4th IMPRS Astronomy Summer School

Drawing Astrophysical Inferences from Data Sets

William H. Press
The University of Texas at Austin

Lecture 3

Central Limit Theorem

Let $S = \frac{1}{N}\sum X_i = \sum \frac{X_i}{N}$ with $\langle X_i \rangle \equiv 0$

Can always subtract off the means, then add back later.

Then

$$\phi_S(t) = \prod_i \phi_{X_i/N}(t) = \prod_i \phi_{X_i}\left(\frac{t}{N}\right)$$

$$= \prod_i \left(1 - \tfrac{1}{2}\sigma_i^2 \frac{t^2}{N^2} + \cdots\right)$$

Whoa! It better have a convergent Taylor series around zero! (Cauchy doesn't, e.g.)

$$= \exp\left[\sum_i \ln\left(1 - \tfrac{1}{2}\sigma_i^2 \frac{t^2}{N^2} + \cdots\right)\right]$$

These terms decrease with N, but how fast?

$$\approx \exp\left[-\frac{1}{2}\left(\frac{1}{N^2}\sum_i \sigma_i^2\right) t^2 + \cdots\right]$$

So, S is normally distributed

$$p_S(\cdot) \sim \text{Normal}(0, \tfrac{1}{N^2}\sum \sigma_i^2)$$

$$p_S(\cdot) \sim \text{Normal}(0, \tfrac{1}{N^2} \sum \sigma_i^2)$$

Moreover, since

$$NS = \sum X_i \quad \text{and} \quad \text{Var}(NS) = N^2 \text{Var}(S)$$

it follows that the simple sum of a large number of r.v.'s is normally distributed, with variance equal to the sum of the variances:

$$p_{\sum X_i}(\cdot) \sim \text{Normal}(0, \sum \sigma_i^2)$$

If N is large enough, and if the higher moments are well-enough behaved, and if the Taylor series expansion exists!

Also beware of borderline cases where the assumptions technically hold, but convergence to Normal is slow and/or highly nonuniform. (This can affect p-values for tail tests, as we will soon see.)

Just as moments are expectations of powers of a single r.v., you can form expectations of products of more than one r.v.

The only really important one is the covariance:

$$\mathrm{Cov}\,(x,y) = \langle (x-\overline{x})(y-\overline{y}) \rangle$$

For multiple r.v.'s, all the possible covariances form a (symmetric) matrix:

$$\mathbf{C} = C_{ij} = \mathrm{Cov}\,(x_i,x_j) = \langle (x_i-\overline{x_i})(x_j-\overline{x_j}) \rangle$$

Notice that the diagonal elements are the variances of the individual variables.

The variance of any linear combination of r.v.'s is a quadratic form in $\mathbf{C}$:

$$\begin{aligned}
\mathrm{Var}\,\left(\sum \alpha_i x_i\right) &= \left\langle \sum_i \alpha_i (x_i-\overline{x_i}) \sum_j \alpha_j (x_j-\overline{x_j}) \right\rangle \\
&= \sum_{ij} \alpha_i \left\langle (x_i-\overline{x_i})(x_j-\overline{x_j}) \right\rangle \alpha_j \\
&= \boldsymbol{\alpha}^T \mathbf{C}\, \boldsymbol{\alpha}
\end{aligned}$$

This also shows that $\mathbf{C}$ is positive definite, so it can be visualized as an ellipsoid in the space of the r.v.'s., where the directions are the different linear combinations.

Multivariate Normal Distributions

Generalizes Normal (Gaussian) to M-dimensions
Like 1-d Gaussian, completely defined by its mean and (co-)variance
Mean is a M-vector, covariance is a M x M matrix

$$N(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{M/2} \det(\boldsymbol{\Sigma})^{1/2}} \exp[-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu}) \cdot \boldsymbol{\Sigma}^{-1} \cdot (\mathbf{x}-\boldsymbol{\mu})]$$

The mean and covariance of r.v.'s from this distribution are

$$\boldsymbol{\mu} = \langle \mathbf{x} \rangle \qquad \boldsymbol{\Sigma} = \langle (\mathbf{x}-\boldsymbol{\mu}) \otimes (\mathbf{x}-\boldsymbol{\mu}) \rangle$$

