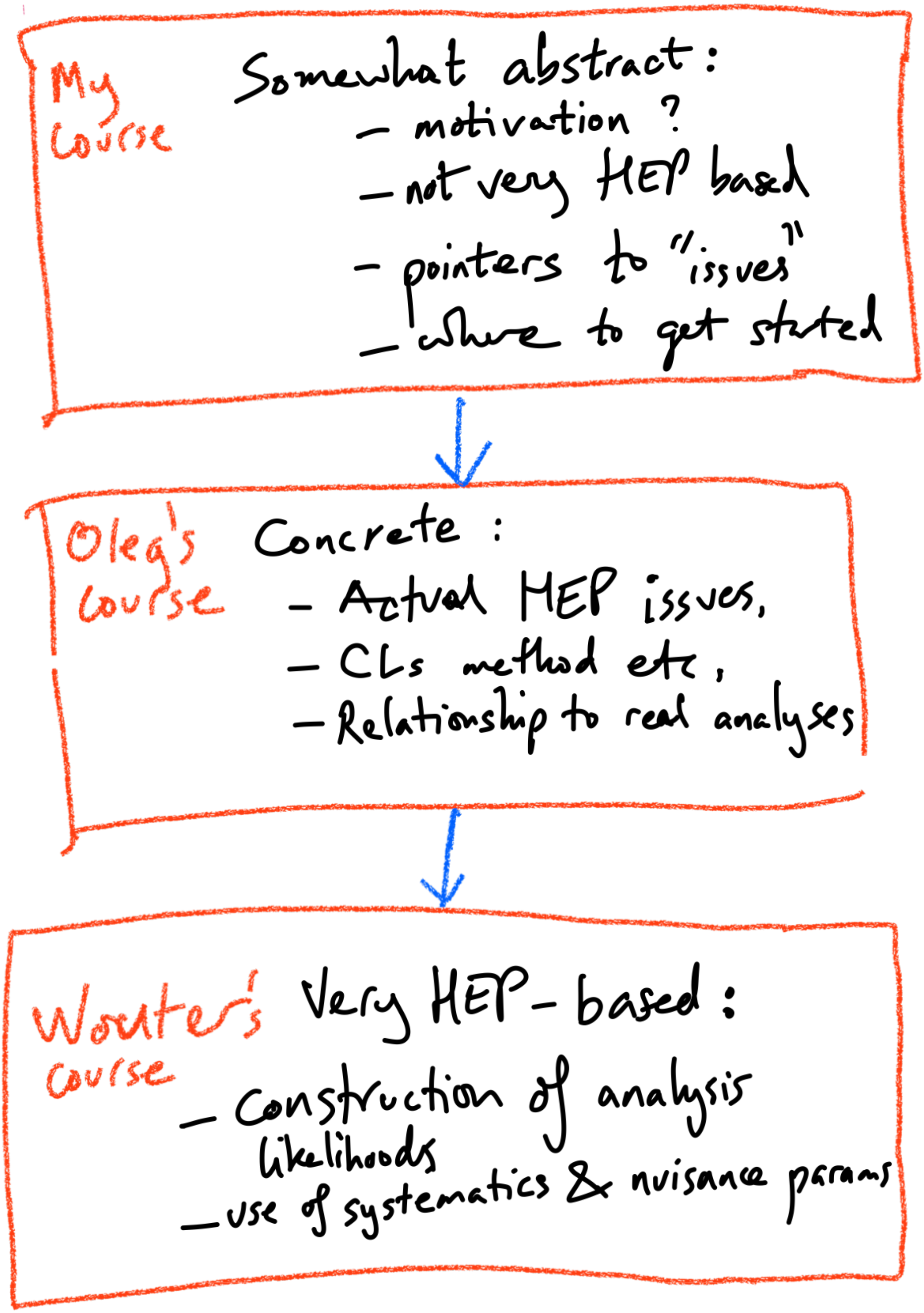
HEP Grad Stats Lectures 1 and 2

Christopher Lester



My course

Somewhat abstract:

- motivation ?
- not very HEP based
- pointers to "issues"
- where to get started

Oleg's course

Concrete :

- Actual HEP issues,
- CL_s method etc,
- Relationship to real analyses

Wouter's course

Very HEP-based:

- construction of analysis likelihoods
- use of systematics & nuisance params

My main goal:

- Make you aware of I.T.I.L.A.

http://www.inference.org.uk/itprnn/book.pdf

(This link has other resources related to ITILA)

Information
Theory
Inference, &
Learning
Algorithms

David MacKay's first best-selling book.

Browsing through it, dipping in and out, will teach you more transferable stats knowledge than any stats course I know

it is LEGALLY FREE as pdf but much more useful as a real book

I recommend you BUY ONE

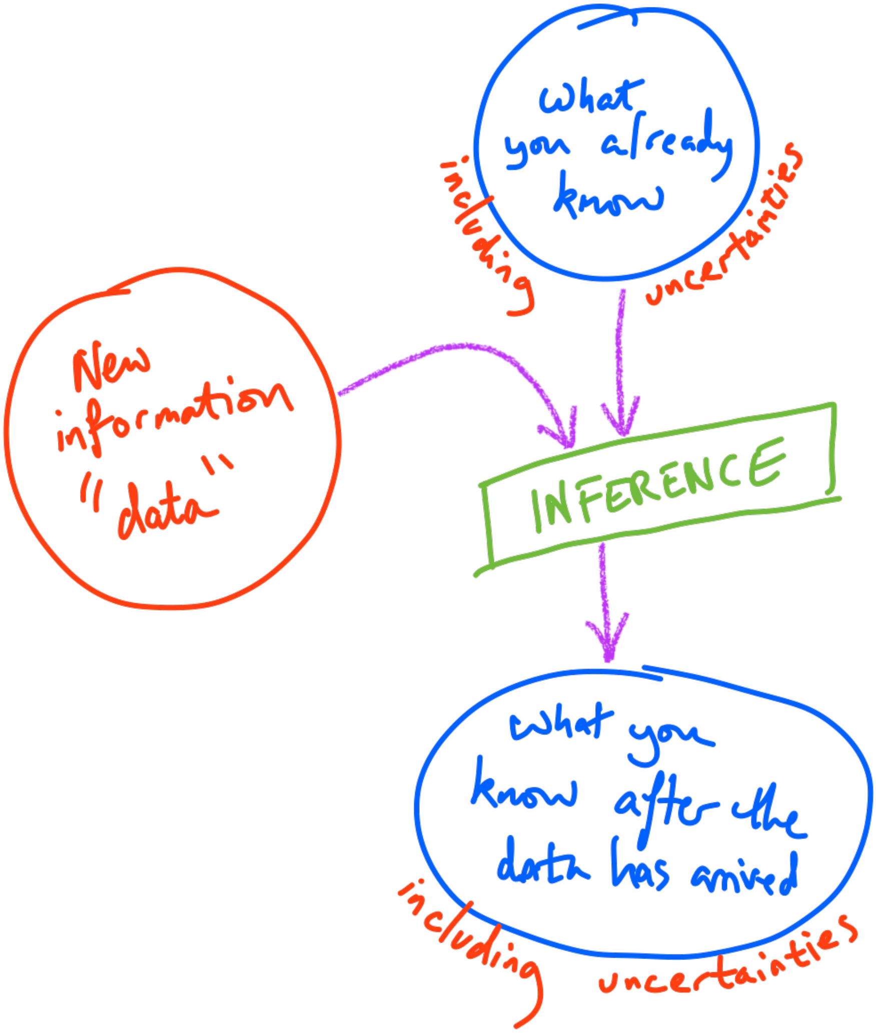
Extra goals:

- I hope you will make efforts to understand the nature of your likelihood, and the "best possible" answer to your problem

  (That is even though the actual inference method you use may be something quite different for practical reasons)

- A desire that you "play" with simple inference problems to stretch your understanding.

Inference problems



Relevant to every branch of science – – and for that matter life & politics too!

The way you solve an inference problem depends on:

- Culture
- Complexity
- Consensus
- Time
- Money / Resources
- How much you care about optimality
- Extent to which problem can be modelled in maths.

HEP is fortunate that:

- most problems are 100% modellable in mathematics, and
- what constitutes data (events?) is usually also well defined

consequence:

The "probability of the data":

$$P\left(\text{data} \,\middle|\, \begin{array}{l}\text{contested}\\\text{assumptions}\\\text{or}\\\text{parameters}\end{array}\right)$$

is, at least in principle, usually well defined

HEP cannot avoid

- Resource problems
- Time problems
- Cultural problems
- Complexity problems

Cannot stress enough that understanding the "probability of the data" for your problem is the single* most important thing to get right...

...before approximations begin.

$$p(\text{data} \mid \text{model})$$

What is it?

units, $\sum_{\text{data}} = 1$, continuous, discrete, mixed?

!

Nested hypotheses?

when is $p(\text{data} \mid \text{params})$ ok?

* RUPERT TOMBS MAY DISAGREE

UNITS UNITS

DISCRETE

$P(4 \mid \text{fair dice}) = \frac{1}{6}$ $\quad [P] = 1$ UNITLESS PROBABILITY

$P(\text{fair dice}) = \frac{9}{10}$ $\quad [P] = 1$ UNITLESS PROBABILITY.

