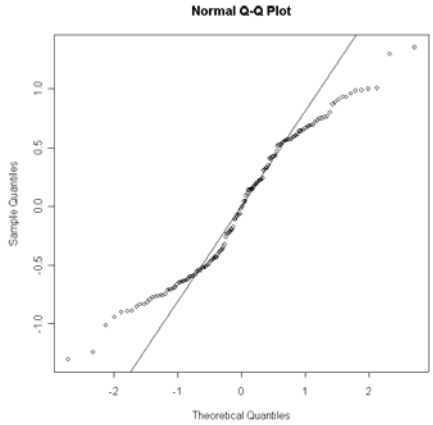
Statistics for Genomic Data Analysis

More about R



http://ludwig-sun2.unil.ch/~darlene/gda/

R: functions for normals

- Generate pseudo-random normals: `> rnorm(…)`
- Probability to the *left* of a value: `> pnorm(…)`
- Quantiles: `> qnorm(…)`
- (Height of the curve: `> dnorm(…)`)
- These 4 fundamental items can be computed for a number of common distributions (e.g. binomial, t, chi-square, etc.): `rbinom(), qt(), pchisq()...`

R: variable types

```
> a <- 49
> sqrt(a)
[1] 7
```

numeric

```
> a <- "The dog ate my homework"
> sub("dog","cat",a)
[1] "The cat ate my homework"
```

character string

```
> a <- (1+1==3)
> a
[1] FALSE
```

logical

R: missing values

- Variables of each data type can also take the value NA (for Not Available)
  - NA is not the same as 0
  - NA is not the same as " " (blank, or empty string)
  - NA is not the same as FALSE
- Any computations involving NA *may or may not* produce NA as a result:

```
> 1+NA
[1] NA
> max(c(NA, 4, 7))
[1] NA
> max(c(NA, 4, 7), na.rm=T)
[1] 7
```

R: functions and operators

- Functions 'do things' with data
  - Input: function *arguments* (0,1,2,…)
  - Output: function *result* (exactly one)
- Exceptions:
  - Functions may also use data that sits in
  - other places, not just in their argument list
  - Functions may also do things other than return a result ("*side effects*")

R: logical and relational operators

- `==` Equal to
- `!=` Not equal to
- `<` Less than
- `>` Greater than
- `<=` Less than or equal to
- `>=` Greater than or equal to
- `is.na(x)` Missing?
- `&` Logical AND
- `|` Logical OR
- `!` Logical NOT

R: vectors

- *vector* : an ordered collection of data of the *same type:*

```
> a <- c(7,5,1)
> a*2
[1] 14 10 2
```

- *Example* : the mean spot intensities of all 15488 spots on a chip; in R, a vector of 15488 numbers
- A single number is the special case of a vector with 1 element
- Other vector types: character strings, logical

R: matrices and arrays

- *matrix* : a rectangular table of data of the *same type*
- *Example* : expression values of 10000 genes for 30 tissue biopsies; in R, a matrix with 10000 *rows* and 30 *columns*
- *array* : a multiply indexed collection of data entries (*e.g.* a 'stack' of matrices)
- *Example* : the red and green foreground and background values for 20000 spots on 120 chips; in R, a 4 x 20000 x 120 (3D) array

R: lists

- *vector* : an ordered collection of data of the *same type;* individual elements accessed with [ ]

```
- > a <- c(7,5,1)
- > a[2]
- [1] 5
```

- *list* : an ordered collection of data of *arbitrary types;* individual elements accessed with $

```
> doe <-
  list(name="john",age=28,married=FALSE)
> doe$name
[1] "john"
> doe$age
[1] 28
```

R: data frames

- *data frame* : the type of R object normally used to store a data set
- A data frame is a rectangular table with rows and columns
  - data *within* each column has the *same type* (*e.g.* number, character, logical)
  - different columns may have *different types*
- *Example:*

```
> tumor.info
      localisation tumorsize progress
XX348     proximal       6.3    FALSE
XX234       distal       8.0     TRUE
XX987     proximal      10.0    FALSE
```

