frac{1}{6} = \frac{1}{12}$, we see that A and C are not independent.

1.2.4 Conditional Probabilities

$P(A \mid B)$ is a so-called *conditional probability*, namely the probability of event A given that event B has occurred, and is defined as

$$P(A \mid B) \;=\; \frac{P(A \cap B)}{P(B)} \tag{2}$$

This is the updated probability of A once we have learned that B has occurred. $P(A)$ is often called the *prior probability* of A since we have no prior information, while $P(A \mid B)$ is called the *posterior probability* of A (knowing B), since it is defined posterior to that, that event B occurred.

Intuitively, after the event B has occurred, an event A is replaced by the event $A \cap B$, and the sample space Ω is replaced by the event B, and the probabilities are renormalized accordingly, which means dividing by $P(B)$.

The probability of an event A can change drastically when one learns that some other event has occurred. The following example is rather extreme, but illustrates this point:

$$\begin{aligned} P(A \mid A) &= 1 \\ P(A \mid \bar{A}) &= 0 \end{aligned}$$

Now, if A and B are independent, then $P(A \cap B) = P(A) \cdot P(B)$ and thus $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)} = \dfrac{P(A) \cdot P(B)}{P(B)} = P(A)$. This means that no new information is gained about A by knowing that B has occurred.

> **Example:** Again continuing the example of the fair die, with the events A of the outcome being divisible by two, B of the outcome being divisible by three, and C of the outcome being divisible by four, we note that $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)} = \dfrac{\frac{1}{6}}{\frac{1}{3}} = \dfrac{1}{2} = P(A)$, as should be, considering that A and B are independent. On the other hand, $P(A \mid C) = \dfrac{P(A \cap C)}{P(C)} = \dfrac{P(\{Four\})}{P(\{Four\})} = 1 \neq P(A) = \dfrac{1}{2}$, since A and C are not independent.

1.2.5 Bayesian Inversion

Either by returning to the definition of conditional probability, Eq. (2), or by noting that if both A and B are to occur, then first B must occur, and then A must occur, or the other way around, we can establish that

$$P(B) \cdot P(A \mid B) \;=\; P(A \cap B) \;=\; P(A) \cdot P(B \mid A)$$

This directly gives us the *Bayesian inversion formula*:

$$P(A \mid B) \;=\; \frac{P(A) \cdot P(B \mid A)}{P(B)} \tag{3}$$

This formula relates the probability of an event A conditional on another event B, i.e. $P(A \mid B)$, with the probability of event B conditional on event A, $P(B \mid A)$, and is useful if the former quantity is not easily determined, while the latter is.

> **Example:** We turn again to the example of the fair die, with the events A of the outcome being divisible by two, B of the outcome being divisible by three, and C of the outcome being divisible by four. If we wish to calculate $P(C \mid A)$, i.e., the probability that the outcome is divisible by four given that it is divisible by two, we observe that the probability $P(A \mid C)$ is 1, since divisibility by four implies divisibility by two, and

recalling that $P(A) = \frac{1}{2}$ and that $P(C) = \frac{1}{6}$, we can establish that $P(C \mid A) = \frac{P(C) \cdot P(A \mid C)}{P(A)} = \frac{\frac{1}{6} \cdot 1}{\frac{1}{2}} = \frac{1}{3}$.

1.2.6 Partitions

Assume that we have a collection $\{A_i : i = 1, \ldots, N\}$ of events (sets) such that

$$
\begin{aligned}
\Omega &= \bigcup_{i=1}^{N} A_i \\
A_i \cap A_j &= \emptyset \quad \text{for } i \neq j.
\end{aligned}
$$

The first equality means that the sets A_i cover the sample space Ω and the second equality states that all events A_i are disjoint. This situation is illustrated in Figure 1. We say that $\{A_i : i = 1, \ldots, N\}$ constitutes a

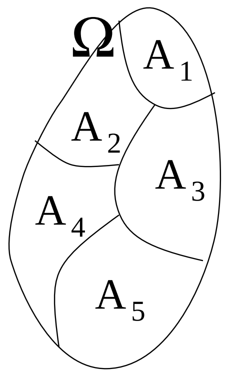

Figure 1: The sets A_1 through A_5 constitute a partition of Ω.

partition of Ω and we can establish the following formula:

$$
\begin{aligned}
P(B) &= P(B \cap \Omega) = P(B \cap \bigcup_{i=1}^{N} A_i) = \qquad (4) \\
&= P(\bigcup_{i=1}^{N} B \cap A_i) = \sum_{i=1}^{N} P(B \cap A_i) = \sum_{i=1}^{N} P(B \mid A_i) \cdot P(A_i)
\end{aligned}
$$