} DISCRETE OUTCOMES

CONTINUOUS

$p(x)\,dx$ is the PROBABILITY that $X \in [x, x+dx)$

PROBABILITIES ARE UNITLESS $\therefore$ $[p(x)\,dx] = 1$

$$\Rightarrow [p(x)][x] = 1$$

$$\Rightarrow [p(x)] = \frac{1}{[x]}$$

} COMMON FEATURE OF CONTINUOUS OUTCOMES

Consistent with $\int p(x)\,dx = \underset{\text{unitless.}}{1}$.

MULTIVARIATE

Know $\iiint p(x,y,z)\,dx\,dy\,dz = 1$

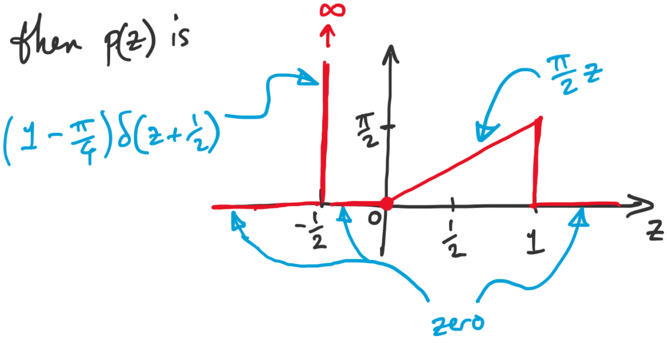
$$\Rightarrow [p(x,y,z)] = \frac{1}{[x]} \cdot \frac{1}{[y]} \cdot \frac{1}{[z]}.$$

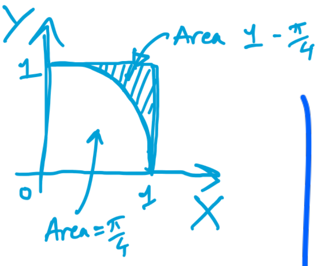
There are distributions which are neither completely discrete nor completely continuous.

For example:

If $\begin{cases} X \sim \text{Unif}[0,1] \\ Y \sim \text{Unif}[0,1] \end{cases}$ and $Z = \begin{cases} \sqrt{X^2+Y^2} & \text{if } X^2+Y^2 \le 1 \\ -\frac{1}{2} & \text{otherwise} \end{cases}$

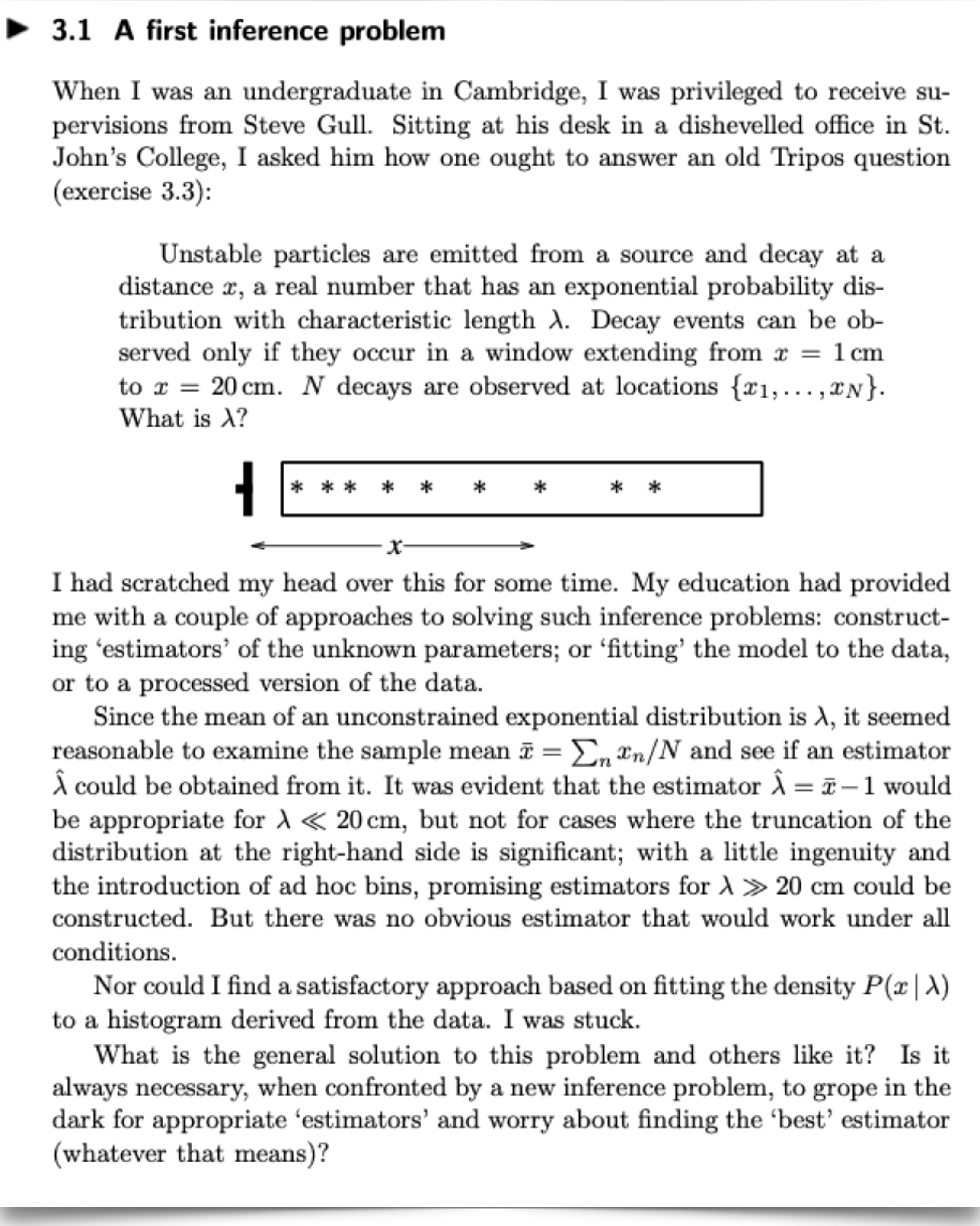
then $p(z)$ is





NEED TO BE CAREFUL WITH THESE

Example from ITILA :



▶ 3.1 A first inference problem

When I was an undergraduate in Cambridge, I was privileged to receive supervisions from Steve Gull. Sitting at his desk in a dishevelled office in St. John's College, I asked him how one ought to answer an old Tripos question (exercise 3.3):

> Unstable particles are emitted from a source and decay at a distance x, a real number that has an exponential probability distribution with characteristic length λ. Decay events can be observed only if they occur in a window extending from $x = 1\,\mathrm{cm}$ to $x = 20\,\mathrm{cm}$. N decays are observed at locations $\{x_1, \ldots, x_N\}$. What is λ?

I had scratched my head over this for some time. My education had provided me with a couple of approaches to solving such inference problems: constructing 'estimators' of the unknown parameters; or 'fitting' the model to the data, or to a processed version of the data.

Since the mean of an unconstrained exponential distribution is λ, it seemed reasonable to examine the sample mean $\bar{x} = \sum_n x_n/N$ and see if an estimator $\hat{\lambda}$ could be obtained from it. It was evident that the estimator $\hat{\lambda} = \bar{x} - 1$ would be appropriate for $\lambda \ll 20\,\mathrm{cm}$, but not for cases where the truncation of the distribution at the right-hand side is significant; with a little ingenuity and the introduction of ad hoc bins, promising estimators for $\lambda \gg 20$ cm could be constructed. But there was no obvious estimator that would work under all conditions.

Nor could I find a satisfactory approach based on fitting the density $P(x \mid \lambda)$ to a histogram derived from the data. I was stuck.

What is the general solution to this problem and others like it? Is it always necessary, when confronted by a new inference problem, to grope in the dark for appropriate 'estimators' and worry about finding the 'best' estimator (whatever that means)?

Steve wrote down the probability of one data point, given λ:

$$P(x \mid \lambda) = \begin{cases} \frac{1}{\lambda} e^{-x/\lambda}/Z(\lambda) & 1 < x < 20 \\ 0 & \text{otherwise} \end{cases}$$

where

$$Z(\lambda) = \int_1^{20} \mathrm{d}x \, \frac{1}{\lambda} e^{-x/\lambda} = \left(e^{-1/\lambda} - e^{-20/\lambda} \right).$$

This seemed obvious enough.

ASIDE:

Dependence of Z on λ comes from acceptance!

Gull's machine has $x_{min} = 1$ cm,

$x_{max} = 20$ cm.

$$Z(\lambda; x_{min}, x_{max}) = \exp\left(\frac{-x_{min}}{\lambda}\right) - \exp\left(\frac{-x_{max}}{\lambda}\right)$$

$$\therefore Z(\lambda; 0, \infty) = \exp(0) - \exp(-\infty)$$

$$= 1 - 0$$

$$= 1$$ independent of λ!

