On some statistical concepts

Laurie Davies

Fakultät Mathematik, Universität Duisburg-Essen

Monday, 17th October 2011

Addendum

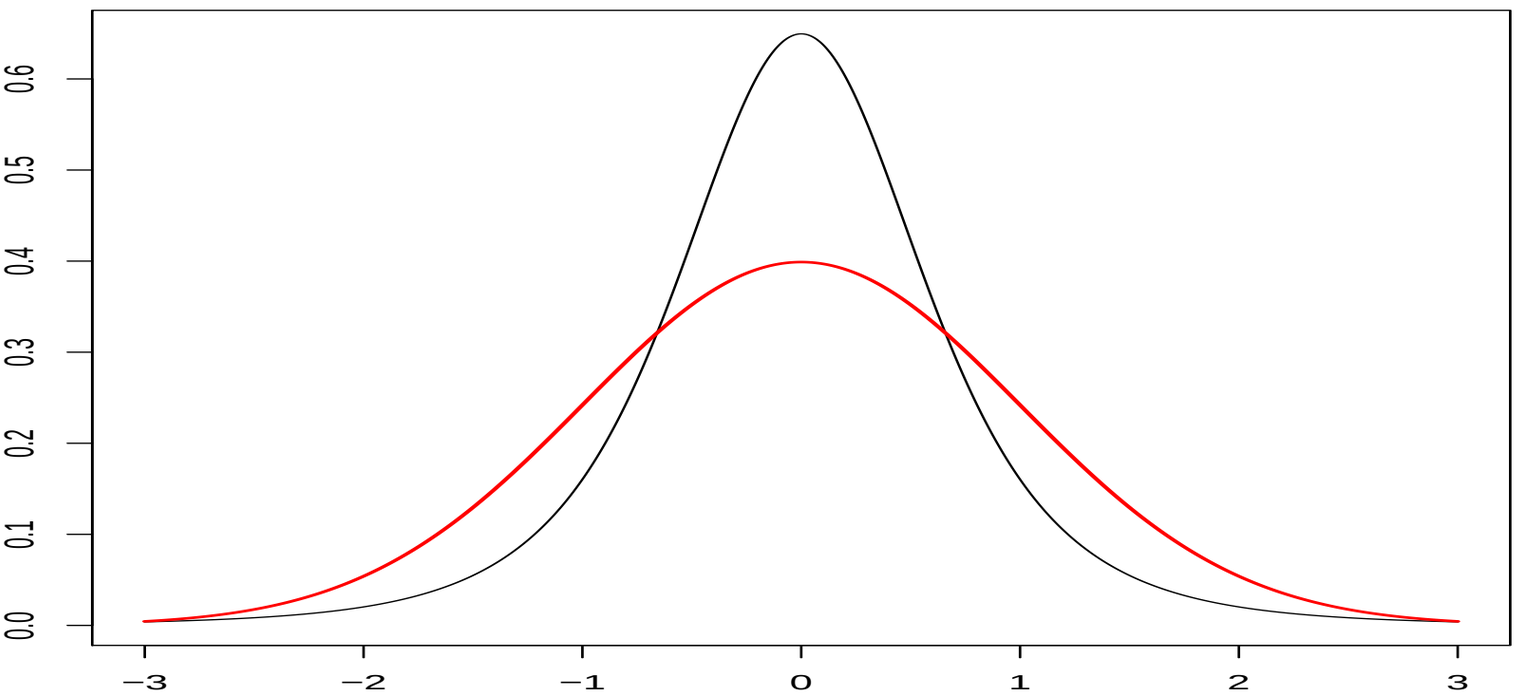
Notation

Lower case letters $\boldsymbol{x}_n$ denote real data.

Upper case letters $\boldsymbol{X}_n$ denote data generated under a model.

