Basisprüfung B. Sc.
Statistics and Probability Theory
SS 2007

Prof. Dr. M.H. Faber

ETH Zürich

Friday 07th of September 2007
14:00 – 16:00

Surname: .....

Name: .....

Stud. Nr.: .....

Course of studies: .....

Basisprüfung B. Sc.: Statistics and Probability Theory
Civil, Environmental and Geomatic Engineering

Date and duration:

Friday 07th of September 2007
Start: 14:00
Duration: 120 minutes

Aids:

- All documentation and manuals allowed (Lecture notes, Exercise tutorials, other books and print-outs etc.)
- Calculators are allowed, but no communication medium (e.g. cell phones, calculators with Bluetooth etc.)

Administration:

- During the 15 minutes reading, it is not allowed to write on the solution sheets.
- Please place your Legi-card on your desk.
- Control first if you have received all the materials:
    - General information and exercises (18 pages)
    - 5 sheets of paper (checkered and stamped)
- Write your **name on every sheet** of paper.
- Use **only** the provided sheets of paper (5 checkered and officially stamped sheets) and use a **new sheet for every exercise**.
- Other sheets will **not** be considered in the corrections!
- When you have finished, place **all** materials in the envelope and leave it on your desk.
- You are allowed to leave until 15:45. After that time, you need to wait until the end of the exam.

Content of the exam:

Content	Description	Page	Points
Exercise 1	Descriptive Statistics	3	25
Exercise 2	Bayes Theorem	7	25
Exercise 3	Probability paper	9	15
Exercise 4	Confidence intervals and Chi-Square Test	12	25
Exercise 5	Sum of two random variables	15	30
Annexes	Cumulative distribution function of the Standard Normal distribution	16	-
	Quantile values of the Chi-square distribution	17	-
Glossary	English-German	18	-
			120

Remarks:

- All exercises 1 to 5 have to be solved.
- If you are having difficulties in a certain question but need a value/number in order to continue, then make an assumption, mark it as an assumption, and continue the calcultions with that value.

Exercise 1:

Descriptive Statistics (25 Points)

The water reservoir of a little town was built in the 1960's and has to be modernized. For redimensioning the reservoir, the town's water consumption is being analyzed. Data has been collected over the last 40 years. For each year the average daily (x) and the maximum daily (y) water consumption (both in $liter/(day \cdot person)$) have been registered.

A) Draw the Tukey box plot in Figure 1 for the town's *mean* daily water consumption x_i, using the data given in Table 1, and indicate in the plot the following values: Median, upper and lower quantiles; interquartile range; upper and lower adjacent values; outside values (if there are no outside values, please remark it, too). What can you say about the skewness of the data?

Table 1: Water consumption during a 40-year period $\left[liter/(day \cdot person)\right]$.

		Mean values		Maximum values	
i	Year	Unordered data set x_i	Ordered data set x_i^0	Unordered data set y_i	Ordered data set y_i^0
1	1969	306	305	470	463
2	1971	317	306	465	465
3	1973	400	310	597	470
4	1975	361	312	643	473
5	1977	360	317	533	477
6	1979	367	323	530	483
7	1981	364	329	545	491
8	1983	426	329	611	530
9	1985	397	353	611	533
10	1987	420	360	562	545
11	1989	378	361	590	559
12	1991	395	364	605	562
13	1993	353	367	559	582
14	1995	329	378	477	590
15	1997	323	395	473	597
16	1999	312	397	491	605
17	2001	329	400	582	611
18	2003	310	420	463	611
19	2005	305	426	483	643

