P

Statistics Applied to Economics

Degree in Economics

F.Tusell

Dpto. Economía Aplicada III (Estadística y Econometría)

Curso 2011–2012

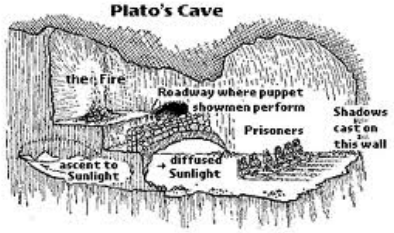
eman ta zabal zazu

Universidad del País Vasco

Euskal Herriko Unibertsitatea

Índice I

The man who started it all



- Plato, 428BC-348BC. Philosopher, disciple of Socrates, bulwark of extreme idealism.
- Ideas are the "true" reality, of which we see "shadows". Cave myth.
- Empirical experience and ideas live in different worlds.
- Western philosophy "...and endless footnote to the work of Plato".

Plato may not have been quite right, but...

- ...he raised some very important points.
- We think in terms of models.
- Geometry is a model of physical reality. Geometrical shapes are idealized real objects.
- Models are pretty much like platonic ideas.
- Probability theory supplies a model for random phenomena.

What do you think "probability" is the model of?

Relative frequency. (But alternative interpretations.)

We formulate problems in terms of models

- Do seat belts reduce mortality in traffic accidents?
- Relative frequencies of deaths within the seat belt users and non seat belt users keeps changing; they are "fluid".
- We expect them to stabilize around fixed, "solid" probabilities: *that's our model.*
- Now we can ask ourselves: is the probability of death among seat belt users smaller than in the other group?

Turning questions into statistical inference problems (I)

- Our models will usually be distributions, some of whose parameters are unknown.
- Our questions can usually be phrased in terms of values of those parameters.
  - What is the average mortality for seat belt users?
    $\Leftrightarrow$ What is p_{Users}? *(estimation problem)*
  - Do seat belts reduce mortality in traffic accidents?
    $\Leftrightarrow$ Is $p_{Users} < p_{NonUsers}$? *(hypothesis test problem)*
- Other problems not quite fitting in either category (e.g., serialization)
- If model is "good", answering questions about the model will enlighten us about the real world.

Is all this *that* new?

- No, it isn't.
- If you think for a moment, many previous examples were phrased in a manner suggesting inferential problems.

Can you think of some instances?

Problem regarding cancer incidence in a school was abnormally high.
Estimating the proportion of people who would vote for a candidate.

Turning questions into statistical inference problems (II)

- Notions such as *independence* relate to models, not to data.
- For instance, if we have

		Hair color	
Eye color		**Blonde** ($p_{\cdot 1}$)	**Brown** ($p_{\cdot 2}$)
Blue	($p_{1\cdot}$)	430 p_{11}	180 p_{12}
Brown	($p_{2\cdot}$)	123 p_{21}	108 p_{22}

  "independence" means $p_{i\cdot} \times p_{\cdot j} = p_{ij}$ for all i, j.
- This has an intuitive meaning in terms of what we expect to see in the data *but is stated in terms of the model.*

We are platonic after all

- Parameters are dancers, we can only glance at shadows.
- Parameters pertain to the *population*.
- The "shadows" we observe are empirical evidence available: samples.
- A *sample* is a collection of elements generated by the population, usually through random sampling.
- From what we observe in the sample, we infer properties of the population, the model.

Point and interval estimation

- Sometimes we are content with a value "close" to the (unattainable) value of the true parameter. Then we have a problem of *point estimation*.
- Sometimes we want an interval that most of the time (with given *confidence*) will cover the true value of the parameter. This is an *interval estimation* problem.
- Common sense will sometimes guide us in choosing an estimator...
- ...but a more principled approach is desirable.

