

Likelihood

- What is likelihood?
- How can it be used to address scientific hypotheses?
- How can we use GLM's in a likelihood context?

Asking the questions we want to know

- Null hypothesis tests are criticized for asking the wrong question
- The right question: “Which hypothesis is best supported by the data?”
- We can use likelihood-based model selection to address this
- Before we learn to do this, we need to understand likelihood first

Likelihoods come from probability distributions, but aren't probabilities

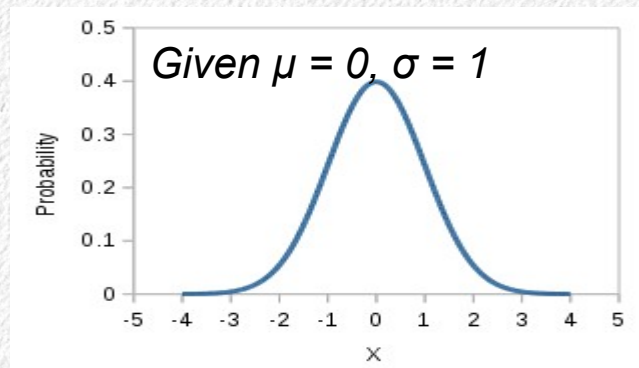
- Likelihood is calculated using a probability distribution, such as the normal →
- To calculate a probability density from the normal distribution, $p(x_i | \mu, \sigma)$:
 - Specify values for the parameters (μ, σ)
 - Calculate probability densities of data values (x_i) given the known parameter values
- To calculate a likelihood from the normal distribution, $L(\mu|x_i, \sigma)$:
 - Specify the value of a data point (x_i)
 - Calculate the likelihood of a possible value of the parameters given the known data
- The estimates of parameters that have the highest likelihood, given the data, are the maximum likelihood estimates

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left[\frac{x_i - \mu}{\sigma}\right]^2}$$

Probability and likelihood for single observations

Probability

$$p(x_i | \mu, \sigma)$$

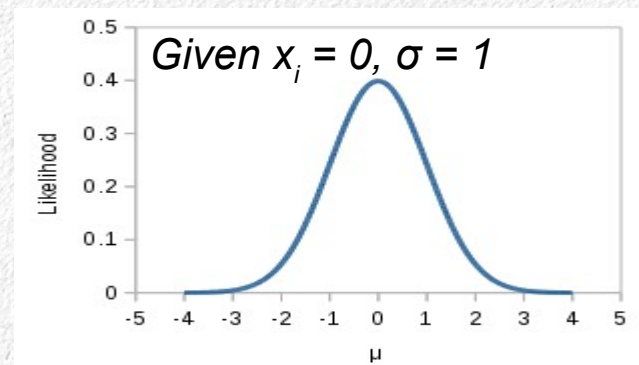


Centered on μ , calculate the probability density at values of x

Given known μ, σ , what is the probability density for a data value, x_i ?

Likelihood

$$L(\mu | x_i, \sigma)$$



Centered on an observed value of x , calculate likelihoods of possible values of μ

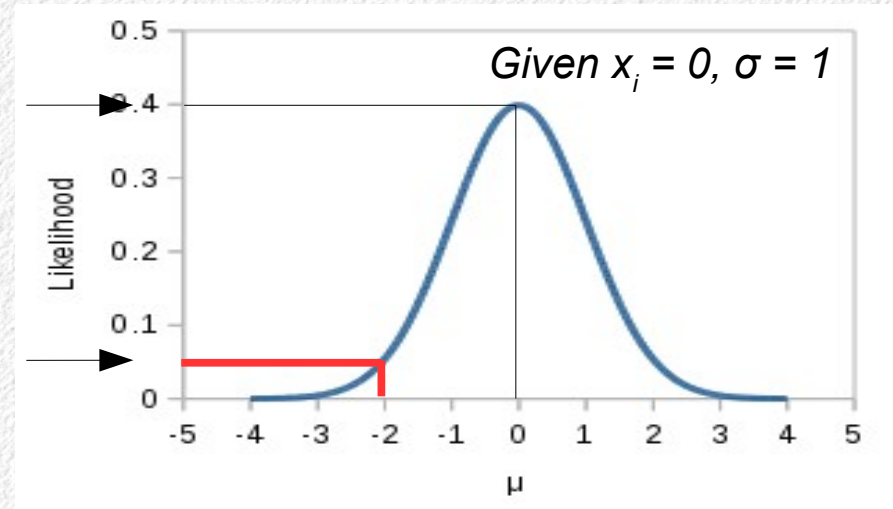
Given known x_i and σ , what is the likelihood of particular values of μ ?

Example: likelihood of two values of μ given x_i and σ

$$L(\mu | x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left[\frac{x_i - \mu}{\sigma}\right]^2}$$

$$L(0 | 0) = \frac{1}{\sqrt{2\pi 1^2}} e^{-\frac{1}{2}\left[\frac{0-0}{1}\right]^2} = 0.4$$

$$L(-2 | 0) = \frac{1}{\sqrt{2\pi 1^2}} e^{-\frac{1}{2}\left[\frac{0+2}{1}\right]^2} = 0.05$$



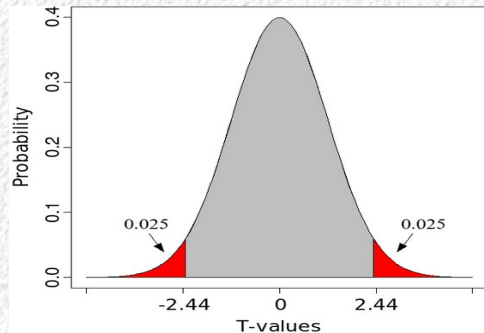
Same equation as a probability distribution

