ASTR 3890 - Selected Topics: Data Science for Large Astronomical Surveys (Spring 2022)

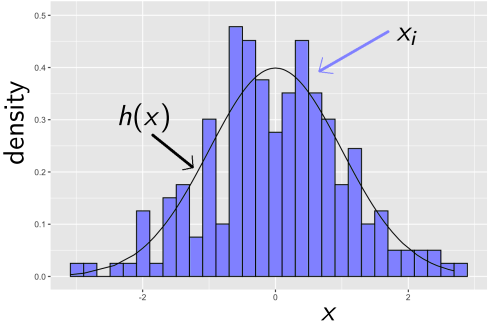
Classical/Frequentist Statistical Inference

Dr. Nina Hernitschek
February 21, 2022

recap: Goal of Statistical Inference

idea:

- measurements are drawn from an underlying probability distribution function (pdf) $h(x)$
- we can only observe the measurements x_i, not the underlying pdf



recap: Goal of Statistical Inference

Statistical inference is about **drawing conclusions from data**, specifically determining the properties of a population by data sampling.

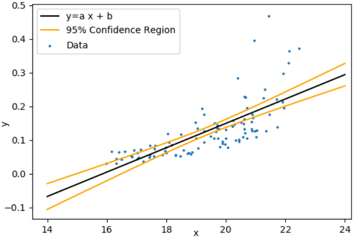
Three examples of inference are:

recap: Goal of Statistical Inference

Statistical inference is about **drawing conclusions from data**, specifically determining the properties of a population by data sampling.

Three examples of inference are:

1. What is the best estimate for a (set of) model parameter(s)?



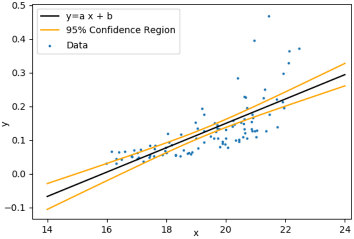
recap: Goal of Statistical Inference

Statistical inference is about **drawing conclusions from data**, specifically determining the properties of a population by data sampling.

Three examples of inference are:

1. What is the best estimate for a (set of) model parameter(s)?

2. How confident we are about our result?

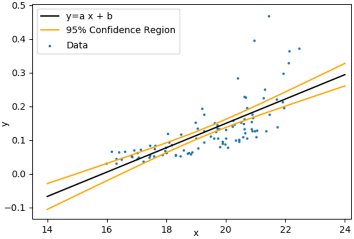


recap: Goal of Statistical Inference

Statistical inference is about **drawing conclusions from data**, specifically determining the properties of a population by data sampling.

Three examples of inference are:

1. What is the best estimate for a (set of) model parameter(s)?
2. How confident we are about our result?
3. Are the data consistent with a particular model/hypothesis?



recap: Some Terminology

We study the properties of some **population** by measuring **samples** from that population. The population doesn't have to refer to different objects.

recap: Some Terminology

We study the properties of some **population** by measuring **samples** from that population. The population doesn't have to refer to different objects.

example: E.g., we may be (re)measuring the position of an object at rest; the population is the distribution of (an infinite number of) measurements smeared by the uncertainty, and the sample are the measurement we've actually taken.

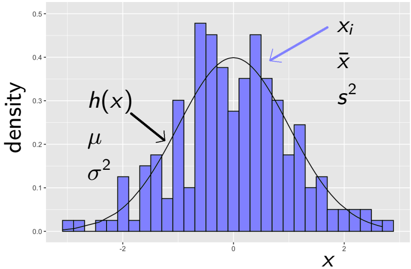
subsequent brightness measurements of a star:

ra	dec	hjd	mag	magErr	filter
347.66112	-7.39883	2458277.96036	20.083	0.135	g
347.66111	-7.39883	2458280.94526	20.49	0.163	g
347.66111	-7.39881	2458283.94197	19.822	0.116	g
347.66113	-7.39883	2458289.93875	20.361	0.155	g
347.66111	-7.39883	2458377.75728	20.103	0.137	r
347.66111	-7.39883	2458380.84366	20.291	0.151	r
347.66111	-7.39883	2458430.66968	20.471	0.162	r

recap: Some Terminology

A **statistic** is any function of the sample. For example, the sample mean is a statistic. But also something like *the value of the first measurement* is also a statistic.

To conclude something about the population from the sample, we use **estimators**. An estimator is a statistic, a rule for calculating an estimate of a given quantity based on observed data.



recap: Some Terminology

There are **point estimators** and **interval estimators**. The point estimators yield single-valued results (example: the position of an object), while with an interval estimator, the result would be a range of plausible values (example: confidence interval for the position of an object).

