Normal Distribution

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

- Arises ubiquitously, via Central limit theorem

mean:	μ
Variance:	σ^2
Range:	$x \in (-\infty, \infty)$, $\mathbb{R}$
Parameters:	μ, (all reals) and σ (>0)

- Revisit Table 4.1, and Fig 4.17, showing relationship btwn distributions.
  - Collectively, read/work through simdata_MLE.R
    - ↳ up to line 87
    - ↳ Demo R probability functions, using -pois()
    - ↳ have students follow allong.
  - Show distribution_demos.cdf tool.

Likelihood

In our discussion of probability theory and distributions, we've written the distributions in the general form

$$P(x|\theta) = f(x),$$ - where x varies and represents our data or events
- and where our statistical parameters $\theta = \{\theta_1, \theta_2, \ldots\}$ etc. are assumed to be fixed.

Likelihood cont.

If we reverse this assumption, so that now we consider our data to be fixed, and we allow our statistical parameters to vary, our probability distribution/density function becomes a likelihood function (and is otherwise unchanged)

$$P(\theta|x) = \mathcal{L}(\theta|x) = f(\theta)$$

And now, instead of talking about the probability of our data, we refer to the likelihood of our parameters, θ.

Let's turn to some specific examples:

Recall $$P(x|\lambda) = P(x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

Then written as a likelihood,

$$\mathcal{L}(\lambda|x) = P(x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$$

If we have 1 observation or data point, and observed that $x = 6$ seeds fell in a square during our observation

What is the likelihood that $\lambda = 4$?

$$\mathcal{L}(\lambda = 4 | x = 6) = \frac{e^{-4}4^6}{6!} \approx 0.104 \quad \text{(via R)}$$

or, generally, $\mathcal{L}(\lambda|x=6) = e^{-\lambda}\lambda^6/6!$

How often do we get just one data point?

Suppose $x = \{6, 7, 5, 8, 3\}$ is our data

Just as with probabilities, if we want to find the joint likelihood that $\lambda = \lambda^*$ over a set of X observations,

$$\mathcal{L}(\lambda = \lambda^* | x = \{6,7,5,8,3\}) = \mathcal{L}(\lambda = \lambda^* | x_1 = 6)\,\mathcal{L}(\lambda = \lambda^* | x_2 = 7) \ldots$$

We can multiply together the likelihoods corresponding to individual data points.

$$\mathcal{L}(\lambda | x = \{\ldots\}) = \frac{e^{-5\lambda}\lambda^{(6+7+5+8+3)}}{6!\,7!\,5!\,8!\,3!}$$

Or, written more generally for some vector of data $x = \{x_1, x_2, \ldots x_n\}$

$$\mathcal{L}(\lambda|x) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$$

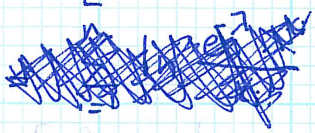
It's often helpful to examine the log of the likelihood function (especially for numerical methods we'll talk about later)

We'll denote this as

$$\mathcal{LL}(\lambda|x) = \ln \mathcal{L}(\lambda|x) = \ln\left[\prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right]$$

which we can simplify to

$$\mathcal{LL}(\lambda|x) = \ln\left[\prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right]$$



Note:

① $\prod_{i=1}^{n} e^{-\lambda} = e^{-n\lambda}$

② $\prod_{i=1}^{n} \lambda^{x_i} = \lambda^{\sum_{i=1}^{n} x_i}$

$$\mathcal{LL}(\lambda|x) = \ln\left[\frac{e^{-n\lambda}\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}\right]$$

$$\mathcal{LL}(\lambda|x) = \ln e^{-n\lambda} + \ln \lambda^{\sum_{i=1}^{n} x_i} - \ln \prod_{i=1}^{n} x_i!$$

$$\mathcal{LL}(\lambda|x) = -n\lambda + \left(\sum_{i=1}^{n} x_i\right)\ln\lambda - \sum_{i=1}^{n} \ln x_i!$$

The Big Question: What value of λ corresponds to the maximum of the likelihood function?

Same? as

↳ What value of λ corresponds to the max of the log likelihood function?

(*) log transformation changes the relative values of the function, but not the location of any max/min values.

Digression: Locating Max/Min of functions via calculus

$\frac{df}{dx} = f'(x) \overset{\text{set}}{=} 0$, Solve for x^* (max, min, or inflection).

