- ALL HW DUE WED in class

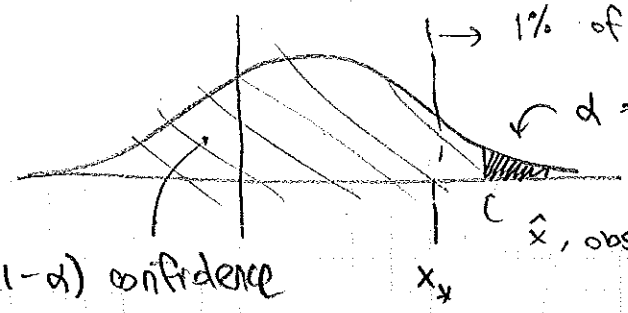
LAST TIME

LIKELIHOOD FUNCTION $\mathcal{L}(H) = P(\hat{x}|H)$ ← eg H can dep on a parameter α

↳ gave p-values, & confidence $(1-\alpha)$

$\hat{x}$: OBSERVED DATA

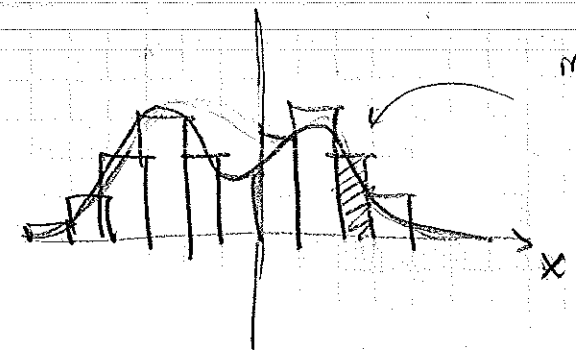
↳ IF P IS GAUSSIAN THEN σ's.



PRACTICAL MATTER: OBS. DATA IS OFTEN BINNED

↳ DISCRETIZED

just like when we talked about discretizing function space



in k^{th} BIN:

$$P(k|H) = \int_{\text{BIN } k} dx\, P(x|H)$$

PROBABILISTIC PROCESS!

eg. unstable particle has lifetime τ ... but how to measure?

$$p(t|\tau)\,dt = \frac{1}{\tau} e^{-t/\tau}\,dt \quad \leftarrow \text{FACT}$$

then take N particles & time their decays

↳ $\hat{t}_1, \hat{t}_2, \ldots, \hat{t}_N = \{\hat{t}_i\}$

$$P(\{\hat{t}_i\}|\tau) = \prod_i^N \left(\frac{1}{\tau} e^{-\hat{t}_i/\tau}\right)$$

LIKELIHOOD, $L(\tau)$

many terms! but can see

$$= \left(\frac{1}{\tau}\right)^N e^{-\sum_i^N \hat{t}_i/\tau}$$

easy to write

$$\mathcal{L}(\tau) = \log L(\tau) = \sum_i^N \left[\log\frac{1}{\tau} - \frac{\hat{t}_i}{\tau}\right]$$

same info as L

SUM EASIER THAN PRODUCT (esp computationally)

HOW TO ESTIMATE τ? $\hat{\tau}$ IS THE VALUE THAT MAXIMIZES L

$\hat{\tau}$ ESTIMATOR

$\leftrightarrow$ MAX $\mathcal{L}$

$$\leftrightarrow \frac{\partial}{\partial\tau}\mathcal{L}\Big|_{\hat{\tau}} = \sum_{i=1}^N \left(-\frac{1}{\hat{\tau}} + \frac{\hat{t}_i}{\hat{\tau}^2}\right) = \boxed{-\frac{N}{\hat{\tau}} + \frac{\sum \hat{t}_i}{\hat{\tau}^2} = 0}$$

$$\Rightarrow \boxed{\hat{\tau} = \frac{1}{N}\sum_{i=1}^N \hat{t}_i}$$ ← no surprise, the average

ACTUALLY, there's another way to approach this:

$$p(t|\tau) = \frac{1}{\tau} e^{-t/\tau} \mapsto p(t|\Gamma) = \Gamma e^{-t\Gamma}$$

$\Gamma = 1/\tau$: DECAY RATE.

EQUIVALENT INFO, DIFFERENT CHOICE OF PARAMETER.

