Modern Astrostatistics, lecture 1 : Getting used to randomness

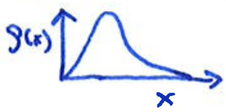
Monday, 4th Feb 2019

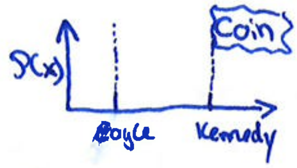
- lecture time: 13:30 – 14:15
  14:30 – 15:15

- https://github.com/elena sellentin

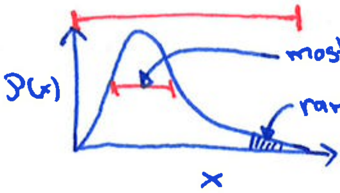
- A random variable, x, is drawn from a probability distribution $\mathcal{P}(x)$:

$$x \sim \mathcal{P}(x)$$





- The „width" of the probability distribution determines the uncertainty of x.

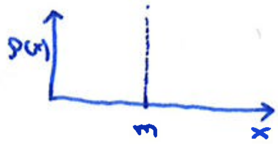


most of the time something like this happens

rarely something like this happens

- A „classical" variable is non-random, and drawn from Dirac's δ-distribution:

$$x \sim \delta_D(x-m)$$

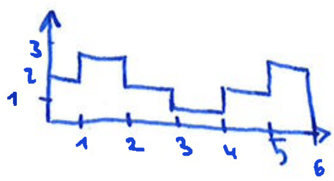


- A „draw" from a probability distribution is also called „a random realization". Information about random variables is encoded in their probability density function: if you don't know it, you have a scientific problem.

Examples:

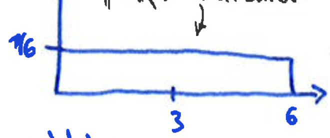
① x be random, and a single draw of x yields x=4. What do you know about x?

② many repeated draws of x yield



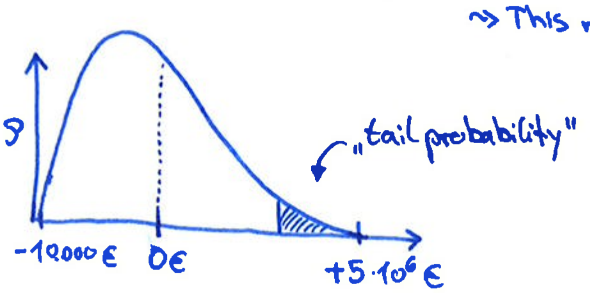
$\xrightarrow{N\to\infty}$

this is called a „Uniform distribution"



⇝ This might be a fair, six-sided die

③ Suspicious carsales(wo)man:



$\Rightarrow$ under all circumstances, try to learn as much as possible about $P(x)$.

- a „univariate" random variable is a random <u>scalar:</u> $x \sim P(x)$
- a „multivariate" random variable is a random <u>vector</u> : $\vec{x} \sim P(\vec{x})$

$$\vec{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \quad , \quad \vec{x} \sim P(\vec{x}) \Rightarrow \begin{array}{l} x_1 \sim P_1(x_1 \; ? \; ? \; ?) \\ x_2 \sim P_2(x_2 \; ? \; ? \; ?) \\ x_3 \sim P_3(x_3 \; \underbrace{? \; ? \; ?}) \end{array}$$

what do these distributions all depend on?

- There are two possibilities:

  1) x_1 has nothing to do with x_2 and x_3:

  $$x_1 \sim P_1(x_1 \; ???) = P_1(x_1) \quad \text{„is independent of } x_2 \text{ and } x_3\text{"}$$

  2) x_1 depends on x_2, x_3:

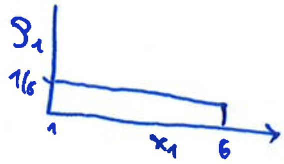
  $$x_1 \sim P_1(x_1 | x_2, x_3)$$

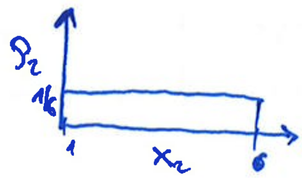
  ↑ „given"

  $\rightsquigarrow$ $P_1(x_1 | x_2, x_3)$ is a „conditional" distribution

  $\rightsquigarrow$ $P(x_1, x_2, x_3)$ is a <u>joint</u> distribution

- Examples: dice





$$P(x_1, x_2) = P(x_1) P(x_2)$$ $\rightsquigarrow$ independent probabilities multiply

$S = x_1 + x_2 \quad P(S | x_1, x_2) \neq P(S) P(x_1) P(x_2)$ because S <u>depends</u> on x_1 and x_2, since it is a <u>function</u> of it

• Rules

- random variables follow probability distributions, $x \sim p(x)$, $\vec{x} \sim \mathcal{P}(\vec{x})$, etc.
- any function of a random variable is again a random variable

$$x \text{ random} \Rightarrow y = f(x) \text{ also random}$$ [Exception: if f is an „expectation operator"]

The probability distribution then transforms as:

$$\left.\begin{array}{l} \cdot\, \mathcal{P}(\vec{x}) \\ \cdot\, f(\vec{x}) = \vec{y} \end{array}\right\} \quad \mathcal{P}(\vec{y}) = \mathcal{P}(\vec{x} = \tilde{\vec{y}}(\vec{x}))\,|\mathcal{J}|$$

with Jacobi-determinant $|\mathcal{J}| = \left|\det\left(\frac{\partial \vec{x}}{\partial \vec{y}}\right)\right|$

and inverse function $\tilde{\vec{y}}$.

