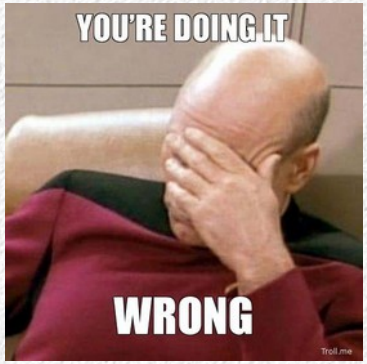
Likelihood

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

Cohen 1994



- Cohen's article lays out a strong case for alternatives to null hypothesis significance testing
- Points out that objections to the procedure have been growing
- Still an issue – in 2015 the journal *Basic and Applied Social Psychology* banned NHST's from papers it published
- What's wrong with them?

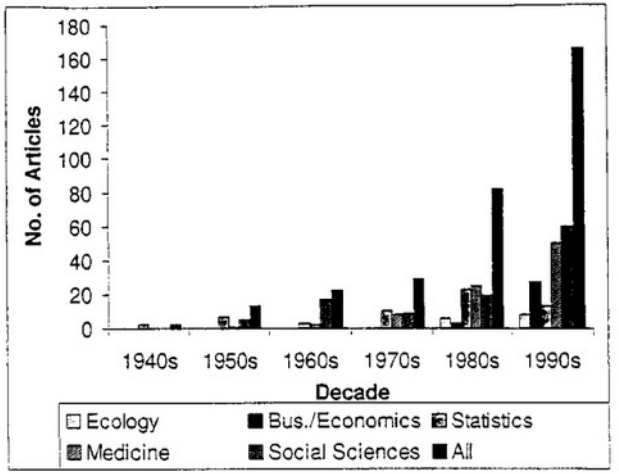


Fig. 1. Sample of articles, based on an extensive sampling of the literature by decade, in various disciplines that questioned the utility of null hypothesis testing in scientific research. Numbers shown for the 1990s were extrapolated based on sample results from volume years 1990–96.

Cohen's case: problems with NHST's

- They don't tell us what we want to know
- They are logically flimsy, and encourage faulty reasoning (logical fallacies)
  - Poor understanding of probability, what statistical test to use to see if the infection rate is different from 0 when 1/30 of the patients have the disease?
- They encourage false dichotomies
  - Rejecting H_o: doesn't support a specific H_a:
- They throw away useful information
  - Effect size, confidence intervals more informative
- If randomization of subjects to treatment groups isn't possible, the null hypothesis is ***never*** true – NHST's are tests of whether n is big enough to detect the difference
- Publication bias against non-significant results causes problems for science

Charge: NHST's don't tell us what we want to know

- We want to know: “Given the data we have collected, what is the probability that some scientific hypothesis is true?”
  - The scientific hypothesis is almost never the null, almost always an alternative hypothesis
  - We want to know $p(H_a|data)$
- What an NHST asks is: “Assuming the null hypothesis is true, what’s the probability of observing the data?”
  - That is, $p(data|H_o)$
- These are not the same
- Cohen: NHST “...does not tell us what we want to know, and we so desperately want to know what we want to know that, out of desperation, we nevertheless believe it does!”

Asking the questions we want to know

- Null hypothesis tests are criticized for asking the wrong question
- The right question is: “Which hypothesis is best supported by the data?”
- We can use likelihood-based model selection to address this
- We will start by learning what likelihood is

Likelihoods come from probability distributions

- Likelihood is calculated using a probability distribution, such as the normal →

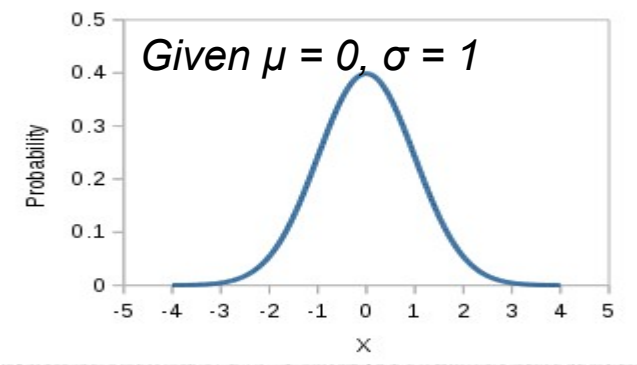
$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left[\frac{x_i-\mu}{\sigma}\right]^2}$$