In this form, only difference between a probability distribution and a likelihood function is in the interpretation

Probability and likelihood for a sample of data points

Probability

$$p(\bar{x} \mid \mu, s_{\bar{x}})$$



Likelihood

$$\prod L(\mu \mid x_i, \sigma)$$

Likelihood of a parameter given a sample is the product of likelihoods of the parameter given each data point

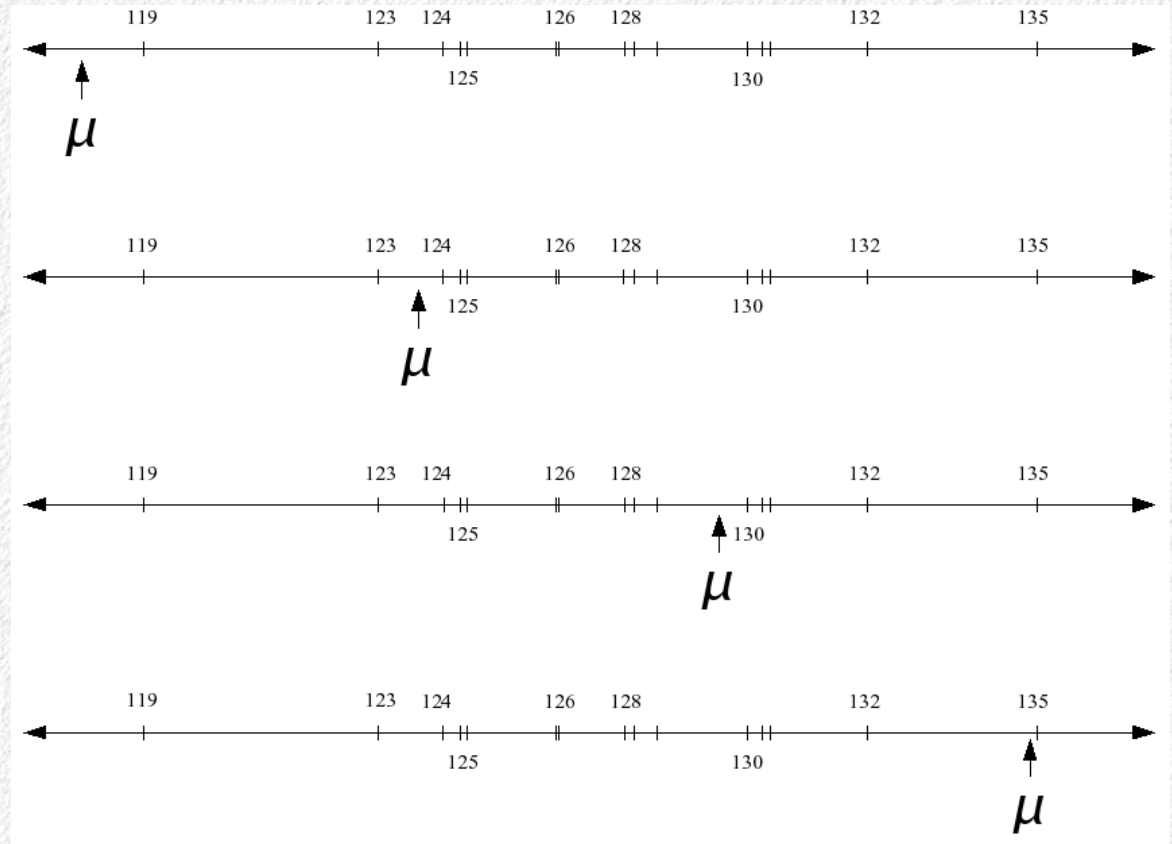
Use a sampling distribution, such as the t

Some nice features of likelihoods

- Likelihoods can be combined
 - Data can be added as it becomes available
 - Can add observations until two treatment groups diverge (big no-no with hypothesis testing)
- No “sampling” distributions
 - Likelihoods of samples are just products of likelihoods of individual observations
 - Parameter estimates and confidence intervals both obtained from the likelihood function
- Solutions can be found numerically
 - Works even when analytical solutions don't exist
 - Can estimate parameters and their SE's/confidence intervals even when they can't be measured directly

Using likelihoods for estimation

	The Data
	123.67
	126.90
	130.78
	125.30
	124.86
	126.96
	135.61
	119.42
	128.74
	132.53
	130.36
	128.31
	130.63
	128.13
	125.17



Infinite number of possible values of μ – which is best?