Why is this so?

⇝ Probability distributions can be normalized: $\int \mathcal{P}(x)\,dx = 1$

$$\rightsquigarrow |\mathcal{P}(x)\,dx| = |\mathcal{P}(y)\,dy| \rightsquigarrow \mathcal{P}(y) = \mathcal{P}(x)\left|\frac{dx}{dy}\right| = \mathcal{P}(\tilde{y}(x))\left|\frac{dx}{dy}\right|$$

to eliminate x fully, the inverse function $\tilde{y}$ is needed

- In the discrete case $\sum_i P_i = 1$, especially $P(A) + P(\bar{A}) = 1$

„not A"

- Consistency of joint and conditional distributions:

$$\mathcal{P}(A,B) = \mathcal{P}(A|B)\,\mathcal{P}(B) = \mathcal{P}(B|A)\,\mathcal{P}(A)$$

$$\mathcal{P}(A|B) = \mathcal{P}(B|A)\,\frac{\mathcal{P}(A)}{\mathcal{P}(B)}$$

Bayes' Theorem: the inversion of conditional statements in statistics.

- Examples

1) Justin Bieber ain't human: My hypothesis.

Proof:
- h: „you are human"
- j: „you are Justin"

- the probability that you are human is $P(h) = 1$.
- there are $\approx 8 \cdot 10^{+9}$ people in the world, hence the probability that you are Justin is $P(j \mid h) = \frac{1}{8 \cdot 10^9} = 1.25 \cdot 10^{-10}$

given you are human

- the probability that Justin is human is:

$P(h \mid j)$

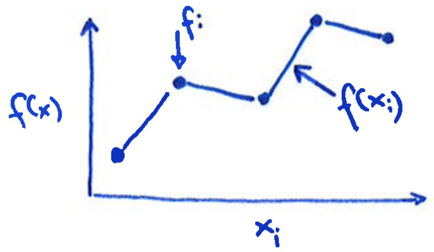
„given you are Justin"...
..."what is the probability that"...
..."you are human."

- We have $P(h|j) = P(j|h)$
$= 1.25 \cdot 10^{-10}$: ridiculously small „p-value" hence I reject the hypothesis that Justin is human.

[Addendum:
$P(h) = 1$
$P(j) = 1.25 \cdot 10^{-10}$ as well

$\left.\right\}$ $P(h|j) = P(j|h) \frac{P(h)}{P(j)} = \frac{1.25 \cdot 10^{-10} \cdot 1}{1.25 \cdot 10^{-10}} = 1$]

2) Cross-validation:



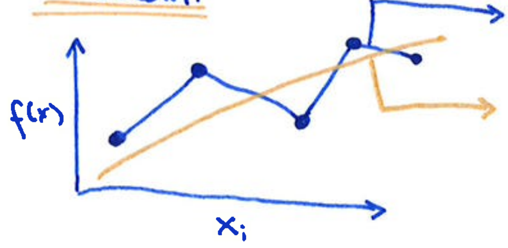
Fitting: define ourselves an error that we want to minimize.

$$\varepsilon_{tot} = \frac{1}{N} \sum_i \| f_i - f(x_i) \|_2^2 = \frac{1}{N} \sum_i (f_i - f(x_i))^2$$

„L2-norm"

We all know this is wrong, but how do we prove it?

Problem:

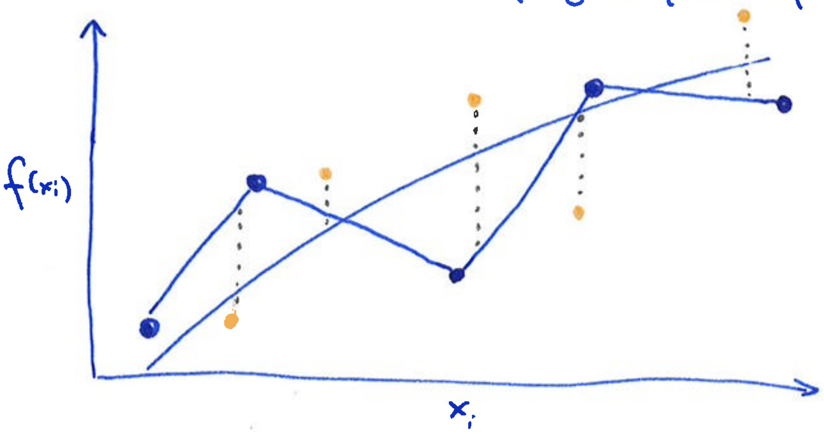


$\varepsilon_{tot} = 0$
$\varepsilon_{tot} \gg 0$

Seems like connect-the-dots it better?