It should *not* be obvious that this covariance in fact obtains from the above distribution! Here's the sketch of a proof (you fill in the words!):

$$\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^T \text{ (Cholesky)}, \quad \boldsymbol{\Sigma}^{-1} = (\mathbf{L}^T)^{-1}\mathbf{L}^{-1}, \quad \mathbf{L}\mathbf{y} \equiv \mathbf{x}$$

$$p(\mathbf{y}) = p(\mathbf{x}) \left| \frac{\partial \mathbf{x}}{\partial \mathbf{y}} \right|$$

$$= \frac{\det(\mathbf{L})}{(2\pi)^{N/2}\det(\boldsymbol{\Sigma})^{1/2}} \exp[-\tfrac{1}{2}(\mathbf{y}^T\mathbf{L}^T)(\mathbf{L}^{T-1}\mathbf{L}^{-1})(\mathbf{L}\mathbf{y})]$$

$$= \prod_i (2\pi)^{-1/2} \exp(-\tfrac{1}{2}y_i^2)$$

$$\langle \mathbf{x}\mathbf{x}^T \rangle = \langle \mathbf{L}\mathbf{y}\mathbf{y}^T\mathbf{L}^T \rangle = \mathbf{L} \langle \mathbf{y}\mathbf{y}^T \rangle \mathbf{L}^T = \mathbf{L}\mathbf{L}^T = \boldsymbol{\Sigma}$$

(I don't know an elementary proof, i.e., without a matrix decomposition. Is there one?)

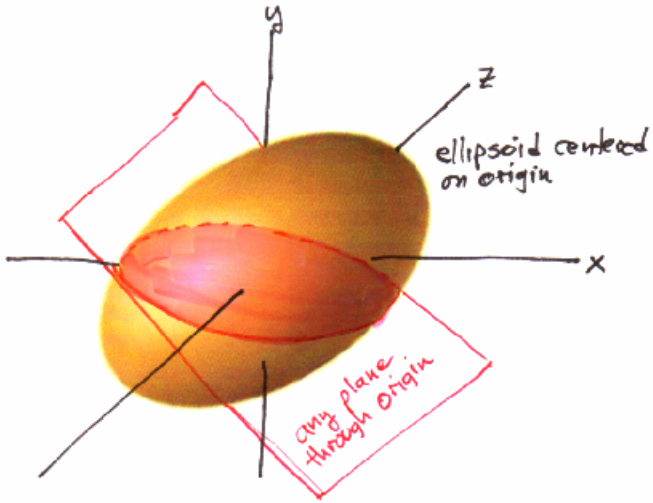
Reduced dimension properties of multivariate normal

$$N(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{M/2} \det(\boldsymbol{\Sigma})^{1/2}} \exp[-\tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu}) \cdot \boldsymbol{\Sigma}^{-1} \cdot (\mathbf{x} - \boldsymbol{\mu})]$$

1. Any slice through a m.v.n. is a m.v.n ("constraint" or "conditioning")

2. Any projection of a m.v.n. is a m.v.n ("marginalization")

You can prove both assertions by "completing the square" in the exponential, producing an exponential in (only) the reduced dimension times an exponential in (only) the lost dimensions. Then the second exponential is either constant (slice case) or can be integrated over (projection case).



Other Cholesky tricks, e.g., generate multivariate normal deviates:

There is a quite general way to construct a vector deviate $\mathbf{x}$ with a specified covariance $\mathbf{\Sigma}$ and mean $\boldsymbol{\mu}$, starting with a vector $\mathbf{y}$ of independent random deviates of zero mean and unit variance: First, use Cholesky decomposition (§2.9) to factor $\mathbf{\Sigma}$ into a left triangular matrix $\mathbf{L}$ times its transpose,

$$\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^T \tag{7.4.2}$$

This is always possible because $\mathbf{\Sigma}$ is positive-definite, and you need do it only once for each distinct $\mathbf{\Sigma}$ of interest. Next, whenever you want a new deviate $\mathbf{x}$, fill $\mathbf{y}$ with independent deviates of unit variance and then construct

$$\mathbf{x} = \mathbf{L}\mathbf{y} + \boldsymbol{\mu} \tag{7.4.3}$$

The proof is straightforward, with angle brackets denoting expectation values: Since the components y_i are independent with unit variance, we have

$$\langle \mathbf{y} \otimes \mathbf{y} \rangle = \mathbf{1} \tag{7.4.4}$$

where $\mathbf{1}$ is the identity matrix. Then,

$$\begin{aligned} \langle (\mathbf{x} - \boldsymbol{\mu}) \otimes (\mathbf{x} - \boldsymbol{\mu}) \rangle &= \langle (\mathbf{L}\mathbf{y}) \otimes (\mathbf{L}\mathbf{y}) \rangle \\ &= \left\langle \mathbf{L}(\mathbf{y} \otimes \mathbf{y})\mathbf{L}^T \right\rangle = \mathbf{L} \langle \mathbf{y} \otimes \mathbf{y} \rangle \mathbf{L}^T \\ &= \mathbf{L}\mathbf{L}^T = \mathbf{\Sigma} \end{aligned} \tag{7.4.5}$$