Then $\left.\frac{d^2f}{dx^2}\right|_{x=x^*}$, if $> 0 \Rightarrow x^*$ min

$< 0 \Rightarrow x^*$ max

To continue our analysis,

$$\frac{d\,\mathcal{L}}{d\lambda} = ? = \frac{d}{d\lambda}\left(-n\lambda + \ln\lambda \sum_{i=1}^{n} x_i + -\sum_{i=1}^{n} \ln x_i!\right)$$

$$= -n + \frac{d}{d\lambda}\left(\ln\lambda \sum_{i=1}^{n} x_i\right) + 0$$

$$= -n + \frac{1}{\lambda}\sum_{i=1}^{n} x_i$$

Set

$$0 = -n + \frac{1}{\lambda}\sum_{i=1}^{n} x_i$$

$$n = \frac{1}{\lambda}\sum_{i=1}^{n} x_i$$

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

So, $\hat{\lambda}$ is our estimate of the value of λ corresponding to the max or min of the log likelihood function.

Check 2nd derivative.

$$\frac{d^2\,\mathcal{L}}{d\lambda^2} = -\lambda^{-2}\sum_{i=1}^{n} x_i, \quad \left.\frac{d^2\,\mathcal{L}}{d\lambda^2}\right|_{\lambda=\hat{\lambda}} = \frac{-\sum_{i=1}^{n} x_i}{\frac{1}{n^2}\left(\sum_{i=1}^{n} x_i\right)^2}$$

$$= \frac{-n^2}{\sum_{i=1}^{n} x_i} \quad \text{which is } < 0, \text{ so } \hat{\lambda} \text{ is a max}$$

We can apply this general solution to the data set we examined earlier

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} = \frac{1}{5}(6+7+5+8+3)$$

$$\hat{\lambda} = \frac{1}{5}\cdot 29 = 5.8$$

< switch back to simdata_MLE.R >

This general approach can be applied to help us estimate the parameters of any probability distrib. given observed data. (although sometimes only w/ the help of computers where analytical solutions aren't possible)

Thus this maximum likelihood approach is valid and useful as a means of estimating parameters that allow us to describe variation in data (ties in to morning's discussion of Stats philosophy).

- in upcoming labs/lectures, we'll discuss how Maximum likelihood can be used to fit more complex/interesting models, as well as being used to compare models, draw inferences, and make predictions.

Model Comparison

We've been bumping into this topic (but skirting it) since the beginning of class (discussing Stats philosophy), and in lab (with AIC comparison). But Now we'll face it head on (especially interesting as we have interesting models to compare now).

(*) When we maximized our likelihood functions and decided that the value(s) of the parameters yielding the highest likelihood were the "best" estimate(s) of statistical params,

- We implicitly made a decision based on differences in likelihood. (or comparing likelihoods).

In the same way that we can compare likelihoods of par. values w/in a model to reach a conclusion, we can compare likelihoods (specif. the max like's) across models to gain a sense of how likely (or "probable") one model or another is relative to the data that we have

- Essential: can only compare likelihoods of models explaining the same Y data.

We also have a sense of parsimony

- Complex models should have ↑ likelihoods due to ↑ complexity alone, but we want a sense of whether an increase in complexity increases likelihood more than due to parameters, indicating it actually has some explanatory traction.

(*) First we'll discuss this kind of comparison / competition between models, then investigate quantifying uncertainty w/in models

(Comparison of / Uncertainty)

Akaike information Criteria, Hirotsugu Akaike, 1974

See EMD book pgs 209 - 212

$$AIC = 2k - 2\ln(\mathcal{L})$$

$$= 2k - 2LL$$

k = # of pars

LL ← how likely the model is given the data investigated

we can arrive @ different AIC values for each model investigated, and compare differences.

Rules of Thumb (vague)

ΔAIC	Support for model w/ lower AIC
0-2	~~Weak~~ equivalent
~~4-7~~	~~Positive~~ clearly distinguishable
~~Strong~~	
>10	~~Very Strong~~ definitely different

Note: AIC doesn't directly provide a test for a classic Null Hypothesis (how well a model fits data in absolute sense, w/ p-values, etc)

→ the "best" model is only judged to be the best of those examined to begin with. If all models are crappy, the best of crap is still crap.

↳ simpler "Null models" can be included in the model set however

(*) AIC doesn't account for small sample sizes (only valid asymptotically)

$$AICc = AIC + \frac{2k(k+1)}{n-k-1}$$ (think abt $n \to \infty$)

Many others exist, including

$$BIC = \ln(n)K - 2LL$$ (more conservative than AIC when $n > 8$ obsvs. or So → $\ln(8) \approx 2$

Model Weights

$$w_i = \frac{e^{-\Delta AIC_i/2}}{\sum_{j=1}^{n} e^{-\Delta AIC_j/2}}$$

} Useful for model averaging.

where $\Delta AIC_i = AIC_i - AIC_{\text{Best Model}}$

As $\Delta AIC_i \to \infty$, $e^{-\Delta AIC_i/2} \to 0$, so $w_i \to 0$

Model Averaging

- What is it?