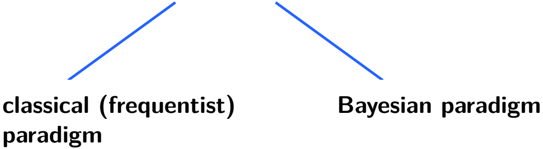
recap: Some Terminology

There are **point estimators** and **interval estimators**. The point estimators yield single-valued results (example: the position of an object), while with an interval estimator, the result would be a range of plausible values (example: confidence interval for the position of an object).

Measurements have **uncertainties** (not errors) and we need to account for these (sometimes they are unknown).

Frequentist vs. Bayesian Statistical Inference

There are two major statistical paradigms that address the statistical inference questions:



classical (frequentist) paradigm

Bayesian paradigm

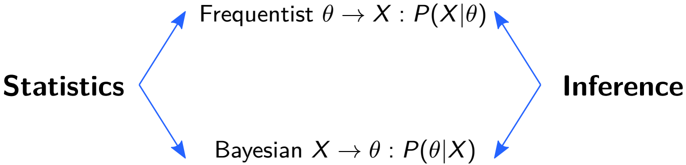
Frequentist vs. Bayesian Statistical Inference

There are two major statistical paradigms that address the statistical inference questions:

Key differences	**classical (frequentist) paradigm**	**Bayesian paradigm**
Definition of probabilities:	relative frequency of events over repeated experimental trials	probabilities quantify our subjective belief about experimental outcomes, model parameters, or models
Quantifying uncertainty:	confidence levels describe the distribution of the measured parameter from the data around the true value	credible regions derived from posterior probabilitiy distributions encode our belief in model parameters

Frequentist vs. Bayesian Statistical Inference

we can summarize this as



Statistics

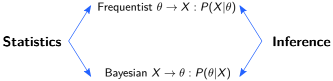
Frequentist $\theta \to X : P(X|\theta)$

Bayesian $X \to \theta : P(\theta|X)$

Inference

Frequentist vs. Bayesian Statistical Inference

an example:



Frequentist $\theta \rightarrow X : P(X|\theta)$

Statistics

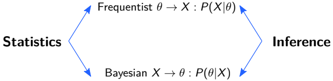
Inference

Bayesian $X \rightarrow \theta : P(\theta|X)$

A person takes an IQ test (which does not give the “real” IQ but is a way to estimate it, and a possible range of values).

Frequentist vs. Bayesian Statistical Inference

an example:



Statistics

Frequentist $\theta \to X : P(X|\theta)$

Bayesian $X \to \theta : P(\theta|X)$

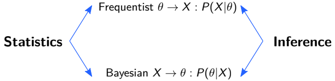
Inference

A person takes an IQ test (which does not give the “real” IQ but is a way to estimate it, and a possible range of values).

For a **frequentist**, the best estimator is the **average** of many test results. So, if 5 IQ tests were taken and the sample mean is of 160, then that would be the estimator of that candidate’s true IQ.

Frequentist vs. Bayesian Statistical Inference

an example:



Statistics

Frequentist $\theta \rightarrow X : P(X|\theta)$

Bayesian $X \rightarrow \theta : P(\theta|X)$

Inference

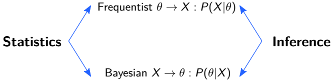
A person takes an IQ test (which does not give the "real" IQ but is a way to estimate it, and a possible range of values).

For a **frequentist**, the best estimator is the **average** of many test results. So, if 5 IQ tests were taken and the sample mean is of 160, then that would be the estimator of that candidate's true IQ.

A **Bayesian** would say: *IQ tests are calibrated with a mean of 100, standard deviation of 15 points* and use this as a **prior** information. The Bayesian estimate of that candidate's person thus would be not 160, but rather 148, or more specifically that $p(141.3 \leq \mu \leq 154.7 \,|\, \overline{x} = 160) = 0.683$.