Samples and statistics

- A sample is a randomly chosen set from the population.
- Capital letters denote random values the members of the sample can yield: $\vec{X} = (X_1, X_2, \ldots, X_n)$.
- Lower case letters, $\vec{x} = (x_1, x_2, \ldots, x_n)$, denote the actual, fixed values obtained in a concrete sample taken.
- A *statistic* is a function of the sample: $S = S(\vec{X})$ or $s = s(\vec{x})$. Before the sample is taken, it is a random variable; after the sample is taken, it becomes a number (or vector of numbers)

Estimators and estimates

- An statistic designed to be "close" to the value of a parameter is an *estimator*.
- The value it takes is an *estimate*.
- Example: $\overline{X} = (X_1 + X_2 + \ldots + X_n)/n$ is a (usually good) estimator of the mean of a distribution. *Given* a concrete sample $x_1, \ldots, x_n$, $\overline{x} = 5.8$ is an estimate.
- With different samples, the same estimator will produce different estimates each time.

Methods for choosing point estimators

- What we choose as an estimator depends on our goals and *loss function* (= how much cost errors).
- For didactical reasons, we will look first at some recipes, then study their properties.
- Two important estimators:
  - Method of moments.
  - Method of maximum likelihood.
- Least squares method is a particular case of the method of moments.

Method of moments (I)

- Equate moments of the distribution (usually function of parameters) to sample moments.
- Solve for the parameters.
- Need as many equations as there are parameters.
- Example: $\mathcal{P}(\lambda)$, sample of n observations.

$$m = \lambda = \frac{X_1 + X_2 + \ldots + X_n}{n} = \overline{X}$$

- Could also use,

$$\lambda + \lambda^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

Usually lower order moments best (and simpler).

Method of moments (II)

- Example: estimate m and σ^2 of $N(m, \sigma^2)$.
- Now we need two equations:

$$\begin{aligned} m &= \frac{X_1 + X_2 + \ldots + X_n}{n} \\ \sigma^2 + m^2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 \end{aligned}$$

from which

$$\begin{aligned} \hat{m} &= \overline{X} \\ \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \end{aligned}$$

Method of moments (III)

- Example: estimate θ in a $U(0, \theta)$.
- The mean is $m = \theta/2$. Therefore,

$$\begin{aligned} \frac{\theta}{2} &= \frac{X_1 + X_2 + \ldots + X_n}{n} \\ \hat{\theta} &= 2\overline{X} \end{aligned}$$

- Not a particularly good estimator, as we will see.

Method of moments (IV)

- Example: estimate λ in a $\exp(\lambda)$.
- The mean is $m = 1/\lambda$. Therefore,

$$\begin{aligned} \frac{1}{\lambda} &= \frac{X_1 + X_2 + \ldots + X_n}{n} \\ \hat{\lambda} &= \frac{1}{\overline{X}} \end{aligned}$$

Method of moments (V)

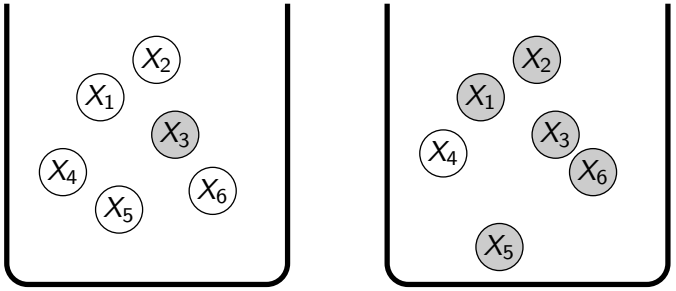
- Example: estimate a and r in a $\gamma(a, r)$.
- Remember that $m = r/a$ and $\sigma^2 = r/a^2$. Therefore,

$$\begin{aligned} \frac{r}{a} &= \frac{X_1 + X_2 + \ldots + X_n}{n} \\ \frac{r}{a^2} + \frac{r^2}{a^2} &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 \end{aligned}$$

- We can solve for a and r to obtain:

$$\begin{aligned} \hat{a} &= r/\overline{X} \\ \hat{r} &= \frac{\overline{X}^2}{n^{-1}\sum_{i=1}^{n} X_i^2 - \overline{X}^{-2}} \end{aligned}$$

Method of maximum likelihood (I)



- We are allowed to sample one of the two urns, but we are not told which one it is. We pick one ball which happens to be grey

What would be your guess?

