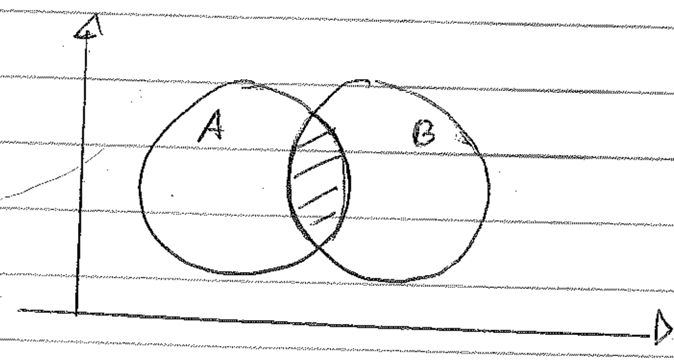
Charles MARGOSSIAN

March 12th 2019
PHC 506 Biometry

Probability and Bayes 1
(continued ...)

Last time you discussed conditional probability.
Consider two events A and B



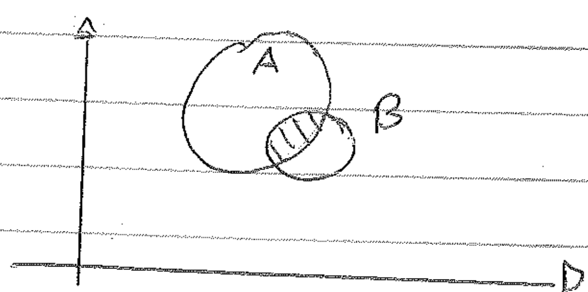
Recall

$$P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)}$$

Note:

$$P(A \mid B) \neq P(B \mid A)$$

E.g



The difference arises because $P(B) < P(A)$

Theorem (Bayes' rule)

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

Proof

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

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

$$= P(A|B) \quad \square$$

Eg Mammogram

The mammogram is a test for breast cancer.
It has ~~an~~ known ~~~~ false positive and negative rates.

Question: given a test that comes back positive, what is the probability of having breast cancer?

Let $A = \{\text{the patient has cancer}\}$
and $B = \{\text{the test is positive}\}$

and $\bar{A}, \bar{B}$ their respective complements.

We know
False neg.: $P(\bar{B} \mid A) = 0.2$

False pos.: $P(B \mid \bar{A}) = 0.1$

From Bayes' rule

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

$$P(B \mid A) = 1 - P(\bar{B} \mid A) = 0.8$$

For $P(A)$, we know the percentage of people with breast cancer, namely

$$P(A) = 0.0004$$

Next $P(B) = P(B \mid A) P(A) + P(B \mid \bar{A}) P(\bar{A})$

$$= (1 - P(\bar{B} \mid A)) P(A) + P(B \mid \bar{A}) (1 - P(A))$$

$$= 0.8 \times 0.0004 + 0.1 \times 0.9996$$

Putting this all together:

$$P(A \mid B) \approx 0.3\ \%$$

Context matters, one can't simply consider the false positive and negative rates

<u>Independence</u>: two events are independent if

$$P(A, B) = P(A)P(B)$$

→ Denote independence with "$\perp\!\!\!\perp$".

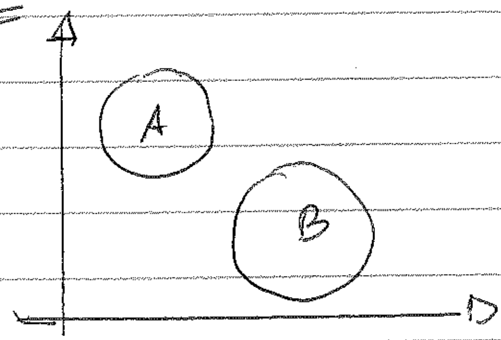
<u>Lemma</u>

If $A \perp\!\!\!\perp B$, then $P(A \mid B) = P(A)$

<u>e.g</u>

If I flip a coin two times, are the two flips independent?