Frequentist vs. Bayesian Statistical Inference

an example:



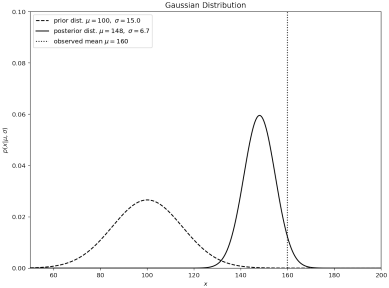
Frequentist $\theta \rightarrow X : P(X|\theta)$

Statistics

Inference

Bayesian $X \rightarrow \theta : P(\theta|X)$

A **Bayesian** would say: *IQ tests are calibrated with a mean of 100, standard deviation of 15 points* and use this as a **prior** information. The Bayesian estimate of that candidate's person thus would be not 160, but rather 148, or more specifically that $p(141.3 \leq \mu \leq 154.7 \,|\, \overline{x} = 160) = 0.683$.



we will see an astronomy example later in `lecture_5.ipynb`

The Null Hypothesis

All statistical tests have a null hypothesis. In inferential statistics, the null hypothesis (often denoted H_0) is that two possibilities are the same and the observed difference is due to chance alone.

The Null Hypothesis

All statistical tests have a null hypothesis. In inferential statistics, the null hypothesis (often denoted H_0) is that two possibilities are the same and the observed difference is due to chance alone.

example: Null and alternative hypothesis
You want to know whether there is a difference in longevity between two groups of mice fed on different diets, diet A and diet B.
Null hypothesis: there is no difference in longevity between the two groups.
Alternative hypothesis: there is a difference in longevity between the two groups.

The p-value

A **p-value** is the calculated **probability of obtaining an effect at least as extreme as the one in your sample data**, assuming the truth of the null hypothesis.

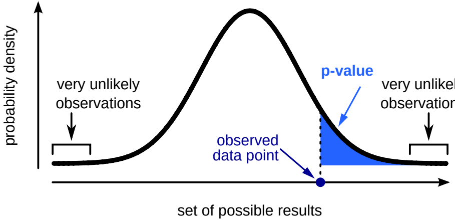
A small p-value means that there is a small chance that the results could be completely random. A large p-value means that the results have a high probability of being random and not due to anything from the experiment. The smaller the p-value, the more statistically significant the result.

The p-value

A **p-value** is the calculated **probability of obtaining an effect at least as extreme as the one in your sample data**, assuming the truth of the null hypothesis.

A small p-value means that there is a small chance that the results could be completely random. A large p-value means that the results have a high probability of being random and not due to anything from the experiment. The smaller the p-value, the more statistically significant the result.

example: A the p-value of 0.05 means that 5% of the time you would see a test statistic at least as extreme as the one you found if the null hypothesis was true.



The p-value

A **p-value** is the calculated **probability of obtaining an effect at least as extreme as the one in your sample data**, assuming the truth of the null hypothesis.

The p-value is often **misinterpreted**.

The p-value is essentially the probability of a false positive based on the data in the experiment. It does not tell the probability of a specific event actually happening and it does not tell the probability that a variant is better than the control. P-values are **probability statements about the data sample** not about the hypothesis itself.

Maximum Likelihood Estimation

The key idea:
Some data are known to be drawn from a certain distribution (e.g.: Gaussian), but we don't know the $\theta = (\mu, \sigma)$ values of that distribution (i.e., the parameters). How to estimate these parameters?

Maximum Likelihood Estimation

The key idea:
Some data are known to be drawn from a certain distribution (e.g.: Gaussian), but we don't know the $\theta = (\mu, \sigma)$ values of that distribution (i.e., the parameters). How to estimate these parameters?

The Maximum Likelihood Estimation (MLE) method tells us to think of the likelihood as a **function of the unknown model parameters**, and to **find the parameters that maximize the value of** L. Those will be the **Maximum Likelihood Estimators** for for the true values of the model.

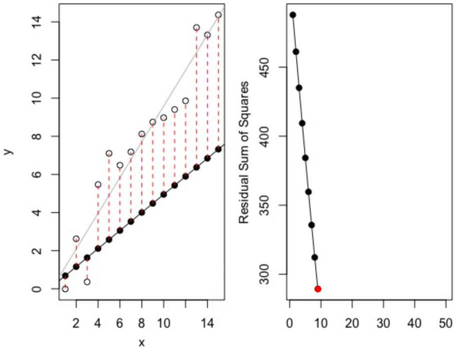
Maximum Likelihood Estimation

example:

trying to fit a line to some data using linear least squares fitting

based on this animation:

`https://yihui.org/animation/example/least-squares/`



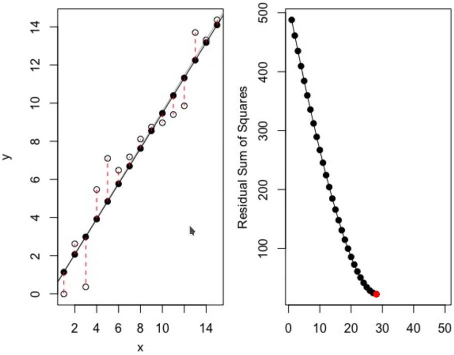
Maximum Likelihood Estimation

example:

trying to fit a line to some data using linear least squares fitting

based on this animation:

`https://yihui.org/animation/example/least-squares/`



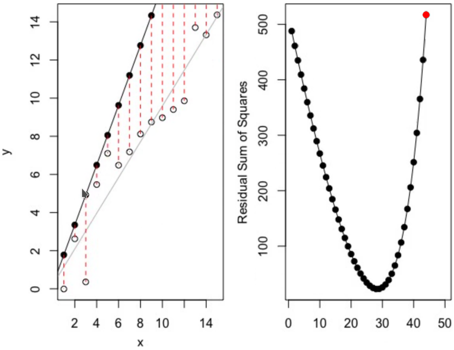
Maximum Likelihood Estimation

example:

trying to fit a line to some data using linear least squares fitting

based on this animation:

`https://yihui.org/animation/example/least-squares/`



Maximum Likelihood Approach

Maximum likelihood estimation follows this blueprint:

1. Hypothesis: Formulate a model, a hypothesis, about how the data are generated.
For example, the data are a measurement of some quantity with Gaussian random uncertainties (i.e., each measurement is equal to the true value, plus a deviation randomly drawn from the normal distribution). Models are described using a set of model parameters $\boldsymbol{\theta}$, and written as $\boldsymbol{M}(\boldsymbol{\theta})$.

2. Maximum Likelihood Estimation: Search for the "best" model parameters $\boldsymbol{\theta}$ maximizing the likelihood $L(\boldsymbol{\theta}) \equiv p(D|M)$.

3. Quantifying Estimate Uncertainty: Determine the confidence region for model parameters, $\boldsymbol{\theta}^{\mathbf{0}}$.

4. Hypothesis Testing: Perform hypothesis tests as needed to make other conclusions about models and point estimates.

Maximum Likelihood Approach

Maximum likelihood estimation follows this blueprint:

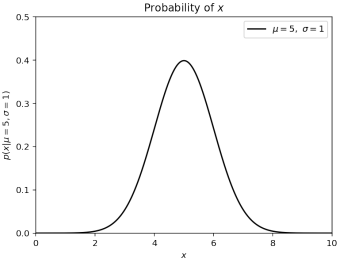
1. **Hypothesis:** Formulate a model, a hypothesis, about how the data are generated.
   For example, the data are a measurement of some quantity with Gaussian random uncertainties (i.e., each measurement is equal to the true value, plus a deviation randomly drawn from the normal distribution). Models are described using a set of model parameters $\boldsymbol{\theta}$, and written as $\boldsymbol{M}(\boldsymbol{\theta})$.
2. **Maximum Likelihood Estimation:** Search for the "best" model parameters $\boldsymbol{\theta}$ maximizing the likelihood $L(\boldsymbol{\theta}) \equiv p(D|M)$.
3. **Quantifying Estimate Uncertainty:** Determine the confidence region for model parameters, $\boldsymbol{\theta}^{\mathbf{0}}$.
4. **Hypothesis Testing:** Perform hypothesis tests as needed to make other conclusions about models and point estimates.

The Likelihood Function

If we know the distribution from which our data were drawn (or make a hypothesis about it), then we can compute the **probability** of our data being generated.

example: If our data are generated by a Gaussian process, then the probability density of a certain value x is

$$p(x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right).$$



The Likelihood Function

If we want to know the total probability of our **entire data set** (as opposed to one measurement) then we must compute the product of all the individual probabilities:

$$L \equiv p(\{x_i\}|M(\theta)) = \prod_{i=1}^{N} p(x_i|M(\theta)),$$

where M is the model and θ refers collectively to the model parameters, which can generally be multi-dimensional.

The Likelihood Function

If we want to know the total probability of our **entire data set** (as opposed to one measurement) then we must compute the product of all the individual probabilities:

$$L \equiv p(\{x_i\}|M(\theta)) = \prod_{i=1}^{N} p(x_i|M(\theta)),$$

where M is the model and θ refers collectively to the model parameters, which can generally be multi-dimensional.

$L(\{x_i\}) \equiv$ the probability of the data given the model parameters.

The Likelihood Function

If we want to know the total probability of our **entire data set** (as opposed to one measurement) then we must compute the product of all the individual probabilities:

$$L \equiv p(\{x_i\}|M(\theta)) = \prod_{i=1}^{N} p(x_i|M(\theta)),$$

where M is the model and θ refers collectively to the model parameters, which can generally be multi-dimensional.

