Preface

Most calculations are either done using Microsoft Excel XP or the free Maple 8 trial version. All worksheets are available online: http://www.stephan-brumme.com/studies/statistik.html

Part I – t-Tests

Problem 1

> **?** *Consider the body temperatures of twenty-five intertidal crabs that we exposed to air at 24.3°C. We wish to ask whether the mean body temperature of members of this species of crab is the same as the ambient air temperature of 24.3°C.*
>
> *Body temperatures (measured in °C):*
>
> *22.9, 25.8, 24.6, 26.1, 22.9, 25.1, 27.3, 24.0, 24.5, 23.9, 26.2, 24.3, 24.6, 23.3, 25.5, 28.1, 24.8, 23.5, 26.3, 25.4 25.5, 23.9, 27.0, 24.8, 25.4*

The first step of each test is to clarify the hypothesis H and its alternative K. According to the problem statement, the hypothesis H is the mean body temperature t_{crabs} of the crabs is the same as the ambient air temperature t_{air}, i.e. $t_{crabs} = t_{air}$. On the other hand, the alternative (called "null hypothesis") K can be written as $t_{crabs} \neq t_{air}$.

There seems to be a correlation between both temperature, thus it's a single sample or paired Student t-test. If we assume that the mean body temperatures are $t_{crabs} \sim N(\mu, \sigma^2)$ distributed then we have to apply the two-sided t-test. The hypothesis H will be discarded in case $|T| > t_{n-1;\,1-\frac{\alpha}{2}}$.

In general,

$$T = \sqrt{n} \cdot \frac{\overline{Z} - \mu_0}{S}$$

$$S^2 = \frac{1}{n-1} \cdot \sum \left(Z_i - \overline{Z}\right)^2$$

If $\mu = \mu_0$ then $T \sim t_{n-1}$. The given scenario defines

$$\mu_0 = t_{air} = 24.3$$

I picked Microsoft Excel XP as my tool-of-choice for this problem. The given data set was imported into a worksheet just within a few seconds and creating a suiting diagram was even easier. In my eyes, the visual understanding of the problem can be slightly enhanced by adding the measured air temperature as a horizontal line at 24.3℃ as one c an see in Figure 1.

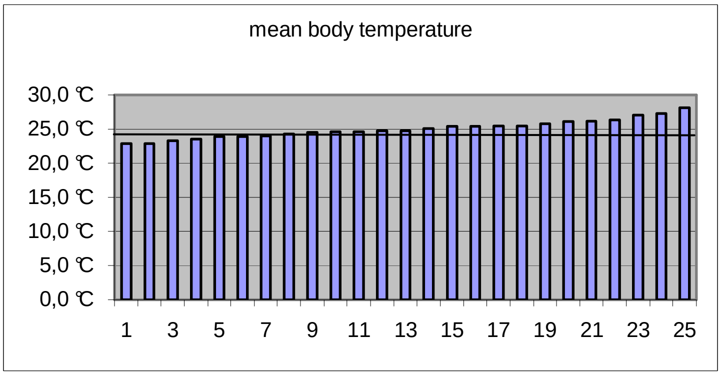

Figure 1 *Measured values*

Excel computes the remaining variables using its built-in functions `COUNT`, `AVERAGE` and `STDEV`:

$$n = 25$$
$$\overline{Z} \approx 25.028$$
$$S \approx 1.318$$

Therefore:

$$T = \sqrt{n} \cdot \frac{\overline{Z} - \mu_0}{S} \approx \sqrt{25} \cdot \frac{25.028 - 24.3}{1.318} \approx 2.7128$$

Looking up a table on the student-t distribution yields

$$t_{24;\ 0.975} \approx 2.0639$$
$$t_{24;\ 0.995} \approx 2.7969$$

Since $T > t_{24;\ 0.975}$ I have to reject the hypothesis H on a 5% level but cannot do so on a 1% level because $T < t_{24;\ 0.995}$. It heavily depends on the level whether the hypothesis H should be rejected or not. However, there are a few indicators supporting the idea that the mean body temperature of craps is correlated to the ambient air temperature.

PROBLEM 2

> **?** *A test has been set up in order the observe whether smoking during pregnancy affects the level of lead (Pb) in the babies' blood. The level of lead has been measured among 202 babies whose mothers are smokers and 333 babies whose mothers are non-smokers. Decide whether an influence can be detected.*

This time I can assume that there is no correlation among the babies and use the ordinary t-test. Their level of lead should be $l \sim N(\mu, \sigma^2)$. If X_i indicates the i-th of n_x babies of the non-smokers and Y_j the j-th of n_y babies of the smokers then one can write:

$$X_i \sim N(\mu_x, \sigma_x^2)$$
$$Y_j \sim N(\mu_y, \sigma_y^2)$$

It is important to emphasize that the equation $\sigma_x^2 = \sigma_y^2 = \sigma^2$ needs to be true. The definition of a helper variable d_0 simplifies the formulas:

$$d_0 = \mu_x - \mu_y$$

Then

$$\overline{X} - \overline{Y} \sim N\left(d_0, \left(\frac{1}{n_x} + \frac{1}{n_y}\right) \cdot \sigma^2\right)$$

σ^2 has to be estimated by

$$\hat{\sigma}^2 = \frac{(n_x - 1) \cdot S_x^2 + (n_y - 1) \cdot S_y^2}{n_x + n_y - 2}$$

The final value of T is

$$T = \frac{\overline{X} - \overline{Y} - d_0}{\sqrt{\left(\frac{1}{n_x} + \frac{1}{n_y}\right) \cdot \hat{\sigma}^2}} \sim t_{n_x + n_2 - 2}$$

I define the hypothesis H to be the case that smoking *does not* affect the level of lead in the babies' blood, i.e. $d_0 = \mu_x - \mu_y = 0$. This hypothesis ought to be rejected if $|T| > t_{n_x + n_2 - 2;\, 1-\frac{\alpha}{2}}$. Here are the parameters:

$$n_x = 333$$
$$n_y = 202$$

$$\overline{X} \approx 8.3571$$
$$\overline{Y} \approx 9.0020$$
$$S_x^2 \approx 3.2892$$
$$S_y^2 \approx 3.0440$$

Furthermore

$$\hat{\sigma}^2 \approx 10.2332$$

And finally

$$T \approx \frac{8.3571 - 9.0020 - 0}{\sqrt{\left(\frac{1}{333} + \frac{1}{202}\right) \cdot 3.1967}} \approx -2.2616$$

The inequality $|T| > t_{n_x + n_2 - 2;\, 1-\frac{\alpha}{2}}$ can be verified right now for the common 5% level:

$$t_{333+202-2;\, 1-\frac{0.05}{2}} = t_{533;\, 0.975} \approx 1.9644$$
$$2.2616 > 1.9644$$

In consequence, I have to deny the hypothesis. That result significantly supports the null hypothesis that smoking *does* affect the level of lead in babies' blood.

