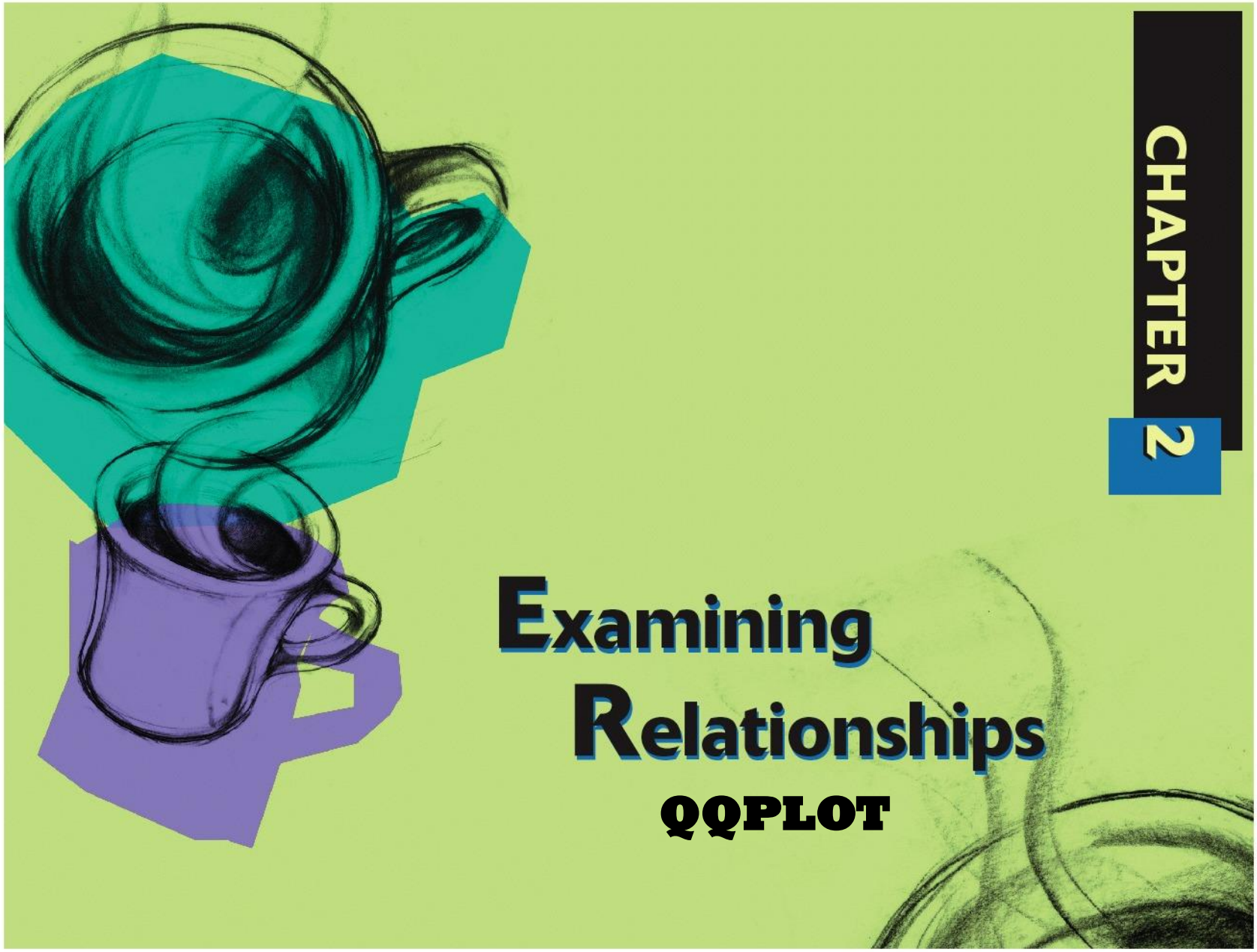
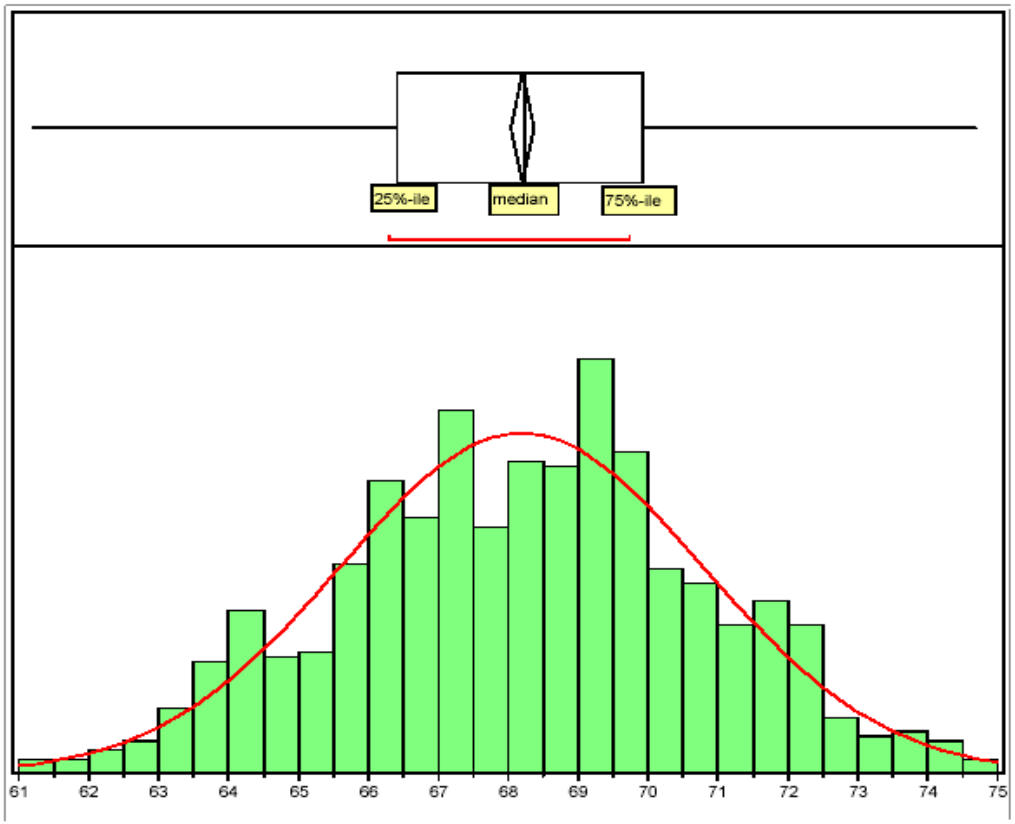

CHAPTER 2
Examining Relationships
QQPLOT

Regression

- Francis Galton *(1822 – 1911)* measured the heights of about 1,000 fathers and sons.
- The following plot summarizes the data on sons' heights.
- The curve on the histogram is a N(68.2, 2.62) density curve.



