Statistics

Statistics has many aspects. These include:

- data gathering
- data analysis
- statistical inference.

For instance, politicians may wish to see whether a new law will be popular. They ask a polling company to sample voters. The polling company telephones 1000 people “at random”. They prepare a list of people, with their ages and addresses, and how they feel about issues like banning wood-chipping in old-growth forests. This data is compressed into a few tables (often presented graphically) and then the number-crunchers are invited to comment, and advise whether the policy should be proposed or not. This will be based on electorates whose vote might change as a result of the new law. The age data will be used to help decide the timing: now, next year, next decade, never? Statistics is about making decisions from data, and also about self-criticism: what is the margin of error in the analysis? How do we balance the costs of, e.g., quality control, against the reliability of the results? How many cars do we need to crash to determine whether the occupants will be safe in case of an accident? How much should we test our new software before release?

This chapter is an introduction to the ideas needed in “serious statistics”. But it is only an introduction.

1. Descriptive statistics

The weather bureau produces lists of temperatures: maxima, minima, temperature at 8 am, and so on. People working for the clothing industry produce lists of height, weight, chest size, and so on, of “randomly chosen” people. Architects and clothes designers need this data in a more “user friendly” form. Histograms, or bar charts are commonly used. For instance, the lists of temperatures can be presented as a “frequency distribution”, e.g.,

<table>
<thead>
<tr>
<th>Max. temp.</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>11–15</td>
<td>17</td>
</tr>
<tr>
<td>16–20</td>
<td>123</td>
</tr>
<tr>
<td>21–25</td>
<td>157</td>
</tr>
<tr>
<td>26–30</td>
<td>58</td>
</tr>
<tr>
<td>31–35</td>
<td>9</td>
</tr>
<tr>
<td>36–40</td>
<td>2</td>
</tr>
</tbody>
</table>

or as a histogram:
An architect might use this kind of data in designing houses or flats.

Another way of presenting the data is via a box and whisker plot. The box contains 50% of the data points. The whiskers show the range of all or most of the data. The “median” is also indicated. For a data set \( \{x_1, x_2, \ldots, x_n\} \), arranged in increasing order, the (sample) median is \( x_{(n+1)/2} \) if \( n \) is odd, and \( (x_{n/2} + x_{(n+1)/2})/2 \) if \( n \) is even. Quartiles and percentiles are defined similarly.

In finance, similar data is available about, for example, returns on shares in the year 1998. There are hundreds of data points, and usually the data is presented as a “smoothed out histogram”:

The vertical axis shows frequency, or often “relative frequency”. If you are asked “what is a typical return on shares”, you might answer giving the mean, or average, the mode, or most likely, or the median, where 50% is greater and 50% worse. Which is best of these depends on the use to which the numbers will be put. In particular, note that modes are not always well-defined. In the diagram above, the mode is to the left of the median, and the mean is to its right.

Cumulative (relative) frequency distributions are other ways of presenting data. These are useful for answering questions like “how likely is it in September that the maximum temperature will be 18° or less?”
2. Populations and sampling

Suppose that the Australian government wants to know what fraction of the population wants to declare war on New Zealand, and employs a statistician to find this out. The statistician will choose a “random sample” of the population and poll these people.

What do we mean by a random sample? If there was an all-Australia phonebook in which everybody in Australia was listed exactly once, then we could open a page and choose a name “at random”, then repeat the process. “Sampling without replacement” means that we exclude the names which have already been chosen, while “sampling with replacement” means that we allow the same name to be chosen more than once. A popular model is choosing balls from an urn containing coloured balls. If the balls correspond to the total population, then sampling with or without replacement corresponds to putting the balls back in or keeping them out when the later balls are chosen.

**Problem 27.** Suppose that an urn contains $N/5$ red balls and $4N/5$ white balls. A sample of $n$ balls is drawn. What is the probability that $k$ of these balls are red? Assume that the sample is drawn

(i) with replacement

(ii) without replacement.

**Answer.** Assume that the balls are labelled $1, 2, 3, \ldots, N$, and that balls $1, 2, \ldots, N/5$ are red.

First we consider case (i), “with replacement”.

The possible outcomes may be represented as strings (lists) of numbers, each in the range $1, 2, \ldots, N$: the first number represents the first ball, etc. There are $N^n$ possible strings, or outcomes, each equally probable.

To have $k$ red balls, $k$ of the numbers in the string must be in the range $1, 2, \ldots, N/5$, and the rest in the range $N/5+1, N/5+2, \ldots, N$. We may choose the $k$ places in the string out of the total of $n$ in $\binom{n}{k}$ ways. Having chosen these $k$ places, there are $N/5$ possible choices of red balls for each of them, so $(N/5)^k$ possible choices of the whole collection. There are $(4N/5)^{n-k}$ possible choices of the remaining numbers. In total, there are

$$\binom{n}{k} (N/5)^k (4N/5)^{n-k}$$

ways of choosing outcomes with $k$ red and $n-k$ white balls. The ratio

$$\frac{\binom{n}{k} (N/5)^k (4N/5)^{n-k}}{N^n} = \binom{n}{k} (1/5)^k (4/5)^{n-k}$$

is the probability of $k$ red balls.

Note that $N$ does not appear in the answer. Note also that we get the same answer by arguing there are $\binom{n}{k}$ ways of choosing which balls will be red, and then the probability of each one of these ways is $(1/5)^k (4/5)^{n-k}$, so the total probability is

$$\binom{n}{k} (1/5)^k (4/5)^{n-k}.$$
Now we consider the second case, “without replacement”.