Right urn, as it can generate grey balls more easily.

Method of maximum likelihood (II)

- Logic underlying previous choice is maximum likelihood logic.

  When confronted to two or more states of nature which may have produced a given evidence, we choose the one(s) with optimal capability to generate such evidence.

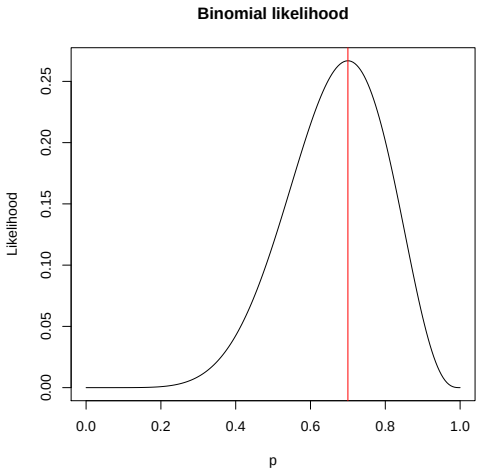
- Both urns could generate a grey ball, but the second one does so much more easily.
- Why assume that something "strange" has happened if we can see the evidence as the outcome of something "common"?

Method of maximum likelihood (III)

- If joint density of a given sample is $f(\vec{x};\theta)$, $\theta \in \Theta$, we call *likelihood function* $f(\vec{x};\theta)$ **seen as a funcion of θ** for given $\vec{x}$.
- To maximize the likelihood is tantamount to choosing the θ which gives maximum density to the observed sample.
- Maximizing θ is *maximum likelihood estimate*, $\hat{\theta}_{MLE}$.
- $f(\vec{x};\theta)$ and $\log f(\vec{x};\theta)$ both achieve their maximum for the same value of θ. Usually easier to maximize the second.

Likelihood example: binomial distribution (I)

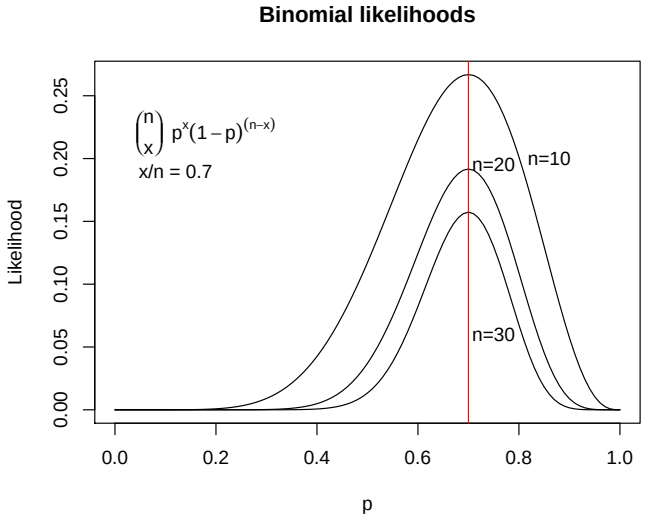
```
> n <- 10 ; x <- 7
> binom <- function(p) {
  l <- choose(n,x) * p^x *
       (1-p)^(n-x)
  return(l)
 }
> curve(binom,from=0.00,
   to=1,n=200,
   ylab="Likelihood",
   xlab="p",
   main="Binomial likelihood")
> abline(v=x/n,col="red")
```



What would happen with different values of x and n?

Maximum always at x/n, sharper peak as n grows.

Likelihood example: binomial (II)



Are the likelihood functions like density functions?

Clearly not; areas below change, not always 1.

Example: MLE of p with $x_1, x_2, \ldots, x_n$ i.i.d. $b(p)$

$$
\begin{aligned}
f(\vec{x}; p) &= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \\
\log f(\vec{x}; p) &= \sum_{i=1}^{n} x_i \log(p) + \left(n - \sum_{i=1}^{n} x_i\right) \log(1-p) \\
\frac{\partial \log f(\vec{x}; p)}{\partial p} &= \frac{\sum_{i=1}^{n} x_i}{p} - \frac{n - \sum_{i=1}^{n} x_i}{1-p} = 0 \\
\hat{p}_{MLE} &= \frac{\sum_{i=1}^{n} x_i}{n}
\end{aligned}
$$

Do we need to know all $x_1, \ldots, x_n$ in order to compute the MLE?