$L(\{x_i\}) \equiv$ the probability of the data given the model parameters.

If we consider L as a function of the model parameters, we refer to it as
$L(\theta) \equiv$ likelihood of the model parameters, given the data.

The Likelihood Function

If we want to know the total probability of our **entire data set** (as opposed to one measurement) then we must compute the product of all the individual probabilities:

$$L \equiv p(\{x_i\}|M(\theta)) = \prod_{i=1}^{N} p(x_i|M(\theta)),$$

where M is the model and θ refers collectively to the model parameters, which can generally be multi-dimensional.

caution:

- While the components of L may be normalized pdfs, their product is not.
- The product can be very small, so we often take the log of L.
- We're assuming the individual measurements are independent of each other.

The Likelihood Function

We can write L out as

$$L = \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x_i - \mu)^2}{2\sigma^2}\right),$$

and simplify to

$$L = \left(\prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}}\right) \exp\left(-\frac{1}{2}\sum\left[\frac{-(x_i - \mu)}{\sigma}\right]^2\right),$$

where we have written the product of the exponentials as the exponential of the sum of the arguments, which will make things easier to deal with later.

To repeat, all we have done is:

$$\prod_{i=1}^{N} A_i \exp(-B_i) = (A_i A_{i+1} \ldots A_N) \exp[-(B_i + B_{i+1} + \ldots + B_N)]$$

The Likelihood Function

If you have done χ^2 analysis (e.g., doing a linear least-squares fit), then you might notice that the argument of the exponential is just

$$\exp\left(-\frac{\chi^2}{2}\right).$$

That is, for our Gaussian distribution

$$\chi^2 = \sum_{i=1}^{N}\left(\frac{x_i - \mu}{\sigma}\right)^2.$$

The Likelihood Function

If you have done χ^2 analysis (e.g., doing a linear least-squares fit), then you might notice that the argument of the exponential is just

$$\exp\left(-\frac{\chi^2}{2}\right).$$

That is, for our Gaussian distribution

$$\chi^2 = \sum_{i=1}^{N}\left(\frac{x_i-\mu}{\sigma}\right)^2.$$

So, **maximizing the likelihood or log-likelihood is the same as minimizing** χ^2. In both cases we are finding the most likely values of our model parameters (here μ and σ).

MLE applied to a Homoscedastic Gaussian

In statistics, a sequence (or a vector) of random variables is homoscedastic if all its random variables have the same finite variance.

The model used here assumes that **all measurements have the same uncertainty**, drawn from $N(0, \sigma)$.

MLE applied to a Homoscedastic Gaussian

In statistics, a sequence (or a vector) of random variables is homoscedastic if all its random variables have the same finite variance.

The model used here assumes that **all measurements have the same uncertainty**, drawn from $N(0, \sigma)$.

homoscedastic uncertainties

(Later we will consider the case where the measurements have different uncertainties (σ_i) which is called **heteroscedastic**.)

MLE applied to a Homoscedastic Gaussian

example: Measuring the Position of a Quasar

Let's assume we wish to estimate the position x of a quasar from a series of individual astrometric measurements.

MLE applied to a Homoscedastic Gaussian

example: Measuring the Position of a Quasar

Let's assume we wish to estimate the position x of a quasar from a series of individual astrometric measurements.

1. We adopt a model where the observed quasar does not move, and has individual measurement uncertainties. We have thus a set of measured positions $D = \{x_i\}$ in 1D with Gaussian uncertainties.

MLE applied to a Homoscedastic Gaussian

example: Measuring the Position of a Quasar

Let's assume we wish to estimate the position x of a quasar from a series of individual astrometric measurements.

1. We adopt a model where the observed quasar does not move, and has individual measurement uncertainties. We have thus a set of measured positions $D = \{x_i\}$ in 1D with Gaussian uncertainties.

2. We derive the expression for the likelihood of there being a quasar at the true position μ that gives rise to our individual measurements. We find the value of $\hat{\mu}$ for which our observations are maximally likely.

MLE applied to a Homoscedastic Gaussian

example: Measuring the Position of a Quasar

Let's assume we wish to estimate the position x of a quasar from a series of individual astrometric measurements.

1. We adopt a model where the observed quasar does not move, and has individual measurement uncertainties. We have thus a set of measured positions $D = \{x_i\}$ in 1D with Gaussian uncertainties.