Data is often normally distributed:

The following table summarizes some aspects of the data:

Quantiles

100.0%	maximum	74.69
90.0%		71.74
75.0%	quartile	69.92
50.0%	median	68.24
25.0%	quartile	66.42
10.0%		64.56
0.0%	minimum	61.20

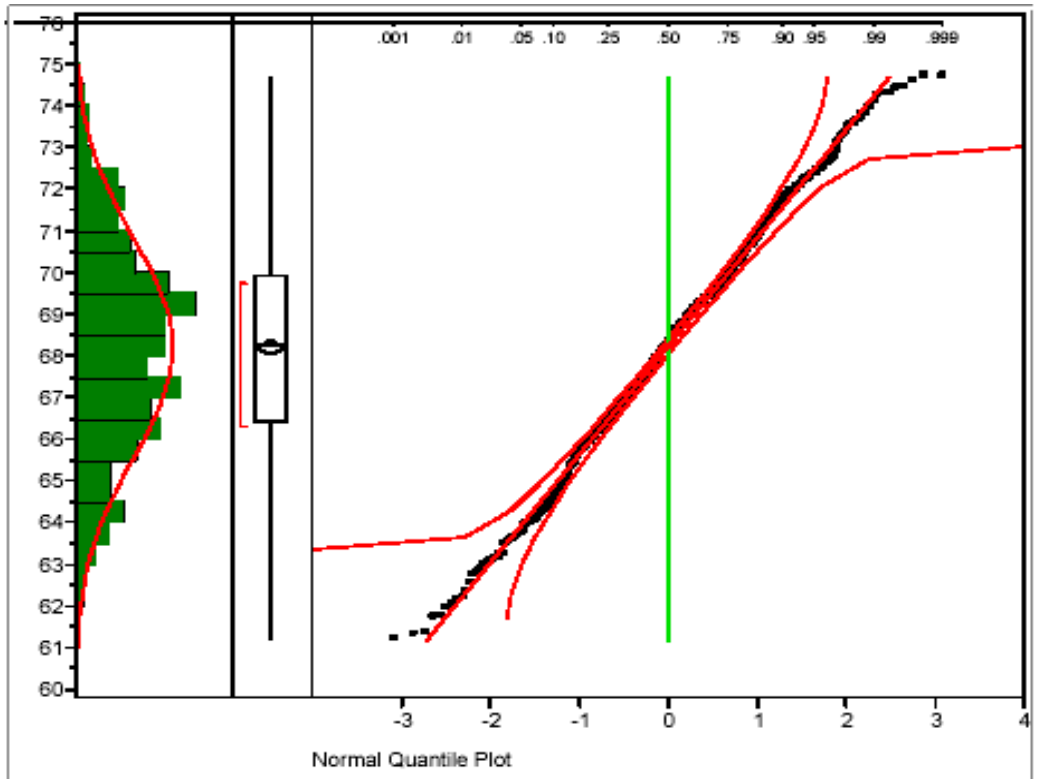
Moments

Mean 68.20	Std Dev 2.60	N= 952

Normal Quantile Plot

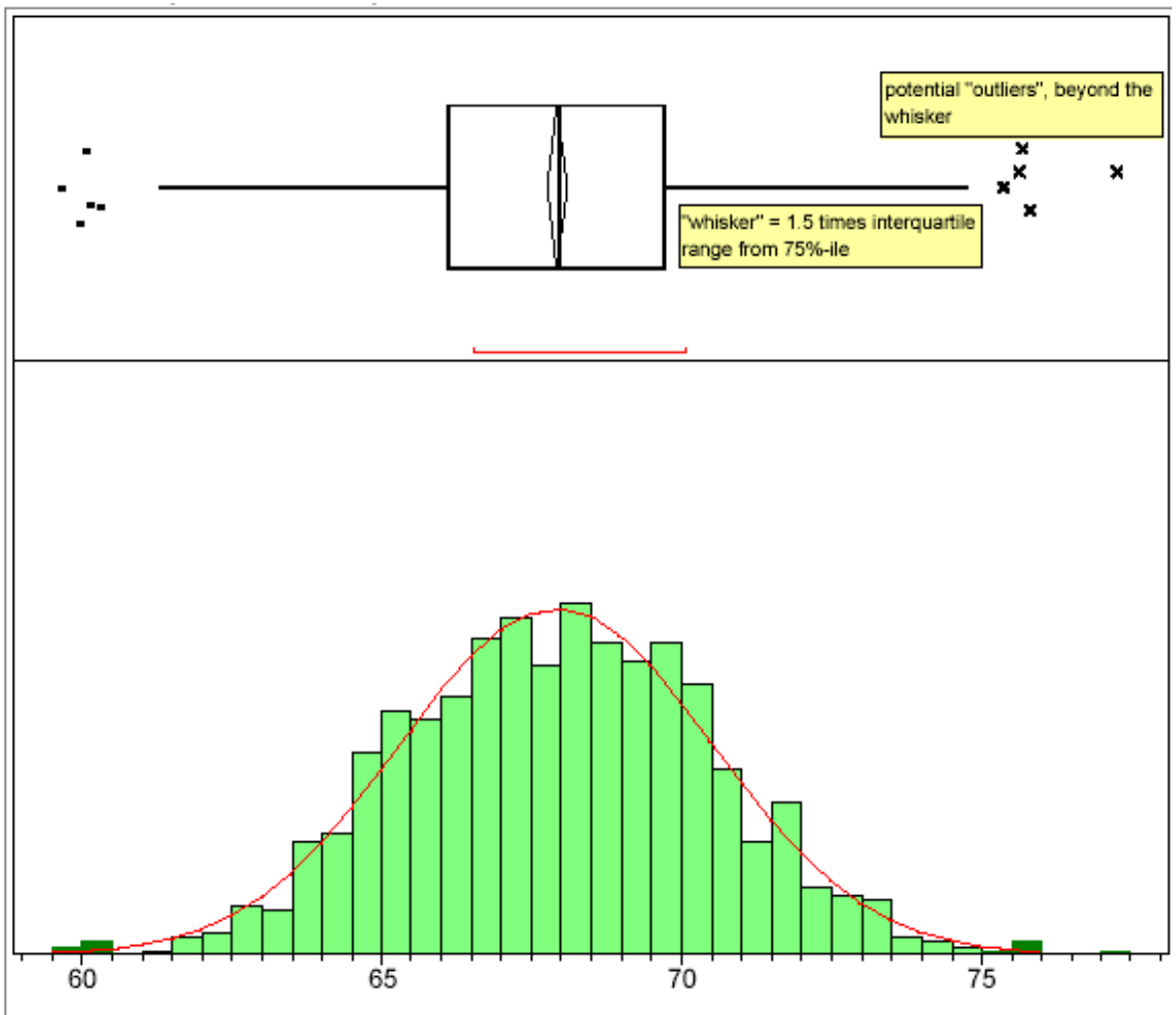
- A "normal quantile plot" provides a better way of determining whether data is well fitted by a normal distribution.
  - How these plots are formed and interpreted?

The plot for the Galton data on sons' heights:



Simulation

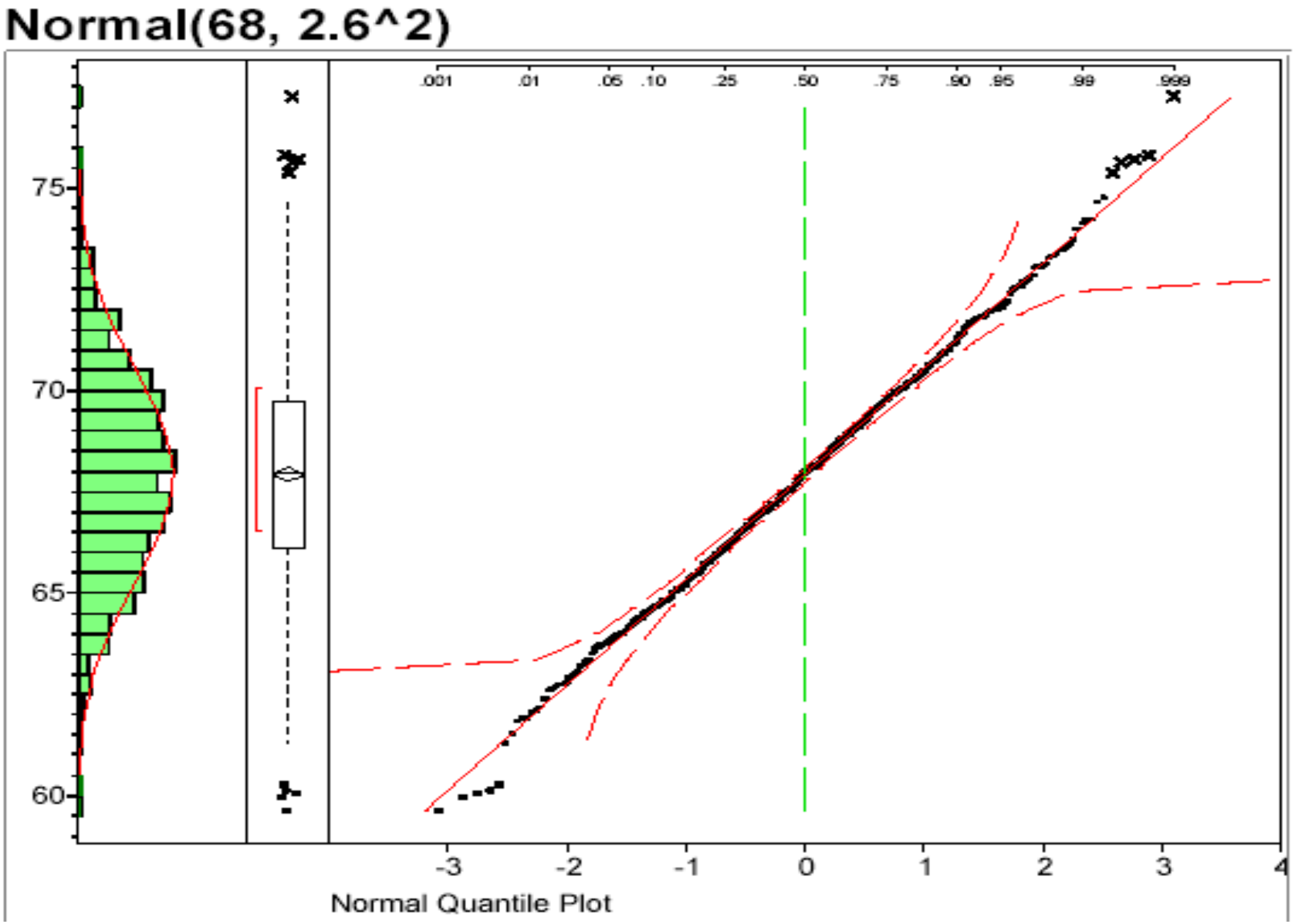
- Here is a histogram and probability plot for a sample of size 1000 from a perfectly normal population with mean = 68 and SD = 2.6.



