PART G. PROBABILITY AND STATISTICS

Content. Data analysis, binomial, Poisson, hypergeometric, normal distributions (Chap. 24)
Random numbers, confidence intervals, tests, regression (Chap. 25)

Statistics package. Load it by typing `with(Statistics):` Type `Mean`, `Variance`, `Median`, etc. Type `?Distributions` to see the binomial, normal, Student's t-, Fisher's, chi-square, and other distributions. Type `CDF` for the cumulative distribution function and `Quantile` for the quantile.

Chapter 24

Data Analysis. Probability Theory

Content. Data analysis of samples (Exs. 24.1, 24.2, Prs. 24.1–24.5)
Discrete distributions (Ex. 24.3, Prs. 24.7–24.12)
Normal distribution (Ex. 24.4, Prs. 24.14, 24.15)

Probability theory provides models (probability distributions) of random experiments in this chapter and the mathematical justification of the statistical methods in the next chapter.

Examples for Chapter 24

EXAMPLE 24.1 DATA ANALYSIS: MEAN, VARIANCE, STANDARD DEVIATION

Given a sample S of size n, that is, consisting of n values (marks, for example),

$$S = [x_1, x_2, ..., x_n],$$

you can arrange these values in ascending or descending order by the command `sort`. For instance,

```
> S := [90, 85, 97, 91, 80, 83, 98, 88, 78, 84, 82, 99, 86, 91, 85];
> sort(%);          # Resp. [78, 80, 82, 83, 84, 85, 85, 86, 88, 90, 91, 91, 97, 98, 99]
> -sort(-S);        # Resp. [99, 98, 97, 91, 91, 90, 88, 86, 85, 85, 84, 83, 82, 80, 78]
```

The **sample mean** $\bar{x}$ is

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j. \tag{1}$$

This is the arithmetic mean of the sample values. It measures the average size of these values. You obtain it (for the above S) by use of the **Statistics package**,

```
> with(Statistics):
> n := Count(S);                                  # Resp. n := 15
> evalf[5](add(S[k], k = 1..n)/n);                # Resp. 87.800
```

or by typing (type `?Mean` for information)

```
> xbar := evalf[5](Mean(S));                      # Resp. xbar := 87.800
```

The **sample variance** s^2 is

$$s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - x)^2. \tag{2}$$

It measures the spread of the sample values. You obtain it (for the above S) by loading the statistics package and typing (type `?Variance` for information)

```
> evalf[5](add((S[k] - xbar)^2, k = 1..n)/(n - 1));
                                41.886
> var := evalf[5](Variance(S));                   # Resp. var := 41.885
```

The **sample standard deviation** is the nonnegative square root of the sample variance. Thus,

```
> sdev := sqrt(var);                              # Resp. sdev := 6.471862174
```

or directly (note that restricting `xbar` and `var` to 5S has resulted in a slight difference due to round-off),

```
> StandardDeviation(S);                           # Resp. 6.47191735776302
```

Similar Material in AEM: Sec. 24.1

EXAMPLE 24.2 DATA ANALYSIS: HISTOGRAMS, BOXPLOTS

These are graphical representations of data. A **histogram** of a sample shows the **absolute frequencies** $a(x)$ = *Number of times the value x occurs in that sample.* Actually, to obtain a better general impression of the essential features of the sample, ***group it into classes***. For instance, take the sample in Example 24.1 and order it,

```
> S := [90, 85, 97, 91, 80, 83, 98, 88, 78, 84, 82, 99, 86, 91, 85]:
> S := sort(%);
```

$$S := [78, 80, 82, 83, 84, 85, 85, 86, 88, 90, 91, 91, 97, 98, 99]$$

Now plot a histogram by typing (type `?Histogram` for information)

```
> with(Statistics):
```

We now group into 5 mark intervals

```
> Histogram(S, binbounds = [75, 80, 85, 90, 95, 100],
    frequencyscale = absolute, labels = [x," "], color = "LightBlue");
```

