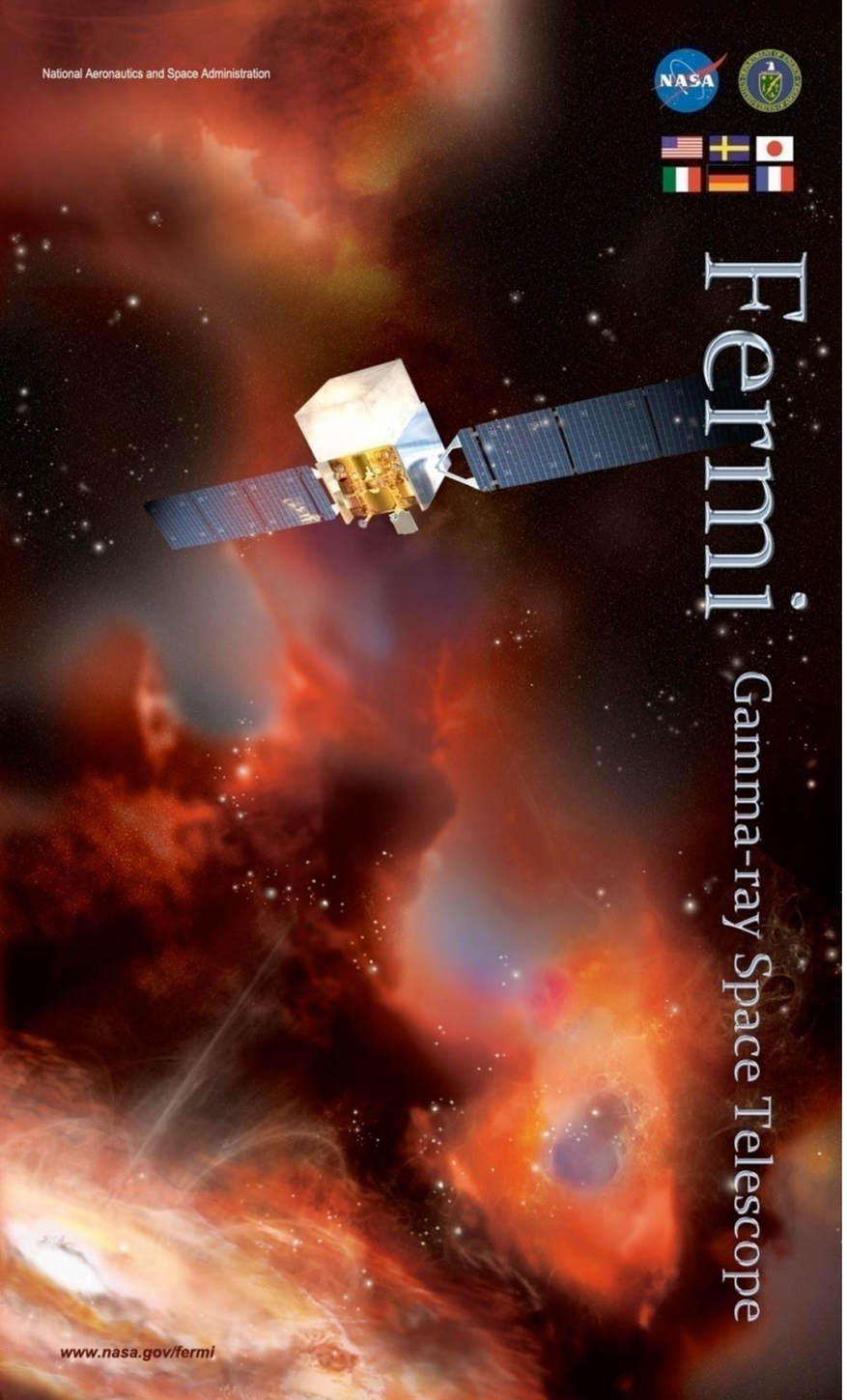


Introduction to Likelihood: Fundamentals to Fermi

Matthew Kerr
Naval Research Lab

2023 Fermi Summer School
Lewes, DE

Philosophical Introduction: What is a Measurement?

- Think back to some undergrad lab: you measured the period of a pendulum, or the temperature of some bath, or the charge on a capacitor…
  - On one level, it was *data*: numbers (with units).
  - But really, it was much more: it was a particular instrument, a technique, an operator, an observer, a recording method… It was a whole *system* that produced a number.

- ***There is no understanding of data without understanding how it was obtained!***

- *Likelihood* is a technique which uses the laws of probability to encapsulate
  - The data.
  - The way the data are distributed intrinsically.
  - The way the "true" distribution is altered by the observing system.

- **Likelihood underpins many (maybe most) modern astrophysical analyses.**

Our Toy Example: The New Baker

- You love baguettes. Love them. And you have purchased your daily bread from a neighborhood baker for many years. Alas! The baker recently retired. You try a baguette from the new baker. Mmm, tasty. But wait! It just seems ever so slightly lighter… HAS THE NEW BAKER SHRUNK MY PRECIOUS BAGUETTES???

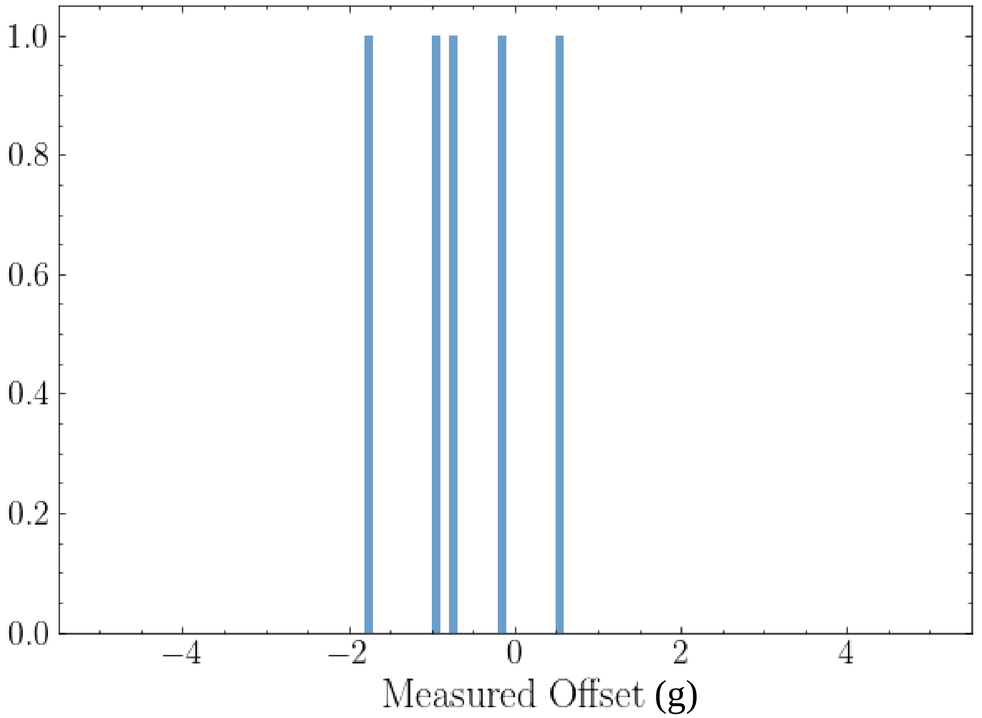
  The baker swears that they measure out the ingredients by mass just like the old baker: each baguette is 100g! You demur, but decide to put it to the test.

- You put in a rush order at your local instrument shop and on Sunday you are the proud new owner of a scale that claims to be accurate to +\- 1g. Next week, you weigh every baguette, obtaining these data:

Monday	Tuesday	Wednesday	Thursday	Friday
98.30g	99.90g	99.24g	100.54g	99.10g

- Oooh, it's looking like SHRINKFLATION!

- But: how sure are you that the baguettes **actually** weigh less than 100g?



Likelihood to the Rescue

- First, we need to understand how our scale is affecting our measurements. Let's assume that it is accurate, and that "+/- 1g" means "measurements will be normally distributed with a standard deviation of 1g". **NOW** we can say something more concrete, because we have a *probability density function* describing how our data should come out as we make a series of measurements: for a given baguette, our *model* says that

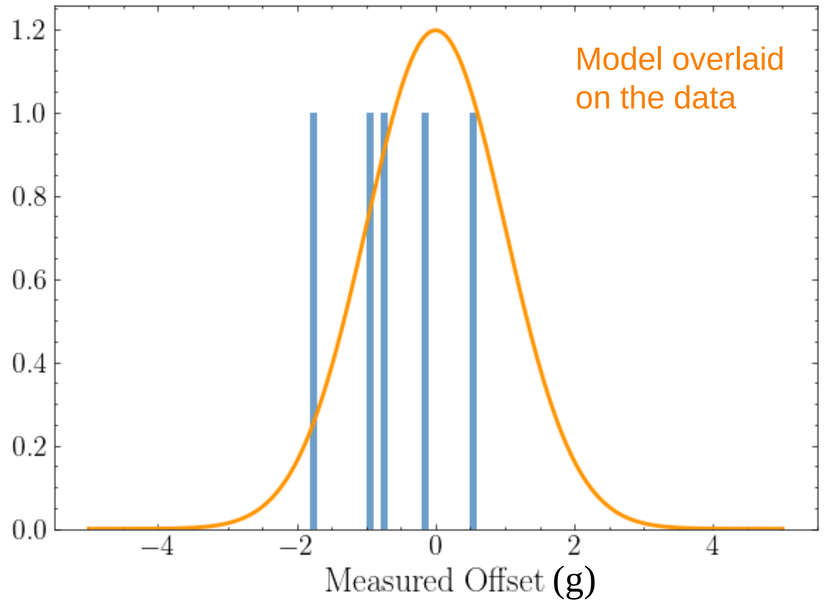
Probability of measuring a baguette mass m ***given*** the model parameters μ, σ.