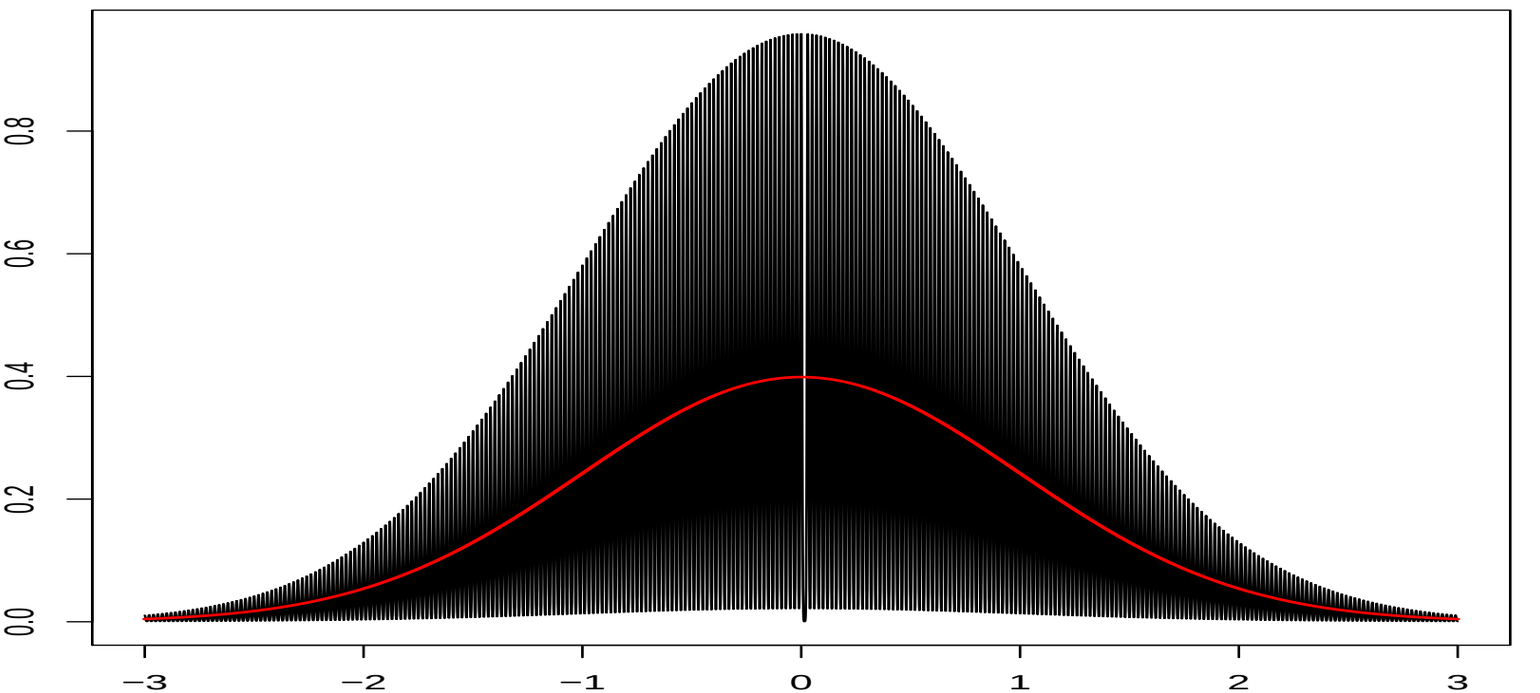
Regularization: Addendum

t3 and Gaussian densities with variance 1.



Regularization: Addendum

comb_{100} density and the $N(0,1)$ density.



Topologies

comb_{100} and the $N(0,1)$ distribution functions

Kolmogorov metric.

$$d_{ko}(F,G) = \max_x |F(x) - G(x)| \leq 1$$

$$d_{ko}(\text{comb}_{100}, N(0,1)) = 0.0035$$

Total variation metric

$$d_{tv}(F,G) = \frac{1}{2}\int |f(x) - g(x)|\, dx \leq 1$$

$$d_{tv}(\text{comb}_{100}, N(0,1)) = 0.375$$

Topologies

Two probability measures P and Q on $\mathbb{R}$.

$$d_{ko}(P,Q) = \sup_{C \in \mathcal{C}_1} |P(C) - Q(C)|, \mathcal{C}_1 = \{(-\infty, x] :\in \mathbb{R}\}$$

$$d_{tv}(P,Q) = \sup_{C \in \mathcal{C}_2} |P(C) - Q(C)|, \mathcal{C}_2 = \text{Borel subsets of } \mathbb{R}$$

The topology generated by d_{ko} is strictly weaker than that generated by d_{tv}

$$\mathcal{O}_{ko} \subset \mathcal{O}_{tv}$$

Topologies

If $\mathcal{C}$ is a Vapnik-Cervonenkis class then

$$d_{\mathcal{C}}(P, Q) = \sup_{C \in \mathcal{C}} |P(C) - Q(C)|$$

generates a weak topology.

