OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG

INF FAKULTÄT FÜR INFORMATIK

Introduction to Simulation

Input Modelling

Input Data

Discrete-event simulations use random variables, e.g.

- Inter-arrival times
- Service times

In order to create a computer model, we need to ...

- Measure the data in the real system,
- Approximate the data by a distribution function,
- Use this function in the simulation program.

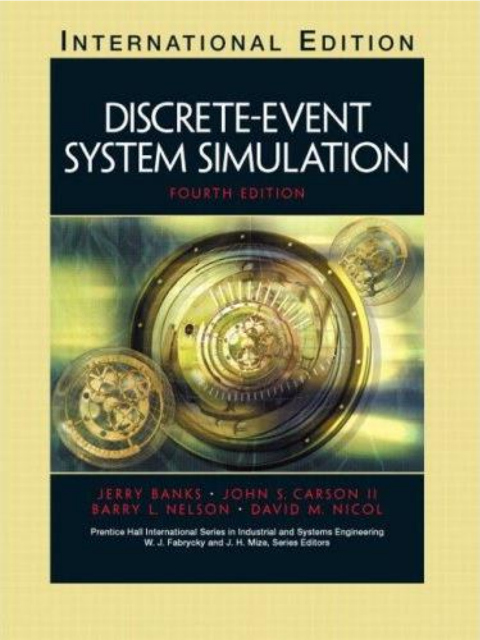
This procedure is called *input data analysis.*

- It is an example of idealisation.

Background Reading

Relevant sections of the book:

- 9.1
- 9.2
- 9.3
- 9.4



Input Data

The four steps involved in input data analysis are:

1. Collect data (measurement or guesswork).
2. Guess a distribution type for the data.
3. Find the specific parameters for the distribution.
4. Check the quality of the distribution obtained.

Collect Data

Collecting data involves lots of work.

It is difficult, expensive and error-prone.

It is where the GIGO principle most commonly applies.

- (GIGO = Garbage In, Garbage Out)

A common approach to collecting data:

- Hire somebody to sit in the bank
  with a stopwatch.

Collect Data

Raw data is a initially just a list of numbers.

- Service times at bank (cashier #1, seconds):

95	23	162	45	144	231	88	...

We wish to replace these numbers by a random variable ...

- that describes how these numbers are distributed.

During the simulation,

- the simulator can take samples from this RV whenever needed.

Collect Data

There are many difficulties in data collection:

- Understanding the structure of the system
- Determining when enough data has been collected
- Avoiding data censoring
- Detecting interdependencies
  (Activities that will be independent in the model may be interdependent in real life.)

Guess Distribution Type

The next step is to guess the distribution type for the data.

There are two sub-steps involved:

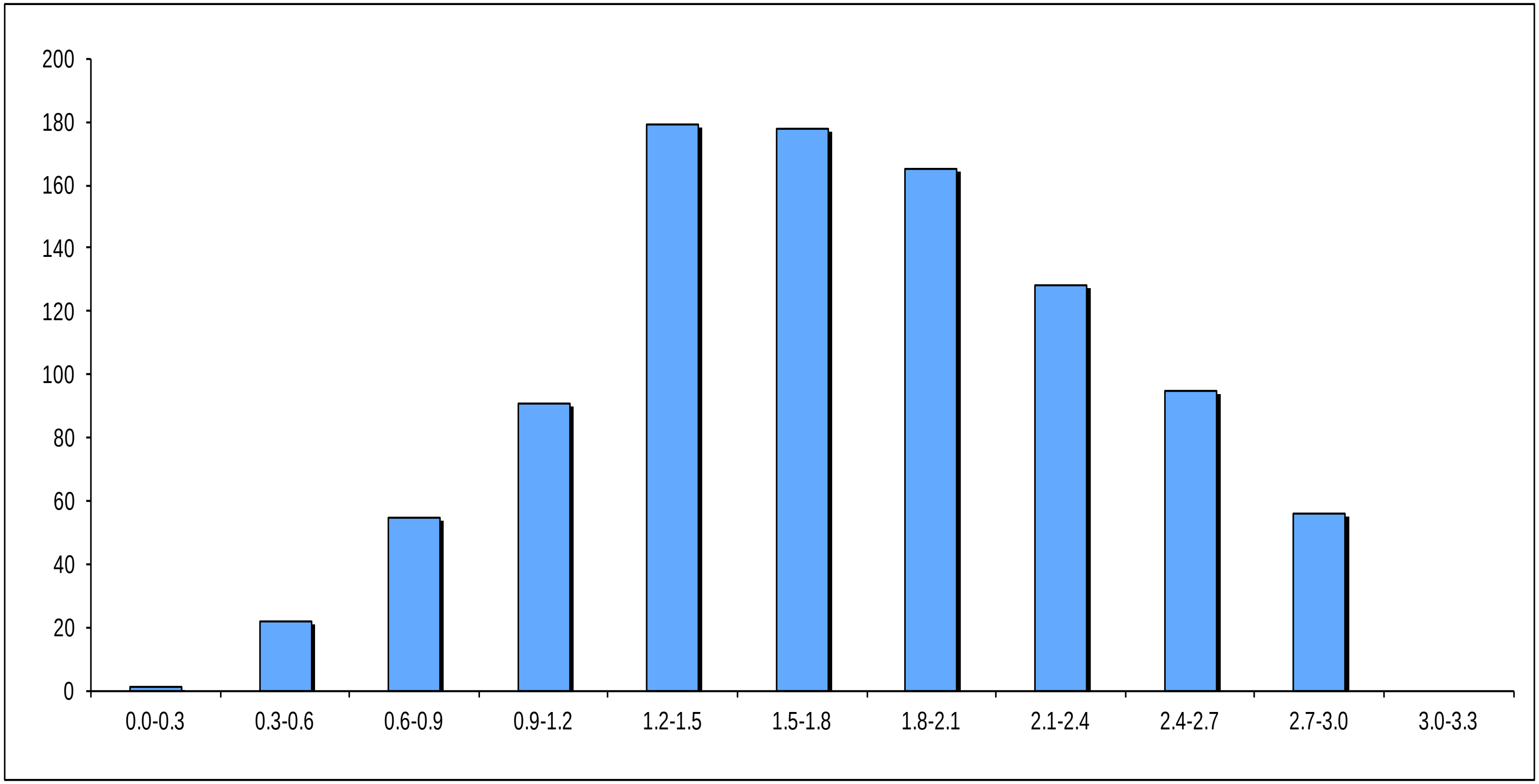
1. Make an educated guess at the distribution.
2. Check the correctness of the guess.

These two sub-steps are achieved by:

1. Drawing a histogram of the data
2. Generating a quantile-quantile plot

Draw a Histogram

A histogram can indicate the type of distribution.



Draw a Histogram

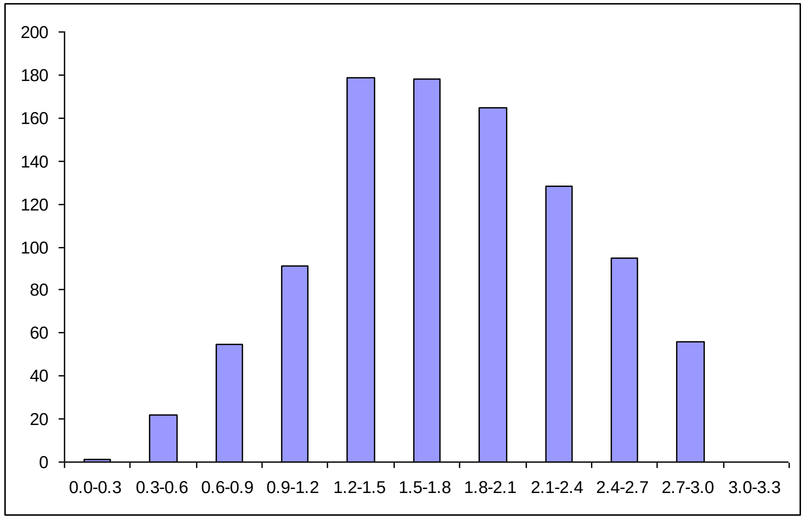
Steps in generating a histogram:

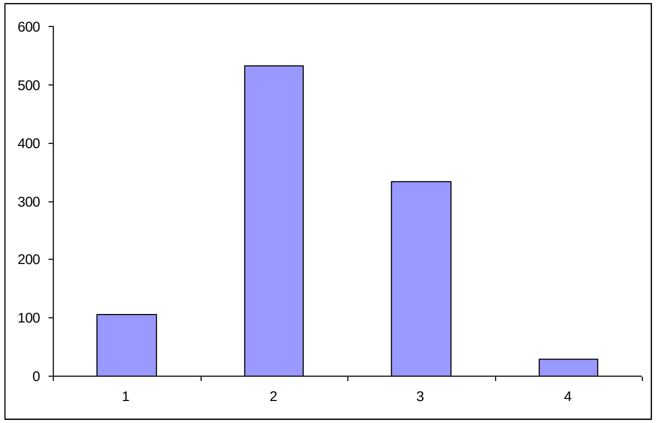
1. Divide the n items of data into k intervals.
2. Plot the number of items of data n_k in each interval.
3. Compare the shape of the histogram to known pdfs.