PROBLEM 3

> ? *Researchers have long been interested in the effects of alcohol on the human body. The authors of the paper "Effects of Alcohol on Hypoxia" examined the relationship between alcohol intake and the time of useful consciousness during high-altitude flight. Ten male subjects were taken to a simulated altitude of 25,000 ft and given several tasks to perform. Each was carefully observed for deterioration in performance due to lack of oxygen, and the time at which useful consciousness ended was recorded. Three days later, the experiment was repeated one hour after the subjects had ingested 0.5 cm^3 of 100-proof whiskey per pound of body weight. The time (in seconds) of useful consciousness was again recorded. The resulting data appears in the accompanying table.*
>
> *Is there sufficient evidence to conclude that ingestion of whiskey reduces the mean time of useful consciousness ?*

Here are the recorded times:

no alcohol	alcohol	difference
261	185	76
565	375	190
900	310	590
630	240	390
280	215	65
365	420	-55
400	405	-5
735	205	530
430	255	175
900	900	0

Table 1 *Recorded times*

The measured values are paired, therefore I have to compute the difference between the recorded times while being not influenced by alcohol and being "drunken". This problem is very similar to problem 1, so I will just present a short outline of the single steps taken to compute the desired result.

Hypothesis H: Ingestion of whiskey reduces the mean time of useful consciousness.

Excel gives us the basic properties of the calculated differences:

$$n = 10$$
$$\overline{Z} = 195.6$$
$$S \approx 230.5265$$

It is possible to get T only from these numbers by using the formula

$$T = \sqrt{n} \cdot \frac{\overline{Z}}{S} \approx 2.6832$$

Now I looked up $t_{9;\,0.95}$:

$$t_{9;\,0.95} \approx 2.2622$$

The hypothesis should be rejected since $T > t_{9;\ 0.95}$ on a level of 5%.

I was quite bored by repeatedly applying the same algorithm three times. Therefore, I decided to compute the two-sided significance in order to get a better impression of the methods implemented in specialized software such as SPSS.

Maple's diagrams do not look as cute as Excel's. On the other hand, Maple solves even very complex equations in a hush. The student-t distribution with nine levels of freedom possesses a similar shape like the normal distribution (both density and distribution function):

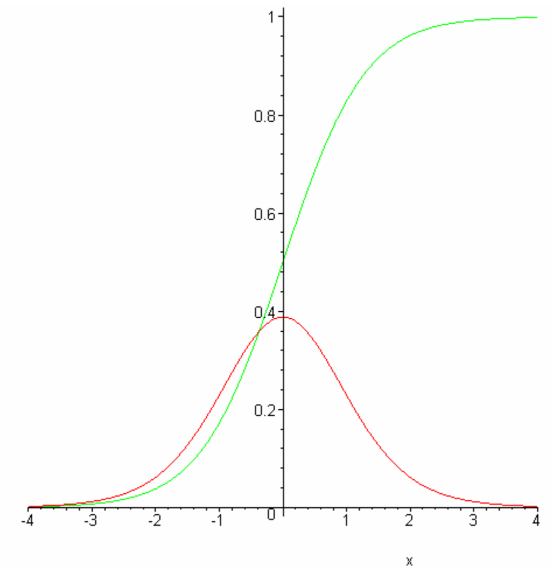

Figure 2 *Student-t distribution*

Just four lines of code are responsible for the diagram:

```
> studentt:=(alpha, size) -> statevalf[icdf, studentst[size]](1-alpha/2):
> tpdf:=statevalf[pdf, studentst[9]]:
> tcdf:=statevalf[cdf, studentst[9]]:
> plot({tpdf(x), tcdf(x)}, x=-4..4);
```

Now comes the tricky part – the significance. That computation relies on one of Maple's core features: solving equations. However, a single command gives me the desired result ($T \approx 2.6832$, two-sided !):

```
> 2*(1-tcdf(2.6832));
  0.025074272
```

Indeed, the result turns out to be the same as SPSS'.

Part II – Binomial Tests

PROBLEM 1

> ? *A commonly used medicine is effective in 40% of all treatments. Decide whether a recently developed medicine is more effective – 20 persons get this innovative medicine under supervision. Describe a test that confirms a significant improvement.*

The problem statement clearly defines two basic parameters of a binomial test:

$$n = 20$$
$$\vartheta = 0.4$$

The level of accuracy in a medical environment needs to be very high since wrong estimations may cause severe injuries or even deaths. Therefore, my proposed test should reach a accuracy level of at least 99.5%. Even though that number may seem quite high at the first glance, I want to underline that just 20 persons do not represent a reliable data set.

An estimator of the efficiency of the new medicine is the arithmetic mean:

$$\hat{\vartheta} = \frac{1}{n} \cdot \sum_{i=1}^{20} Z_i$$

where $\sum Z_i \sim B(n, \vartheta)$.

The only way to prove that the new medicine actually heals better than the old one does is to state that the *new one is not better*. This decision is caused by the fact that you cannot not surely verify a hypothesis H but its alternative K. So we have: H is "the new one is not better" where K stands for "the old one is not better". In mathematical terms:

$$H : \vartheta \leq \vartheta_0$$
$$K : \vartheta > \vartheta_0$$

The closer $\hat{\vartheta}$ gets to 1, the higher is the number of healed persons k and the better the new medicine works. A sufficient k fulfils the condition

$$P_{\vartheta_0}\left(\sum Z_i > k\right) \leq \alpha$$

which means that all observations using the *old* medicine and getting more than k successes have a lower probability than α. According to the hypothesis ($\vartheta \leq \vartheta_0$)

$$P_{\vartheta}\left(\sum Z_i > k\right) \leq P_{\vartheta_0}\left(\sum Z_i > k\right)$$

I can refuse the hypothesis if $\sum Z_i > k_{1-\alpha}$ holds true. Excel generates a table of all possible k without any expensive calculations just by invoking `BINOMDIST`:

k	P(X<=k)	P(X>k)
0	0.004%	99.996%
1	0.052%	99.948%
2	0.361%	99.639%
3	1.596%	98.404%
4	5.095%	94.905%
5	12.560%	87.440%
6	25.001%	74.999%
7	41.589%	58.411%
8	59.560%	40.440%
9	75.534%	24.466%
10	87.248%	12.752%
11	94.347%	5.653%
12	97.897%	2.103%
13	99.353%	0.647%
14	**99.839%**	**0.161%**
15	99.968%	0.032%
16	99.995%	0.005%
17	99.999%	0.001%
18	100.000%	0.000%
19	100.000%	0.000%
20	100.000%	0.000%