The outcomes are strings of \( n \) distinct integers from the range 1, 2, \ldots, \( N \). The number of these is \( N(N-1)(N-2) \ldots (N-n+1) = \binom{N}{n} n! \). We can think of this as saying that there are \( \binom{N}{n} \) ways to choose the integers in the string and \( n! \) possible orders.

To count the strings with \( k \) numbers in the “red range” 1, 2, \ldots, \( N/5 \) and the other \( n-k \) numbers in the “white range” \( N/5+1, N/5+2, \ldots, N \), we observe that there are \( \binom{N/5}{k} \) ways to choose the \( k \) numbers from the red range and \( \binom{4N/5}{n-k} \) to choose the numbers from the white range. There are \( n! \) ways to put these in order. The total number of strings which correspond to \( k \) red balls is

\[
\binom{N/5}{k} \binom{4N/5}{n-k} n!
\]

Then the probability of \( k \) red balls is

\[
\frac{\binom{N/5}{k} \binom{4N/5}{n-k} n!}{\binom{N}{n} n!} = \frac{\binom{N/5}{k} \binom{4N/5}{n-k}}{\binom{N}{n}}.
\]

In this case, \( N \) does appear in the answer. Indeed, if \( n > N \), then there are no ways of choosing a sample of \( n \) items without replacement from the population of \( N \), and the probability must be 0. For this to be possible, the probability must depend on \( N \). Note also that we need not have bothered about the order of the balls in this argument. \( \Box \)

Observe that, if \( N \gg n \), then the probability of \( k \) red balls in a sample of size \( n \) without replacement is

\[
\frac{\binom{N/5}{k} \binom{4N/5}{n-k}}{\binom{N}{n}} = \frac{(N/5)! (4N/5)! n! (N-n)!}{N! k! (N/5-k)! (n-k)! (4N/5-n+k)! n! (N/5)! (4N/5)! (N-n)!}
\]

\[
= \frac{k! (n-k)! (N/5-k)! (4N/5-n+k)! N!}{N(N-1)(N-2) \cdots (N-n+1)}
\]

\[
= \binom{n}{k} \frac{(N/5)(N/5-1) \cdots (N/5-k+1)(4N/5)(4N/5-1) \cdots (4N/5-n+k+1)}{N N N \cdots N}
\]

\[
= \binom{n}{k} (1/5)^k (4/5)^{n-k},
\]

and there is not much difference between sampling with replacement or without. It is an interesting exercise to try to see how big the difference between the probabilities of
sampling with and without replacement is. By this argument, we can show that
\[ \lim_{N \to \infty} \left( \frac{N}{k} \right) \left( \frac{4N}{5} \right) \left( \frac{N}{n} \right)^{-1} = \left( \frac{n}{k} \right) (1/5)^k (4/5)^{n-k}. \]

More generally, we can show that, if we have \( Np \) red balls and \( Nq \) white balls, where \( p + q = 1 \), then the probabilities that samples of \( n \) contain \( k \) white balls are
\[ \left( \frac{n}{k} \right) p^k q^{n-k} \] when we sample with replacement
\[ \left( \frac{n}{k} \right) \left( \frac{N-n}{Np-k} \right) \left( \frac{N}{Np} \right)^{-1} \] when we sample without replacement.

Further,
\[ \lim_{N \to \infty} \left( \frac{n}{k} \right) \left( \frac{N-n}{Np-k} \right) \left( \frac{N}{Np} \right)^{-1} = \left( \frac{n}{k} \right) p^k q^{n-k}. \]

When \( N \) is big, we always use such approximations, because calculating \( \left( \frac{N}{Np} \right) \) and other similar expressions is very messy.

**Challenge Problem.** Suppose \( N = 2 \times 10^8 \) and \( p = 1/2 \). Compute \( \left( \frac{N}{Np} \right) \).

There is another approximation, called the Poisson approximation, which is important when \( n \) is large and \( p \) is small.

**Theorem 21.** Suppose that \( \lambda \) is fixed, and that \( np = \lambda \). Then
\[ \lim_{n \to \infty} \left( \frac{n}{k} \right) p^k (1-p)^{n-k} = \frac{e^{-\lambda} \lambda^k}{k!}. \]

**Proof.** Since \( p = \lambda/n \),
\[ \left( \frac{n}{k} \right) p^k (1-p)^{n-k} = \frac{n!}{(n-k)!} \frac{(\lambda/n)^k}{k!} \left( \frac{1-\lambda/n}{n} \right)^{n-k} \]
\[ = \frac{n(n-1) \cdots (n-k+1)}{n \cdots n} \frac{1}{k!} \lambda^k \left( \frac{1-\lambda/n}{n} \right)^{n-k}. \]

Now \( \lim_{n \to \infty} (1-\lambda/n)^n = e^{-\lambda} \) and \( \lim_{n \to \infty} (1-\lambda/n)^{-k} = 1 \), and the result follows.

Consider \( N \) atoms of uranium \((U^{235})\). How can we assign a probability that any given atom will decay in the next minute (or hour, or day)?