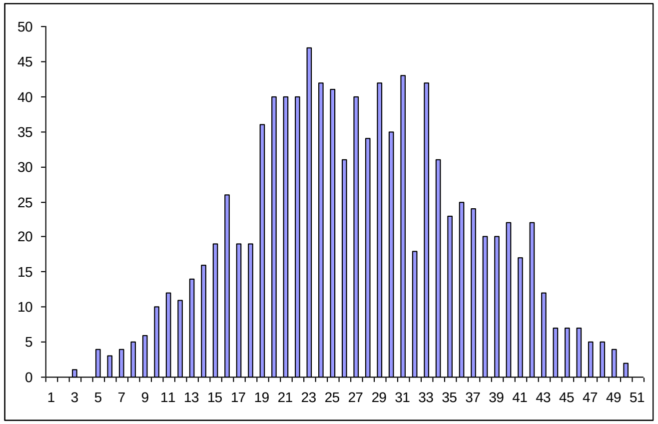
The number of intervals k must be chosen carefully.

Draw a Histogram

Choose $k \approx \sqrt{n}$ for best results:







Quantile–Quantile Plots

Q–Q plots test whether the type of distribution is correct.

They are useful when $n < 30$, when histograms are messy.

Given a random variable X with cdf F...

- then γ is the q – quantile of X: $F(\gamma) = P(X \leq \gamma) = q$

Quantile–Quantile Plots

Steps in generating a Q–Q plot:

1. Given the measured data x_i from the random variable X
2. Sort the x_i to give y_j, $y_1 \leq y_2 \ldots \leq y_n$
3. Plot y_j against $F^{-1}((j - ½) / n)$

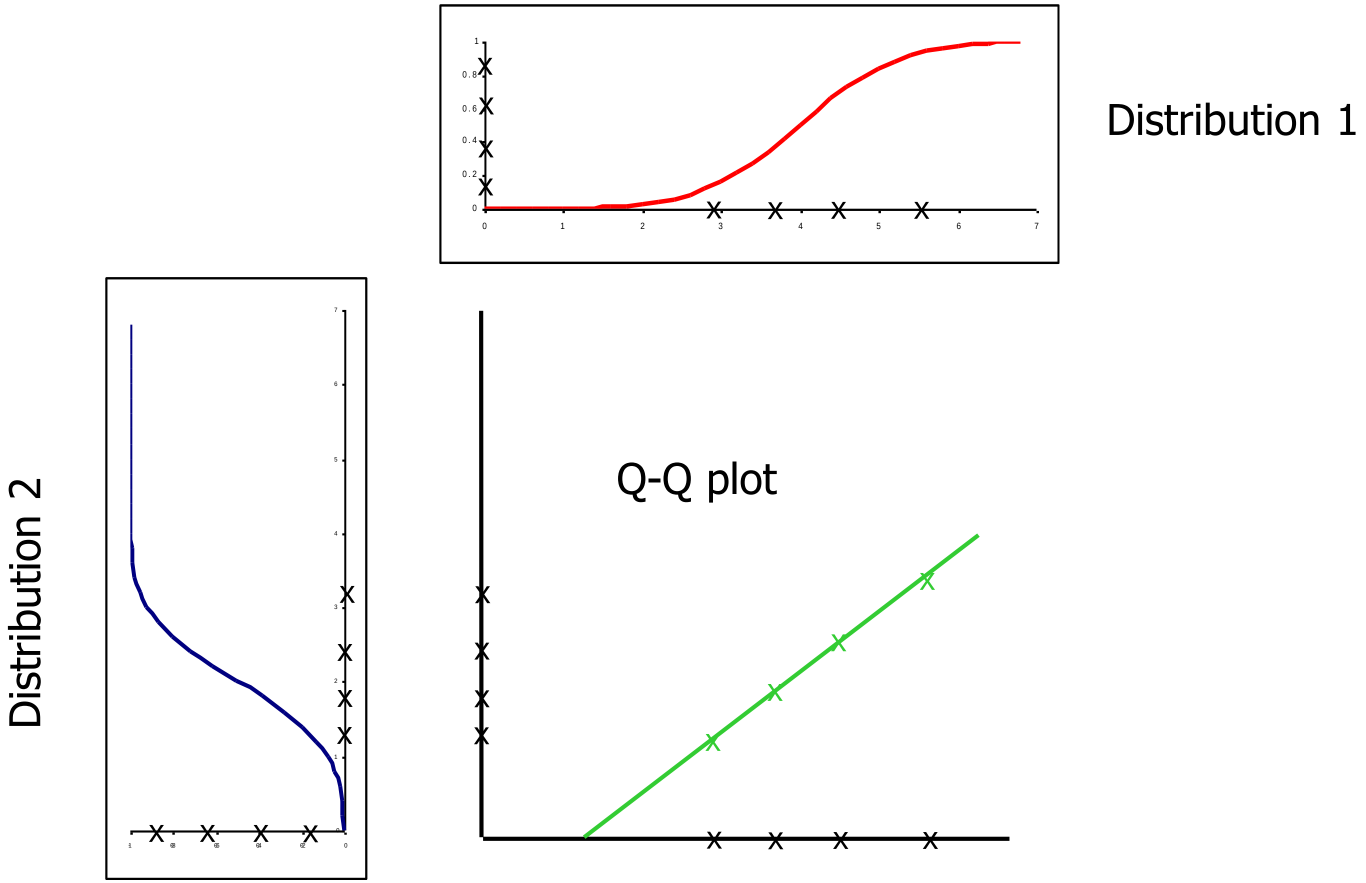
If the graph yields a straight line, then X is of type F.

In addition, the parameters are correct, if

- the slope of the line is 1, and
- the line passes through the origin.

Quantile–Quantile Plots

What are we doing here?

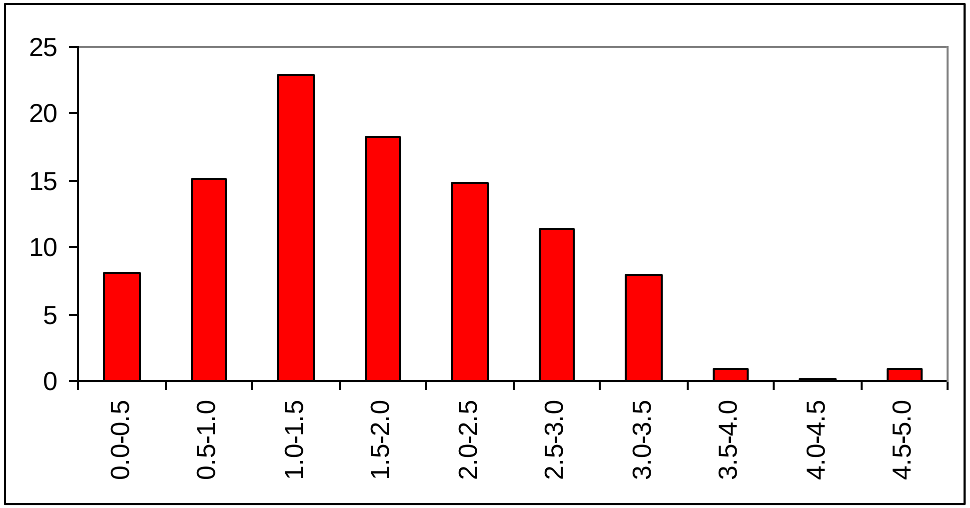


Example Q–Q Plot

Sort 100 failure times y_i for a mechanical component:

$$0.30 \qquad 1.21 \quad 0.88 \quad 3.02 \quad 1.45 \quad \ldots$$

Draw a histogram using $k = 10$:

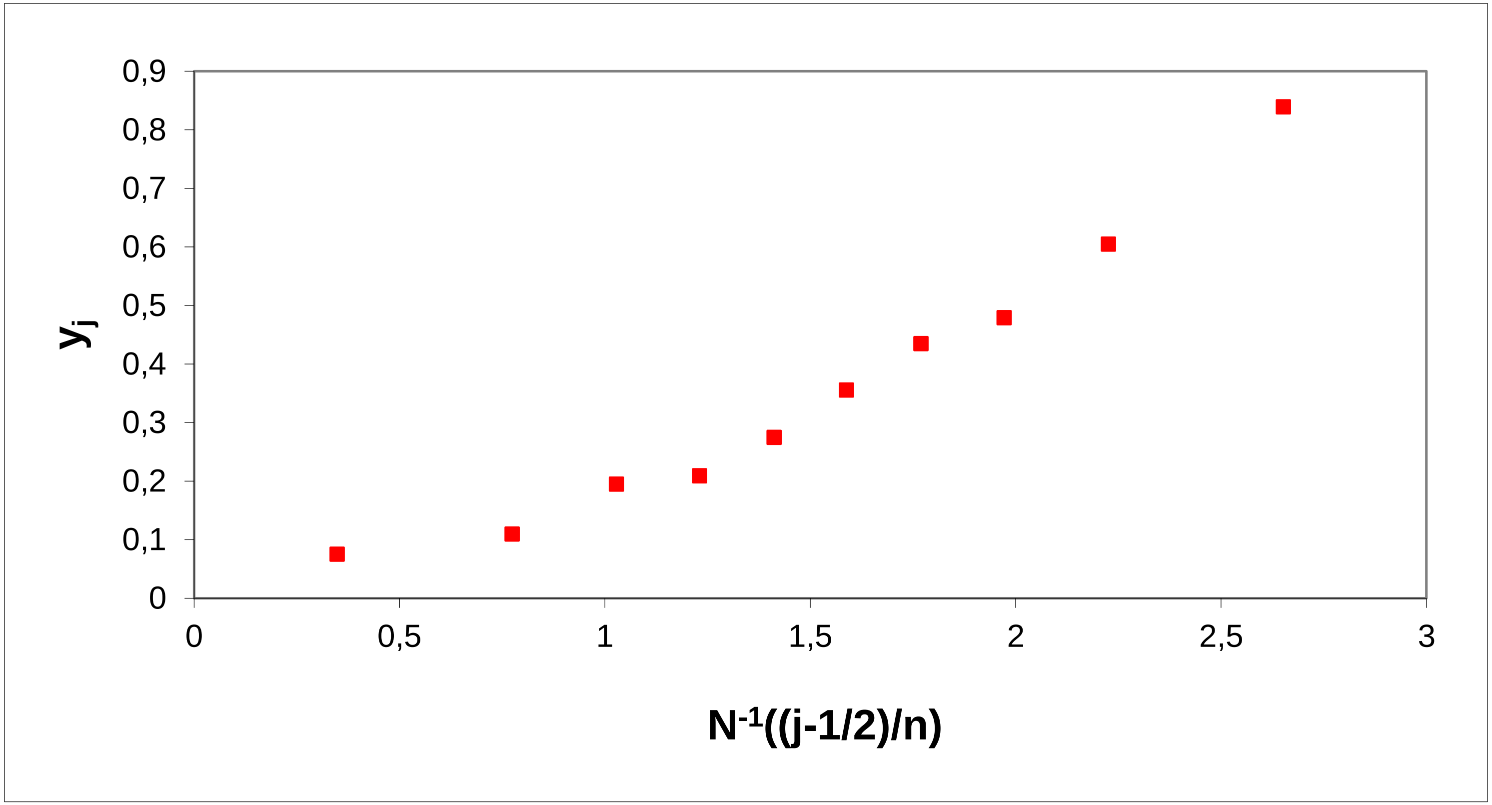


Guess at a normal distribution

Example Q–Q Plot

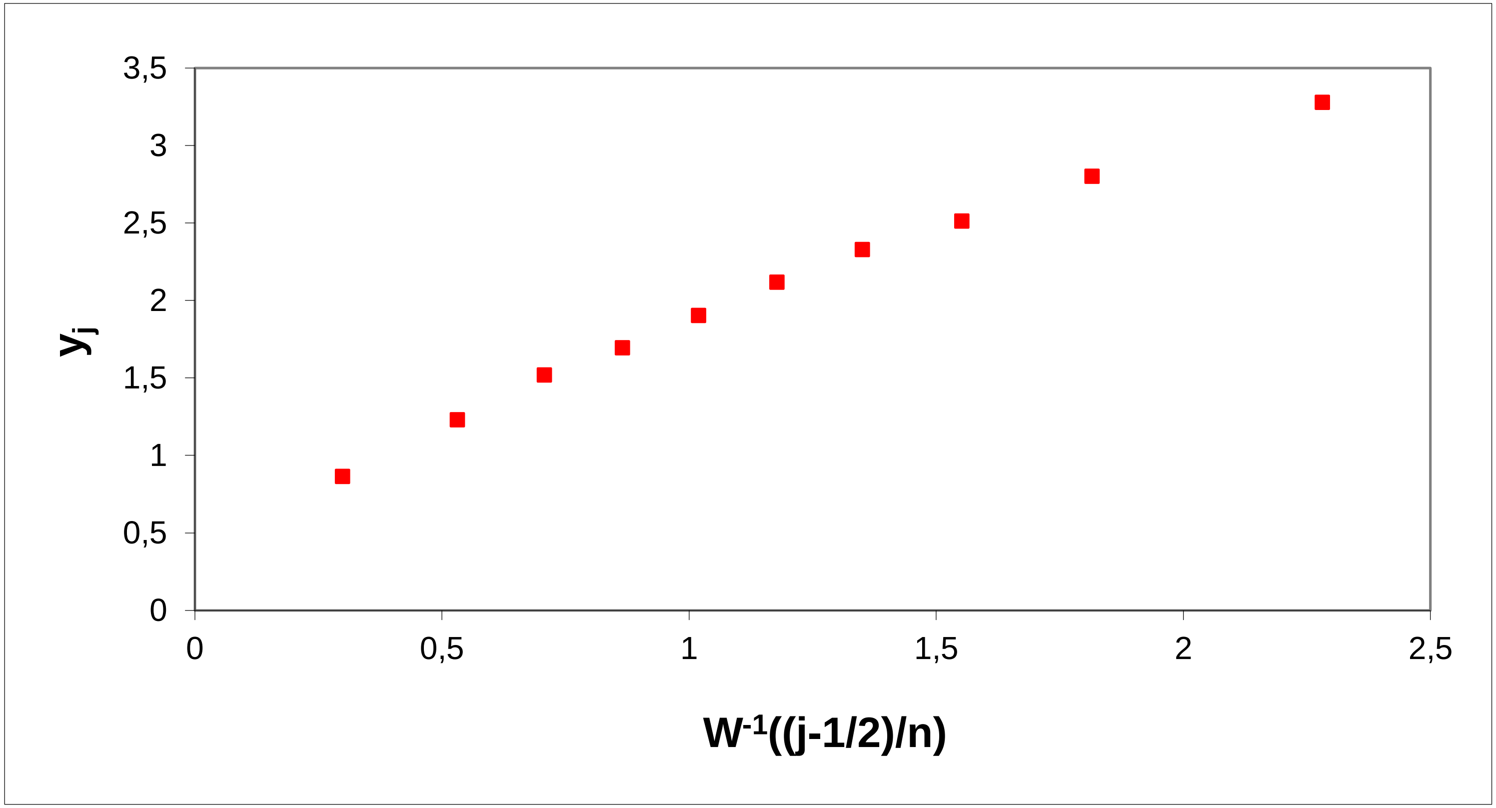
Choose a normal distribution for F: $F \sim N(1.5, 0.7)$

- Plot the y_j values against $F^{-1}((j - ½) / n)$:



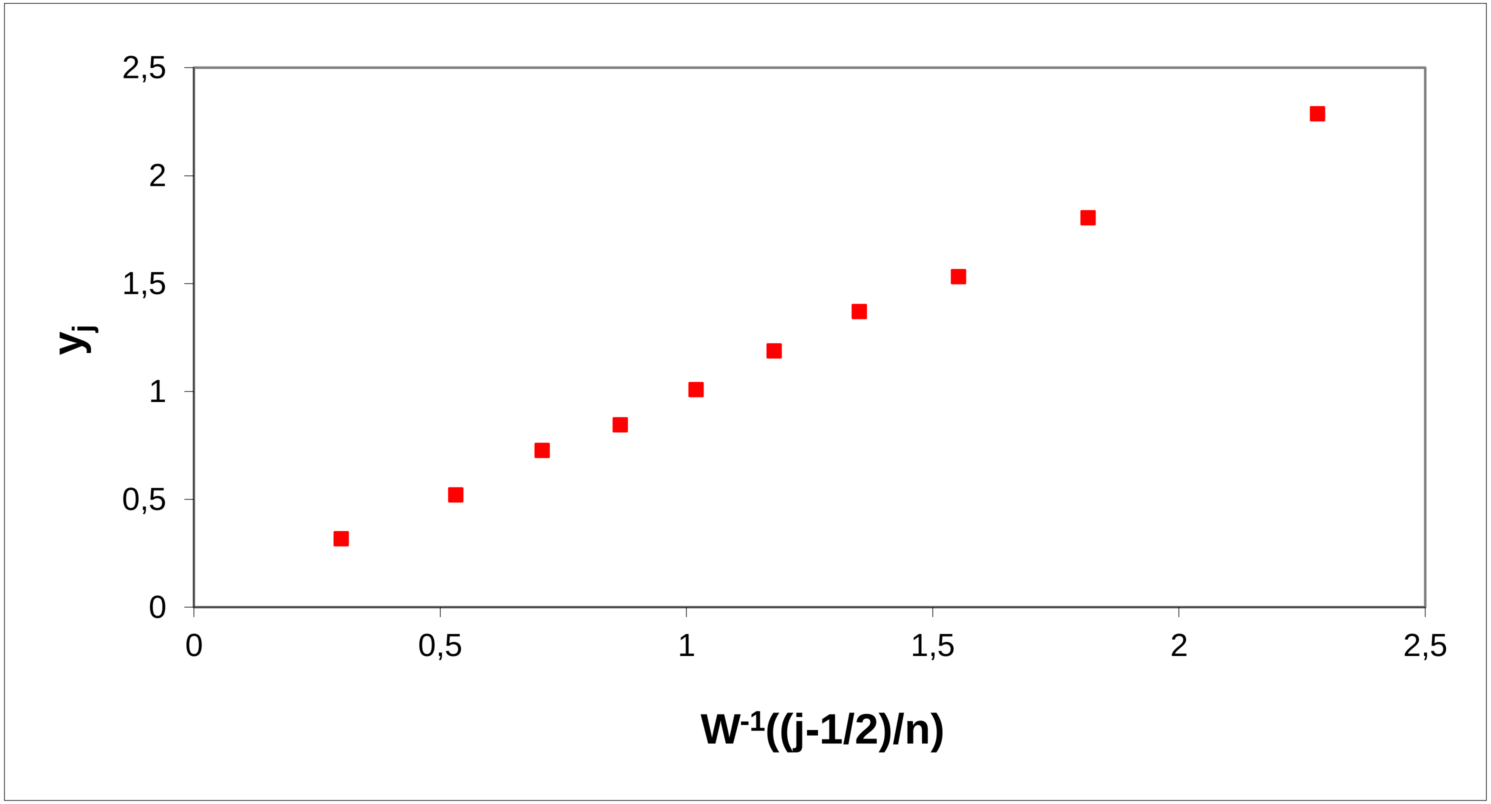
Example Q–Q Plot

Try again: Choose F to be $W(3.0, 3.0)$:



Example Q–Q Plot

Try again: Choose F to be $W(2.0, 2.0)$:



Parameter Estimation

Now we must estimate the parameters of the distribution

- Or: experiment with the Q–Q plot parameters!

Given samples X_i of the random variable X

Definition: Sample mean $\bar{X} = \frac{1}{n}\sum_{1}^{n} X_i$

Definition: Sample variance $S^2 = \frac{1}{n-1}\sum_{1}^{n}(X_i - \bar{X})^2$

Parameter Estimation

Use sample mean and sample variance

Exponential distribution: $\lambda = \frac{1}{\overline{X}}$

Normal distribution: $\mu = \overline{X}$

$$\sigma^2 = S^2$$

Uniform distribution on (0, b): $b = \frac{n+1}{n} \max\{X_i\}$

Weibull(α, β) distribution: Complicated – refer to literature!

Goodness-of-Fit Tests

Goodness-of-fit tests can be used on our distribution hypothesis:

- Kolmogorov-Smirnoff test
- Chi-squared test

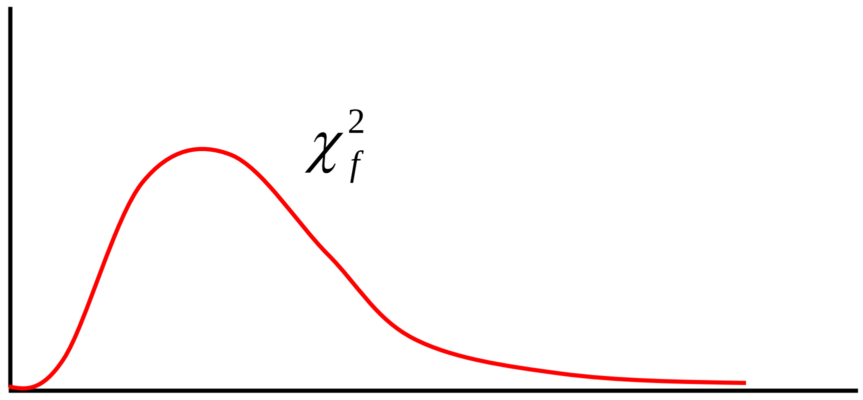
They test correspondence of data to hypothesised distribution.

They must be used with caution:

- They will accept almost anything for small n.
- They will reject almost everything for large n.

Chi–Square Test

The chi–squared distribution>



It has one parameter f (number of degrees of freedom).

The chi–squared test is used to compare two sets of data.

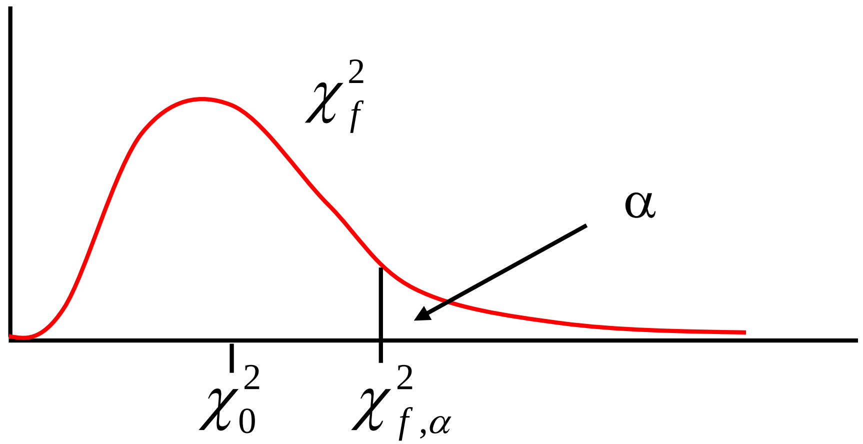
Chi-Square Test

Measure the difference χ_0^2 between guessed *F* and data.

- If F and data have the same distribution, then χ_0^2 ~ chi-squared distribution.

Hypothesis: Our data is ~ *F*

Choose $\alpha = P$(reject, although in fact correct)



If $\chi_0^2 > \chi_{f,\alpha}^2$ then reject the hypothesis, otherwise accept it.