Table 2 *Binomial distribution*

The smallest k I accept is $k=14$ because the probability that the old medicine reaches that level is below 0.05%:

$$P_{\vartheta_0}\left(\sum Z_i > 14\right) \leq 0.05\%$$

PROBLEM 2

> ? *Suppose you played 50 tennis matches against your favourite opponent. You won 29 and lost 21. Now your opponent proposed that you are not a significantly better player than he is since that distribution is more or less random. Decide whether he is true or not.*

The algorithm does not differ from the one used in the previous problem. Of course, the parameters slightly changed:

$$n = 50$$

$$\vartheta = 0.5$$

Excel reveals these numbers:

k	P(X<=k)	P(X>k)
0	0.000%	100.000%
1	0.000%	100.000%
2	0.000%	100.000%
3	0.000%	100.000%
4	0.000%	100.000%
5	0.000%	100.000%
6	0.000%	100.000%
7	0.000%	100.000%
8	0.000%	100.000%
9	0.000%	100.000%
10	0.001%	99.999%
11	0.005%	99.995%
12	0.015%	99.985%
13	0.047%	99.953%
14	0.130%	99.870%
15	0.330%	99.670%
16	0.767%	99.233%
17	1.642%	98.358%
18	3.245%	96.755%
19	5.946%	94.054%
20	10.132%	89.868%
21	16.112%	83.888%
22	23.994%	76.006%
23	33.591%	66.409%
24	44.386%	55.614%
25	55.614%	44.386%

k	P(X<=k)	P(X>k)
26	66.409%	33.591%
27	76.006%	23.994%
28	83.888%	16.112%
29	**89.868%**	**10.132%**
30	94.054%	5.946%
31	96.755%	3.245%
32	98.358%	1.642%
33	99.233%	0.767%
34	99.670%	0.330%
35	99.870%	0.130%
36	99.953%	0.047%
37	99.985%	0.015%
38	99.995%	0.005%
39	99.999%	0.001%
40	100.000%	0.000%
41	100.000%	0.000%
42	100.000%	0.000%
43	100.000%	0.000%
44	100.000%	0.000%
45	100.000%	0.000%
46	100.000%	0.000%
47	100.000%	0.000%
48	100.000%	0.000%
49	100.000%	0.000%
50	100.000%	0.000%

Table 3 *Binomial distribution*

As one can see from the table, winning 29 out of 50 games indeed does not necessarily mean to be the significantly better player since the level of confidence is below 90%.

PROBLEM 3

? *A woman who smokes during pregnancy increases health risks to the infant. Suppose that a sample of 300 pregnant women who smoked prior to pregnancy contained 51 who quit smoking during pregnancy. Does this data support the theory that fewer than 25% of female smokers quit smoking during pregnancy ?*

My beloved Excel worksheet can be reused for the second time ☺

$$n = 300$$
$$\vartheta = 0.25$$

The whole consumes too much space; hence, I concentrate on the most interesting parts of it.

k	P(X<=k)	P(X>k)
46	0.003%	99.997%
47	0.006%	99.994%
48	0.011%	99.989%
49	0.020%	99.980%
50	0.034%	99.966%
51	**0.057%**	**99.943%**
52	0.095%	99.905%
53	0.153%	99.847%
54	0.242%	99.758%
55	0.374%	99.626%
56	0.567%	99.433%
57	0.842%	99.158%
58	1.226%	98.774%
59	1.752%	98.248%
60	2.456%	97.544%
61	3.378%	96.622%
62	4.564%	95.436%
63	6.057%	93.943%
64	7.901%	92.099%
65	10.131%	89.869%
66	12.779%	87.221%
67	15.861%	84.139%
68	19.382%	80.618%
69	23.327%	76.673%
70	27.667%	72.333%
71	32.354%	67.646%
72	37.323%	62.677%
73	42.495%	57.505%
74	47.785%	52.215%
75	53.098%	46.902%

Table 4 *Binomial distribution*

The observed group strongly supports the theory at a level of more than 99.9%, which is sufficient even in a medical context. However, the women did not give birth until the data has been collected. Thus, there is a probability that some of the still smoking mother-to-be quit smoking. If more than six women do so then my statement will be proven wrong.

Part III – Comparison of Probabilities

PROBLEM 1

? *A medical analysis reveals that the blood of 34 out of 113 boys and 54 out of 139 girls contains some kind of an anaphylactic protecting the children from a flu virus. Is there a significant dissimilarity among boys and girls ?*

In order to solve that problem, I compute the mean value independently for both boys and girls:

$$n_{boys} = 113$$
$$n_{girls} = 139$$
$$a_{boys} = 34$$
$$a_{girls} = 54$$

$$\hat{\vartheta} = \frac{n}{a}$$
$$\hat{\vartheta}_{boys} \approx 0.3009$$
$$\hat{\vartheta}_{girls} \approx 0.3885$$

If there is no significant discrepancy then

$$\vartheta_{boys} = \vartheta_{girls} = \vartheta$$

and

$$N(0,1) \sim \frac{\hat{\vartheta}_{boys} - \hat{\vartheta}_{girls}}{\sqrt{\left(\frac{1}{n_{boys}} + \frac{1}{n_{girls}}\right) \cdot \vartheta \cdot (1 - \vartheta)}}$$

Parameter ϑ can be estimated by the weighted mean of $\hat{\vartheta}_{boys}$ and $\hat{\vartheta}_{girls}$:

$$\hat{\vartheta} = \frac{n_{boys} \cdot \hat{\vartheta}_{boys} + n_{girls} \cdot \hat{\vartheta}_{girls}}{n_{boys} + n_{girls}}$$

One ought to reject the hypothesis if $|T| > u_{1-\frac{\alpha}{2}}$:

$$T = \frac{\hat{\vartheta}_{boys} - \hat{\vartheta}_{girls}}{\sqrt{\left(\frac{1}{n_{boys}} + \frac{1}{n_{girls}}\right) \cdot \hat{\vartheta} \cdot (1 - \hat{\vartheta})}}$$

We get

$$\hat{\vartheta} \approx 0.3492$$
$$T \approx -1.4508$$

Because of

$$u_{0.975} \approx 1.9600$$
$$|T| < u_{0.975}$$

I have to reject the hypothesis. There is no significant dissimilarity concerning the level of anaphylactic among boys and girls.