Standard deviation σ, **assumed** to be 1g

Mean μ, **assumed** to be 100g

$$p(m|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(m-\mu)^2}{\sigma^2}}$$

2pis don't matter much, but I have them here for now for completeness

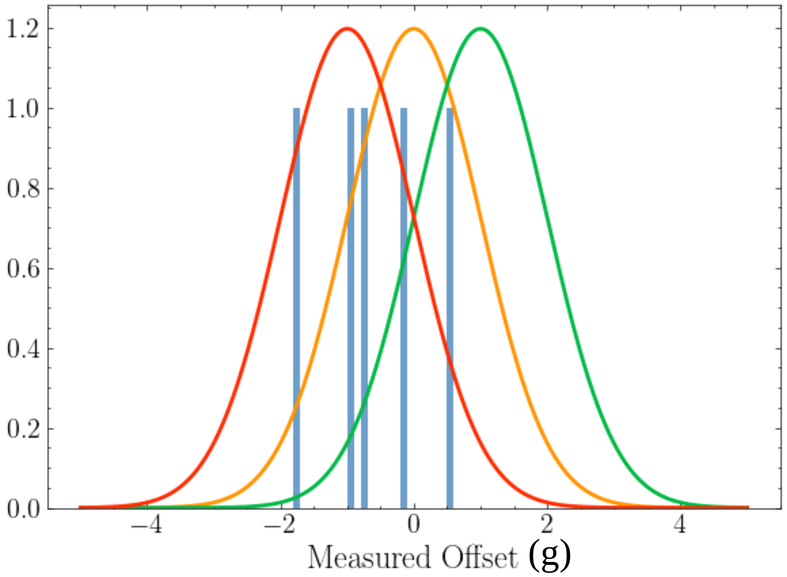


- This model tells us about the range of measurements we can make. But we want to know: what is the true mass of these baguettes? Thus, we turn the problem around: if data we observe have a low probability of having occurred, the model might not be right. Conversely, if they have a high probability, we have confidence in the model.

- Going further, we can *adjust* the model (e.g. the mean) so that the probability of the data becomes higher (or lower). Given the data, evaluating the probability as a function of model parameters is called the likelihood.
  - It is *exactly* the probability density function, but with data substituted for random variables.
  - This changes the interpretation. Likelihood is a function of *model parameters.*

Parameter Estimation: Maximum Likelihood

- Here are the same data with 3 different models applied (a shift in the mean of 1g):
  - The red model (mean of 99g) puts more of the measurements in a high probability region. However, we can't make the mean too low, because the gaussian shape falls off rapidly, and the low probability assigned to the largest measurement will eventually "win" out over the high probability associated with the lower ones.
- Instead of considering individual models, we can let μ take on a continuous range from -5g to 5g. The likelihood as a function of μ is then:



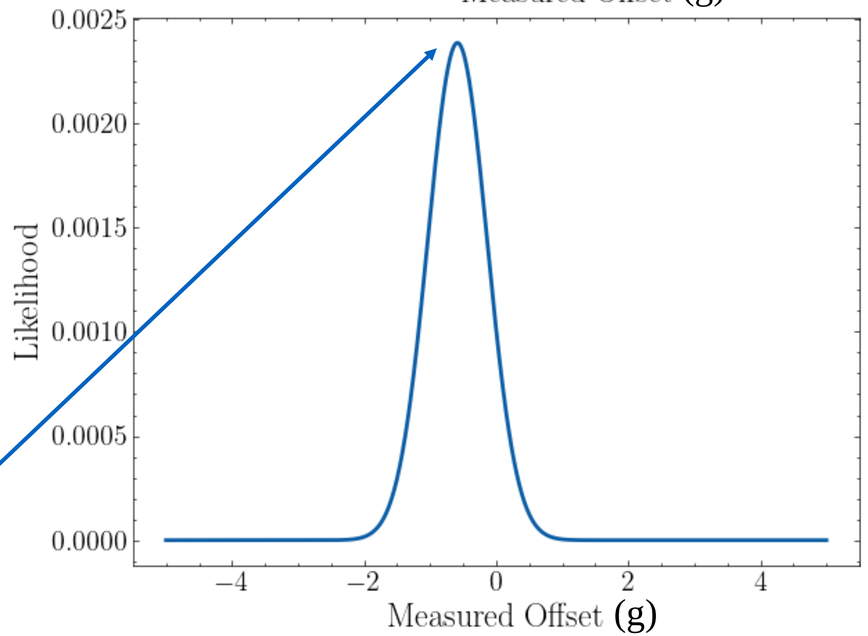
Monday	Tuesday	Wednesday	Thursday	Friday
98.30g	99.90g	99.24g	100.54g	99.10g

$$L(\mu|\text{data}) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(98.30g-\mu)^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(99.90g-\mu)^2}{2\sigma^2}} \ldots \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(99.10g-\mu)^2}{2\sigma^2}}$$

$$= \prod_{i=1}^{i=5} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(m_i-\mu)^2}{2\sigma^2}}$$

*Writing the probability as a product like this is done when the data are independent.

This value of the mean μ corresponds to the model that assigns the highest probability to the data we observed.



It is the **MAXIMUM LIKELIHOOD ESTIMATOR** for the mean.

Parameter Estimation: Maximum Likelihood

- Rather than numerically determine the maximum likelihood, let's brush up our calculus. First, it's almost universal to use the "log likelihood" because
  - the probabilities we deal with can be quite small (but are always positive); logs mitigate numerical issues
  - max(L) occurs at the same place as max(log(L))
  - the natural log greatly simplifies gaussian likelihoods, which come up very frequently. Thus:

$$\log L(\mu|\text{data}) = -\sum_{i=1}^{5} \frac{(m_i - \mu)^2}{2\sigma^2} - \log\sigma + \text{constants}$$

$$\frac{\partial \log L}{\partial \mu} = \sum_{i=1}^{5} \frac{(m_i - \mu)}{\sigma^2} \quad \rightarrow \quad 0 = \sum_{i=1}^{5} \frac{(m_i - \mu)}{\sigma^2} \quad \rightarrow \quad \hat{\mu} = \frac{1}{5}\sum_{i=1}^{5} m_i = \bar{m} = 99.41\text{g}$$

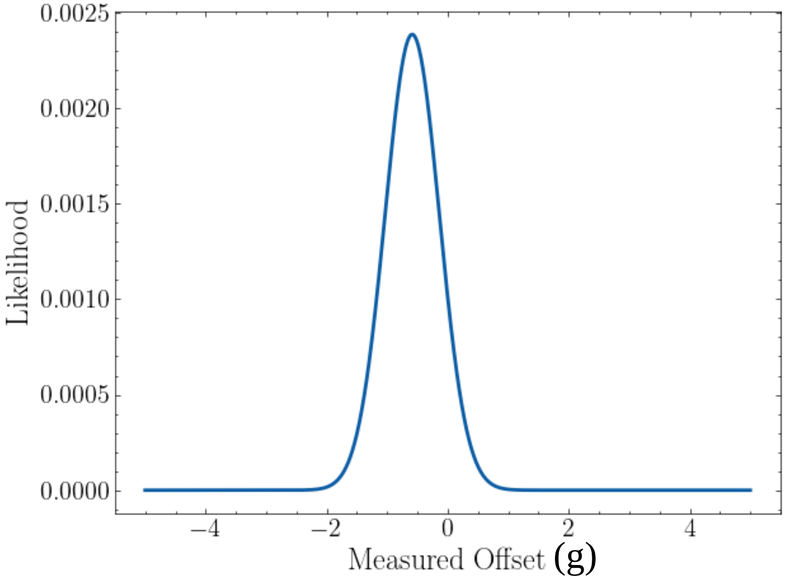
- The maximum likelihood estimator for the mean is just the "normal formula" for the sample mean.
  - According to this, it sure looks like the baker is shortchanging us by 0.6g!
  - But because we're good scientists, we need to ask:
    - (1) What is the uncertainty on the maximum likelihood estimate? Or put another way,
    - (2) With what confidence can we rule out the "null hypothesis" (mean=100g)?