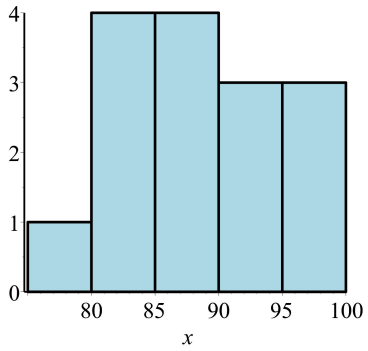
We can also see the groupings (values on the bin boundaries go to the 'higher' bin).

```
> TallyInto(S, Vector([75, 80, 85, 90, 95, 100]));
```

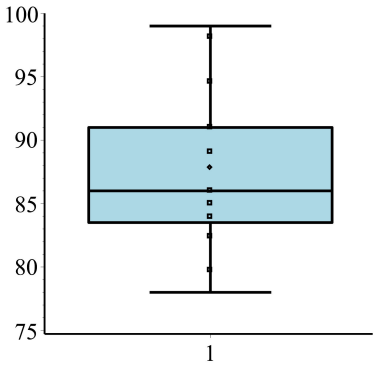
$$[75..80. = 1, 80..85. = 4, 85..90. = 4, 90..95. = 3, 95..100. = 3]$$

Boxplots. A boxplot of a sample shows the smallest sample value that is not an outlier (lower end of the plot), the **lower quartile** (bottom of the box), the **median**, the **upper quartile** (top of the box), and the largest sample value that is not an outlier (upper end of the plot). To get a boxplot for the above sample S, type the following. (Type `?BoxPlot`.)

```
> BoxPlot(S, view = [0.5..1.5, 75..100], color = "LightBlue");
```



Example 24.2. Histogram of the grouped sample



Example 24.2. Boxplot of the sample

Similar Material in AEM: Sec. 24.1

EXAMPLE 24.3 DISCRETE PROBABILITY DISTRIBUTIONS

In this chapter you will need the binomial, Poisson, and hypergeometric distributions, which are discrete, and the uniform and normal distributions, which are continuous. All these are available in Maple. Also the chi-square, t–, and F–distributions needed in Chap. 25 are available – along with other distributions. To see this, type `?Distributions`.

Discrete distributions are given by their **probability function** $f(x)$. You call these functions as shown below in terms of the distributions to be discussed.

Binomial distribution. The probability function is

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, ..., n.$$

This is the probability of x successes in n independent trials when the probability of success in a single trial is p. For instance, if $p = 0.55$ and $n = 5$, you obtain numeric values of f for $x = 0, ..., 5$ by typing

```
> with(Statistics):  Digits := 5:
```

```
> f := seq(ProbabilityFunction('Binomial'(5, 0.55), x), x = 0..5);
```

$$f := 0.018453, 0.11277, 0.27565, 0.33692, 0.20589, 0.050328$$

Note that `'Binomial'(5, 0.55)` is a short form for `RandomVariable(Binomial(5, 0.55)), x)`

Remember that individual values can be accessed by typing `f[1]`, `f[2]`, etc. For instance,

```
> f[4];                                                    # Resp. 0.33692
```

Note that this is $f(3)$ because x begins with 0, whereas counting terms of the sequence begins with 1, say, $j = 1$. Thus $x = j - 1$ in connection with the plotting below. This will give you the right numbers on the x-axis.

Similarly, you obtain values of the distribution function F by typing `CumulativeDistributionFunction` or `CDF`.

```
> F := seq(CDF('Binomial'(5, 0.55), x), x = 0..5);
```

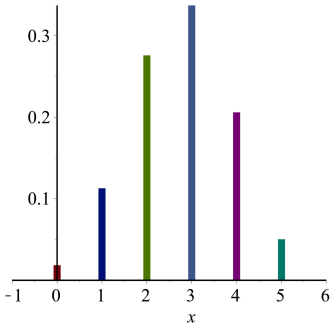
$$F := 0.018453,\ 0.13122,\ 0.40687,\ 0.74378,\ 0.94967,\ 1$$

Bar graphs of probability functions can be obtained by generating pairs of endpoints of the bars and then plotting them. Using the values $f(0) =$ `f[1]`, $f(1) =$ `f[2]`, ..., $f(5) =$ `f[6]`, type