Part IV – Kolmogorov Tests

PROBLEM 1

> **?** *Generate 100 samples of a normal distribution $N(2,4)$. Apply the Kolmogorov test to verify whether these numbers are observations of a*
>
> *a)* $N(2,4)$ *distribution*
>
> *b)* $N(1,4)$ *distribution*
>
> *Are the values provided in the file "uniform.txt" observations of an uniform distribution* $U \in [0,1]$ *?*

The latter problem is solved first – these numbers presented below were given in uniform.txt:

N	random number	N	random number	N	random number
1	0.382000183	34	0.816522721	67	0.555162206
2	0.100680563	35	0.972502823	68	0.181157872
3	0.596484268	36	0.466322825	69	0.970274972
4	0.899105808	37	0.300210578	70	0.686941130
5	0.884609516	38	0.750206000	71	0.528794214
6	0.958464309	39	0.351481674	72	0.796685690
7	0.014496292	40	0.775658437	73	0.805658132
8	0.407422102	41	0.074343089	74	0.262215033
9	0.863246559	42	0.198431349	75	0.177953429
10	0.138584552	43	0.064058351	76	0.866756188
11	0.245033113	44	0.358348338	77	0.114841151
12	0.045472579	45	0.487044893	78	0.059511093
13	0.032380139	46	0.511215552	79	0.761558885
14	0.164128544	47	0.373455000	80	0.738395337
15	0.219611194	48	0.985900449	81	0.986297189
16	0.017090365	49	0.040711692	82	0.925595874
17	0.285042879	50	0.230719932	83	0.903866695
18	0.343089084	51	0.004974517	84	0.544969024
19	0.553636280	52	0.926145207	85	0.500778222
20	0.357371746	53	0.100314341	86	0.674977874
21	0.371837519	54	0.256691183	87	0.489822077
22	0.355601672	55	0.775688955	88	0.145786920
23	0.910306101	56	0.679647206	89	0.037965026
24	0.466017640	57	0.809106723	90	0.796258431
25	0.426160466	58	0.724326304	91	0.671559801
26	0.303903317	59	0.085055086	92	0.731681265
27	0.975707266	60	0.132267220	93	0.584521012
28	0.806665242	61	0.756157109	94	0.152226325
29	0.991241188	62	0.626514481	95	0.892178106
30	0.256263924	63	0.173650319	96	0.377819147
31	0.951689199	64	0.404797510	97	0.200476089
	0.053437910	65	0.552323984	98	0.205786309
33	0.705038606	66	0.711508530	99	0.333964049
				100	0.325144200

Table 5 *Uniformly distributed random numbers*

The total number of random values is big enough to apply an approximation:

$$n = 100 > 40$$

$$P\left(\sqrt{n} \cdot D_n \le x\right) \approx 1 - 2 \cdot \sum_{k=1}^{\infty} (-1)^{k-1} \cdot e^{-2k^2x^2}$$

D_n is one important parameter:

$$D_n = \sup_{t \in R} \left| \hat{F}_n(t) - F_0(t) \right|$$

A uniform distribution U of n on the interval $[0,1]$ should subdivide that interval into n equally sized partitions as shown in Figure 3:

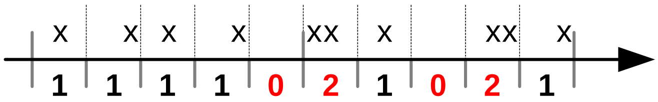

Figure 3 *Exemplary uniform distribution of ten arbitrary elements*

Obviously, not all of the n intervals cover the "perfect" number of just one element. In Figure 3 all red numbers symbolize intervals with not exactly one single element, i.e. no element or two (or even more) elements.

I subdivided the "real" data set into 100 intervals, each 0.01 wide. Maple 8 trial helped me by determining the frequencies for all 100 intervals. These few lines of code did all the work (I omit most of the input data required for **uniformData**):

```
> uniformData:=[0.382000183,…,0.3251442]:
> partitions:=[seq(i/100..(i+1)/100, i=0..99)]:
> weighted:=tallyinto(uniformData, partitions):
> frequency(weighted);
```

[0, 0, 2, 0, 1, 2, 0, 1, 1, 1, 0, 2, 2, 0, 1, 2, 0, 1, 1, 2, 0, 1, 1, 1, 1, 1, 1, 2, 0, 1, 1, 1, 0, 1, 1, 1,
3, 2, 0, 0, 1, 1, 2, 4, 2, 0, 2, 0, 1, 1, 0, 1, 0, 0, 3, 0, 1, 1, 0, 1, 1, 0, 0, 0, 1, 2, 2, 2, 0, 2, 2,
0, 1, 2, 0, 0, 0, 2, 2, 0, 3, 1, 2, 0, 0, 0, 1, 2, 0, 0, 2, 1, 1, 1, 1, 0, 1, 3, 0, 3]