Only $\sum_{i=1}^{n} x_i$ is necessary to compute the MLE.

Example: MLE of λ with $x_1, x_2, \ldots, x_n$ i.i.d. $\mathcal{P}(\lambda)$

$$
\begin{aligned}
f(\vec{x}; \lambda) &= \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \\
\log f(\vec{x}; \lambda) &= -n\lambda + \sum_{i=1}^{n} x_i \log(\lambda) - \sum_{i=1}^{n} \log(x_i!) \\
\frac{\partial \log f(\vec{x}; \lambda)}{\partial \lambda} &= -n + \frac{\sum_{i=1}^{n} x_i}{\lambda} = 0 \\
\hat{\lambda}_{MLE} &= \frac{\sum_{i=1}^{n} x_i}{n} = \overline{x}
\end{aligned}
$$

Do we need to know all $x_1, \ldots, x_n$ in order to compute the MLE?

Only $\sum_{i=1}^{n} x_i$ is necessary to compute the MLE.

Example: MLE of m, σ^2 with $x_1, \ldots, x_n$ i.i.d. $N(m, \sigma^2)$

$$
\begin{aligned}
f(\vec{x}; m, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_i-m)^2/2\sigma^2} \\
\log f(\vec{x}; m, \sigma^2) &= -\frac{n}{2}\log(\sigma^2) - \frac{n}{2}\log(2\pi) - \sum_{i=1}^{n} \frac{(x_i-m)^2}{2\sigma^2} \\
\frac{\partial \log f(\vec{x}; m, \sigma^2)}{\partial m} &= \frac{\sum_{i=1}^{n}(x_i - m)}{\sigma^2} = 0 \\
\frac{\partial \log f(\vec{x}; m, \sigma^2)}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^{n}(x_i-m)^2}{2\sigma^4} = 0
\end{aligned}
$$

whence

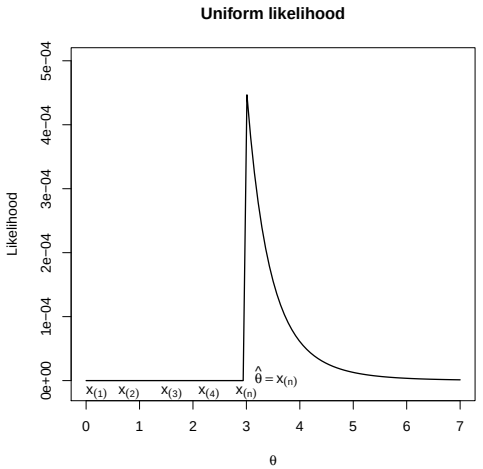
$$
\hat{m}_{MLE} = \frac{\sum_{i=1}^{n} x_i}{n} = \overline{x} \qquad \hat{\sigma^2}_{MLE} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}
$$

Do we need to know all $x_1, \ldots, x_n$ in order to compute the MLE?

Only $\sum_{i=1}^{n} x_i$ and $\sum_{i=1}^{n}(x_i - \overline{x})^2$ necessary.

Example: MLE of θ with $X_1, \ldots, X_n$ i.i.d. $U(0, \theta)$

- Likelihood function not as usual.
- Not differentiable.
- Pick maximum by inspection.
- $x_{(1)}, \ldots, x_{(n)}$ called "order statistics".

