

Wednesday, November 19, 2014

Maximum likelihood estimation

We never have access to
the **true** probability distribution of population
neither
the **true** sampling probability distribution

We can only get estimates !

What is a good estimator ?

What are the good estimators for the mean, the
standard deviation, or for the proportion ?

Consider a data vector y of individual observations y_i :

$$y = (y_1, y_2, y_3, \dots, y_m)$$

which is a random sample from an unknown population.

Let $f(y|w)$ be the probability density function (PDF) of the vector y given the parameter vector w

$$w = (w_1, w_2, \dots, w_k)$$

If the individual observations are independent then, according to probability theory, the PDF for the data y can be expressed as a multiplication of the PDFs for individual observations:

$$f(y|w) = f(y = (y_1, y_2, \dots, y_m)|w) = f_1(y_1|w)f_2(y_2|w)\dots f_m(y_m|w)$$

Consider a simplest case where $m=k=1$

and that y represents the success in 10 Bernoulli trials and the success in each individual trials is given by w .

Under these conditions the PDF for y is a binomial distribution:

$$f(y|n=10, w) = \frac{10!}{y!(10-y)!} w^y (1-w)^{10-y}$$



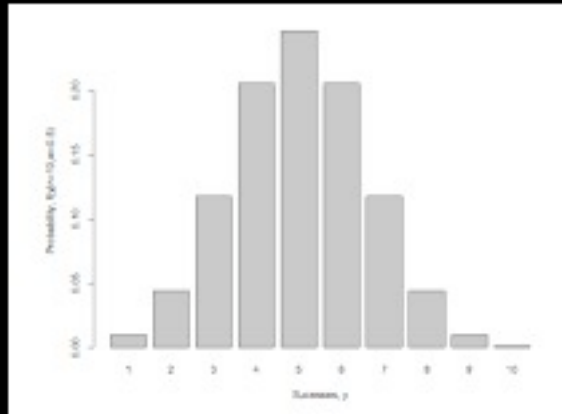
Tossing a 1 euro fair coin 10 times and success getting "1" face

The PDF for y is given by:

$$f(y|n=10, w=0.5) = \frac{10!}{y!(10-y)!} 0.5^y (1-0.5)^{10-y}$$

$$y \in \{1, 2, \dots, 10\}$$

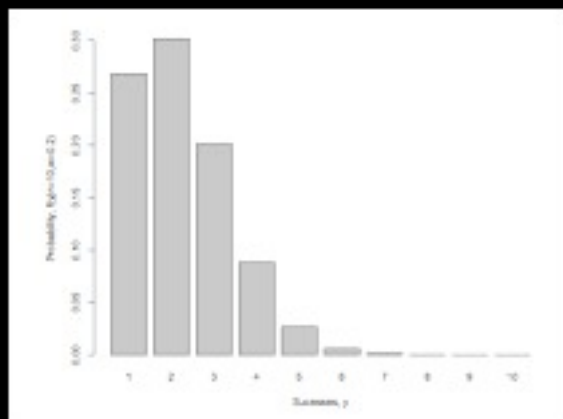
$$f(y|n=10, w=0.5) = \frac{10!}{y!(10-y)!} 0.5^y (1-0.5)^{10-y}$$



```
n=10
y=1:n
w=0.5
fy=factorial(n)/(factorial(y)*factorial(n-y))*w^y*(1-w)^(n-y)
barplot(fy,names.org=y,xlab="Successes, y",ylab="Probability, f(y|n=10,w=0.5)")
```



What if the coin is "unfair"?



```
n=10
y=1:n
w=0.2
fy=factorial(n)/(factorial(y)*factorial(n-y))*w^y*(1-w)^(n-y)
barplot(fy,names.org=y,xlab="Successes, y",ylab="Probability, f(y|n=10,w=0.2)")
```



Suppose we take a coin (fair or unfair) and toss it 10 times and get:

$$y=4$$

What is the best estimate for the value of w ?