So, just take: $y_i \sim \mathrm{N}(0,1)$ and you get a multivariate normal deviate!

A related, useful, Cholesky trick is to draw error ellipses (ellipsoids, …)

$$\Sigma = \mathbf{L} \cdot \mathbf{L}^T$$

So, locus of points at 1 standard deviation is

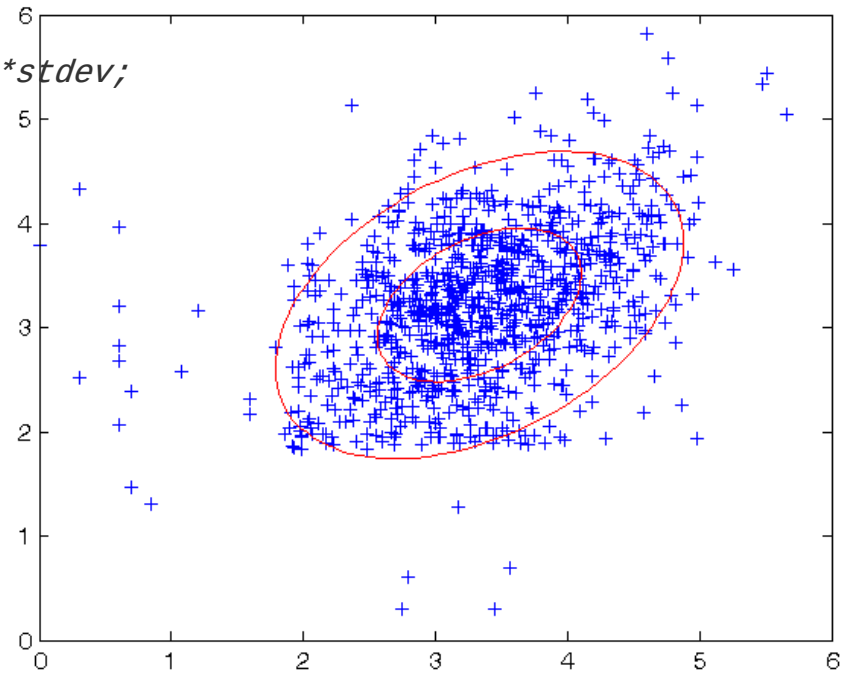
$$1 = (\mathbf{x} - \mu) \cdot \Sigma^{-1} \cdot (\mathbf{x} - \mu) \quad \Rightarrow \quad \left| \mathbf{L}^{-1} \cdot (\mathbf{x} - \mu) \right| = 1$$

If z is on the unit circle (sphere, …) then

$$\mathbf{x} = \mathbf{L} \cdot \mathbf{z} + \mu$$

```
function [x y] = errorellipse(mu,sigma,stdev,n)
L = chol(sigma,'lower');
circle = [cos(2*pi*(0:n)/n); sin(2*pi*(0:n)/n)].*stdev;
ellipse = L*circle + repmat(mu,[1,n+1]);
x = ellipse(1,:);
y = ellipse(2,:);
```

```
plot(i1llen,i2llen,'+b');
hold on
[xx yy] = errorellipse(mu,sig,1,100);
plot(xx,yy,'-r');
[xx yy] = errorellipse(mu,sig,2,100);
plot(xx,yy,'-r')
```



Weighted Nonlinear Least Squares Fitting
a.k.a. χ^2 Fitting
a.k.a. Maximum Likelihood Estimation of Parameters (MLE)
a.k.a. Bayesian parameter estimation
(with uniform prior and maybe some other normality assumptions)

these are not all exactly identical,
but they're real close!

$$y_i = y(\mathbf{x}_i|\mathbf{b}) + e_i$$

measured values supposed to be a model, plus an error term

$$e_i \sim N(0, \sigma_i)$$

the errors are Normal, either independently…

$$\mathbf{e} \sim N(0, \mathbf{\Sigma})$$

… or else with errors correlated in some known way (e.g., multivariate Normal)

We want to find the parameters of the model **b** from the data.