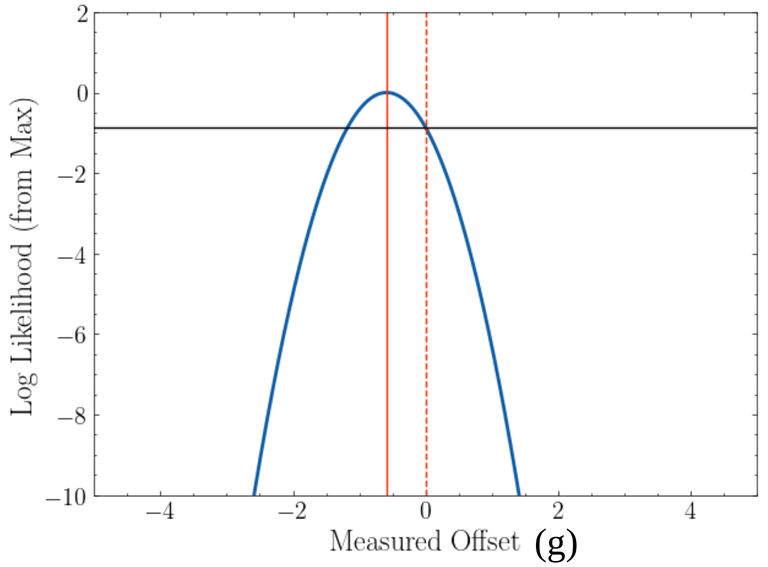
Approach I: Parameter Uncertainty Estimation

- We noted that the likelihood has a maximum at the sample mean, and that it falls off (the probability to observe the data decreases) as the model μ changes.
  - Intuitively, the rate at which the likelihood falls off must be related to the uncertainty. If it falls off very quickly, we feel comfortable ruling out μs which are far away. If it falls off slowly, then we have to consider that we don't have enough data to really pin it down.
- Again, it's better to work with the log likelihood. How can we encapsulate "rate of falloff"?
  - The maximum value itself doesn't mean much.
  - At the maximum, the derivative is always 0!
  - What about the 2nd derivative? Remember the 2nd derivative is related to curvature, and note that it has units mass^{-2}. So does the variance!
  - Proposal: $\text{var}^{-1}(\hat{\mu}) \equiv -\frac{\partial^2 \log L}{\partial \mu^2}$.

$$-\frac{\partial^2 \log L}{\partial \mu^2} = \sum_{i=1}^{5} \frac{1}{\sigma^2} = \frac{5}{\sigma^2} \rightarrow \quad \text{std}(\hat{\mu}) = \frac{\sigma}{\sqrt{\text{N}}}$$

  - This makes intuitive sense: it involves the precision of our scale (σ), and it improves as we make more measurements ($\propto \sqrt{N}$), which should ring some bells!
  - Thus, we obtain $\hat{\mu} = 99.41 \pm \frac{1}{\sqrt{5}}\text{g} = 99.41 \pm 0.45\text{g}$.





Matthew:
Breeze Over This!

Parameter Uncertainty: Justification and Generalization

- This approach isn't just a guess! It's related to the Fisher information, which is the expectation of the 2nd derivative of the log likelihood. For general parameters α (and subject to some conditions…)

$$F_{ij} = < \frac{\partial^2 \log L}{\partial \alpha_i \partial \alpha_j} >$$

- (1) the Cramer-Rao lower bound tells us that the uncertainty on an unbiased parameter cannot be less than the inverse Fisher information.
  - MLEs are generally "efficient", i.e. they meet the Cramer-Rao bound.
- (2) Asymptotically, MLEs are distributed as a normal variable whose covariance matrix is the inverse Fisher information.
- Putting these together motivates *a general purpose technique for estimating uncertainty: approximate the Fisher information from the "hessian", the matrix of 2nd derivatives, and invert it*. So maximum likelihood estimation boils down to
  - Finding the maximum (often using 1st derivatives) to find the parameters alpha
  - Evaluating the hessian (2nd derivatives) to estimate the covariance C_{ij} for the parameters
- This approach is basically approximating the likelihood as being gaussian, so it may fail badly in cases where that isn't true: very non-gaussian model, few data.
  - When in doubt, MC it out: simulate data according to your model and look at the MLE distribution.

Approach II: Hypothesis Testing / Model Comparison

- It might also be nice if we could estimate some kind of probability of being wrong.
  - The null hypothesis is $\mu = \mu_0 = 100\mathrm{g}$, while one alternative hypothesis is $\mu = \hat{\mu}\ 99.41\mathrm{g}$.

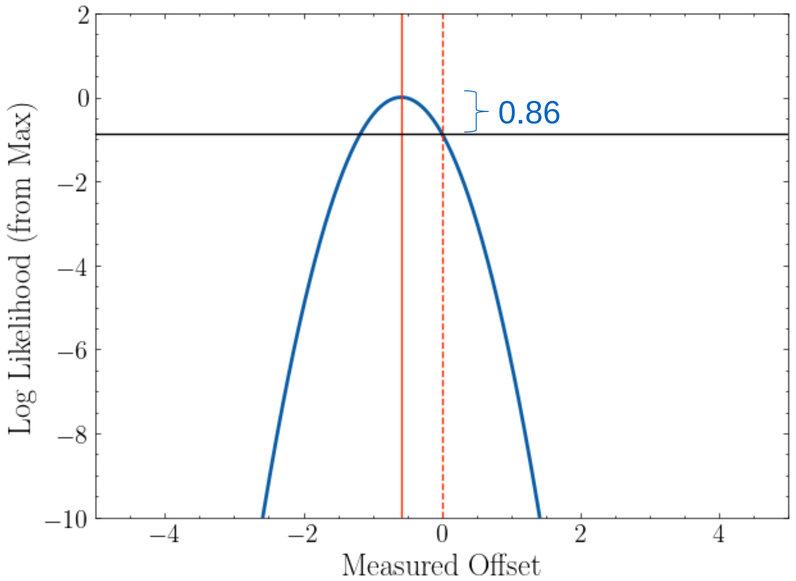
- Experiment has 4 possible outcomes:
  - NH is correct and we accept it. (Baguettes are the same as before, we move on with our life.)
  - NH is correct but we reject it! Type I error. (Baguettes are the same, we raise a fuss, the baker gets very upset.)
  - AH is correct and we accept it. (Baguettes changed, we report it, the baker fixes their scales.)
  - AH is correct but we reject it. Type II error. (Baguettes are lighter, but we don't realize/report it.)

- Experiments are designed to produce some probability of Type I and Type II error based on the "importance" of the result and available experimental resources.
  - Reducing errors requires more data, more observing time, more money, etc. But if the result is really important, you will expend the resources. Tradeoff.
  - Generally, people are more concerned about Type I errors because they tend to result in fake "discoveries".
    - In astronomy and astrophysics, these are also the ones we care most about. A Type I error might result in an erratum, awkward questions at a conference, a reputation for sloppy work…
    - With a Type II error, maybe you fail to find something that would produce a Nobel prize. But more likely, someone will come along in ten years with a bigger telescope and they'll get the result and clear up the record.

Approach II: Hypothesis Testing / Model Comparison

- The likelihood tells us how to compare probability for *data*, not for models. However, it stands to reason that models that predict low probabilities are less good. Can we use that to gauge our probability of Type I error?

- The *Likelihood Ratio Test* compares the likelihoods for two different hypotheses. For us,
  - Remember, the null hypothesis is $\mu = \mu_0 = 100\text{g}$, while
  - The alternative hypothesis is $\mu = \hat{\mu}\ 99.41\text{g}$.
  - Again, it's nice to use logs. Going back to our definition, we find $\log L(\mu = \hat{\mu}) - \log L(\mu = \mu_0) = 0.86$. This means the data are about 2.4x more likely under the alternative hypothesis. How significant is it?



- This LRT is a *statistic* (based on the data).
  - It will have two different distributions, depending on whether the NH or the AH is correct.
    - (It is also possible neither is correct!)
  - Since we are most worried about Type I error, we want to know the "***distribution of the LRT in the null hypothesis***." If we know that, we can calculate the probability of finding any value of the LRT and thus the chance of making a Type I error.
  - Unfortunately, there isn't a general rule for this.
    - Fortunately, there is a reasonably broad class of models where we DO know the distribution.

Wilks' Theorem

- Nested model: the null model can be obtained by setting (some) parameters of the alternative model to a specific fixed value.
  - In our toy model, the null model is obtained by fixing $\mu = \mu_0 = 100\mathrm{g}$. They are nested.
- If a model is nested, and IF every parameter is well defined in the null hypothesis (more on this later!), then *asymptotically*

$$2(\log L_{\mathrm{alt}} - \log L_0) \equiv \mathrm{TS} \sim \chi^2_n.$$

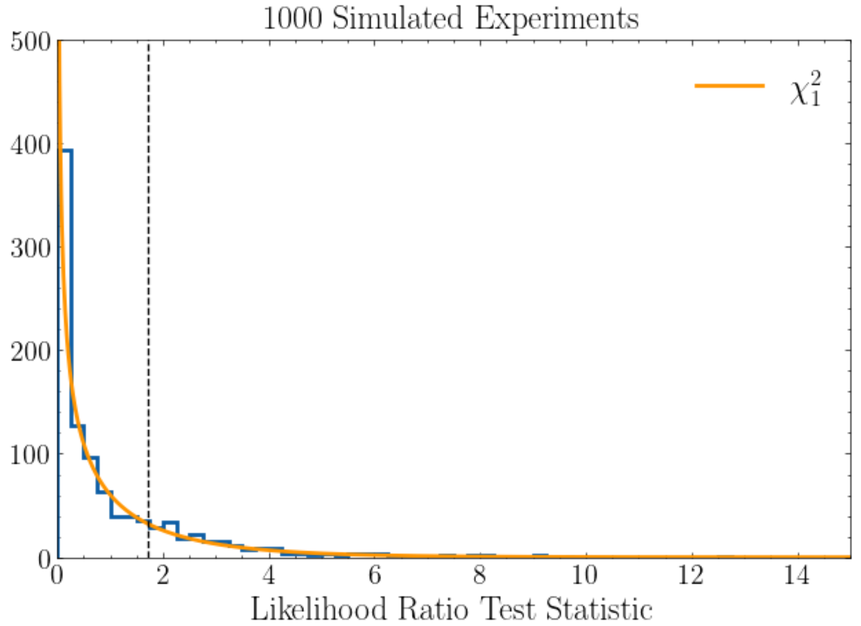
This defines TS, the "Test Statistic" for the likelihood ratio test, and states that *if the null hypothesis is true,* then TS will be distributed as chi-square variable with *n* degrees of freedom. Here, *n* is also the difference in the number of free parameters between the models.

  - For gaussians, TS is actually typically called "chi squared", so this is why that test (sometimes) works.