This can be seen as follows: Since the sample space Ω must contain the event B, B equals $B \cap \Omega$. Thus the probability $P(B)$ is the same as the probability $P(B \cap \Omega)$. As the sample space consists of the union of the disjoint events A_i, we have $\Omega = \bigcup_{i=1}^{N} A_i$, and we can thus write $P(B \cap \bigcup_{i=1}^{N} A_i)$ instead of $P(B \cap \Omega)$. Now, $B \cap \bigcup_{i=1}^{N} A_i$ is equivalent to $\bigcup_{i=1}^{N} B \cap A_i$, and we can thus rewrite $P(B \cap \bigcup_{i=1}^{N} A_i)$ as $P(\bigcup_{i=1}^{N} B \cap A_i)$. Furthermore, the probability of the union of disjoint events equals the sum of the probability of the individual events. Thus $P(\bigcup_{i=1}^{N} B \cap A_i) = \sum_{i=1}^{N} P(B \cap A_i)$. The last equality of the equation follows from the definition of conditional probabilities, Eq. (2).

Equation 4 is very useful when there is a natural partition $\{A_i\}$ of Ω, and it is easy to calculate the probabilities $P(A_i)$ and the conditional probabilities $P(B \mid A_i)$.

> **Example:** In the example of the fair die, one way one could determine the probability of event B, that the outcome is divisible by three, is to use the natural partition $A_1 = \{One\}, \ldots, A_6 = \{Six\}$ where $P(A_i) = \frac{1}{6}$ for all i, and where $P(B \mid A_i)$ is 1 if i is divisible by 3 and 0 if it is not. Thus $P(B) = 0 \cdot \frac{1}{6} + 0 \cdot \frac{1}{6} + 1 \cdot \frac{1}{6} + 0 \cdot \frac{1}{6} + 0 \cdot \frac{1}{6} + 1 \cdot \frac{1}{6} = \frac{1}{3}$.

1.2.7 Combinatorics

And now for something completely different: This section contains some results from elementary combinatorics that will be needed later, e.g., in Section 1.5.1.

The following table summarizes the number of different ways k elements can be selected from n elements. This is in fact an instance the much dreaded "urn and balls" scenario, which has plagued statistics as long as anyone can remember. This particular incarnation can be viewed as the even more feared "Lotto" scenario where the urn contains n balls numbered from 1 to n, and where we will at random draw a ball k times. In doing this, we may or may not replace selected elements, i.e., we may or may not put the drawn ball back into the urn, and we may or may not be interested in the order in which they are selected. This gives us four different cases:

	Without replacement	With replacement
Ordered	$(n)_k$	n^k
Unordered	$\binom{n}{k}$	$\binom{n+k-1}{k}$

Here

$$\begin{aligned} n! &= n \cdot (n-1) \cdot \ldots \cdot 1 \\ (n)_k &= n \cdot (n-1) \cdot \ldots \cdot (n-k+1) = \frac{n!}{(n-k)!} \\ n^k &= \overbrace{n \cdot \ldots \cdot n}^{k \text{ factors}} \\ \binom{n}{k} &= \frac{(n)_k}{k!} = \frac{n!}{(n-k)!k!} \end{aligned}$$

$n!$ is read out "n factorial". By definition $0! = 1$. In particular, the number of *permutations* of n elements is $(n)_n = n!$. $\binom{n}{k}$ is read out "n over k".

We devote the rest of the section to proofs of these formulas:

Ordered, without replacement:

We wish to select k items from n items, where selected items are not replaced, and where we are interested in the order in which they are selected. First we have n possible choices, then we have $n-1$ choices left, since we have removed one item, etc., down to the kth choice, where we have $n-(k-1) = n-k+1$ alternatives. Thus $n \cdot (n-1) \cdot \ldots \cdot (n-k+1) = (n)_k$ possibilities in total. □

Ordered with replacement:

We wish to select k items from n items, where selected items are replaced, and where we are interested in the order in which they are selected. First we have n possible choices, then we have again n possible choices, since we put back the selected item, etc., down to the kth choice, where we still have n alternatives. Thus n^k possibilities in total. □

Unordered without replacement:

We wish to select k items from n items, where selected items are not replaced, and where we are not interested in the order in which they are selected. We will first select k items paying attention to the order. Then there are $(n)_k$ possibilities. For each *set* of k different items, there will be $k!$ permutations

of them. Thus among these $(n)_k$ sequences of k items there are for each such set $k!$ sequences that should be considered to be the same, since we are not interested in the order in which items are selected. We will then factor out this by dividing $(n)_k$ by $k!$. In short:
$\#\text{unordered without replacement} = \dfrac{\#\text{ordered without replacement}}{\#\text{permutations}} =$
$\dfrac{(n)_k}{k!} = \binom{n}{k}$. Thus $\binom{n}{k}$ possibilities in total. $\Box$

Unordered with replacement:

We wish to select k items from n items, where selected items are replaced, and where we are not interested in the order in which they are selected. To this end, we make a list of the n items, and make a tick beside "i" each time we select item i, $i = 1, \ldots, n$:

k	1	...	i	j	...	n
n	|	...	||	|	...	|

We can represent this by a sequence of n zeros and k ones, where the number of ones immediately preceding the ith zero indicates the number of times item i is selected:

$$\underbrace{1 \overbrace{0 \ldots 0 0 \ldots 0}^{n} \ldots 11010 \ldots 10}_{n+k}$$

Thus, this reduces to selecting k elements (ones) from $n + k - 1$ elements without replacement, and where the order is not significant. (The "minus one" is since the last digit cannot be a one.) Thus $\binom{n+k-1}{k}$ possibilities in total. $\Box$

The last two proofs used the well-known tactics of mathematicians of reducing to Case One.

> **Story time:** A mathematician is faced with the following scenario: A bucket, a well and a house on fire. He immediately finds the correct solution of using the bucket to fetch water from the well and extinguish the fire. Next, the mathematician is faced with the scenario of a bucket, a well and a house that is not on fire. He rapidly sets fire to the house and reduces the problem to Case One.

1.3 Stochastic Variables

A *stochastic*, or *random*, *variable* ξ is a function from a sample space Ω to R, the set of real numbers. Thus, if $u \in \Omega$, then $\xi(u) \in R$. Figure 2 illustrates this. It may seem strange to call a function a variable, but this convention will die hard, if ever. There are two dominant conventions for denoting stochastic variables — one using Greek letters like $\xi, \eta, \zeta, \ldots$ and the other using Roman capital letters $X, Y, Z, \ldots$ We will use the former convention.

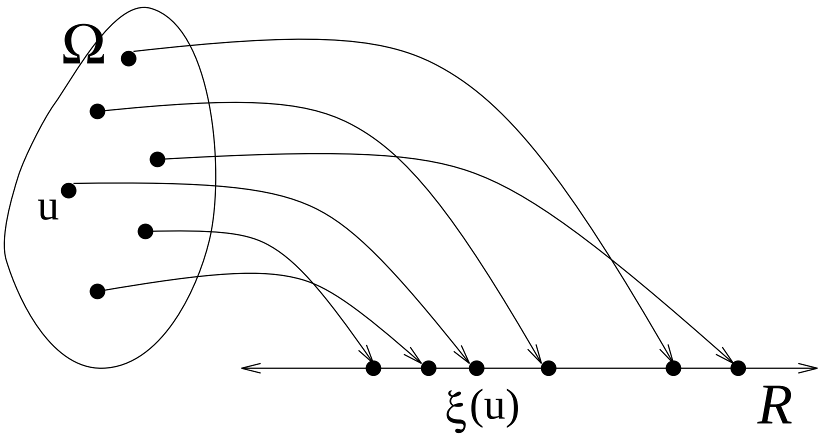

Figure 2: A random variable is a function from Ω to R.

Example: The running example of the fair die does not illustrate this very well, since there is a mapping form the sample space $\Omega = \{One,\ Two,\ Three,\ Four,\ Five,\ Six\}$ to the subset $\{1, 2, 3, 4, 5, 6\}$ of R that is so obvious, that it seems artificial to distinguish between the two. However, let us define the stochastic variable ζ as the function from Ω to R such that $\zeta(One) = 1, \zeta(Two) = 2, \ldots, \zeta(Six) = 6$.

1.3.1 Distribution Function

Let A be a subset of R, and consider the inverse image of A under ξ, i.e., $\xi^{-1}(A) = \{u : \xi(u) \in A\} \subseteq \Omega$. We will let $P(\xi \in A)$ denote the probability of this set, i.e., $P(\xi^{-1}(A)) = P(\{u : \xi(u) \in A\}) = P(\xi \in A)$. See Figure 3.

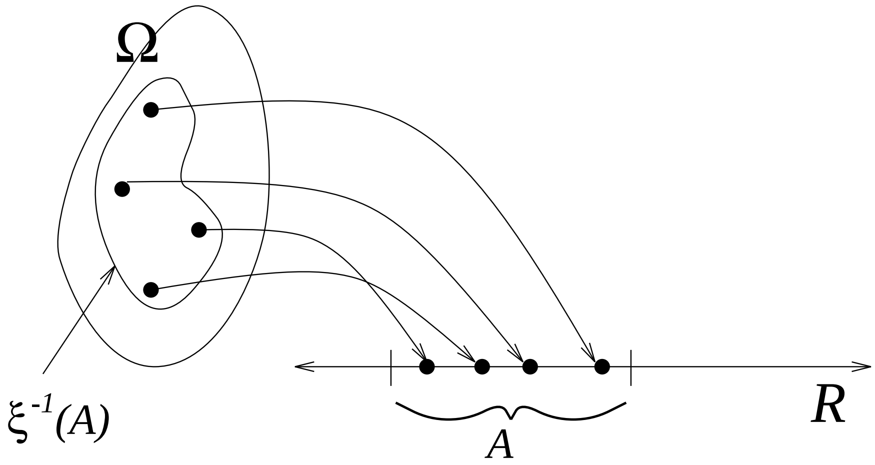

Figure 3: $P(\xi \in A)$ is defined through $P(\xi^{-1}(A)) = P(\{u : \xi(u) \in A\})$.