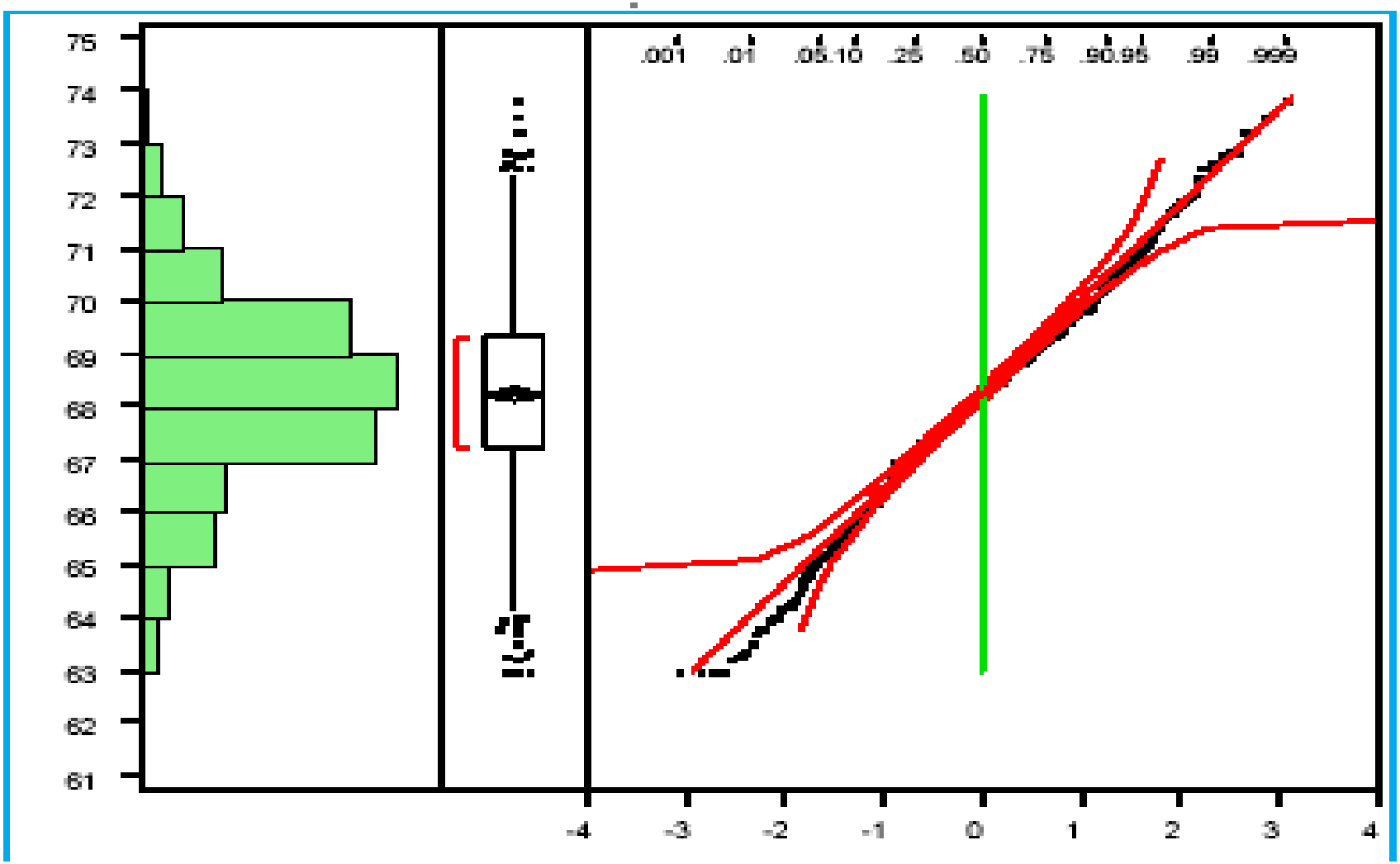
Moments

Mean	67.92
Std Dev	2.60
N	1000

Simulation

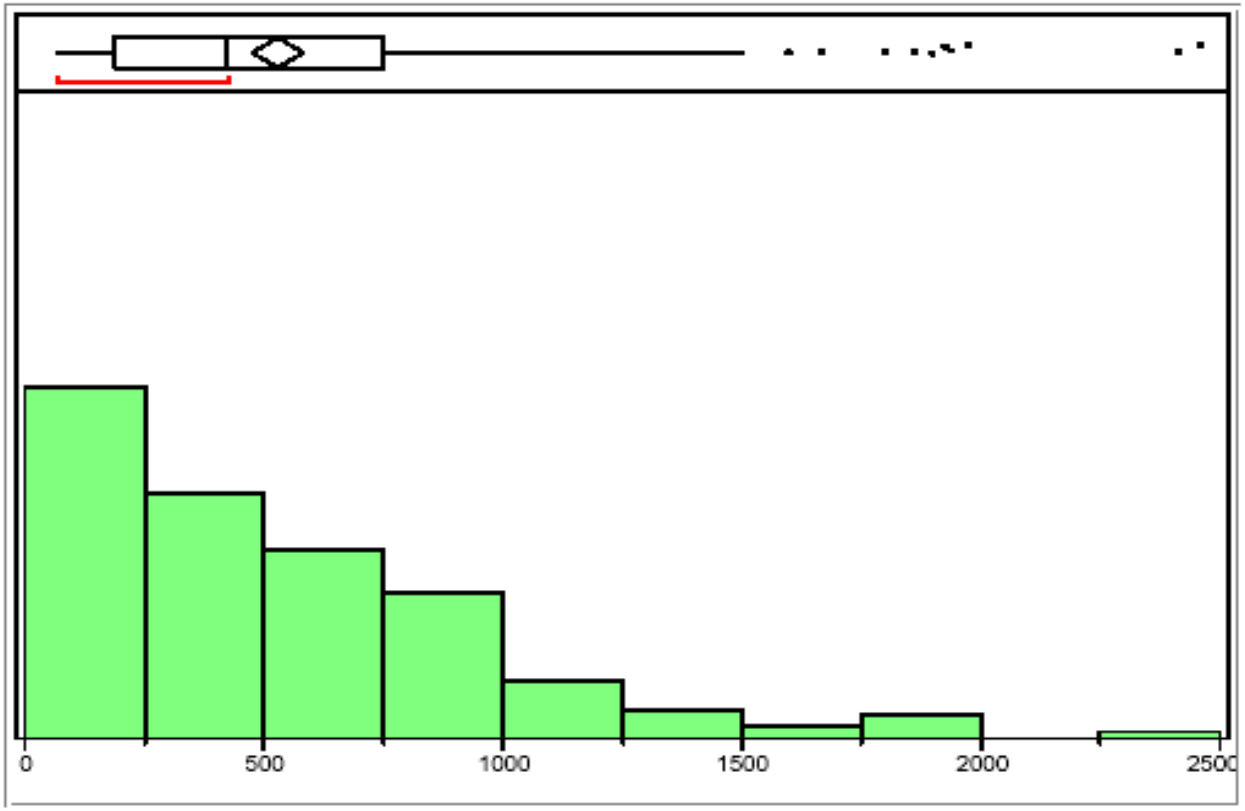

Normal(68, 2.6^2)
.001
.01
.05
.10
.25
.50
.75
.90
.95
.99
.999
75
70
65
60
-3
-2
-1
0
1
2
3
4
Normal Quantile Plot

Summary on parents' heights



Not all real data is approximately normal:

- Histogram and normal probability plot for the salaries (in $1,000) of all major league baseball players in 1987.
  - Only position players – not pitchers – who were on a major league roster for the entire season are included.



Moments

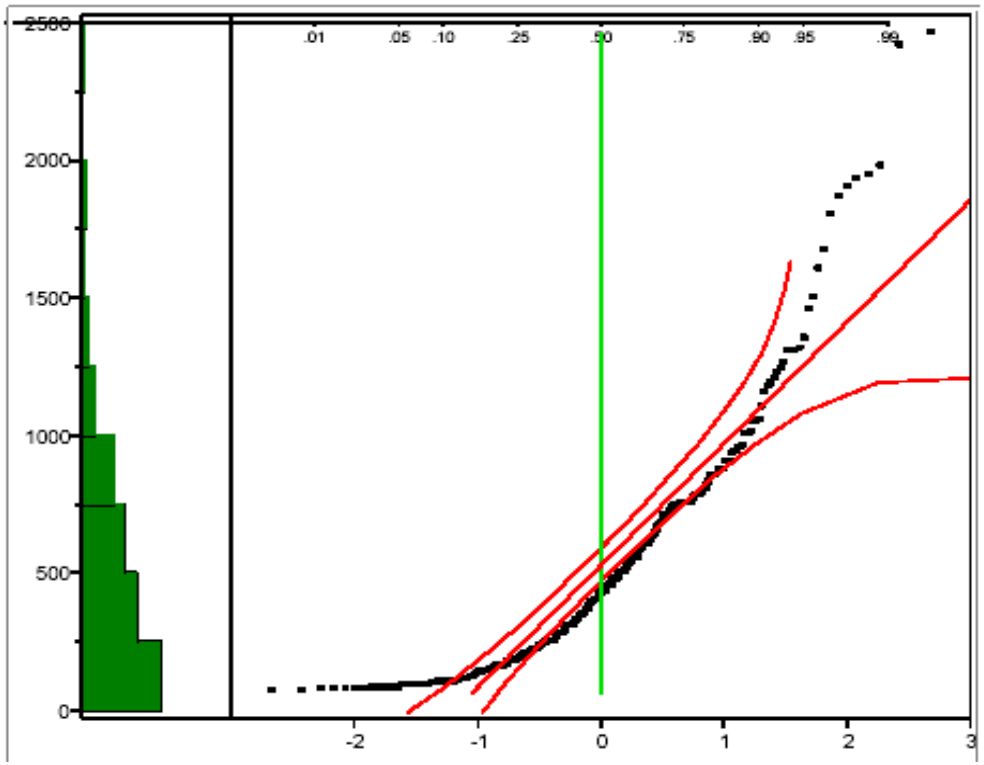
Mean 529.7

S.D. 441.6

N 260

Normal Quantile Plot

- This distribution is “skewed to the right”.
  - How this skewness is reflected in the normal quantile plot?
  - Both the largest salaries and the smallest salaries are much too large to match an ideal normal pattern. (They can be called “outliers”.)
  - This histogram seems something like an exponential density. Further investigation confirms a reasonable agreement with an exponential density truncated below at 67.5.



Judging whether a distribution is approximately normal or not

- Personal incomes, survival times, etc are usually skewed and not normal.

- Risky to assume that a distribution is normal without actually inspecting the data.

- Stemplots and histograms are useful.
- Still more useful tool is the normal quantile plot.

Normal quantile plots

- Arrange the data in increasing order. Record percentiles of each data value.
- Do normal distribution calculations to find the z-scores at these same percentiles.
- Plot each data point x against the corresponding z.
- If the data distribution is standard normal, the points will lie close to the 45-degree line x=z.
- If it is close to any normal distribution, the points will lie close to some straight line.

Example

Q Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution.

A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

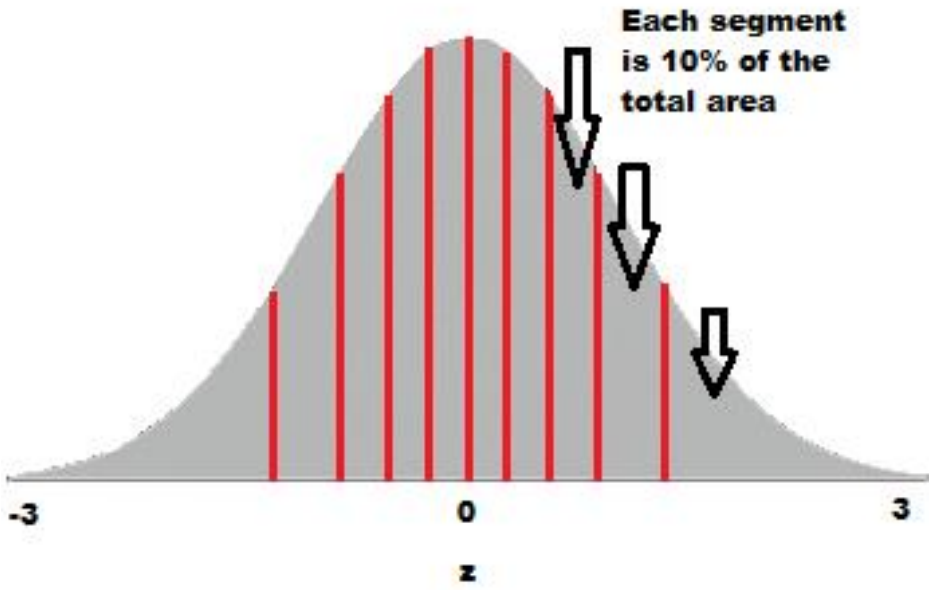
How to Make a Q Q Plot

Sample question:
Do the following values come from a normal distribution?
7.19, 6.31, 5.89, 4.5, 3.77, 4.25, 5.19, 5.79, 6.79.

