Chapter 25

Mathematical Statistics

Content. Random numbers, samples (Ex. 25.1, Pr. 25.1)
Confidence intervals (Exs. 25.2–25.4, Prs. 25.2–25.5)
Tests (Exs. 25.5–25.10, Prs. 25.6–25.13)
Regression (Ex. 25.11, Prs. 25.14, 25.15)

The **binomial**, **Poisson**, **hypergeometric**, and **normal distributions**, discussed in Chap. 24 of this Guide, will occur again. For their Maple commands see Examples 24.3 and 24.4. Type `?Distributions`. New distributions in this chapter are **Student's *t*-distribution** (see Example 25.3), the **chi-square distribution** (Example 25.4), and **Fisher's *F*-distribution** (Example 25.9; also known as **variance-ratio distribution**). These will be needed in connection with confidence intervals, tests, etc. Commands (in terms of typical examples) are as follows.

```
> with(Statistics):
> evalf[6](Quantile('StudentT'(5), 0.95));          # Resp. 2.01504
> evalf[6](CDF('StudentT'(5), 2.0150));             # Resp. 0.949997
> evalf[6](Quantile('ChiSquare'(15), 0.99));        # Resp. 30.5779
> evalf[6](CDF('ChiSquare(15)', 30.578));           # Resp. 0.990000
> evalf[6](Quantile('FRatio'(7, 12), 0.95));        # Resp. 2.91336
> evalf[6](CDF('FRatio'(7, 12), 2.9134));           # Resp. 0.950002
```

Explanations. For Student's t-distribution with 5 df (**degrees of freedom**) the 95%-point is at 2.015.... That is, the distribution function F satisfies $F(2.015...) = 0.95$. For the chi-square distribution with 15 df the 99%-point is at 30.577... For Fisher's F-distribution with (7, 12) df the 95%-point is at 2.913...

Examples for Chapter 25

EXAMPLE 25.1 RANDOM NUMBERS

Random numbers can be used for obtaining samples (this always means ***random samples***, by definition) from populations. Type `?rand`. Suppose you want to draw a sample of 20 items from a population of 90 items (screws, animals, humans or whatever). Number the items and type

```
> rn := rand(1..90):
> seq(rn(j), j = 1..20);
      45, 6, 59, 44, 38, 69, 27, 17, 90, 34, 18, 52, 56, 43, 83, 25, 90, 60, 14, 50
```

If you call the generator again, you will get another sample

```
> seq(rn(j), j = 1..20);
      47, 8, 46, 44, 9, 77, 59, 16, 1, 70, 77, 39, 71, 67, 78, 51, 53, 12, 19, 63
```

and so on. If you just want a single random number (a single item), type

```
> rn();                                                    # Resp. 40
```

If you have reasons to obtain the same sample (or sequence of samples) with your generator, you can use the following.

```
> with(RandomTools):
> SetState(state = 1234567):
> Ri1 := Generate(list(posint(range = 90), 20));
    Ri1 := [20, 71, 70, 14, 79, 36, 76, 46, 3, 6, 28, 31, 48, 84, 19, 72, 23, 36, 34, 33]
> Ri2 := Generate(list(posint(range = 90), 20));
    Ri2 := [88, 47, 87, 85, 43, 39, 59, 13, 1, 16, 51, 23, 77, 26, 77, 38, 80, 41, 77, 40]
> SetState(state = 1234567):
> Generate(list(posint(range = 90), 20));
        [20, 71, 70, 14, 79, 36, 76, 46, 3, 6, 28, 31, 48, 84, 19, 72, 23, 36, 34, 33]
```

Note that you must set the state before generating the next set of numbers.

Mean, variance, and standard deviation of samples are obtained as explained in Example 24.1 in this Guide. They vary from sample to sample.

```
> with(Statistics):
> n := Count(Ri1);                                  # Resp. n := 20
> evalf[6](add(Ri1[i], i = 1..n)/n);                # Resp. 41.4500
> m1 := evalf(Mean(Ri1));                           # Resp. m1 := 41.4500
> evalf[6](Mean(Ri2));                              # Resp. 50.4000
> evalf[6](add((Ri1[i] - m1)^2, i = 1..n)/(n - 1)); # Resp. 650.995
> evalf[6](Variance(Ri1));                          # Resp. 651.0
> evalf[6](Variance(Ri2));                          # Resp. 733.619
> evalf[6](sqrt(Variance(Ri1)));                    # Resp. 25.5147
> evalf[6](StandardDeviation(Ri1));                 # Resp. 25.5147
> evalf[6](StandardDeviation(Ri2));                 # Resp. 27.0854
```

Similar Material in AEM: Sec. 25.1

EXAMPLE 25.2 CONFIDENCE INTERVAL FOR THE MEAN OF THE NORMAL DISTRIBUTION WITH KNOWN VARIANCE

Find a confidence interval for the mean of the normal distribution with known variance $\sigma^2 = 9$. Use a sample of 100 values with sample mean $\bar{x} = 5$. Choose the confidence level $\gamma = 95\%$.

