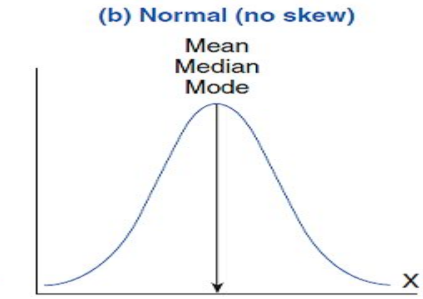
*Exercise 6*

## Homework exercise

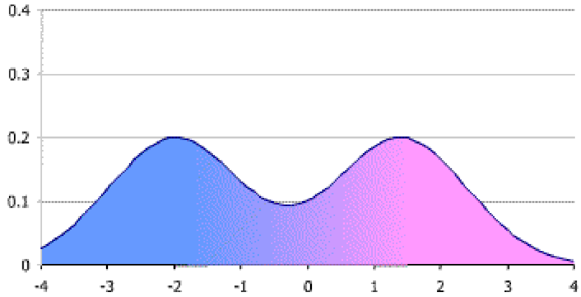
*To be solved at home before the exercise session.*

1. a. Assume that we have an iid. random sample $x_1, \ldots, x_{1000}$ and we'd like to use the normal Q-Q plot to assess whether the sample came from a normal distibution. How do you expect the normal Q-Q plot to roughly look like (i.e. what general features do you expect it to have and *why*), if the true distribution of the data is
   i. a normal distribution,
   ii. a right-skew distribution,
   iii. a left-skew distribution,
   iv. a bimodal distribution,
   v. a distribution with light tails,
   vi. a distribution with heavy tails?