If we suppose that this probability exists, and write it as \( p \), then we might:

- suppose that each atom has exactly the same probability of decaying, and that they decay independently
- compute the probability that exactly \( k \) atoms decay in the time period: it is
\[ \left( \frac{N}{k} \right) p^k (1-p)^{N-k} \]
- use the theorem above to deduce that this is approximately equal to
\[ \frac{\lambda^k e^{-\lambda}}{k!} \]
• observe the same sample many times (or different samples once) with a Geiger counter to see how many atoms decay, and plot the number of decays observed in a histogram:

\[
\text{relative frequency of number of times observed}
\]

\[k\]

• find the value of \(\lambda\) for which \(\frac{\lambda^k e^{-\lambda}}{k!}\) fit the data best

• keep on observing to see if our hypotheses are invalidated by subsequent observations.

This is the process by which physicists estimate the probability that are given atom will decay in a given time period. The “Poisson distribution” fits the observed data very well. Based on these calculations, “half lives” of radio-active substances are estimated.

Very similar calculations are done by people who run “call centres”. There are millions of potential callers, but the probability that they will call is low. Again, the number of people who call in a given time period is described by a probability \(\frac{\lambda^k e^{-\lambda}}{k!}\). “Queuing theory” is then used to ensure that there is a good match between the number of callers and the number of operators. This considers how the calls arrive in a time interval as well as how many.

Given a sample from a population, and numbers defined on the sample, \(x_1, x_2, \ldots, x_n\) say, we define the sample mean \(\bar{x}\) by

\[
\bar{x} = \frac{1}{n} (x_1 + x_2 + \cdots + x_n),
\]

and the sample variance \(s^2\) by

\[
s^2 = \frac{1}{n} \left( (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2 \right).
\]

In summation notation,

\[
\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i
\]

\[
s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2.
\]
Note that
\[ s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i^2 - 2\bar{x}x_i + \bar{x}^2) \]
\[ = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \frac{2\bar{x}}{n} \sum_{i=1}^{n} x_i + \bar{x}^2 \]
\[ = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \bar{x}^2. \]

The variance \( s^2 \) measures the “spread” of the sample. The sample standard deviation \( s \) is defined to be \( \sqrt{s^2} \).

3. Discrete random variables and expectations

The roll of a dice, or the number of letters posted in a given letter box, or the way that people vote, or the number of shares sold are all “random”. Philosophically, they may or may not be random, but even if they are not, we cannot predict them, and the only effective way we have to deal with them is by treating them as if they were random. They are examples of “discrete random variables”: in this course, a discrete random variable means something which varies in an apparently random way, whose possible values are integers, or whose possible values can be indexed by the integers—e.g., drawing a card at “random” from a pack of cards, where we label the cards by the numbers 1 to 52, or a “yes” or “no” answer, where we label “yes” by 1 and “no” by 0. We assume that there exists a collection of numbers \( p_k \) such that the probability that the random variable takes the value \( k \) is \( p_k \).

Usually we denote random variables by uppercase letters: \( X \), \( Y \), \( Z \), and so on. The probability \( p_k \) that \( X \) takes the value \( k \) is written \( \mathbb{P}(X = k) \). We require that the numbers \( p_k \) satisfy the conditions
\[ p_k \geq 0 \]
\[ \sum_{k} p_k = 1. \]

Then \( \{p_k\} \) is called a (discrete) probability distribution. Given a random variable \( X \) with probability distribution \( \{p_k\} \), we define the mean of \( X \), or expected value of \( X \), to be
\[ \mathbb{E}(X) = \sum_{k} kp_k. \]

If we find values for \( X \) repeatedly, then the value \( k \) will come up with relative frequency \( p_k \), and the average value for \( X \) will be \( \mathbb{E}(X) \). We also define the moments of \( X \):
\[ \mathbb{E}(X^2) = \sum_{k} k^2 p_k \]
\[ \mathbb{E}(X^3) = \sum_{k} k^3 p_k \]
and so forth, the variance of $X$:

$$\text{Var}(X) = \sum_k (k - \mathbb{E}(X))^2 p_k,$$

and the standard deviation of $X$:

$$\sigma(X) = \sqrt{\text{Var}(X)}.$$

When we have a discrete probability distribution, the variance and standard deviation measure the “spread” of the distribution. Indeed, if $X$ is a random variable, then $\text{Var}(X)$ is an “average” of the squares of the distances of $X$ from the mean. Note that

$$\text{Var}(X) = \sum_k k^2 p_k - 2\mathbb{E}(X)\sum_k kp_k + \mathbb{E}(X)^2$$

$$= \sum_k k^2 p_k - 2\mathbb{E}(X)\sum_k kp_k + \mathbb{E}(X)^2$$

$$= \sum_k k^2 p_k - \mathbb{E}(X)^2$$

$$= \mathbb{E}(X^2) - \mathbb{E}(X)^2.$$

**Problem 28.** Compute the mean and variance for a binomially distributed random variable, for which

$$\mathbb{P}(X = k) = \binom{n}{k} p^k q^{n-k}.$$

**Answer.** First, we compute the mean:

$$\mathbb{E}(X) = \sum_{k=0}^{n} k \binom{n}{k} p^k q^{n-k}$$

$$= \sum_{k=1}^{n} \frac{n!}{(k-1)! (n-k)!} p^k q^{n-k}$$

$$= np \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)! (n-k)!} p^{k-1} q^{n-k}$$

$$= np \sum_{j=0}^{n-1} \frac{(n-1)!}{j!(n-1-j)!} p^j q^{n-1-j}$$

$$= np \sum_{j=0}^{n-1} \binom{n-1}{j} p^j q^{n-1-j}$$

$$= np (p+q)^{n-1}$$

$$= np.$$

For $\mathbb{E}(X^2)$, we use the fact that $k^2 = k(k-1) + k$. Then a similar calculation shows that $\mathbb{E}(X^2) = n(n-1)p^2 + np$, whence $\text{Var}(X) = npq$. \(\triangle\)
Problem 29. Suppose that
\[ P(X = k) = \begin{cases} \frac{1}{n} & k = 1, 2, \ldots, n \\ 0 & \text{otherwise} \end{cases} \]

Find the mean and the variance of \( X \).