- In our case, *n* is 1, and the TS=1.72. We can now use the chi-square distribution to ask "what is the probability of getting a value this large or larger?" If it's a low probability (TS is large), we would *reject* the null hypothesis with good confidence.
  -
```
from scipy.stats import chi2
chi2.sf(1.72,1)
0.1896930449612064 3
```

- Thus, there's a 19% chance of seeing a TS this large or larger by chance even if the null hypothesis if true. That's a relatively large probability of making a Type I error.

- Caveat: results are asymptotic and the LRT based on less-than-infinite data may differ.



Back to Baguettes

- So, with only one week of data:
  - Using Approach 1, we measured $\hat{\mu} = 99.41 \pm 0.45\text{g}$. That's <2$\sigma$ from 100g.
  - Using Approach 2, Wilks' theorem told us the probability of Type I error (incorrectly rejecting null hypothesis) was 19%.
- Conclusion: let's not be too hasty!

- I hope it's not too much of a stretch to see parallels between this and some research projects… It's important to know when not to push borderline results too far! (And how to determine whether or not they **are** borderline.)

- Question time:
  - Suppose we collect 1 month (4 weeks) of data. How much smaller will our uncertainty be?

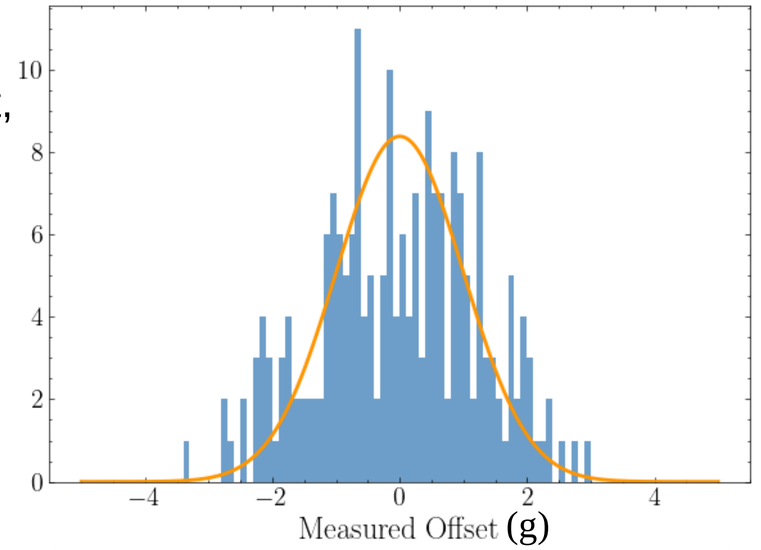
Uncertainty Models: Intrinsic and Experimental

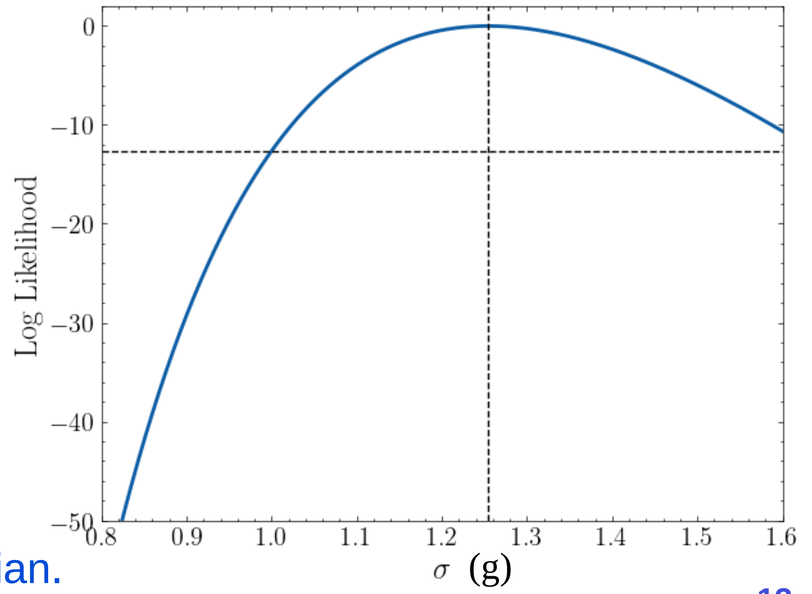
- Let's suppose we keep on taking data for a year: 210 measured baguettes (!)
- The mean still looks close to 100g, but some of the measured values are rather far out, around 3σ. This is unlikely to happen by chance if our model is OK.
  - Maybe our error model isn't quite right. We assumed our scale would produce a gaussian distribution with std=1g, but maybe the manufacturer meant something else, or maybe the distribution isn't normal at all.
  - Or, maybe the problem is in the baker's scale and the error is intrinsic to the data.
- Likelihood can tackle this too! We just need to use a more flexible model.
  - The simplest possible thing we can do is make σ a free parameter:

$$\log L(\mu, \sigma|\text{data}) = -\sum_{i=1}^{5} \frac{(m_i - \mu)^2}{2\sigma^2} - \log\sigma + \text{constants}$$

$$\frac{\partial \log L}{\partial \sigma} = \sum_{i=1}^{5} \frac{(m_i - \mu)^2}{\sigma^3} - \frac{1}{\sigma} \rightarrow \hat{\sigma} = \sqrt{\frac{1}{5}\sum_{i=1}^{5}(m_i - \mu)^2} = \text{std}(m) = 1.25\text{g}$$

- As before, we see that the maximum likelihood estimator for sigma gives us the "usual" formula for the population standard deviation.
  - NB: we have to choose our model: is μ a free parameter, or is it fixed to 100g?
  - Most generally you would optimize both at the same time, and calculate the hessian.



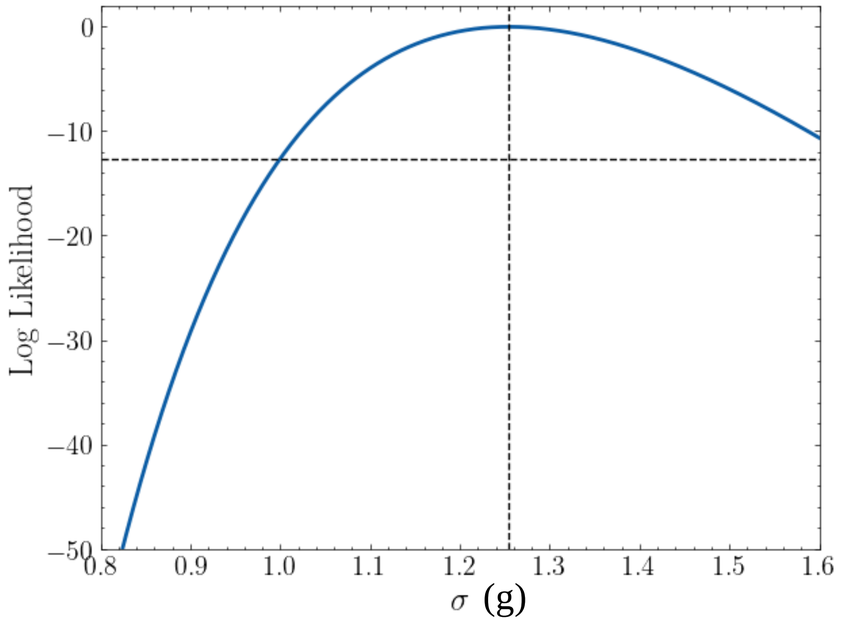


Significance and Interpretation

- First, let's estimate the uncertainty:

$$\frac{\partial^2 \log L}{\partial \sigma^2} = \sum_{i=1}^{5} \frac{-3(m_i - \mu)^2}{\sigma^4} + \frac{1}{\sigma^2} \rightarrow \quad \mathrm{std}(\hat{\sigma}) = \frac{\hat{\sigma}}{\sqrt{2\mathrm{N}}}$$

This equation has an interesting form: the **fractional** error goes like $1/\sqrt{2N}$, so we can quickly estimate a significance for N=210: 4.8%. The MLE we calculated, 1.25g, differs by about 25%, so this looks like it's about 5σ away! This is a **really useful thing to remember** when you're thinking about experiments: how much data do you need to measure something vs. how much do you need to **calibrate** your experiment if you can't "trust" it.



  - NB that the actually uncertainty needs to use the measured value, so is 1.25 ± 0.06g.
  - Also, note asymmetry in log likelihood: remember that this is an approximation based on the curvature.

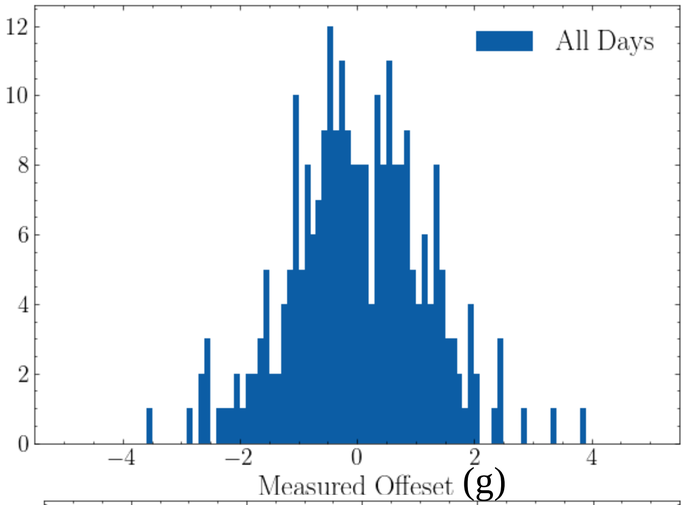
- Now, what about Wilks' Theorem? Is the model nested?
  - $\log L\,(\sigma = \hat{\sigma}) - \log L\,(\sigma = \sigma_0) = -12.63.$
  - $P(\chi_1^2 \geq 2 \times 12.63) = 5 \times 10^{-7}$. (about "5 sigma")

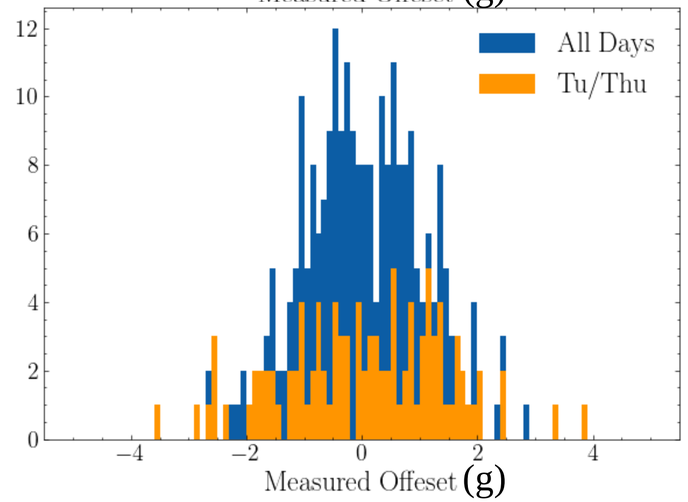
General point: very elaborate "uncertainty models" can be used to characterize both intrinsic and experimental error sources. But be careful in your assumptions: not everything is a gaussian!