Chi–Square Test

1. Divide the data into k classes.

2. Count the number of items of data O_i in each class i.

3. Determine the expected number of items of data E_i in each class i according to the hypothesised distribution.

4. Compute the test statistic $$\chi_0^2 = \sum_1^k \frac{(O_i - E_i)^2}{E_i}$$

…

Chi-Square Test

5. Choose a level of significance α.

6. Look up the value of $\chi^2_{f,\alpha}$, where $f = k - s - 1$.
   (s is the number of parameters of the distribution.)

7. If $\chi_0^2 > \chi^2_{f,\alpha}$ then reject the hypothesis.

N.B. Acceptance is *not* proof that the data ~ F!

- It only means we have *not been able to disprove it.*

Interpretation:

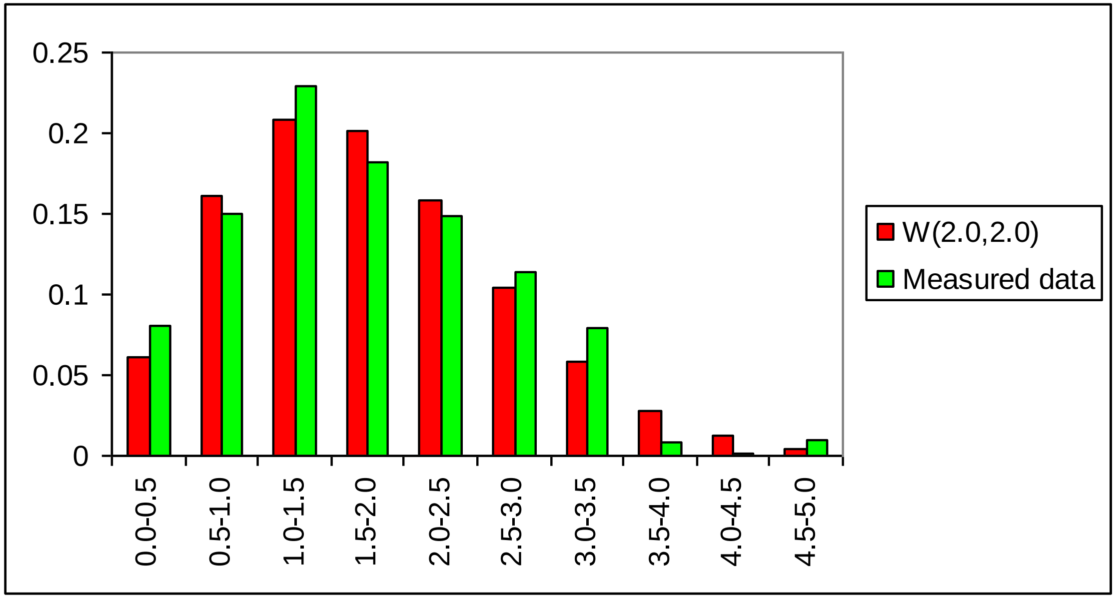
- *The difference between the two sets of data is not large enough for it to be unlikely that they have the same distribution.*

Example

Choose the data from slides 15–18 ($\sim W(2.0, 2.0)$)

Divide into $k = 10$ classes

Plot actual and expected number in each class



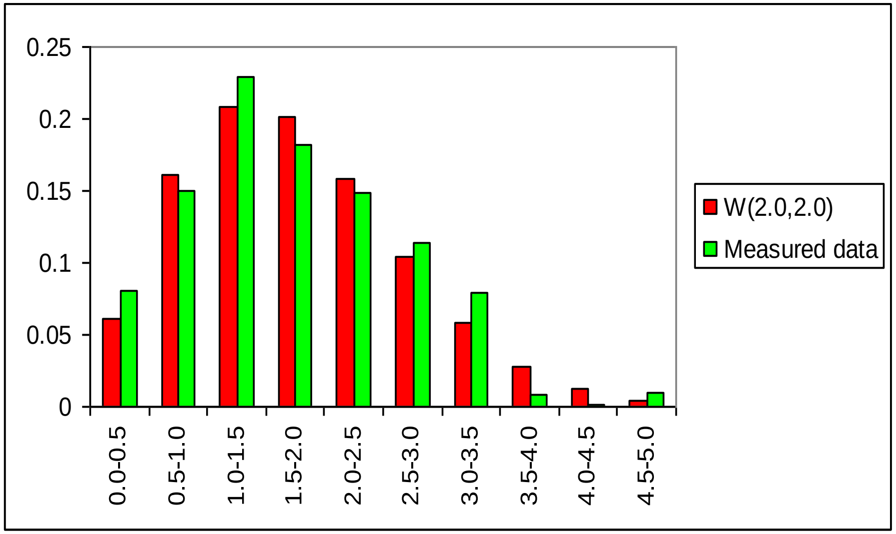
Example

Compute the test statistic $\chi_0^2 = 5.10$

Choose $\alpha = 0.01$

Set $f = k - s - 1$
$= 10 - 2 - 1 = 7$

Look up $\chi^2_{7,0.01} = 18.5$



We therefore accept the hypothesis

Chi-Square Test

Choice of k :

- $n < 20$: Do not use this test!
- $50 < n < 100$: $k = n/10 \ldots n/5$
- $n > 100$: $k = \sqrt{n} \ldots n/5$

$E_i < 5$ should be avoided

- Classes need not be of equal width!

Ideally, classes should be chosen to have equal probability.

Background

Argument:

- Consider a set of N randomly chosen people.
- You assume that males and females are equal in number.
- You would expect to count about $N/2$ males and $N/2$ females.

However,

- In general you will count $n_{male} \mathrel{!}= N/2$ and $n_{female} \mathrel{!}= N/2$.

The question

- How much may n_{male} and n_{female} differ from $N/2$ but still allow you to assume that males and females are equally distributed?

Background

Consider the quantities $\frac{\left(n_{male}-\frac{N}{2}\right)^2}{\frac{N}{2}}$ and $\frac{\left(n_{female}-\frac{N}{2}\right)^2}{\frac{N}{2}}$

If males and females are equally distributed, then

- it can be assumed that both terms are $N(0,1)$ distributed.

In that case, the sum $\chi_0^{\,2}=\frac{\left(n_{male}-\frac{N}{2}\right)^2}{\frac{N}{2}}+\frac{\left(n_{female}-\frac{N}{2}\right)^2}{\frac{N}{2}}$

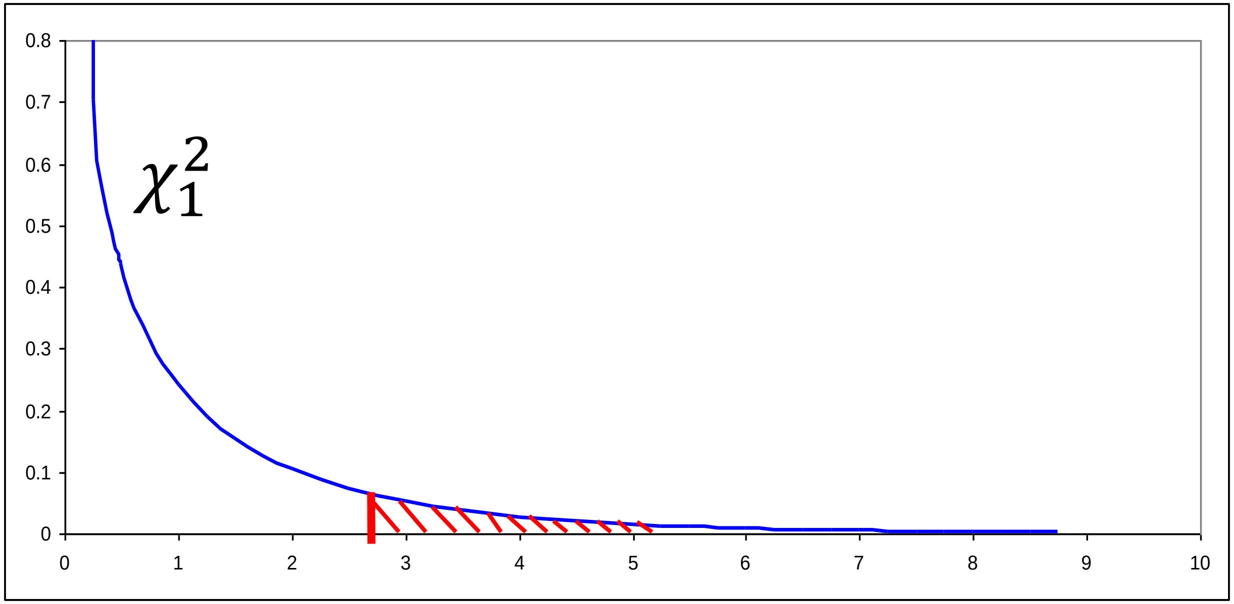
- is χ^2 distributed with one degree of freedom.