Solution. Regard $\bar{x}$ as an observed value of a random variable $\bar{X} = (1/n)(X_1 + \ldots + X_n)$, where $X_1, \ldots, X_n$ are independent random variables all having the same

distribution of some random variable X. It can be shown that if X is normal with mean μ and variance σ^2, then $\bar{X}$ is normal with mean μ and variance σ^2/n.

Type the given data, denoting the confidence level γ by g.

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

```
> n:= 100:  xbar := 5:  var := 9:  g := 0.95:
```

Consider the standardized normal distribution `Normal(0, 1)` (see Example 24.4). You obtain the shortest interval on the x-axis corresponding to $\gamma = 95\%$ of the area under the density curve (that is, corresponding to the probability 0.95) if you choose that interval symmetrically located with respect to the mean 0. Then its endpoints $-c$ and $+c$ correspond to the probabilities 2.5% and 97.5%. You get the latter by typing

```
> X := RandomVariable(Normal(0, 1)):
```

```
> c := evalf[6](Quantile('Normal'(0, 1) ,0.975));  # Resp. c := 1.95996
```

```
> CDF('Normal'(0, 1), 1.9600);                     # Resp. 0.975002104851780
```

For $\bar{X}$ this corresponds to $\mu - k$ and $\mu + k$, where $k = c\sigma/\sqrt{n}$. Accordingly, type

```
> k := c*sqrt(var)/sqrt(n);                        # Resp. k := 0.5879880000
```

You now get the confidence interval by replacing $\mu - k$ with $\bar{x} - k$ and $\mu + k$ with $\bar{x} + k$

```
> conf1 := xbar - k;                               # Resp. conf1 := 4.412012000
```

```
> conf2 := xbar + k;                               # Resp. conf2 := 5.587988000
```

The confidence interval is $\text{CONF}_{0.95}(4.41 \le \mu \le 5.59)$.

In this approach you used the standardized normal distribution, as though it has been tabulated. On the computer you can proceed more directly by noting that the midpoint of the confidence interval is $\bar{x}$ and the endpoints are the 2.5%-point and the 97.5%-point of the distribution of $\bar{X}$, which is normal with variance $\sigma^2/n = 9/100$, hence with standard deviation $\sigma/\sqrt{n} = 0.3$. You thus obtain directly

```
> Quantile('Normal'(xbar, 0.3), 0.025);            # Resp. 4.41201080463817
```

```
> Quantile('Normal'(xbar, 0.3), 0.975);            # Resp. 5.58798919536183
```

Similar Material in AEM: Sec. 25.3

EXAMPLE 25.3 CONFIDENCE INTERVAL FOR THE MEAN OF THE NORMAL DISTRIBUTION WITH UNKNOWN VARIANCE. *t*-DISTRIBUTION

Find a confidence interval for the mean of the normal distribution with unknown variance σ^2. Use the sample 242, 251, 248, 245, 250, 247, 244. Choose the confidence level $\gamma = 99\%$.

Solution. Regard $\bar{x}$ as an observed value of a random variable $\bar{X}$, as in the previous example. $k = c\sigma/\sqrt{n}$ can no longer be used because σ is no longer known. The idea is to replace σ by the sample standard deviation s as defined in Example 24.1 in this Guide, and to regard s as an observed value of a random variable S. To use S, one must know its probability distribution. Student (pseudonym for W. S. Gosset) has shown that if the population random variable X is normal with mean μ and variance σ^2, then $t = \dfrac{\bar{x} - \mu}{s/\sqrt{n}}$ is an observed value of a random variable T which has a ***t-distribution*** with $n - 1$ **degrees of freedom (df)**, and he gave the formula for the probability density of T, which is symmetric with respect to 0. Accordingly, type the given data, and calculate $\bar{x}$ and s.

```
> sample := [242, 251, 248, 245, 250, 247, 244]:
```