<u>e.g 2</u>



Are A and B independent?

Charles MARGOSSIAN

March 12th 2019
PHC 506 Biometry

Probability and Bayes II

1/ Probabilistic models

What is a model?
There are several perspectives we can adopt:
- approximation of reality
- predictive model
- ...

All these are data generating processes.

In theory, they can be deterministic, but when analyzing data, it is useful to make their output random.

The randomness accounts for:
- noise in our measurements
- variations due to unknown factors
- ...

Often times, the model ~~contains~~ has **parameters** and allows **inputs**.

Formally, the model $\mathcal{M}$ is a map:

$$\mathcal{M} : (\theta, x) \longmapsto Y$$

Since Y is random, $\mathcal{M}$ is characterized by a distribution

$$Y \sim P_\theta(\cdot \mid x)$$

E.g 1 Ball in free fall

Suppose our data is the velocity of the ball at different time points. Physics tells us its acceleration is constant.

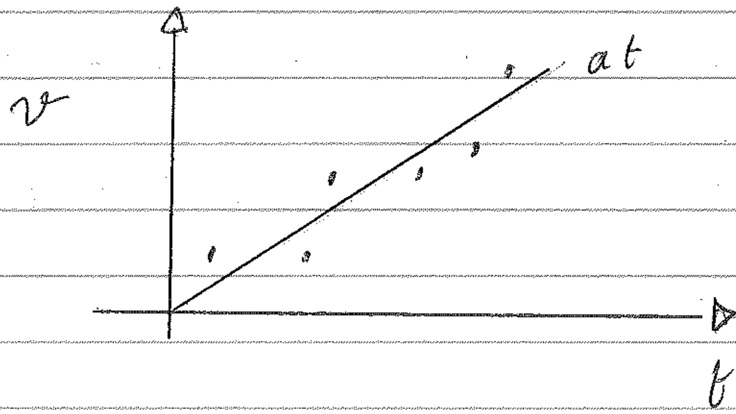
Thus $v(t) = a\,t$

But because of measurement errors and unacounted air resistance we pick up an error term, ε.

ε is random.
We propose $\varepsilon \sim N(0, \sigma^2)$.

Then
$$v_i \sim N(at_i, \sigma^2)$$

Here, the parameters are a and σ.
The input (or covariate) is $\vec{t} = (t_1, t_2, \ldots, t_n)$



Simulated data may look as above. ||

E.g 2 PK model

Our data is the plasma drug concentration.
We have a complicated functional relationship (say a pharmacokinetic model).

$$c(t) = f(t, \theta)$$

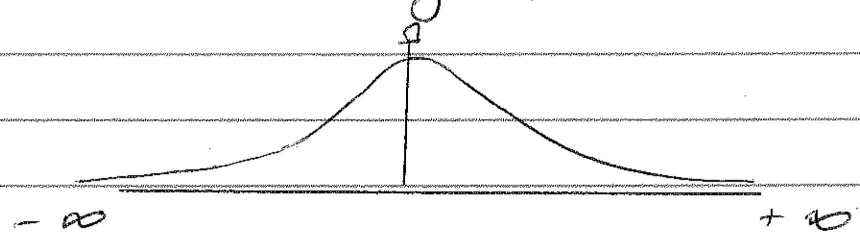
We can then add an error term.
$$c_i \sim N(f(t_i, \theta), \sigma^2).$$ ||

Note that this reasoning applies to linear regression, logistic regression, etc.

<u>Remark</u>: in both examples, my noise was normally distributed.
The normal is mathematically convenient, and sometimes arises due to the Central Limit Theorem. (CLT).
Recall: asymptotically, an average follows a normal.
However, the normal may not always be appropriate.

<u>E.g 3</u> <u>Revisit examples 1 and 2</u>

What is the range of the normal ?