Principle of maximum likelihood

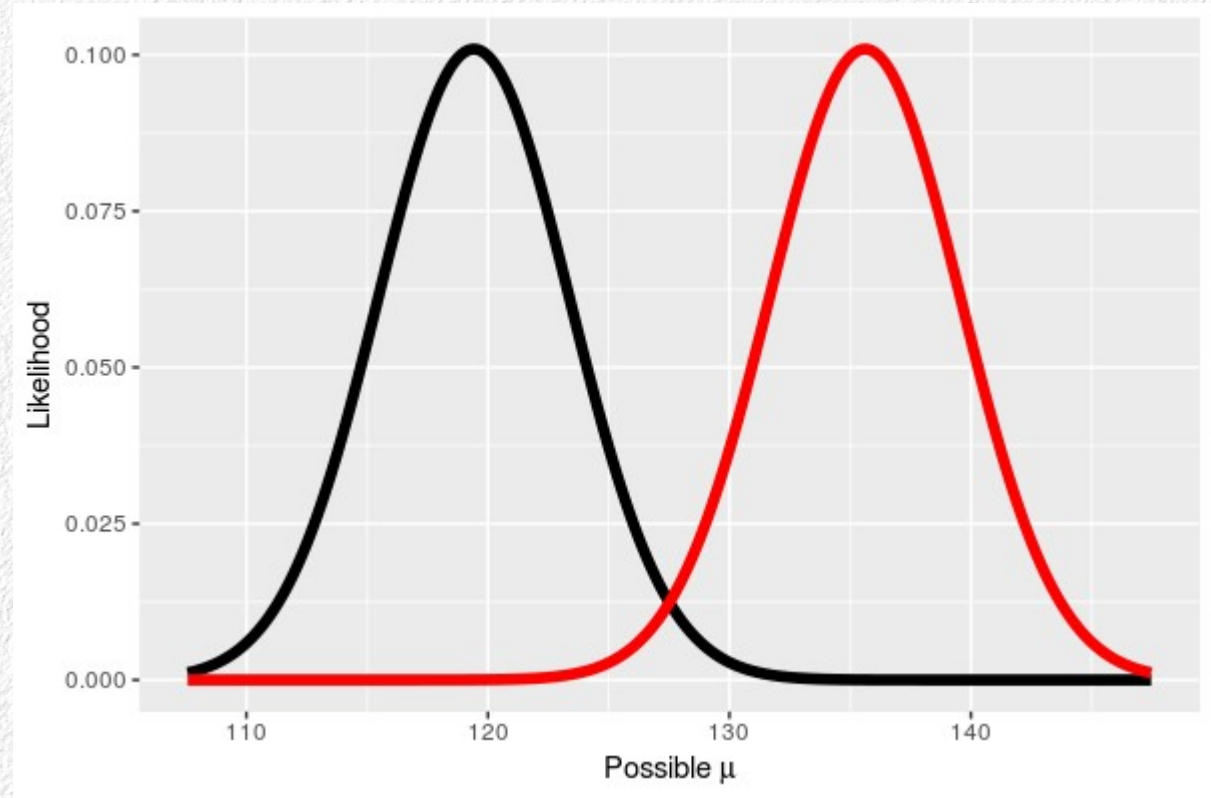
- The parameter value that is most likely given the data is the best estimate of the parameter
 - Whichever value of μ that maximizes the likelihood function is the best estimate of μ
 - Interpreted as the value of μ that is most likely to have given rise to the data
- Which value of μ maximizes the likelihood function for this sample of data?

Likelihood function for biggest and smallest data values

Centered on data values → curves peak at each data value

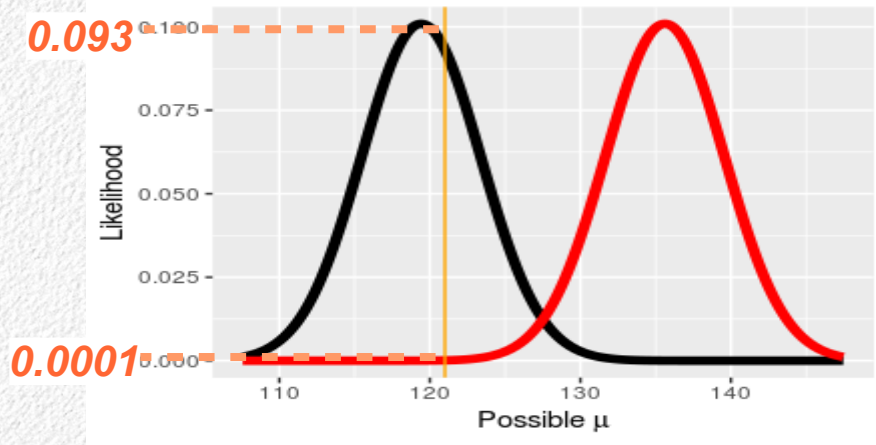
For each data value, the maximum likelihood estimate of μ is equal to the data value

What about the combination of both data values?