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

```
> n := Count(sample);                                  # Resp. n := 7
```

```
> xbar := Mean(sample);                  # Resp. xbar := 246.714285714286
```

```
> s := StandardDeviation(sample);          # Resp. s := 3.25137333621172
```

Now obtain the 99%-point for the t-distribution with $n - 1$ degrees of freedom and calculate $K = Cs/\sqrt{5n}$, the counterpart of k in the previous example.

```
> C := Quantile('StudentT'(n - 1), 0.99); # Resp. C := 3.14263811305830
```

```
> K := evalf[4](C*s/sqrt(n));                          # Resp. K := 3.863
```

You now get the confidence interval by replacing $\mu - K$ with $\bar{x} - K$ and $\mu + K$ with $\bar{x} + K$

```
> conf1 := xbar - K;                  # Resp. conf1 := 242.851285714286
```

```
> conf2 := xbar + K;                  # Resp. conf2 := 250.577285714286
```

This gives the confidence interval $\text{CONF}_{0.99}(242.85 \le \mu \le 250.58)$. This is rather large, but keep in mind that your sample was small and σ was unknown. If it were known and equal to s, you should get a shorter interval because you use more information. Can you calculate this?

Similar Material in AEM: Sec. 25.3

EXAMPLE 25.4 CONFIDENCE INTERVAL FOR THE VARIANCE OF THE NORMAL DISTRIBUTION. χ^2-DISTRIBUTION

Find a confidence interval for the unknown variance σ^2 of the normal distribution, using the following sample and choosing the confidence level $\gamma = 95\%$. (The mean μ need not be known.)

89 84 87 81 89 86 91 90 78 89 87 99 83 89

Solution. It can be shown that, under the normality assumption, the quantity

$$y = (n-1)s^2/\sigma^2$$

is an observed value of a random variable Y that has a **chi-square distribution** with $n-1$ degrees of freedom. Here, s^2 is the sample variance, as before. Type the sample and then $(n-1)s^2$.

```
> with(Statistics):
> sample := [89, 84, 87, 81, 89, 86, 91, 90, 78, 89, 87, 99, 83, 89]:
> n := Count(sample);                                  # Resp. n := 14
> nssquare := (n - 1)*Variance(sample);    # Resp. nssquare := 326.857...
```

Determine the 2.5%-point and the 97.5%-point of the chi-square distribution (which is not symmetric) with $n-1 = 13$ degrees of freedom

```
> c1 := Quantile('ChiSquare'(n-1), 0.025);      # Resp. c1 := 5.00875...
> c2 := Quantile('ChiSquare'(n-1), 0.975);      # Resp. c2 := 24.7356...
```

From this you obtain the endpoints of the confidence interval

```
> conf1 := evalf[4](nssquare/c1);                # Resp. conf1 := 65.26
> conf2 := evalf[4](nssquare/c2);                # Resp. conf2 := 13.21
```

This gives the confidence interval $\text{CONF}_{0.95}(13.21 \le \sigma^2 \le 65.26)$.

Similar Material in AEM: Sec. 25.3

EXAMPLE 25.5 TEST FOR THE MEAN OF THE NORMAL DISTRIBUTION

You want to buy 500 coils of wire. Test the manufacturer's claim that the breaking limit X of the wire is $\mu = \mu_0 = 200$ lb (or more). Assume that X has a normal distribution.

Solution. Test the **hypothesis** $\mu = \mu_0 = 200$ against the **alternative** $\mu = \mu_1 < 200$, an undesirable weakness. Hence this test is **left-sided**, the **rejection region** extends from a **critical point** c to the left.

To obtain a sample, select some of the coils, say, 25, at random. Cut a piece from each coil and determine the breaking limit experimentally. Suppose that this sample of $n = 25$ values has the mean $\bar{x} = 197$ lb (somewhat less than the claim!) and the standard deviation $s = 6$ lb. Then (as in Example 25.3 in this Guide)

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{197 - 200}{6/\sqrt{25}} = -2.50$$

is an observed value of a random variable T that has a $t-$distribution with $n-1$ degrees of freedom. Choose a **significance level** of the test, say $\alpha = 5\%$. Find the critical c by typing (similarly as in Example 25.3 in this Guide)

```
> with(Statistics):
> n := 25:
> c := Quantile(RandomVariable(StudentT(n - 1)), 0.05);
```

$$c := -1.71088207939271$$

Reason as follows. If the hypothesis is true, the probability of obtaining a $t < c$ is very small, namely, equal to $\alpha = 5\%$, so that it would happen only about once in 20 tests. Hence if it happens, as in the present case, where $-2.5 < -1.7109$, cast doubt on the truth of the hypothesis. Hence **reject the hypothesis** and assert that $\mu < 200$ and the manufacturer had promised too much.