2. We derive the expression for the likelihood of there being a quasar at the true position μ that gives rise to our individual measurements. We find the value of $\hat{\mu}$ for which our observations are maximally likely.

3. We determine the uncertainties (confidence intervals) on our measurement.

MLE applied to a Homoscedastic Gaussian

example: Measuring the Position of a Quasar

Let's assume we wish to estimate the position x of a quasar from a series of individual astrometric measurements.

1. We adopt a model where the observed quasar does not move, and has individual measurement uncertainties. We have thus a set of measured positions $D = \{x_i\}$ in 1D with Gaussian uncertainties.

2. We derive the expression for the likelihood of there being a quasar at the true position μ that gives rise to our individual measurements. We find the value of $\hat{\mu}$ for which our observations are maximally likely.

3. We determine the uncertainties (confidence intervals) on our measurement.

4. We test whether what we've observed is consistent with our adopted model. For example, is it possible that the quasar was really a misidentified star with measurable proper motion?

MLE applied to a Homoscedastic Gaussian

example: Measuring the Position of a Quasar

We have a the set of measured positions $D = \{x_i\}$ in 1D with Gaussian uncertainties, and therefore:

$$L \equiv p(\{x_i\}|\mu, \sigma) = \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x_i - \mu)^2}{2\sigma^2}\right).$$

Note: This is $p(\{x_i\})$, the probability of the full data set, not not $p(x_i)$ of just one measurement. If σ is both constant and known, then this is a one parameter model with number of model parameters $k = 1$ and model parameter $\theta_1 = \mu$.

MLE applied to a Homoscedastic Gaussian

example: Measuring the Position of a Quasar

As we found above, likelihoods can be really small, so let's define the **log-likelihood function** as $\ln L = \ln[L(\theta)]$. The maximum of this function happens at the same place as the maximum of L. Note that any constants in L have the same effect for all model parameters, so constant terms can be ignored.
In this case we then have

$$\ln L = \mathrm{const} - \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^2}$$

by using

$$L = \prod_{i=1}^{N} \left(\frac{1}{\sigma\sqrt{2\pi}} \right) \exp\left(-\frac{1}{2} \sum \left[\frac{-(x_i - \mu)}{\sigma} \right]^2 \right).$$

MLE applied to a Homoscedastic Gaussian

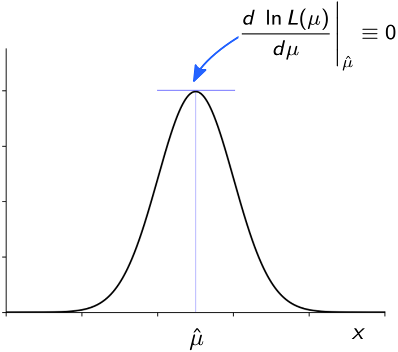
example: Measuring the Position of a Quasar

We finally determine the **maximum** in the usual way by setting the derivative of $\ln L$ to zero:

$$\left.\frac{d\ \ln L(\mu)}{d\mu}\right|_{\hat{\mu}} \equiv 0.$$

That gives

$$\sum_{i=1}^{N} \frac{(x_i - \hat{\mu})}{\sigma^2} = 0.$$



MLE applied to a Homoscedastic Gaussian

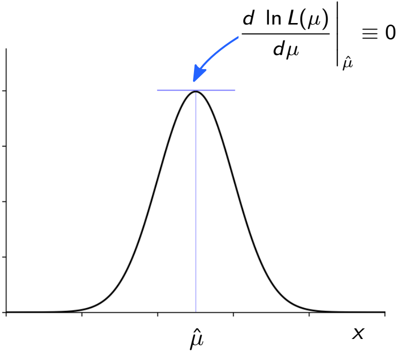
example: Measuring the Position of a Quasar

We finally determine the **maximum** in the usual way by setting the derivative of ln L to zero:

$$\left.\frac{d\ \ln L(\mu)}{d\mu}\right|_{\hat{\mu}} \equiv 0.$$

That gives

$$\sum_{i=1}^{N} \frac{(x_i - \hat{\mu})}{\sigma^2} = 0.$$



We should also check that the 2^{nd} derivative is negative, to ensure this is the **maximum** of L.