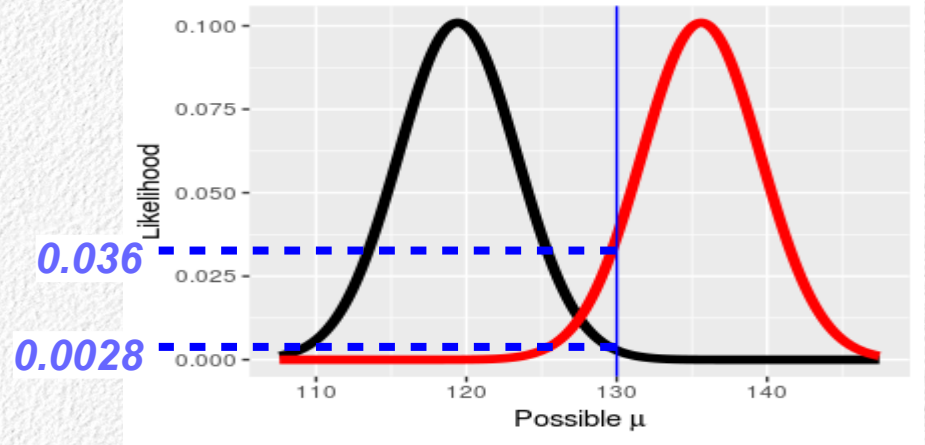


Two possible values of μ

$$L(121 \mid x=119.42, 135.61, \sigma=3.95)$$



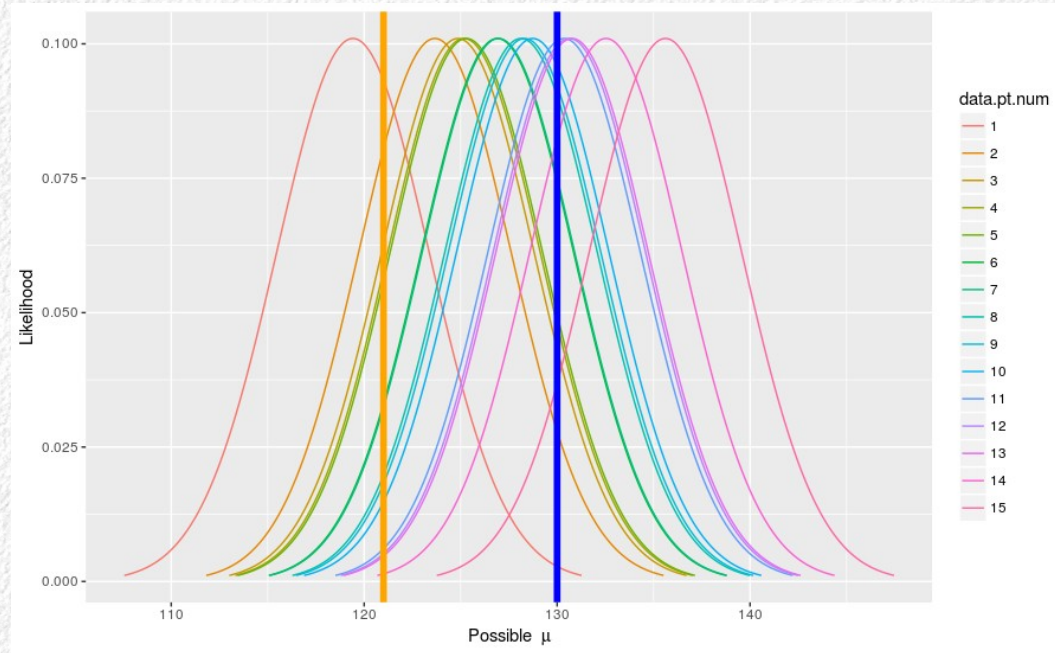
$$L(130 \mid x=119.42, 135.61, \sigma=3.95)$$



Likelihood of 121 is the product of likelihoods in orange: $0.093 \times 0.0001 = 0.0000093$

Likelihood of 130 is the product of likelihoods in blue: $0.0028 \times 0.036 = 0.0001$

Likelihood of two hypothetical values of μ given the data



Likelihood of 121 for each data point is where orange line intersects each likelihood function
Likelihood of 130 for each data point is where the blue line intersects each likelihood function

Likelihood of 121, 130 given the entire data set

		Likelihood of individual data points	
	The Data	Mean 121	Mean 130
	123.67	0.08	0.03
	126.90	0.03	0.07
	130.78	0.00	0.10
	125.30	0.06	0.05
	124.86	0.06	0.04
	126.96	0.03	0.08
	135.61	0.00	0.04
	119.42	0.09	0.00
	128.74	0.01	0.10
	132.53	0.00	0.08
	130.36	0.01	0.10
	128.31	0.02	0.09
	130.63	0.01	0.10
	128.13	0.02	0.09
	125.17	0.06	0.05
Mean	127.82		
Std. Dev.	3.95		