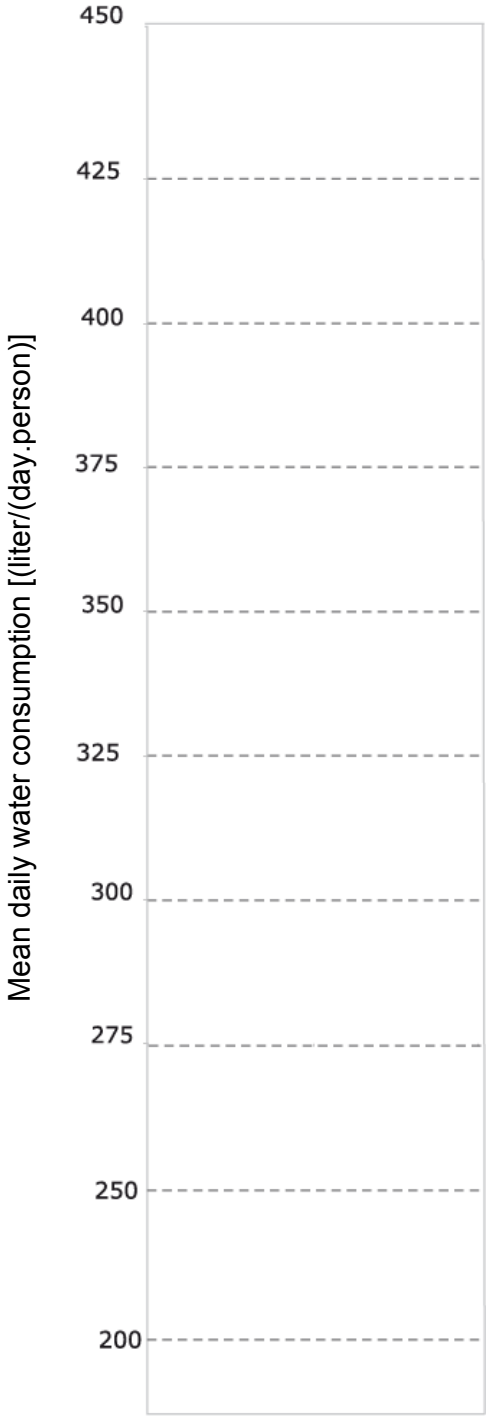


Figure 1: Tukey Box Plot.

B) Assess the correlation between the mean and the maximum value, based on the Quantile-Quantile-Plot in Figure 2.

1. What can you say by just looking at the Figure 2 (mark the right answer(s)):

- ☐ There is no correlation.
- ☐ There is a negative correlation.
- ☐ There is a positive correlation.

2. Calculate the correlation coefficient based on the values given in Table 2. Fill out the table *completely* with an accuracy of one decimal place.

