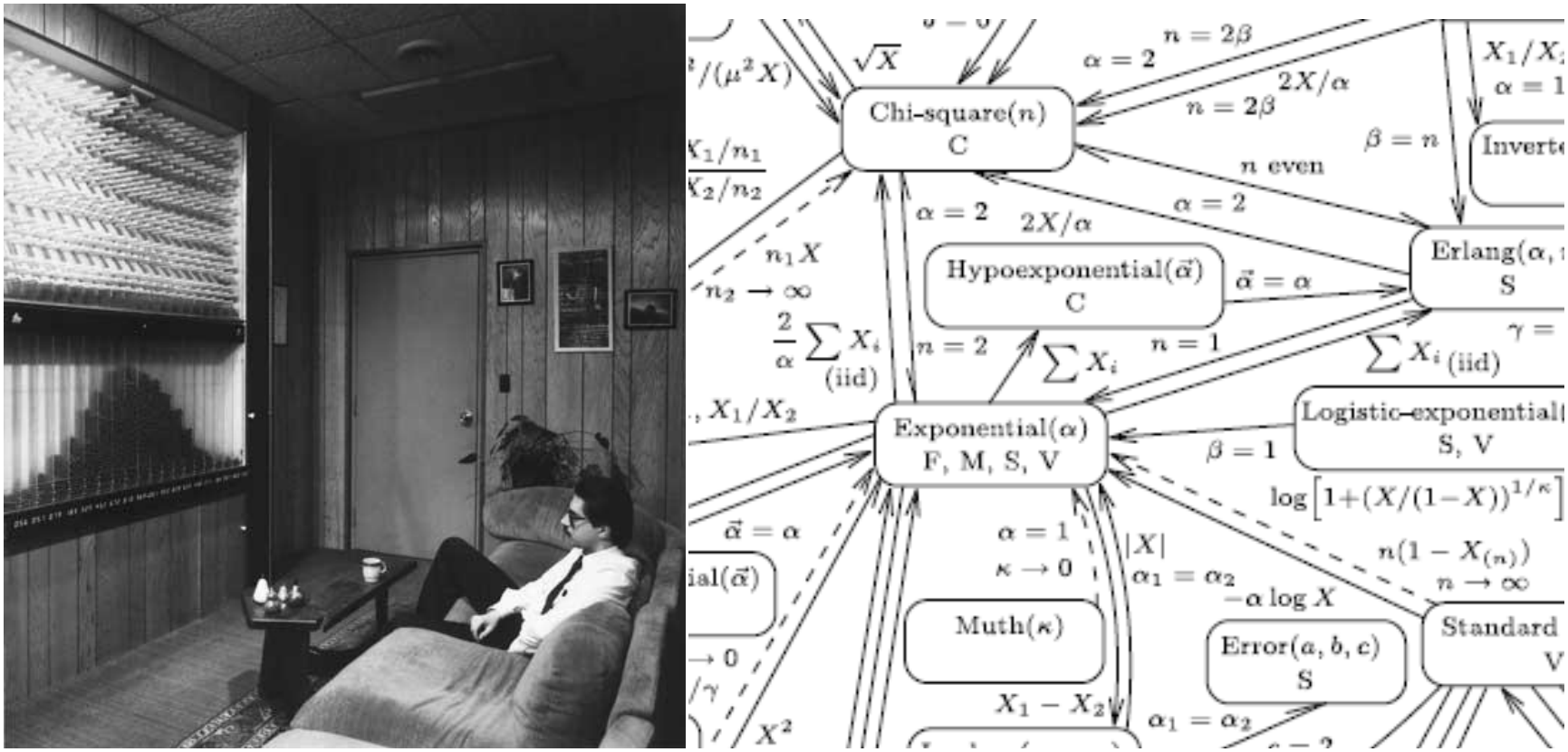
# Probability distributions




The Random Mechanical Cascade in action in the PEAR laboratory. Courtesy the PEAR Archives.

**Probability distributions -in DATA**

In each analysis of biological data,

There is always at least one

**response** or **dependent variables**

and usually

**explanatory** or **independent variables**

The values of explanatory variables can provide information which can be used to predict values of response variables

**Sources of stochasticity - randomness in response variables**

Quantitative response variables are measured with a certain precision (often *ignored*).

Sampling effects: Responses can be samples of a population which is variable. Different samples can have different compositions.

The data are generated by an inherently stochastic process.

**Probability distributions**

**Response variables** are **random variables**
Explanatory variables can be random variables or not

General goal of an analysis:

Understand the **data generating mechanism**

Approach:

**statistical modelling**
of the probability distributions of response variables

**including model comparison**
**hypothesis testing** or another approach (**AIC**, **cross-validation**), to compare different models

# ASIDE: Usage of probability distributions in inference

## Example: Is Euro a fair coin?

Soon after the Euro was introduced as currency in Europe, it was widely reported that someone had spun a Euro 250 times and gotten heads 140 times.

a. Estimate the true proportion of heads using a 95% confidence interval. (remember to check conditions)

$$CI: \hat{p} \pm z^* \sqrt{\frac{\hat{p}\hat{q}}{n}} = .56 \pm 1.96 \sqrt{\frac{(.56)(.44)}{250}} = .56 \pm .062$$

$$CI: (.488, .622)$$

b. Does your confidence interval provide evidence that the coin is unfair when spun? Explain.

c. What is the significance level?