$$L(121|data) = 3.4 \times 10^{-28}$$

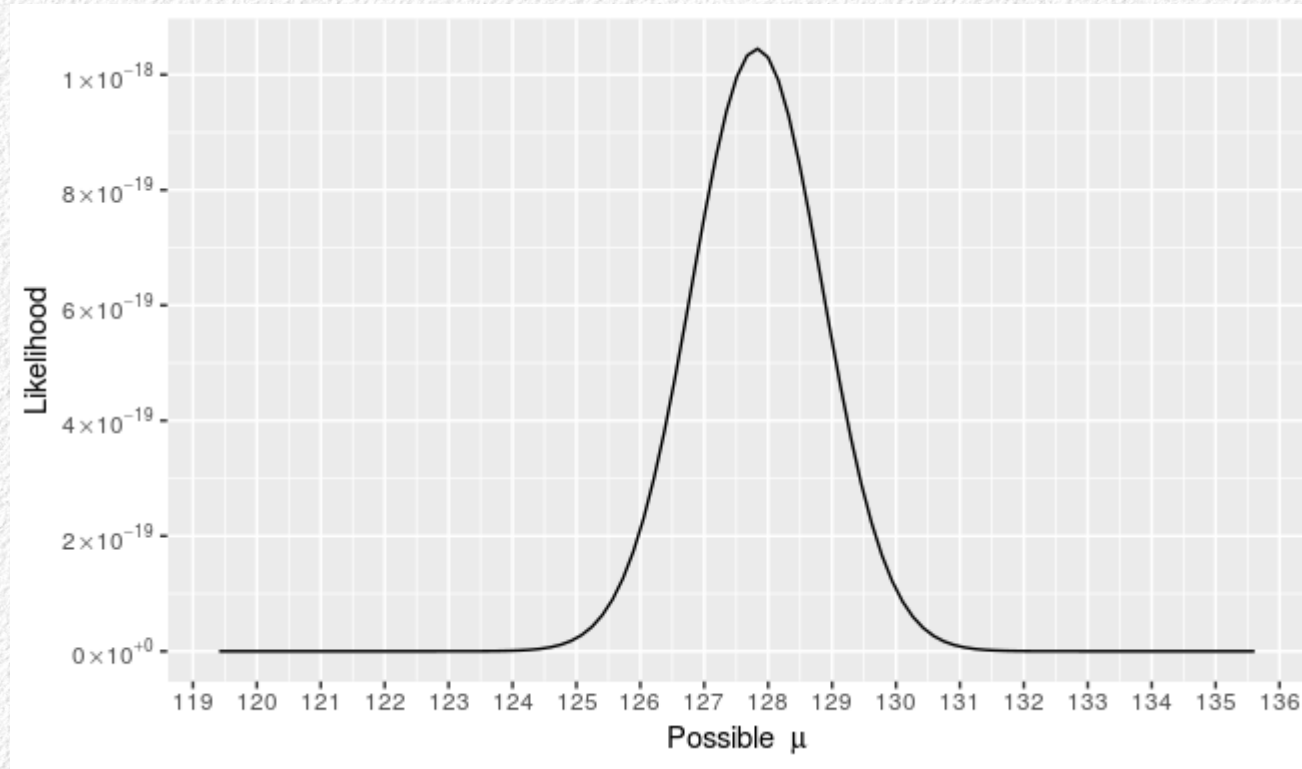
$$L(130|data) = 1.12 \times 10^{-19}$$

Which is bigger?

Individual likelihoods for two possible means

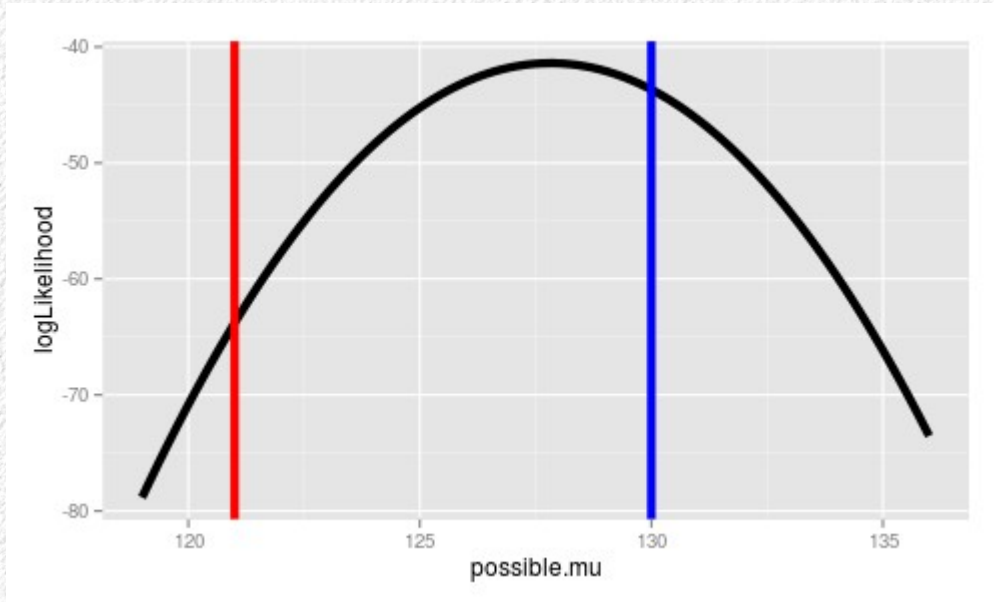
Product of these is the $L(\mu | data)$ for each

Likelihoods of possible μ from 119 to 136



What's the maximum likelihood estimate for μ ?

Ln(likelihood) changes the scale, makes likelihoods additive



$$Likelihood = \prod \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left[\frac{x_i - \mu}{\sigma}\right]^2}$$

$$Loglik = -0.5n \ln(2\pi) - 0.5n \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum (x_i - \mu)^2$$