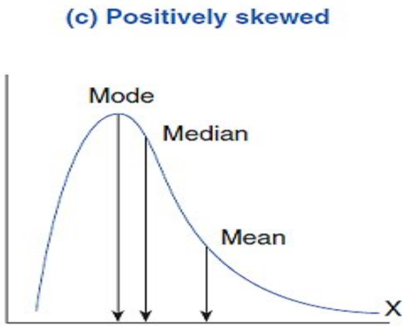
i. a normal distribution



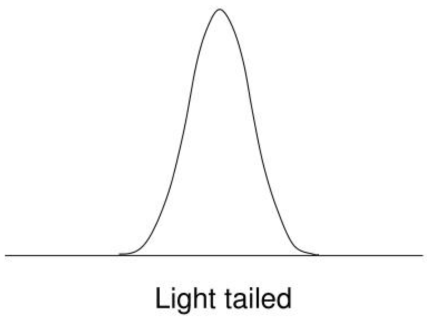
iv. a bimodal distribution



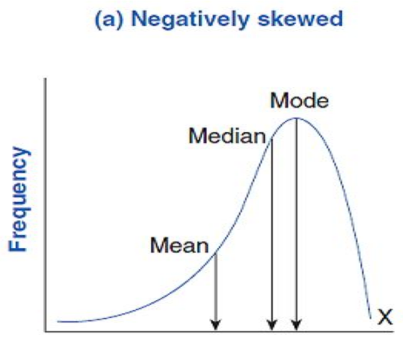
ii. a right-skew distribution



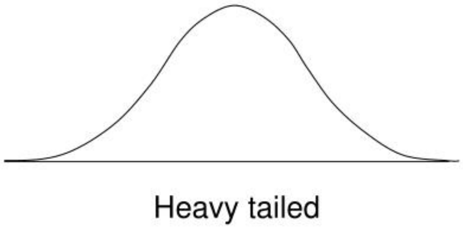
v. a distribution with light tails



iii. a left-skew distribution

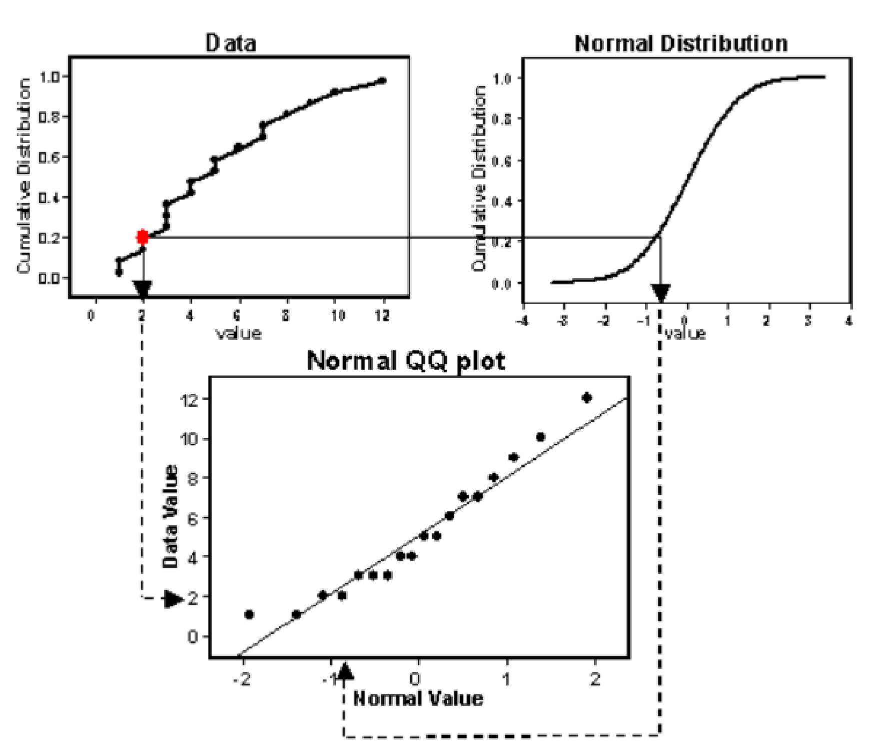


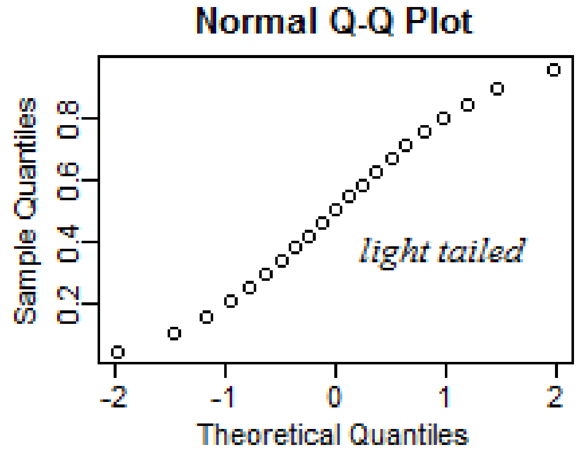
vi. a distribution with heavy tails

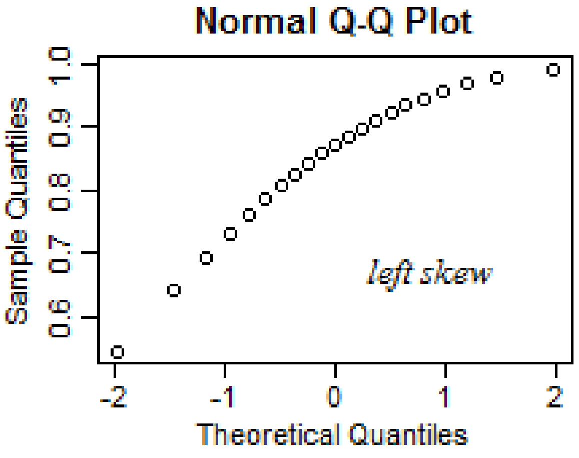


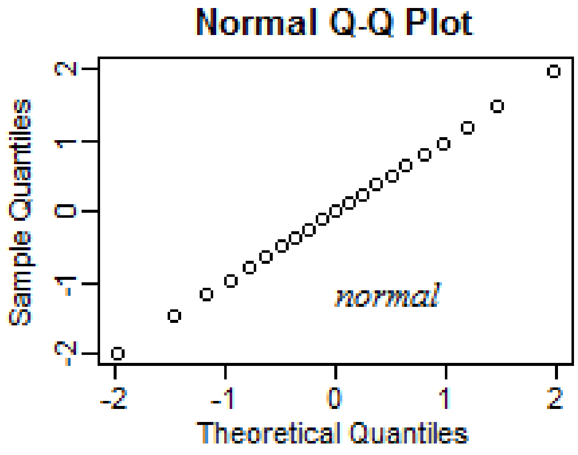
# How the Normal QQ plot is constructed

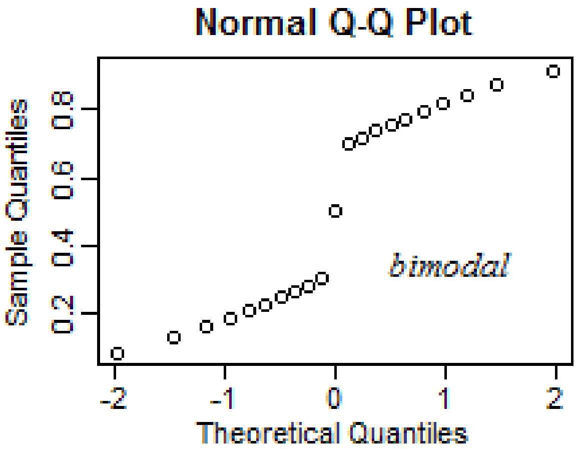
First, the data values are ordered and cumulative distribution values are calculated as (*i*– 0.5)/*n* for the *i*th ordered value out of *n* total values (this gives the proportion of the data that falls below a certain value). A