R: subsetting (or indexing)

- A very powerful feature of R is the *subset operator* `[ ]`
- Allows access to *individual elements* of a vector, matrix, array or data frame are by specifying their index(es), or their name(s)
- *Examples* :

```
> tumor.info[3, 2]
[1] 10
> tumor.info["XX987", "tumorsize"]
[1] 10
> tumor.info["XX987",]
      localisation tumorsize progress
XX987     proximal        10    FALSE
```

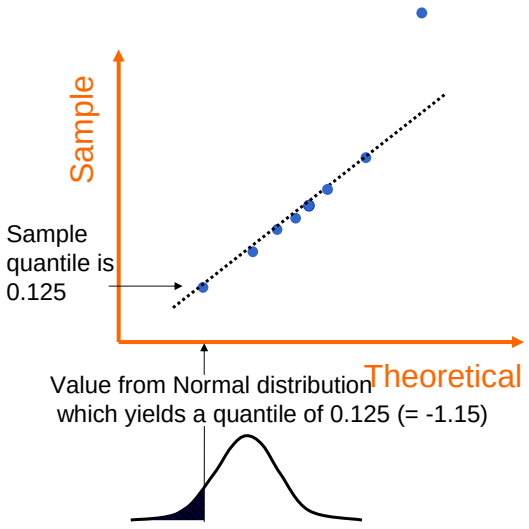
R: importing and exporting data

- *Many ways* to get data into and out of R
- One straightforward way is to use *tab-delimited text files* (e.g. save an Excel sheet as tab-delimited text, for easy import into R)
- Useful R functions: `read.delim(), read.table(), read.csv(), write.table()`
- *Example* :

```
> x = read.delim("filename.txt")
> write.table(x, file="x.txt",
  sep="\t")
```

QQ-Plot

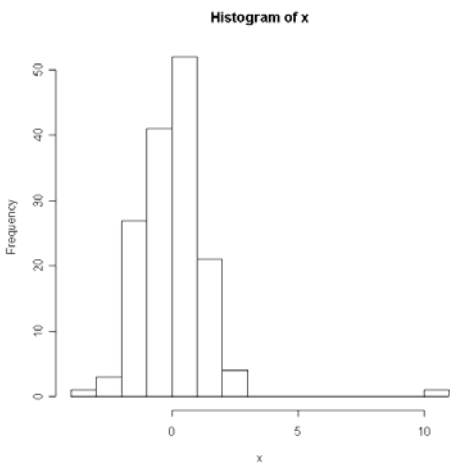
- Quantile-quantile plot
- Used to assess whether a sample follows a particular (e.g. normal) distribution (or to compare two samples)
- A method for looking for outliers when data are mostly normal

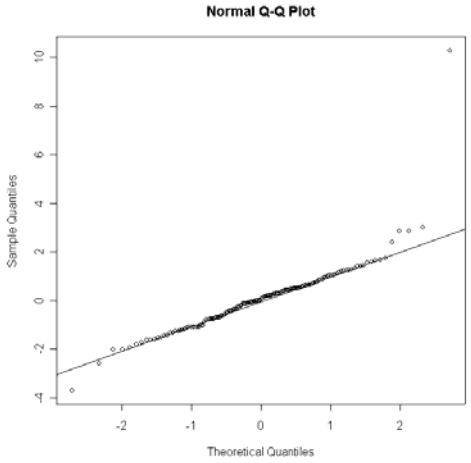


Typical deviations from straight line patterns

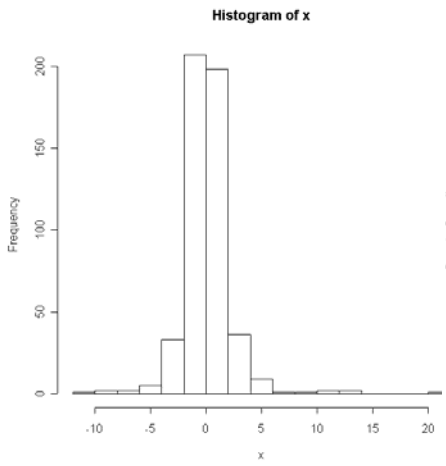
- Outliers
- Curvature at both ends (long or short tails)
- Convex/concave curvature (asymmetry)
- Horizontal segments, plateaus, gaps