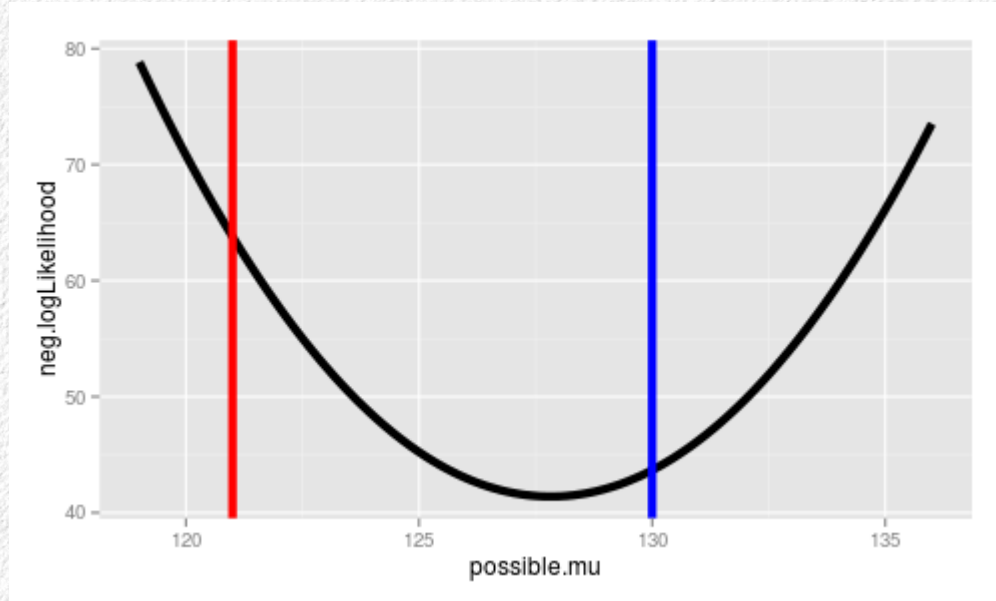
$$L(121|data) = 3.4 \times 10^{-28}$$

$$\longrightarrow \ln(L(121|data)) = -63.79$$

$$L(130|data) = 1.12 \times 10^{-19}$$

$$\longrightarrow \ln(L(130|data)) = -43.68$$

-Ln(likelihood) changes the direction



$$-\ln(L(121|data)) = 63.79$$

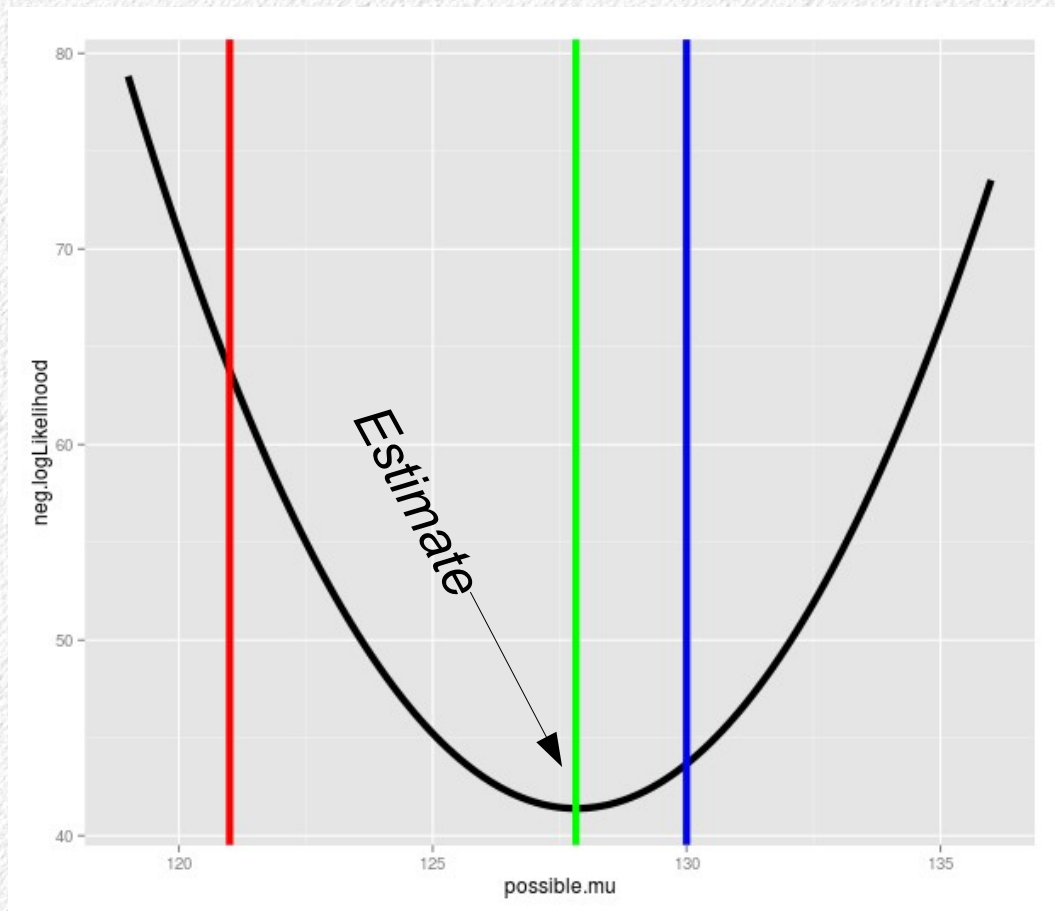
$$-\ln(L(130|data)) = 43.68$$

Why? Convenience (we are bad at negative numbers, and $-\ln(L)$ has some nice properties we'll meet later)

*But, to find the **maximum likelihood estimate**, we need to find the estimate with the **minimum** -logLikelihood value*

Likelihood of means given the data

Maximum likelihood estimate
of the mean is at the minimum of
this function, at 127.82



Sometimes there are analytical solutions for ML estimates

- We can find the minimum of the $-\log\text{Likelihood}$ function to come up with an analytical solution
 - Not always possible
 - When an analytical solution isn't possible, the estimates are derived **numerically**
 - Sophisticated form of trial and error
- We'll look at how it's done for the population mean, assuming the data are normally distributed
 - Find the first partial derivative of the likelihood function, set it to 0, and solve for μ

$$Loglik = -0.5 n \ln(2\pi) - 0.5 n \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum (x_i - \mu)^2$$

$$\frac{\partial Loglik}{\partial \mu} = \frac{-1}{\sigma^2} \sum (x_i - \mu)$$

$$0 = \frac{-1}{\sigma^2} \sum (x_i - \mu)$$

ML estimator for μ

$$0 = \sum (x_i - \mu)$$

$$0 = \sum (x_i) - n\mu$$

$$\hat{\mu} = \bar{x} = \frac{\sum (x_i)}{n} = \frac{1917.37}{15} = 127.82$$

So, \bar{x} is a maximum likelihood estimator for μ

Simplifying the likelihood function

- We can drop any term that is a constant, or that the parameters being estimated don't depend on
 - The values will be the same up to an additive constant \rightarrow shapes will be the same
 - Maximum will be at the same place
- If we drop terms, we still have a normal likelihood function, but it is no longer the normal probability distribution