If A is the interval $(-\infty, x]$,[4] then the real-valued function F defined by

$$F(x) \;=\; P(\{u : \xi(u) \leq x\}) \;=\; P(\xi \leq x) \quad \text{for all } x \in R$$

is called the *distribution function* of the random variable ξ. Sometimes F is denoted F_ξ to indicate that it is the distribution function of the particular random variable ξ.

> **Example:** In the reappearing example of the fair die, the distribution function F_ζ is defined by

[4] (a, b) denotes an open interval, while $[a, b]$ denotes a closed interval, i.e., in the former case a and b do not belong to the interval, while in the latter case they do.

$$F_\zeta(x) = \begin{cases} \frac{0}{6} = 0 & \text{for} \quad x < 1, \\ \frac{1}{6} & \text{for} \quad 1 \le x < 2, \\ \frac{2}{6} = \frac{1}{3} & \text{for} \quad 2 \le x < 3, \\ \frac{3}{6} = \frac{1}{2} & \text{for} \quad 3 \le x < 4, \\ \frac{4}{6} = \frac{2}{3} & \text{for} \quad 4 \le x < 5, \\ \frac{5}{6} & \text{for} \quad 5 \le x < 6, \\ \frac{6}{6} = 1 & \text{for} \quad 6 \le x. \end{cases}$$

The graph of this function is depicted in Figure 4.

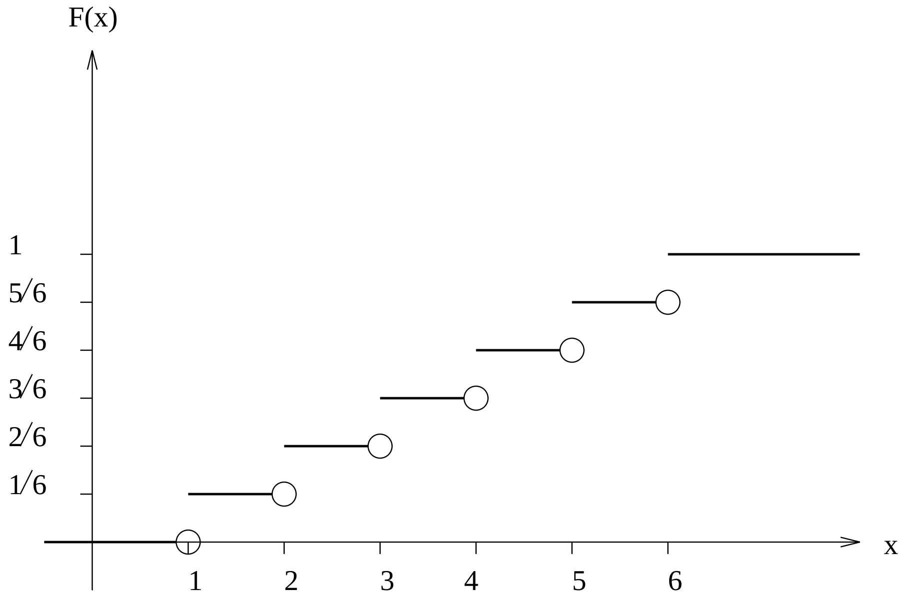

Figure 4: Fair die: Graph of the distribution function.

Some rather useful corollaries that are worth remembering include:

$a)$ $P(\xi > x) \; = \; 1 - F(x)$

$b)$ $P(a < \xi \leq b) \; = \; F(b) - F(a)$

$c)$ F is a nondecreasing function, i.e. if $x_1 < x_2$, then $F(x_1) \leq F(x_2)$.

$d)$ $\lim_{x \to -\infty} F(x) \; = \; 0$

$e)$ $\lim_{x \to \infty} F(x) \; = \; 1$

1.3.2 Discrete Stochastic Variables

The image of the sample space Ω in R under the random variable ξ, i.e., the range of ξ, is called the *sample space of the stochastic variable* ξ and is denoted Ω_ξ. In short $\Omega_\xi = \xi(\Omega)$.

Stochastic variables come in several varieties. Two common types are discrete and continuous stochastic variables:

- A random variable is *discrete* iff Ω_ξ is finite or countable.
- A random variable is *continuous* iff
  1. F is continuous, and
  2. F is differentiable with a continuous derivative except in at most a finite number of points.