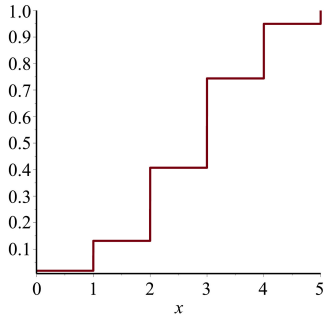
```
> s := seq([[j - 1, 0], [j - 1, f[j]]], j = 1..6);
```

$$s := [[0,0],[0,0.018453]],\ [[1,0],[1,0.11277]],\ [[2,0],[2,0.27565]],\ [[3,0],[3,0.33692]],\ [[4,0],[4,0.20589]],\ [[5,0],[5,0.050328]]$$

```
> plot(s, x = -1..6, thickness = 5);                       # A
```



Example 24.3.A Probability function of the binomial distribution with $n = 5$ and $p = 0.55$



Example 24.3.B Distribution function of the binomial distribution with $n = 5$ and $p = 0.55$

Graphs of distribution functions. In the discrete case, these are step functions, with the stepsizes equal to the values of the corresponding probability function. You can obtain the graph of the distribution function by typing

```
> plot(CDF('Binomial'(5, 0.55), x), x = 0..5);             # B
```

Poisson distribution. The probability function is

$$f(x) = \frac{\mu^x}{x!} e^{-\mu}, \qquad x = 0, 1, \ldots .$$

It is the limiting case of the binomial distribution if one lets $n \to \infty$ and $p \to 0$ so that the mean $\mu = np$ approaches a finite value. Its graph has infinitely many bars whose lengths decrease to zero very quickly. For instance, to plot f with $\mu = 5$, type

```
> poi := evalf(seq(ProbabilityFunction('Poisson'(5), x), x = 0..15));
```

$$poi := 0.0067379,\ 0.033690,\ 0.084224,\ 0.14037,\ 0.17547,\ 0.17547,\ 0.14622,\ 0.10444,\ 0.065277,\ 0.036265,\ 0.018132,\ 0.0082418,\ 0.0034342,\ 0.0013208,\ 0.00047173,\ 0.00015724$$

```
> bars := seq([[j - 1, 0], [j - 1, poi[j]]], j = 1..16);
```

$$bars := [[0,0],[0,0.0067379]],\ [[1,0],[1,0.033690]],\ [[2,0],[2,0.084224]],$$
$$[[3,0],[3,0.14037]],\ [[4,0],[4,0.17547]],\ [[5,0],[5,0.17547]],\ [[6,0],[6,0.14622]],$$
$$[[7,0],[7,0.10444]],\ [[8,0],[8,0.065277]],\ [[9,0],[9,0.036265]],\ [[10,0],[10,0.018132]],$$
$$[[11,0],[11,0.0082418]],\ [[12,0],[12,0.0034342]],\ [[13,0],[13,0.0013208]],$$
$$[[14,0],[14,0.00047173]],\ [[15,0],[15,0.00015724]]$$

```
> plot(bars, labels = [x, " "], thickness = 5);                    # C
```

Hypergeometric distribution. The probability function is

$$f(x) = \frac{\binom{M}{n}\binom{N-M}{n-x}}{\binom{N}{n}}$$

This is the probability of obtaining x red balls in drawing n balls from a lot of N balls, M of which are red, and the drawing is done one-by-one ***without replacement*** (that is, balls drawn are not returned to the lot). Instead of red balls you could think of defective screws in a lot of screws produced. For instance, when $N = 100$ and $M = 20$, and you draw 5 balls, you can plot f by typing the following. Type `?plot[options]`.) Type the following.

```
> N := 100:  M := 20:
```

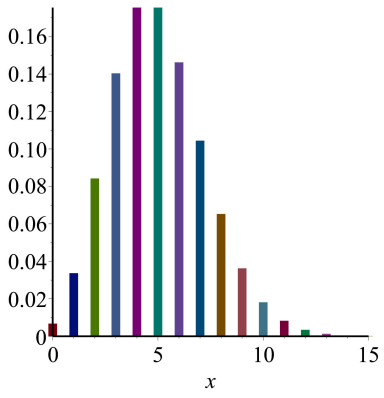
```
> hyp := evalf(seq(ProbabilityFunction('Hypergeometric'(N, M, 5), x),
    x = 0..5));
```