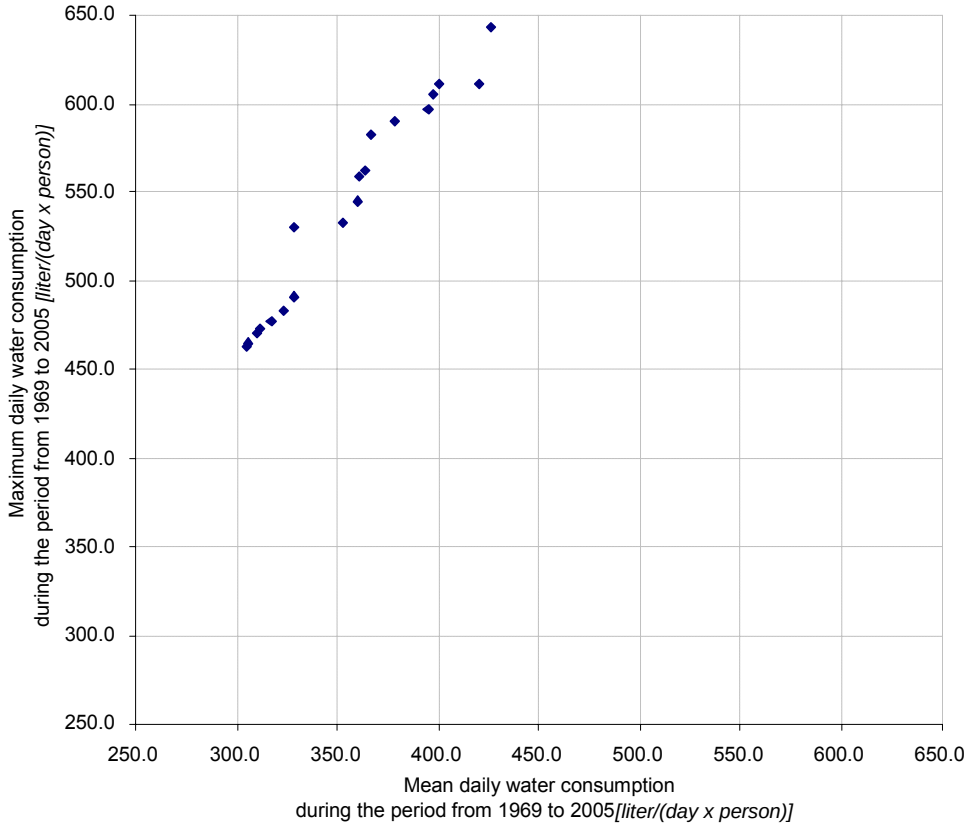


Figure 2: Quantile-Quantile-Plot of the annual mean and maximum water consumption $\left[liter/(day \cdot person)\right]$*; period from 1969 to 2005.*

Table 2: Calculation table for the assessment of the correlation coefficient .

Year	i	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1969	1	306	470					
1977	2	360	533					
1985	3	397	611					
1989	4	378	590					
1993	5	353	559					
1997	6	323	473					
2001	7	329	582					
2005	8	305	483					
Σ	-	2751.0	4301.0	-	-			
Σ / n	-	343.9	537.6	-	-			
$\sqrt{\Sigma / n}$	-	-	-	-	-			-

Exercise 2:

Bayes' theorem (25 Points)

Total Dissolved Solids (TDS) in water is a measure of the amount of minerals, salts, metals and any other solid particles that are dissolved in water. The measurement of TDS is essential to evaluate the quality of water, as water with high TDS often has a bad taste and/or a high level of hardness. The amount of TDS in water is generally classified in one of the following three categories :

C_1 : less than 100 parts per million (ppm)

C_2 : between 100 ppm and 300 ppm and

C_3 : greater than 300 ppm

A new testing equipment to measure the amount of TDS in water is now being considered for use in the public water supply system for a city.

Based on the available historical records from the water supply to the city, the prior probabilities that the amount of TDS belongs to each of the three categories have been determined and are listed below, going from the lowest category C_1 to the highest category C_3:

$$P(C_1) = 0.53$$
$$P(C_2) = 0.35$$
$$P(C_3) = 0.12$$

The manufacturer of this new testing equipment provides information about the accuracy of the equipment in the form of the following table.

Table 3: Accuracy of the new testing equipment.

Amount of TDS in water	Indications I_1, I_2, I_3 shown by the testing equipment		
	I_1: TDS < 100 ppm	I_2: 100 ppm ≤ TDS ≤ 300 ppm	I_3: TDS > 300 ppm
C_1: TDS < 100 ppm	0.81		0.12
C_2: 100 ppm ≤ TDS ≤ 300 ppm		0.87	0.09
C_3: TDS > 300 ppm	0.01		0.95

A) Table 3 is incomplete. Please fill in the missing values.

B) The new testing equipment was used to measure the TDS in a given water sample. The test indicated the amount of TDS in the water to be between 100 ppm and 300 ppm. Determine the probabilities that the amount of TDS in water is i) < 100 ppm, ii) between 100 ppm and 300 ppm and iii) > 300 ppm.

C) What is the probability that the indication shown by the testing equipment correctly corresponds to the actual category of TDS in water?

(Note: Part **C)** can be solved independently from parts **B)** and **D)**)

D) What is the probability that the indication shown by the testing equipment corresponds to the correct or a higher category compared to the actual category of TDS in water?

(Note: Part **D)** can be solved independently from parts **B)** and **C)**)

Exercise 3:

Probability paper **(15 Points)**

The sewage water system in a city is planned to be renewed due to frequent flood events. In order to assess the rainwater discharge flow, the records of the annual maximum amount of rainfall (precipitation) [mm/hour] for the last 10 years are considered, see Table 4. An engineer is modeling the annual maximum precipitation using a log-Normal distribution. It is assumed that the annual maximum precipitations can be considered as independent realizations from the same distribution. Please answer the following questions.