Similar Material in AEM: Sec. 25.4

EXAMPLE 25.6 TEST FOR THE MEAN: POWER FUNCTION

You make an **error of the first kind** if you reject a hypothesis although it is true. You do this with probability α, the significance level of the test. You make an **error of the second kind** if you accept a hypothesis although the alternative is true. The corresponding probability is denoted by β. The quantity $\eta = 1 - \beta$ (thus the probability of avoiding an error of the second kind) is called the **power** of the test. β depends on the alternative μ. One calls $\beta(\mu)$ the **operating characteristic (OC)** and $\eta(\mu) = 1 - \beta(\mu)$ the **power function** of the test.

Let X be normal with mean μ and known variance $\sigma^2 = 9$. Then $\bar{X}$ is normal with mean μ and variance σ^2/n, where n is the size of the sample used in the test (see Example 25.2 in this Guide). Let the hypothesis be $\mu_0 = 24$. Choose $\alpha = 5\%$. If $n = 10$ then the standard deviation of $\bar{X}$ is $sd = 3/\sqrt{10}$.

1. Left-sided test (as in the previous example). The **critical region** (rejection region) extends from the critical $c = c_1$ to the left. Obtain the critical c_1 by typing

```
> with(Statistics):  Digits := 5:
> n := 10:  mu0 := 24:  var := 9:  sd := sqrt(var/n):
> c1 := Quantile('Normal'(mu0, sd), 0.05);          # Resp. c1 := 22.440
```

The power is the area under the density curve of $\bar{X}$ with the alternative μ being true, from $-\infty$ to c_1. This curve, and hence the area, depends on μ. The power is practically 1 at $\mu = 20$, decreases monotone to 0.05 at $\mu = 24$ and practically to 0 at $\mu = 26$; see the figure.

```
> powerleft := CDF('Normal'(mu, sd), c1);
```

$$powerleft := 0.50000 - 0.50000\,\mathrm{erf}\,(-16.725 + 0.74536\,\mu)$$

2. Right-sided test. The critical region now extends from the critical $c = c_2$ to the right. Obtain the critical $c = c_2$ by typing

```
> c2 := Quantile('Normal'(mu0, sd), 0.95);          # Resp. c2 := 25.560
```

The power is the area under the density curve of $\bar{X}$ with the alternative μ being true, from c_2 to ∞, and again depends on μ.

```
> powerright := 1 - CDF('Normal'(mu, sd), c2);
```

$$powerright := 0.50000 + 0.50000\,\mathrm{erf}\,(-19.052 + 0.74536\,\mu)$$

3. Two-sided test. The critical region now consists of two parts, from $-\infty$ to a lower critical point $c = c_{3a}$ and from an upper critical point $c = c_{3b}$ to ∞. These are the 2.5%- and 97.5%-points of the distribution of $\bar{X}$ with the hypothesis $\mu_0 = 24$ being true. Obtain these points by typing

```
> c3a := Quantile('Normal'(mu0, sd), 0.025);      # Resp. c3a := 22.141
```

```
> c3b := Quantile('Normal'(mu0, sd), 0.975);      # Resp. c3b := 25.859
```

The power is the area under the density curve of $\bar{X}$ with the alternative being true, from $-\infty$ to c_{3a} and from c_{3b} to ∞. This is the U-shaped curve in the figure.

```
> powertwosided := CDF('Normal'(mu, sd), c3a) + 1
                   - CDF(RandomVariable(Normal(mu, sd)), c3b);
```

$$powertwosided := 1.0 - 0.50000\,\mathrm{erf}\,(-16.503 + 0.74536\,\mu) + 0.50000\,\mathrm{erf}\,(-19.274 + 0.74536\,\mu)$$

Plotting all three curves on common axes is now quite simple.