Do the Bayes thing!

$$P(\mathbf{b}|\{y_i\}) \propto P(\{y_i\}|\mathbf{b})P(\mathbf{b})$$
$$\propto \prod_i \exp\left[-\tfrac{1}{2}\left(\frac{y_i - y(\mathbf{x}_i|\mathbf{b})}{\sigma_i}\right)^2\right] P(\mathbf{b})$$
$$\propto \exp\left[-\tfrac{1}{2}\sum_i \left(\frac{y_i - y(\mathbf{x}_i|\mathbf{b})}{\sigma_i}\right)^2\right] P(\mathbf{b})$$
$$\propto \exp[-\tfrac{1}{2}\chi^2]P(\mathbf{b})$$

Now the idea is: Find (somehow!) the parameter value $\mathbf{b}_0$ that minimizes χ^2 .

For linear models, you can solve linear "normal equations" or, better, use Singular Value Decomposition. See NR3 section 15.4

In the general nonlinear case, you have a general minimization problem, for which there are various algorithms, none perfect.

Those parameters are the MLE. (Note that it is implictly Bayes with uniform prior.)

How "accurate" are they?

How accurately are the fitted parameters determined?
As Bayesians, we would instead say, what is their posterior distribution?

Taylor series:

$$-\tfrac{1}{2}\chi^2(\mathbf{b}) \approx -\tfrac{1}{2}\chi^2_{\min} - \tfrac{1}{2}(\mathbf{b}-\mathbf{b}_0)^T \left[\tfrac{1}{2}\frac{\partial^2\chi^2}{\partial\mathbf{b}\partial\mathbf{b}}\right](\mathbf{b}-\mathbf{b}_0)$$

So, while exploring the χ^2 surface to find its minimum, we must also calculate the Hessian (2nd derivative) matrix at the minimum.

Then

$$P(\mathbf{b}|\{y_i\}) \propto \exp\left[-\tfrac{1}{2}(\mathbf{b}-\mathbf{b}_0)^T\boldsymbol{\Sigma}_b^{-1}(\mathbf{b}-\mathbf{b}_0)\right]P(\mathbf{b})$$

with

$$\boldsymbol{\Sigma}_b = \left[\tfrac{1}{2}\frac{\partial^2\chi^2}{\partial\mathbf{b}\partial\mathbf{b}}\right]^{-1}$$

covariance (or "standard error") matrix of the fitted parameters

We have obtained the covariance structure of all the parameters, and indeed (at least in CLT normal approximation) their entire joint distribution!

But what if we want confidence limits on one parameter at a time?
Or maybe a confidence ellipse on two parameters?

Condition or Marginalize uninteresting parameters? (Know the difference!)

$$P(\mathbf{b}|\{y_i\}) \propto \exp\left[-\tfrac{1}{2}(\mathbf{b}-\mathbf{b}_0)^T\boldsymbol{\Sigma}_b^{-1}(\mathbf{b}-\mathbf{b}_0)\right] P(\mathbf{b})$$

Condition: (this is rare!) Fix uninteresting parameters at specified values.

In $\boldsymbol{\Sigma}_b^{-1} = \left[\frac{1}{2}\frac{\partial^2\chi^2}{\partial\mathbf{b}\partial\mathbf{b}}\right]$ submatrix of *interesting* rows and columns is new $\boldsymbol{\Sigma}_b^{-1}$

Take matrix inverse if you want their covariance $\boldsymbol{\Sigma}_b$

(If you fix parameters at other than $\mathbf{b_0}$, the mean also shifts – exercise for reader!)

Marginalize: (this is usual) Ignore (integrate over) uninteresting parameters.