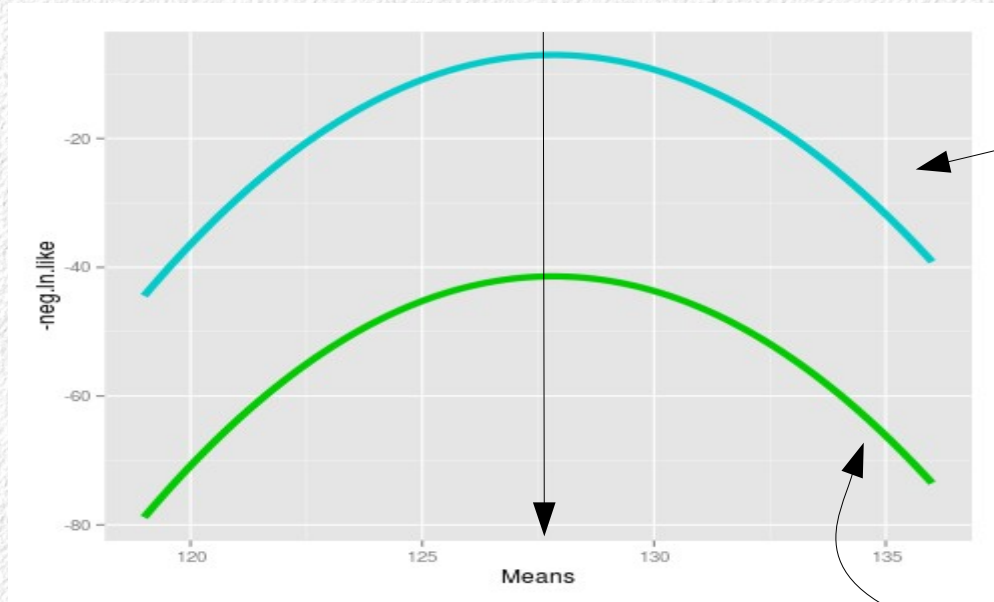
So, this \longrightarrow $-0.5 n \ln(2\pi) - 0.5 n \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum (x_i - \mu)^2$

and this

$\longrightarrow -\frac{1}{2\sigma^2} \sum (x_i - \mu)^2$

will give the same estimate of μ

$$-0.5 n \ln(2\pi) - 0.5 n \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum (x_i - \mu)^2$$



$$-\frac{1}{2\sigma^2} \sum (x_i - \mu)^2$$

Same shapes, differ by a constant amount across the whole curve
Both versions identify the same best value for the estimate of μ

Law of Likelihood

“Within the framework of a statistical model, a particular set of data supports one statistical hypothesis better than another if the likelihood of the first hypothesis is greater than the likelihood of the second hypothesis” *Edwards, 1992*

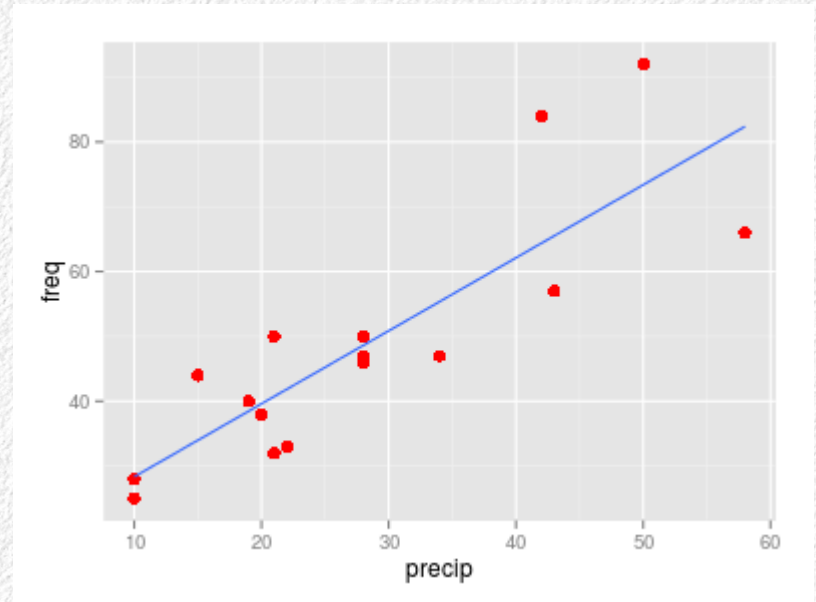
- We can treat a fitted model as a scientific hypothesis
- Compare the support for different hypotheses by comparing likelihoods of different models

Example: temperature and gene frequencies in butterflies

- Gene frequencies are variable among populations, suspect this is due to climatic conditions
- Elevation affects temperature and precipitation, but there will be some variation in temp and precipitation independent of elevation as well
- Collect data on four variables
 - Maximum temp
 - Minimum temp
 - Precipitation
 - Elevation
- Question: do the two temperature variables matter, above and beyond the effects of elevation and precipitation?

Likelihood of a model given the data

- Linear models predict mean values of y at a given value of x
- Assume random variation around the predicted values is normally distributed
 - We can check this (model criticism)
 - If it's true, then the maximum likelihood solution is also the least squares solution
- Can calculate likelihoods of residuals, combine them to be the likelihood of the model
- Simple example using just precip as a predictor first...



Modeled as a linear relationship with:

Intercept = 17.0956

Slope = 1.1258

What's the likelihood of this model given the data?