Background

Example: Set $N = 100$, $n_{male} = 45$, $n_{female} = 55$

- Then $\chi_0^{\ 2} = \frac{(-5)^2}{50} + \frac{5^2}{50} = 1$

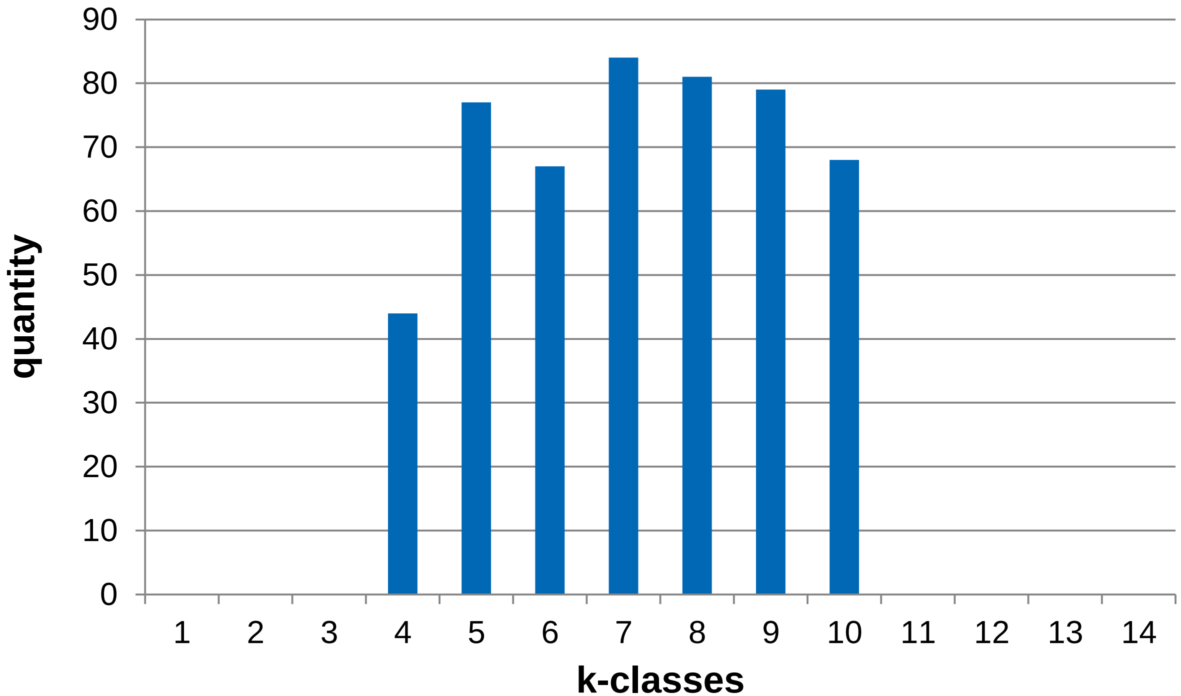
A probability of 10% is reached at about 2.75: $\chi^2_{1,0.1} \approx 2.75$



The Sims – Almost Normal Family Life

List of samples of employments duration of IT–consultants from last year, units are days:

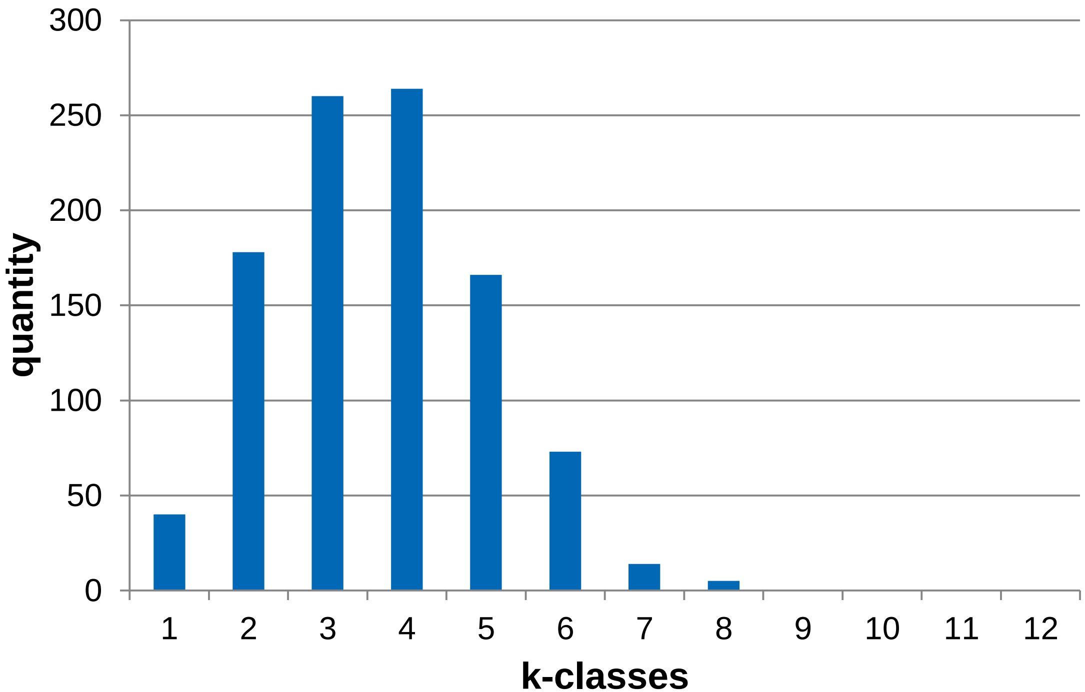
- 13.8, 24.02, 29.65, 14.85, 13.93, 13.31, …



Star Trek – Enterprise in Danger

Recorded times for treating crew members who have been exposed to engine-room plasma:

- 0.03, 1.6, 4.01, 2.82, 3.92, 0.46, 1.8, …



Simulation Project

Goals of the course:

- Put your knowledge of simulation into practice
- Practice many soft skills
- Experience a real-life project

FIN Bachelor students can count the course as ...

- WPF FIN SMK
- Wahlpflichtfach

Past Projects

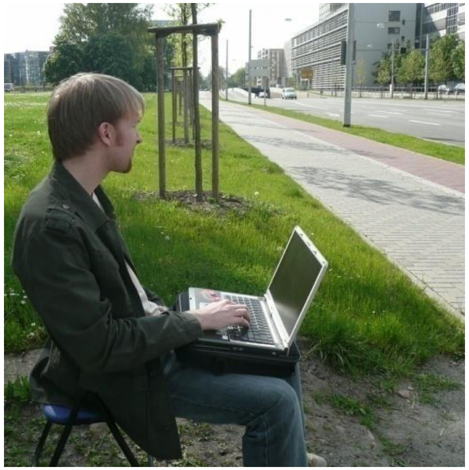
- Town planning in Magdeburg
- Production facilities
- University: Mensa, Library, Fast-Food Restaurant…

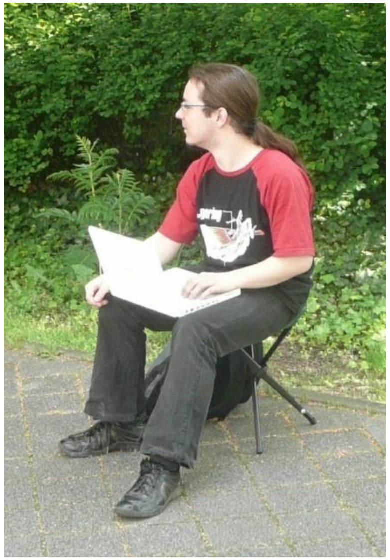
Simulation Project

Milestone 4 – Data Analysis

Gustav-Adolf-Str. Richtung Altstadt

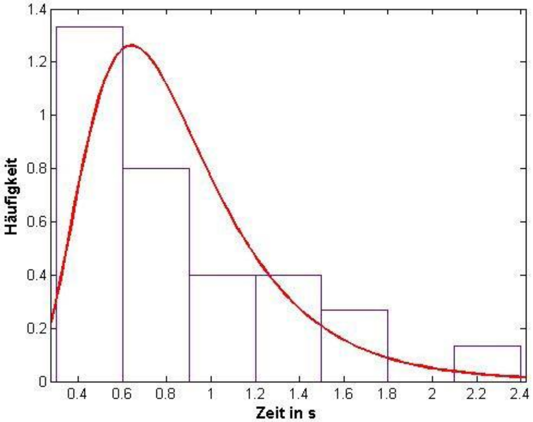
min	max	Observed	Expected	
0	10000	62	57,2	0,4027972
10000	20000	24	29,2	0,9260274
20000	30000	11	14,9	1,02080537
30000	40000	10	7,6	0,75789474
40000	60000	10	5,9	2,84915254
				5,95667725