• Solution: The fitter forgot that the drawn random numbers don't have a meaning. Their probability distribution is the quantity of importance.

Imagine I was mean and only gave you half of the points:

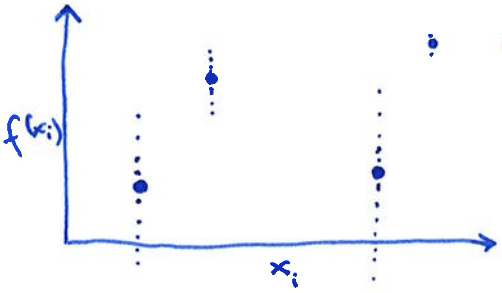


$\Rightarrow$ for the new points $\quad \varepsilon_{smooth} \ll \varepsilon_{zigzag}$

> $\rightsquigarrow$ Your fit does not only need to explain the data you have. It must also explain data which you don't have, but which are just as valid a realizations as those at hand.
>
> $g(x)$ is the important quantity.

• **The correct error estimate:** we invented $\varepsilon_{est} = \frac{1}{N}\sum_i (f_i - \underbrace{f(x_i)}_{\mu_i})^2 = \frac{1}{N}(\vec{f} - \vec{\mu})^T \mathbb{1} (\vec{f} - \vec{\mu})$

What if the individual points scatter differently?

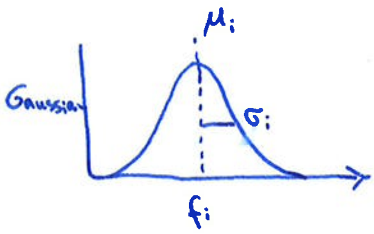


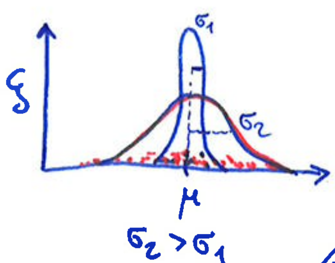
• Let each $f_i \sim \mathcal{G}(\mu_i, \sigma_i^2)$

$$= \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{1}{2} \frac{(f_i - \mu_i)^2}{\sigma_i^2}\right)$$

• then $g(f_1)\, g(f_2) \dots g(f_N) = \prod_i g_i(f_i)$

$$= \prod_i \left[\frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{1}{2} \frac{(f_i - \mu_i)^2}{\sigma_i^2}\right)\right]$$





- take the logarithm:

$$\log\left[\prod_i \rho_i(f_i)\right] \propto \sum_i \frac{(f_i-\mu_i)^2}{\sigma_i^2} = (\vec{f}-\vec{\mu})^T \underbrace{\begin{pmatrix} 1/\sigma_1^2 & & 0 \\ & 1/\sigma_2^2 & \\ 0 & & \ddots & 1/\sigma_5^2 \end{pmatrix}}_{\text{this matrix replaces } \mathbb{1} \text{ from before}} (\vec{f}-\vec{\mu})$$

$$C^{-1} = \begin{pmatrix} 1/\sigma_1^2 & & \\ & \ddots & \\ & & 1/\sigma_5^2 \end{pmatrix}$$

is called „inverse covariance matrix"; it takes the role of a <u>metric</u> in $\chi^2 = (\vec{f}-\vec{\mu})^T C^{-1} (\vec{f}-\vec{\mu})$.

⇝ This metric measures compatibility between data ($\downarrow \vec{f}$) and theory ($\downarrow \vec{\mu}$).

- $\vec{\mu}$ is called „mean", its univariate equivalent is μ.
- C is called „covariance matrix", its univariate equivalent is σ^2, the „variance".
- We have

$$\vec{\mu} = \int \vec{x}\, \rho(\vec{x})\, dx \quad \text{or} \quad \mu = \int x\, \rho(x)\, dx$$

$$C = \int (\vec{x}-\vec{\mu})(\vec{x}-\vec{\mu})^T \rho(\vec{x})\, dx \quad \text{or} \quad \sigma^2 = \int (x-\mu)^2 \rho(x)\, dx$$

These are „expectation values"
⇝ they are <u>not</u> random

the diagonal elements of C are „variances" σ^2, and the offdiagonals are „<u>co</u>variances"

↓

co: latin for „with": covariance: sth that varies <u>with</u> something else

$$\mathrm{Cov}(x_1, x_2) = \int (x_1-\mu_1)(x_2-\mu_2)\, \rho(\vec{x})\, dx$$

The „correlation matrix" has elements

$$r_{ij} = \frac{\mathrm{Cov}(x_i, x_j)}{\sigma_i \sigma_j}.$$

Mandatory exercises:

„BasicsTutorial.pdf": 1, 3, 4, 5, 8 and 8.2 (χ^2-test)

The other exercises are for you to see how it gets astronomically relevant, but they are not needed for the exam (too long.)

The tutors are allowed to give hints starting on Monday 11th of Feb.