The attentive reader may have noticed that Calculus is entering into this presentation through the back door. This is really necessary for a good understanding of the subject, in particular of continuous random variables, and we refer those interested in this to some introductory book on calculus. In the following, we will only present descrete cases, as these are relevant in text processing.

> **Example:** In the persistent example of the fair die, $\Omega_\zeta = \{1, 2, 3, 4, 5, 6\}$, and ζ is thus a discrete random variable.

1.3.3 Frequency Function

Another convenient way of characterizing a random variable is by its *frequency function* f:

- For a discrete random variable, $f(x) = P(\xi = x)$.

Sometimes f is denoted f_ξ to indicate that it is the frequency function of the particular random variable ξ. The frequency function of a discrete stochastic variable is often referred to as the *probability function.*

Example: Continuing the perannial example of the fair die, we find that the frequency function f_ζ is

$$f_\zeta(1) = f_\zeta(2) = f_\zeta(3) = f_\zeta(4) = f_\zeta(5) = f_\zeta(6) = \frac{1}{6}$$

This is an example of a uniform distribution, i.e., $f(x)$ has the same value for all elements in Ω_ζ. Note that a uniform distribution can only be assigned to a finite or bounded sample space. Figure 5 shows the graphical representation of the frequency function f_ζ.

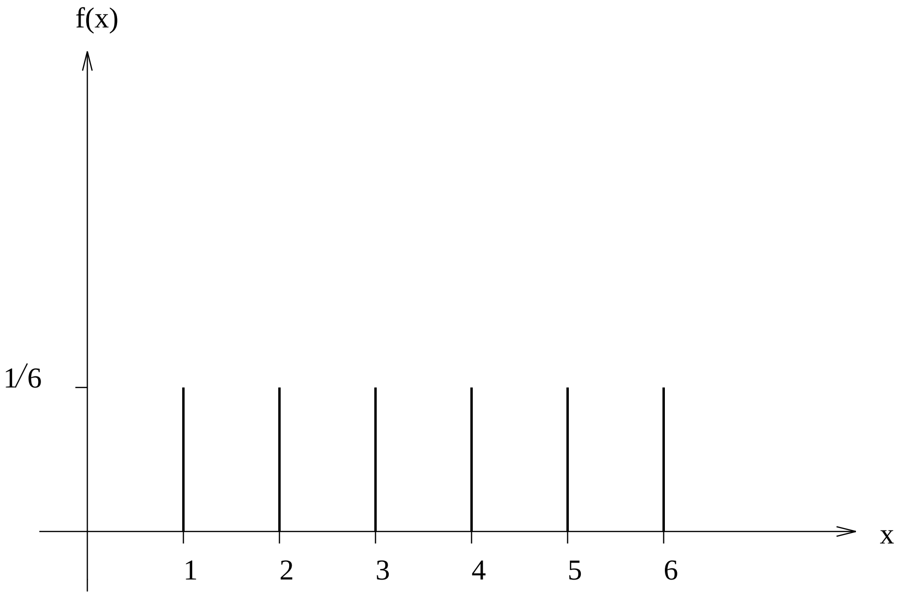

Figure 5: Fair die: Graph of the frequency function

The probabilities of events can be calculated from the frequency function:

- For a discrete random variable

$$P(\xi \in A) = \sum_{x \in A} f_\xi(x)$$

and in particular

$$P(\xi \leq x) \;=\; F_\xi(x) \;=\; \sum_{i:x_i \leq x} f_\xi(x_i)$$

Since all probabilities must sum to one, we have:

- For a discrete random variable

$$P(\Omega_\xi) \;=\; \sum_{x \in \Omega_\xi} f_\xi(x) \;=\; 1$$

1.3.4 Expectation Value

The *expectation value* or *statistical mean* of a stochastic variable ξ, denoted $\mathrm{E}[\xi]$, is defined as follows:

- For a discrete random variable

$$\mathrm{E}[\xi] \;=\; \sum_{x \in \Omega_\xi} x \cdot f(x) \;=\; \sum_i x_i \cdot f(x_i)$$

Expectation values are often denoted μ. The expectation value is the average value of the outcome of the random variable weighted by probability, indicating the center of gravity of ξ.

Example: Continuing the notorious example of the fair die,

$$\mathrm{E}[\zeta] \;=\; \sum_{i=1}^{6} i \cdot \frac{1}{6} \;=\; \frac{21}{6} \;=\; \frac{7}{2}$$

Note that this is not a possible value for ζ.

A random variable can be a function of another random variable, i.e., $\eta = g(\xi)$. The expectation value of the random variable η can be calculated from the frequency function f_ξ of ξ:

- For a discrete random variable

$$\mathrm{E}[\eta] \;=\; \mathrm{E}[g(\xi)] \;=\; \sum_{x \in \Omega_\xi} g(x) \cdot f_\xi(x) \;=\; \sum_i g(x_i) \cdot f_\xi(x_i)$$

As we shall soon see, this trick can come in quite handy.