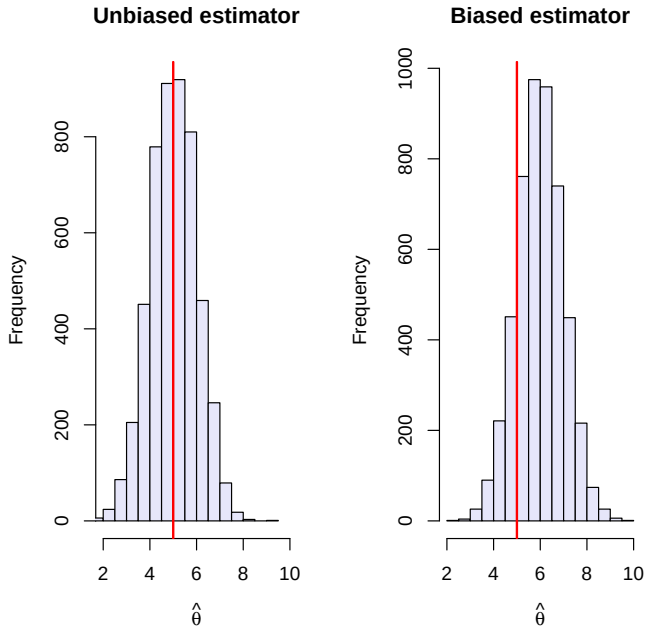


Do we need to know all $x_1, \ldots, x_n$ in order to compute the MLE?

Only one ($x_{(n)}$, the largest) is necessary!

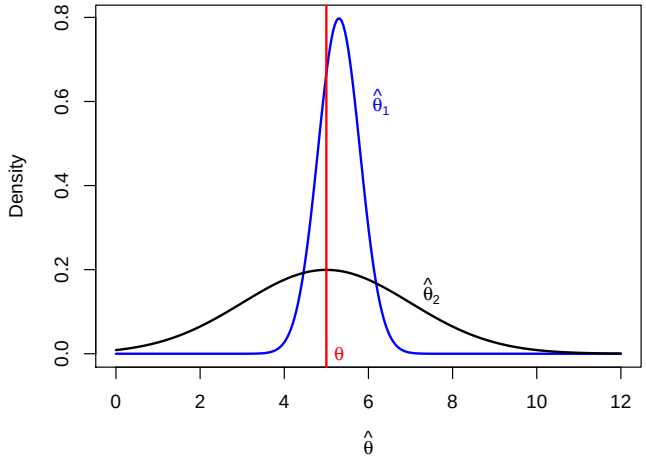
Unbiasedness (I)

- $\hat{\theta}$ unbiased for θ means that $E[\hat{\theta}] = \theta$.



Unbiasedness (II)

- In principle a desirable property...
- ...but sometimes we may prefer a biased estimator.



Which one would you prefer?

If squared error loss, we might prefer $\hat{\theta}_1$, even if biased.

Unbiasedness (III)

- In a $\mathcal{P}(\lambda)$, $\hat{\lambda} = \overline{X}$ is unbiased.
- In a $N(m, \sigma^2)$, $\hat{m} = \overline{X}$ is unbiased.
- In a $N(m, \sigma^2)$, $\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ is biased.
- $\hat{\sigma}^2_* = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ is unbiased; $\hat{\sigma}^2_{MLE}$ is *asymptotically unbiased*.
- In a $\exp(\lambda)$, $\hat{\lambda} = \frac{1}{\overline{X}}$ is biased;

$$E[\hat{\lambda}] = E\left[\frac{1}{\overline{X}}\right] \neq \frac{1}{E[\overline{X}]}$$

Unbiasedness (IV)

- Among two unbiased estimators, we would prefer the one with smaller variance.
- If any of both are biased, we have to take this into account.
- One way is to select the one with minimum mean squared error (MSE).

$$\begin{aligned} \mathrm{MSE}(\hat{c}) &= E[(\hat{c} - c)^2] \\ &= E[(\hat{c} - E(\hat{c}) + E(\hat{c}) - c)^2] \\ &= E[\hat{c} - E(\hat{c})]^2 + E[E(\hat{c}) - c)]^2 \\ &= \sigma^2_{\hat{c}} + (\mathrm{bias}(\hat{c}))^2 \end{aligned}$$

What implicit assumption does MSE about gravity of estimation error?

"Twice as large, four times as bad." Arbitrary, mathematically convenient.