In $\boldsymbol{\Sigma}_b = \left[\frac{1}{2}\frac{\partial^2\chi^2}{\partial\mathbf{b}\partial\mathbf{b}}\right]^{-1}$ submatrix of *interesting* rows and columns is new $\boldsymbol{\Sigma}_b$

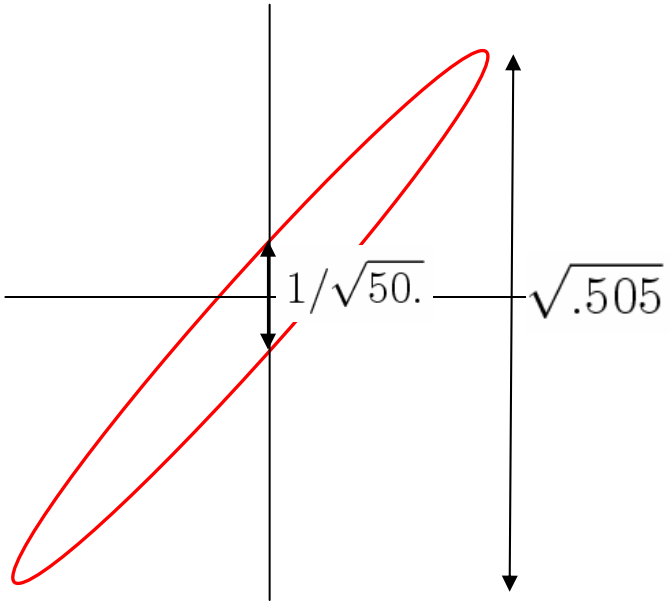
Special case of one variable at a time: Just take diagonal components in $\boldsymbol{\Sigma}_b$

Take matrix inverse if you want the distribution of interestings (see top line).

Why does this work? Roughly: (1) any marginalization of Gaussian is Gaussian [complete the square to separate interesting vs. uninteresting], and (2) variances are pairwise expectations and don't depend on whether other parameters are interesting or not.

Example:

$$P(\mathbf{b}|\{y_i\}) \propto \exp\left[-\tfrac{1}{2}(\mathbf{b}-\mathbf{b}_0)^T\boldsymbol{\Sigma}_b^{-1}(\mathbf{b}-\mathbf{b}_0)\right] P(\mathbf{b})$$



$$\boldsymbol{\Sigma}_b^{-1} = \left[\tfrac{1}{2}\frac{\partial^2\chi^2}{\partial\mathbf{b}\partial\mathbf{b}}\right] = \begin{pmatrix} 50. & -49. \\ -49. & 50. \end{pmatrix}$$

$$\boldsymbol{\Sigma}_b = \begin{pmatrix} .505 & .495 \\ .495 & .505 \end{pmatrix}$$

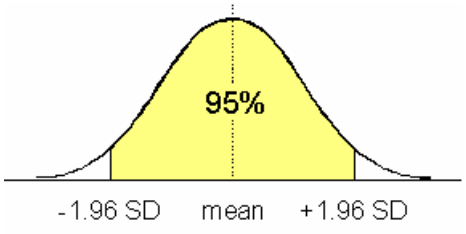
By the way, don't confuse the "covariance matrix of the fitted parameters" with the "covariance matrix of the data". For example, the data covariance is often diagonal (uncorrelated σ_i's), while the parameters covariance is essentially never diagonal!

If the data has correlated errors, then the starting point for χ^2 is:

$$\chi^2 = [\mathbf{y}_{\{\mathbf{i}\}} - \mathbf{y}(\mathbf{x}_{\{i\}}|\mathbf{b})]^T\boldsymbol{\Sigma}^{-1}[\mathbf{y}_{\{\mathbf{i}\}} - \mathbf{y}(\mathbf{x}_{\{i\}}|\mathbf{b})] \quad \text{instead of} \quad \sum_i \left(\frac{y_i - y(\mathbf{x}_i|\mathbf{b})}{\sigma_i}\right)^2$$

Confidence intervals or regions

The variances of *one parameter* at a time imply confidence intervals as for an ordinary 1-dimensional normal distribution:

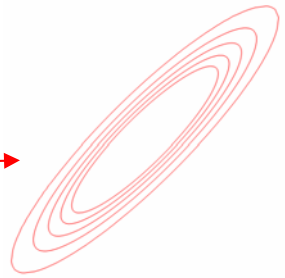


(Remember to take the square root of the variances to get the standard deviations!)

If you want to give confidence regions for *more than one parameter* at a time, you have to decide on a shape, since any shape containing 95% (or whatever) of the probability is a 95% confidence region!