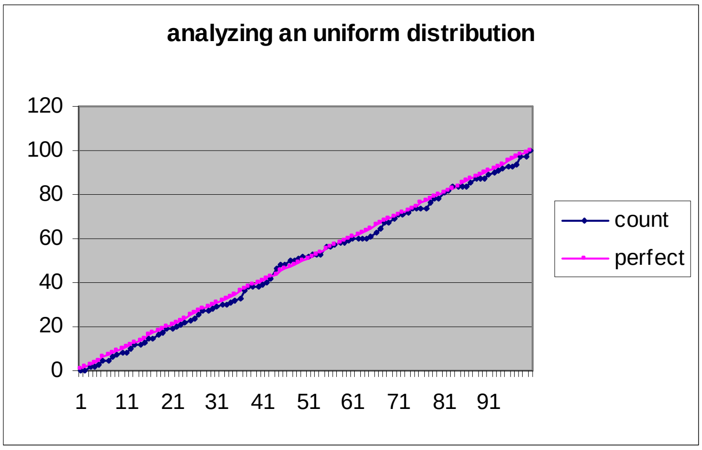

Figure 4 *Distribution of the Data Set*

from	to	elements	"error"
0.00	0.01	-	1
0.01	0.02	-	1
0.02	0.03	2	1
0.03	0.04	-	1
0.04	0.05	1	-
0.05	0.06	2	1
0.06	0.07	-	1
0.07	0.08	1	-
0.08	0.09	1	-
0.09	0.10	1	-
0.10	0.11	-	1
0.11	0.12	2	1
0.12	0.13	2	1
0.13	0.14	-	1
0.14	0.15	1	-
0.15	0.16	2	1
0.16	0.17	-	1
0.17	0.18	1	-
0.18	0.19	1	-
0.19	0.20	2	1
0.20	0.21	-	1
0.21	0.22	1	-
0.22	0.23	1	-
0.23	0.24	1	-
0.24	0.25	1	-
0.25	0.26	1	-
0.26	0.27	1	-
0.27	0.28	2	1
0.28	0.29	-	1
0.29	0.30	1	-
0.30	0.31	1	-
0.31	0.32	1	-
0.32	0.33	-	1
0.33	0.34	1	-
0.34	0.35	1	-
0.35	0.36	1	-
0.36	0.37	3	2
0.37	0.38	2	1
0.38	0.39	-	1
0.39	0.40	-	1
0.40	0.41	1	-
0.41	0.42	1	-
0.42	0.43	2	1
0.43	**0.44**	**4**	**3**
0.44	0.45	2	1
0.45	0.46	-	1
0.46	0.47	2	1
0.47	0.48	-	1
0.48	0.49	1	-
0.49	0.50	1	-
0.50	0.51	-	1
0.51	0.52	1	-
0.52	0.53	-	1
0.53	0.54	-	1
0.54	0.55	3	2
0.55	0.56	-	1
0.56	0.57	1	-
0.57	0.58	1	-
0.58	0.59	-	1
0.59	0.60	1	-
0.60	0.61	1	-
0.61	0.62	-	1
0.62	0.63	-	1
0.63	0.64	-	1
0.64	0.65	1	-
0.65	0.66	2	1
0.66	0.67	2	1
0.67	0.68	2	1
0.68	0.69	-	1
0.69	0.70	2	1
0.70	0.71	2	1
0.71	0.72	-	1
0.72	0.73	1	-
0.73	0.74	2	1
0.74	0.75	-	1
0.75	0.76	-	1
0.76	0.77	-	1
0.77	0.78	2	1
0.78	0.79	2	1
0.79	0.80	-	1
0.80	0.81	3	2
0.81	0.82	1	-
0.82	0.83	2	1
0.83	0.84	-	1
0.84	0.85	-	1
0.85	0.86	-	1
0.86	0.87	1	-
0.87	0.88	2	1
0.88	0.89	-	1
0.89	0.90	-	1
0.90	0.91	2	1
0.91	0.92	1	-
0.92	0.93	1	-
0.93	0.94	1	-
0.94	0.95	1	-
0.95	0.96	-	1
0.96	0.97	1	-
0.97	0.98	3	2
0.98	0.99	-	1
0.99	1.00	3	2

Table 6 *Examined intervals*

The biggest difference between an optimal density of 1 and the observed random numbers occurred in the interval [0.43, 0.44]. Four numbers – three more than expected – fall into that interval. Now I can write down the formula of $D_n = \sup_{t \in R} \left| \hat{F}_n(t) - F_0(t) \right|$ specialized for the uniform distribution:

$$
\begin{aligned}
k &\in \{1,..,100\} \\
n &= 100 \\
I_k &= \left[\frac{k-1}{n}, \frac{k}{n} \right[\\
D_n &= \sup_{k \in \{1,...,n\}} \left| \frac{\max elements(I_k)}{n} - \frac{1}{n} \right| \\
&= \frac{4}{100} - \frac{1}{100} \\
&= 0.03
\end{aligned}
$$

Bronstein's famous book contains a precomputed table of the Kolmogorov distribution. Two noteworthy values are:

$$
\begin{aligned}
Q(\lambda_\alpha) &= 1 - \alpha \\
Q(1.36) &\approx 0.9505 \\
Q(1.63) &\approx 0.9902
\end{aligned}
$$

If the inequality $\sqrt{n} \cdot D_n > \lambda_\alpha$ holds true then I can conclude that the distribution is not uniformly distribution. The statement is valid with an error probability of α. Applying the inequality to the looked up values $Q(\lambda_\alpha)$:

$$
\begin{aligned}
\sqrt{100} \cdot 0.03 &= 0.3 \\
0.3 &< 1.36 \\
0.3 &< 1.63
\end{aligned}
$$

According to the Kolmogorov test, the random numbers can be treated as uniformly distributed which corresponds to my assumption drawn from Figure 4.

Now the first part of problem will be solved. Unfortunately, I did not read the manuals thoroughly and oversaw the exact definition of the functions generating and analysing the normal distribution. They expect σ not σ^2 - but I realized it *after* solving the problem. Therefore, the distributions used on the following pages are not $N(2,4)$ and $N(1,4)$, instead, they are $N(2,4^2)$ and $N(1,4^2)$. I am too lazy to correct that flaw of mine, the solution remains unchanged.

Maple produced the one listed on the next page:

	N(2,4²)
1	-8.548158370
2	-7.248662614
3	-6.645504657
4	-6.289195253
5	-5.639590245
6	-5.121919059
7	-4.742507115
8	-4.578897211
9	-4.514975148
10	-3.486728812
11	-3.299444217
12	-3.289474735
13	-3.123069152
14	-3.027110171
15	-2.813438348
16	-2.650812448
17	-2.551267865
18	-2.461531122
19	-2.384103920
20	-2.300249850
21	-1.713149517
22	-1.664410449
23	-1.631743721
24	-1.608626846
25	-1.464242246
26	-1.348859216
27	-1.346134098
28	-1.318409479
29	-1.095392287
30	-0.948517946
31	-0.941512025
32	-0.819561854
33	-0.626849761
34	-0.608254673
35	-0.586935584
36	0.371592381
37	0.420819502
38	0.541681695
39	0.544190973
40	0.608728166
41	0.686214012
42	0.941927984
43	0.988638495
44	1.001693247
45	1.023144123
46	1.368722432
47	1.555903964
48	1.807520282
49	1.834715733
50	1.990961603
51	2.129620936
52	2.302976655
53	2.600958033
54	2.694486540
55	2.696360988
56	2.742515604
57	2.837068064
58	2.873683514
59	3.210959215
60	3.235463676
61	3.272288708
62	**3.361773114**
63	3.398777348
64	**3.576075366**
65	3.656392614
66	3.729704603
67	3.927045500
68	3.932515903
69	4.293757605
70	4.376072563
71	4.719890270
72	4.803535259
73	5.166928912
74	5.335042903
75	5.421508275
76	5.441276652
77	5.495871990
78	5.518409644
79	5.527496197
80	5.547629093
81	5.626198514
82	5.706761207
83	6.108838892
84	6.599344495
85	6.637350036
86	6.647513932
87	6.664543969
88	6.743719873
89	6.769169234
90	6.903709492
91	6.961043448
92	7.488623325
93	7.591387106
94	8.167324734
95	8.376247965
96	8.437075666
97	8.698212761
98	8.951948263
99	11.526490990
100	11.543703330

Table 7 *N(2,4²) random numbers*

I discovered a powerful function called `COUNTIF` while playing with Excel. It counts the total number of all cells in a given region satisfying a specified condition.

After generating equally sized intervals, the Excel worksheet determines how many of the 100 random numbers fit into these intervals. The next step is to compute the relative frequency – I just have to divide the absolute frequencies by 100. These values are compared against an idealized $N(2,4^2)$ or $N(1,4^2)$ distribution. To do so, I find out the absolute value of the difference between observed and idealized distribution.