$$w=?$$



Maximum likelihood estimation of w

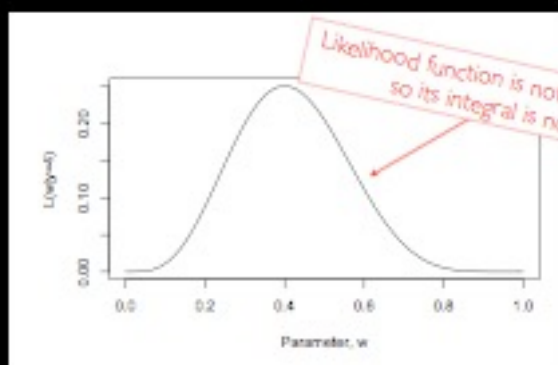
We define the likelihood function of the parameter w given the data y as:

$$L(w|y) = f(y|w)$$

For the one parameter binomial distribution this is:

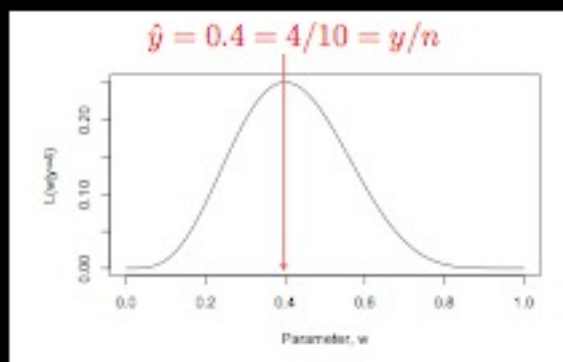
$$L(w|y=4) = f(y=4|n=10, w) = \frac{10!}{4!(10-4)!} w^4 (1-w)^{10-4}$$

How does the likelihood function look like?



```
n=10
y=4
w=seq(0,1,by=0.01)
fy=factorial(n)/(factorial(y)*factorial(n-y))*w^y*(1-w)^(n-y)
plot(w,fy,main="Likelihood function",xlab="Parameter, w",ylab="L(w|y=4)",type="l")
```

What is the maximum of the likelihood function ?



Maximum likelihood estimation

For convenience MLE are obtained by maximizing the log-likelihood function

$$\ln L(w|y)$$

Maximum likelihood estimation

At maximum of a function its first derivative vanishes ...

$$\frac{\partial \ln L(w|y)}{\partial w} = 0$$

... and its second derivative is negative

$$\frac{\partial^2 \ln L(w|y)}{\partial w^2} < 0$$

Maximum likelihood estimation

Setting $y=4$ in the log-likelihood yields:

$$\begin{aligned}\ln L(w|y=4) &= \ln \left(\frac{10!}{4!(10-4)!} w^4 (1-w)^{10-4} \right) \\ &= \ln \left(\frac{10!}{4!6!} \right) + 4\ln(w) + 6\ln(1-w)\end{aligned}$$

Maximum likelihood estimation

The first derivative of the log-likelihood is

$$\frac{\partial \ln L(w, |y=4)}{\partial w} = \frac{4}{w} - \frac{6}{1-w}$$

Setting this to zero...

$$\frac{4}{w} - \frac{6}{1-w} = 0$$

...and solving to w yields:

$$w = 0.4$$

Maximum likelihood estimation

The second derivative of the log-likelihood at $w = 0.4$ is negative

$$\frac{\partial^2 \ln L(w|y=4)}{\partial w^2} = -\frac{4}{w^2} - \frac{6}{(1-w)^2} < 0$$

Wednesday, November 19, 2014



Uncertainty and uncertainty propagation

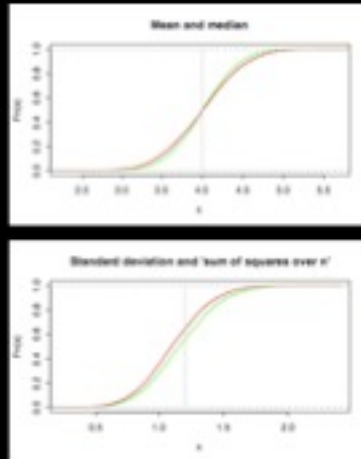
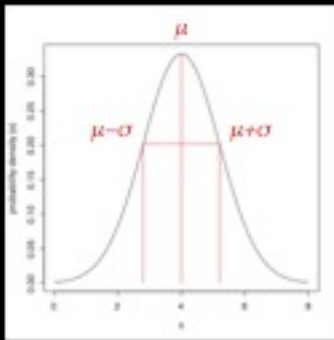
We never have access to
the **true** probability distribution of population
neither
the **true** sampling probability distribution

We can only get estimates !