  Compete $\theta \sim$ model 1 or $\theta \sim$ model 2

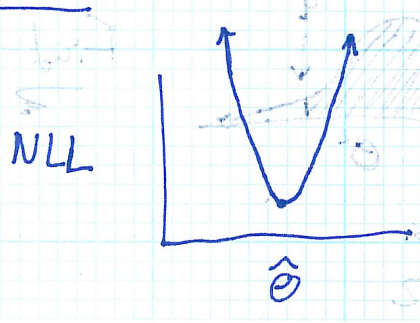
  ↳ If both are somewhat equivalent such that we can't justify choosing 1 or 2 over the other, we don't have to give up / pretend we know nothing

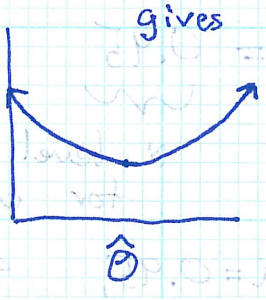
  - Can make predictions by model averaging given the AIC weights
  - Important, pragmatic solution allowing action in the face of scientific uncertainty.

$\theta \sim w_1 \cdot$ Model 1 $+ w_2 \cdot$ Model 2 ... and so on.

The model comparison techniques we've just investigated allow us to compare a version of likelihoods across models. Now let's return to considering uncertainty within a model in our estimate of parameters.

Motivation!



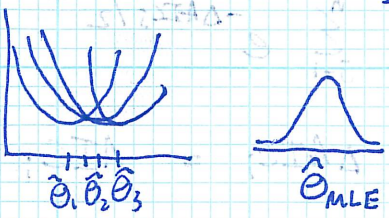


Curvature of NLL function around $\hat{\theta}$ gives us information about how certain we can be in our estimate of the point value $\hat{\theta}$.

It's known that the distribution of MLE estimates $\hat{\theta}$ (say across repeated, duplicate data sets) follows a normal distribution for large enough sample sizes

EMD book pg 191

- Assuming that $-LL(\theta|x)$ is Quadratic around $\hat{\theta}_{MLE}$

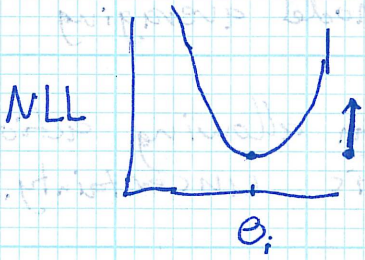


then changes in NLL resulting from variation in $\hat{\theta}_{MLE}$ Will follow

$$\sum_{i=1}^{r} \hat{\theta}_i^2$$ (for r # of parameters being estimated)

when $\hat{\theta}$ is normally distributed, $\sum_{i=1}^{r} \hat{\theta}_i^2 \sim \chi^2(r)$

The punch line is that if we look @ Δ NLL w/ variation in some particular $\hat{\theta}_i$:



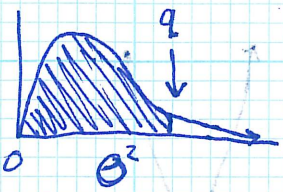
(1 df because in 1 case we fix $\theta_i = \hat{\theta}_i$ and otherwise are considering variation in θ)

This means that we can use the χ^2 distribution to tell us in some probabilistic sense when changing θ_i results in a substatial enough shift in NLL as to be significantly different from our $\hat{\theta}_i$

↳ Ie, we can estimate CI's for $\hat{\theta}_i$

$$\int_0^q \chi^2(1)\, d\theta^2 = 0.95$$

α level for our CI's.

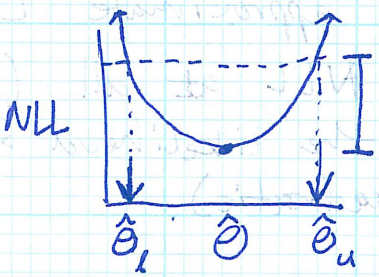


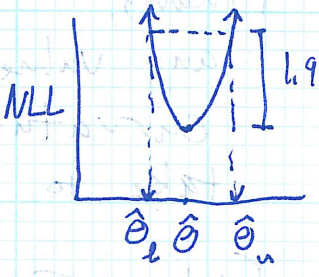
Find q

For $\chi^2(1)$ and $\alpha = 0.95$, $q = 1.92$

So, if we can find

95% CI for $\hat{\theta}$ is $(\hat{\theta}_l, \hat{\theta}_u)$





(Smaller CI's = steeper NLL)

For models with multiple parameters, $\hat{\Theta} = \{\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_n\}$

We can look @ the likelihood surface in 2 different ways:

① Likelihood Slices

Plot $-\mathbb{L}(\theta_1, \hat{\theta}_2, \hat{\theta}_3, \ldots, \hat{\theta}_n \mid x)$

ie, fix all θ_i at their $\hat{\theta}_i$ except θ_1 and calculate the $-\mathbb{L}$ value,

→ determine CI's as above.