cumulative distribution graph is produced by plotting the ordered data versus the cumulative distribution values (graph on the top left in the figure below). The same process is done for a standard normal distribution (a Gaussian distribution with a mean of 0 and a standard deviation of 1, shown in the graph on the top right of the figure below). Once these two cumulative distribution graphs have been generated, data values corresponding to specific quantiles are paired and plotted in a QQ plot (bottom graph in the figure below).

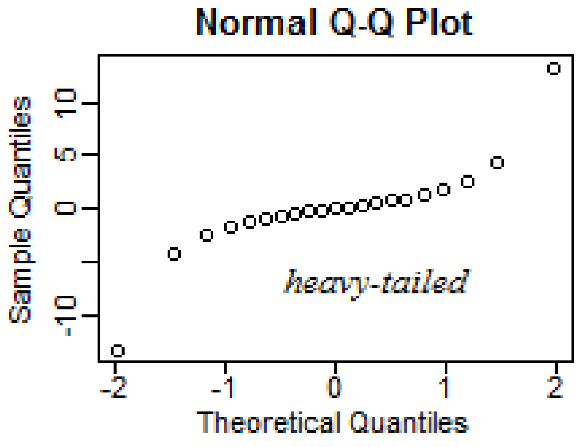


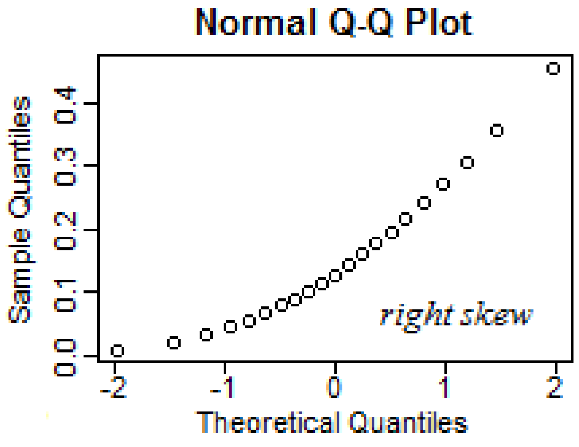












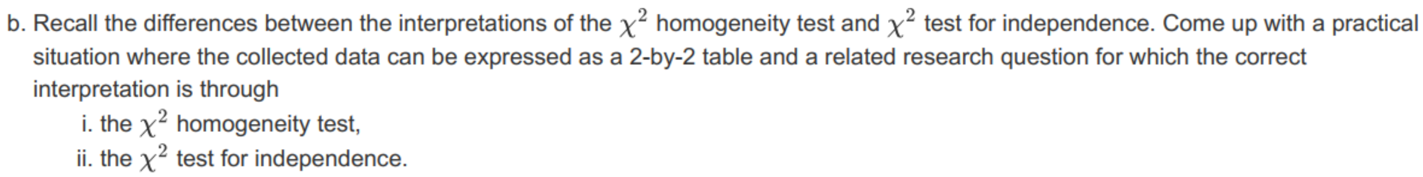
i. a normal distribution,

Since the data has normal distribution, it will coincide directly with normal Q-Q plot, therefore a straight line of points is expected