A) Plot the precipitation data from Table 4 on the provided probability paper for a log-Normal distribution (Figure 3). The horizontal and vertical axes should correspond to $\ln x$ and $\Phi^{-1}(F(x))$. Use Table 5 for the necessary calculations.

B) Assume that the log-Normal distribution is an appropriate distribution for modeling the annual maximum precipitation.

1. How can you express, for the line in Figure 3, the gradient and the intersection with the horizontal axis ($\Phi^{-1}(F(x)) = 0$) using the two parameters λ and ζ ?

2. Estimate graphically the parameters λ and ζ .

Hint: The cumulative distribution function of a log-Normal distribution may be expressed as:

$$F_X(x) = \Phi\left(\frac{\ln x - \lambda}{\zeta}\right),\ x,\ \zeta > 0 \qquad (3.1)$$

where $\Phi(\cdot)$ is the Standard Normal cumulative distribution function, and λ and ζ are the parameters of the distribution. Note that equation (3.1) can be reformulated as:

$$\Phi^{-1}(F_X(x)) = \frac{\ln x - \lambda}{\zeta},\ x,\ \zeta > 0 \qquad (3.2)$$

where $\Phi^{-1}(\cdot)$ is the inverse function of $\Phi(\cdot)$.

Table 4: Annual maximum precipitation in the last 10 years.

Year (i)	Precipitation x_i [mm/hour]
1	37.5
2	20.5
3	31.3
4	21.3
5	39.0
6	53.9
7	19.5
8	41.2
9	70.8
10	24.3

Table 5: Calculation sheet.

i	$i/11$	$\Phi^{-1}(i/11)$	Precipitation x_i [mm/hour] (sorted)	Logarithm of precipitation x_i
1			19.5	
2			20.5	
3			21.3	
4			24.3	
5			31.3	
6			37.5	
7			39.0	
8			41.2	
9			53.9	
10			70.8	