Q: DO WE GET THE SAME ESTIMATOR

$\hat{\Gamma} = 1/\hat{\tau}$?

$$\mathcal{L} = \sum_i^N \left[\log\Gamma - \Gamma \hat{t}_i \right]$$

$$\left.\frac{\partial \mathcal{L}}{\partial \Gamma}\right|_{\hat{\Gamma}} = \frac{N}{\Gamma} - \sum_i^N \hat{t}_i = 0 \Rightarrow \hat{\Gamma} = \left[\frac{1}{N} \sum_i^N \hat{t}_i \right]^{-1}$$

$\hat{\tau}$ ✓

THIS DEMONSTRATES A GENERAL PROPERTY:

the MAXIMUM of the likelihood is the same, no matter what parameterization.

$\hat{\Gamma} = \hat{\tau}^{-1}$ in the same way that $\Gamma = \tau^{-1}$

from $\max \mathcal{L}$

BUT EVEN THOUGH THE ESTIMATORS $\hat{\Gamma}$ & $\hat{\tau}$ ARE CONSISTENT, THEY ARE NOT EQUALLY USEFUL!

eg. EXPECTATION OF THE MEASUREMENTS $\hat{t}_i$

$$\langle \hat{t}_i \rangle = \int \hat{t}_i \, \frac{1}{\tau} e^{-\hat{t}_i/\tau} \, d\hat{t}_i$$

$$= \int \left(\frac{\hat{t}_i}{\tau}\right) e^{-(\hat{t}_i/\tau)} \, d\left(\frac{\hat{t}_i}{\tau}\right) \tau$$

$$= \tau$$

⇒ EXPECTATION OF ESTIMATOR:

$$\langle \hat{\tau} \rangle = \left\langle \frac{1}{N} \sum_i^N \hat{t}_i \right\rangle = \frac{1}{N} \sum_i^N \langle \hat{t}_i \rangle = \tau \quad \checkmark$$

AGREES $\forall N$.

on the other hand.

$$\left\langle \hat{\Gamma} \right\rangle = \left\langle \frac{N}{\sum_i \hat{t}_i} \right\rangle = N \left\langle \int_0^\infty dw \prod_i^N e^{-w\hat{t}_i} \right\rangle$$

Trick: $\frac{1}{x} = \int_0^\infty dw\, e^{-wx}$ for $x > 0$

$$= N \int_0^\infty dw \left[\prod_i^N \int_0^\infty dt_i \left| e^{-wt_i} \right| \Gamma e^{-\Gamma t_i} \right]$$

(by def of $\langle \cdots \rangle$)

$$= N \int_0^\infty dw \left(\frac{\Gamma}{\Gamma + w} \right)^N$$

$$= N \Gamma^N \int_\Gamma^\infty du\, u^{-N}$$

$$= N \Gamma^N \frac{1}{1-N} u^{1-N} \Big|_{u=\Gamma}^{u=\infty}$$

$$= N \Gamma^N \cdot \frac{-1}{1-N} \left(\Gamma^{1-N} \right)$$

$$= \left[\frac{N}{N-1} \Gamma \right] \neq \Gamma \quad !!$$

↑ BIASED ESTIMATOR OF $\hat{\Gamma}$

LESSON: MAX OF LIKELIHOOD: DOESN'T MATTER WHAT PARAMETR.

BUT BIAS DOES CARE, @ LEAST FOR FINITE N.

ANOTHER EXAMPLE : GAUSSIANS

$$P(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

$$\mathcal{L} = \log L = \sum_i \left[\log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{(\hat{x}_i - \mu)^2}{2\sigma^2} \right]$$

L = $P(\{x_i\}|\mu,\sigma^2)$ ← measurements

$$\left.\frac{\partial \mathcal{L}}{\partial \mu}\right|_{\hat{\mu},\hat{\sigma}} = +\frac{1}{\hat{\sigma}^2} \sum_i (\hat{x}_i - \hat{\mu}) = 0 \Rightarrow \boxed{\hat{\mu} = \frac{1}{N}\sum_i x_i}$$

unbiased.

$$\left.\frac{\partial \mathcal{L}}{\partial \sigma^2}\right|_{\hat{\mu},\hat{\sigma}} = \sum_i \left[\sqrt{2\pi\hat{\sigma}^2} \cdot \frac{1}{\sqrt{2\pi}} \left(-\frac{1}{2}(\hat{\sigma}^2)^{-3/2}\right) + \frac{(\hat{x}_i - \hat{\mu})^2}{2(\hat{\sigma}^2)^2} \right]$$