Two diagrams visualize the results:

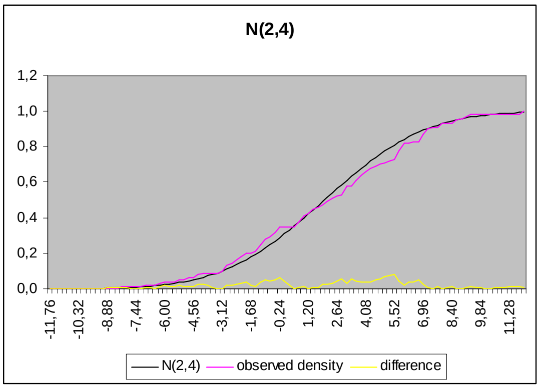

Figure 5 *N(2,4²) hypothesis*

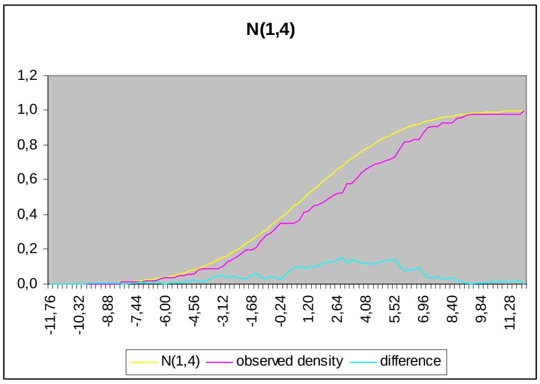

Figure 6 *N(1,4²) hypothesis*

I got a far higher difference – that is actually the desired D_n - for the $N(1,4^2)$ hypothesis. Its upper limit is higher so I intentionally suppose I have to deny it at a certain level. The exact values are:

$$D_{n\,2,4} \approx 0.0806$$
$$D_{n\,1,4} \approx 0.1508$$

Maybe you remember the thresholds taken from the Bronstein:

$$Q(\lambda_\alpha) = 1 - \alpha$$
$$Q(1.36) \approx 0.9505$$
$$Q(1.63) \approx 0.9902$$

And the inequality did not change, too …

$$\sqrt{n} \cdot D_n > \lambda_\alpha$$

Now it is time to replace the variables by concrete numbers:

$$10 \cdot 0.0806 \overset{wrong\ !!!}{>} 1.63$$
$$10 \cdot 0.0806 \overset{wrong\ !!!}{>} 1.36$$
$$10 \cdot 0.1508 \overset{wrong\ !!!}{>} 1.63$$
$$10 \cdot 0.1508 \overset{TRUE}{>} 1.36$$

Since the inequality is wrong at a level of 5% for both the $N(2,4^2)$ and the $N(1,4^2)$ hypothesis, I cannot refuse these hypotheses. Something different happens at a 1% level – the $N(1,4^2)$ hypothesis must be rejected whereas the $N(2,4^2)$ hypothesis still can be accepted.

Part V – Rank Tests

PROBLEM 1

? *Two assay methods for measuring the level of vitamin B12 in red blood cells were compared in the paper "Noncobalimin Vitamin B12 Analogues in Human Red Cells, Liver and Brain"(American Journal of Clinical Nutrition, 1983). Blood samples were taken from 15 healthy adults, and for each blood sample, the B12 level was determined using both methods.*

A rank test is built upon differences:

method 1	method 2	difference
204	205	1
240	238	-2
209	198	-11
277	253	-24
197	180	-17
227	209	-18
207	217	10
205	204	-1
131	137	6
282	250	-32
76	82	6
194	165	-29
120	79	-41
92	100	8
114	107	-7
150	140	-10

Table 8 *Data obtained from measurements*

The absolute values of the differences are ordered and grouped. Each table entry "requires" one rank. If some entries contain the same value then the rank has to be shared.

absolute	rank
1	1.5
1	
2	3
6	4.5
6	
7	6
8	7
10	8.5
10	
11	10
17	11
18	12
24	13
29	14
32	15
41	16

Table 9 *Ranking*

All differences are denoted by D_i. Next, I sum up all ranks related to positive differences and store the result in $w_n^+ = \sum_{D_i > 0} R(|D_i|)$. The same goes for all negative differences, thus I got:

$$\begin{aligned} w_n^+ &= 1.5 + 4.5 + 4.5 + 7 + 8.5 \\ &= 26 \\ w_n^- &= 1.5 + 3 + 6 + 8.5 + 10 + 11 + 12 + 13 + 14 + 15 + 16 \\ &= 110 \end{aligned}$$

I was unable to locate a precomputed table containing the exact thresholds of rank tests. Therefore, this problem is solved using an approximation since the total number of random numbers is sufficiently large enough (>10) and all preconditions of the central limit theorem are fulfilled.

$$Q_i = \begin{cases} 0 & D_i > 0 \\ 1 & D_i < 0 \end{cases}$$

$$\begin{aligned} Ew_n^+ = Ew_n^- &= \sum_{i=1}^{n} Q_i \cdot i \\ &= \frac{1}{2} \cdot \sum_{i=1}^{n} i = \frac{n \cdot (n+1)}{4} \\ &= \frac{16 \cdot 17}{4} = 68 \end{aligned}$$

$$\begin{aligned} Var\, w_n^+ = Var\, w_n^- &= Var \sum_{i=1}^{n} Q_i \cdot i \\ &= \sum_{i=1}^{n} i^2 \cdot Var\, Q_i = \frac{n \cdot (n+1) \cdot (2n+1)}{4 \cdot 6} \\ &= \frac{16 \cdot 17 \cdot 33}{24} = 374 \end{aligned}$$

The final Z:

$$\begin{aligned} Z &= \frac{w_n^+ - Ew_n^+}{\sqrt{Var\ w_n^+}} \\ &= \frac{26-68}{\sqrt{374}} \\ &\approx -2.1718 \end{aligned}$$

I reject the hypothesis H if $|Z| > u_{1-\frac{\alpha}{2}}$ where $U \sim N(0,1)$. A dedicated table gives:

$$\begin{aligned} u_{0.975} &\approx 1.9600 \\ u_{0.995} &\approx 2.5758 \end{aligned}$$

H is accepted at a 5% level but rejected at a 1% level.

Part VI – χ^2 Tests

PROBLEM 1

? *Be the descendants of beans of three different types, the distribution scheme is 1:2:1. An experiment examines 100 of these descendants and found 29 times type 1, 44 times type 2 and 27 times type 3. Is there a significant discrepancy ?*

When applying the χ^2 test one has to ensure to observe only discrete events, e.g. $z_j \in \{red, blue, green\}$.

Let's rewrite the problem statement in a more mathematical style:

$$P(Z = z_j) = p_j$$
$$P(Z = 1) = 0.25$$
$$P(Z = 2) = 0.5$$
$$P(Z = 3) = 0.25$$

The hypothesis is $p_j = \hat{p}_j$ for all j. Then:

$$T = n \cdot \sum_{j=1}^{k} \frac{(\hat{p}_j - p_j)^2}{p_j}$$
$$= \sum_{j=1}^{k} \frac{(H_j - n \cdot p_j)^2}{n \cdot p_j}$$

if H_j be the frequency of event j. Furthermore:

$$n = 100$$
$$k = 3$$
$$H_1 = 29$$
$$H_2 = 44$$
$$H_3 = 27$$
$$n \cdot p_1 = 25$$
$$n \cdot p_2 = 50$$
$$n \cdot p_3 = 25$$

The term $n \cdot p_j$ is sometimes called *residual*. It is allowed to utilize the χ^2 test because $n \cdot p_j \geq 5$ is true for all three kinds of beans.