Steve wrote down the probability of one data point, given λ:

$$P(x \mid \lambda) = \begin{cases} \frac{1}{\lambda}\, e^{-x/\lambda}/Z(\lambda) & 1 < x < 20 \\ 0 & \text{otherwise} \end{cases}$$

where

$$Z(\lambda) = \int_1^{20} \mathrm{d}x \, \frac{1}{\lambda}\, e^{-x/\lambda} = \left(e^{-1/\lambda} - e^{-20/\lambda}\right).$$

This seemed obvious enough. Then he wrote *Bayes' theorem*:

$$\begin{aligned} P(\lambda \mid \{x_1, \ldots, x_N\}) &= \frac{P(\{x\} \mid \lambda) P(\lambda)}{P(\{x\})} && (3.3) \\ &\propto \frac{1}{(\lambda Z(\lambda))^N} \exp\left(-\textstyle\sum_1^N x_n/\lambda\right) P(\lambda). && (3.4) \end{aligned}$$

Suddenly, the straightforward distribution $P(\{x_1, \ldots, x_N\} \mid \lambda)$, defining the probability of the data given the hypothesis λ, was being turned on its head so as to define the probability of a hypothesis given the data. A simple figure showed the probability of a single data point $P(x \mid \lambda)$ as a familiar function of x, for different values of λ (figure 3.1). Each curve was an innocent exponential, normalized to have area 1. Plotting the same function as a function of λ for a fixed value of x, something remarkable happens: a peak emerges (figure 3.2). To help understand these two points of view of the one function, figure 3.3 shows a surface plot of $P(x \mid \lambda)$ as a function of x and λ.

Steve wrote down the probability of one data point, given λ:

$$P(x \mid \lambda) = \begin{cases} \frac{1}{\lambda} e^{-x/\lambda}/Z(\lambda) & 1 < x < 20 \\ 0 & \text{otherwise} \end{cases}$$

where

$$Z(\lambda) = \int_1^{20} \mathrm{d}x \, \frac{1}{\lambda} e^{-x/\lambda} = \underbrace{\left(e^{-1/\lambda} - e^{-20/\lambda} \right)}_{\text{Depends on } \lambda \text{ only because of finite acceptance.}}.$$

This seemed obvious enough.

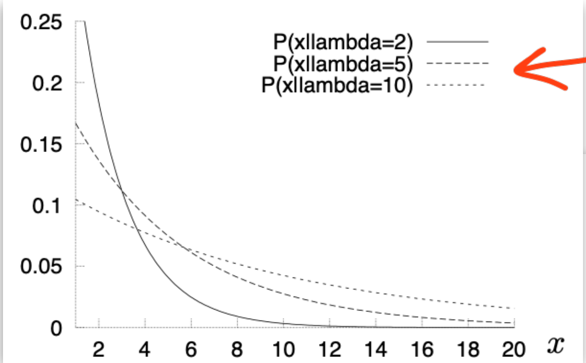


Figure 3.1. The probability density $P(x \mid \lambda)$ as a function of x.

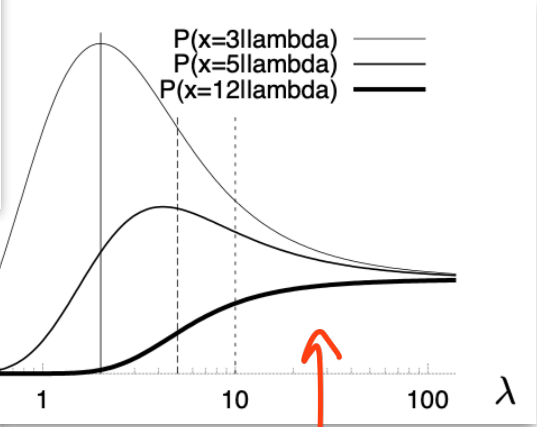
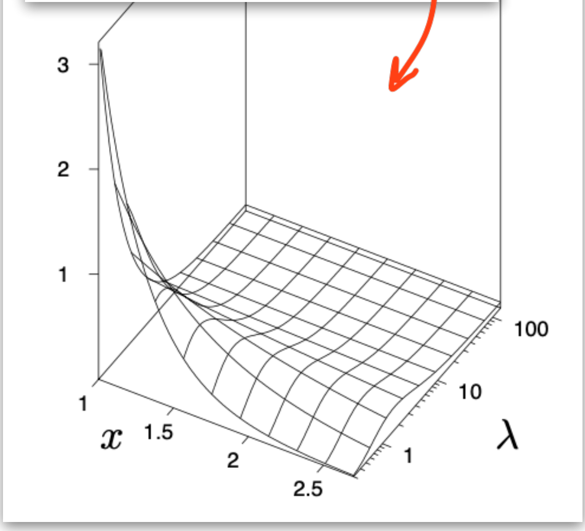


Figure 3.2. The probability density $P(x \mid \lambda)$ as a function of λ, for three different values of x. When plotted this way round, the function is known as the *likelihood* of λ. The marks indicate the three values of λ, $\lambda = 2, 5, 10$, that were used in the preceding figure.

Figure 3.3. The probability density $P(x \mid \lambda)$ as a function of x and λ. Figures 3.1 and 3.2 are vertical sections through this surface.



$$\int p(x \mid \lambda)\, dx = 1$$

$$\int p(x \mid \lambda)\, d\lambda \neq 1$$

(in general)

This seemed obvious enough. Then he wrote *Bayes' theorem*:

$$P(\lambda \,|\, \{x_1,\ldots,x_N\}) = \frac{P(\{x\}\,|\,\lambda)P(\lambda)}{P(\{x\})}$$
$$\propto \frac{1}{(\lambda Z(\lambda))^N} \exp\left(-\textstyle\sum_1^N x_n/\lambda\right) P(\lambda).$$

For a dataset consisting of several points, e.g., the six points $\{x\}_{n=1}^N = \{1.5, 2, 3, 4, 5, 12\}$, the likelihood function $P(\{x\}\,|\,\lambda)$ is the product of the N functions of λ, $P(x_n\,|\,\lambda)$ (figure 3.4).

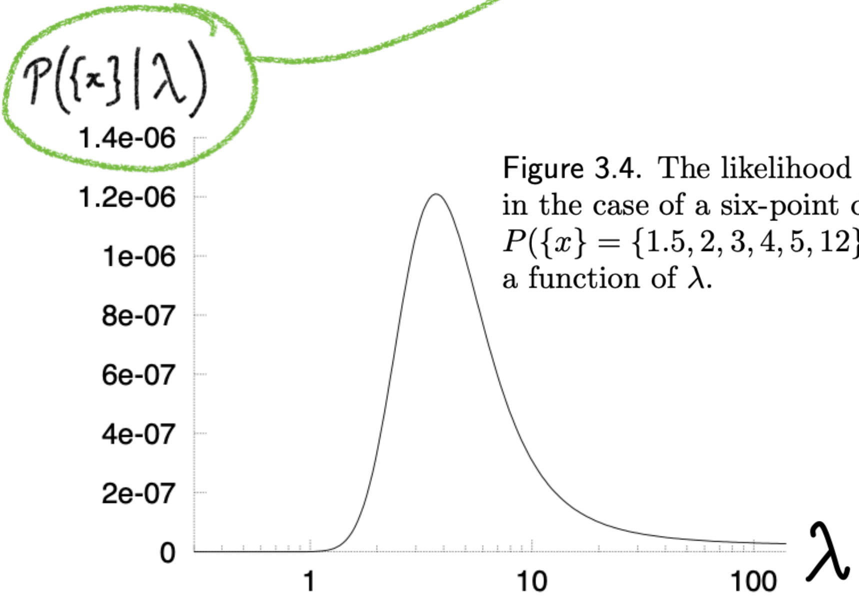


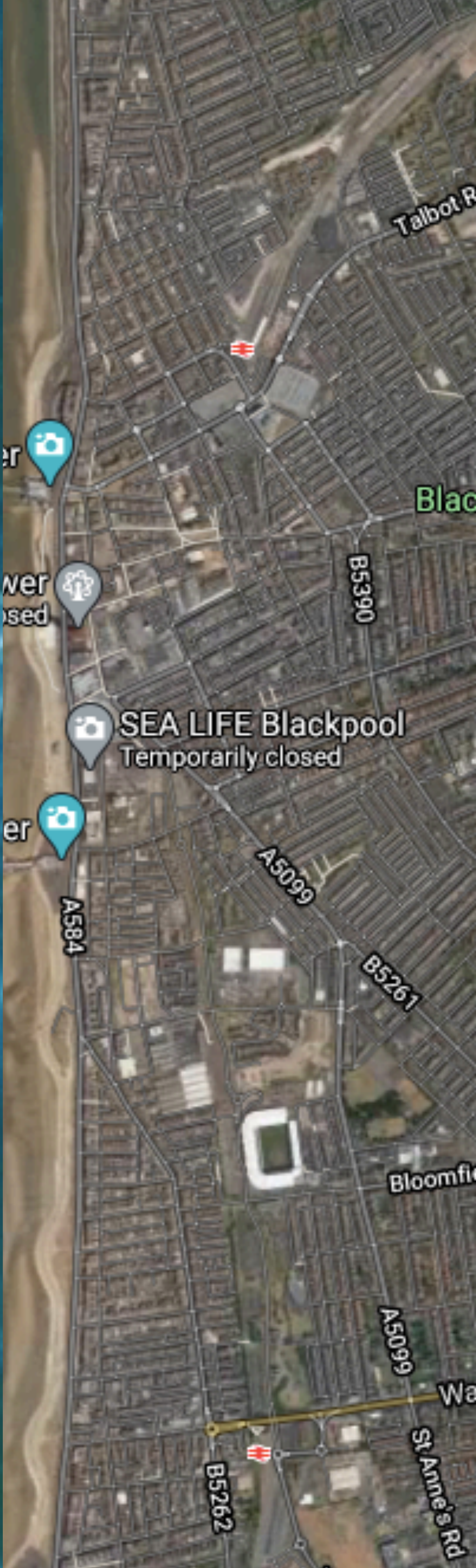
Figure 3.4. The likelihood function in the case of a six-point dataset, $P(\{x\} = \{1.5, 2, 3, 4, 5, 12\}\,|\,\lambda)$, as a function of λ.