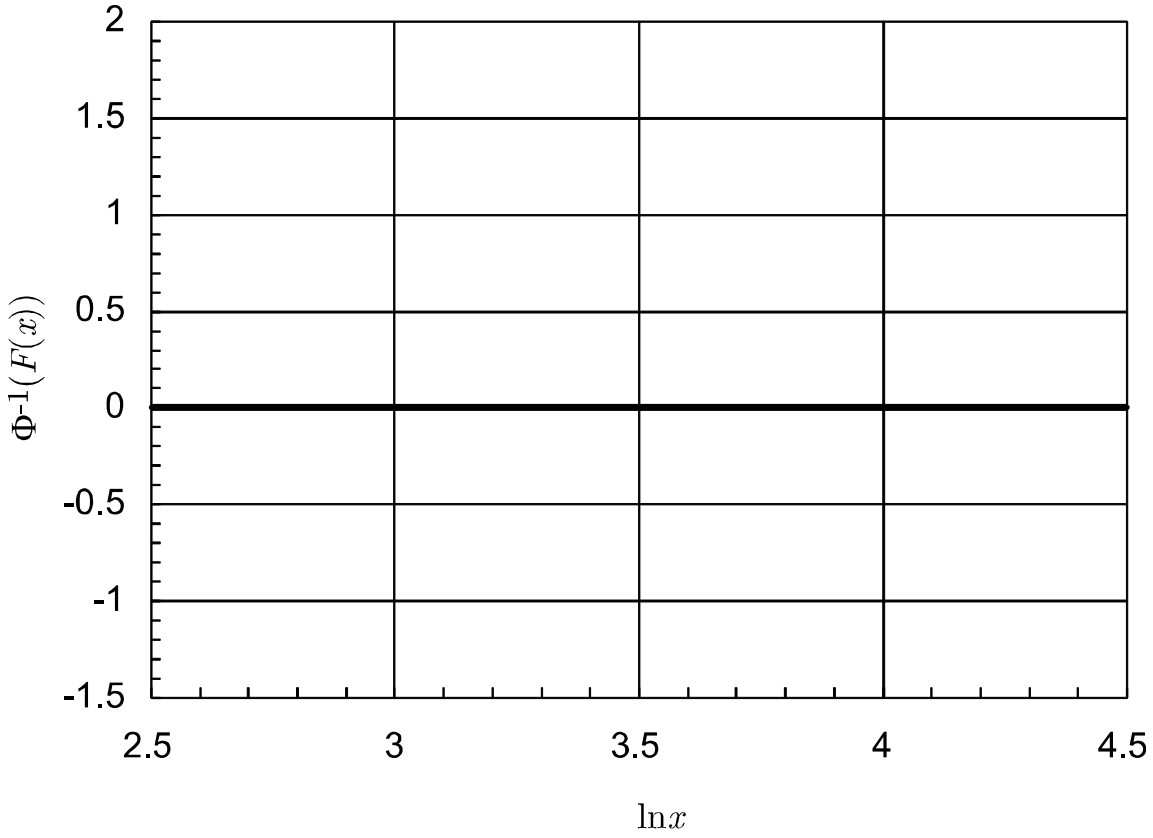


Figure 3: Probability paper.

Exercise 4:

Confidence intervals and Chi-Square Test (25 Points)

For the expansion of the water supply system in the area of the airport of Zurich the depth of a soil layer is of interest. The depth of the soil layer of interest is to be measured at n different locations using a sonic instrument. Based on experience from measurements with this instrument, the site engineer models each depth measurement X as a Normal distributed random variable with mean μ_X and standard deviation $\sigma_X = 0.4$ meters.

A) Express the mean and standard deviation of the estimator $\overline{X}$ in terms of μ_X, σ_X and n.

B) $n = 50$ depth measurements are carried out and the sample average is calculated equal to $\bar{x} = 5$ meters. Estimate the double sided interval that will contain the mean μ_X with a probability of 90%.

C) The soil depth measurements have been arranged into intervals as shown in Table 6. Carry out a Chi-square test, at the 5% significance level, of the null hypothesis that the depth measurements follow a Normal distribution with the assumed parameters $\mu_X = 5$ meters for the mean value and $\sigma_X = 0.4$ meters for the standard deviation. Use Table 6 for the necessary calculations.

The sample statistic used in the Chi-square test is written as: $\varepsilon_m^2 = \sum_{j=1}^{k} \frac{\left(N_{o,j} - N_{p,j}\right)^2}{N_{p,j}}$ where k is the number of intervals containing the observed values.

Table 6: Chi-square test.

Interval of depth (meters) x_j	Number of observed values $N_{o,j}$	Predicted probability $p(x_i)$	Predicted number of observations $N_{p,j} = np(x_j)$	Sample statistic
0-4.5	4			
4.5-5.0	15			
5.0-5.5	24			
5.5-6.0	6			
6.0-∞	1			
				ε_m^2 =

D) In question **C)**, a Normal distribution with mean $\mu_X = 5$ meters and standard deviation $\sigma_X = 0.4$ meters has been postulated as representative for the soil depth data. Let us call this model I. Assume now that a model II is postulated as representative for the soil depth data. The model assumes that a Normal distribution with mean $\mu_X = 5.1$ meters and standard deviation $\sigma_X = 0.5$ meters is representative for the soil depth data. Both distribution parameters in model II are calculated using the soil depth data. The site engineer carries out again the Chi-square test, at the 5% significance level and estimates the Chi-square sample statistic. Furthermore, she calculates the sample likelihood for both models. This information is summarized in Table 7.

Table 7: Sample likelihood for the two models I and II.

Model	**Degrees of freedom**	**Chi-square sample statistic**	**Sample likelihood**
I		4.114	0.634
II		2.936	0.575

1. Please fill out the missing information in the table.

2. Can the engineer accept at the 5% significance level the null hypothesis that the distribution described by model II is representative for the soil depth data?

3. Which one of the two models is more suitable for modeling the soil depth in the area of the Zurich airport?

Hint: A), **B)**, **C)** and **D)** can be solved independently from one another.