Weak metrics allow direct comparisons between empirical distributions and models

$$d_{\mathcal{C}}(\mathbb{P}_n, P) = \mathsf{O}(1/\sqrt{n}),\ d_{tv}(\mathbb{P}_n, P) = 1.$$

In many cases $\mathcal{C}$ is closed under affine transformations A, or can be taken to be so. In such cases

$$d_{\mathcal{C}}(P^A, Q^A) = d_{\mathcal{C}}(P, Q)$$

Distance independent of unit of measurement.

Topologies

The following operate in a weak topology:

EDA, histograms, distribution functions, q-q-plots, scatter plots, outliers, goodness-of-fit tests, ...

Much of (formal) inference operates in a strong density based topology

Likelihood

Topologies

Distribution function F with density f

$$F(x) = \int_{-\infty}^{x} f(u)\, du$$

Put

$$\mathcal{F} = \{F : \text{absoluely continuous distribution function}\}$$

$(\mathcal{F}, d_{ko})$ is a metric space

$$\mathcal{D} = \{f : f \geq 0, \int f(u)\, du = 1\}$$

$(\mathcal{D}, \|\cdot\|_1)$ is a metric space

Topologies

Differential operator D

$$F(x) = \int_{-\infty}^{x} f(u)\, du, \quad D(F) = f.$$

$$D : (\mathcal{F}, d_{ko}) \to (\mathcal{D}, \|\cdot\|_1)$$

D is pathologically discontinuous.

Likelihood

(a) Likelihood reduces the measure of fit between a data set $\boldsymbol{x}_n$ and a statistical model P_θ to a single number irrespective of the complexity of both.

(b) Likelihood is dimensionless and imparts no information about closeness.

(c) Likelihood is blind. Given the data and the model or models, it is not possible to deduce from the values of the likelihood whether the models are close to the data or hopelessly wrong.

(d) Likelihood does not order models with respect to their fit to the data.

Likelihood

(e) Likelihood based procedures for model choice (AIC, BIC, MDL, Bayes) give no reason for being satisfied or dissatisfied with the models on offer.

(f) Likelihood does not contain all the relevant information in the data $\boldsymbol{x}_n$ about the values of the parameter θ.

(g) Given the model, the sample cannot be reduced to the sufficient statistics without loss of information.

(h) Likelihood is based on the differential operator and is consequently pathologically discontinuous.

(i) Likelihood is evanescent: a slight perturbation of the model P_θ to a model P_θ^* can cause it to vanish.

Likelihood

On the positive side:

(j) Likelihood delimits the possible.

Solomonoff-Kolmogorov complexity

A sequence of 0s and 1s

$$1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 1, \ldots, 0, 1$$

Write the shortest computer programme which reproduces the sequence
Exploit any regularities in the sequence

$$0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, \ldots, 1, 0, 1$$

```
      do 10 i=1,n/2
      write(*,*) 0,1
10    continue
```

Length of the programme is a measure of complexity of the sequence

Solomonoff-Kolmogorov complexity

Complex sequences are those for which the length of the shortest programme is the length of the sequence

Random sequences are complex
No better than simply writing the sequence

write(*,*) 0
write(*,*) 1
write(*,*) 0
etc

Martin-Löf: complex sequences pass tests of randomness

Minimum Description Length MDL

Encode the data using a prefix code.

Objects 1,2,3,... with code lengths $n(1), n(2), n(3)...$
Kraft's inequality

$$\sum_j 2^{-n(j)} \leq 1$$

$= 1$ a probability measure

A prefix code corresponds to a probability measure (model) and vice versa.

Minimum Description Length MDL

Idea behind MDL

Several models for the data

Encode the data with each model

The model with the shortest code length is the best of the models.

Minimum Description Length MDL

Data $\boldsymbol{x}_n$ of finite precision

Use model P with $P(\{\boldsymbol{x}_n\}) = 1$.

The corresponding prefix code is 1 of length 1

This is not liked by proponents of MDL

To avoid this it is now required to decode the code to recover the data.

No problem, send the decoder P and then the code 1.

Minimum Description Length MDL

To avoid it the models must be chosen before seeing the data.

This argument does not apply to Solomonoff-Kolmogorov complexity

How does one encode the model?

Typically by encoding parameters, (μ, σ) for Gaussian models

This simply ignores the complexity of the Gaussian model.

Nothing alters if the Gaussian model is replaced by the much more complex Gaussian comb

Minimum Description Length MDL

The weakness of MDL is the lack of any prescription for encoding models