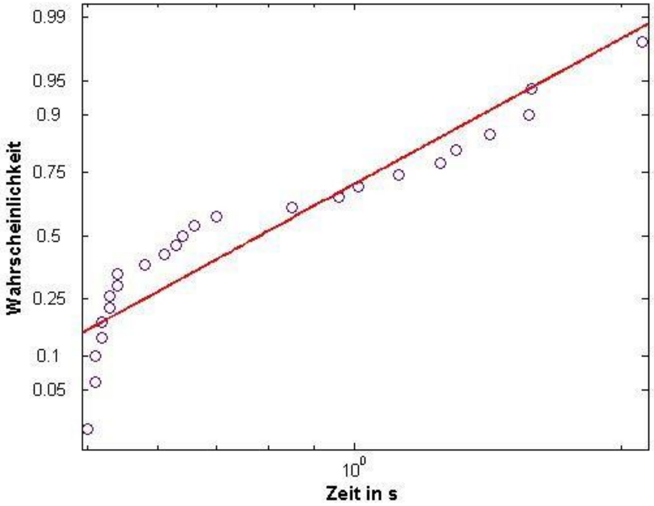




Exp(0,000672) <—

f	3
alpha	0,05
chisq	7,81472776
	ACCEPT





Simulation Project

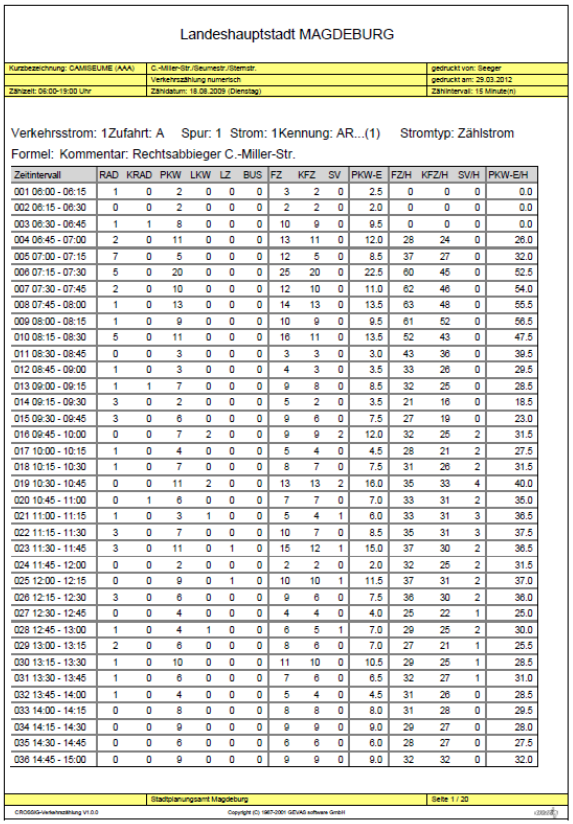
Real-world data is never easy to interpret!

Equipment	Tag	Dauer (h)	Häufigkeit	MIN(DATSTAT.TS_	MAX(DATSTAT.TS	TEXT
Q06-A-NB02	14.04.2009	0.06	11	14.04.2009 01:54	14.04.2009 21:25	M02 Entleerung nicht möglich-anderer Ablauf
Q06-A-NB02	14.04.2009	0.01	2	14.04.2009 07:07	14.04.2009 17:08	M07 Ansatz 01 Modul 07 aktiv
Q06-A-NB02	14.04.2009	67.54	10	14.04.2009 06:06	14.04.2009 19:07	M01 Überbrückung Leerfahren
Q06-A-NB02	14.04.2009	0.07	18	14.04.2009 15:56	14.04.2009 23:54	M03 Schiebescheiben offen und überbrückt M03
Q06-A-NB02	14.04.2009	0.2	3	14.04.2009 06:16	14.04.2009 15:57	M01 Störung Sensor Längenüberwachung Bahn 3
Q06-A-NB02	14.04.2009	1.2	21	14.04.2009 04:35	14.04.2009 11:31	M02 Fühlerbruch PT100
Q06-A-NB02	14.04.2009	0.01	3	14.04.2009 04		
Q06-A-NB02	14.04.2009	0.01	2	14.04.2009 14		
Q06-A-NB02	14.04.2009	0.06	8	14.04.2009 14		
Q06-A-NB02	14.04.2009	0	1	14.04.2009 15		
Q06-A-NB02	14.04.2009	1.2	21	14.04.2009 04		
Q06-A-NB02	14.04.2009	0.01	3	14.04.2009 07		
Q06-A-NB02	14.04.2009	0.11	7	14.04.2009 04		
Q06-A-NB02	14.04.2009	0.04	9	14.04.2009 15		
Q06-A-NB02	14.04.2009	0.11	7	14.04.2009 04		
Q06-A-NB02	14.04.2009	0.01	2	14.04.2009 14		
Q06-A-NB02	14.04.2009	41.77	92	14.04.2009 02		

Wafer	Start	VLG	MSW	BL	NB M01
----------------	----------------	----------------	----------------	----------------	----------------
0000214966A	13.04.2009 23:55	13.04.2009 23:55	13.04.2009 23:55	13.04.2009 23:55	13.04.2009 23:55
0000214967A	13.04.2009 23:55	13.04.2009 23:55	13.04.2009 23:55	13.04.2009 23:55	13.04.2009 23:55
0000214968A	13.04.2009 23:55	13.04.2009 23:55	13.04.2009 23:55	13.04.2009 23:55	13.04.2009 23:55
0000214969A	13.04.2009 23:55	13.04.2009 23:55	13.04.2009 23:55	13.04.2009 23:55	13.04.2009 23:55
0000214970A	13.04.2009 23:55	13.04.2009 23:55	13.04.2009 23:55	13.04.2009 23:55	13.04.2009 23:55
0000214971A	13.04.2009 23:55	13.04.2009 23:55	13.04.2009 23:55	13.04.2009 23:56	13.04.2009 23:55
0000214972A	13.04.2009 23:55	13.04.2009 23:55	13.04.2009 23:55	13.04.2009 23:56	13.04.2009 23:55
0000213049A	13.04.2009 23:55	13.04.2009 23:50	13.04.2009 23:50	13.04.2009 23:56	13.04.2009 23:55
0000214973A	13.04.2009 23:55	13.04.2009 23:55	13.04.2009 23:55	13.04.2009 23:56	13.04.2009 23:55
0000214974A	13.04.2009 23:55	13.04.2009 23:55	13.04.2009 23:55	13.04.2009 23:56	13.04.2009 23:55
0000214975A	13.04.2009 23:55	13.04.2009 23:55	13.04.2009 23:55	13.04.2009 23:56	13.04.2009 23:55
0000213053A	13.04.2009 23:55	13.04.2009 23:55		13.04.2009 23:56	13.04.2009 23:55
0000213054A	13.04.2009 23:55	13.04.2009 23:55		13.04.2009 23:56	13.04.2009 23:55
0000214976A	13.04.2009 23:55	13.04.2009 23:55	13.04.2009 23:55	13.04.2009 23:56	13.04.2009 23:55
0000214977A	13.04.2009 23:55	13.04.2009 23:55	13.04.2009 23:55	13.04.2009 23:56	13.04.2009 23:55
0000213056A	13.04.2009 23:55	13.04.2009 23:55		13.04.2009 23:56	13.04.2009 23:55
0000214980A	13.04.2009 23:55	13.04.2009 23:55	13.04.2009 23:55	13.04.2009 23:56	13.04.2009 23:55

Simulation Project

Real-world data is never easy to interpret!