But velocity and drug concentration <u>cannot</u> be negative !

Indeed $c \in [0; +\infty)$.

Other limitation: what if I have more variance when I measure high values? Again this is <u>not</u> captured by our previous models.

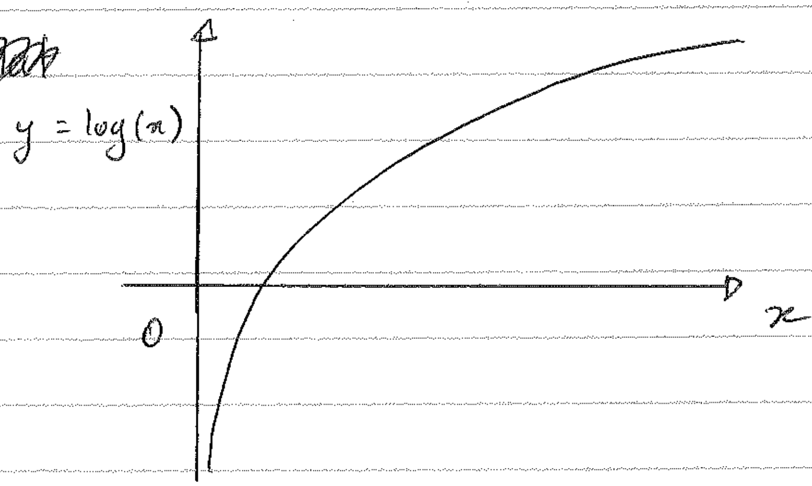
||

The lognormal distribution

Often, it is convenient to work on an unconstrained scale, i.e. $\mathbb{R}$ or $(-\infty ; +\infty)$.

This can be achieved with the log function, since

$$\log : \mathbb{R}^+ \longmapsto \mathbb{R}.$$



Indeed, while $c \in \mathbb{R}^+$, $\log(c) \in \mathbb{R}$.

E.g. 4

Take our drug concentration model.

$$c(t) = f(t, \theta)$$

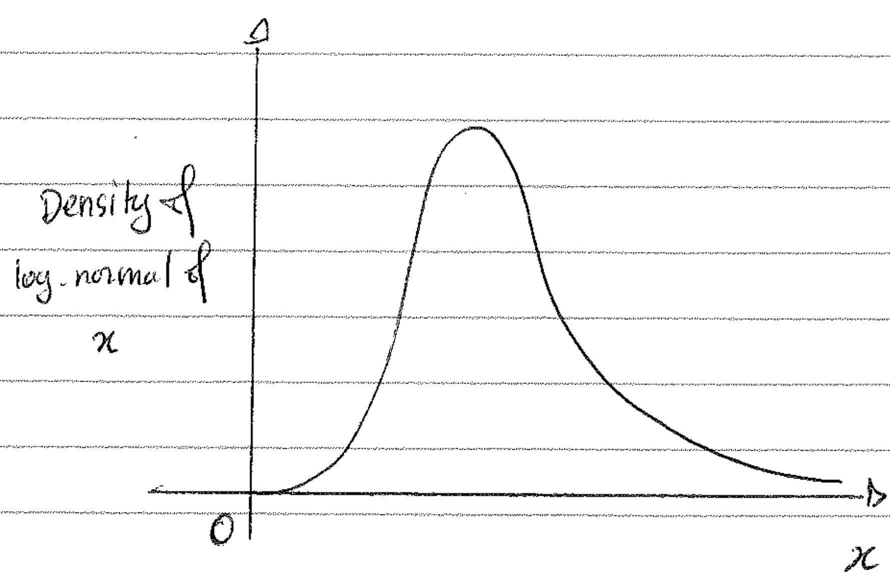
$$\Leftrightarrow \quad \log c(t) = \log f(t, \theta)$$

Add an error term here,

$$\log c(t) \sim N\left(\log f(t,\theta),\ \tau^2\right)$$

This motivates the log normal distribution:

$$c(t) \sim \log N\left(\log f(t,\theta),\ \tau^2\right)$$



As required, $c(t)$ now has to be positive.