**Probability distributions**

Many types of response variables have typical probability distributions
Probability distributions can be grouped in families

- Distributions for **counts**

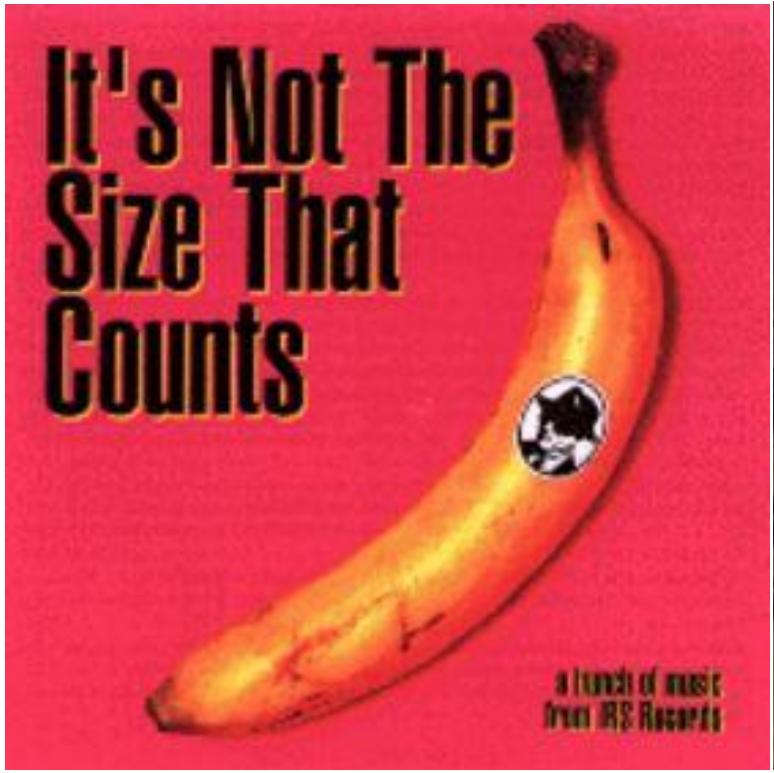
e.g. *number of males, number of pollinators / plant, number of reads*

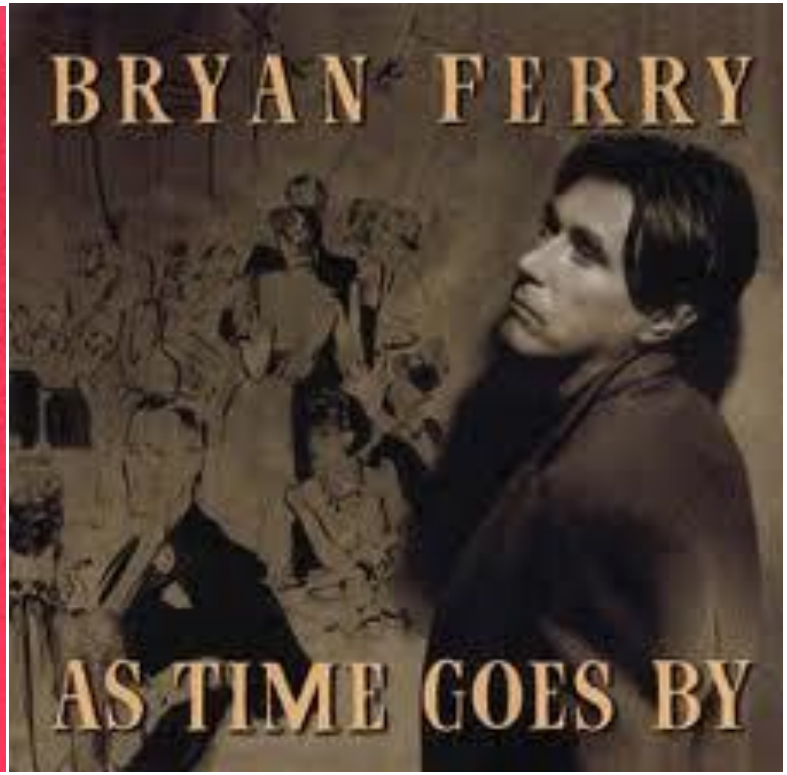
- The **Gaussian** family

e.g. *body size measurements, quantitative data & continuous values*

- Distributions for **durations**

e.g. *lifespan in a given environment*


It's Not The
Size That
Counts


BRYAN FERRY
AS TIME GOES BY

**Counts**

Discrete uniform: All values of the response variable are equally likely

Bernoulli - Binomial: success 1 – failure 0
Multinomial: more categories of outcome

Poisson: independent events happening at random to an individual

Negative Binomial:
A) waiting time to *c* events of a type (two possible types of events)
B) clustered (overdispersed) count data ≡ non-independence

Geometric: waiting time to the first event

Zeta: convenient distribution for ranks

**The Gaussian family**, continuous variables

Normal

Power-Transformed Normal: Box Cox transformation

Log Normal: The log data values are normally distributed

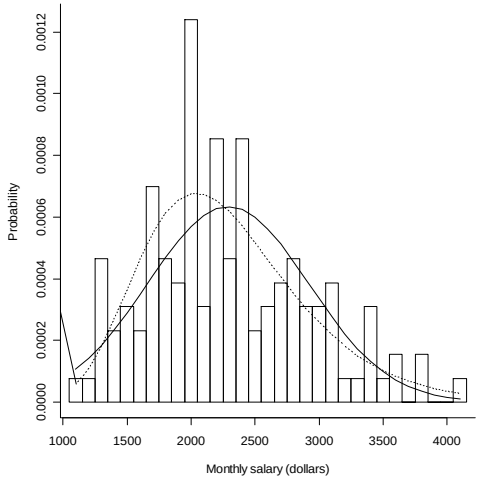
Inverse Gaussian: Waiting time for hitting a "ceiling" in a random walk