$\left(-\frac{1}{2}\frac{1}{\hat{\sigma}^2}\right)$

$$= \frac{-N}{2\hat{\sigma}^2} + \frac{\sum_i (\hat{x}_i - \hat{\mu})^2}{2(\hat{\sigma}^2)^2} = 0$$

$$\Rightarrow \boxed{\hat{\sigma}^2 = \frac{1}{N}\sum_i (\hat{x}_i - \hat{\mu})^2}$$

BUT THIS is BIASED : $\hat{\mu}$ is built out of $\hat{x}_i$

so if we had a vector

$\underline{\hat{x}} = (\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_N)$

then $\sum_i (\hat{x}_i - \mu)$ subtracts out fluctuations along $(1,1,1,\ldots,1)$

So only fluctuations in $(N-1)$ other dir. are gaussian → VARIANCE IN LENGTH OF VECTOR IS $(N-1)\sigma^2$

↳ unbiased estimate : $S^2 = \frac{1}{N-1}\sum (x_i - \hat{\mu})^2$

A LOOK AT CONFIDENCE INTERVALS

eg measure $\hat{x}$. THIS IS THE ESTIMATOR FOR x
eg by LIKELIHOOD.

from ROOS COUSINS ApJ 95

BUT THE TRUE VALUE x_t IS PROBABLY NOT $\hat{x}$.
WHAT WINDOW AROUND $\hat{x}$ ARE YOU REASONABLY SURE THAT x_t IS IN?

eg 68% SURE THAT $x_1 < x_t < x_2$

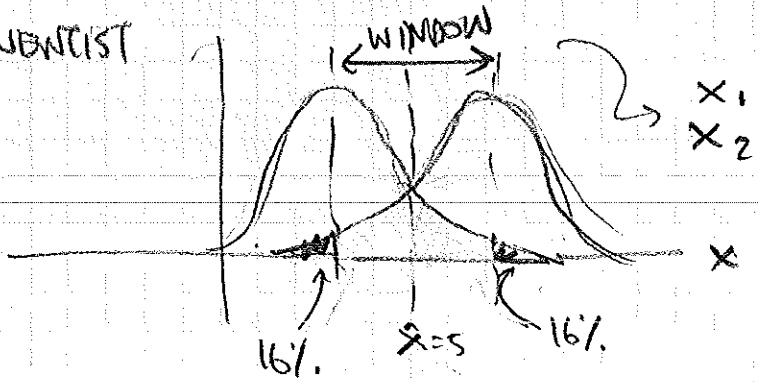
CONSIDER FIRST : GAUSSIAN

SUPPOSE $\sigma^2 = 1$
$\hat{x} = 5$ ← measured

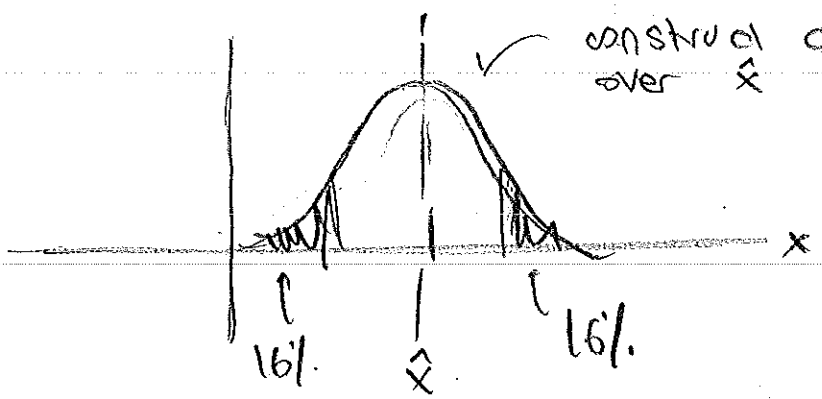
No matter what x_t is 68% of WINDOWS BY AN ENSEMBLE OF EXPTS WILL INCLUDE x_t.