In addition

$$\mathrm{Var}\big(c(t)\big) = \left(e^{\tau^2} - 1\right) e^{2\log(f(t,\theta)) + \tau^2}$$

which increases as $f(t,\theta)$ increases !

//

When constructing a model, we want to make sure it generates the data we measure (and captures the phenomenon of interest).

2/ Inference

Usually, we have data, $Z := (X, Y)$, but not θ.
The goal of inference is to reverse-engineer the data generating process.

That is what are values of θ that are consistent with (X, Y) and $\mathcal{M}$.

2.1 / Maximum Likelihood Estimator

The MLE is $\hat{\theta} := \underset{\theta}{\operatorname{argmax}}\, p_\theta(y \mid x)$

E.g

Suppose $X_1, \dots, X_n \overset{iid}{\sim}$ Bernoulli(p)

That is

$$X = \begin{cases} 1, & \text{with prob. } p \\ 0, & \text{with prob } 1-p \end{cases}$$

The probability mass function is

$$p(X=x) = p^{x}(1-p)^{1-x}$$

Does this distribution make sense?

Well

$$p(X=0) = 1-p \quad \checkmark$$

$$p(X=1) = p \quad \checkmark$$

The X's are independent.

Thus

$$p(\underline{x}) = p(x_1, \dots, x_n) = p(x_1)\,p(x_2) \cdots p(x_n)$$

$$= p^{x_1}(1-p)^{1-x_1} \cdots p^{x_n}(1-p)^{1-x_n}$$

$$= p^{\sum_{i=1}^{n} x_i}(1-p)^{\sum_{i=1}^{n} 1-x_i}$$

How do we find the value of p which maximizes $p(\underline{x})$?

The logarithm is monotone.
Thus maximizing $p(\underline{x}_n)$ is the same as maximizing $\log p(\underline{x}_n)$.

And
$$\log p(\underline{x}_n) = \sum_{i=1}^{n} x_i \log(p) + \sum_{i=1}^{n} (1-x_i) \log(1-p)$$

Then
$$\frac{\partial}{\partial p} \log p(\underline{x}_n) = \frac{1}{p} \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} (1-x_i) \frac{1}{1-p}$$

$$= \frac{1}{p(1-p)} \left(\left(\sum_{i=1}^{n} x_i\right)(1-p) - pn + p\left(\sum_{i=1}^{n} x_i\right) \right)$$

$$= \frac{1}{p(1-p)} \left(\sum_{i=1}^{n} x_i - pn \right)$$

At an optimum point, $\frac{\partial}{\partial p} \log p(\underline{x}_n) = 0$.

$$\Leftrightarrow \quad \sum_{i=1}^{n} x_i - \hat{p}n = 0$$

$$\Leftrightarrow \quad \hat{p} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

||

E.g 2

Suppose $X_1, \ldots, X_n \overset{iid}{\sim} N(\theta, 1)$

A similar derivation shows

$$\hat{\theta}_{MLE} = \frac{1}{n} \sum_{i=1}^{n} X_i . \qquad //$$

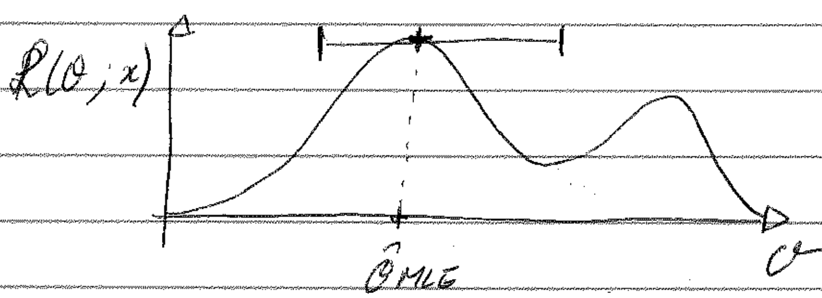
E.g 3 Suppose

for $i = 1, \ldots, n$, $Y_i \sim N(\beta_0 + \beta X_i, \sigma^2)$

This is the setting of the linear regression. Then, the MLE for β_0 and β is the coefficient of the ordinary least square fit.