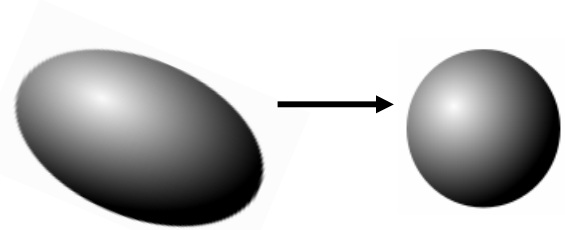
It is *conventional* to use contours of probability density as the shapes (= contours of $\Delta\chi^2$) since these are maximally compact.

But which $\Delta\chi^2$ contour contains 95% of the probability?

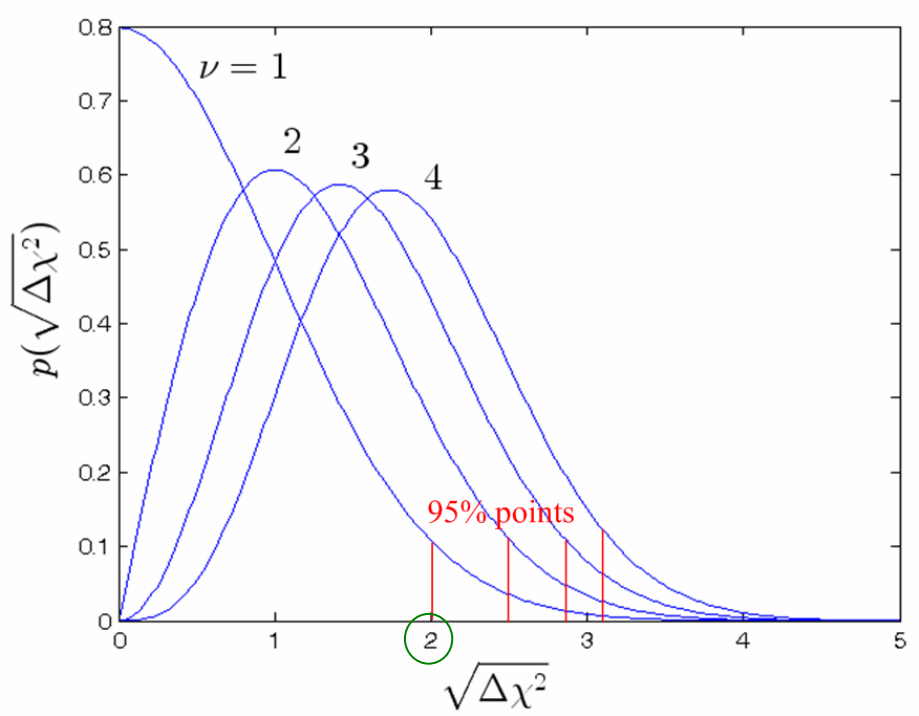


What $\Delta\chi^2$ contour in ν dimensions contains some percentile probability?

Rotate and scale the covariance to make it spherical.
(Linear, so contours still contain same probability.)



Now, each dimension is an independent Normal, so Chisquare(ν) is by definition the distribution of radius squared (sum of ν individual t^2 values)!



$\Delta\chi^2$ as a Function of Confidence Level p and Number of Parameters of Interest ν						
	ν					
p	1	2	3	4	5	6
68.27%	1.00	2.30	3.53	4.72	5.89	7.04
90%	2.71	4.61	6.25	7.78	9.24	10.6
95.45%	4.00	6.18	8.02	9.72	11.3	12.8
99%	6.63	9.21	11.3	13.3	15.1	16.8
99.73%	9.00	11.8	14.2	16.3	18.2	20.1
99.99%	15.1	18.4	21.1	23.5	25.7	27.9

Frequentists love MLE estimates (and not just the case with a Normal error model) because they have provably nice properties asymptotically

- Consistency: converges to true value of the parameters
- Equivariance: estimate of function of parameter = function of estimate of parameter
- asymptotically Normal
- asymptotically efficient (optimal): among estimators with the above properties, it has the smallest variance

The "Fisher Information Matrix" is another name for the Hessian of the log probability (or, rather, log likelihood):

$$\mathbf{I}(\mathbf{b}) = -\left\langle \frac{\partial^2 \log P(\{y_i\}|\mathbf{b})}{\partial \mathbf{b} \partial \mathbf{b}} \right\rangle \approx 2\boldsymbol{\Sigma}_b^{-1}$$

except that, strictly speaking, it is an expectation over the population

Bayesians tolerate MLE estimates because they are almost Bayesian – even better if you put the prior back into the minimization.

But Bayesians keep in mind that we live in a non-asymptotic world!