Answer. First,
\[
\mathbb{E}(X) = \sum_{k=1}^{n} k \frac{1}{n} = \frac{1}{n} \sum_{k=1}^{n} k = \frac{n+1}{2}.
\]
Also
\[
\mathbb{E}(X^2) = \sum_{k=1}^{n} k^2 \frac{1}{n} = \frac{1}{n} \sum_{k=1}^{n} k^2 = \frac{(n+1)(2n+1)}{6},
\]
whence \( \text{Var}(X) = \frac{n^2 - 1}{12} \). \( \triangle \)

Problem 30. Suppose that
\[ P(X = k) = \frac{e^{-\lambda} \lambda^k}{k!} \quad k = 0, 1, 2, \ldots \]

Find the mean and the variance of \( X \).

Answer. First,
\[
\mathbb{E}(X) = \sum_{k=0}^{\infty} k \frac{e^{-\lambda} \lambda^k}{k!} = \sum_{k=1}^{\infty} \frac{e^{-\lambda} \lambda^{k-1+1}}{(k-1)!} = \lambda \sum_{j=0}^{\infty} \frac{e^{-\lambda} \lambda^j}{j!} = \lambda.
\]
Next,

\[ \mathbb{E}(X^2) = \sum_{k=0}^{\infty} k^2 \frac{e^{-\lambda} \lambda^k}{k!} \]

\[ = \sum_{k=0}^{\infty} k(k-1)\frac{e^{-\lambda} \lambda^k}{k!} + \sum_{k=0}^{\infty} \frac{e^{-\lambda} \lambda^k}{k!} \]

\[ = \sum_{k=2}^{\infty} \frac{e^{-\lambda} \lambda^{k-2+2}}{(k-2)!} + \lambda \]

\[ = \lambda^2 \sum_{j=0}^{\infty} \frac{e^{-\lambda} \lambda^{j}}{j!} + \lambda \]

\[ = \lambda^2 + \lambda, \]

whence \( \text{Var}(X) = \lambda. \) \( \triangle \)

For calculations with discrete random variables, remember that

\[ \sum_{k=1}^{n} k = \frac{n(n+1)}{2}, \]

\[ \sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}, \]

\[ \sum_{k=1}^{n} k^3 = \frac{n^2(n+1)^2}{4}, \]

\[ \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = 1, \]

\[ \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^\lambda. \]

4. Continuous random variables

Many important random variables take real values, not just integer values, e.g., heights of randomly chosen people, \( x \) coordinates of random points in the plane, lifetimes of mechanical or electrical products. The probability that such a random variable takes any particular value is usually 0, but the probability that it lies in some interval can be nonzero.

A continuous random variable is defined to be a random variable \( X \) for which there exists a function \( f: \mathbb{R} \to [0, \infty) \), called the probability density function, such that

\[ \mathbb{P}(a < X < b) = \int_{a}^{b} f(x) \, dx. \]
It is required that
\[ \int_{-\infty}^{\infty} f(x) \, dx = 1. \]

**Example.** The random variable \( X \) is said to be uniformly distributed in the interval \([c, d]\) if
\[ f(x) = \begin{cases} 
\frac{1}{d-c} & c < x < d \\
0 & \text{otherwise.} 
\end{cases} \]

**Example.** The random variable \( X \) is said to be exponentially distributed with parameter \( \alpha > 0 \) if
\[ f(x) = \begin{cases} 
\alpha e^{-\alpha x} & x > 0 \\
0 & x \leq 0. 
\end{cases} \]

**Example.** Suppose that \( \mu \in \mathbb{R} \) and \( \sigma \in \mathbb{R}^+ \). The random variable \( X \) is said to be normally distributed with mean \( \mu \) and variance \( \sigma^2 \) (and we write \( N(\mu, \sigma) \)) if
\[ f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \]

**Problem 31.** Check that
\[ \int_{-\infty}^{\infty} f(x) \, dx = 1 \]
for each of the previous examples.

For continuous random variables, we define
\[ \mathbb{E}(X) = \int_{-\infty}^{\infty} x \, f(x) \, dx \]
\[ \mathbb{E}(X^2) = \int_{-\infty}^{\infty} x^2 \, f(x) \, dx \]
\[ \mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x) \, f(x) \, dx. \]

In particular,
\[ \mathbb{E} \left( (X - \mathbb{E}(X))^2 \right) = \int_{-\infty}^{\infty} (x - \mathbb{E}(X))^2 \, f(x) \, dx \]
\[ = \int_{-\infty}^{\infty} [x^2 - 2x\mathbb{E}(X) + \mathbb{E}(X)^2] \, f(x) \, dx \]
\[ = \int_{-\infty}^{\infty} x^2 \, f(x) \, dx - 2\mathbb{E}(X)^2 + \mathbb{E}(X)^2 \]
\[ = \mathbb{E}(X^2) - \mathbb{E}(X)^2. \]

We also define \( \text{Var}(X) \) to be \( \mathbb{E} \left( (X - \mathbb{E}(X))^2 \right) \), and \( \sigma(X) \) to be \( \text{Var}(X)^{1/2} \).
5. Examples of continuous probability distributions

Problem 32. Find the mean and variance of the random variable $X$ which is uniformly distributed on $[c, d]$.