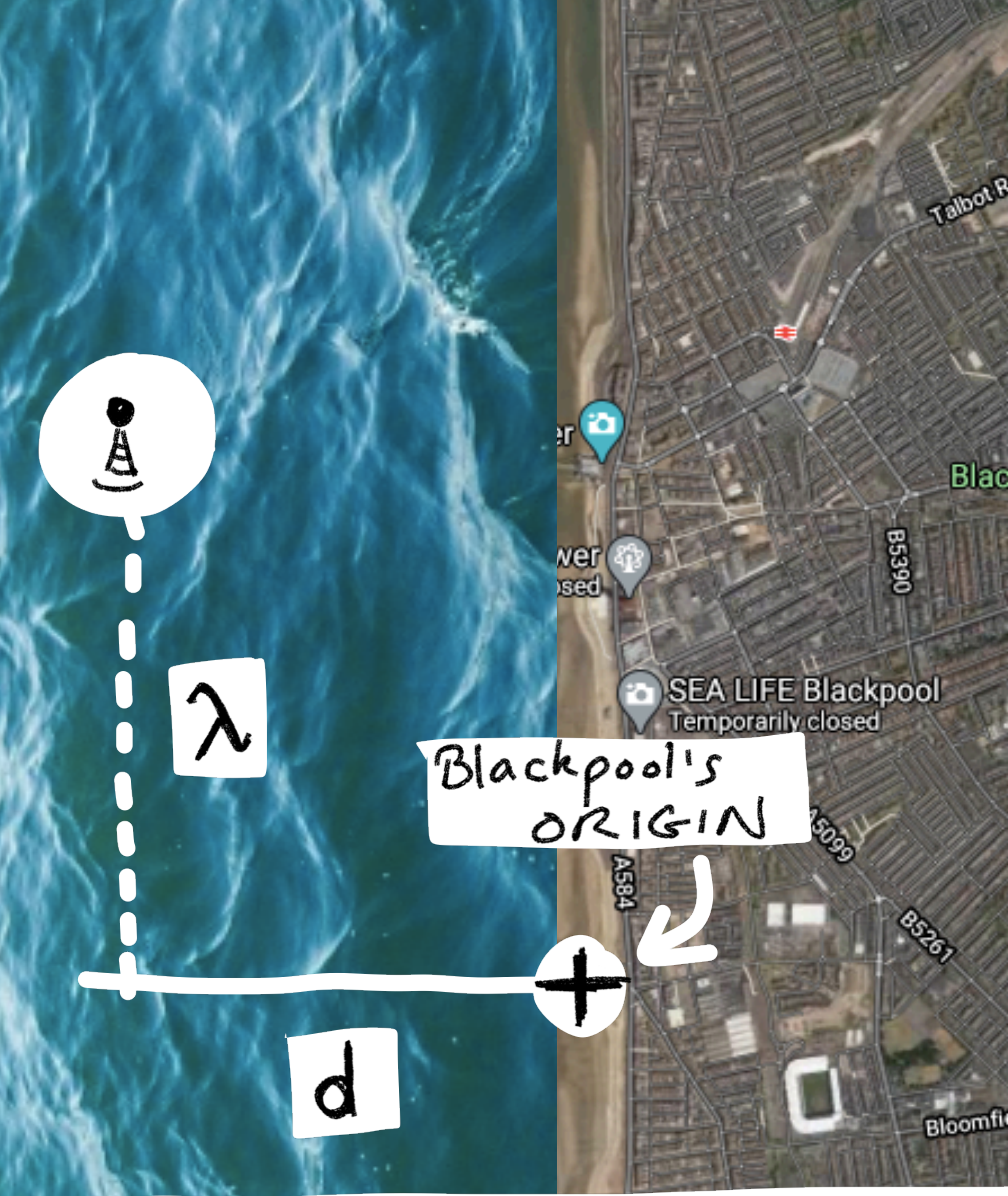
Don't disagree with $P(\{x\}|\lambda)$.
Consider dis~~agree~~ing with $P(\lambda)$ or
perhaps on whether you like $P(\lambda|\{x\})$.

Let's gain some experience from PLAYING with a similar example we invent for ourselves.

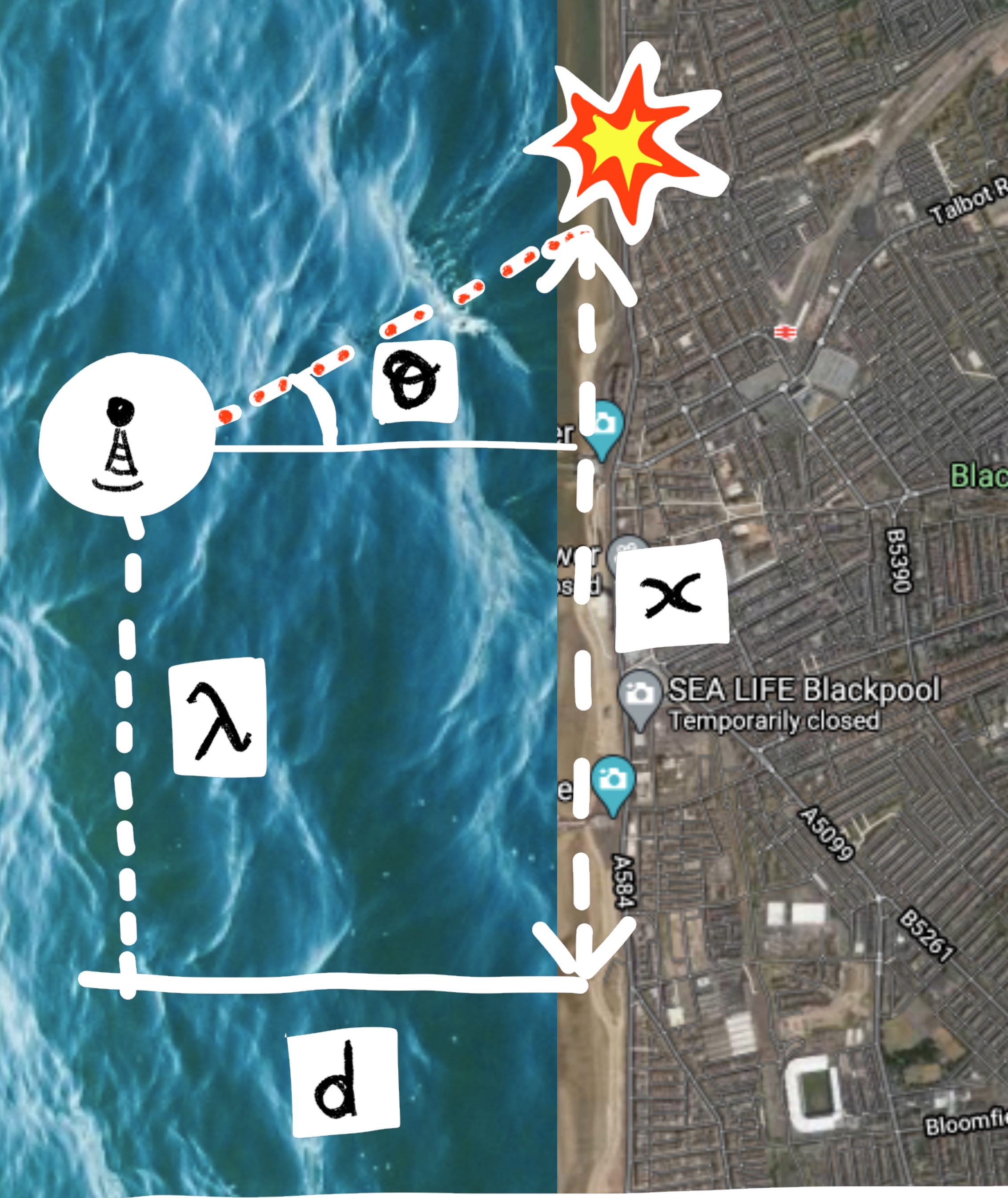
1. We could try IGNORING the advice
2. We could try TAKING the advice
3. We could try to LEARN something

Blackpool's
d-mile
lighthouse


Talbot R
Blac
er
wer
osed
B5390
SEA LIFE Blackpool
Temporarily closed
er
A5099
A584
B5261
Bloomfie
A5099
Wa
St Anne's Rd
B5262



The d-mile lighthouse



The d-mile lighthouse DISASTER

$\boxed{\text{Attempt 1}}$ Ignore advice

— Follow our noses —

HEP Physicist Special

Attempt 1

Ignore advice

— Follow our noses —

Consider

MEAN

value of $\{x_1, x_2, \ldots, x_N\}$

JO USED SOMETHING SIMILAR

SHRUG

NEED PLOT TOMORROW

GUESS

→ MATHEMATICA DEMO

HEP Physicist Special

Attempt 1

Ignore advice

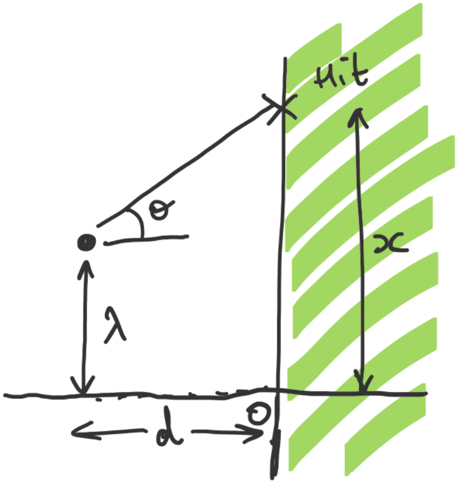
— Follow our noses —

Why is increasing the amount of data not helping our estimate of λ?