Exercise 5:

Sum of Exponential distributed random variables (30 Points)

The amount of water in a reservoir is relevant for the assessment of a drinking water supply system. To assess the risk of water shortage due to lack of water in the reservoir, the duration between two rainfall events is considered. The duration between two successive rainfall events is assumed to follow an Exponential distribution with the mean value of 10 days. Furthermore, it is assumed that durations between events are independent. Let T_1 represent the duration between the 1st and the 2nd rainfall events, T_2 the duration between the 2nd and the 3rd rainfall events, respectively, and $T = T_1 + T_2$. Answer the following questions.

Hint: The cumulative distribution function of an Exponential distributed random variable is expressed as:

$$F_T(t) = 1 - \exp(-\lambda t),\ \lambda, t \geq 0 \qquad (5.1)$$

where λ is the parameter, and the mean value of the random variable is $1/\lambda$.

A) Calculate the mean value of the random variable T.

B) To which distribution family does the random variable T belong?

C) Calculate the probability density function of the random variable T.

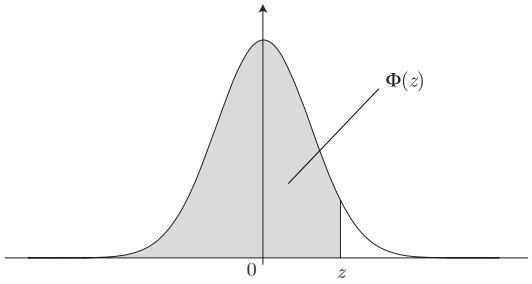
D) Calculate the cumulative distribution function of the random variable T.

E) Calculate the probability of $T \leq 20$ days.

F) For a series of n rainfall events, the sum of all durations between successive rainfall events is now being considered. This can be represented by the variable $Y = T_1 + T_2 + T_3 + \ldots\ldots\ldots\ldots\ldots + T_{n-2} + T_{n-1}$, where T_i is the duration between the i^{th} and the $(i+1)^{th}$ rainfall events. Assuming that n is large, derive an expression for the exceedance probability $P(Y > y),\ y \geq 0$ in terms of y and n.

Annexes: Tables

Cumulative distribution function of the Standard Normal distribution $\Phi(z)$.

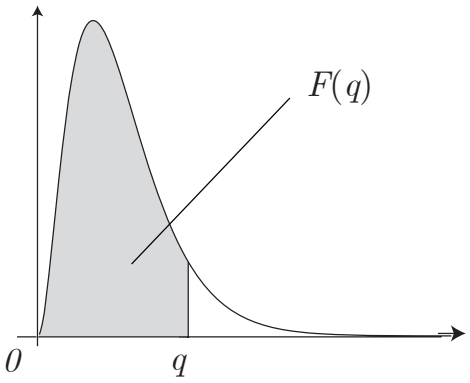


Probability density function of the standard normal random variable.

