2.2. Distribution of data

1. QQ plot:

Example:

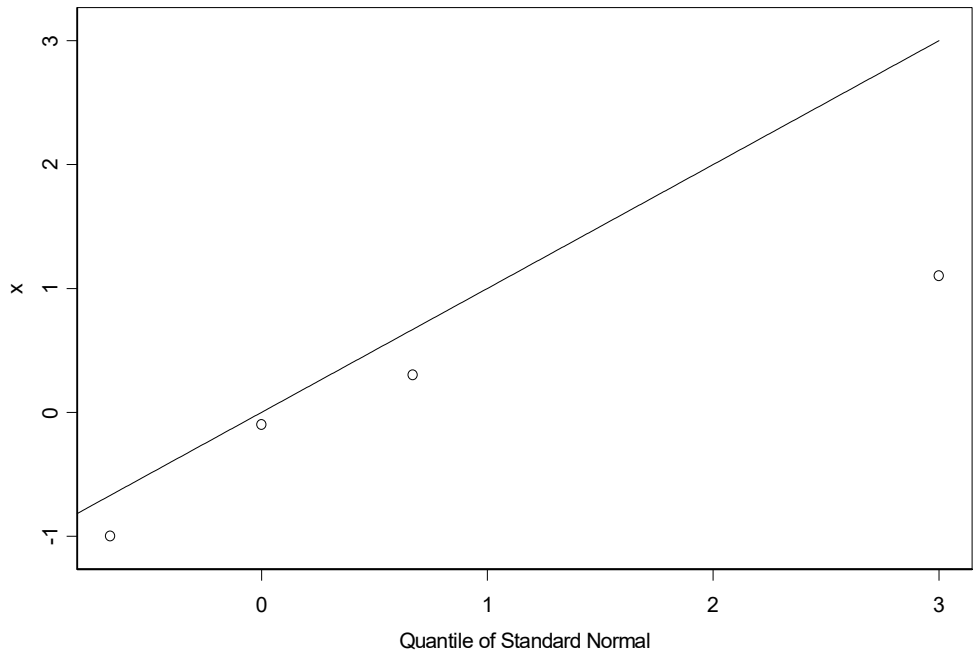
$\mathbf{n = 4}$

x	-1.0	-0.1	0.3	1.1
i	1	2	3	4
Empirical CDF (i/n)	0.25	0.5	0.75	1

Therefore,

Percent	25	50	75	100
Percentile of x	-1.0	-0.1	0.3	1.1
Percentile of $N(0,1)$	-0.67	0	0.67	3

Intuitively, if x is nearly standard normal, the percentile of x should be very close to the corresponding percentile of the standard normal random variable. Then, if we plot these percentiles of the standard normal variable in the above table versus the corresponding percentile of x, the points $(-0.67, -1)$, $(0, -0.1)$, $(0.67, 0.3)$, and $(3, 1.1)$ should fall around the line y=x.



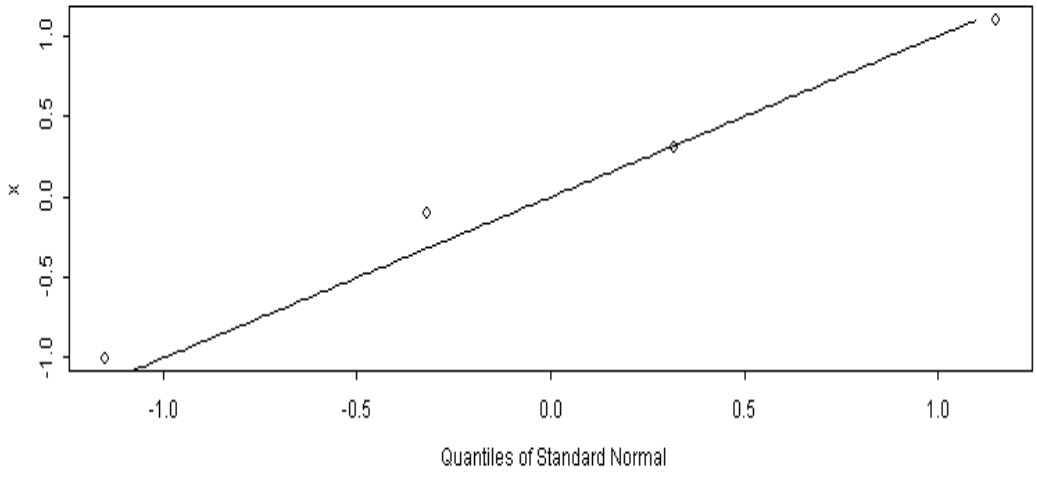
In S-plus, the above process has been modified. The empirical CDF formula used in S-plus is $\frac{i-a}{n+1-2a}$. As $a = 0.5$, the empirical CDF is $\frac{i-0.5}{n}$. Therefore, in S-plus,