MLE applied to a Homoscedastic Gaussian

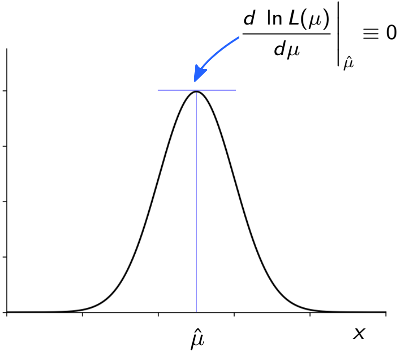
example: Measuring the Position of a Quasar

We finally determine the **maximum** in the usual way by setting the derivative of ln L to zero:

$$\left.\frac{d\ \ln L(\mu)}{d\mu}\right|_{\hat{\mu}} \equiv 0.$$

That gives

$$\sum_{i=1}^{N} \frac{(x_i - \hat{\mu})}{\sigma^2} = 0.$$



$$\left.\frac{d\ \ln L(\mu)}{d\mu}\right|_{\hat{\mu}} \equiv 0$$

Constants in ln L disappear when differentiated, so constant terms can typically be ignored. This will change if we select between different models, rather than parameter estimation.

MLE applied to a Homoscedastic Gaussian

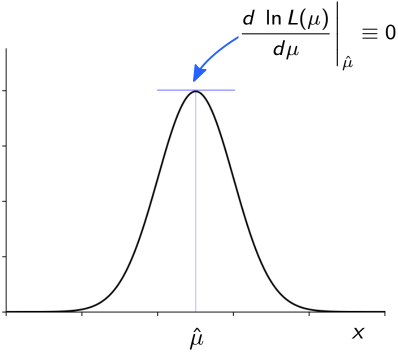
example: Measuring the Position of a Quasar

Since $\sigma = \mathrm{const}$ (in our case), that says

$$\sum_{i=1}^{N} x_i = \sum_{i=1}^{N} \hat{\mu} = N\hat{\mu}.$$

Thus our result is:

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i$$



which is just the sample arithmetic mean of all the measurements!

MLE applied to a Homoscedastic Gaussian - Quantifying Estimate Uncertainty

The uncertaintly of the estimate $\hat{\mu}$ is captured by the shape and distribution of the likelihood function, but we'd like to capture that with a few numbers.

The **asymptotic normality of MLE** is invoked to approximate the likelihood function as a Gaussian (or the $\ln L$ as a parabola), i.e. we take a Taylor expansion around the MLE, keep terms up 2^{nd} order, then define the uncertainty as:

$$\sigma_{jk} = \sqrt{[F^{-1}]_{jk}},$$

where

$$F_{jk} = -\frac{d^2}{d\theta_j}\frac{\ln L}{d\theta_k}\bigg|_{\theta=\hat{\theta}}.$$

The matrix F is known as the **observed Fisher information matrix**. The elements σ^2_{jk} are known as the **covariance matrix**.

MLE applied to a Homoscedastic Gaussian - Quantifying Estimate Uncertainty

observed Fisher information matrix F with

$$F_{jk} = -\frac{d^2}{d\theta_j}\frac{\ln L}{d\theta_k}\bigg|_{\theta=\hat{\theta}}$$

The **marginal error bars** for each parameter, θ_i are given by the diagonal elements, σ_{ii}. These are the "error bars" that are typically quoted with each measurement. Off diagonal elements, σ_{ij}, arise from any correlation between the parameters in the model.

For the homoscedastic Gaussian, the uncertainly on the mean is

$$\sigma_\mu = \left(-\frac{d^2 \ln L(\mu)}{d\mu^2}\bigg|_{\hat{\mu}} \right)^{-1/2}$$

We find

$$\frac{d^2 \ln L(\mu)}{d\mu^2}\bigg|_{\hat{\mu}} = -\sum_{i=1}^{N} \frac{1}{\sigma^2} = -\frac{N}{\sigma^2} \Rightarrow \sigma_\mu = \frac{\sigma}{\sqrt{N}}.$$

$\Rightarrow$ the estimator of μ is $\bar{x} \pm \frac{\sigma}{\sqrt{N}}$, which you should be familiar with

Properties of Maximum Likelihood Estimators

Assuming the data truly are drawn from the model, ML estimators have the following useful properties:

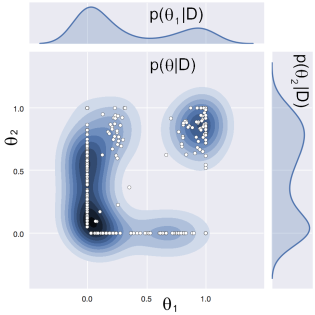
Properties of Maximum Likelihood Estimators

Assuming the data truly are drawn from the model, ML estimators have the following useful properties:

- They are **consistent estimators**: They converge to the true parameter value as $N \rightarrow \infty$.
- They are **asymptotically normal estimators**: As $N \rightarrow \infty$ the distribution of the parameter estimate approaches a normal distribution, centered at the MLE, with a certain spread.
- They **asymptotically achieve the theoretical minimum possible variance**, called the Cramér-Rao bound. They achieve the best possible uncertainty given the data at hand; no other estimator can do better in terms of efficiently using each data point to reduce the total error of the estimate (see Eq. 3.33 in the textbook).