What is a good estimator ?

What are the good estimators for the mean, the
standard deviation, or for the proportion ?

sampling



(for samples of Gaussian distributed data)

sample mean gives better estimates of the true mean than median:

sample mean is more precise than median

standard deviation of the sample gives better estimates of the true standard deviation than the 'non-standard :-)' deviation

sample standard deviation is more accurate

(however for non-gaussian distribution median can be more robust)

high accuracy

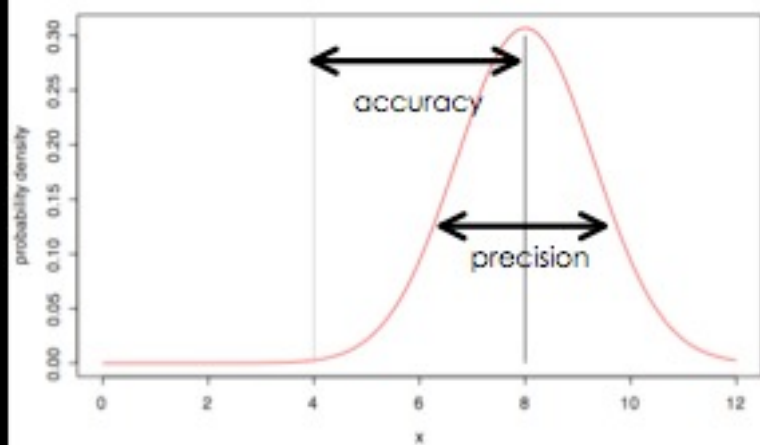
low accuracy

low precision



high precision



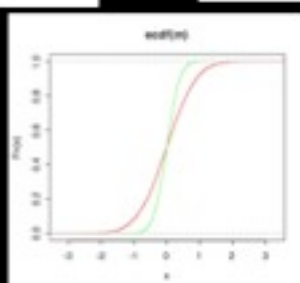
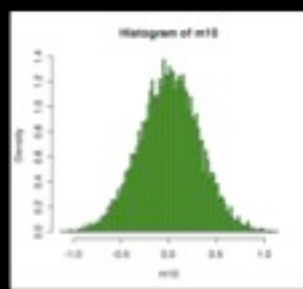
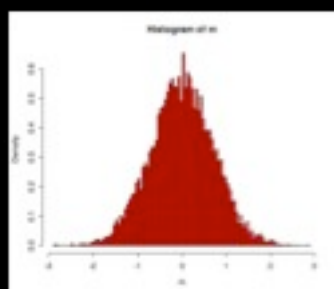


Uncertainty in the estimate of the mean

$$\bar{Y} = \frac{\sum_{i=1}^n Y_i}{n}$$

$$\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}, \quad SE_{\bar{Y}} = \frac{s}{\sqrt{n}}$$

$$\bar{Y} = \sum_{i=1}^n Y_i f(Y_i)$$



Only for large n (large samples) the sampling distribution of the mean is really a normal distribution with standard deviation SE_Y

For small n , small sample sizes, the sample mean distribution is related to a Student or t-distribution with $n-1$ degrees of freedom