(1) ONE WAY TO DO THIS $(100\% - 68\%)\frac{1}{2} = 16\%$

FREQUENTIST



USUALLY WE THINK OF THIS AS



this mental construction is justified by likelihood

$$L = p(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-x_t)^2/2\sigma^2}$$

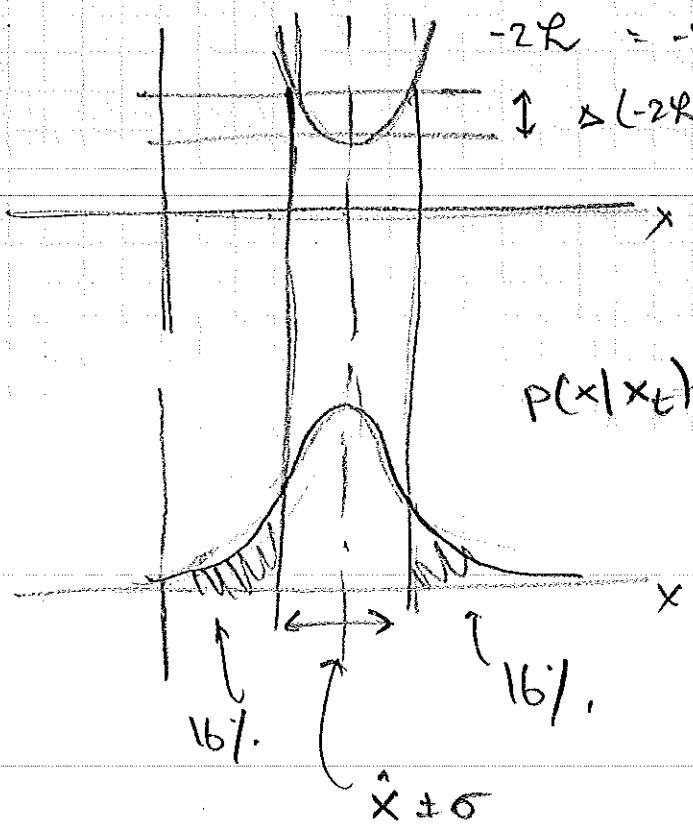
$$\mathcal{L} = \log L = \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{(x-x_t)^2}{2\sigma^2}$$ ← ASSUME σ KNOWN

to construct conf. interval, take difference w/r to maximum likelihood

USEFUL TO WORK WITH

$$-2\mathcal{L} = \frac{1}{\sigma^2}(x-x_t)^2 + (\text{indep of } x_t)$$

(2) Then: when $-2\mathcal{L}$ changes by 1 unit, that sets boundary of window
↳ $\pm\sigma$ for GAUSSIAN $p(x \mid x_t)$



FREQUENTIST: DON'T THINK ABOUT $p(x|x_t)$ AS A PDF!

so $-2\log L$ org supports "draw L" picture
↳ BUT THAT TREATS L AS A PDF

THE PROBLEM: $L(x_t)$ $\neq$ $p(x_t|\hat{x})$

$p(x_t|\hat{x})$: this is a pdf over x_t shows up as $p(x_t|\hat{x})\,dx_t$.

$L(x_t)$: nothing about this is a pdf

$L = p(\hat{x}|x_t)$ ← pdf w/r.t $\hat{x}$ BUT NOT x_t will NEVER see $p(\hat{x}|x_t)(dx_t)$

... at least as FREQUENTIST.

AS A BAYESIAN:

$$p(x_t|\hat{x}) = \underbrace{p(\hat{x}|x_t)}_{L(x_t)}\ p(x_t) \Big/ p(\hat{x})$$

this prior is a pdf in x_t

$p(\hat{x}) = 1$

maybe uniform! (subj to caveat: what parameterize?)

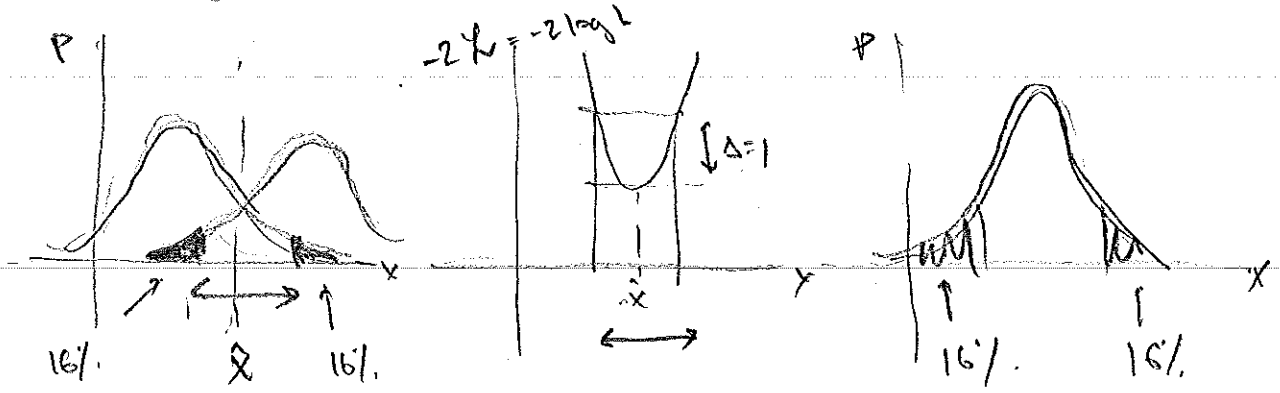
effectively →

$$p(x_t|\hat{x}) = p(\hat{x}|x_t)\,p(x_t) \Big/ \int dx_t\ p(\hat{x}|x_t)p(x_t)$$