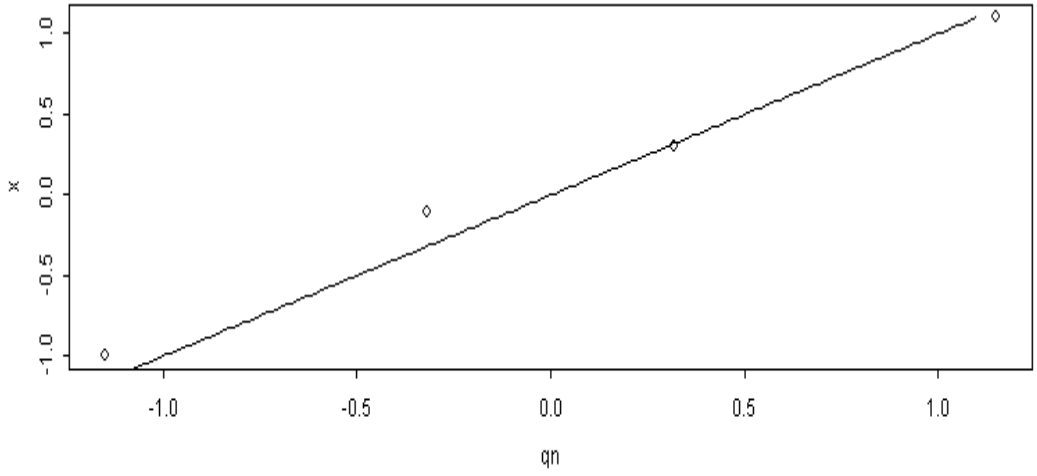
x	-1.0	-0.1	0.3	1.1
i	1	2	3	4
Empirical CDF	0.125	0.375	0.625	0.875

Thus,

Percent	12.5	37.5	62.5	87.5
Percentile of x	-1.0	-0.1	0.3	1.1
Percentile of $N(0,1)$	-1.15	-0.32	0.32	1.15

The plot of the percentile of $N(0,1)$ versus the percentile of *x* is





In S-plus, the qqnorm(x) will generate the same plot as the one of the percentile of $N(0,1)$ versus the percentile of x as the plot. The following S-plus commands can compare the plot generated by qqnorm(x) directly and the plot using the above process.

Example (Splus):

```
par(mfrow=c(2,1))
x=c(-1.0,-0.1,0.3,1.1)
qqnorm(x)                                  # the qq-normal plot
lines(seq(-1.1,1.1,by=0.01),seq(-1.1,1.1,by=0.01))    # the line y=x

px=ppoints(x)                              # px: the empirical CDF for x
qn=qnorm(px)                               # the percentile for the standard normal
plot(qn,x)                                 # distribution. The plot is the same as the one
lines(seq(-1.1,1.1,by=0.01),seq(-1.1,1.1,by=0.01))     # generated by qqnorm(x)
```

2. Chi-square test:

Let

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i}$$

The chi-square test with level of significance α for testing

$$H_0: p_1 = a_1, p_2 = a_2, \cdots, p_k = a_k$$

is to

reject H_0 as $\chi^2 > \chi^2_{k-1,\alpha}$.

Example :

The following are the number of wrong answers for the number of the students.

Number of wrong answers	0	1	2	3
Number of the students	21	31	12	0

Suppose X is the random variable representing the number of wrong answers. Please test X is distributed as $Binomial(3, 0.25)$ with $\alpha = 0.05$.

[solutions:]

As H_0 is true, the distribution for the number of wrong answers is

$$p_1 = P(X = 0) = \binom{3}{0} 0.25^0 0.75^3 = \frac{27}{64}$$

$$p_2 = P(X = 1) = \binom{3}{1} 0.25^1 0.75^2 = \frac{27}{64}$$

$$p_3 = P(X = 2) = \binom{3}{2} 0.25^2 0.75^1 = \frac{9}{64}$$

$$p_3 = P(X = 3) = \binom{3}{3} 0.25^3 0.75^0 = \frac{1}{64}.$$

Since the sample size $n = 21 + 31 + 12 + 0 = 64$, the expected numbers under H_0 are

$$e_1 = np_1 = 64 \cdot \frac{27}{64} = 27, e_2 = np_2 = 64 \cdot \frac{27}{64} = 27,$$

$$e_3 = np_3 = 64 \cdot \frac{9}{64} = 9, e_4 = np_4 = 64 \cdot \frac{1}{64} = 1.$$

Therefore,

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i}$$

$$= \frac{(21-27)^2}{27} + \frac{(31-27)^2}{27} + \frac{(12-9)^2}{27} + \frac{(0-1)^2}{27}$$

$$= 3.92$$

Since $\chi^2 = 3.92 < 7.81 = \chi^2_{3,0.05}$, we do *not* reject H_0.

Example (Splus):

```
x=c(rep(0,21),rep(1,31),rep(2,12),rep(3,0))
breaks=-1:3
chi=chisq.gof(x,cut.points=breaks,dist="binomial",size=3,prob=0.25)
chi$count
chi$expected
### Other distribution
x=rcauchy(50)
chisq.gof(x) # hypothesize a normal distribution
chisq.gof(x,dist="cauchy")    # hypothesize a Cauchy distribution
```

Note:

As expected counts ＜ 5, the code is as follows:

Example (Splus):

```
x=c(rep(0,21),rep(1,31),rep(2,12))
breaks=c(-1:1,3)
chi=chisq.gof(x,cut.points=breaks,dist="binomial",size=3,prob=0.25)
chi
chi$count
chi$expected
```