$$hyp := 0.31931,\ 0.42014,\ 0.20734,\ 0.047849,\ 0.0051483,\ 0.00020593$$

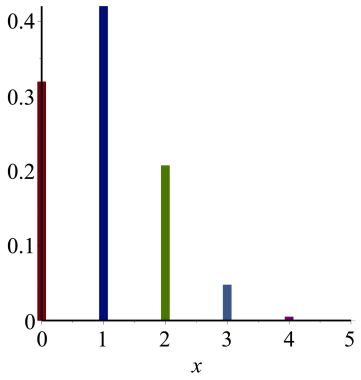
```
> barhyp := seq([[j - 1,0], [j - 1, hyp[j]]], j = 1..6);
```

$$barhyp := [[0,0],[0,0.31931]],\ [[1,0],[1,0.42014]],\ [[2,0],[2,0.20734]],$$
$$[[3,0],[3,0.047849]],\ [[4,0],[4,0.0051483]],\ [[5,0],[5,0.00020593]]$$

```
> plot(barhyp, labels = [x, " "], thickness = 5);                  # D
```



Example 24.3.C Probability function of the Poisson distribution with mean $\mu = 5$



Example 24.3.D Probability function of the hypergeometric distribution with $N = 100$, $M = 20$, and $n = 5$

Similar Material in AEM: Sec. 24.7

EXAMPLE 24.4 NORMAL DISTRIBUTION

The **standardized normal distribution** (that is, the normal distribution with mean 0 and variance 1) is a **continuous distribution** with **density**

$$\text{(A)} \qquad f(x) = \frac{1}{\sqrt{2\,\pi}} \exp\left(-\frac{x^2}{2}\right) \qquad (-\infty < x < \infty).$$

Hence its distribution function is

$$\text{(B)} \qquad \Phi(z) = \frac{1}{\sqrt{2\,\pi}} \int_{-\infty}^{z} e^{-u^2/2}\, du.$$

Maple knows this function, as well as its inverse (see below), and various numeric tables of $\Phi(z)$ can be found in the literature.

Setting $u = (v - \mu)/\sigma$, you have $du = dv/\sigma$ and $z = (x - \mu)/\sigma$. You thus obtain the distribution function of the normal distribution with mean μ and variance σ^2

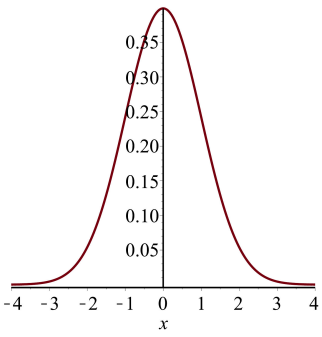
$$\text{(C)} \qquad F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sigma\sqrt{2\,\pi}} \int_{-\infty}^{x} \exp\left[-\frac{1}{2}\left(\frac{v-\mu}{\sigma}\right)^2\right] dv.$$

To plot the density (A), type the following, which shows that `PDF` denotes the density (‘**probability density function**’) and `'Normal'`(μ, σ) the **normal distribution.** (Type `?Distributions`.)

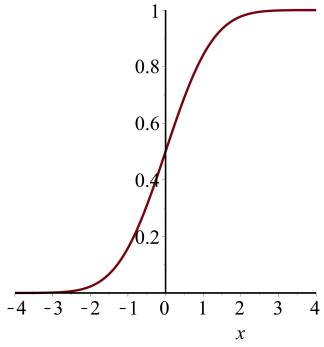
```
> with(Statistics):
> plot(PDF('Normal'(0, 1), x), x = -4..4);
```

To plot the distribution function (B), type

```
> plot(CDF('Normal'(0, 1), x), x = -4..4);
```



Example 24.4 Density (A) of the standardized normal distribution



Example 24.4 Distribution function (B) of the standardized normal distribution

Two basic tasks in connection with any distribution. Illustration in terms of a random variable X that is normal with mean 5.0 and variance 0.09.

First task. Given $x = 5.4$, find the probability that, in a trial (an experiment), X will assume any value not exceeding 5.4.

Solution. You need the **distribution function** `CDF`. Note that `'Normal'` involves the standard deviation 0.3, not the variance. The answer is 90.9%, approximately.

```
> CDF('Normal'(5.0, 0.3), 5.4);                    # Resp. 0.908788780274132
```

Second task. Given the probability $P = 0.95 = 95\%$, find $x = c$ such that, with probability 0.95, the random variable X will assume any value not exceeding c. (This task will arise in Chap. 25 for several distributions.)