Answer. For this random variable,

$$f(x) = \begin{cases} 
\frac{1}{d-c} & \text{if } c < x < d \\
0 & \text{otherwise.} 
\end{cases}$$

Then

$$\mathbb{P}(X \in [a, b]) = \int_a^b f(x) \, dx = \begin{cases} 
1 & \text{if } a \leq c < d \leq b \\
0 & \text{if } b \leq c \text{ or } d \leq a \\
\frac{b-a}{d-c} & \text{if } c \leq a \leq b \leq d \\
\ldots & \ldots \ldots \ldots 
\end{cases}$$

For this random variable,

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x) \, dx$$

$$= \frac{1}{d-c} \int_c^d x \, dx$$

$$= \frac{d^2 - c^2}{2(d-c)}$$

$$= \frac{d + c}{2}.$$ 

Further,

$$\mathbb{E}(X^2) = \int_{-\infty}^{\infty} x^2 f(x) \, dx$$

$$= \frac{1}{d-c} \int_c^d x^2 \, dx$$

$$= \frac{d^3 - c^3}{3(d-c)}$$

$$= \frac{d^2 + cd + c^2}{3}.$$ 

It follows that $\text{Var}(X) = (d - c)^2 / 12.$ \hfill \triangle$

Problem 33. Show that

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2 / 2\sigma^2}$$

is a probability density function.
5. EXAMPLES OF CONTINUOUS PROBABILITY DISTRIBUTIONS

Answer. Clearly, \( f \geq 0 \). Further, substituting \( z = (x - \mu)/\sigma \),

\[
\int_{-\infty}^{\infty} f(x) \, dx = \frac{1}{\sqrt{2\pi\sigma}} \int_{-\infty}^{\infty} e^{-z^2/2\sigma^2} \, dz \\
= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2/2} \, dz.
\]

Define

\[
I = \int_{-\infty}^{\infty} e^{-z^2/2} \, dz \\
= \int_{-\infty}^{\infty} e^{-w^2/2} \, dw.
\]

Then

\[
I^2 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-(w^2+z^2)/2} \, dz \, dw.
\]

This integral represents a volume with circular symmetry around the vertical axis and an infinite base. If we calculate the volume using cylindrical shells, we get

\[
I^2 = \int_{0}^{\infty} 2\pi r e^{-r^2/2} \, dr \\
= 2\pi \int_{0}^{\infty} e^{-r^2/2} r \, dr \\
= 2\pi \int_{0}^{\infty} e^{-s} \, ds \\
= 2\pi.
\]

Therefore \( I = \sqrt{2\pi} \), and \( f \), as defined above, is a probability density function. \( \triangle \)

Problem 34. Find the mean and variance of the random variable \( X \) with normal distribution \( N(\mu, \sigma) \).

Answer. For this random variable,

\[
f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-(x-\mu)^2/2\sigma^2}.
\]

The substitution \( z = (x - \mu)/\sigma \) gives

\[
\mathbb{E}(X) = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi\sigma}} e^{-(x-\mu)^2/2\sigma^2} \, dx \\
= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (\mu + \sigma z) e^{-z^2/2} \, dz \\
= \frac{\mu}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2/2} \, dz + \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} ze^{-z^2/2} \, dz \\
= \mu.
\]
Also,

\[
\mathbb{E}(X^2) = \int_{-\infty}^{\infty} x^2 \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \, dx
\]

\[
= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (\mu + \sigma z)^2 e^{-\frac{z^2}{2}} \, dz.
\]

Actually, it is easier to find

\[
\mathbb{E}((X-\mu)^2) = \int_{-\infty}^{\infty} (x-\mu)^2 \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \, dx
\]

\[
= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (\sigma z)^2 e^{-\frac{z^2}{2}} \, dz
\]

\[
= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-\frac{z^2}{2}} \, dz
\]

\[
= -\frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z \cdot \left[-ze^{-\frac{z^2}{2}}\right] \, dz
\]

\[
= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} 1 \cdot e^{-\frac{z^2}{2}} \, dz
\]

\[
= \sigma^2.
\]

(Really, we should consider \(\lim_{R \to \infty} \int_{-R}^{R} \cdots \, dz\) when we integrate by parts). It follows that \(\text{Var}(X) = \sigma^2\). \(\triangle\)

To calculate probabilities with the normal distribution, we use the same change of variables: if \(X \sim N(\mu, \sigma)\),

\[
\mathbb{P}(a < X < b) = \int_{a}^{b} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \, d\sigma
\]

\[
= \int_{\mu+\sigma a}^{\mu+\sigma b} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \, dz.
\]

This integral cannot be computed in terms of standard functions, but it can be computed numerically to arbitrary high degrees of precision.

It is standard to tabulate

\[
\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \, dz
\]

for \(x \geq 0\). By symmetry, or change of variables,

\[
\int_{a}^{b} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \, dz = \int_{-b}^{-a} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \, dz.
\]

With the tables of \(\Phi(x)\) and various arguments, we can find \(\mathbb{P}(a < Z < b)\) for any \(a\) and \(b\).
Problem 35. Suppose that \( Z \) is \( N(0, 1) \). (Such a random variable is called standard normal.) Find \( P(Z < -0.2) \), \( P(1 < Z < 2) \) and \( P(-1 < Z < 2) \).