- So, we conclude that the alternative model with $\sigma = 1.25$g is a much better description of the data and we **reject the null hypothesis.** Can we make any conclusions about the source of the additional uncertainty?

Joint Likelihood

- Sometimes, we have different classes of data. Suppose our total data looks something like this.
  - Again, 210 measurements, and again we see some outliers relative to our null model.
- Now, we go through our log book and separate the measurements by day. We plot some different combinations, and M/W/F seem to follow our null distribution, while Tu/Th seem to be broader.
  - Be careful! If you try enough combinations, eventually you'll find something unusual by chance.
  - It's best if you have physical reasoning: e.g., you know a different crew bakes on Tu & Thu!
- If we have data with distinct properties, we can use "joint likelihood": we use a model that is appropriate for each part of the data.
  - This technique is very valuable, because often some properties are universal, so you can use both sets of data to make inferences. (Think "Front"/"Back" for Fermi, or PSF0…PSF3)
- Here, let's adopt a model where μ is linked between the sets, and we assume that the M/W/F data follow the null hypothesis ($\sigma = 1$) while the Tu/Thu data have $\sigma \neq 1$. Here is the likelihood:





$$\log L_{tot} = \log L_{\mathrm{MWF}} + \log L_{\mathrm{TTh}} = \sum_{i\ in\ "MWF"} -0.5\frac{(m_i-\mu)^2}{\sigma_{\mathrm{MWF}}^2} - \log\sigma_{\mathrm{MWF}} + \sum_{i\ in\ "TTh"} -0.5\frac{(m_i-\mu)^2}{\sigma_{\mathrm{TTh}}^2} - \log\sigma_{\mathrm{TTh}}$$

- You can see by inspection that:
  - The MLE for σ_{TTh} only involves T/Th data: it's just the std!
  - On the other hand, μ is "linked", so all data contribute to its measurement.
  - The models are nested (we recover the NH by setting $\sigma_{\mathrm{TTh}} = 1$), and we can use Wilks' Theorem to estimate the significance. Conclusion: the Tu/Th crew are sloppier with their measurements!

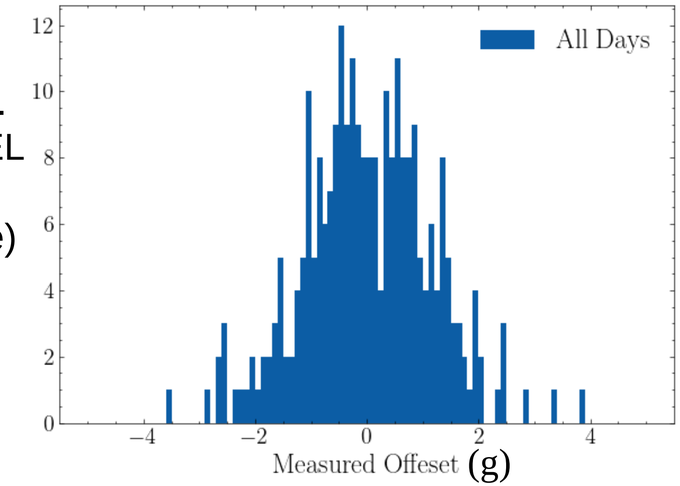
Mixture Model (I)

- Now, let's consider an example where we **can't** use Wilks' Theorem. Let's take the same data, but, due to a bookkeeping error, we accidentally deleted all of our data about days of the week. Oops!
- The distribution is still broader than our null model. As before we could fit an overall σ, but suppose we know there are two bakery crews, and we want to test an idea similar to that on the previous slide. Now, rather than having a mixture of two types of data, we can make a mixture of two types of MODEL

  Specifically, we suppose that for any given baguette, theirs is a probability f from one of two (or more) different probability distributions. This is a **mixture model**:

$$P(m|\sigma_1, \mu_1, \sigma_2, \mu_2) = f \times n_1(m|\sigma_1, \mu_1) + (1 - f) \times n_2(m|\sigma_2, \mu_2)$$

  where we are denoting the normal distributions as n_1 and n_2.



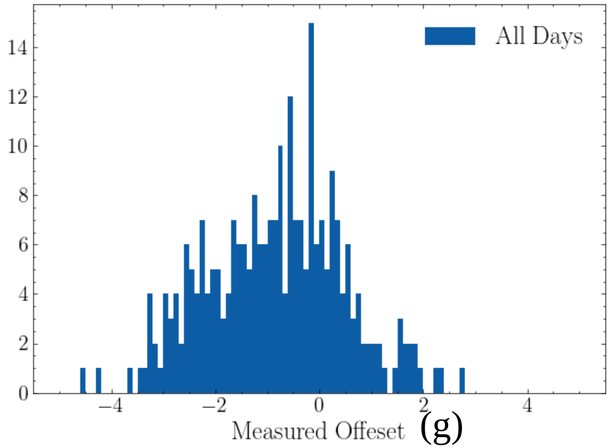
- This model has 5 parameters, and we can make a whole family of nested models by assuming e.g. $\mu_1 = \mu_2, \mu_1 = \mu_0 \ldots$ If we assume $\mu_1 = \mu_2 = \mu_0$, $\sigma_1 = \sigma_2 = \sigma_0$, and $f = 1$, we recover our original null hypothesis.
- We can go through and make likelihood calculations as before, computing estimators, estimator uncertainties, and LRT test statistics.
- **Although all of these models are nested, *many pairs do not satisfy the criteria for Wilks' Theorem!***
  - Wilks' Theorem requires that all parameters be well-defined "in the null hypothesis". What does this mean? An example:
  - If we choose a null model with $f = 1$, there is no 2nd distribution, so σ_2 and μ_2 no have impact on the model. ANY value of σ_2 and μ_2 would produce the same P(m). They are **degenerate**, and thus Wilks' Theorem does not apply!
    - Maximum likelihood estimators are still fine, the LRT still exists. You just don't necessarily know what it means.
  - Other cases are fine. E.g., if you had a hypothesis about the Tu/Th team, you could **fix** f to be 3/5, then all of the models you explored would be OK for WT. (But you would need to be careful that your assumptions were OK.)

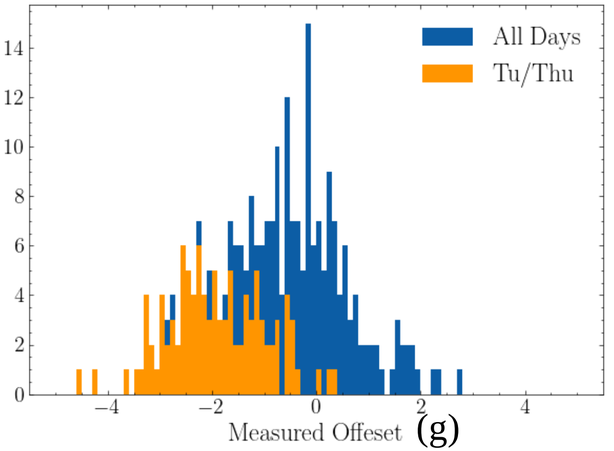
Mixture Model (II)

- Mixture distributions are often used to describe NEW COMPONENTS (read: new astrophysical sources) and the literature is **CHOCK FULL** of people who incorrectly cite Wilks' Theorem.
  - This is true for Fermi-LAT analysis, too!
- Consider this unlabeled data: it looks a little biased – the mean is definitely < 100g.
  - We could model this very simply as a single gaussian with a mu (and/or sigma).
  - Looking at the labeled data, we clearly see what's going on: The Tu/Thu crew are baking lighter baguettes! IF we had these labels, we should do a joint likelihood with linked σ (our scale is the same) and then estimate the two means.
- If we don't have those labels, we can still look for a "new source" of lighter baguettes with a mixture model:

$$\log L = \sum_i \log\left\{\frac{f}{\sigma}\exp\left(-\frac{(m_i-\mu_1)^2}{2\sigma^2}\right)+\frac{1-f}{\sigma}\exp\left(-\frac{(m_i-\mu_2)^2}{2\sigma^2}\right)\right\}$$

- NB one negative development: the log can't apply to both exponentials! Makes things a lot harder.
  - Generally you can't solve for MLEs analytically. However, numerical optimization still likes derivatives.
- The relative "strength" of the sources is parameterized by f. Source 1 emits baguettes of mass μ_1, while sources 2 emits baguettes of mass μ_2, and we measure each source with our "baguette camera" with a resolution 1g (normal distribution).
- Now, these sources are rather "close together", barely within the capability of our "camera" to resolve. So we'd like to test the two-source hypothesis against the one-source hypothesis.
- Unfortunately, Wilks' Theorem does not apply! In the null hypothesis, $f = 1$, so μ_2 is degenerate. (I.e. the log likelihood doesn't change no matter what value of μ_2 we pick, in the null hypothesis.)