Logistic: Bit narrower than normal but with thicker tails

Log-Logistic

f2 <- c(1,1,6,3,4,3,9,6,5,16,4,11,6,11,3,4,5,6,4,4,5,1,1,4,1,2,0,2,0,0,1**) # monthly salaries**

y2 <- seq(1100,4100,by=100) # categories, per 100 dollar




```
z2 <- fit.dist(y2,f2,"normal",delta=100,plot=T,
        #delta gives measurement precision
xlab="Monthly salary (dollars)",main="",bty="L")
z2a <- fit.dist(y2,f2,"logistic",delta=100)
z2b<- fit.dist(y2,f2,"log normal",delta=100,plot=T,
z2c <- fit.dist(y2,f2,"gamma",delta=100)
z2d<- fit.dist(y2,f2,"inverse Gauss",delta=100)
```

**Durations**

*Discrete Time*: Negative Binomial, Geometric, zeta
*Continuous Time*: Inverse Gaussian, log normal, log logistic

Exponential: Constant intensity of events, continuous analogue of the geometric

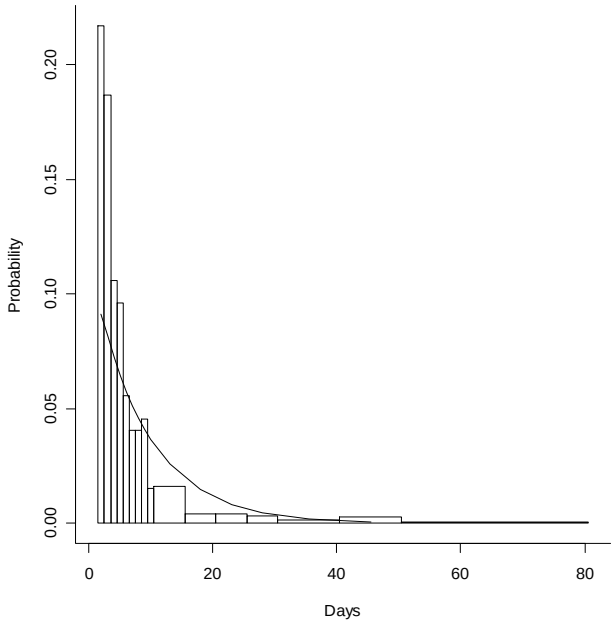
Pareto: Tail of a complex distribution, continuous analogue of the zeta

Gamma: Total duration is the sum over periods with different but constant intensities

Weibull: Several processes in parallel

Extreme Value: extreme phenomena

```
 f2 <- c(43,37,21,19,11,8,8,9,3,16,4,4,3,3,5,4) # frequencies of durations of strikes
length(f2)
y2 <- c(seq(1.5,10.5,by=1),seq(15.5,30.5,by=5),40.5,50.5,80.5)
# y2 contains breaks, not midpoints
z2<- fit.dist(y2,f2,"exponential",breaks=T,censor=T,plot=T ,
        xlab="Days",main="",bty="L")
```



How do you know from which distribution data points are generated?

- Expectations from existing theory, previous analysis

- Kolmogorov-Smirnov test & other tests for specific distributions

- Graphical inspection QQ Plots

- Likelihood comparisons

Expectations from existing theory

**The law of large numbers**
**Central limit Theorem**


THÉORIE

ANALYTIQUE

DES PROBABILITES;

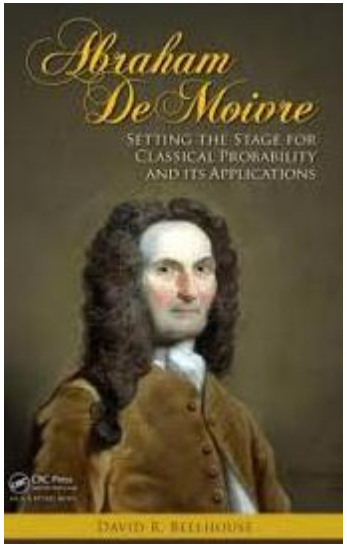
PAR M. LE COMTE LAPLACE,

Chancelier du Sénat-Conservateur, Grand-Officier de la Légion d'Honneur; Membre de l'Institut impérial et du Bureau des Longitudes de France; des Sociétés royales de Londres et de Gottingue; des Académies des Sciences de Russie, de Danemarck, de Suède, de Prusse, de Hollande, d'Italie, etc.

PARIS,

Mme Ve COURCIER, Imprimeur-Libraire pour les Mathématiques, quai des Augustins, n° 57.

1812.