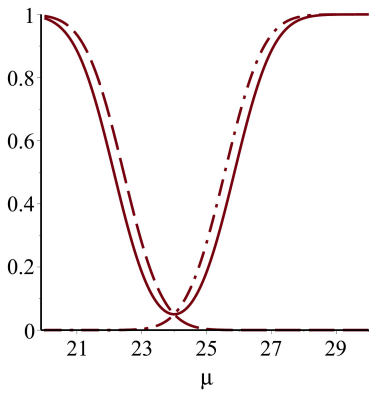
```
> P1 := plot(powerleft, mu = 20..30, linestyle = dash):
```

```
> P2 := plot(powerright, mu = 20..30, linestyle = dashdot):
```

```
> P3 := plot(powertwosided, mu = 20..30):
```

```
> with(plots):
```

```
> display(P1, P2, P3);
```



Example 25.6. Power functions of tests of $\mu_0 = 24$ against $\mu < 24$, $\mu > 24$, and $\mu \neq 24$

Similar Material in AEM: Sec. 25.4

EXAMPLE 25.7 TEST FOR THE VARIANCE OF THE NORMAL DISTRIBUTION

Using a sample of size $n = 15$ and sample variance $s^2 = 13$ from a normal population, test the hypothesis $\sigma^2 = \sigma_0^2 = 10$ against the alternative $\sigma^2 = \sigma_1^2 = 20$.

Solution. It can be shown that, under the normality assumption and under the

hypothesis, the quantity

$$y = (n-1)s^2/{\sigma_0}^2 = 14 \times 13/10 = 18.2$$

is an observed value of a random variable $Y = 14\,S^2/10$ that has a **chi-square distribution** with $n - 1 = 14$ degrees of freedom. (See also Example 25.4 in this Guide.) Because the alternative is greater than the hypothesis, the test is right-sided. Choose a significance level, say, $\alpha = 5\%$. Determine the 95%-point c of that distribution from

```
> with(Statistics):  n := 15:  Digits := 5:
> c := Quantile('ChiSquare'(n - 1), 0.95);          # Resp. c := 23.685
```

Because $y < c$ (and the test is right-sided), accept the hypothesis.

If the alternative is true, $Y_1 = (n-1)S^2/{\sigma_1}^2 = 14\,S^2/20 = Y/2$ has a chi-square distribution with $n - 1$ degrees of freedom. Hence the **power** is the area under the density curve of Y_1 from $c/2$ to ∞. That is,

```
> power := 1 - CDF('ChiSquare'(n - 1), c/2);
                          power := 0.618990111037540
```

This leaves a probability of 38% for committing an error of the second kind. This is too large, and you should repeat the test with a larger sample (if available!).

Similar Material in AEM: Sec. 25.4

EXAMPLE 25.8 COMPARISON OF MEANS

Test the hypothesis that two normal distributions with the same variance have the same mean, $\mu_1 = \mu_2$, against the alternative that they have different means. Choose the significance level $\alpha = 5\%$. Use two samples $x_1, x_2, ..., x_{n_1}$ and $y_1, y_2, ..., y_{n_2}$ which are independent. (Dependence would mean that some x-values are related to some y-values, for instance, if they came from the two front tires of the same car, from test scores of the same student, etc. Equality of variances will be tested in the next example.)

x:	64	63	47	58	57	56	62	52		
y:	34	40	31	45	59	61	68	47	48	32

Solution. Calculate the mean $\bar{x}$ and the variance s_x^2 of the first sample by typing

```
> with(Statistics):  Digits := 5:
> Sa1 := [64, 63, 47, 58, 57, 56, 62, 52]:
> n1 := Count(Sa1);                                  # Resp. n1 := 8
> xbar := Mean(Sa1);                                 # Resp. xbar := 57.375
> xvar := Variance(Sa1);                             # Resp. xvar := 33.697
> Sa2 := [34, 40, 31, 45, 59, 61, 68, 47, 48, 32]:
> n2 := Count(Sa2);                                  # Resp. n2 := 10
> ybar := Mean(Sa2);                                 # Resp. ybar := 46.500
> yvar := Variance(Sa2);                             # Resp. yvar := 164.71
```

It can be shown that, if the hypothesis is true, the quantity

$$t_0 = \sqrt{\frac{n_1\, n_2\, (n_1 + n_2 - 2)}{n_1 + n_2}}\, \frac{\bar{x} - \bar{y}}{\sqrt{(n_1 - 1)\, {s_x}^2 (n_2 - 1)\, {s_y}^2}}$$

is an observed value of a random variable T that has a t-distribution with $n_1 + n_2 - 2$ degrees of freedom. The test is two-sided. Because $\alpha = 5\%$, determine the 2.5%-point c_1 and the 97.5%-point c_2 of the t-distribution with $n_1 + n_2 - 2 = 8 + 10 - 2 = 16$ degrees of freedom by typing

```
> Quantile('StudentT'(n1 + n2 - 2), 0.025);          # Resp. −2.1199
> Quantile('StudentT'(n1 + n2 - 2), 0.975);           # Resp. 2.1199
```

If t_0 lies between these values (inclusively), accept the hypothesis. Otherwise reject it. Calculate

```
> t0 := evalf( sqrt(n1*n2*(n1 + n2 - 2)/(n1 + n2))*(xbar-ybar)/
   sqrt((n1 - 1)*xvar + (n2 - 1)*yvar) );
```

$$t0 := 2.2123$$

Reject the hypothesis and assert that the populations from which the samples were drawn have different means.