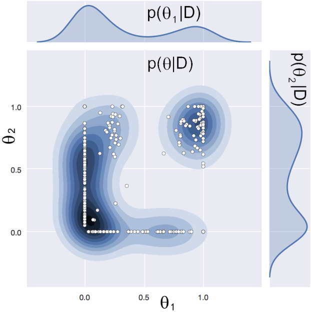
Maximizing the Likelihood - Practical Implications

For the Gaussian distribution we solved for the maximum liklehood analytically.
For many likelihoods we cannot solve for the maximum analytically, and we have to resort to **numerical solutions**.



Maximizing the Likelihood - Practical Implications

For the Gaussian distribution we solved for the maximum liklehood analytically.
For many likelihoods we cannot solve for the maximum analytically, and we have to resort to **numerical solutions**.



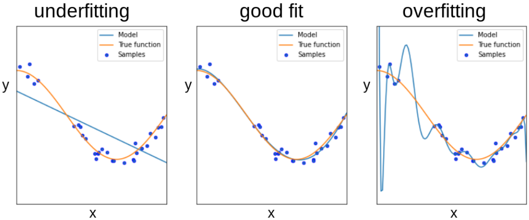
⇨ We'll treat these in detail later using MCMC and robust statistics that account for outliers.

Goodness Of Fit & Model Comparison

The MLE approach tells us what the best-fitting model parameters are, but not how good the fit actually is.
If the model isn't well suited for the data, then we should not expect a good fit.

example:

N points drawn from a linear distribution can always be fitted perfectly with an $N-1$ order polynomial - which won't help to predict future measurements



Goodness Of Fit & Model Comparison

we can describe the **goodness of fit** in words as simply the following:

The goodness of fit tells us whether or not it is likely to have obtained the maximum (log-)likelihood $\ln L^0$ by randomly drawing from the data.

Using the best-fit parameters of a model, the maximum likelihood value L^0 should not be an unlikely occurence. Otherwise: model is not describing the data well.

Goodness Of Fit & Model Comparison

For the **Gaussian distribution**:

With a standard transform of variables, we compute the z score for each data point:

$$z_i = (x_i - \mu)/\sigma$$

Then

$$\ln L = \text{constant} - \frac{1}{2}\sum_{i=1}^{N} z_i^2 = \text{constant} - \frac{1}{2}\chi^2.$$

⇨ for Gaussian uncertainties, $\ln L$ is distributed as χ^2

Goodness Of Fit & Model Comparison

The χ^2 distribution has a mean of $N - k$ and a standard deviation of $\sqrt{2(N-k)}$.
We define the χ^2 per degree of freedom, χ^2_{dof}, as

$$\chi^2_{\mathrm{dof}} = \frac{1}{N-k}\sum_{i=1}^{N} z_i^2 .$$

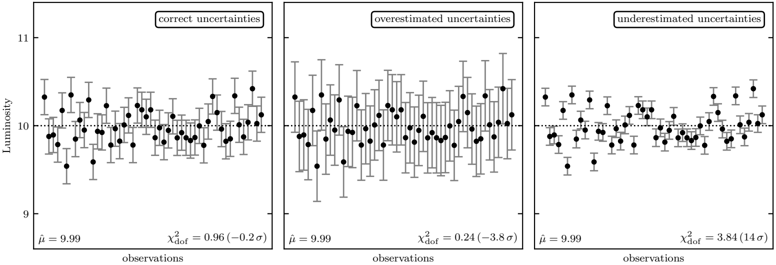
where again k is the number of model parameters determined from the data.

Goodness Of Fit & Model Comparison

For a good fit, we would expect that $\chi^2_{\mathrm{dof}} \approx 1$.

If χ^2_{dof} is significantly larger than 1, or $(\chi^2_{\mathrm{dof}} - 1) >> \sqrt{2/(N-k)}$, then it is likely that we are not using the correct model.

If data uncertainties are (over)under-estimated then this can lead to improbably (low) high χ^2_{dof}, as seen below.



Break & Questions

afterwards we continue with `lecture_5.ipynb` from the `github` repository