Shouldn't precision on the mean of x go down like $\frac{1}{\sqrt{N}}$?

$$x_i \sim G(\mu, \sigma^2) \Rightarrow \left(\begin{matrix}\text{mean of} \\ N\ x_i\end{matrix}\right) \sim G\left(\mu, \frac{\sigma^2}{N}\right)$$

HEP Physicist Special



$$p_\theta(\theta) = \begin{cases} \frac{1}{\pi} & \theta \in [-\frac{\pi}{2}, \frac{\pi}{2}] \\ 0 & \text{otherwise} \end{cases}$$

$$p_\theta(\theta)\, d\theta = p_x(x)\, dx$$

$$\langle \theta \rangle = \int \theta\, p_\theta(\theta) = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \theta \frac{1}{\pi} d\theta = 0 \quad \text{(by inspection)}$$

$$\langle \theta^2 \rangle = \int \theta^2 p_\theta(\theta)\, d\theta = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \theta^2 \frac{1}{\pi} d\theta = \left[\frac{1}{3}\theta^3\right]_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \frac{1}{\pi} = \frac{1}{\pi}\frac{2}{3}\left(\frac{\pi}{2}\right)^3 = \frac{\pi^2}{12}$$

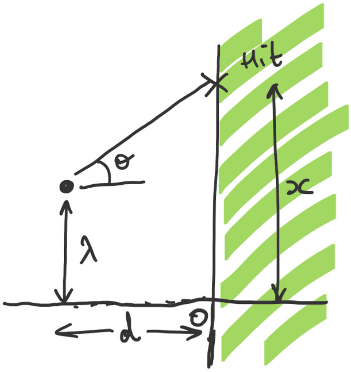
$$\therefore \mathrm{Var}(\theta) = \langle \theta^2 \rangle - \langle \theta \rangle^2 = \frac{\pi^2}{12} - 0 = \frac{\pi^2}{12}$$

Tracking Resolution = (sensor pitch)/$\sqrt{12}$ ✓✓

(KNOWN TO MOST PHYSICISTS WHO WORK ON TRACKERS)

What about $\langle x \rangle$, $\langle x^2 \rangle$ & $\mathrm{Var}(x)$?

Will need to work out $p_x(x|\lambda)$ first ....



$$p_\theta(\theta) = \begin{cases} \frac{1}{\pi} & \theta \in [-\frac{\pi}{2}, \frac{\pi}{2}] \\ 0 & \text{otherwise} \end{cases}$$

$$p_\theta(\theta)\, d\theta = p_x(x)\, dx$$

$$\therefore p_x(x) = p_\theta(\theta) \frac{d\theta}{dx} \qquad (2)$$

Need $\frac{d\theta}{dx}$. First relate θ & x:

$$\frac{x-\lambda}{d} = \tan\theta \qquad \therefore x = \lambda + d\tan\theta \qquad (1)$$

$$(1) \Rightarrow 1 = d\sec^2\theta \frac{d\theta}{dx}$$

$$\Rightarrow \frac{d\theta}{dx} = \frac{1}{d\sec^2\theta}$$

$$= \frac{1}{d(1+\tan^2\theta)}$$

$$= \frac{1}{d\left(1+\frac{(x-\lambda)^2}{d^2}\right)}$$

$$= \frac{d}{d^2+(x-\lambda)^2}$$

$$\therefore (2) \Rightarrow p_x(x|\lambda) = \frac{1}{\pi} \cdot \frac{d}{d^2+(x-\lambda)^2}$$

Check: $\int_{-\infty}^{\infty} p_x(x)\, dx = \left[\frac{d}{\pi}\frac{1}{d}\tan^{-1}\left(\frac{x-\lambda}{d}\right)\right]_{-\infty}^{\infty} = \frac{1}{\pi}\left(\frac{\pi}{2} - \left(-\frac{\pi}{2}\right)\right) = 1$ ✓✓

$$\langle x \rangle = \int x\, p_x(x)\, dx$$

$$= \int_{-\infty}^{\infty} \frac{x d}{d^2 + (x-\lambda)^2}\, dx \qquad y = x - \lambda$$

$$= d \int_{-\infty}^{\infty} \frac{y + \lambda}{d^2 + y^2}\, dy$$

$$= d \int_{-\infty}^{\infty} \frac{y}{d^2+y^2}\, dy + d\lambda \int_{-\infty}^{\infty} \frac{1}{d^2+y^2}\, dy$$

$$= d \left[\frac{1}{2} \ln\left(d^2+y^2\right) \right]_{-\infty}^{\infty} + d\lambda \left[\frac{1}{d} \tan^{-1}\left(\frac{y}{d}\right) \right]_{-\infty}^{\infty}$$

$$= d \left\{ \underbrace{\frac{1}{2}\left(\ln\infty - \ln\infty\right)}_{\substack{\uparrow \\ \text{OH DEAR}}} \right\} + \lambda \left(\underbrace{\frac{\pi}{2} - \left(-\frac{\pi}{2}\right)}_{\text{☺}} \right)$$

$$\langle x^2\rangle = \int x^2 p_x(x)\,dx$$

$$= \int_{-\infty}^{\infty} \frac{x^2 d}{d^2+(x-\lambda)^2}\,dx$$

$y = x - \lambda$

$$= d\int_{-\infty}^{\infty}\frac{(y+\lambda)^2}{d^2+y^2}\,dy$$

$$= d\int_{-\infty}^{\infty}\frac{d^2+y^2+2\lambda y+(\lambda^2-d^2)}{d^2+y^2}\,dy$$

$$= d\underbrace{\int_{-\infty}^{\infty}1\,dy}_{+\infty\,!} + \underbrace{2\lambda d\int_{-\infty}^{\infty}\frac{y}{d^2+y^2}\,dy}_{\text{The } \ln(\infty)-\ln(\infty) \text{ term we saw before}} + d(\lambda^2-d^2)\underbrace{\int\frac{1}{d^2+y^2}\,dy}_{\text{The ☺ term we saw before}}$$

$$\therefore \mathrm{Var}(x) = \langle x^2\rangle - \langle x\rangle^2 = \infty$$ ☹

Attempt 1

Ignore advice

— Follow our noses —

Oh dear. The innocuous looking x distribution had:

- INFINITE variance, ☹
- NO mean at all! ☹

MORAL:

The mean is not as simple as they told you in kindergarten. It doesn't always exist, and "$\infty - \infty$" is not zero.

$$x_i \sim G(\mu, \sigma^2) \Rightarrow \begin{pmatrix}\text{mean of}\\ N\ x_i\end{pmatrix} \sim G\left(\mu, \frac{\sigma^2}{N}\right)$$

BVT

$$x_i \sim \text{Blackpool}(\lambda, \infty) \Rightarrow \overset{\text{Non-existent}}{\begin{pmatrix}\text{mean of}\\ N\ x_i\end{pmatrix}} \sim \left(\mu, \frac{\infty}{N}\right)$$

HEP Physicist Special

Details matter!

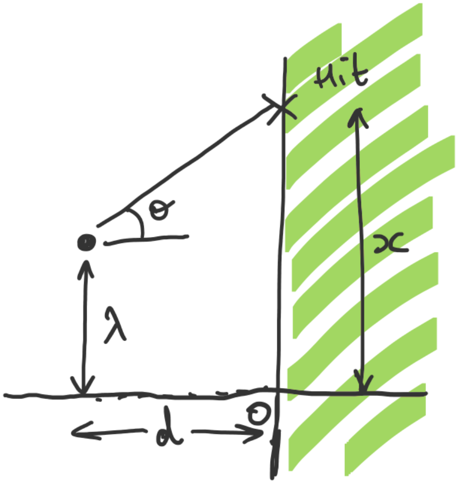
Attempt 2

Take advice

— Follow the Likelihood —

$$p_x(x|\lambda) = \frac{1}{\pi}\cdot\frac{d}{d^2+(x-\lambda)^2}$$

Cardsharp / smart investor special



$$p_\theta(\theta) = \begin{cases} \frac{1}{\pi} & \theta \in [-\frac{\pi}{2}, \frac{\pi}{2}] \\ 0 & \text{otherwise} \end{cases}$$

$$p_\theta(\theta)\,d\theta = p_x(x)\,dx$$

$$p_x(x|\lambda) = \frac{1}{\pi}\cdot\frac{d}{d^2+(x-\lambda)^2}$$

Bayes Theorem:

$$p(\lambda|\underset{\uparrow\,\underline{x}}{\text{data}}) = \frac{p(\overset{\underline{x}}{\text{data}}|\lambda)p(\lambda)}{p(\text{data})} \underset{\text{keeping } \lambda \text{ dependence only}}{\propto} p(\overset{\underline{x}}{\text{data}}|\lambda)p(\lambda)$$

$$= p(\lambda)\,p(\{x_1, x_2, x_3, \dots, x_N\}|\lambda)$$

$$= p(\lambda) \prod_i p_x(x_i|\lambda)$$

ignoring factors that don't depend on λ ↘

$$\propto p(\lambda) \prod_i \frac{1}{d^2+(x_i-\lambda)^2}$$

→ VISUALISE IN MATHEMATICA.