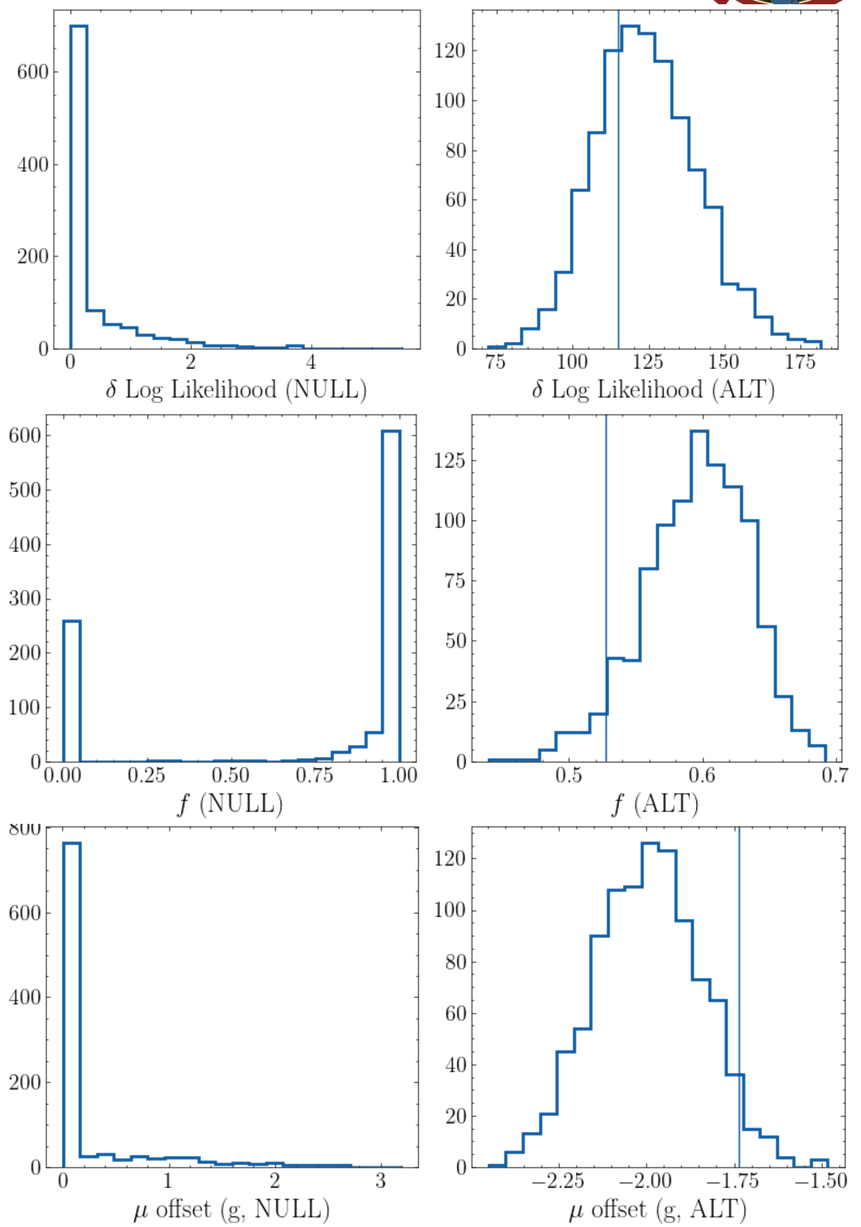




Matthew:
Breeze Over This!

Checking with Monte Carlo

- But, we can still calculate the maximum likelihood estimators and the LRT (Test Statistic). To determine the uncertainties and the significance, we need to do simulations.
  - Simulations of the null hypothesis let us determine the significance.
  - Simulations of the alternative hypothesis let us gauge uncertainties and do sanity checks.
    - NB that simulations of the alternative hypothesis can be a good idea in other cases, too, because we aren't guaranteed the hessian is a good approximation of the parameter uncertainty.
- Here are simulations in each hypothesis along with the MLEs. (I'll post the code if you want it.) The observed values from the data set are shown as vertical lines.
  - The LRT is *very* significant, TS~240. We could have skipped the sims.
  - The MLE distributions are approximately gaussian but have slightly longer tails.
- Caveat: simulations can be more accurate than the analytic approach, but they still depend on having the right models. Garbage in, garbage out. They are also much more computationally intensive.
  - Imagine if you need to simulate the entire gamma-ray sky, do source finding, make a logN-logS fit, etc. just for one realization!



Extension to Bayesian Concepts

- Basic probability rule: suppose you have a two-variate distribution, $p(m,d)$. Then, $p(m,d) = p(m|d) \times p(d)$. Likewise, $p(d,m) = p(d|m) \times p(m)$. Solving for this, we have Bayes' Theorem:

$$p(m|d) = p(d|m) \times \frac{p(m)}{p(d)}$$

the "posterior" probability density (for model parameters)

the "prior" probability density (for model parameters)

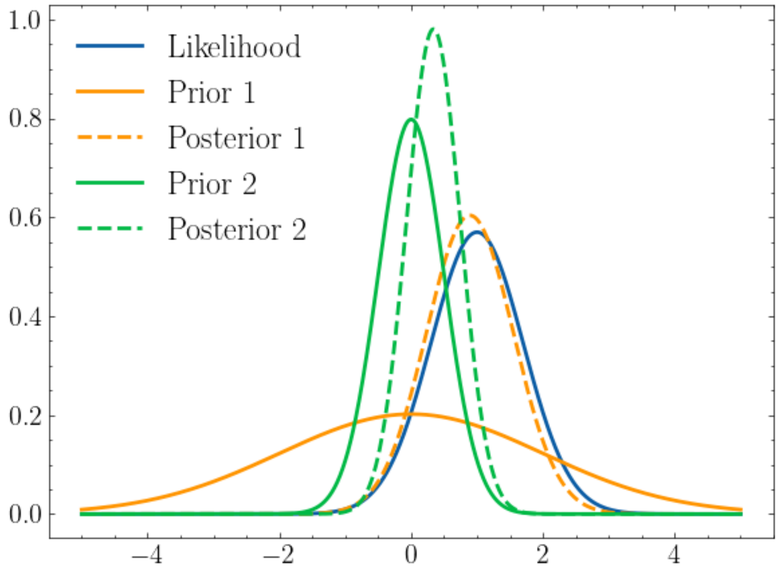
the likelihood we've been talking about

the probability of the data; "evidence"

- The key is in the interpretation: before, we had data (random variables) and model parameters (just numbers). Bayes' Theorem re-interprets those parameters as random variables, too. Therefore, let m be the parameters from some model, and d be the data.
  - People have different opinions about the validity of this approach. But, by converting model parameters to random variables, it lets us use the whole machinery of statistics to address their properties, to calculate uncertainties, to do model selection…
  - The "prior", $p(m)$, is critical. We can (and must) use it to encode our beliefs/information about the model parameters. Just like with ML, it's important to gauge the impact of our assumptions on our results.

Parameter Estimation: Bayesian Credible Intervals

- Bayesian credible intervals are parameter ranges which contain some specified fraction of the posterior distribution, e.g. 68%.
  - They are very much analogous to “error bars”, at least colloquially.
  - They are more flexible because we have the entire distribution.
- They essentially “promote” our likelihood directly to a probability density. Previously, we appealed to arguments of plausibility (“well, a model that gives a high probability to the data is good!”) and asymptotic distributions (the MLE follows the Fisher information).
- Bayes makes this all concrete; it doesn’t matter how much data we have, or what the shape of the posterior is. Likelihood + prior = result



- The COST is it is even MORE model dependency. Not only must we specify the model for our likelihood, we must specify a prior range for parameters.
  - There is a wide body of literature on priors, beyond the scope of this lecture.
  - Shown above are two examples of a prior: an “informative” one, and an “uninformative” one. The posteriors differ strongly. (In either case, it is easy to construct a credible interval: make a cumulative distribution and read off the answer.)
  - Generally, you can at least put some wide bounds on parameters. A VERY COMMON TACTIC is to adopt a uniform prior. This makes it real easy, because the likelihood IS the posterior then.

Hypothesis Testing: Bayesian Model Selection

- Just like Bayes promoted our likelihood to a probability density function, it lets us directly compare models via probability (rather than comparing the probabilities of **data**).
  - The most classic examples of these are from medicine and crime: suppose you have a pretty good test (90% efficacy, 5% false positive) and you test for a very rare thing many times.

	True Positive	True Negative
Test Positive	0.90	0.05
Test Negative	0.10	0.95

If we have a positive test (data), the True Positive Model is 0.90/0.05 = 18x more *likely*. But, what is the *total* probability of the True Positive.

  - For this, we need to know the incidence rate. Suppose only 1:100,000 people are sick (True Positive). The prior probability for TP is thus is 1e-5, so the posterior probability of TP, *given a positive test result,* is still only 0.9*1e-5 = 0.0009%. The test has basically updated the "odds" of being sick from 1:100,000 to 1:11,111. But it is still VERY UNLIKELY that the positive tester is actually sick.