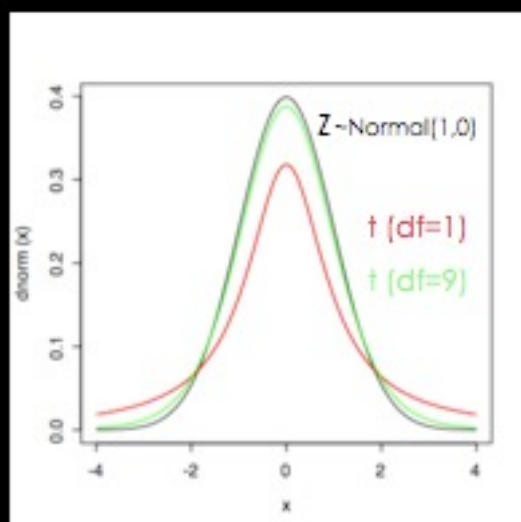
The difference between the sample mean and the true mean $\bar{Y} - \mu$

divided by the estimate of the standard error $SE_{\bar{Y}} = \frac{s}{\sqrt{n}}$

has a student's t-distribution with $n-1$ degrees of freedom

$$t = \frac{\bar{Y} - \mu}{SE_{\bar{Y}}}$$

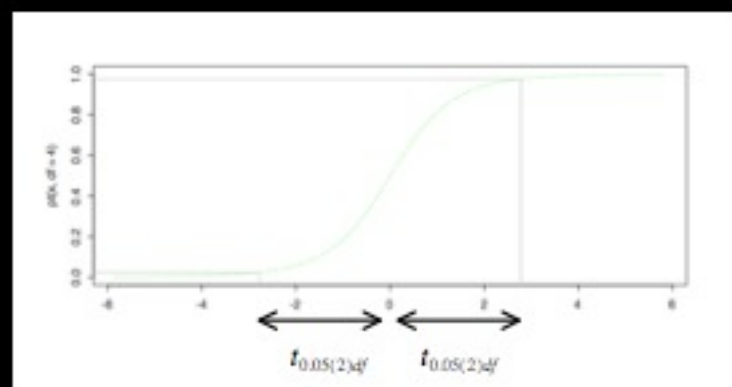
[Practical work: use random sampling in R to study the distribution of the quantity t obtained from the samples and the "known distribution" μ (used in generating data) for several values of sample size n , overlay a t-distribution and a normal, and also make a Kolmogorov-Smirnov test]



The interval of confidence of 95% is the interquantile range centred on the sample mean that includes 95% of the sampling distribution mass:

$$-t_{0.05(2)df} < \frac{\bar{Y} - \mu}{SE_{\bar{Y}}} < +t_{0.05(2)df}$$

$$\bar{Y} - t_{0.05(2)df} SE_{\bar{Y}} < \mu < \bar{Y} + t_{0.05(2)df} SE_{\bar{Y}}$$



$$-t_{0.05(2)df} < \frac{\bar{Y} - \mu}{SE_{\bar{Y}}} < +t_{0.05(2)df}$$

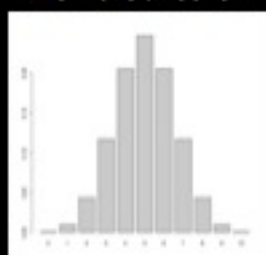
$$\bar{Y} - t_{0.05(2)df} SE_{\bar{Y}} < \mu < \bar{Y} + t_{0.05(2)df} SE_{\bar{Y}}$$



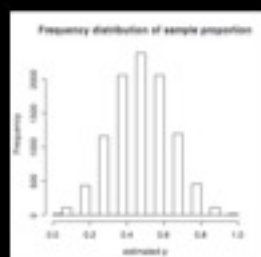
sampling

$$\hat{p} = \frac{X}{n}$$

Binomial distribution



trials=10
p=0.5



Uncertainty in the proportion estimate

$$\hat{p} = \frac{X}{n}$$

$$\text{Mean}(\hat{p}) = p$$

$$\sigma(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$$

The 95% confidence interval is calculated in an approximated way by :