Step 1: Order the items from smallest to largest.

- 3.77
- 4.25
- 4.50
- 5.19
- 5.89
- 5.79
- 6.31
- 6.79
- 7.19

Step 2: Draw a normal distribution curve.
Divide the curve into n+1 segments.
We have 9 values, so divide the curve into 10 equally-sized areas.
For this example, each segment is 10% of the area
(because 100% / 10 = 10%).



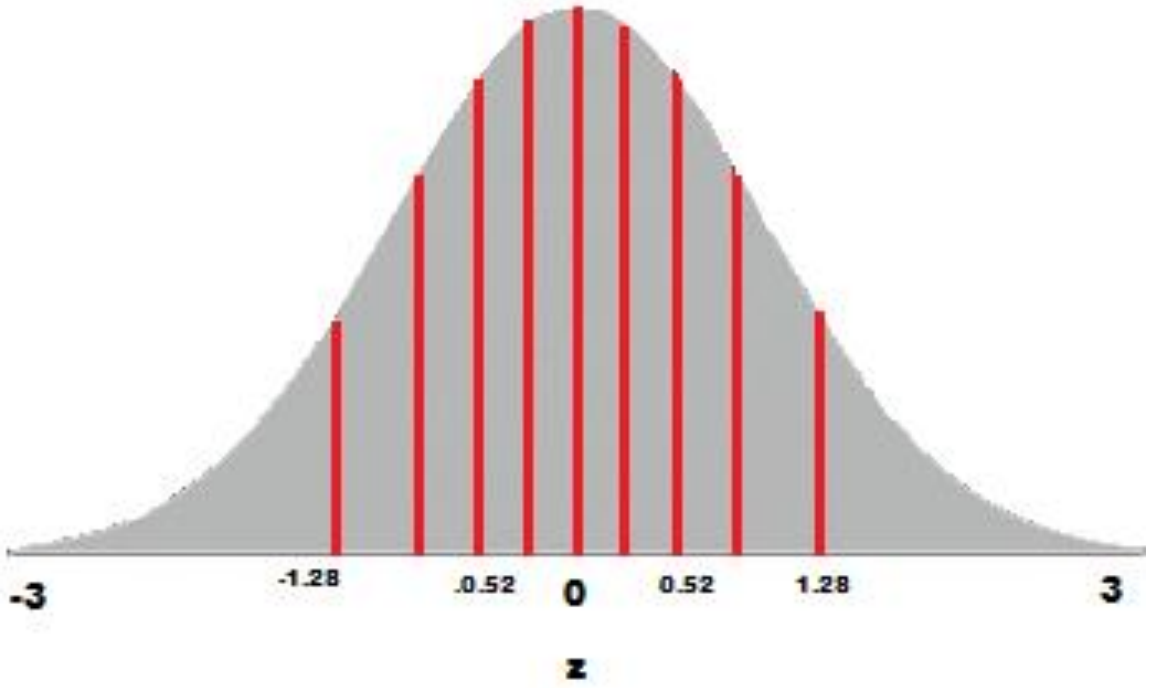
Step 3: Find the z-value (cut-off point) for each segment in Step 3. These segments are areas, so refer to a z-table (or use software) to get a z-value for each segment.
The z-values are:

- 10% = -1.28
- 20% = -0.84
- 30% = -0.52
- 40% = -0.25
- 50% = 0
- 60% = 0.25
- 70% = 0.52
- 80% = 0.84
- 90% = 1.28
- 100% = 3.0