ii. a right-skew distribution

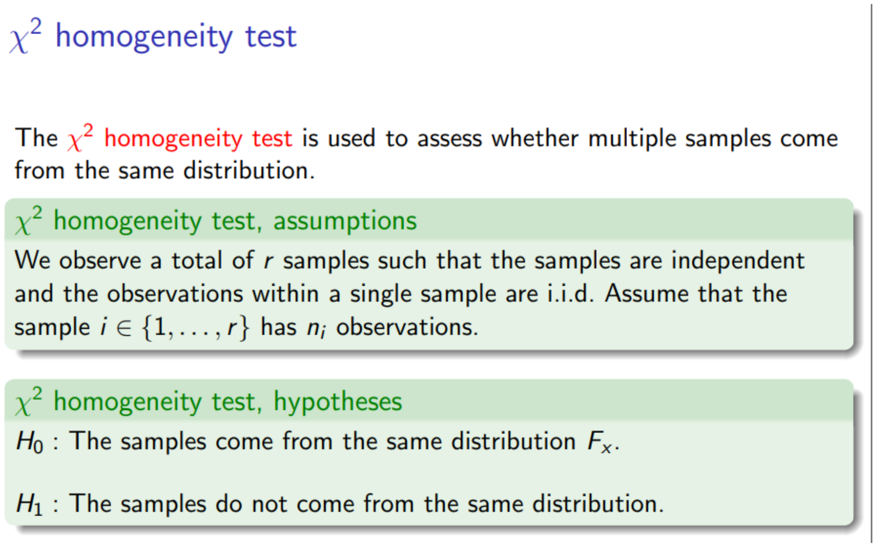
The data has bigger concentration  on the left side and tails away towards the right side, which makes the lower quantile much more concentrated. Therefore, the line of points is curved down towards the lower quantile of the sample axis

iii. a left-skew distribution
The data has bigger concentration on the right side and tails away towards the left side, which makes the higher quantile much more concentrated. Therefore, the line of points is curved up towards the upper quantile of the sample axis
iv. a bimodal distribution
It is the combination of both left and right skew, and the middle part is sparsely concentrated. Thus, the line is broken into three parts, the first part is the below curved down of right skew, the second part is the above curved up of the left skew and the third part is the middle least concentrated data, making it has few data points
v. a distribution with light tails,
It is not skewed so the line is symmetric. Since the tails are light, there would be far fewer data points in the sample that will fall into the tails, thus making the normal QQ plots of the two tails quite thin over the upper and lower quantile
vi. a distribution with heavy tails?
It is not skewed so the line is symmetric. Since the tails are heavy, there would be more data points in the sample that will fall into the tails, thus making the QQ plots of the middle quantile heavily concentrated in the normal QQ plot



b. Recall the differences between the interpretations of the $\chi^2$ homogeneity test and $\chi^2$ test for independence. Come up with a practical situation where the collected data can be expressed as a 2-by-2 table and a related research question for which the correct interpretation is through

i. the $\chi^2$ homogeneity test,
ii. the $\chi^2$ test for independence.



$\chi^2$ homogeneity test

The $\chi^2$ homogeneity test is used to assess whether multiple samples come from the same distribution.

$\chi^2$ homogeneity test, assumptions

We observe a total of $r$ samples such that the samples are independent and the observations within a single sample are i.i.d. Assume that the sample $i \in \{1, \dots, r\}$ has $n_i$ observations.

$\chi^2$ homogeneity test, hypotheses

$H_0$ : The samples come from the same distribution $F_x$.

$H_1$ : The samples do not come from the same distribution.