z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$
0.00	0.5000	0.50	0.6915	1.00	0.8413	1.50	0.9332	2.00	0.9772
0.01	0.5040	0.51	0.6950	1.01	0.8438	1.51	0.9345	2.10	0.9821356
0.02	0.5080	0.52	0.6985	1.02	0.8461	1.52	0.9357	2.20	0.9860966
0.03	0.5120	0.53	0.7019	1.03	0.8485	1.53	0.9370	2.30	0.9892759
0.04	0.5160	0.54	0.7054	1.04	0.8508	1.54	0.9382	2.40	0.9918025
0.05	0.5199	0.55	0.7088	1.05	0.8531	1.55	0.9394	2.50	0.9937903
0.06	0.5239	0.56	0.7123	1.06	0.8554	1.56	0.9406	2.60	0.9953388
0.07	0.5279	0.57	0.7157	1.07	0.8577	1.57	0.9418	2.70	0.9965330
0.08	0.5319	0.58	0.7190	1.08	0.8599	1.58	0.9429	2.80	0.9974449
0.09	0.5359	0.59	0.7224	1.09	0.8621	1.59	0.9441	2.90	0.9981342
0.10	0.5398	0.60	0.7257	1.10	0.8643	1.60	0.9452	3.00	0.9986501
0.11	0.5438	0.61	0.7291	1.11	0.8665	1.61	0.9463	3.10	0.9990324
0.12	0.5478	0.62	0.7324	1.12	0.8686	1.62	0.9474	3.20	0.9993129
0.13	0.5517	0.63	0.7357	1.13	0.8708	1.63	0.9484	3.30	0.9995166
0.14	0.5557	0.64	0.7389	1.14	0.8729	1.64	0.9495	3.40	0.9996631
0.15	0.5596	0.65	0.7422	1.15	0.8749	1.65	0.9505	3.50	0.9997674
0.16	0.5636	0.66	0.7454	1.16	0.8770	1.66	0.9515	3.60	0.9998409
0.17	0.5675	0.67	0.7486	1.17	0.8790	1.67	0.9525	3.70	0.9998922
0.18	0.5714	0.68	0.7517	1.18	0.8810	1.68	0.9535	3.80	0.9999277
0.19	0.5753	0.69	0.7549	1.19	0.8830	1.69	0.9545	3.90	0.9999519
0.20	0.5793	0.70	0.7580	1.20	0.8849	1.70	0.9554	4.00	0.9999683
0.21	0.5832	0.71	0.7611	1.21	0.8869	1.71	0.9564	4.10	0.9999793
0.22	0.5871	0.72	0.7642	1.22	0.8888	1.72	0.9573	4.20	0.9999867
0.23	0.5910	0.73	0.7673	1.23	0.8907	1.73	0.9582	4.30	0.9999915
0.24	0.5948	0.74	0.7704	1.24	0.8925	1.74	0.9591	4.40	0.9999946
0.25	0.5987	0.75	0.7734	1.25	0.8944	1.75	0.9599	4.50	0.9999966
0.26	0.6026	0.76	0.7764	1.26	0.8962	1.76	0.9608	4.60	0.9999979
0.27	0.6064	0.77	0.7794	1.27	0.8980	1.77	0.9616	4.70	0.9999987
0.28	0.6103	0.78	0.7823	1.28	0.8997	1.78	0.9625	4.80	0.9999992
0.29	0.6141	0.79	0.7852	1.29	0.9015	1.79	0.9633	4.90	0.9999995
0.30	0.6179	0.80	0.7881	1.30	0.9032	1.80	0.9641	5.00	0.9999997
0.31	0.6217	0.81	0.7910	1.31	0.9049	1.81	0.9649		
0.32	0.6255	0.82	0.7939	1.32	0.9066	1.82	0.9656		
0.33	0.6293	0.83	0.7967	1.33	0.9082	1.83	0.9664		
0.34	0.6331	0.84	0.7995	1.34	0.9099	1.84	0.9671		
0.35	0.6368	0.85	0.8023	1.35	0.9115	1.85	0.9678		
0.36	0.6406	0.86	0.8051	1.36	0.9131	1.86	0.9686		
0.37	0.6443	0.87	0.8078	1.37	0.9147	1.87	0.9693		
0.38	0.6480	0.88	0.8106	1.38	0.9162	1.88	0.9699		
0.39	0.6517	0.89	0.8133	1.39	0.9177	1.89	0.9706		
0.40	0.6554	0.90	0.8159	1.40	0.9192	1.90	0.9713		
0.41	0.6591	0.91	0.8186	1.41	0.9207	1.91	0.9719		
0.42	0.6628	0.92	0.8212	1.42	0.9222	1.92	0.9726		
0.43	0.6664	0.93	0.8238	1.43	0.9236	1.93	0.9732		
0.44	0.6700	0.94	0.8264	1.44	0.9251	1.94	0.9738		
0.45	0.6736	0.95	0.8289	1.45	0.9265	1.95	0.9744		
0.46	0.6772	0.96	0.8315	1.46	0.9279	1.96	0.9750		
0.47	0.6808	0.97	0.8340	1.47	0.9292	1.97	0.9756		
0.48	0.6844	0.98	0.8365	1.48	0.9306	1.98	0.9761		
0.49	0.6879	0.99	0.8389	1.49	0.9319	1.99	0.9767		

Annexes: Tables

Quantile values of the Chi-square distribution q.