Similar Material in AEM: Sec. 25.4

EXAMPLE 25.9 COMPARISON OF VARIANCES. *F*-DISTRIBUTION

Test the hypothesis that the variances of the two normal distributions in the previous example are equal against the alternative that they are different. Choose the significance level 5%.

Solution. Type the samples and calculate their variances.

```
> with(Statistics):  Digits := 5:
> Sa1 := [64, 63, 47, 58, 57, 56, 62, 52]:
> n1 := Count(Sa1);                                   # Resp. n1 := 8
> xvar := Variance(Sa1);                          # Resp. xvar := 33.697
> Sa2 := [34, 40, 31, 45, 59, 61, 68, 47, 48, 32]:
> n2 := Count(Sa2);                                   # Resp. n2 := 10
> yvar := Variance(Sa2);                          # Resp. yvar := 164.71
```

It can be shown that, if the hypothesis is true, the ratio $v_0 = s_x^2/s_y^2$ is an observed value of a random variable V which has an ***F*-distribution** with $(n_1 - 1,\ n_2 - 1) = (7,\ 9)$ degrees of freedom. Because $\alpha = 5\%$ and the test is two-sided, determine the 2.5%-point c_1 and the 97.5%-point c_2 of this distribution by typing

```
> c1 := Quantile('FRatio'(n1 - 1, n2 - 1), 0.025);
```

$$c1 := 0.20733$$

```
> c2 := Quantile('FRatio'(n1 - 1, n2 - 1), 0.975); # Resp. c2 := 4.1970
```

Reject the hypothesis because v_0 does not lie between c_1 and c_2. Indeed,

```
> v0 := evalf(xvar/yvar);                             # Resp. v0 := 0.20458
```

Similar Material in AEM: Sec. 25.4

EXAMPLE 25.10 CHI-SQUARE TEST FOR GOODNESS OF FIT

With this test you find out how well the distribution of a sample fits the hypothetical distribution of the population. For instance, can you claim, on the 5%-level, that a die is fair if, in 20,000 trials, you obtain $x = 1, 2, ..., 6$ with the following absolute frequencies (actual classical data obtained by R. Wolf in Switzerland).

3407 3631 3176 2916 3448 3422

Solution. If the die is fair, each of the 6 values is equally likely, hence the expected absolute frequency is $e = 3333.33$. For each x calculate the observed value minus e, square it and divide the result by e. The sum of the 6 numbers thus obtained is an observed value of a random variable χ which is (asymptotically) chi-square distributed with $6 - 1 = 5$ degrees of freedom. Call this sum χ_0 . Clearly, it measures the discrepancy between observations and expectation. Accordingly, type

```
> with(Statistics):  Digits := 6:
> data := [3407, 3631, 3176, 2916, 3448, 3422]:
> n := Count(data);                                  # Resp. n := 6
> e := evalf(20000/n);                             # Resp. e := 3333.33
> chi0 := sum((data[j] - e)^2/e, j = 1..6);     # Resp. chi0 := 94.1890
```

The test is right-sided. The rejection region (critical region) extends from the critical c to the right and corresponds to a probability of 5%. Hence c is the 95%-point of the chi-square distribution with 5 degrees of freedom,

```
> c := Quantile('ChiSquare'(n - 1), 0.95);          # Resp. c := 11.0705
```

You see that χ_0 is much larger than c. Reject the hypothesis and assert that Wolf's die was not fair or there were flaws in throwing and/or counting.

Similar Material in AEM: Sec. 25.7

EXAMPLE 25.11 REGRESSION

In regression analysis you choose values $x_1, x_2, ..., x_n$ of an ordinary variable (for instance, x may be time) and observe corresponding values $y_1, y_2, ..., y_n$ of a random variable Y (for instance, temperature at some place). This gives a sample of n pairs $(x_1, y_1), ..., (x_n, y_n)$. You assume that the mean of Y depends linearly on x, say, $\mu(x) = \kappa_0 + \kappa_1 x$. This is the **regression line of the population.** Let the sample be (x = pressure in atmospheres, y = decrease of volume of leather in

x:	4000	6000	8000	10000
y:	2.3	4.1	5.7	6.9

From the sample you obtain the **sample regression line** $y = k_0 + k_1 x$. You fit this line through the given points (given pairs of coordinate values in the xy-plane) by the **least squares principle**, as follows.