So: Attempt 2 was much better than Attempt 1.

Does that mean following the likelihood is always best?

No:

- Tractability still important
- simple solutions sometimes still exist which keep the Frequentists happy

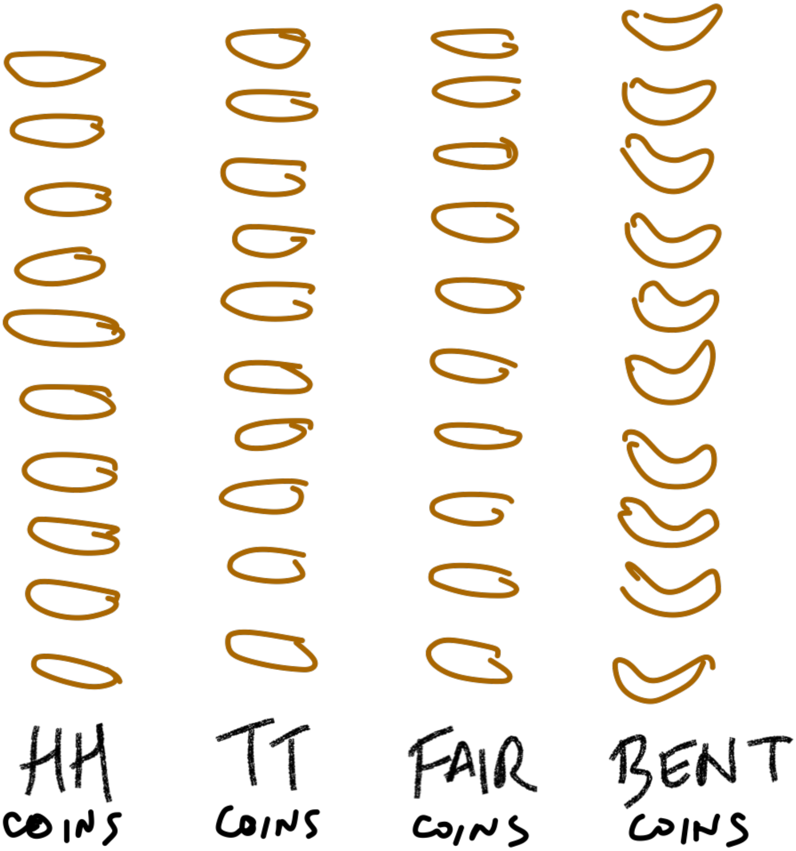
- SEE MATHEMATICA (MEDIAN DEMO) →

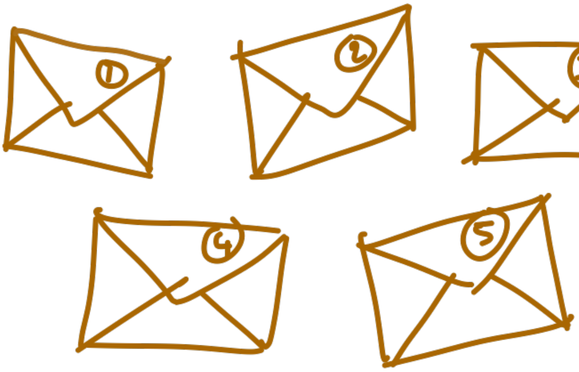
- NOTE EXTRA WORK WOULD BE NEEDED TO GIVE UNCERTAINTY TO MEDIAN ? BULK?

- DISCUSS MAX-LIKELIHOOD IN THIS EXAMPLE

- SHALL NOT DISCUSS MAX-LIKELIHOOD ISSUES (REPARAMETRISATION)

Time to play the coin game?





THIS IS ABOUT PRIORS

f = fraction HEADS in biased coin

data = time-ordered coin toss history
= "{H, H, T, H, T, T, H}" (for example).

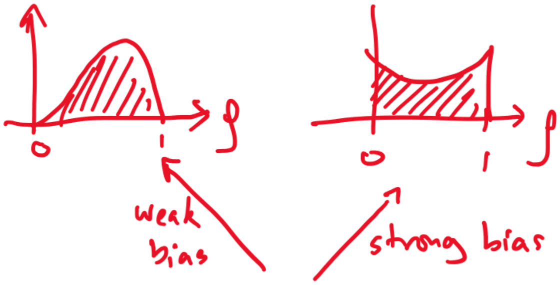
N_H = # of heads in data
N_T = # of tails in data
$N = N_H + N_T$ = # of tosses

N is assumed fixed and "given".

Priors:

$p(HH)$

$p(TT)$



$p(\text{fair})$

$p(\text{biased})$ & $p(f \mid \text{biased})$

Constraints on Priors:

$$p(HH) + p(TT) + p(\text{fair}) + p(\text{biased}) = 1$$

$$\int_0^1 p(f \mid \text{biased})\, df = 1$$

$$p(\text{data} \mid HH) = \begin{cases} 0 & \text{if } N_T > 0 \\ 1 & \text{otherwise} \end{cases}$$

$$p(\text{data} \mid TT) = \begin{cases} 0 & \text{if } N_H > 0 \\ 1 & \text{otherwise} \end{cases}$$

$$p(\text{data} \mid \text{fair}) = \left(\tfrac{1}{2}\right)^N$$

$$p(\text{data} \mid \text{biased}; f) = f^{N_H} \cdot (1-f)^{N_T}$$

$$p(\text{data} \mid \text{biased}) \equiv \int_0^1 p(\text{data} \mid \text{biased}, f)\, p(f \mid \text{biased})\, df$$

Bayes theorem:

$$p(HH \mid \text{data}) = \frac{p(\text{data} \mid HH)\, p(HH)}{p(\text{data})}$$

$$p(TT \mid \text{data}) = \frac{p(\text{data} \mid TT)\, p(TT)}{p(\text{data})}$$

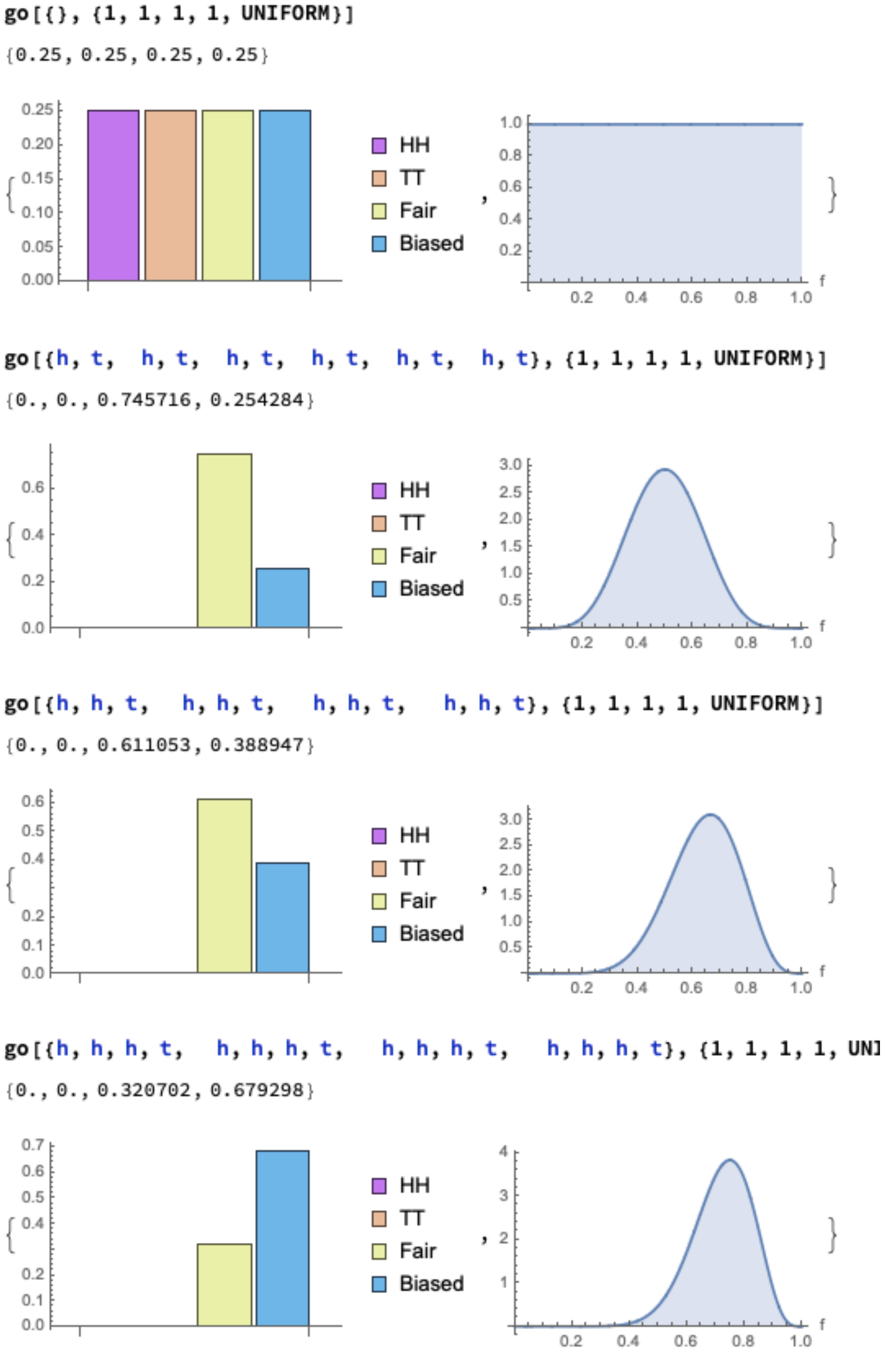
$$p(\text{fair} \mid \text{data}) = \frac{p(\text{data} \mid \text{fair})\, p(\text{fair})}{p(\text{data})}$$