For example, in a survey of TV viewing preferences, we might ask respondents to identify their favorite program. We might ask the same question of two different populations, such as males and females. We could use a chi-square test for homogeneity to determine whether male viewing preferences differed significantly from female viewing preferences.
The table could be as follows:

| R ↓       K —> | Horror | Comedy | Total |
|---|---|---|---|
| Girls | 30 | 20 | 50 |
| Boys | 80 | 70 | 150 |
| Total | 110 | 90 | 200 |

$E_{1,1}$ = 50 x 110 / 200 = 27.5

$E_{1,2}$  = 50 x 90 / 200 = 22.5

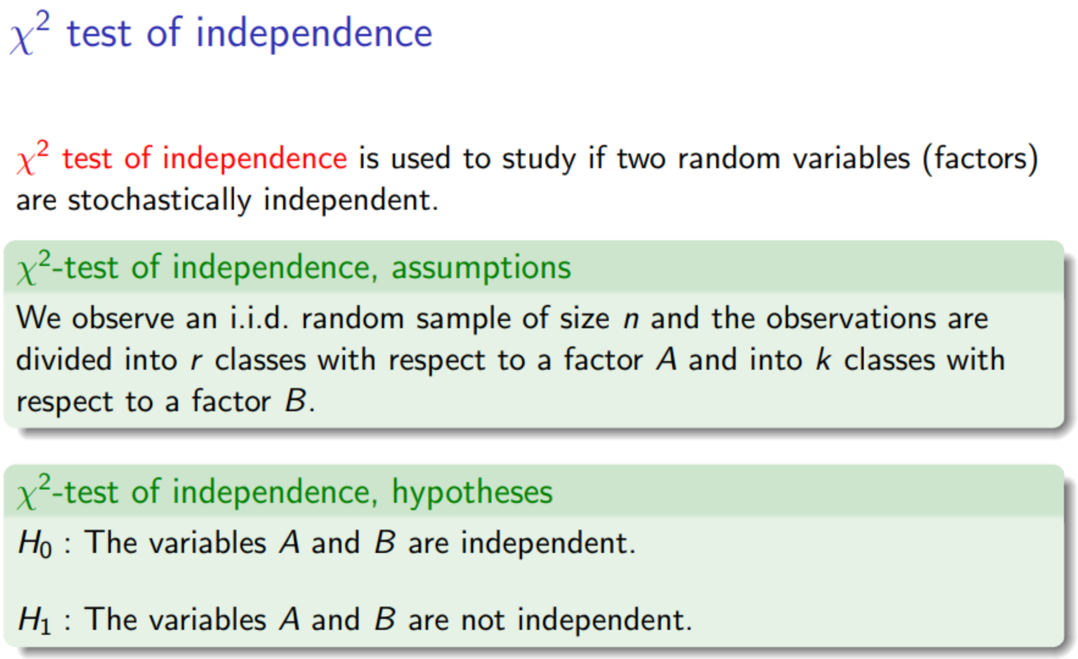
$E_{2,1}$ = 150 x 110 / 200 = 82.5

$E_{2,2}$ = 150 x 90 / 200 = 67.5

Chi-squared = (30 - 27.5)^2/30  + (80 - 82.5)^2 / 80 + ((20 - 22.5)^2) / 20 + (70 - 67.5)^2/70
= 0.208 + 0.078125 + 0.3125 + 0.0892 = 0.6805

Degree of freedom: (r - 1)(k - 1) = (2 - 1)(2 - 1) = 1

Since 0.6805 and 1 are close, H0 is accepted: The distribution of horror and comedy lovers are identical between bois and girls



## $\chi^2$ test of independence

$\chi^2$ test of independence is used to study if two random variables (factors) are stochastically independent.

**$\chi^2$-test of independence, assumptions**

We observe an i.i.d. random sample of size $n$ and the observations are divided into $r$ classes with respect to a factor $A$ and into $k$ classes with respect to a factor $B$.

**$\chi^2$-test of independence, hypotheses**

$H_0$ : The variables $A$ and $B$ are independent.

$H_1$ : The variables $A$ and $B$ are not independent.

We have a list of movie genres; this is our first variable. Our second variable is whether or not the patrons of those genres bought snacks at the theater. Our idea (or, in statistical terms, our null hypothesis) is that the type of movie and whether or not people bought snacks are unrelated. The owner of the movie theater wants to estimate how many snacks to buy. If movie type and snack purchases are unrelated, estimating will be simpler than if the movie types impact snack sales.

| R ↓ K —> | Snacks | No snacks | Total |
|---|---|---|---|
| Horror | 30 | 20 | 50 |
| Comedy | 80 | 70 | 150 |
| Total | 110 | 90 | 200 |

Chi square and degree of freedom are like above: 0.6805 and 1. Since they are close, H0 is accepted: Movie genre and whether snacks are bought are not related