Residuals used to calculate log-likelihood of the model

- Log-likelihood of the model given each residual is calculated
 - Using normal likelihood function
 - \hat{y} = predicted allele freq.
 - = 17.0956 + 1.1258 (precip)
 - y_i are the data values
 - $\sigma^2 = \text{SSE}/n$
- Log-likelihoods of residuals summed to get the log-likelihood of the model given all the data

Frequency	y.hats	residuals	log.likelihood
57	65.50367	-8.5036709	-3.575789
38	39.61099	-1.6109893	-3.206743
46	48.61714	-2.6171395	-3.229262
47	48.61714	-1.6171395	-3.206848
50	48.61714	1.3828605	-3.203128
44	33.98215	10.0178545	-3.724245
50	40.73676	9.2632419	-3.647226
25	28.35330	-3.3533017	-3.252528
28	28.35330	-0.3533017	-3.193666
40	38.48522	1.5147794	-3.205151
33	41.86253	-8.8625269	-3.608778
66	82.39020	-16.3902024	-4.615038
47	55.37175	-8.3717520	-3.564005
32	40.73676	-8.7367581	-3.597061
84	64.37790	19.6220978	-5.231135
92	73.38405	18.6159477	-5.027478

Log likelihood = -59.088

$$-0.5n \ln(2\pi) - 0.5n \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum (y_i - \hat{y})^2$$

Residuals

With likelihoods, need more than one model to reach a conclusion

- Likelihoods only tell us anything in comparison with other likelihoods
- If we want to know if temperature measurements are important, beyond the effects of temperature and precipitation, we must:
 - Fit a model with all of the predictors, calculate its likelihood
 - Fit a model with the two temperature variables omitted, calculate its likelihood
 - Compare the likelihoods
- We know how to fit models, what about those likelihoods?

Likelihoods from standard GLM output


- R will give you log likelihoods if you ask nicely, but not all packages do
- You can calculate log likelihood from error SS, which is found in any ANOVA table:

$$\ln(L(model|data)) = -\frac{1}{2}n \ln\left(\frac{SSE}{n}\right)$$

- Note that n is not error degrees of freedom, it's sample size (so, SSE/n is **not** the error MS)

Response: freq

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
alt	1	3439.6	3439.6	67.775	1.632e-06	***
precip	1	1134.1	1134.1	22.346	0.0003954	***
Residuals	13	659.8	50.8			



$$-\frac{1}{2}(16)\ln\left(\frac{659.8}{16}\right)=-29.75$$

First hypothesis (H1)

Log-likelihoods for butterfly models

Response: freq

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
alt	1	3439.6	3439.6	106.6503	5.354e-07	***
precip	1	1134.1	1134.1	35.1628	9.876e-05	***
max.temp	1	296.3	296.3	9.1872	0.01143	*
min.temp	1	8.7	8.7	0.2695	0.61398	
Residuals	11	354.8	32.3			


$$-\frac{1}{2}(16)\ln\left(\frac{354.8}{16}\right)=-24.79$$

Second hypothesis (H2)

Which hypothesis has the higher likelihood, give the data?

Problem: the more complex model will have a higher likelihood

- The closer to the data values (i.e. the smaller the residuals) the higher the likelihood
- Taking a simple model and adding predictors improves fit, reduces the size of residuals → higher likelihood
- True even for terrible predictors, like random numbers
- Consequently, even though the model with temps had a higher likelihood, we can't be certain it's better supported by the data
- We can solve this problem by testing if the likelihood for the model with temps is statistically significantly better than the one without them

Testing differences in support for competing hypotheses

- **Statistical support** is defined as the log of the ratio of likelihoods for the two hypotheses

$$Support = \ln \left(\frac{L(H\ 1)}{L(H\ 2)} \right) = \ln(L(H\ 1)) - \ln(L(H\ 2))$$

- Negative twice the support follows a chi-square distribution that has d.f. equal to the difference in residual d.f. between the models
- So, we can use $-2(\text{support})$ as a χ^2 test statistic

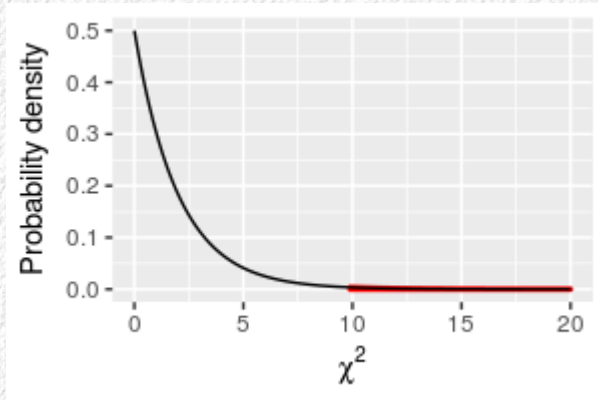
Likelihood ratio test comparing the two models

- In general, a support ratio can be used for a **likelihood ratio test**
- The test statistic is:

$$\chi^2 = -2 \ln \left(\frac{L(H_1)}{L(H_2)} \right) = -2 \ln(L(H_1)) - 2 \ln(L(H_2))$$

- d.f. is 2, because the residual d.f. for the full model (H_2) is 11, and for the model with temps dropped (H_1) is 13
- LR tests only valid for **nested** models (i.e. one model has a subset of the terms in the other)

Butterfly alleles – with and without temperatures included



Multiply log-likelihoods by -2, calculate the difference to obtain Chi-square test statistic

The models differ by 2 residual d.f.

The difference in fit is significant

Model	SSE	Resid. df	Ln(likelihood)	-2Ln(likelihood)	Chisq	df	p
Without temps	659.8	13	-29.75	59.51	9.93	2	0.007
With temps	354.8	11	-24.79	49.58			

Next steps

- Next we will learn an approach to evaluating competing hypotheses called the “Method of Support”
 - Based on likelihoods of models given data
 - Can be applied to non-nested models
 - Does not assume a null hypothesis
 - No p-values
 - Asks “which model (i.e. hypothesis) is best supported by the data?”, which is what we want to know
- Good stuff!