② Likelihood Profiles

Plot $-\mathbb{L}(\theta_1, \hat{\theta}_2^*, \hat{\theta}_3^*, \ldots, \hat{\theta}_n^*)$ (fancy θ's)

where $\hat{\theta}_j^*$ is a new estimate of $\hat{\theta}_j$ given the particular value of θ_1 being queried

↳ essentially re-optimize over all other θ_j for each diff value of θ_i investigated.

Accounts for covariation btwn parameters (changing $\hat{\theta}_i$ makes some other choice of θ_j better than the current $\hat{\theta}_j$).

→ determine CI's as above.

* Less biased, much more computation

Finally, We can approximate CI's without calculating any values of NLL at all (based only on the local curvature of the likelihood surface, which we again take to be quadratic).

Fisher Information CI's

EMD Book pg 196-200

remember that $\frac{\partial^2 f}{\partial \theta^2}$ tells us about the curvature of function f w/ respect to changes in dimension θ.

Applied to a likelihood surface,

$$\frac{df(p)}{dp} = \frac{f(p+\Delta p) - f(p)}{\Delta p}$$

$$\left.\frac{d^2 f(p)}{dp^2}\right|_{p=\hat{p}} = \frac{f(\hat{p}+\Delta p) - 2f(\hat{p}) + f(\hat{p}-\Delta p)}{(\Delta p)^2}$$

$\left.\frac{d^2 l}{d\theta^2}\right|_{\theta=\hat{\theta}}$ tells us about the curvature of the likelihood of θ given data x around our MLE estimate of $\hat{\theta}$

When we have a multiparameter $\theta = \{\theta_1, \theta_2, \ldots, \theta_n\}$ we'll actually have a matrix of partial 2nd derivatives called a Hessian

$$\text{Hessian} = \begin{pmatrix} \frac{\partial^2 -LL}{\partial \theta_1^2} & \frac{\partial^2 -LL}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 -LL}{\partial \theta_1 \partial \theta_n} \\ \frac{\partial^2 -LL}{\partial \theta_1 \partial \theta_2} & \frac{\partial^2 -LL}{\partial \theta_2^2} & & \vdots \\ \cdots & & \ddots & \\ \frac{\partial^2 -LL}{\partial \theta_1 \partial \theta_n} & \cdots & & \frac{\partial^2 -LL}{\partial \theta_n^2} \end{pmatrix}$$

HESSIAN = Fisher Information

$-E(H)\big|_{\theta_{MLE}}$ = Observed information

Inverse of the Hessian is the estimate of the Variance-Covariance of the matrix of parameters. Don't want to invert these by hand

Vcov( ) helps us understand CI's for parameters in a manner similar to profiles

- when assumptions are satisfied.

width of the CI's should follow

- diagonals of VCov give parameter variance
- assumed normality
- sqrt gives std. dev / std. error values per parameter

$$\text{lower} = \hat{\theta} - 1.96 \cdot \text{s.e.}$$
$$\text{upper} = \hat{\theta} + 1.96 \cdot \text{se.}$$

↳ these are what you see in mle outputs, when it can estimate the Hessian

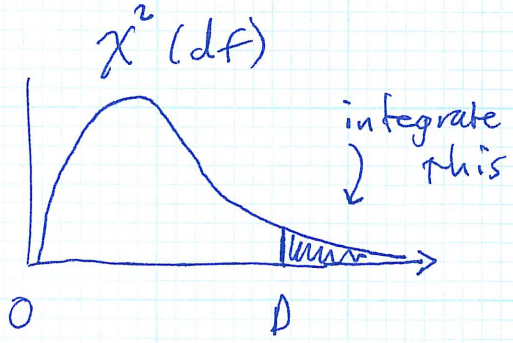
Likelihood Ratio Test – what is nested model

Likelihood ratio, $D = -2 \ln \left(\frac{\text{likelihood of reduced model}}{\text{likelihood of complex model}} \right)$

$$D = -2 \ln(\text{like reduced}) + 2 \ln(\text{alternative})$$

D is χ^2 distributed test statistic
w/ df equal to dif. btwn alternative & reduced

So we can do the same thing as w/ slices & profiles



... Viola, "p-value"

- examples of nested models, context of LTER fert & disturb.