Answer. First, by symmetry, \( P(Z < -0.2) = P(Z > 0.2) \). Next, \( P(Z > 0.2) = 1 - P(Z \leq 0.2) \). Finally \( P(Z \leq 0.2) \) can be found from the tables: it is .5693. Thus
\[
P(Z < -0.2) = 1 - P(Z \leq 0.2)
= 1 - \Phi(0.2)
= 0.4307.
\]

By symmetry, we see that the area under the curve where \( z < -0.2 \) is equal to the area under the curve where \( z > 0.2 \), which is the total area under the curve, i.e., 1, less the area under the curve where \( z < 0.2 \).

\( P(1 < Z < 2) \) corresponds to the difference \( \Phi(2) - \Phi(1) \), i.e.,
\[
P(1 < Z < 2) = .9772 - .8413 = .1359.
\]

Finally,
\[
P(-1 < Z < 2) = P(Z < 2) - P(Z \leq -1)
= \Phi(2) - (1 - P(Z < 1))
= \Phi(2) + \Phi(1) - 1
= .8185.
\]
Note that $\mathbb{P}(-1 < Z < 1) = \mathbb{P}(-1 < Z < 2) - \mathbb{P}(1 < Z < 2) = 0.6826$, that is, about two thirds of the area under the standard normal curve lies less than one standard deviation from the mean.

In MAPLE, we can use the numerical integration to compute $\Phi(x)$. Alternatively, MAPLE knows the “error function”:

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} \, dt.$$ 

6. Changes of variables

Given a random variable $X$, we also deal with functions of $X$, e.g., $X^2$, $\frac{X-\mu}{\sigma}$, or $g(X)$, which are also random variables. Suppose that $g$ is an increasing function, and that $\lim_{t \to \pm\infty} g(t) = \pm\infty$. What can we say about $g(X)$?

It is useful to introduce the cumulative distribution function: $F(x) = \mathbb{P}(X \leq x)$. Since $g$ is increasing, $X \leq x$ if and only if $g(X) \leq g(x)$. Thus

$$\mathbb{P}(g(X) \leq g(x)) = \mathbb{P}(X \leq x) = F(x).$$

Since $g$ is increasing, $g^{-1}$ is defined. Put $y = g(x), x = g^{-1}(y)$. Then

$$\mathbb{P}(g(X) \leq y) = F(g^{-1}(y)).$$

Thus, in principle, we know the cumulative probability distribution of $g(X)$. If we assume temporarily that $X$ and $g(X)$ have probability density functions $f$ and $h$, then

$$\mathbb{P}(g(X) \leq y) = \mathbb{P}(X \leq g^{-1}(y)) = \int_{-\infty}^{g^{-1}(y)} f(s) \, ds$$

$$= \int_{-\infty}^{y} h(t) \, dt.$$ 

Then

$$h(y) = \frac{d}{dy} \int_{-\infty}^{y} h(t) \, dt$$

$$= \frac{d}{dy} \int_{-\infty}^{g^{-1}(y)} f(s) \, ds$$

$$= \frac{dg^{-1}(y)}{dy} f(g^{-1}(y)).$$

Thus we can determine $h$. In particular, if $g(x) = \frac{x-\mu}{\sigma}$, then

$$h(y) = \sigma f(\mu + \sigma y).$$

This can be used to prove the following general result.

**Theorem 22.** Suppose that $X$ is a (continuous) random variable, and that $\mu = \mathbb{E}(X)$ and $\text{Var}(X) = \sigma^2 = \mathbb{E}(X^2) - \mathbb{E}(X)^2 < \infty$. Then if $Y = \frac{X-\mu}{\sigma}$,

$$\mathbb{E}(Y) = 0$$

$$\text{Var}(Y) = 1.$$
This theorem also holds for discrete random variables.

7. Independence

We say that random variables are independent if “one has no bearing on the other”. Mathematically, \(X\) and \(Y\) are independent if
\[
P(X \leq x, Y \leq y) = P(X \leq x) \cdot P(Y \leq y).
\]

If \(X\) and \(Y\) are not independent, then they are dependent. For example, if I toss two coins, and set \(X = 0\) or \(1\) as the first coin is heads or tails, and \(Y = 0\) or \(1\) as the second coin is heads or tails, then I presume that the possible outcomes are equally likely. Then \(X\) and \(Y\) are independent.

\[
\begin{array}{ccc}
X & 0 & 1 \\
Y \backslash & & \\
0 & 0.25 & 0.25 \\
1 & 0.25 & 0.25 \\
\end{array}
\]

On the other hand, height and weight (of a randomly chosen person) are dependent. If I measure a “random person”, and set \(X = 0\) if they weigh less than 60 kilos, and \(X = 1\) if they weigh more, and set \(Y = 0\) or \(1\) as they are less or more than 1650 mm in height, then I find empirically relative frequencies like those shown:

\[
\begin{array}{ccc}
X & 0 & 1 \\
Y \backslash & & \\
0 & 0.2 & 0.2 \\
1 & 0.2 & 0.4 \\
\end{array}
\]

You can easily convince yourself that \(X\) and \(Y\) are dependent.