$$p(\text{biased} \mid \text{data}) = \frac{p(\text{data} \mid \text{biased})\, p(\text{biased})}{p(\text{data})}$$

subsidiary result

$$p(f \mid \text{data}, \text{biased}) = \frac{p(\text{data} \mid f, \text{biased})\, p(f \mid \text{biased})}{p(\text{data} \mid \text{biased})}$$

$$\propto p(\text{data} \mid f, \text{biased})\, p(f \mid \text{biased})$$



```
go[{}, {1, 1, 1, 1, UNIFORM}]
{0.25, 0.25, 0.25, 0.25}
```

```
go[{h, t,  h, t,  h, t,  h, t,  h, t,  h, t}, {1, 1, 1, 1, UNIFORM}]
{0., 0., 0.745716, 0.254284}
```

```
go[{h, h, t,   h, h, t,   h, h, t,   h, h, t}, {1, 1, 1, 1, UNIFORM}]
{0., 0., 0.611053, 0.388947}
```

```
go[{h, h, h, t,   h, h, h, t,   h, h, h, t,   h, h, h, t}, {1, 1, 1, 1, UNI
{0., 0., 0.320702, 0.679298}
```

CavMyDocs/teaching/GradStats/2020/*.nb

You are FORCED to use cut and count

Should you use a BDT ?

Should you eyeball the best cut ?

What is the best cut _anyone_ could make?

You are FORCED to use cut and count

Should you use a BDT?

Should you eyeball the best cut?

What is the best cut anyone could make?

The Neyman-Pearson Lemma tells us!

The

Neyman-Pearson Lemma

tells us that:

The **best** **possible** cut is a cut on the likelihood ratio:

$$\rho = \frac{p(\text{event} \mid \text{signal})}{p(\text{event} \mid \text{background})}$$

.... in more detail:

Put all the data from an event into $\underline{x}$.

$\underline{x}$ may have many dimensions :

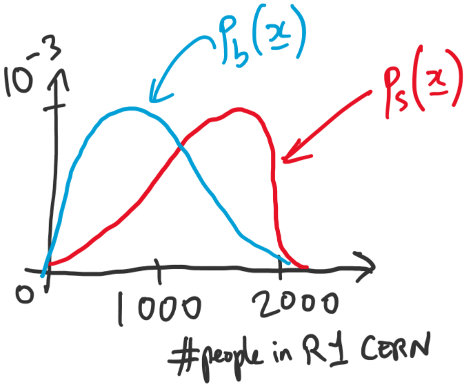
$$\underline{x} = (a^{\mu}, b^{\mu}, \text{\#people in R1 CERN}, \text{run-number}, \ldots)$$

There is some multidimensional probability density for $\underline{x}$ for signal events, and likewise for background events:

$$p(\underline{x} \,|\, \text{signal}) \longrightarrow P_s(\underline{x}) \qquad \text{(for short)}$$

$$p(\underline{x} \,|\, \text{background}) \longrightarrow P_b(\underline{x})$$

E.g. If signal events are (by defn) those on days where R1 sells GOOD PIZZA and background events are those from any other day, then



× (factors to do with rain and ATLAS weeks)

Put <u>all</u> the data from an event into $\underline{x}$.

$\underline{x}$ may have many dimensions:

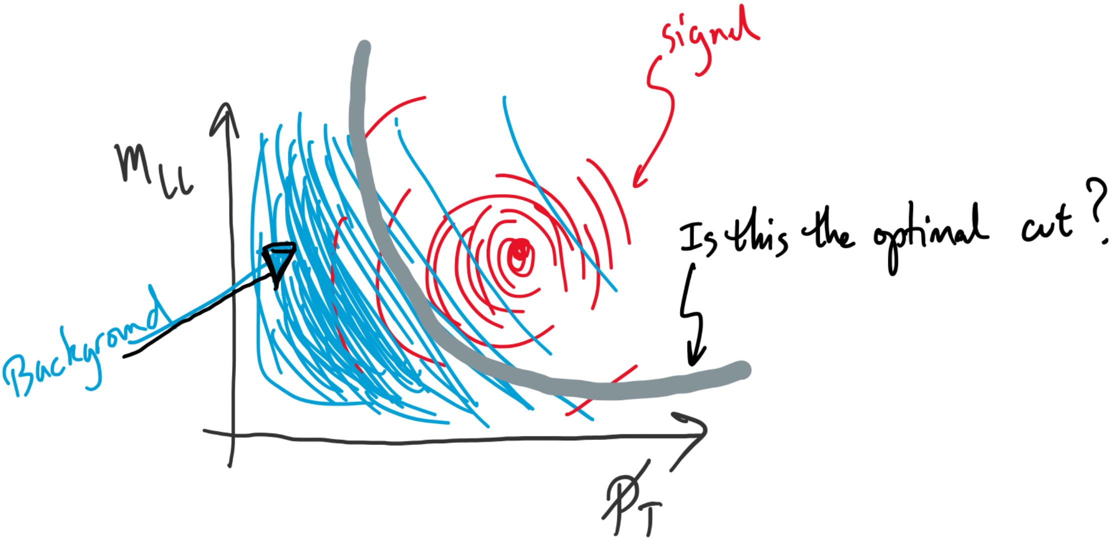
$$\underline{x} = (a^\mu, b^\mu, \#\text{people in R1 CERN}, \text{run-number}, \ldots)$$

There is some multidimensional probability density for $\underline{x}$ for signal events, and likewise for background events:

$$p(\underline{x} \mid \text{signal}) \longrightarrow P_s(\underline{x})$$
$$p(\underline{x} \mid \text{background}) \longrightarrow P_b(\underline{x})$$

(for short)

Two dimensional example:

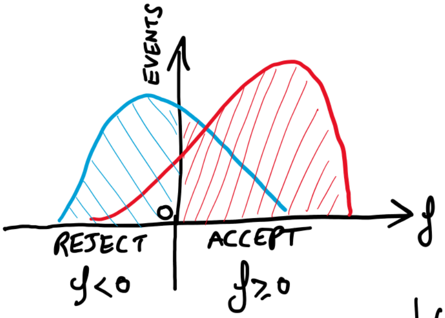


What is optimal depends on what you are optimizing for:

What do you want to maximise? $\frac{N_s}{N_b}$, $\frac{N_s}{\sqrt{N_b}}$, $\frac{N_s+N_b}{\sqrt{N_b}+0.3N_b}$, ... ?

Let $K(N_s,N_b)$ be the thing you want to optimize.

Let $f(\underline{x})$ define an OPTIMAL cut like this:



$h(\underline{x})$ is an arbitrary perturbing event variable.

Let $g(\underline{x},\mu) = f(\underline{x}) + \mu h(\underline{x})$

Heaviside step function

Define $$D_s(\mu) = \int \Theta(g(\underline{x},\mu))\, p_s(\underline{x})\, d\underline{x}$$

$$D_b(\mu) = \int \Theta(g(\underline{x},\mu))\, p_b(\underline{x})\, d\underline{x}$$

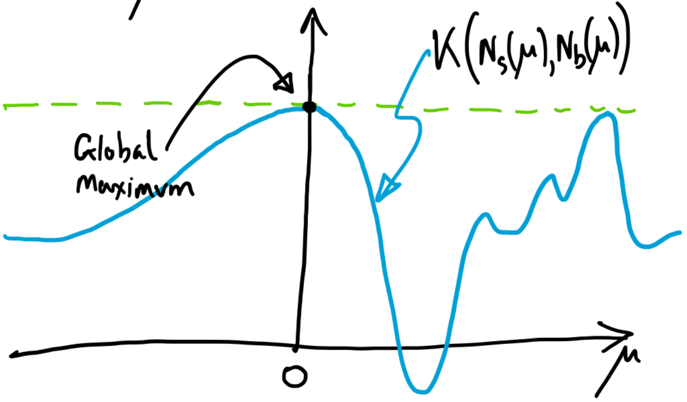
Suppose also that λ is the fraction of signal events among N and that $1-\lambda$ is the fraction of background events.

abbreviation

Then clearly $$\langle N_s \rangle = N \lambda D_s(\mu) = N_s$$
$$\langle N_b \rangle = N(1-\lambda) D_b(\mu) = N_b$$

are the number of signal and background events we would select if we used $g(\underline{x},\mu) \geq 0$ as our selection.