//

The MLE enjoys many nice mathematical properties.
But it is a point estimate; can build confidence intervals.



In the above example, can a point estimate accurately describe the set of ϑ consistent with Z and M?

2.2/ Bayesian Inference

Proposition: treat the parameter as a random variable and estimate its distribution, given some data.

Want

$$p(\vartheta | Z).$$

From Bayes' rule,

$$p(\vartheta | Z) = \frac{p(Z|\vartheta)\, p(\vartheta)}{p(Z)}$$

$p(Z|\vartheta)$: the well known likelihood function.

$p(\vartheta)$: the prior distribution.
It encodes information about the parameter known before observing the data, based on:
- theoretical constraints
- Results from previous data analysis
- mathematical convenience

$p(z)$: the evidence.
Acts as a normalizing constant.

E.g

Suppose $X_1, X_2, \dots, X_n \overset{iid}{\sim} N(\theta, \sigma^2)$

and $\theta \sim N(\mu, \tau^2)$

These define our likelihood and our prior.

Lemma:

Given the above, the posterior distribution of θ is

$$p(\theta \mid \underline{x}) = N\left(\frac{\mu/\tau^2 + \frac{\bar{X} n}{\sigma^2}}{1/\tau^2 + n/\sigma^2}, \frac{1}{1/\tau^2 + n/\sigma^2} \right)$$

Let's look at the mean

$$E(\theta \mid x) = \frac{\mu/\tau^2 + \frac{\bar{X} n}{\sigma^2}}{1/\tau^2 + n/\sigma^2}$$

It is the weighted average between the prior mean, μ, and the sample mean $\bar{X}$, Recall $\mathrm{Var}(\bar{X}_n) = \sigma^2/n$.

The weights are the inverse variance.

What happens to $\mathbb{E}(\theta|X)$ as $n \to +\infty$?

//

This ~~information~~ example gives us some sense of how much information is encoded in the prior and in the data.

E.g. Bayesian learning

Suppose we observe a first set of data $Z_1 = (X_1, Y_1)$, and compute $p(\theta | Z_1)$ given some initial prior $p(\theta)$.
Our new prior is then $\tilde{p}(\theta) = p(\theta | Z_1)$.
~~Now we observe~~
Suppose we observe a second set of data $Z_2 = (X_2, Y_2)$.
We can then update the posterior.

But what if we had updated our initial prior, $p(\theta)$, simultaneously using Z_1 and Z_2?

First procedure:

$$\tilde{p}(\theta | Z_2) = \frac{p(Z_2|\theta)\, p(\theta | Z_1)}{p(Z_2)}$$

$$= \frac{p(Z_2|\theta)}{p(Z_2)} \frac{p(Z_1|\theta)\, p(\theta)}{p(Z_1)}$$

Now we assume Z_2 and Z_1 are independent, conditional on θ.
Hence

$$p(Z_1, Z_2 | \theta) = p(Z_1|\theta)\, p(Z_2|\theta)$$

Additionally, assume Z_1 and Z_2 are independent.
<u>Remark</u>: conditional independence and independence are <u>not</u> equivalent.