**example: genotypic value of a pig**

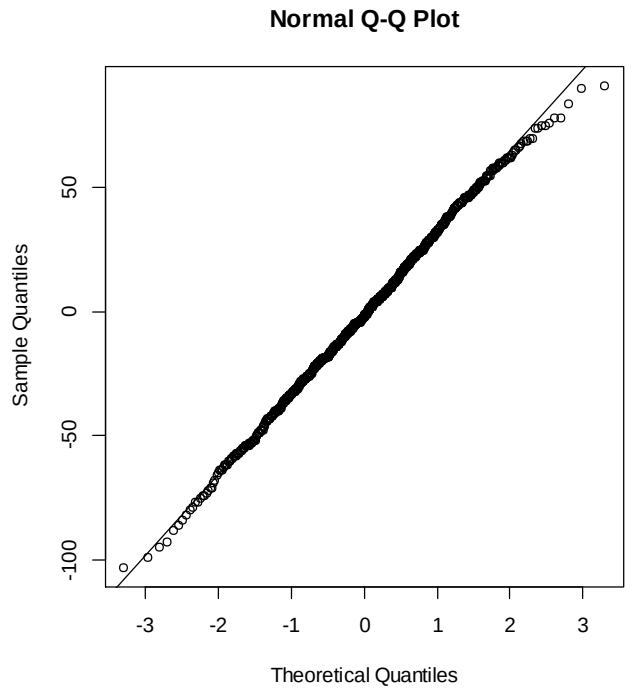
Additive contributions of many loci with small effect to phenotype
We center genotypic value around zero, by substracting the expected genotypic value.

```
gvalue<-function(n,p)sum(rbinom(n,1,p))-n*p
```

Simulate them:

```
individuals<-numeric(1000)
for(i in 1:1000)individuals[i]<-gvalue(5,0.7)
hist(individuals)
qqnorm(individuals); qqline(individuals)
```

```
individuals2<-numeric(1000)
for(i in 1:1000)individuals2[i]<-gvalue(500,0.7)
hist(individuals2)
qqnorm(individuals2); qqline(individuals2)
```



**ASIDE: Parallel between data simulations and simulations used in inference:**

Statistics calculated from data which are random variables are themselves random variables - they can be simulated.

```
nsuccess<-function(n,p)sum(rbinom(n,1,p)) # n euro tosses
```

The summed response over *n* observations of a bernoulli trial with *p* the probability of success per trial.

```
statistics<-numeric(100) # 100 experiments
for(i in 1:100)statistics[i]<-nsuccess(5,0.5)
hist(statistics)
qqnorm(statistics); qqline(statistics)
```

## ASIDE: Statistics: their most recurring probability distributions

- Normal distribution
- student t distribution
- Chisq distribution
- F distribution

How do you know from which distribution data points are generated?

- Expectations from existing theory

These can be tested

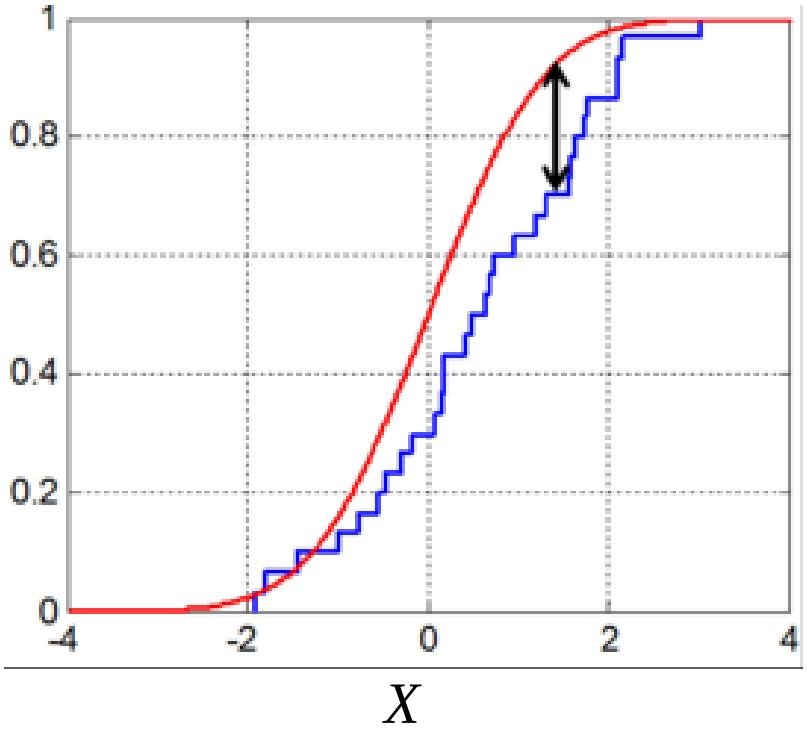
- Kolmogorov-Smirnov test - other tests for specific distributions

- QQ Plots graphical inspection

- Likelihood comparisons

**Kolmogorov Smirnov test**
**Compare an empirical (data) distribution with a reference**

Cumulative probability functions reference $F(x)$ (red) and data $F_n(x)$ (blue)



**Kolmogorov Smirnov test**

**a data distribution with a reference**

The *Kolmogorov–Smirnov statistic* quantifies a distance between the *empirical distribution function* of the sample and the *cumulative distribution function of the reference distribution*

Empirical distribution function $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{X_i \leq x}$

with *I* the indicator function (0 or 1)

Kolmogorov-Smirnov statistic: $D_n = \sup_x |F_n(x) - F(x)|$

$\sqrt{n}D_n$ is compared to the Kolmogorov distribution to which this statistic converges for *n* going to infinity

## Kolmogorov Smirnov test

Comparison of distances between cumulative distribution functions of a data sample and a reference (theory), **of two samples**

Advantages:

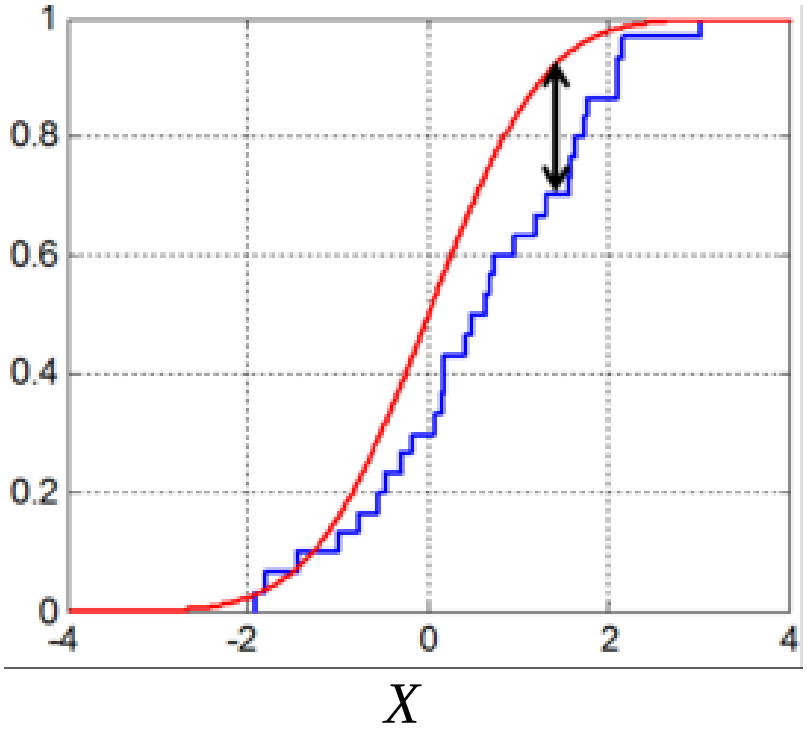
- The distribution of the K-S test statistic does not depend on the underlying cumulative distribution function being tested
- It has asymptotic power 1
- It can be used to get a confidence band for $F(x)$

Limitations:

- It only applies to continuous distributions
- more sensitive near the center of the distribution than at the tails
- the distribution must be fully specified. That is, if location, scale, and shape parameters are estimated from the data, the critical region of the K-S test is no longer valid. Few corrected tables exist for that situation.

## QQ Plots

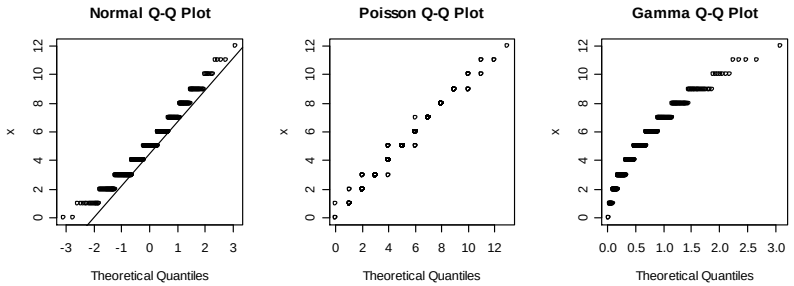
Cumulative probability functions $F(x)$ (red) and $F_n(x)$ (blue)




**Can we compare these in more detail than with the KS test?**

# QQ Plots

Qualitative, no statistic calculated, no hypothesis test, can be made for any theoretical distribution
Question that can be answered: Is the approximation acceptable?



## QQ Plots

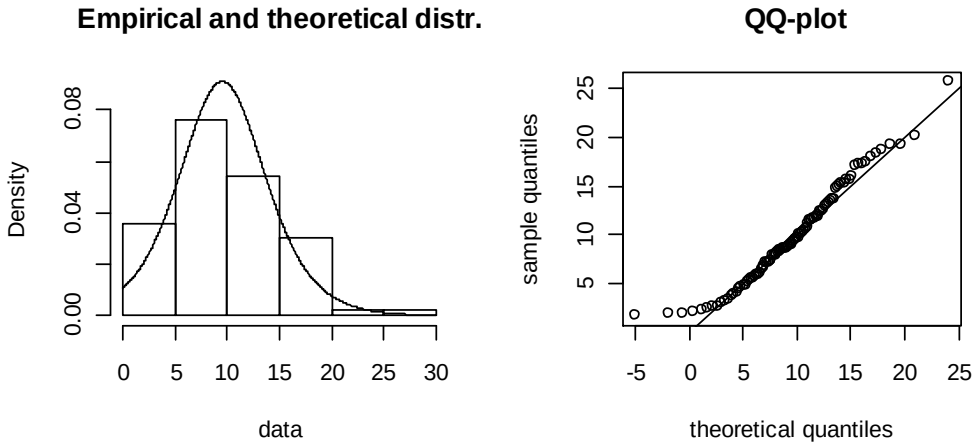
- Qualitative, no statistic calculated, no hypothesis test
- A plot of the quantiles of two distributions against each other, or a plot based on "predictions" of these quantiles for a reference distribution.

Quantiles?
Theoretical distribution → Invert cumulative distribution function $F(x)$ (maybe with interpolation added)
Empirical distribution function → similar, use rules for plotting *positions*