For information type `?Fit`. Then type

```
> xSa := [4000, 6000, 8000, 10000]:
> ySa := [2.3, 4.1, 5.7, 6.9]:
> with(Statistics):  Digits := 6:
```

```
> line := Fit(a + b*x, xSa, ySa, x);
```

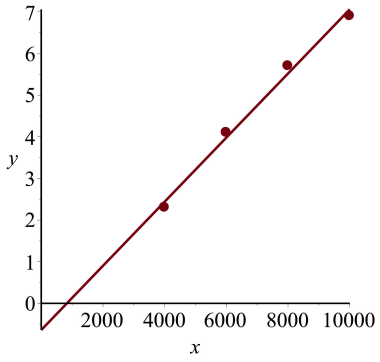
$$line := -0.640000 + 0.000770000000000000\,x$$

It is also possible to use

```
> with(CurveFitting):
> LeastSquares(xSa, ySa, x);          # Resp. −0.640000 + 0.000770000 x
```

This is the sample regression line. Obtain the figure by the subsequent commands. From the figure you see that for $x = 4000 \ldots 10000$ the line fits the data reasonably well (but would be useless near $x = 0$ because there is no decrease of volume when the pressure is 0).

```
> P1 := plot(line, x = 0..10000):
> P2 := plot(xSa, ySa, style = point, symbol= solidcircle,
    symbolsize = 20):
> with(plots):
> display(P1, P2, labels = [x, y]);
```



Example 25.11. Sample regression line and given data (four points)

The least squares principle is purely geometric. Probability enters through random variables. To obtain confidence intervals or tests for the **regression coefficient** κ_1, you must make assumptions about the probability distribution of Y. To apply the theory on the normal distribution (as in the previous examples), make the reasonable assumption that Y is normal and its variance σ^2 is the same for all x.

For obtaining a confidence interval for κ_1 you will need the variance of the x-values, the variance of the y-values, and the covariance. Accordingly, type

```
> n := Count(xSa);                          # Resp. n := 4
> xvar := Variance(xSa);                    # Resp. xvar := 6.66667 10^6
> yvar := Variance(ySa);                    # Resp. yvar := 3.98334
> cov := add((xSa[k] - Mean(xSa))*(ySa[k] - Mean(ySa)), k = 1..n)/n;
```

$$cov := 3850.0$$

```
> xycov := Covariance(xSa, ySa);            # Resp. xycov := 5133.33
```

Then type the formula for the sample regression coefficient k_1 and for an auxiliary quantity q_0.

```
> k1 := xycov/xvar;                         # Resp. k1 := 0.000769999
```

```
> q0 := (n - 1)*(yvar - k1^2*xvar);          # Resp. q0 := 0.09204
```

It can be shown that, under your assumptions, the random variable Y has a Student's t-distribution with $n - 2 = 2$ degrees of freedom. Choose a confidence level, say, 95%. Then type the commands for the 2.5%-point and the 97.5%-point of that distribution (actually, only the latter, c, because the former is $-c$, by the symmetry of the distribution).

```
> c := Quantile('StudentT'(n - 2), 0.975);          # Resp. c := 4.30265
```

Half the length of the confidence interval is

```
> k := c*sqrt(q0/((n - 2)*(n - 1)*xvar));          # Resp. k := 0.000206393
```

With this you obtain the confidence interval $\text{CONF}_{0.95}(0.00056 \leq \kappa_1 \leq 0.00098)$ because its endpoints are

```
> conf1 := k1 - k;                          # Resp. conf1 := 0.000563606
```

```
> conf2 := k1 + k;                          # Resp. conf2 := 0.000976392
```

Similar Material in AEM: Sec. 25.9

Problem Set for Chapter 25

Pr.25.1 (Experiment on sample mean) Find out experimentally how sample means vary from sample to sample. Plot a histogram of their frequency function. *Suggestion:* 20-100 samples of size 5 obtained by the random generator from a population of size 50. Use `SetState`, so that you can reproduce your samples if needed.
(*AEM* Sec. 25.1)

Pr.25.2 (Confidence interval for the mean) Find a 99% confidence interval for the mean of a normal population with standard deviation 2.7, using the sample 25.5, 24.7, 24.6, 24.8, 26.4, 28.7. (*AEM* Sec. 25.3)

Pr.25.3 (Length of confidence interval) Plot the length of a 95% confidence interval as a function of sample size n and measured in multiples of σ, for the mean of the normal distribution with known variance (*AEM* Sec. 25.3)

Pr.25.4 (Confidence interval for the mean) What confidence interval would you obtain in Example 25.3 in this Guide if σ were known and equal to $s = 3.2514$ (the value in that example), the other data being as before?