From the above assumption, $Z_1 \perp\!\!\!\perp Z_2$,

$$p(Z_1, Z_2) = p(Z_1)\, p(Z_2)$$

Thus

$$\tilde{p}(\theta | Z_2) = \frac{p(Z_1, Z_2|\theta)\, p(\theta)}{p(Z_1, Z_2)}$$

$$= p(\theta | Z_1, Z_2)$$

Thus, the two procedure yield the same result.

e.g

Suppose $X \sim N(\theta, \sigma^2)$
$\theta \sim N(\mu, \tau^2)$

where

$\mu = 3.0$, $\sigma^2 = 5$

$\tau^2 = 2$,

and we observe $X = \{-1, 4, 3, 2\}$

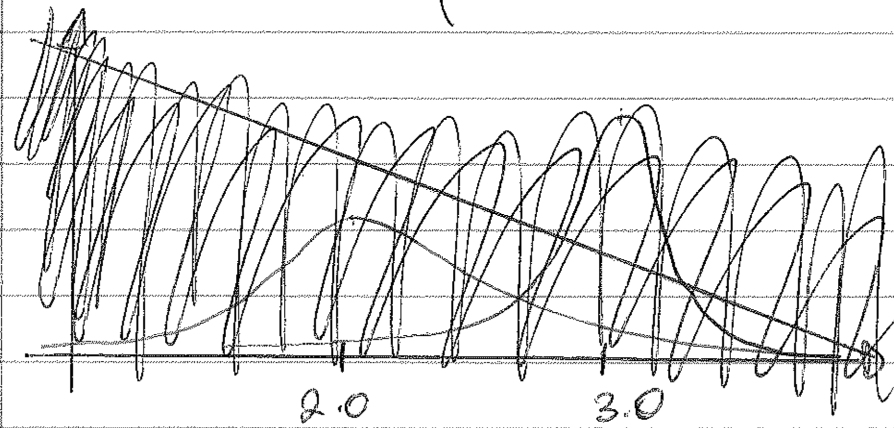
Calculate $p(\theta|x)$.

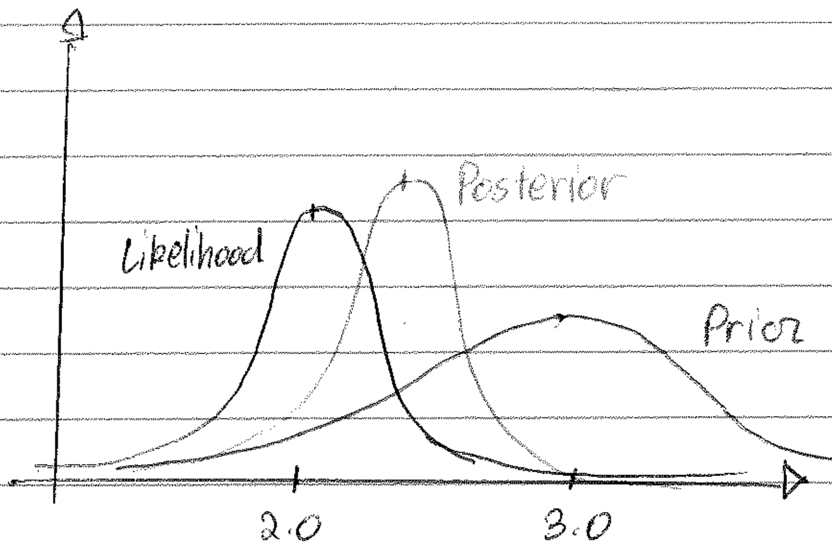
First $n = 4$ and $\bar{X} = \dfrac{-1+4+3+2}{4} = 2$

From conjugacy,

$$p(\theta|x) = N\left(\frac{3.0/2 + 4\times2/5}{1/2 + 4/5}, \frac{1}{1/2 + 4/5}\right)$$

$$\approx \left(2.385, 0.769\right)$$





What if we observe a new data point $Y = \{3.5\}$?
Need to update the prior:

$$E(\theta|X,Y) = \frac{2.385/0.769 + 3.5/5}{1/0.769 + 1/5}$$

$$= 2.533$$

$$Var(\theta|X,Y) = \frac{1}{1/0.769 + 1/5} = 0.667$$

||