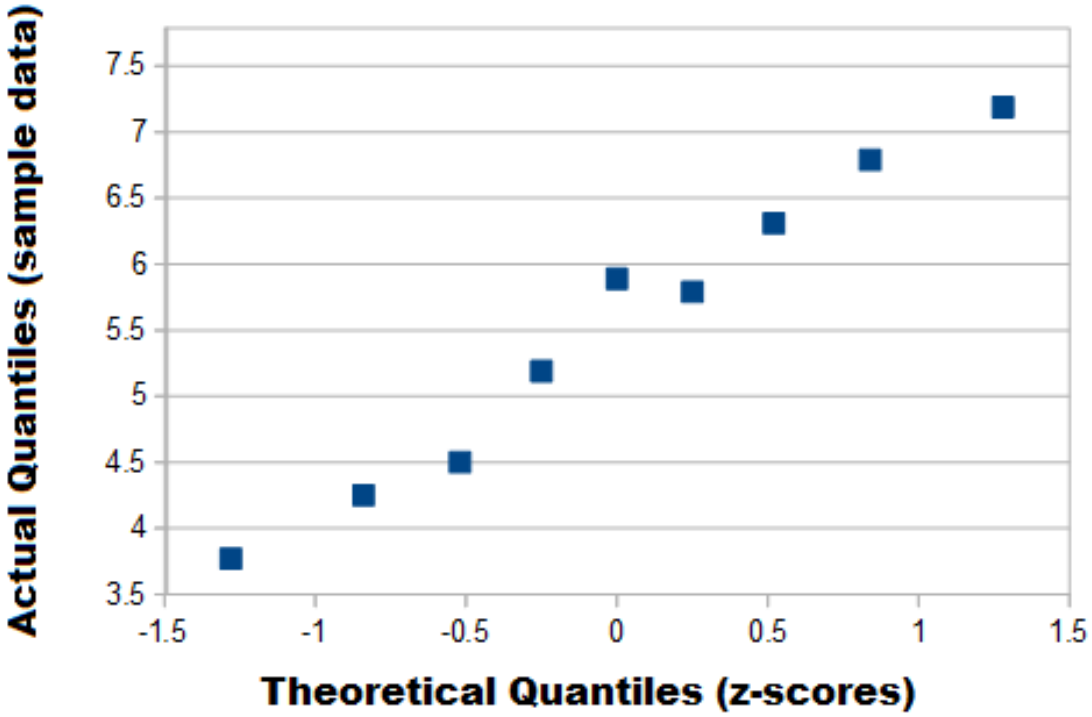
- Using the following table

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767

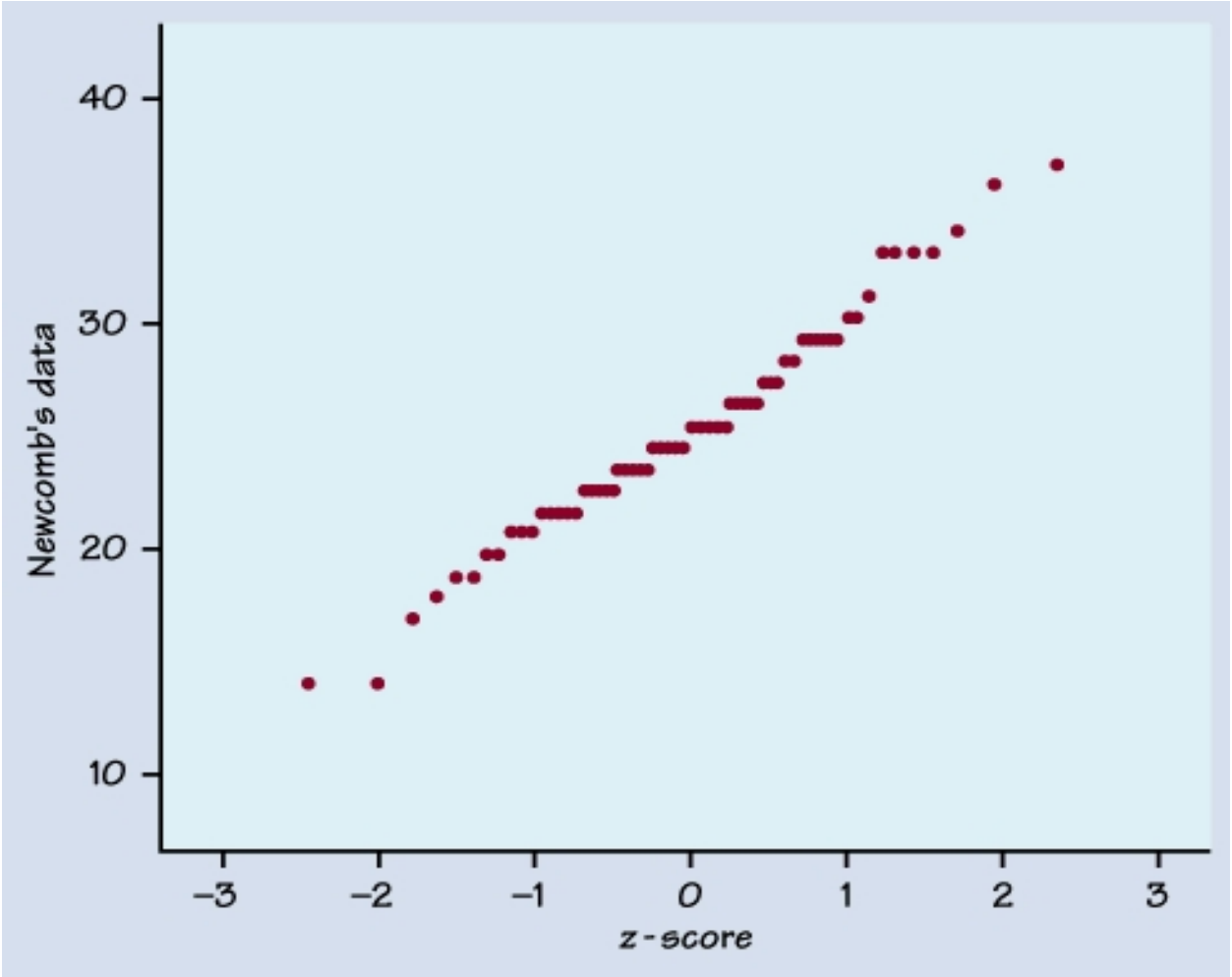
Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998
3.5	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998
3.6	0.4998	0.4998	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.7	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.8	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999


-3
-1.28
.0.52
0
0.52
1.28
3
z

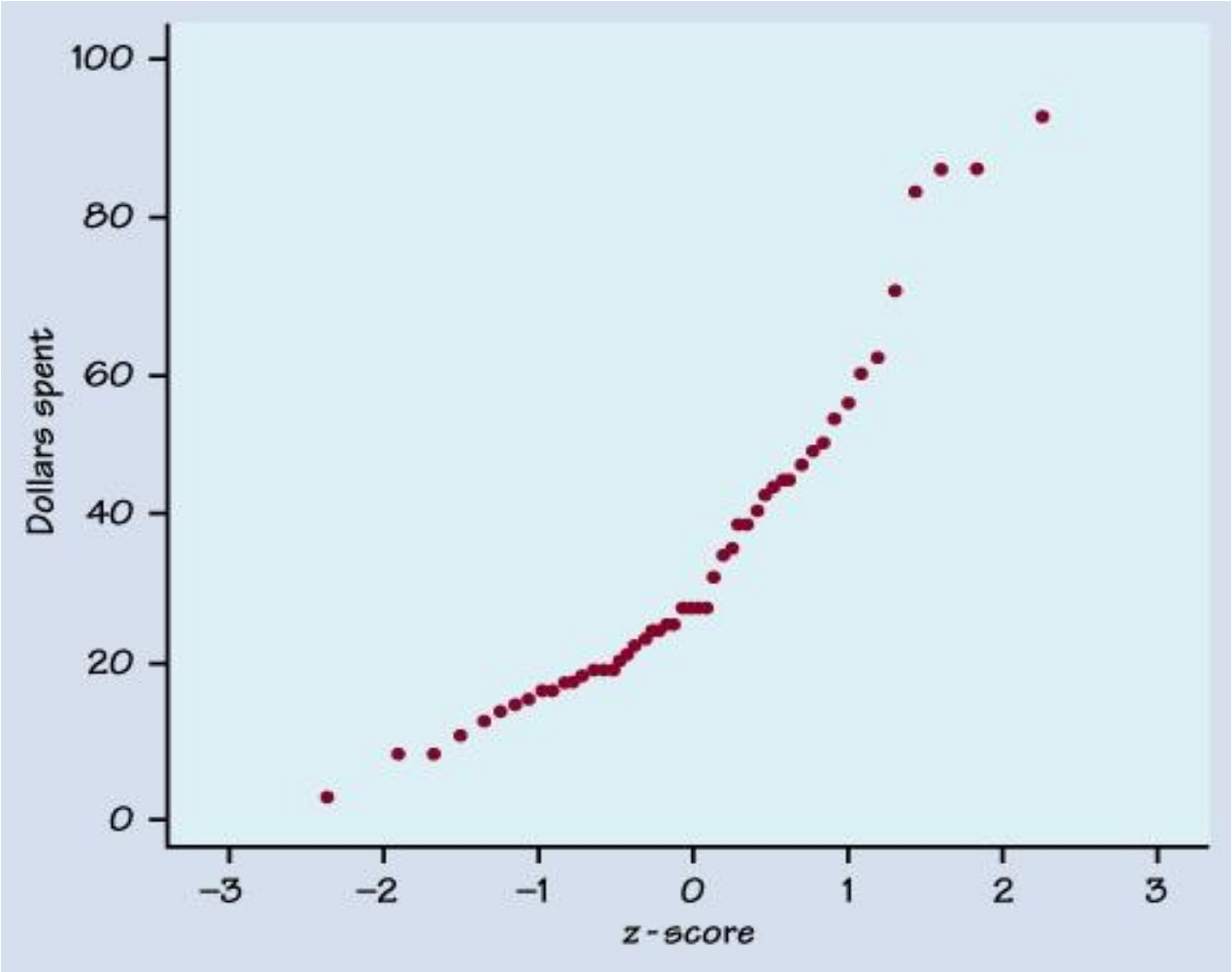
Step 4: Plot your data set values (Step 1) against your normal distribution cut-off points (Step 3). I used Open Office for this chart:



Examples of some QQplots



- *granularity*

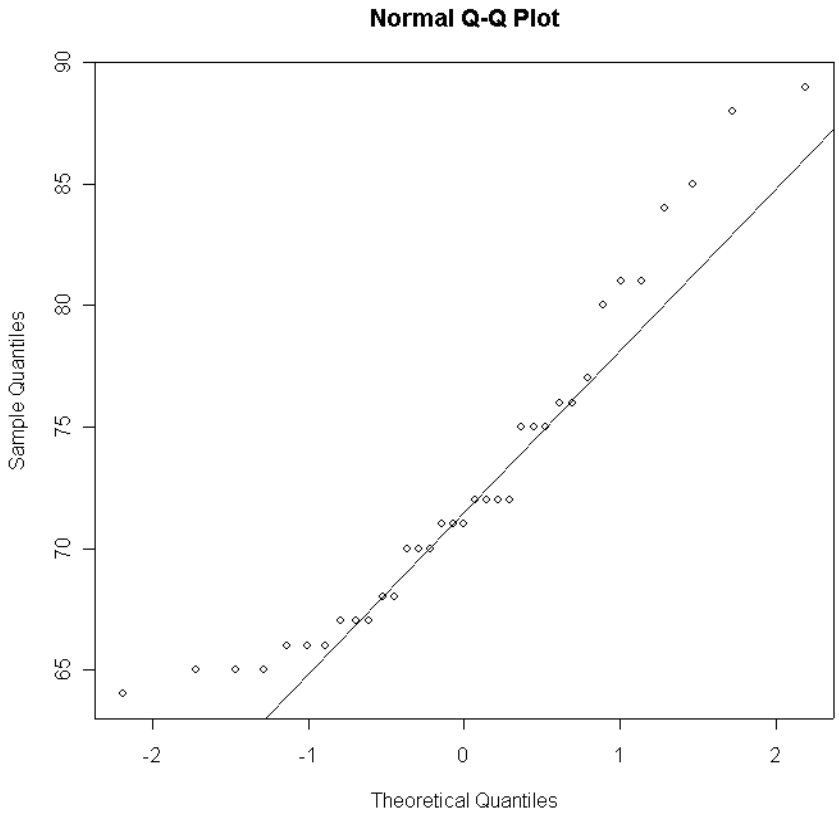


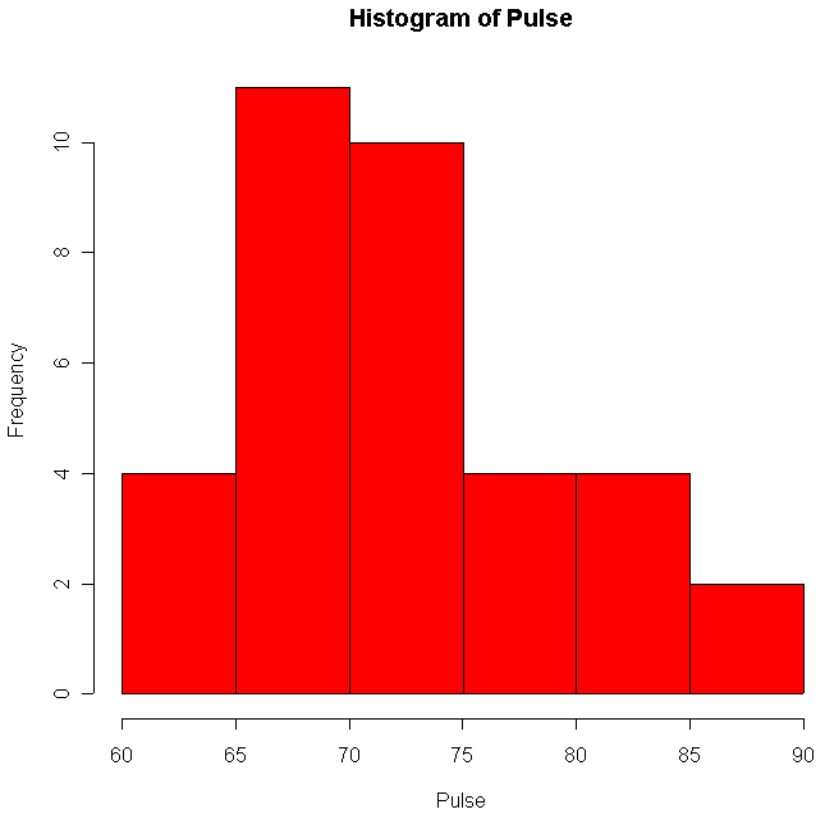
- Right-skewed distribution

qqline (R-function)

- Plots a line through the first and third quartile of the data, and the corresponding quantiles of the standard normal distribution.
- Provide a good ‘straight line’ that helps us see whether the points lie close to a straight line.

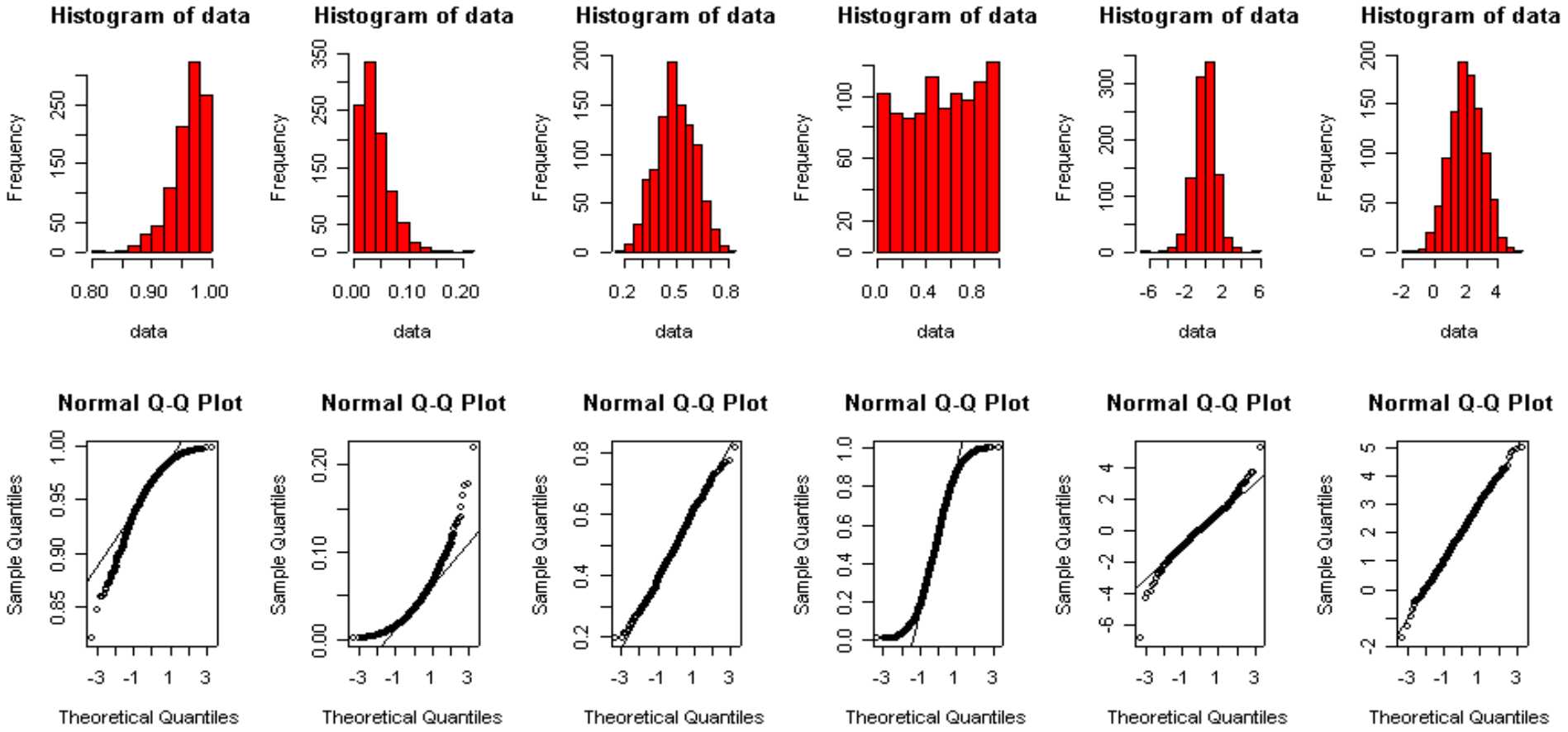
Quantiles
−2
−1
0
1
2


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
65
70
75
80
85
90
-2
-1
0
1
2