Solution. You need the **inverse** of `CDF` i.e. `Quantile` of the distribution function . The answer is $x = c = 5.49$, approximately.

```
> Quantile('Normal'(5.0, 0.3), 0.95);                # Resp. 5.4935
```

Similar Material in AEM: Sec. 24.8

Problem Set for Chapter 24

Pr.24.1 (Mean and variance) Find the mean filling [grams] and the variance of the sample of fillings 204, 197, 194, 206, 198, 203, 203. (*AEM* Sec. 24.1)

Pr.24.2 (Mean, median, standard deviation) Find the mean, median, and standard deviation of the sample 4, 2, 4, 5, 3, 7, 5, 4. (*AEM* Sec. 24.1)

Pr.24.3 (Histogram, bar graph) Plot a histogram for the sample in Pr.24.1. For information type `?Histogram`. (*AEM* Sec. 24.1)

Pr.24.4 (Histogram, boxplot) Find the five-number summary and make a histogram and a boxplot of the sample in Pr.24.2. (*AEM* Sec. 24.1)

Pr.24.5 (Boxplot) Find the median and the other two quartiles of the sample in Pr.24.1 and make a boxplot. Type `Median`, `Quartile`, and `Quantile` to see the details of the commands not shown in Example 24.2 in this Guide. (*AEM* Sec. 24.1)

Pr.24.6 (Probability) A circuit contains 10 automatic switches. You want, with a probability of 95%, that during a given time interval all the switches are working. What probability of failure per time interval can you admit for a single switch? (*AEM* Sec. 24.3)

Pr.24.7 (Experiment on binomial distribution) For what values of p will $f(x)$, with constant n, be (a) large for small x, (b) large for large x? Experiment with plots and n's of your choice (10, 20, 100, or whatever). (*AEM* Sec. 24.7)

Pr.24.8 (Binomial distribution) Find and plot (as a bar graph) the probabilities of x successes in 40 independent trials with probability of success 1/2 in a single trial. Why is the figure symmetric? Does the figure remind you of something you have seen in connection with the normal distribution? For which x's are these probabilities very small? (*AEM* Secs. 24.7, 24.8)

Pr.24.9 (Permutations) In how many ways can you assign 9 workers to 9 jobs? First guess, then calculate. (*AEM* Sec. 24.4)

Pr.24.10 (Experiment on Stirling formula) A convenient approximation for (inconvenient) large factorials is $n! \approx \sqrt{2\,\pi\,n}\,\left(\frac{n}{e}\right)^n$. Find conjectures about the absolute error and the relative error for growing n by experimentation. (*AEM* Sec. 24.4)

Pr.24.11 (Poisson distribution) Investigate the Poisson distribution graphically. What happens if you let μ increase? μ is the mean as well as the variance. Do the graphs give you that impression? Does the distribution approach some kind of symmetry? For what μ? (*AEM* Sec. 24.7)

Pr.24.12 (Hypergeometric distribution) If a carton of 30 fuses contains 5 defectives and 4 fuses are randomly drawn from it, without replacement, what are the probabilities of obtaining 0, 1, 2, 3, 4 defective fuses? (*AEM* Sec. 24.7)

Pr.24.13 (Uniform distribution) The uniform distribution on an interval $a \leq x \leq b$ has the density $f(x) = 1/(b-a)$ if $a < x < b$ and 0 otherwise. Find the mean and the variance on the computer by using the definitions. (*AEM* Sec. 24.6)

Pr.24.14 (Normal distribution) Let X be normal with mean 116 and variance 36. Find $P(X < 122.5)$, $P(X > 110)$, $P(120.5 < X < 121.25)$ on the computer. (*AEM* Sec. 24.8)

Pr.24.15 (Normal distribution) If sick-leave time X used, by employees of some company, in one month is (very roughly) normal with mean 1000 hours and standard deviation 100 hours, how much time t should be budgeted for sick leave during the next month if t is to be exceeded with a probability of only 20%? (*AEM* Sec. 24.8)