**Theorem 23.** *If \(X\) and \(Y\) are independent, then*
\[
\begin{align*}
\mathbb{E}(X + Y) &= \mathbb{E}(X) + \mathbb{E}(Y) \\
\text{Var}(X + Y) &= \text{Var}(X) + \text{Var}(Y).
\end{align*}
\]

Actually, the first of these always holds, but the second does not. The proof of this theorem involves integration in two variables.
3. STATISTICS

As a corollary of this, if \( X_1, X_2, \ldots, X_n \) are independent, identically distributed random variables, with mean \( \mu \) and variance \( \sigma^2 \),

\[
S_n = X_1 + X_2 + \cdots + X_n
\]

is a random variable with mean \( n\mu \) and variance \( n\sigma^2 \), and from Theorem 22,

\[
A_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)
\]

is a random variable with mean \( \mu \) and variance \( \sigma^2 / n \), and then \( A_n - \mu \) is a random variable with mean 0 and variance \( \sigma^2 / n \), and finally

\[
\frac{A_n - \mu}{\sigma/\sqrt{n}} = \frac{S_n - n\mu}{\sigma/\sqrt{n}}
\]

is a random variable with mean 0 and variance 1. The amazing fact is that these “normalized averages” can be described very well by the normal distribution.

**Theorem 24 (The Central Limit Theorem).** Suppose that \( X_1, X_2, \ldots, X_n \) are independent, identically distributed random variables, with mean \( \mu \) and variance \( \sigma^2 \), and that \( S_n \) denotes \( X_1 + \cdots + X_n \). Then

\[
\lim_{n \to \infty} \mathbb{P} \left( a \leq \frac{S_n - n\mu}{\sqrt{n}\sigma} \leq b \right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-z^2/2} \, dz = \Phi(b) - \Phi(a).
\]

**Proof.** We will consider the proof of this in one relatively simple case. Suppose that \( X \) is a discrete random variable, \( \mathbb{P}(X = 1) = p \) and \( \mathbb{P}(X = 0) = q = 1 - p \). Then \( \mathbb{E}(X) = p \) and \( \mathbb{V}(X) = pq \). Suppose also that \( X_1, X_2, \ldots, X_n \) are independent random variables, each with the same distribution as \( X \), and set \( S_n = X_1 + X_2 + \cdots + X_n \). We know that

\[
\mathbb{P}(S_n = k) = \binom{n}{k} p^k q^{n-k}.
\]

Fix \( a, b \in \mathbb{R} \), with \( a < b \). Then

\[
\mathbb{P} \left( a < \frac{S_n - np}{\sqrt{npq}} < b \right) = \sum_{k:a<\frac{S_n - np}{\sqrt{npq}}<b} \binom{n}{k} p^k q^{n-k}
\]

\[
= \sum_{k:np+a\sqrt{npq}<k<np+b\sqrt{npq}} \binom{n}{k} p^k q^{n-k}.
\]

First, we replace \( \binom{n}{k} \) by an approximation: Stirling’s formula says that

\[
n! \approx \sqrt{2\pi n} n^{n+1/2} e^{-n}.
\]

Then

\[
\binom{n}{k} \approx \frac{1}{\sqrt{2\pi}} \frac{n^{n+1/2}}{k^{k+1/2}(n-k)^{n-k+1/2}} \frac{e^{-n}}{e^{-k}e^{-(n-k)}}
\]

\[
= \frac{n^{1/2}}{\sqrt{2\pi k(n-k)}} \frac{1}{\sqrt{2\pi}} \frac{e^{-n}}{k^{1/2}(n-k)^{1/2}} \left( \frac{n}{n-k} \right)^{n-k},
\]
and so
\[\binom{n}{k} p^k q^{n-k} \approx \frac{n^{1/2}}{[2\pi k(n-k)]^{1/2}} \left(\frac{np}{k}\right)^k \left(\frac{nq}{n-k}\right)^{n-k}.\]

We put this onto the sum, and replace it by an integral:
\[
P\left(a < \frac{S_n - np}{\sqrt{npq}} < b\right) \approx \int_{np+a/\sqrt{npq}}^{np+b/\sqrt{npq}} \frac{n^{1/2}}{[2\pi k(n-k)]^{1/2}} \left(\frac{np}{k}\right)^k \left(\frac{nq}{n-k}\right)^{n-k} \, dk.
\]

Now we change variables: set \(z = \frac{k-np}{\sqrt{npq}}\), i.e. \(k = np + z \sqrt{npq}\), to get
\[
P\left(a < \frac{S_n - np}{\sqrt{npq}} < b\right) \approx \int_a^b \frac{1}{\sqrt{2\pi}} \left(\frac{np}{k}\right)^{k+1/2} \left(\frac{nq}{n-k}\right)^{n-k+1/2} \, dz;
\]

it suffices to show that
\[
\lim_{n \to \infty} \left(\frac{np}{k}\right)^{k+1/2} \left(\frac{nq}{n-k}\right)^{n-k+1/2} = e^{-z^2/2}.
\]

To do this we change the exponents \(k\) and \(n-k\) into expressions involving \(z\):
\[
\left(\frac{np}{k}\right)^{k+1/2} \left(\frac{nq}{n-k}\right)^{n-k+1/2} = \left(1 + z \sqrt{\frac{q}{np}}\right)^{-np-z \sqrt{npq}-1/2} \left(1 - z \sqrt{\frac{p}{nq}}\right)^{-nq+z \sqrt{npq}-1/2}.
\]

A tricky limit argument shows that this expression tends to \(e^{-z^2/2}\) as \(n \to \infty\). \(\Box\)

Problem 36. Assume that the population has height which is normally distributed with mean 165 and standard deviation 15. What proportion of the population is over 195 in height?