Probability density function of Chi-square distribution.

ν	***F(q)*=0.75**	0.90	0.95	0.98	0.99	0.995	0.999
1	1.3233	2.7055	3.8415	5.4119	6.6349	7.8794	10.8276
2	2.7726	4.6052	5.9915	7.8240	9.2103	10.5966	13.8155
3	4.1083	6.2514	7.8147	9.8374	11.3449	12.8382	16.2662
4	5.3853	7.7794	9.4877	11.6678	13.2767	14.8603	18.4668
5	6.6257	9.2364	11.0705	13.3882	15.0863	16.7496	20.5150
6	7.8408	10.6446	12.5916	15.0332	16.8119	18.5476	22.4577
7	9.0371	12.0170	14.0671	16.6224	18.4753	20.2777	24.3219
8	10.2189	13.3616	15.5073	18.1682	20.0902	21.9550	26.1245
9	11.3888	14.6837	16.9190	19.6790	21.6660	23.5894	27.8772
10	12.5489	15.9872	18.3070	21.1608	23.2093	25.1882	29.5883
11	13.7007	17.2750	19.6751	22.6179	24.7250	26.7568	31.2641
12	14.8454	18.5493	21.0261	24.0540	26.2170	28.2995	32.9095
13	15.9839	19.8119	22.3620	25.4715	27.6882	29.8195	34.5282
14	17.1169	21.0641	23.6848	26.8728	29.1412	31.3193	36.1233
15	18.2451	22.3071	24.9958	28.2595	30.5779	32.8013	37.6973
16	19.3689	23.5418	26.2962	29.6332	31.9999	34.2672	39.2524
17	20.4887	24.7690	27.5871	30.9950	33.4087	35.7185	40.7902
18	21.6049	25.9894	28.8693	32.3462	34.8053	37.1565	42.3124
19	22.7178	27.2036	30.1435	33.6874	36.1909	38.5823	43.8202
20	23.8277	28.4120	31.4104	35.0196	37.5662	39.9968	45.3147
21	24.9348	29.6151	32.6706	36.3434	38.9322	41.4011	46.7970
22	26.0393	30.8133	33.9244	37.6595	40.2894	42.7957	48.2679
23	27.1413	32.0069	35.1725	38.9683	41.6384	44.1813	49.7282
24	28.2412	33.1962	36.4150	40.2704	42.9798	45.5585	51.1786
25	29.3389	34.3816	37.6525	41.5661	44.3141	46.9279	52.6197
26	30.4346	35.5632	38.8851	42.8558	45.6417	48.2899	54.0520
27	31.5284	36.7412	40.1133	44.1400	46.9629	49.6449	55.4760
28	32.6205	37.9159	41.3371	45.4188	48.2782	50.9934	56.8923
29	33.7109	39.0875	42.5570	46.6927	49.5879	52.3356	58.3012
30	34.7997	40.2560	43.7730	47.9618	50.8922	53.6720	59.7031

ν : Degree of freedom.

Glossary

Assumption	Annahme
Accuracy	Genauigkeit, Präzision
Amount	Menge
Appropriate	passend
Depth	Tiefe
Decimal places	Nachkommastellen
Exceed	Überschreiten, grösser sein als
Expansion	Erweiterung
Gradient	Steigung
Intersection	Schnittpunkt, -linie, -menge
Locations	Ort, Position
Postulate	Postulieren, Vorschlagen
Rainwater discharge flow	Regenwasserabfluss
(to be) rebuilt	Umgebaut werden
Sewage water system	Kanalisation
Soil layer	Bodenschicht
Sonic	(Ultra-)Schall
Successive	Aufeinanderfolgend
Suitable	Angemessen, passend
(to be) stuck	Steckenbleiben
Total Dissolved Solids	TDS-Wert=Summe der gelösten Salze im Wasser; mg/l (Milligramm pro Liter)
Water Shortage	Wassermangel, Wasserknappheit
Water Supply System	Wasserversorgungssystem