We know that since $f(\underline{x})$ is OPTIMAL and since $f(\underline{x}) = g(\underline{x},0)$ then $K(N_s(\mu), N_b(\mu))$ attains its global maximum at $\mu = 0$:



and hence at that maximum:

$$0 = \left.\frac{dK}{d\mu}\right|_{\mu=0} = \left(\frac{\partial K}{\partial N_s} \frac{dN_s}{d\mu} + \frac{\partial K}{\partial N_b} \frac{dN_b}{d\mu} \right)_{\mu=0}$$

$$0 = \left.\frac{dK}{d\mu}\right|_{\mu=0} = \left(\frac{\partial K}{\partial N_s}\frac{dN_s}{d\mu} + \frac{\partial K}{\partial N_b}\frac{dN_b}{d\mu}\right)_{\mu=0}$$

ABBREVIATION →

$$= \underbrace{\frac{\partial K}{\partial N_s} N \lambda}_{K_s} D_s'(0) + \underbrace{\frac{\partial K}{\partial N_b} N (1-\lambda)}_{K_b} D_b'(0)$$

$$= K_s D_s'(0) + K_b D_b'(0) \qquad \circledast$$

So, what is $D_i'(\mu)$?

Heaviside step

$$D_i'(\mu) = \frac{d}{d\mu}\int \Theta\big(g(\underline{x},\mu)\big)\, p_i(\underline{x})\, d\underline{x}$$

differentiation under $\int$ sign.

$$= \int \frac{\partial}{\partial \mu}\Big(\Theta\big(g(\underline{x},\mu)\big)\Big)\, p_i(\underline{x})\, d\underline{x}$$

MAGIC

$$= \int \delta\big(g(\underline{x},\mu)\big)\, \frac{\partial g(\underline{x},\mu)}{\partial \mu}\, p_i(\underline{x})\, d\underline{x}$$

BY DEF OF $g(\underline{x},\mu)$

$$= \int \delta\big(g(\underline{x},\mu)\big)\, h(\underline{x})\, p_i(\underline{x})\, d\underline{x}$$

$$\therefore\ D_i'(0) = \int \delta\big(f(\underline{x})\big)\, h(\underline{x})\, p_i(\underline{x})\, dx$$

When we defined $h(\underline{x})$ we said it was <u>arbitrary</u>. Everything we have done up to now could use any $h(\underline{x})$. We now use that freedom to set

$$h(\underline{x}) = \delta^{(n)}(\underline{x} - \underline{m})$$

some arbitrary constant event

With that choice:

$$D_i'(0) = \int \delta(f(\underline{x}))\, \delta^{(n)}(\underline{x} - \underline{m})\, p_i(\underline{x})\, dx$$

$$= \delta(f(\underline{m}))\, p_i(\underline{m})$$

Substituting this back into $\circledast$ our optimality condition becomes:

$$0 = K_s\, \delta(f(\underline{m}))\, p_s(\underline{m}) + K_b\, \delta(f(\underline{m}))\, p_b(\underline{m})$$

$$\Rightarrow \quad 0 = \delta(f(\underline{m}))\left[K_s\, p_s(\underline{m}) + K_b\, p_b(\underline{m})\right]. \quad \oplus$$

Recall that $\oplus$ must be true for ALL $\underline{m}$. Because the δ-function is zero when $f(\underline{m}) \neq 0$, the second term in $\oplus$ only constrains values of $\underline{m}$ for which $f(\underline{m}) = 0$. These are events $\underline{m}$ lying on the OPTIMAL CUT.

$\therefore$ Every event $\underline{m}$ on the OPTIMAL CUT satisfies:

$$K_s\, p_s(\underline{m}) + K_b\, p_b(\underline{m}) = 0$$

or equivalently

meaning independent of $\underline{m}$

$$\frac{p_s(\underline{m})}{p_b(\underline{m})} = -\frac{K_b}{K_s} = \text{CONST}$$

Q.E.D.

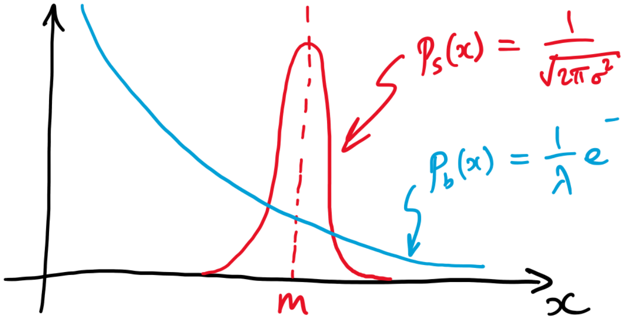
Take home messages:

If you are doing cut-and-count:

- A BDT may be easier to implement than a cut on the full likelihood ratio,
- But a BDT can never beat the full likelihood ratio — — or put another way;
- What a good BDT has to do is "discover" surfaces of constant likelihood ratio.

Toy example over page!

→



$$p_s(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

$$p_b(x) = \frac{1}{\lambda} e^{-\frac{x}{\lambda}}$$

$$\rho = \frac{p_s}{p_b} \propto \exp\left(-\frac{(x-m)^2}{2\sigma^2} + \frac{x}{\lambda}\right)$$

$$\therefore \rho = \text{const} \implies \frac{(x-m)^2}{2\sigma^2} - \frac{x}{\lambda} = \text{const}$$

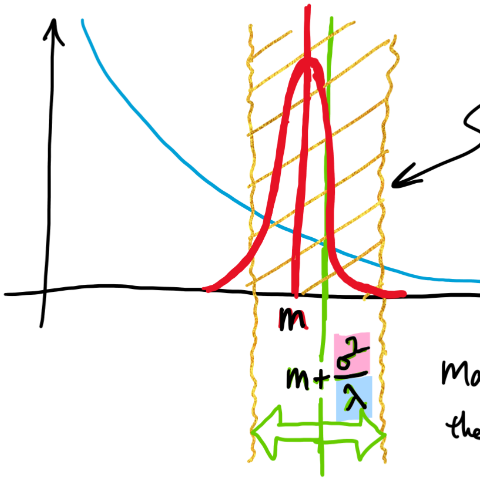
$$\implies (x-m)^2 - \frac{2\sigma^2}{\lambda}x = \text{const}$$

$$\implies x^2 - 2\left(m + \frac{\sigma^2}{\lambda}\right)x = \text{const}$$

$$\implies \left(x - \left(m + \frac{\sigma^2}{\lambda}\right)\right)^2 = \text{const}$$

$$\implies x = \left(m + \frac{\sigma^2}{\lambda}\right) \pm \text{const}$$

$\therefore$ the OPTIMAL CUT is not centred on m, but is in fact centred slightly to the right:

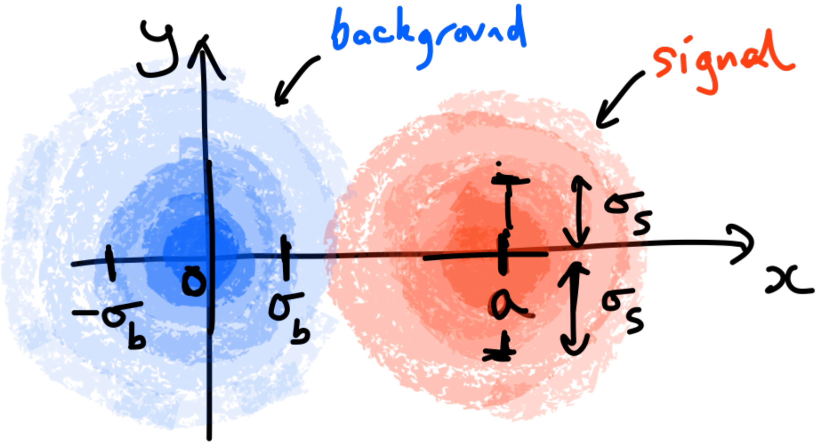


optimal selection should extend further to right than to left by $\frac{\sigma^2}{\lambda}$!

Makes sense! There is less background on the right of the signal than on its left.

Exercise :

Suppose that background events have a 2D-gaussian distribution centred on $(0,0)$ with variance σ_b in the x-direction and in the y-direction, while signal events are 2D-gaussian distributed centred on $(a,0)$ with variance σ_s in each direction. (See diagram)



In the above scenario, show that OPTIMAL CUTS are:

- Lines of constant x if $\sigma_s = \sigma_b$
- circles centred on $\left(\frac{a\sigma_b^2}{\sigma_b^2-\sigma_s^2}, 0\right)$ if $\sigma_s \neq \sigma_b$.

END