The median is the 0.5 quantile $q_{0.5}$

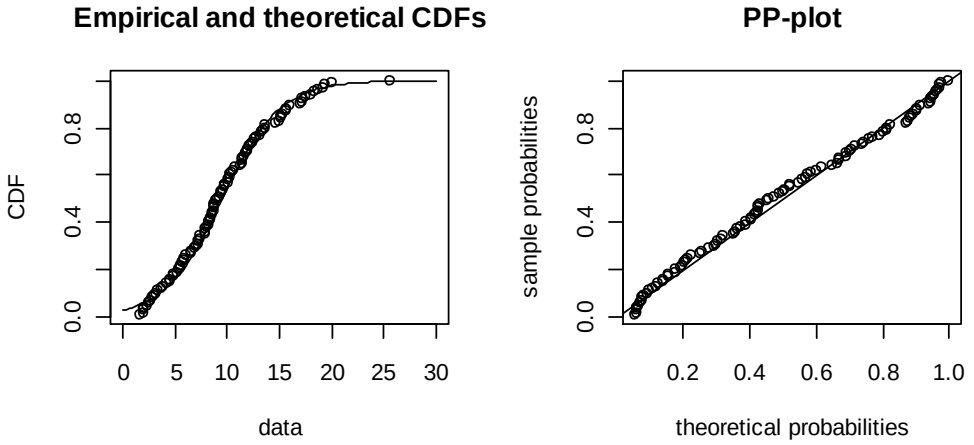
Given two cumulative probability distribution functions $F$ and $G$, with associated quantile functions $F^{-1}$ and $G^{-1}$ (the inverse function of the CDF is the quantile function), the Q–Q plot draws the $q$th quantile of $F$ against the $q$th quantile of $G$ for a range of values of $q$. Thus, the Q–Q plot is a parametric curve indexed over [0, 1] with values in the real plane $\mathbf{R}^2$.

# QQ Plots, example using the fitdistrplus library.







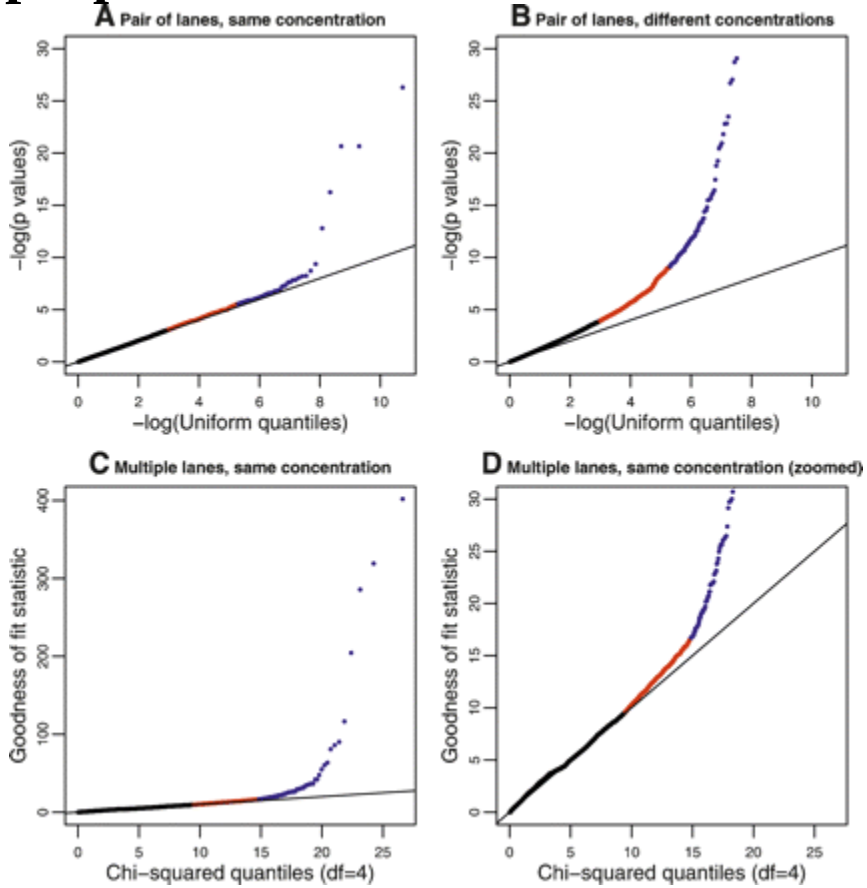






**Note the PP-plot**

# ASIDE: QQ plot approaches are applied similarly to study properties of statistics




RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays *Genome Res. 2008. 18: 1509-1517*

**Likelihood comparisons**

Likelihood is an essential concept in modern statistics

Given a data generating mechanism *B*, we can use the **conditional probability** Pr(*A*|*B*) to reason about the relative frequency at which data *A* occur

Alternatively,
given the data *A*, we can use the **likelihood function,** written as $L(B|A)$ or $L(A|B)$, or the **likelihood** $L(B)$ to reason about the plausibility of *B*.

The likelihood is in fact $L(B) = L(A|B) = \Pr(A|B)$

A likelihood function $L_f(x \mid \theta)$ describes the probability or probability density for the occurrence of a sample (data) configuration $x$ given that a probability density $f$ with parameter $\theta$ is assumed to specify the data generating mechanism. This probability is seen as a function where $x$ is fixed and $\theta$ varies.

Likelihood large $\rightarrow$ The assumed $f$ with parameter $\theta$ is plausible

Given $x$, the likelihood $L_f(\theta)$ or $L(\theta)$ is the probability or probability density for the occurrence of a sample configuration $x$ given that parameter $\theta$ is assumed to specify the data generating mechanism ($f$ explicit or implicit).

Likelihoods can be used to rank different $\theta$ or $B$ according preference
Likelihood large $\rightarrow$ The assumed $f$ is plausible $\rightarrow$ preferred
Likelihood small $\rightarrow$ The assumed $f$ is implausible

## Likelihood

A sample of great tit nestlings, with 3 males and 6 females.