$$p' = \frac{X + 2}{n + 4}$$

$$p' - Z \sqrt{\frac{p'(1-p')}{n+4}} < p < p' + Z \sqrt{\frac{p'(1-p')}{n+4}}$$

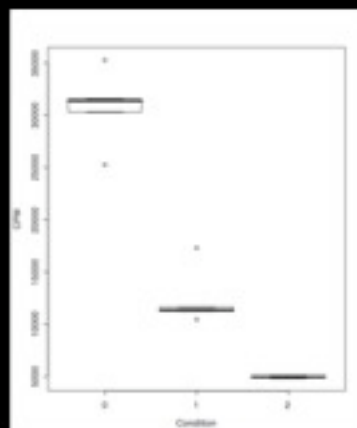
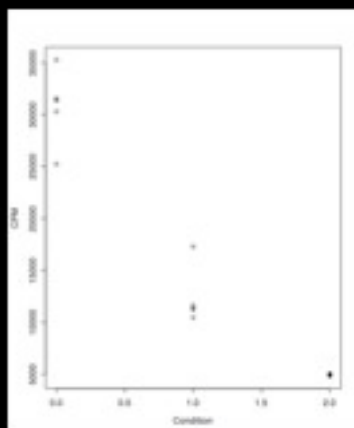
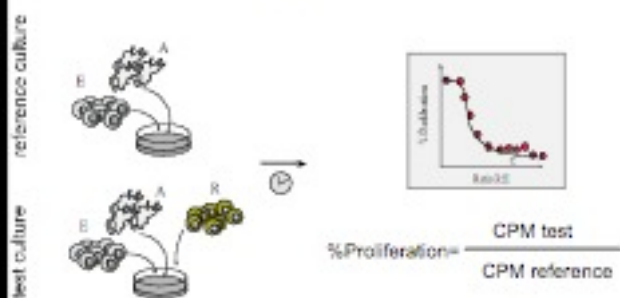
(Agresti-Coull method)

$$p' - Z \sqrt{\frac{p'(1-p')}{n+4}} < p < p' + Z \sqrt{\frac{p'(1-p')}{n+4}}$$

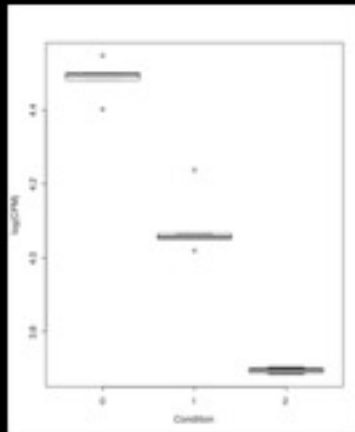
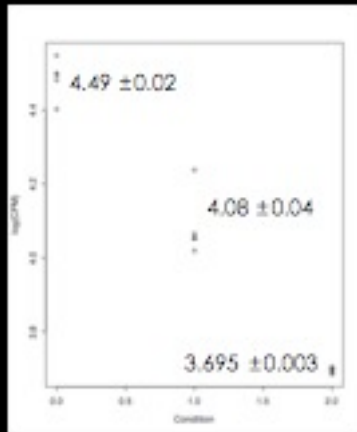
How does uncertainty propagate ?

An Example

Tolerance/Suppression *in vitro*



Anything funny about the graph ?



How would you **calculate** and **report** the ratio between the CPMs in condition 1 and 2 divided by the CPMs in condition 0 ?

General expression for uncertainty propagation

Consider a quantity R

$$R = R(x, y, \dots)$$

which is a function of a series of variables x, y, \dots

with uncertainties $\delta_x, \delta_y, \dots$

The uncertainty in R is estimated as :

$$\delta_R = \sqrt{\left(\frac{\partial R}{\partial x} \delta_x\right)^2 + \left(\frac{\partial R}{\partial y} \delta_y\right)^2 + \dots}$$

Reporting the results with uncertainty

It is important to report the results with the correct number of figures.

Begin by rounding the uncertainty in the result to one significant figure, then quote the value to the same decimal place.

$$g = 9.98 \pm 0.07 \text{ m/s}^2$$

not

$$g = 9.9878384 \pm 0.07239234 \text{ m/s}^2$$

$$g = 9.9878384 \pm 0.07 \text{ m/s}^2$$

$$g = 10 \pm 0.07239234 \text{ m/s}^2$$