Landeshauptstadt MAGDEBURG

Kurzbezeichnung: CAMISEUME (AAA)	C.-Miller-Str./Seumestr./Sternstr.	gedruckt von: Seeger
	Verkehrszählung numerisch	gedruckt am: 29.03.2012
Zählzeit: 06:00-19:00 Uhr	Zähldatum: 18.08.2009 (Dienstag)	Zählintervall: 15 Minute(n)

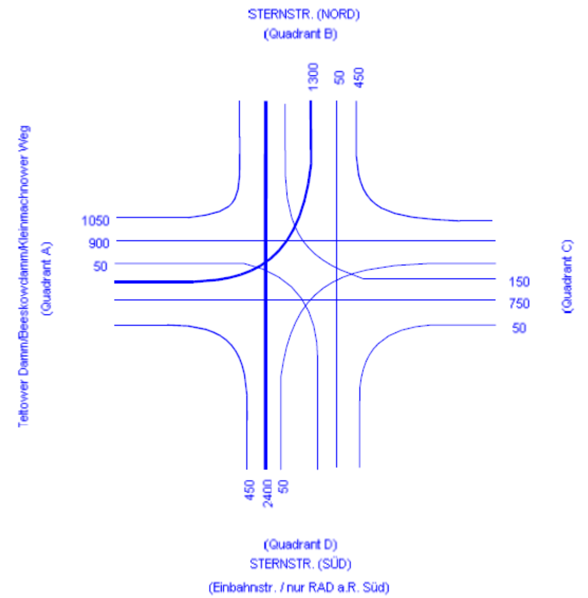
Verkehrsstrom: 1Zufahrt: A Spur: 1 Strom: 1Kennung: AR...(1) Stromtyp: Zählstrom

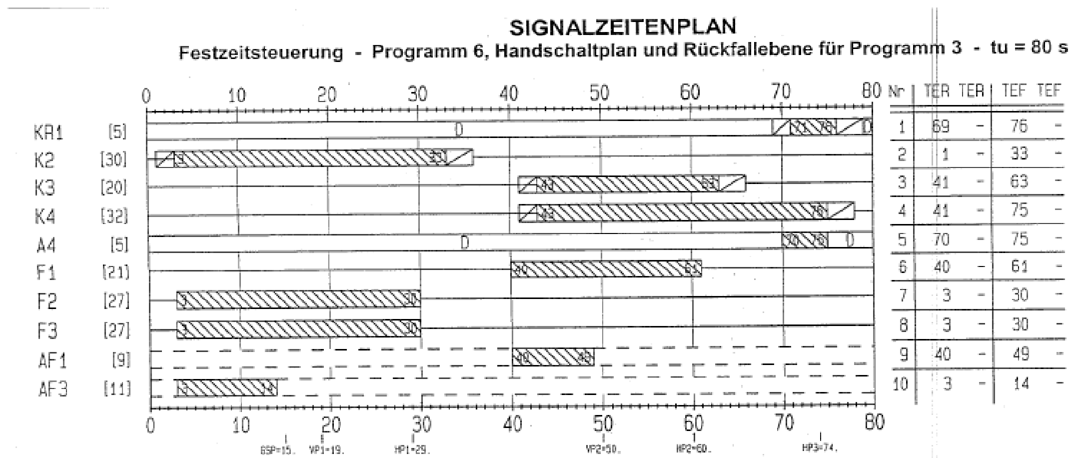
Formel: Kommentar: Rechtsabbieger C.-Miller-Str.

Zeitintervall	RAD	KRAD	PKW	LKW	LZ	BUS	FZ	KFZ	SV	PKW-E	FZ/H	KFZ/H	SV/H	PKW-E/H
001 06:00 - 06:15	1	0	2	0	0	0	3	2	0	2.5	0	0	0	0.0
002 06:15 - 06:30	0	0	2	0	0	0	2	2	0	2.0	0	0	0	0.0
003 06:30 - 06:45	1	1	8	0	0	0	10	9	0	9.5	0	0	0	0.0
004 06:45 - 07:00	2	0	11	0	0	0	13	11	0	12.0	28	24	0	26.0
005 07:00 - 07:15	7	0	5	0	0	0	12	5	0	8.5	37	27	0	32.0
006 07:15 - 07:30	5	0	20	0	0	0	25	20	0	22.5	60	45	0	52.5
007 07:30 - 07:45	2	0	10	0	0	0	12	10	0	11.0	62	46	0	54.0
008 07:45 - 08:00	1	0	13	0	0	0	14	13	0	13.5	63	48	0	55.5
009 08:00 - 08:15	1	0	9	0	0	0	10	9	0	9.5	61	52	0	56.5
010 08:15 - 08:30	5	0	11	0	0	0	16	11	0	13.5	52	43	0	47.5
011 08:30 - 08:45	0	0	3	0	0	0	3	3	0	3.0	43	36	0	39.5
012 08:45 - 09:00	1	0	3	0	0	0	4	3	0	3.5	33	26	0	29.5
013 09:00 - 09:15	1	1	7	0	0	0	9	8	0	8.5	32	25	0	28.5
014 09:15 - 09:30	3	0	2	0	0	0	5	2	0	3.5	21	16	0	18.5
015 09:30 - 09:45	3	0	6	0	0	0	9	6	0	7.5	27	19	0	23.0
016 09:45 - 10:00	0	0	7	2	0	0	9	9	2	12.0	32	25	2	31.5
017 10:00 - 10:15	1	0	4	0	0	0	5	4	0	4.5	28	21	2	27.5
018 10:15 - 10:30	1	0	7	0	0	0	8	7	0	7.5	31	26	2	31.5
019 10:30 - 10:45	0	0	11	2	0	0	13	13	2	16.0	35	33	4	40.0
020 10:45 - 11:00	0	1	6	0	0	0	7	7	0	7.0	33	31	2	35.0
021 11:00 - 11:15	1	0	3	1	0	0	5	4	1	6.0	33	31	3	36.5
022 11:15 - 11:30	3	0	7	0	0	0	10	7	0	8.5	35	31	3	37.5
023 11:30 - 11:45	3	0	11	0	1	0	15	12	1	15.0	37	30	2	36.5
024 11:45 - 12:00	0	0	2	0	0	0	2	2	0	2.0	32	25	2	31.5
025 12:00 - 12:15	0	0	9	0	1	0	10	10	1	11.5	37	31	2	37.0
026 12:15 - 12:30	3	0	6	0	0	0	9	6	0	7.5	36	30	2	36.0
027 12:30 - 12:45	0	0	4	0	0	0	4	4	0	4.0	25	22	1	25.0
028 12:45 - 13:00	1	0	4	1	0	0	6	5	1	7.0	29	25	2	30.0
029 13:00 - 13:15	2	0	6	0	0	0	8	6	0	7.0	27	21	1	25.5
030 13:15 - 13:30	1	0	10	0	0	0	11	10	0	10.5	29	25	1	28.5
031 13:30 - 13:45	1	0	6	0	0	0	7	6	0	6.5	32	27	1	31.0
032 13:45 - 14:00	1	0	4	0	0	0	5	4	0	4.5	31	26	0	28.5
033 14:00 - 14:15	0	0	8	0	0	0	8	8	0	8.0	31	28	0	29.5
034 14:15 - 14:30	0	0	9	0	0	0	9	9	0	9.0	29	27	0	28.0
035 14:30 - 14:45	0	0	6	0	0	0	6	6	0	6.0	28	27	0	27.5
036 14:45 - 15:00	0	0	9	0	0	0	9	9	0	9.0	32	32	0	32.0

	Stadtplanungsamt Magdeburg	Seite 1 / 20

CROSSIG-Verkehrszählung V1.0.0 — Copyright (C) 1987-2001 GEVAS software GmbH





Learning Goals

Learning questions:

- Construct a Q-Q Plot using a given dataset
- What conclusions can you draw from a Q-Q Plot?
- How does the chi-square test work?
- What is the meaning of the parameter Alpha in the chi-square test?
- What does the result $\chi_0^2 > \chi_{f,\alpha}^2$ in the chi-square test mean?
- What does the result $\chi_0^2 < \chi_{f,\alpha}^2$ in the chi-square test mean?