The likelihood that a binomial distribution with $p_{males} = 0.5$ is the data generating mechanism

$$\binom{9}{3} 0.5^3 (1-0.5)^6 = 0.164$$

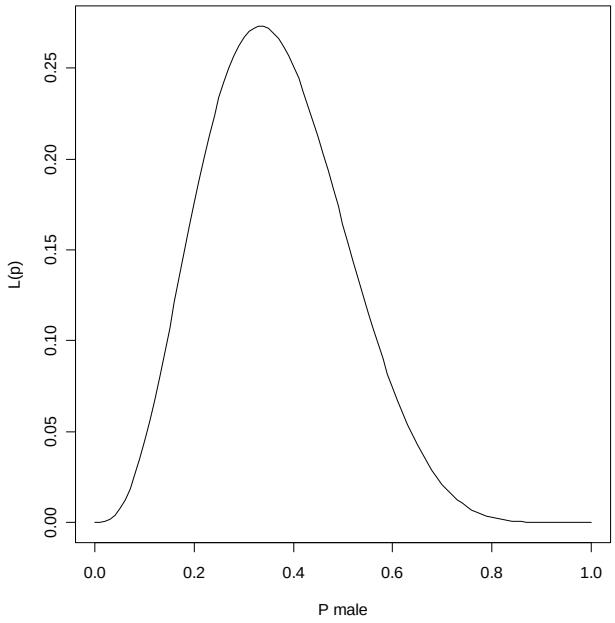
The likelihood that a binomial distribution with $p_{males} = 0.2$ is the data generating mechanism

$$\binom{9}{3} 0.2^3 (1-0.2)^6 = 0.176$$

in R?

# Likelihood

For $p = 1/3$, the likelihood is largest. That model is in a sense the most plausible binomial one, given the data. $\hat{p} = 1/3$ is called the **maximum likelihood estimate** of the parameter of the binomial distribution



**Likelihood comparisons of distributions**

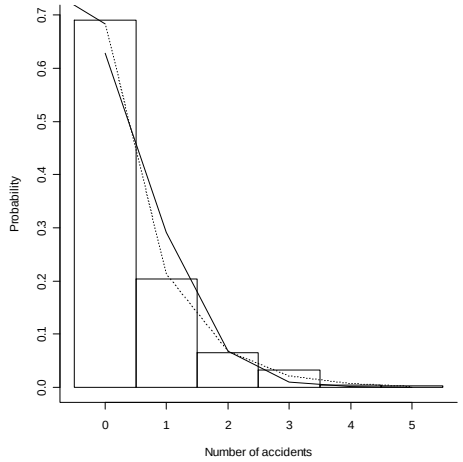
Compare likelihoods of different values of the parameter of the same type of distribution

Compare likelihoods of different types of distributions, with parameter values per type at the values which give the maximum likelihood.

e.g., *Poisson* versus *negative binomial*

```
library(MASS)

help(fitdistr) # fits univariate probability distributions to data

# we will now different function for illustrating and comparing probability distributions
# it is a bit more tedious to use than fitdistr(), but more versatile

library(gnlm) # install -only if you feel the need- from zip file
http://www.commanster.eu/rcode.html

f3 <- c(447,132,42,21,3,2) # Data Lindsey table Exercise 1.2 Car Accidents
y3 <- seq(0,5) # categories histogram
z3 <- fit.dist(y3,f3,"Poisson",plot=T,xlab="Number of accidents", main="",bty="L")

Poisson distribution,  n = 647

     mean  variance    mu.hat
0.4652241 0.6908308 0.4652241

-log likelihood           AIC (-log likelihood +number of parameters)
       27.54904        28.54904
```

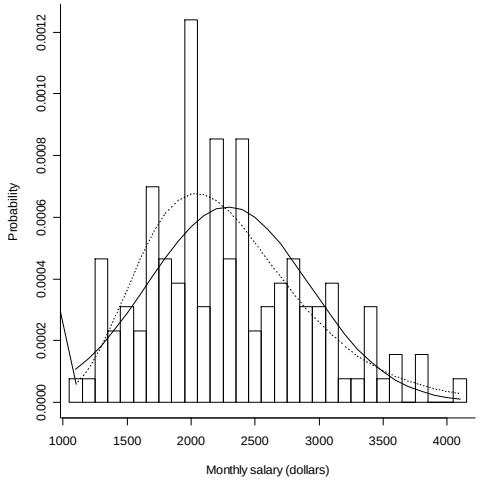
```
z3a <- fit.dist(y3,f3,"negative binomial",exact=F,plot=T,add=T,lty=3)

negative binomial distribution,  n = 647

     mean  variance    nu.hat gamma.hat
0.4652241 0.6908308 0.6734270 0.9593397

-log likelihood           AIC
      2.742104       4.742104
```