Evaluating the formula leads to:

$$
\begin{aligned}
T &= \sum_{j=1}^{3} \frac{\left(H_j - n \cdot p_j\right)^2}{n \cdot p_j} \\
&= \frac{(29-25)^2}{25} + \frac{(44-50)^2}{50} + \frac{(27-25)^2}{25} \\
&= 0.64 + 0.72 + 0.08 \\
&= 1.52
\end{aligned}
$$

There are just two degrees of freedom:

$$
\begin{aligned}
\chi^2_{2,\ 0.95} &\approx 6.0 \\
\chi^2_{2,\ 0.99} &\approx 9.2
\end{aligned}
$$

I accept the hypothesis.

PROBLEM 2

> ? *Generate 100 $B(1, 0.6)$ distributed random numbers. Apply the χ^2 test to verify whether these data are $B(1, 0.6)$, $B(1, 0.9)$ or / and $B(1, 0.5)$ distributed.*

Basically, there are just two events: 0 and 1. The generated random numbers were quite close to my expectations (I omit the table to save some space):

$$n = 100$$
$$k = 2$$
$$H_0 = 41$$
$$H_1 = 59$$

For

$$T = \sum_{j=1}^{k} \frac{(H_j - n \cdot p_j)^2}{n \cdot p_j}$$

we get

ϑ	$n \cdot p_1$	$n \cdot p_0$	T
0.6	60	40	0.0417
0.9	90	10	106.7778
0.5	50	50	3.2400

Table 10 χ^2 *test of binomial distributions*

Some interesting $\chi^2_{1,1-\alpha}$:

$$\chi^2_{1,0.95} \approx 3.8$$
$$\chi^2_{1,0.99} \approx 6.6$$

The random numbers may be $B(1, 0.6)$ or $B(1, 0.5)$ but are definitely not $B(1, 0.9)$ distributed.

PROBLEM 3

? *Generate 50 $Poisson(\lambda = 0.5)$ distributed random numbers. Apply the χ^2 to verify whether these numbers are $Poisson(\lambda = 0.5)$ distributed.*

Teamwork par excellence – Maple generated the numbers, Excel analysed them:

nr	r.v.
1	0
2	0
3	0
4	0
5	0
6	1
7	0
8	1
9	0
10	1

nr	r.v.
11	0
12	0
13	1
14	0
15	0
16	2
17	1
18	0
19	1
20	2

nr	r.v.
21	2
22	0
23	0
24	0
25	0
26	0
27	0
28	1
29	1
30	1

nr	r.v.
31	0
32	1
33	1
34	1
35	0
36	0
37	0
38	1
39	0
40	1

nr	r.v.
41	0
42	3
43	0
44	1
45	0
46	0
47	1
48	3
49	0
50	0

The according observed and expected frequencies H_j and $n \cdot p_j$:

value	observed	expected
0	29	30.3265
1	16	15.1633
2	3	3.7908
3	2	0.6318
>3	0	0.0875

Table 11 *Classification of events*

Parameters:

$$n = 50$$
$$k = 5$$

Hence:

$$T = \sum_{j=1}^{k} \frac{(H_j - n \cdot p_j)^2}{n \cdot p_j}$$
$$\approx 3.3196$$

There are five categories, i.e. four degrees of freedom:

$$\chi^2_{4,0.95} \approx 9.5$$
$$\chi^2_{4,0.99} \approx 13.3$$

I accept the hypothesis without any doubt.

PROBLEM 4

? *A dice is tossed 3000 times. Verify whether it is unbiased.*

The following scheme emerged:

number	occurences
1	511
2	472
3	572
4	498
5	513
6	434

Table 12 *Tossing a dice 3000 times*

A regular or unbiased dice is expected to produce each number with a probability of $1/6$. Even though I have a basic understanding of probability, Figure 7 does not quite look as expected:

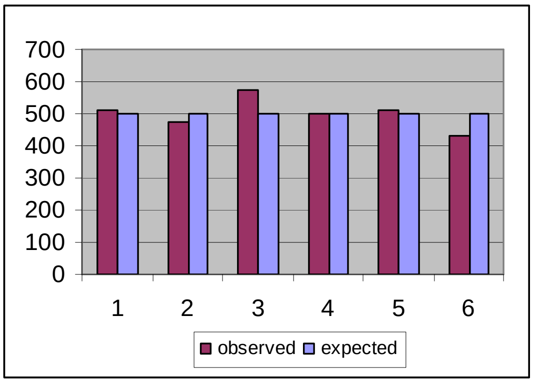

Figure 7 *Tossing a Dice*

Number 3 occurred about 15% too often while number 6 should appear 13 more often. Now the χ^2 test will prove or disprove my supposition.

$$n = 3000$$
$$k = 6$$
$$T \approx 21.236$$

There are five degrees of freedom:

$$\chi^2_{5,0.95} \approx 11.1$$
$$\chi^2_{5,0.99} \approx 15.1$$

The χ^2 test confirms my supposition: the hypothesis will be rejected, hence the dice is not regular, it is biased.

PROBLEM 5

> ? *The results of an experiment to assess the effects of crude oil on fish parasites were described in the paper "Effects of Crude Oil on Gastrointestinal Parasites of Two Species of Marine Fish". Three treatments were compared: (1) no contamination, (2) contamination by 1-year-old weathered oil, and (3) contamination by new oil. For each treatment condition, a sample of fish was taken, and then each fish was classified as either parasitized or not parasitized.*

The χ^2 homogeneity test's task is to compare independent, discrete random variables. All three groups of fish do not correlate in any way, they are independent. Their amounts are discrete.

group	parasitized	nonparasitized
no oil	30	3
old oil	16	8
new oil	16	16

Table 13

One can conclude:

$$\begin{aligned} H_p &= H_{p,no\,oil} + H_{p,old\,oil} + H_{p,new\,oil} \\ &= 30 + 16 + 16 = 62 \\ H_n &= H_{n,no\,oil} + H_{n,old\,oil} + H_{n,new\,oil} \\ &= 3 + 8 + 16 = 27 \\ H_{no\,oil} &= H_{p,no\,oil} + H_{n,no\,oil} \\ &= 30 + 3 = 33 \end{aligned}$$

$$\begin{aligned} H_{old\,oil} &= 24 \\ H_{new\,oil} &= 32 \end{aligned}$$

$$\begin{aligned} H &= H_p + H_n \\ &= 62 + 27 = 89 \end{aligned}$$

The expected amount of parasitized fishes in water not contaminated by oil is:

$$\begin{aligned} H_{e,p,no\,oil} &= n \cdot p_{p,no\,oil} = H_{no\,oil} \cdot \frac{H_p}{H} \\ &\approx 22.99 \end{aligned}$$