Answer. Let \(X\) denote the random variable height, and \(Z\) denote \(\frac{X-165}{15}\). Then
\[
P(X > 195) = P(X - 165 > 195 - 165 = 30)
= P\left(\frac{X - 165}{15} > \frac{30}{15} = 2\right)
= P(Z > 2).
\]

Now \(Z\) is approximately \(N(0, 1)\), so \(P(Z > 2) = 1 - P(Z \leq 2) = 1 - \Phi(2) = 0.0228\).
\(\triangle\)

Problem 37. A coin is tossed 1000 times. Using the normal approximation, find (approximately) the probability that the number of heads is between 400 and 600.

Answer. Let \(X\) denote the number of heads. Then
\[
P(400 \leq X \leq 600) = \sum_{k=400}^{600} \binom{1000}{k} \left(\frac{1}{2}\right)^{1000}.
\]
Put \( Z = \frac{X - 500}{\sqrt{1000 \cdot \frac{1}{2}}} = \frac{X - 500}{5\sqrt{10}} \). Then \( Z \sim N(0, 1) \). Further,

\[
P(400 \leq X \leq 600) = P(-100 \leq X - 500 \leq 100)
= P\left(\frac{-100}{5\sqrt{10}} \leq \frac{X - 500}{5\sqrt{10}} \leq \frac{100}{5\sqrt{10}}\right)
= P(-2\sqrt{10} \leq Z \leq 2\sqrt{10})
= 2\Phi(2\sqrt{10}) - 1
\approx 1
\]
to 4 decimal places. \( \triangle \)

**Problem 38.** Assume that the probability that a random person is born in October is \( \frac{1}{12} \). Estimate the probability that, in a random group of 100 people, at most 5 have birthdays in October.

**Answer.** Let \( X \) be the random variable which is 1 if someone is born in October and 0 otherwise. Then

\[
\mathbb{E}(X) = \frac{1}{12}, \\
\mathbb{V}(X) = \frac{11}{144}.
\]

Let \( X_i \) be 1 or 0 as person \( i \) is born in October or not, and write \( S = S_{100} = X_1 + \cdots + X_{100} \). Then if \( Z \) is a standard normal variable,

\[
P(S \leq 5) = P\left(S - \frac{100}{12} \leq 5 - \frac{100}{12}\right)
= P\left(S - \frac{100}{12} \leq \frac{5 - 100}{\sqrt{\frac{1100}{144}}}\right)
\approx P\left(Z \leq -\frac{40}{\sqrt{1100}}\right)
= P\left(Z \leq -\frac{4}{\sqrt{11}}\right)
= P\left(Z \geq \frac{4}{\sqrt{11}}\right)
= 1 - P\left(Z < \frac{4}{\sqrt{11}}\right)
\approx 1 - P(Z < 1.2060 \cdots)
= 1 - .8657
= 0.1343,
\]
Problem 39. 100 balloons filled with helium are tied around a car dealer’s yard. The probability that any given balloon bursts in a one-hour period is 0.01. What is the probability that 5 or more balloons burst in an hour? Can you estimate the probability that at most 10 balloons burst in the next eight-hour period?

**Answer.** The probability that at least 5 balloons burst in an hour is

\[ \sum_{k=5}^{100} \binom{100}{k} (0.01)^k (0.99)^{100-k}. \]

A more workable expression is 1 minus the probability that at most 4 balloons burst in an hour, which is

\[ 1 - \sum_{k=0}^{4} \binom{100}{k} (0.01)^k (0.99)^{100-k}. \]

This is numerically messy, but is about 0.0034.

We can also use the Poisson approximation: \( \lambda = np = 1 \). Then the probability that \( k \) balloons burst in 1 hour is approximately

\[ \frac{e^{-\lambda} (0.01)^k}{k!}, \]

whence the probability that at least 5 balloons burst is given by

\[ 1 - \Pr(\text{at most 4 balloons burst}) \]
\[ = 1 - \left( e^{-1} + e^{-1} + \frac{e^{-1}}{2!} + \frac{e^{-1}}{3!} + \frac{e^{-1}}{4!} \right) \]
\[ = 1 - e^{-1} \left( 2 + \frac{1}{2} + \frac{1}{6} + \frac{1}{24} \right) \]
\[ = 0.0037. \]

To deal with an 8-hour period, let us estimate the probability that 1 balloon lasts more than 8 hours. This will be approximately

\( (0.99)^8 = 0.9227 \cdots = p, \)

say. Then the probability that at most 10 balloons burst in an 8-hour period is

\[ \binom{100}{0} p^{100} (1-p)^0 + \binom{100}{1} p^{99} (1-p) + \cdots + \binom{100}{10} p^{90} (1-p)^{10}. \]

You could use a Poisson approximation here too, but it would be less accurate, though easier to compute.

**Problem 40.** A company making light globes believes that 0.05% of its product is defective. In a sample of 100 globes, 10 are defective. What is the probability of 10 or more globes being defective in a sample?
Answer. Here \( N = 100, p = 0.05 \) and \( Np = 5 \). The Poisson distribution will give a good approximation. Let \( X \) be the number of defective items in a sample of 100. Then

\[
P(X = k) \cong \frac{e^{-5}5^k}{k!},
\]

and so

\[
P(X \geq 10) = 1 - P(X \leq 9)
= 1 - e^{-5} \left[ 1 + 5 + \frac{25}{2!} + \cdots + \frac{5^8}{8!} + \frac{5^9}{9!} \right]
= 0.03.
\]

\( \triangle \)