normalize

only makes sense wrt $p(x_t)$

& 3 eg's of CONFIDENCE INTERVAL



FREQUENTIST

PROFILE w/ LIKELIHOOD

BAYES w/ FLAT PRIOR

WHY THIS MATTERS · things are more subtle when the pdf is not Gaussian.

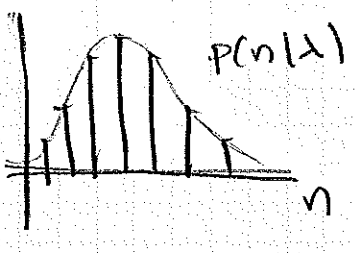
eg POISSON STATISTICS

$$P(n|\lambda) = \frac{1}{n!} e^{-\lambda} \lambda^n$$

n events; λ expected

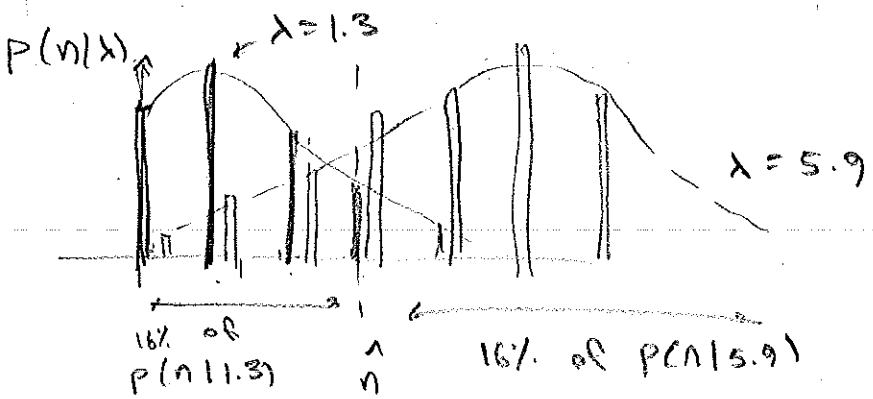
eg rare decay → PROBABILISTIC; "WOULD BE BINOMIAL" BUT I DON'T HAVE A SENSE OF N OF TRIALS.

OBS: DISCRETE AS A "DISTRIBUTION" IN n



SUPPOSE YOU MEASURE $\hat{n} = 3$ events.
WANT TO GIVE CONFIDENCE INTERVAL FOR λ.

① FREQUENTIST : 16% BELOW, 16% ABOVE



(2) LIKELIHOOD RATIO

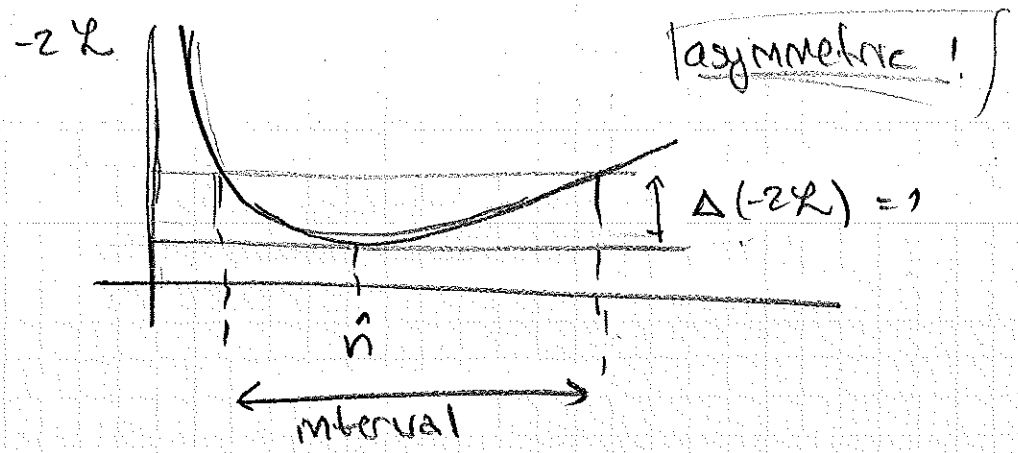
$$L(\lambda) = \frac{\lambda^{\hat{n}}}{\hat{n}!} e^{-\lambda}$$