- In astrophysics, our models are both more complicated and more flexible. We also generally adhere to the principle that a simpler model is better: Occam's razor. This motivates the concept of "model evidence":

$$\int dm\, p(d|m)p(m)$$

It is the integral of the likelihood times the prior over the full parameter space.

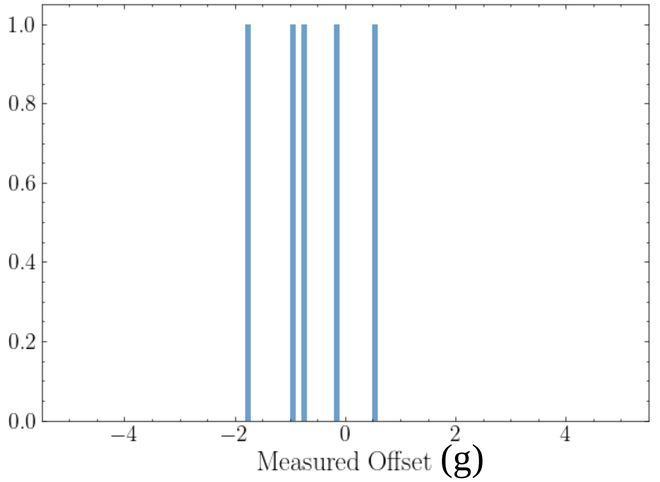
Matthew:
Breeze Over This!

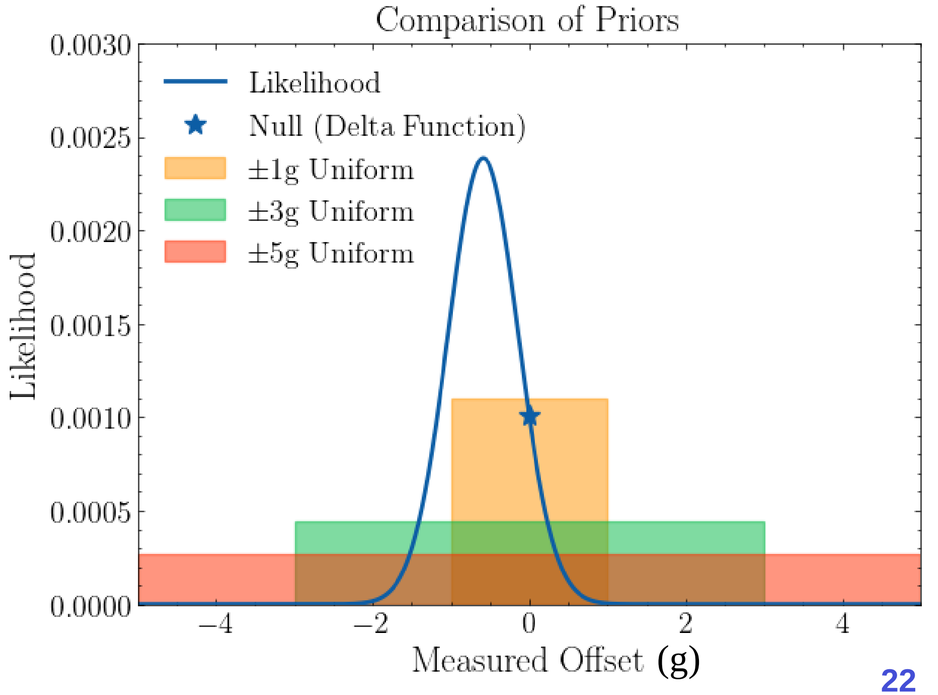
Bayesian Evidence: Intuition and Evaluation

- Come back to our first “five baguette” data set and our question about model selection: did the baguettes change mass or not? We considered Wilks’ Theorem. Here is the same decision process using “Bayesian model evidence”:

- We compute the evidence using 4 priors: a Delta function giving the null hypothesis, and then a uniform (flat) distribution expanding to wider possible values of the mean.
  - Intuitively, the model that has the smallest range while also best describing the data (orange) gets the highest evidence. The other priors “dilute” their agreement by also considering larger parameter spaces.
  - The model evidence for the +/-1g span is only 10% higher than the null model. Thus, we have no reason to pick the more complicated models.

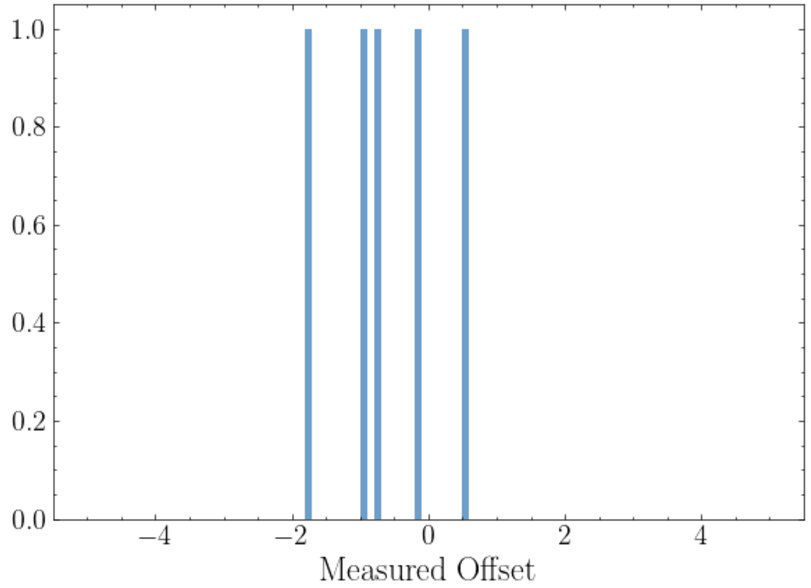
- This 1-d gaussian model is easy, but generally, computing the evidence is REALLY computationally intensive. Integral over many dimensions.
  - “Nested samplers” are good for evaluating evidence. But it’s still heinously expensive. I suggest: don’t burn the planet, save it for very important and/or tricky problems you can’t sort out any other way.
    - In general, always try to profile your code before going to more cores!





From Baguettes to Bosons

- Let's go back to our original data set, and let each baguette be a photon.
- Our scale becomes a gamma-ray (or X-ray, or optical) detector with 0.1-deg pixels, and the +/-1g precision becomes a 1-degree point spread function.
- How can we answer basic questions about the possible gamma-ray source, like its brightness, its position, its multiplicity. Does our PSF model look OK…?



- We already have all the answers! We just need to use a slightly more complicated statistical model for the likelihood: Poisson statistics.
- Slight change in philosophy: while we only observe 5 photons, we actually have more information than that. Any given pixel **could** have had one or more photons observed in it. Therefore, our data set is actually the whole set of pixels and the counts in them. Zeros have value!
  - You can ask me about "unbinned" statistics, which basically do allow you to throw away empty bins, which can be a savings for sparse data. However, as you'll see, you still need to account for empty pixels. So we'll just use bins.
- The other big change in philosophy is that we KNOW that the Poisson distribution is the correct one here. The "uncertainty" part of our model is associated with the source, not the instrument.

Poisson Log Likelihood

$$p(n|\lambda) = \frac{\lambda^n}{n!} e^{-\lambda} \qquad \log L = \sum_i n_i \log \lambda_i - \lambda_i$$

- The Poisson distribution gives the probability of observing *n* counts given a predicted λ counts. The exponentials make the log likelihood very attractive!
  - Pro tip: λ **must be dimensionless.** You can never take the log of a dimensionful thing.
    - Therefore, you will be **integrating** some **rate** (cts/s/cm^2) over time and effective area.
  - The Poisson distribution becomes gaussian as $\lambda \gg 1$.
  - It is very far from gaussian otherwise! (Basically exponential.)
  - λ is strictly positive.
- All of the difficulty lies in evaluating λ:
  - Consider a general (but 1-d) case of a PSF $f(x)$. (Recall the integral of the PSF must be 1). And a source with an amplitude of A cts/s and a position mu. Then, for a pixel of width dx,

$$\lambda_i = A \times T \times \int_{x_i - \delta x/2}^{x_i + \delta x/2} f(x|\mu_1)\,.$$

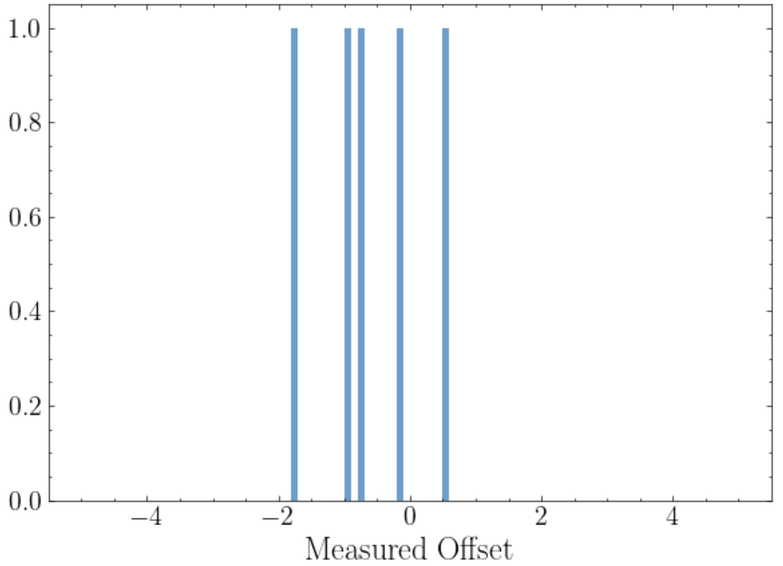
  - For two sources, we would need 2 rates: $\lambda_i = T \times \int_{x_i-\delta x/2}^{x_i+\delta x/2} A_1 \times f(x|\mu_1) + A_2 \times f(x|\mu_2) = \lambda_{1i} + \lambda_{2i}$
  - For Fermi-LAT, our integrals extend over energy and time, too!