Histogram of Pulse
Frequency
Pulse
0
2
4
6
8
10
60
65
70
75
80
85
90

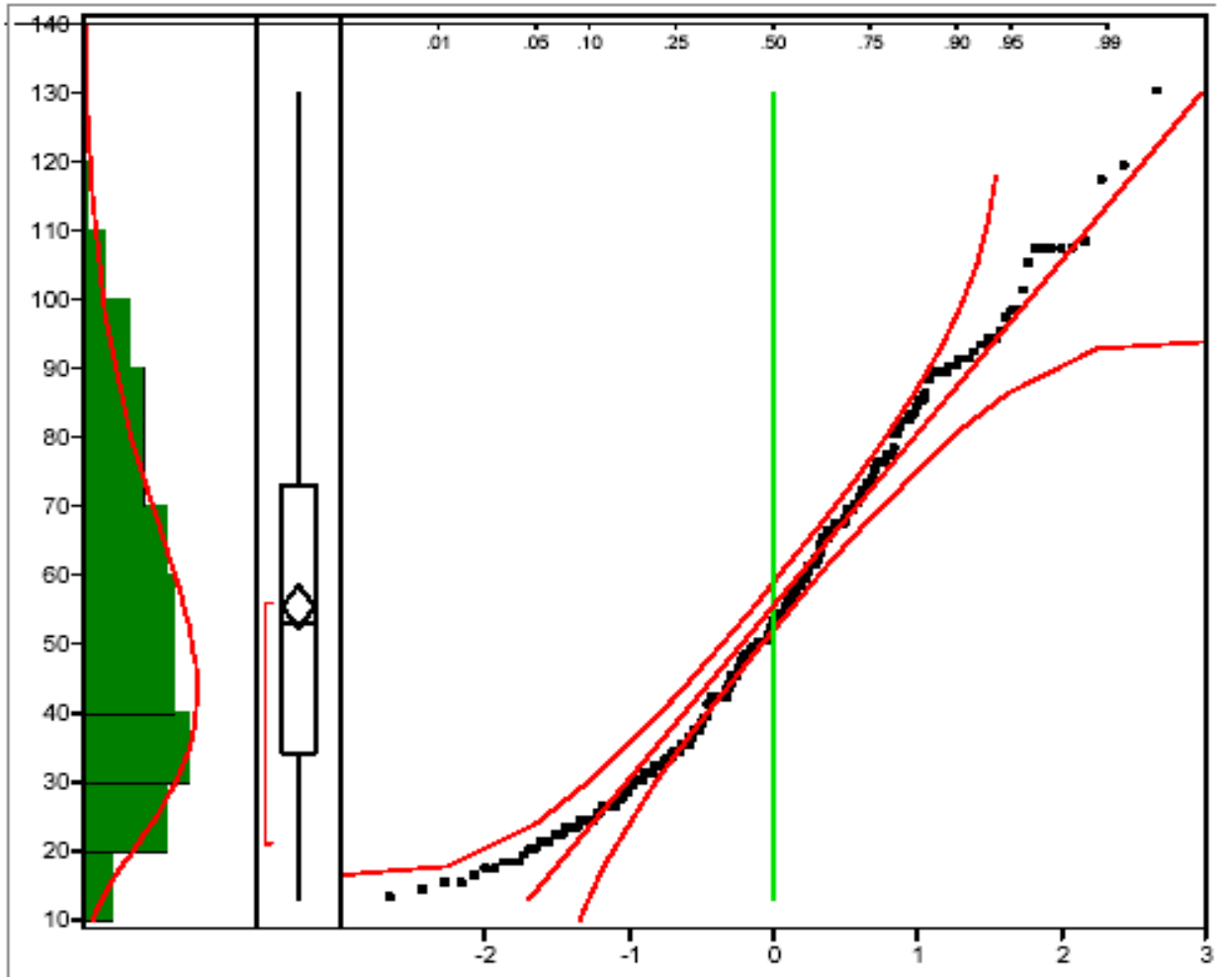
- Pulse data

‘Simulations’



Another non-normal pattern

- The data here is the number of runs scored in the 1986 season by each of the players in the above data set.



Moments

Mean 55.33

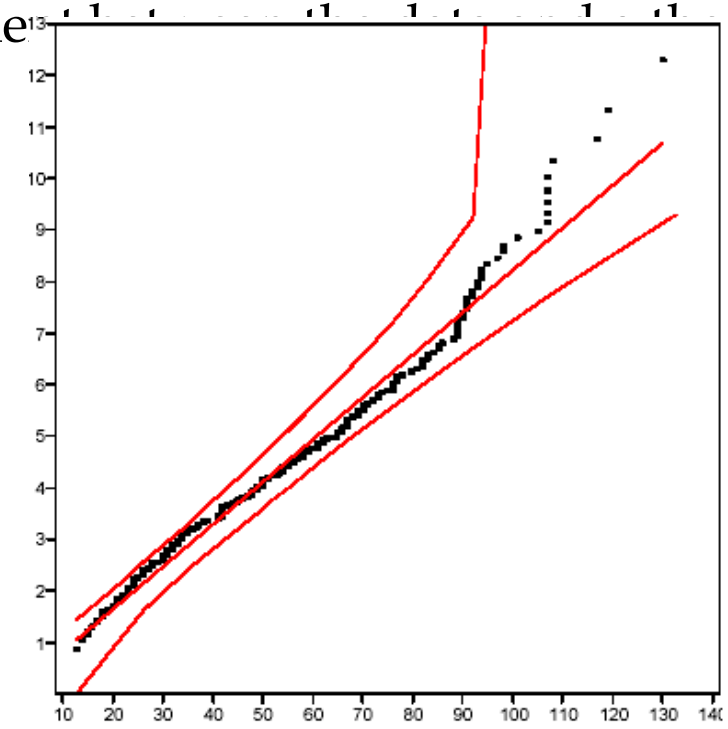
S.D. 25.02

N 261

Note: n = 261 here, but in the preceding data n = 260. The discrepancy results from the fact that one player in the data set has a missing salary figure.

Gamma Quantile Plot

- This data is fairly well fit by a gamma density with parameters $\alpha = 4.55$ and $\lambda = 12.16$. (How do we find those two numbers?)
  - What is the gamma density curve?
  - How do we plot a quantile plot to check on gamma density?
  - The data points form a fairly straight line on this plot; hence there is reasonable agreeme[illegible] retical $\Gamma(4.55,12.16)$ distribution.



QQplot

- x <- qgamma(seq(.001, .999, len = 100), 1.5) # compute a vector of quantiles

- plot(x, dgamma(x, 1.5), type = "l") # density plot for shape 1.5

- QQplots are used to assess
  - whether data have a particular distribution, or
  - whether two datasets have the same distribution.

- If the distributions are the same, then the QQplot will be approximately a straight line.
  - The extreme points have more variability than points toward the center.
  - A plot with a "U" shape means that one distribution is skewed relative to the other.
  - An "S" shape implies that one distribution has longer tails than the other.
  - In the default configuration a plot from qqnorm that is bent down on the left and bent up on the right means that the data have longer tails than the Gaussian.
  - plot(qlnorm(ppoints(y)), sort(y)) # log normal qqplot