Likewise, the expected amount of non-parasitized fishes in water contaminated by old oil:

$$\begin{aligned} H_{e,n,old\,oil} &= n \cdot p_{n,old\,oil} = H_{old\,oil} \cdot \frac{H_n}{H} \\ &\approx 7.28 \end{aligned}$$

A cross table visualizes the relationships:

	type	parasitized	nonparasitized	total
no oil	observed	30	3	**33**
	expected	***22.99***	*10.01*	*33.00*
old oil	observed	16	8	**24**
	expected	*16.72*	***7.28***	*24.00*
new oil	observed	16	16	32
	expected	*22.29*	*9.71*	*32.00*
total	observed	**62**	**27**	**89**
	expected	*62.00*	*27.00*	*89.00*

Table 14 *Cross table*

The final step computes the sum of the normalized squared differences between observed and expected fishes:

$$T = \sum_{g \in \{no\,oil,\,old\,oil,\,new\,oil\}} \left(\sum_{k \in \{parasitized,\,nonparasitized\}} \frac{\left(H_{e,k,g} - \dfrac{n_{k,g} \cdot H_{k,g}}{n} \right)^2}{\dfrac{n_{k,g} \cdot H_{k,g}}{n}} \right)$$

Maybe the bulky formula becomes clearer when looking at a slightly enhanced version of the cross table where I added the normalized squared differences. The highlighted value is the outcome of:

$$\frac{(30 - 22.99)^2}{22.99} \approx 2.14$$

	type	parasitized	nonparasitized	total
no oil	observed	30	3	33
	expected	*22.99*	*10.01*	*33.00*
	diff^2	**2.14**	**4.91**	
old oil	observed	16	8	24
	expected	*16.72*	*7.28*	*24.00*
	diff^2	**0.03**	**0.07**	
new oil	observed	16	16	32
	expected	*22.29*	*9.71*	*32.00*
	diff^2	**1.78**	**4.08**	
total	observed	62	27	89
	expected	*62.00*	*27.00*	*89.00*

Table 15 *Normalized squared differences*

The sum of all **bold** values is T :

$$T \approx 13.0047$$

From three observed group one infers only two degrees of freedom:

$$\chi^2_{2,\ 0.95} \approx 6.0$$
$$\chi^2_{2,\ 0.99} \approx 9.2$$

The hypothesis – there are no differences – must be refused. The presence or absence of different kinds of oil significantly influences the rate of infections caused by parasites among fishes.

PROBLEM 6

? *A study examines whether method A cures significantly better than method B does.13 out of 15 persons treated with method A were successfully cured while method B reached a quantity of only 10 out of 15 persons.*

It is late in the evening and I am getting quite sleepy. That is the main reason why I strip down my solution of problem 6 to the bare minimum. Vive la cut'n'paste !

	type	cured	not cured	total
A	observed	13	2	15
	expected	*11.50*	*3.50*	*15.00*
B	observed	10	5	15
	expected	*11.50*	*3.50*	*15.00*
total	observed	23	7	30
	expected	*23.00*	*7.00*	*30.00*

Table 16 *Cross table*

$$T \approx 1.6770$$
$$\chi^2_{1,0.95} \approx 3.8$$
$$\chi^2_{1,0.99} \approx 6.6$$

Method A does not show a considerable improvement in comparison to method B.

PROBLEM 7

? *A study examines the frequency of marihuana consumption among 445 students depending on the drug consumption (such as alcohol) of their parents. Is there a significant relationship ?*

The relationship can be shown (or not) with the χ^2 test of independence. If two events are independent then the equation $P(X = x, Y = y) = P(X = x) \cdot P(Y = y)$ is always true.

	type	no parent	one parent	both parents	total
never	observed	141	68	17	226
	expected	*119.35*	*82.78*	*23.87*	*143.22*
seldom	observed	54	44	11	109
	expected	*57.56*	*39.93*	*11.51*	*69.07*
regularly	observed	40	51	19	110
	expected	*58.09*	*40.29*	*11.62*	*69.71*
total	observed	235	163	47	445
	expected	*235.00*	*163.00*	*47.00*	*282.00*

Table 17 *Cross table*

I can apply the same formula I did in problems 5 and 6.

$$T = \sum_{g \in student's\,consumption} \left(\sum_{k \in parents'\,consumption} \frac{\left(H_{e,k,g} - \frac{n_{k,g} \cdot H_{k,g}}{n} \right)^2}{\frac{n_{k,g} \cdot H_{k,g}}{n}} \right)$$

A minor changed algorithm guides us to the same result we would achieve with the method used in problems 5 and 6: so-called residuals are the difference between observed and estimated occurrences of an event. They can be standardized by dividing by the square root of the estimated occurrences. Finally yet importantly, one has to add the squared standardized residuals. Let us take a look at the table:

	type	no parent	one parent	both parents	total
never	observed	141	68	17	226
	expected	*119.35*	*82.78*	*23.87*	*143.22*
	residual	***21.65***	***-14.78***	***-6.87***	
	standardized	***1.98***	***-1.62***	***-1.41***	
seldom	observed	54	44	11	109
	expected	*57.56*	*39.93*	*11.51*	*69.07*
	residual	***-3.56***	***4.07***	***-0.51***	
	standardized	***-0.47***	***0.64***	***-0.15***	
regularly	observed	40	51	19	110
	expected	*58.09*	*40.29*	*11.62*	*69.71*
	residual	***-18.09***	***10.71***	***7.38***	
	standardized	***-2.37***	***1.69***	***2.17***	
total	observed	235	163	47	445
	expected	*235.00*	*163.00*	*47.00*	*282.00*

Table 18 *Cross table & (standardized) residuals*

The residual of students consuming no marihuana but being the child of two parents doing so was computed this way:

$$\begin{aligned} residual_{never,both} &= 17 - 23.87 \\ &= -6.87 \\ standardizedresidual_{never,both} &= \frac{-6.87}{\sqrt{23.87}} \\ &= -1.41 \end{aligned}$$

Then:

$$\begin{aligned} T &= \sum_{g \in student's\,consumption} \left(\sum_{k \in parents'\,consumption} standardizedresiduals_{g,k}^2 \right) \\ &\approx 22.3731 \\ \chi^2_{4,0.95} &\approx 9.5 \\ \chi^2_{4,0.99} &\approx 13.3 \end{aligned}$$

There are many sign indicating that the drug consumption of parents seriously influences the “drug career” of their children.

The “new” algorithm seems to be more suitable when doing all the calculations without a computer. Nowadays, the first algorithm is cheaper to set up – hence, I prefer it.