Example: Continuing the long-lasting example of the fair die, if $\eta = \zeta^2$, then

$$\mathrm{E}[\eta] \;=\; \mathrm{E}[\zeta^2] \;=\; \sum_{i=1}^{6} i^2 \cdot \frac{1}{6} \;=\; \frac{91}{6}$$

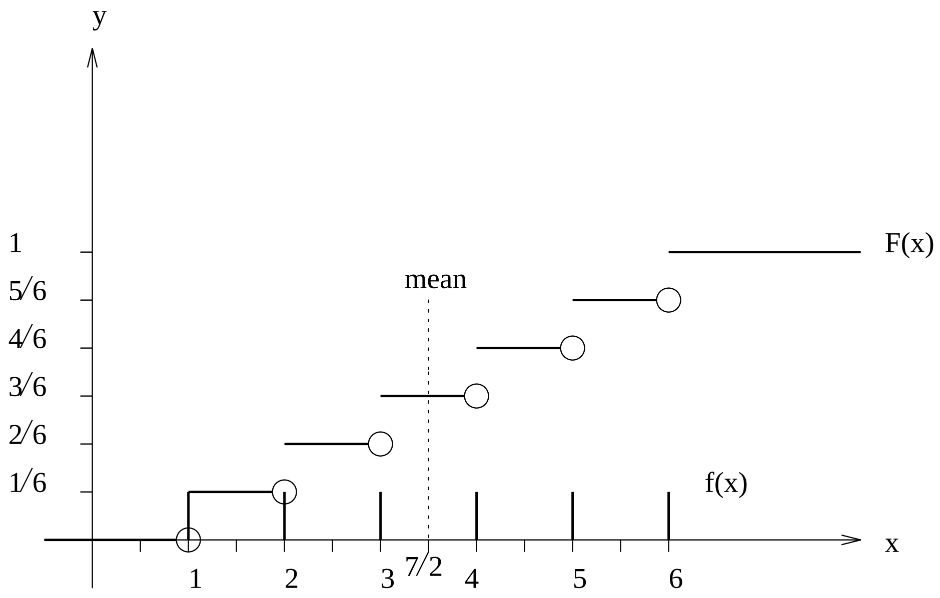

Figure 6: Fair die: Expectation value (mean)

1.3.5 Variance

The *variance* of a stochastic variable ξ, denoted $\mathrm{Var}[\xi]$, is defined as the expectation value of $(\xi - \mu)^2$, where $\mu = \mathrm{E}[\xi]$, i.e., $\mathrm{Var}[\xi] = \mathrm{E}[(\xi - \mu)^2]$.

- For a discrete random variable, this means that

$$\mathrm{Var}[\xi] \quad = \quad \sum_{x \in \Omega_\xi} (x - \mu)^2 \cdot f(x)$$

Variances are often denoted σ^2. The variance is a measure of how spread out the probability mass is from the center of gravity μ. σ itself is referred to as the *standard deviation*.

> **Example:** Continuing the inextinguishable example of the fair die,
>
> $$\mathrm{Var}[\zeta] \quad = \quad \sum_{i=1}^{6} (i - \frac{7}{2})^2 \cdot \frac{1}{6} \quad = \quad \frac{35}{12}$$

1.4 Two-dimensional Stochastic Variables

Let ξ and η be two random variables defined on the same sample space Ω. Then (ξ, η) is a two-dimensional random variable from Ω to $\Omega_{(\xi,\eta)} = \{(\xi(u), \eta(u)) : u \in \Omega\} \subseteq R^2$. Here, $R^2 = R \times R$ is the Cartesian product of the set of real numbers R with itself. This is illustrated in Figure 7. We can pull off the same sort of stunts with two-dimensional stochastic variables as with one-dimensional stochastic variables.