full : Poisson maximum likelihood model
dotted: negative binomial ML model

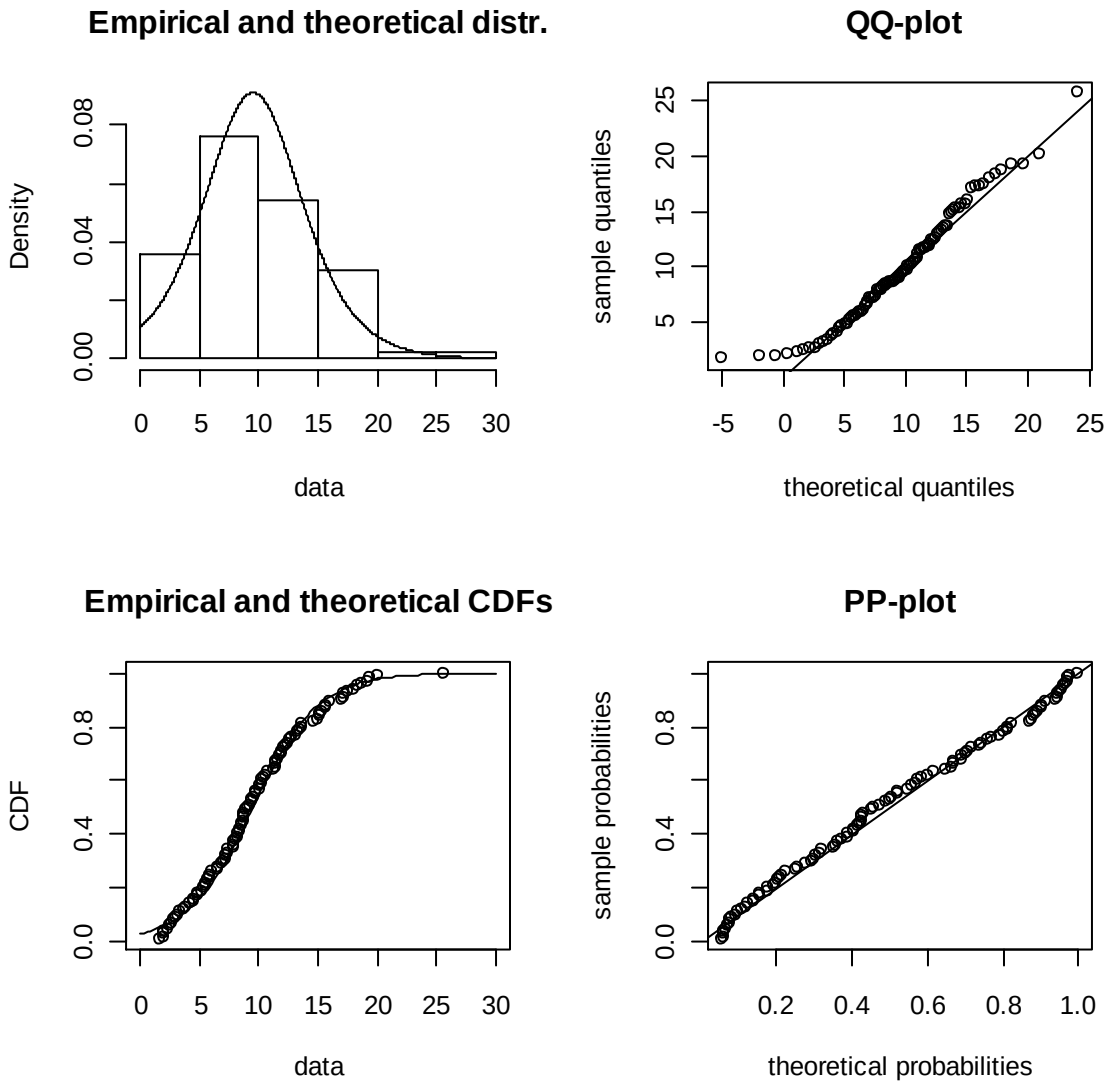
f2 <- c(1,1,6,3,4,3,9,6,5,16,4,11,6,11,3,4,5,6,4,4,5,1,1,4,1,2,0,2,0,0,1**) # monthly salaries**

y2 <- seq(1100,4100,by=100) # categories, per 100 dollar
make graphs for all five




| | Normal | Logistic | Log-Normal | Gamma | Inverse Gaussian |
|---|---|---|---|---|---|
| -log Likelihood | 23.32734 | 25.03046 | 19.82553 | **19.61159** | 19.69305 |

```
# using the fitdistrplus package WARNING: it uses function name fit.dist also!
library(fitdistrplus)
# Example for a logistic distribution

xn<-rnorm(n=100,mean=10,sd=5) # here the function starts from raw data, not a histogram
summary(logisfit<-fitdist(xn,"logis"))

Fitting of the distribution ' logis ' by maximum likelihood
Parameters :
         estimate Std. Error
location 9.558479  0.4905856
scale    2.800738  0.2319048
Loglikelihood:  -301.5674   AIC:  607.1349   BIC:  612.3452
Correlation matrix:
          location      scale
location 1.00000000 0.03713437
scale    0.03713437 1.00000000

plotdist(xn,"logis",para=list(location=logisfit$estimate[1],scale= logisfit$estimate [2]))
```


Empirical and theoretical distr.
Density
0.00
0.04
0.08
0
5
10
15
20
25
30
data
QQ-plot
sample quantiles
5
10
15
20
25
-5
0
5
10
15
20
25
theoretical quantiles
Empirical and theoretical CDFs
CDF
0.0
0.4
0.8
0
5
10
15
20
25
30
data
PP-plot
sample probabilities
0.0
0.4
0.8
0.2
0.4
0.6
0.8
1.0
theoretical probabilities

How does the QQ plot work here?

```
sort(xn) # sort all data values
length(xn) # how many observations?

sort(xn)[1] # the smallest value. 1/100 of data values are not larger than this

# Given the theoretical distribution, what is the value of the random variable
# where 1/100 of values are not  larger than that value? = they are smaller or equal

qlogis(0.5/100,logisfit$estimate[1],logisfit$estimate[2])
# evaluated at midpoint of the first bin
# see ?ppoints on this choice

sort(xn)[2] # the second smallest value. 2/100 of data values are not larger than this
qlogis(1.5/100,logisfit$estimate[1],logisfit$estimate[2]) # midpoint of the second bin

# and so on...
```

**References**

Lindsey, JK (1995) Introductory Statistics: A Modelling Approach. Oxford University Press.

http://cran.r-project.org/web/views/Distributions.html

http://cran.r-project.org/web/packages/bbmle/vignettes/mle2.pdf