← continuous function of λ; still not a pdf

$P(\hat{n}|\lambda)$ ← nothing about this allows integration over $d\lambda$

Not Gaussian! no longer any sense that $\Delta(-2 \log L) = \Delta(-2\mathcal{L}) \Leftrightarrow 1\sigma$

BUT CAN STILL PLOT & USE these guidelines



WHAT IF WE'RE BAYESIAN?

STICK TO FLAT PRIOR: $P(\lambda) = \text{const}$.

↳ non-flat changes things a lot — but let's see how even a flat prior changes the analysis

now we have:

$$P(\lambda|\hat{n}) = \frac{P(\hat{n}|\lambda)\, P(\lambda)}{N}\, d\lambda$$

$P(\hat{n}|\lambda)$: $L(\lambda)$ not a pdf in λ; $P(\lambda)$: this is a pdf in λ

$P(\lambda|\hat{n})\,d\lambda$: this is a pdf

N: normaliz.

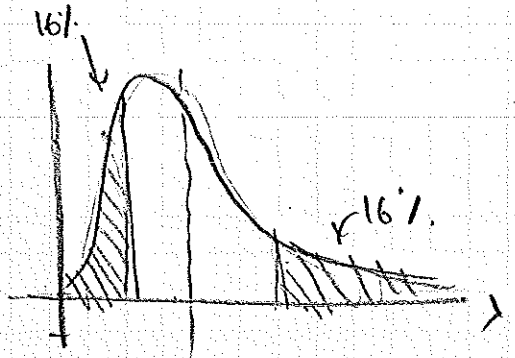
WE'RE STILL LOOKING FOR 68% CONFIDENCE

↳ some window around $\hat{n}$ where true value λ is 68% likely to fall into

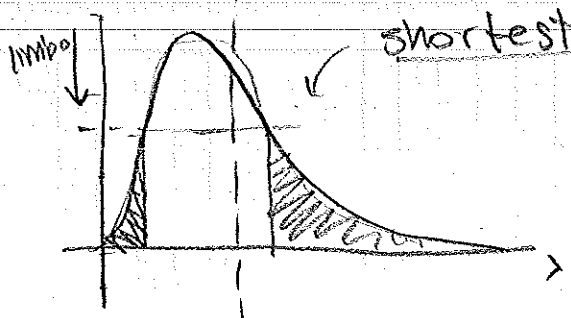
↑ if we applied this procedure to many different labs, the true value λ will fall into the windows of 68% of them.

BUT $p(\lambda|\hat{n})$ is asymmetric... so we have a choice of how to do this

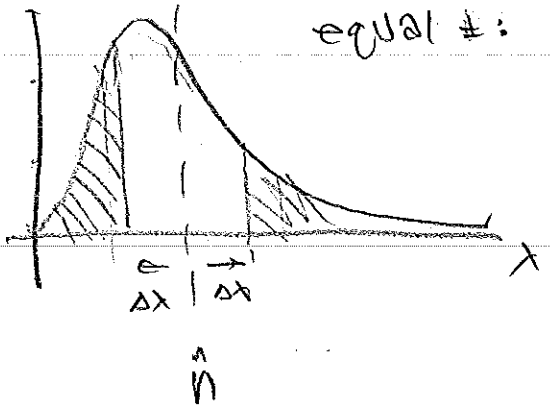
↳ BOTTOM LINE: SAY EXACTLY WHAT YOU DID



central: 16% to left, 16% to right



shortest: $p(\lambda|\hat{n})$ outside interval should be less than $p(\lambda|\hat{n})$ inside!



equal ±: interval is $\hat{n} \pm \delta\lambda$

this is just to illustrate some of the nuances when giving confidence intervals

→ to highlight the operational differences between FREQUENTIST & BAYESIAN.

seems kind of academic, until you have an experiment that says

"the best fit for the neutrino mass is $m^2 = (-50 \pm 30)\ eV^2$ "

↑

m^2 cannot be negative !!
the theory upon which these measurements were fit assume $m > 0$.

so HERE A BAYESIAN PRIOR OF

$$P(m^2) = \Theta(m^2)$$

MAY BE USEFUL.