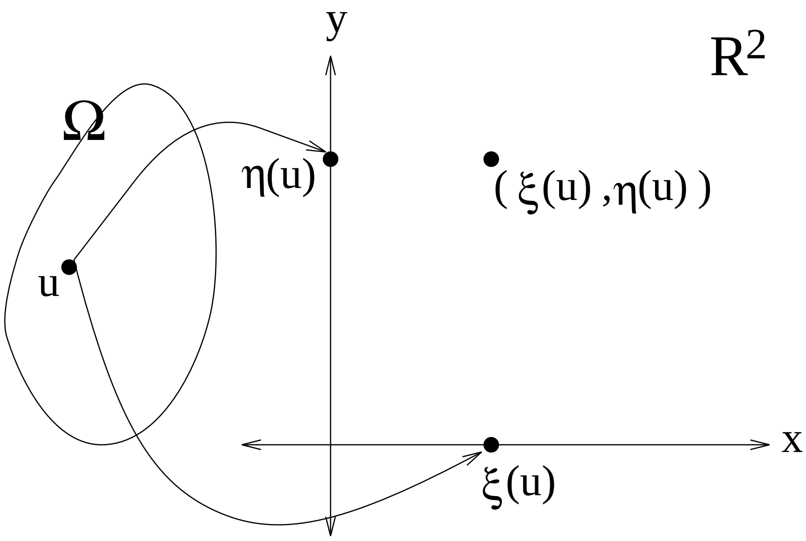

Figure 7: A two-dimensional random variable is a function from Ω to R^2.

1.4.1 Distribution Function

The distribution function of the two-dimensional random variable (ξ, η) is defined as follows:

$$F(x, y) \quad = \quad P(\xi \leq x, \eta \leq y) \quad = \quad P(\{u : \xi(u) \leq x, \eta(u) \leq y\}) \text{ for all } (x, y) \in R^2$$

This is called the *joint*, or *bivariate*, *distribution* of ξ and η and $P(\xi \in A, \eta \in B)$ is called the *joint probability* of $\xi \in A$ and $\eta \in B$.

1.4.2 Frequency Function

There are discrete and continuous versions of two-dimensional random variables, and they all have frequency functions:

- A two-dimensional random variable (ξ, η) is discrete iff $\Omega_{(\xi,\eta)}$ is finite or countable. The frequency function f of (ξ, η) is then defined by

$$f(x,y) \;=\; P(\xi = x, \eta = y) \;=\; P((\xi,\eta) = (x,y)) \text{ for } (x,y) \in R^2$$

Thus, for all two-dimensional stochastic variables we have $f(x,y) \geq 0$.

Furthermore, for $A \subseteq \Omega_{(\xi,\eta)}$

$$P((\xi,\eta) \in A) \;=\; \sum_{(x,y)\in A} f(x,y) \text{ if } (\xi,\eta) \text{ is discrete.}$$

Intuitively, a two-dimensional stochastic variable defines a mass distribution in the real plane. For a discrete variable, $f(x,y)$ is the mass at point (x,y).

For discrete stochastic variables, we can recover the frequency functions of either of the individual variables by summing over the other:
If (ξ, η) is discrete, we have

$$f_\xi(x) \;=\; \sum_{y\in\Omega_\eta} f(x,y) \quad \text{and} \quad f_\eta(y) \;=\; \sum_{x\in\Omega_\xi} f(x,y)$$

In this context, f_ξ is often referred to as the *marginal distribution* of ξ, and similarly for f_η.

1.4.3 Independence

Two stochastic variables ξ and η, defined on the same sample space, are said to be independent iff

$$P(\xi \in A, \eta \in B) \;=\; P(\xi \in A) \cdot P(\eta \in B) \tag{5}$$

for all subsets A and B of R.

Not very surprisingly, ξ and η being independent is equivalent to

- $F_{(\xi,\eta)}(x,y) = F_\xi(x) \cdot F_\eta(y)$ for all $(x,y) \in R^2$; or
- $f_{(\xi,\eta)}(x,y) = f_\xi(x) \cdot f_\eta(y)$ for all $(x,y) \in R^2$.

1.5 Selected Probability Distributions

Some particular types of stochastic variables pop up quite often, and their distributions are given special names, such as Binomial and Normal (or Gaussian) distribution. In the following, we will discuss the Binomial distribution.

1.5.1 Binomial Distribution

A common situation is when we repeat an experiment a number of times and see how many times some particular outcome occurs. For example, assume that we flip a coin a number of times, and count the number of times "heads" comes up.

Let $\Omega = \{u, v\}$ be a basic sample space (where in our example say u is "heads") and let ξ be the number of times u occurs in n independent trials. The stochastic variable ξ is then defined on the sample space

$$\overbrace{\{u, v\} \times \{u, v\} \times \ldots \times \{u, v\}}^{n \text{ factors}}$$

and an outcome in this sample space can be described as a string of length n over the alphabet $\{u, v\}$, e.g., if $n = 5$, $uvvuu$. This is mapped to a real number, in this case 3, since this string contains three us. Thus, the value of ξ is a natural number between 0 and n, i.e. $\Omega_\xi = \{0, 1, \ldots, n\}$, and ξ is a discrete random variable, since this set is finite.