Pr.25.5 (Confidence interval for the variance) Find a 95% confidence interval for the variance of the normal distribution, using the sample of carbon monoxide emission (grams/mile) of a passenger car cruising at a speed of 60 mph. 14.7, 14.5, 14.9, 14.6, 14.2,15.1, 14.3, 15.0. (*AEM* Sec. 25.3)

Pr.25.6 (Test for the mean) Test the hypothesis $\mu_0 = 24$ against the alternative $\mu_1 = 27$, choosing $\alpha = 5\%$ and using a sample of size 10 with mean 25.8 from a normal population with variance 9. Is the power of the test sufficiently large?
(*AEM* Sec. 25.4)

Pr.25.7 (Test for the mean) If a standard treatment cures about 75% of patients suffering from a certain disease, and a new treatment cured 310 of the first 400 patients on whom it was tried, can you conclude that the new treatment is better? First guess. Then calculate, choosing $\alpha = 5\%$ and using the fact that $X =$ *Number of cases cured in 400 cases* is about normal with mean np and variance $np\,(1-p)$.
(*AEM* Sec. 25.4)

Pr.25.8 (Dependence of power on sample size) How does the figure in Example 25.6 in this Guide change if you take a larger sample (of size $n = 500$, for instance)? Give the reason. Plot a new figure for $n = 500$. (*AEM* Sec. 25.4)

Pr.25.9 (Test for the variance) Suppose that in the past the standard deviation of weights of certain 100.0-oz packages filled by a machine was 0.8 oz. Test the hypothesis $H_0 : \sigma = 0.8$ against the alternative $H_1 : \sigma > 0.8$ (an undesirable increase), using a sample of 20 packages with standard deviation 1.0 oz, assuming normality and choosing $\alpha = 5\%$. (*AEM* Sec. 25.4)

Pr.25.10 (Comparison of means) Will an increase of temperature increase the yield (measured in grams/min) of some chemical process? Test this, using the following independent samples, assuming normality, and choosing $\alpha = 5\%$. (*AEM* Sec. 25.4)

Yield x at 55^o C	97	108	115	103	113	117	130	127	111	107
Yield y at 70^o C	115	123	138	118	105	130	132	127		

Pr.25.11 (Paired comparison of means) Measure the electric voltage in a circuit at the same instants simultaneously by two kinds of voltmeters. Test the hypothesis that there is no difference in the calibration of the two kinds of instruments against the alternative that there is a difference. In this case you merely need a sample of differences of corresponding measurements ("**paired comparison**"), say, 0.5, −0.7, 0.3, 1.1, 0.9, −1.2, 0.5, 1.3,1.0. Assume normality and choose $\alpha = 5\%$. (*AEM* Sec. 25.4)

Pr.25.12 (Comparison of variances) Test that the variances of the populations in Pr.25.10 are equal against the alternative that they are different. Choose $\alpha = 5\%$.
(*AEM* Sec. 25.4)

Pr.25.13 (Goodness of fit) Can you assert that the traffic on the three lanes of an expressway (in one direction) is about the same on each lane if a count gives 920, 870, 750 cars on the right, middle, and left lanes, respectively, during the same interval of time? (*AEM* Sec. 25.7)

Pr.25.14 (Linear regression) If a sample of 9 pairs of values $[x_j,\, y_j]$ has the variance of the x-values 118.000, the variance of the y-values 215.125, and the covariance −155.750, what can you say about the sample regression line? What about a 95% confidence interval for the regression coefficient (the slope) κ_1 if you assume Y to be normal with variance independent of x? (*AEM* Sec. 25.9)

Pr.25.15 (Quadratic regression parabola) Fit a quadratic parabola through the following data. Plot the curve and the given data (as points) on common axes. Type `?fit` for information.

x:	1	2	4	5	7	8
y:	7	5	2	1	2	4