Maximum Likelihood with Simple Model

- For our case, let's assume there's only one source, one spatial dimension and that our PSF is gaussian with sigma=1 deg.

$$\log L = \sum_i n_i \log \lambda_i - \lambda_i$$

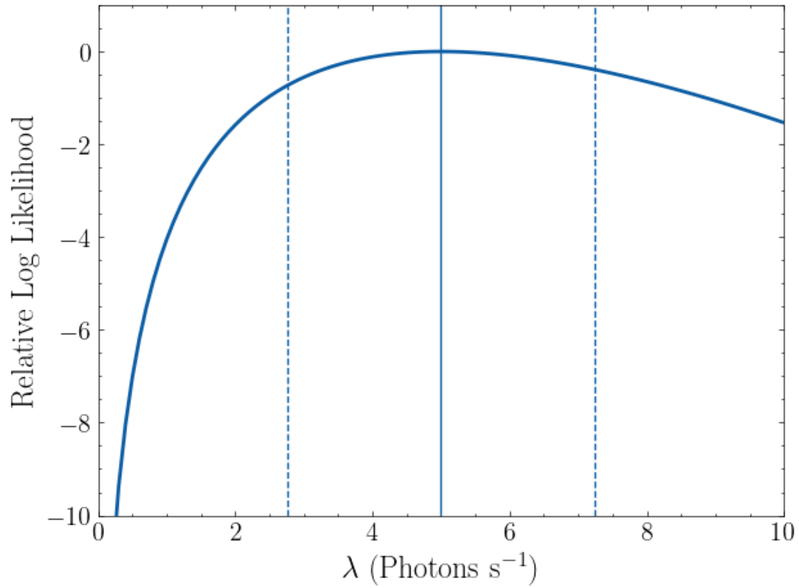
$$\lambda_i = A \times T \times \int_{x_i-0.05}^{x_i-0.05} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \equiv A \times T \times F_i$$

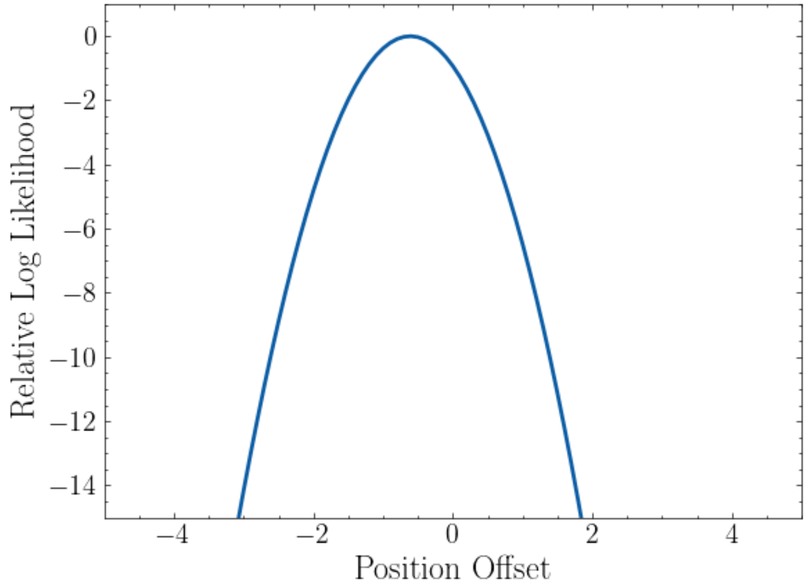


- Even this simple case is much more complicated than the "baguettes". We have to integrate a potentially complicated PSF over **every** pixel. (And every time the source position changes, we have to re-do the integrals.) If we have many sources, we'd have a lot of integrals to do!
- Source parameters are almost always buried inside of logarithms or integrals, meaning <u>maximum likelihood typically must be done numerically</u>.
  - As always, evaluating derivatives analytically can help an optimizer. Can also do productive things with approximations for fast source finding. In this case, our gaussian PSF can actually be evaluated analytically to avoid an actual quadrature.

Maximum Likelihood with Simple Model

- Here are the results for a single-source model:
- First, we fix the source at the true position.
  - (We wouldn't usually know this!)
  - Note that the likelihood/posterior is VERY non-gaussian: we can't actually have $\lambda \to 0$ because $\log \lambda \to \infty$. (Intuitively, we have observed counts, so there **must** be a source.)
  - Consequently, we probably shouldn't rely just on the gaussian errors, but use a Bayesian-type approach and provide two-sided uncertainties.
- Second, we fix the source flux to the known value (or to the maximum likelihood value) and scan for position.
  - We see the peak is slightly offset from 0 but that the difference in log likelihood isn't large.
  - Does this follow Wilks' Theorem?
    - If so, the difference in likelihood has a significance of about 17%, almost the same as the baguette case!





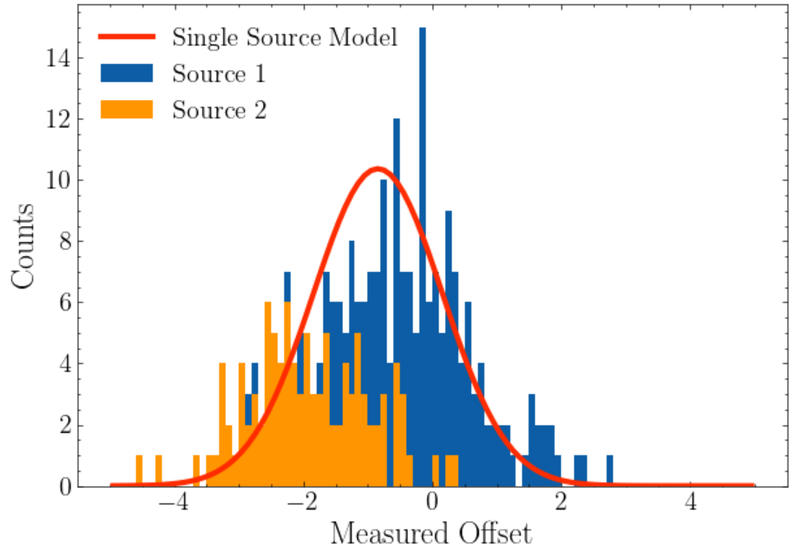
Additional Sources

- Let's come back to our "two-source" data. As before, we can consider a "null" model with a single source and an alternative model with two sources.

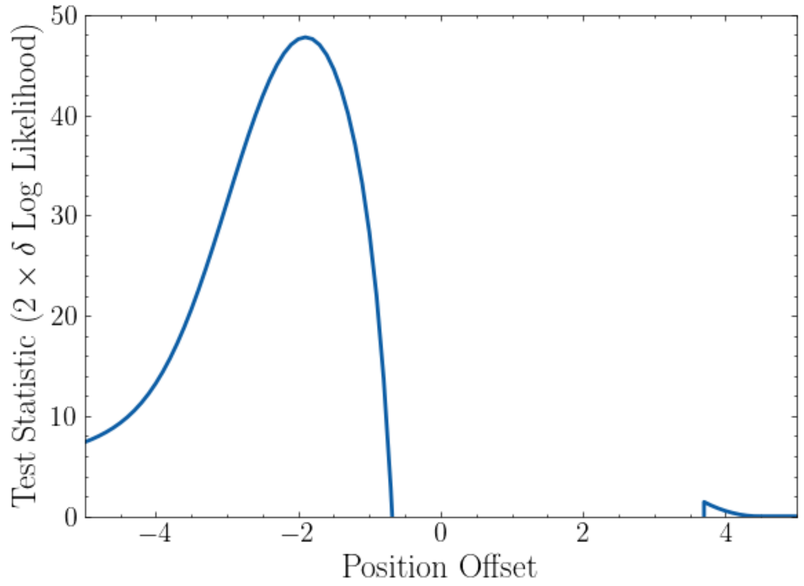
$$\log L = \sum_i n_i \log(\lambda_{1i} + \lambda_{21}) - (\lambda_{1i} + \lambda_{2i})$$

$$\lambda_{ji} = A_j \times T \times \int_{x_i - 0.05}^{x_i - 0.05} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu_j)^2}{2\sigma^2}\right] \equiv A_j \times T \times F_i$$

The null model is $\lambda_2 \to 0$. The plot shows both the data (with labels) and the best-fit (maximum likelihood) model with a single source overlaid. It looks… OK, but not great.



- Now, we can evaluate the presence of an additional source. We don't know its position a priori, and because the sources overlap, they will definitely affect each other. Therefore, use "profile likelihood": report the log likelihood as a function of $\mu\_$ at the maximum values of the other parameters.
  - This is a "TS map"!
  - Does the TS reported here satisfy Wilks' Theorem? Think about the definition of a source "position" when its flux is 0.



Summary

- Likelihood can be used for parameter estimation and model testing, especially with extensions to Bayesian methodology.
- You must always think and check carefully your choice of statistical distributions to represent your data.
  - The most expensive Bayesian model evidence comparison is useless if none of your models is a good fit to the data. (Conversely, some models are flexible enough to fit anything and should perhaps be avoided.)
- With “raw” Fermi-LAT data, you will always be using Poisson statistics, and the complexity lies in your source model choices.
- Use the “cheapest” way of interpreting your data you can.
  - If Wilks’ applies, use it and move on with life. But beware when it doesn’t.
  - If your error bars aren’t so critical, just use the hessian and move on with life. If they are, consider simulations and/or full evaluations of the likelihood/posterior.
- Likelihood can be very computationally expensive (lots of integrals), so if your analysis requires a lot of it, look into whether you can optimize parts of it.