In general, let p be the probability of an event u. If ξ is the number of times u occurs in n independent trials, then ξ has a *Binomial distribution* with parameters n and p. This is often denoted $\xi \sim bin\{n,p\}$. We will now try to establish the frequency function of a stochastic variable with a Binomial distribution. Assume that event u has probability p and event v has probability $q = 1 - p$. Then for example the sequence $uvvuu$ has probability $pqqpp = p^3q^2$. In fact, any sequence with three us and two vs has probability p^3q^2. Now, there are ten different such strings, i.e., there are ten outcomes in $\{u, v\} \times \{u, v\} \times \{u, v\} \times \{u, v\} \times \{u, v\}$ that map to 3, namely $uuuvv$, $uuvuv$, $uuvvu$, $uvuuv$, $uvuvu$, $uvvuu$, $vuuuv$, $vuuvu$, $vuvuu$ and $vvuuu$.
Each of these outcomes has probability p^3q^2. Thus the probability of getting exactly three "heads" in five flips is $10 \cdot (\frac{1}{2})^3 \cdot (\frac{1}{2})^2 = \frac{10}{32} = 0.3125$.

Similarly, we can compute the probability of getting exactly zero, one, two, four or five heads in five independent trials of the coin-flipping experiment. Figure 8 shows the resulting probabilities, and Figure 9 gives a graphical representation of the distribution.

k	0	1	2	3	4	5	$\sum$
$P(\xi = k)$	$\frac{1}{32}$	$\frac{5}{32}$	$\frac{10}{32}$	$\frac{10}{32}$	$\frac{5}{32}$	$\frac{1}{32}$	1

Figure 8: The probability of getting $0, 1, 2, 3, 4, 5$ heads in 5 independent trials.

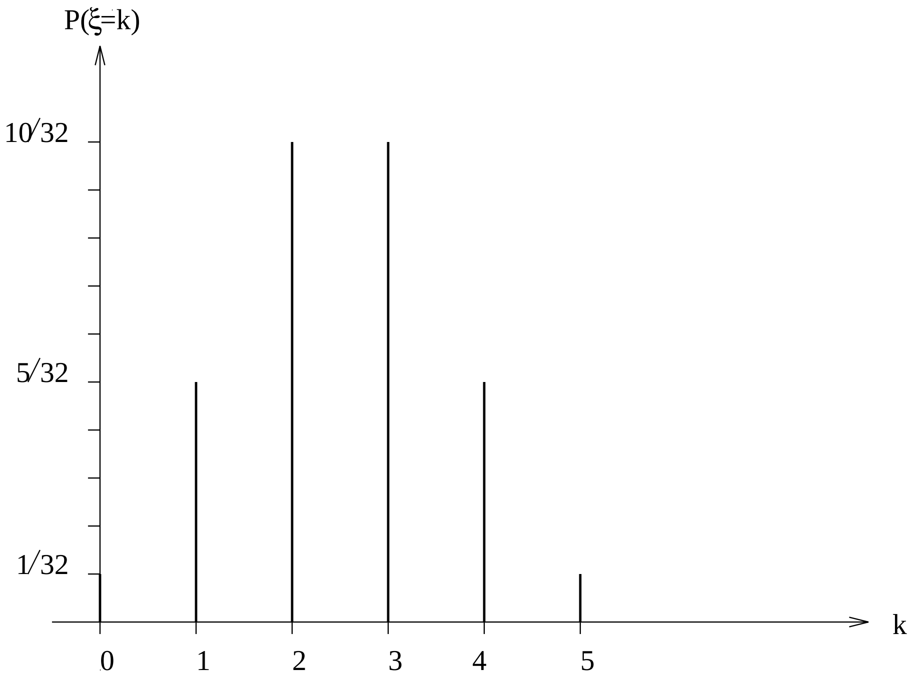

Figure 9: The probability distribution when flipping a coin five times.

In general, any string of k us and $n-k$ vs will have probability $p^k q^{n-k}$. Now, there are $\binom{n}{k} = \frac{n!}{(n-k)!k!}$ distinct such strings, as explained in detail in Section 1.2.7. Thus, the probability $P(\xi = k) = f_\xi(k) = \binom{n}{k} p^k q^{n-k}$.

In words, the probability $P(\xi = k)$, defining the frequency function $f_\xi(k)$, is the product of the number of possible ways of selecting k items from n items, $\binom{n}{k}$, and the probability of each such possibility, $p^k q^{n-k}$. This is valid for $k = 0, \ldots, n$.

This gives us the frequency function of a binomially distributed variable:

$$f_\xi(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k \in \{0, 1, \ldots, n\}, \quad 0 < p < 1 \tag{6}$$

This distribution is parameterized by the probability p of the particular outcome u, and n, the number of trials.

1.5.2 Maximum-Likelihood Estimators

How do we find good estimators? One method is to use *Maximum-Likelihood Estimators*, MLEs. This method is based on a simple, but important idea:

> *Choose the alternative that maximizes the probability of the observed outcome.*

The most important example of a maximum-likelihood estimator is the relative frequency f_n of an outcome, which is the maximum-likelihood estimator of the probability p of that outcome.