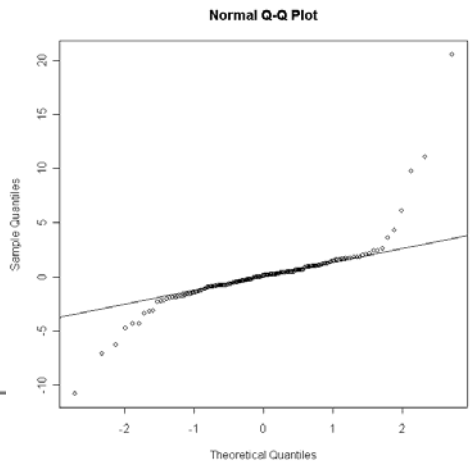
Outliers



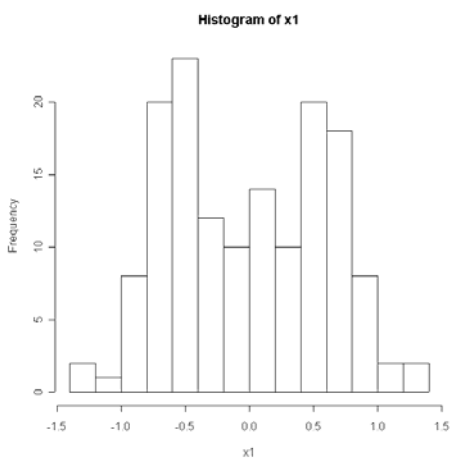


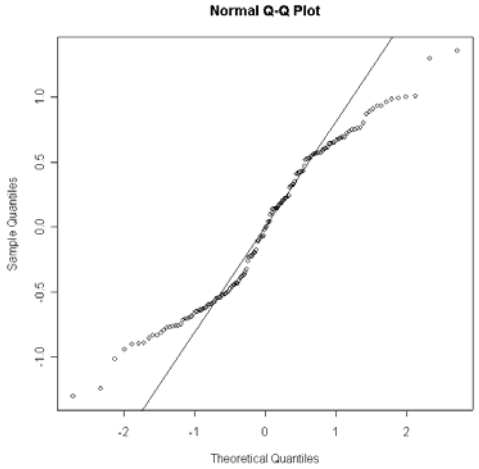
Long Tails



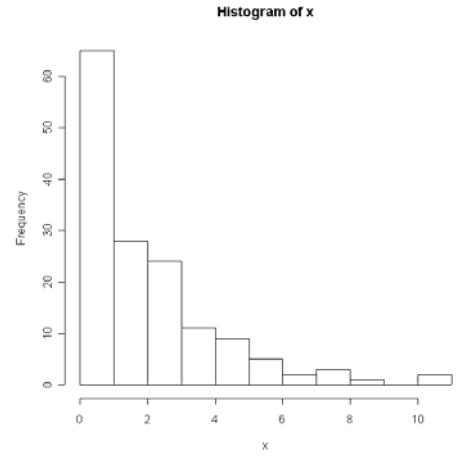


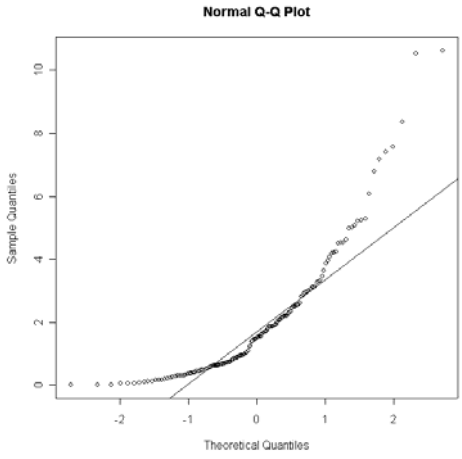
Short Tails





Asymmetry





Plateaus/Gaps