- The population parameters for the mean and standard deviation are μ and σ
- To calculate a probability density from the normal distribution, $p(x_i \mid \mu, \sigma)$:
  - Specify known values for the parameters (μ, σ)
  - Calculate probability density of a data value (x_i) given the known parameter values
- To calculate a likelihood from the normal distribution, $L(\mu|x_i, \sigma)$:
  - Specify the known value of a data point (x_i), and any known parameter (σ)
  - 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**

Probability density and likelihood for single observations

Probability density

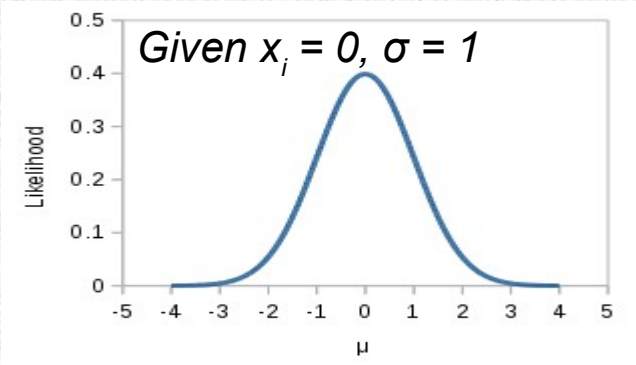
$$p(x_i \mid \mu, \sigma)$$



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

Likelihood

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



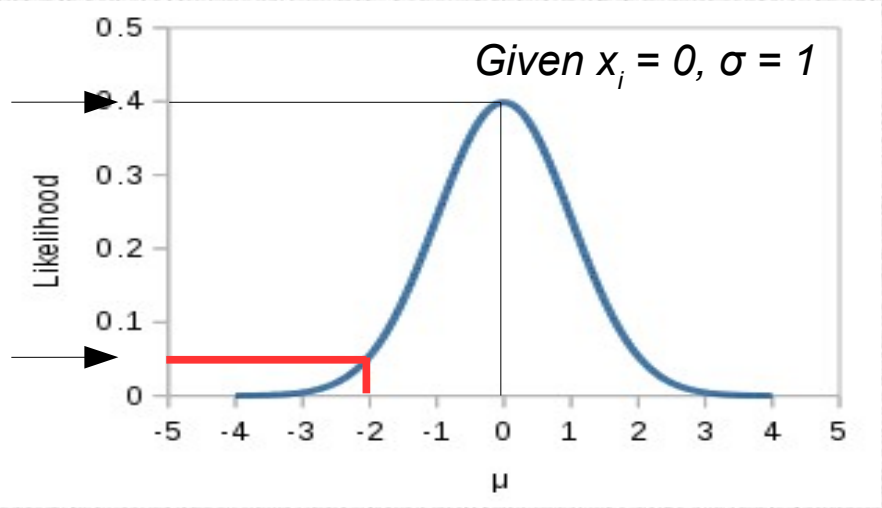
Centered on an observed value of x, calculate likelihoods of possible values of μ

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

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

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

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

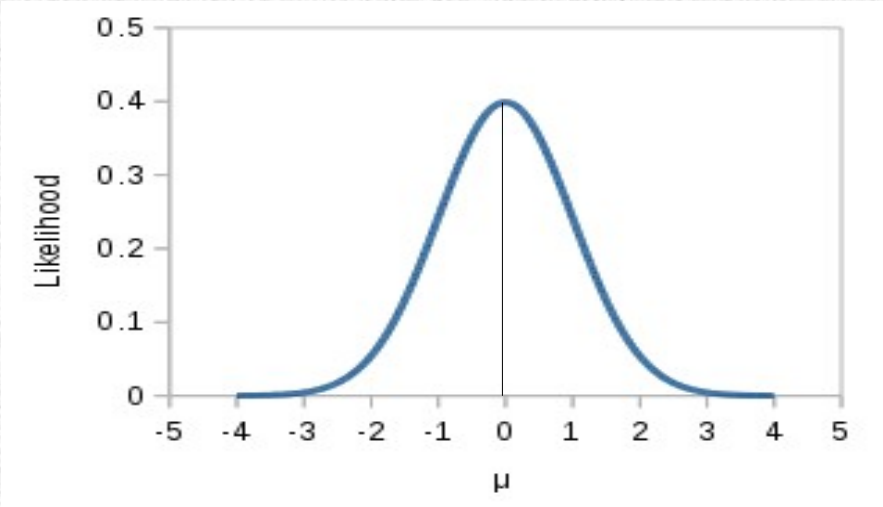


In this form, only difference between a probability density and a likelihood is in the interpretation

Principle of maximum likelihood

- The paremeter 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
- With a single data value, the likelihood function is at its maximum over the data value – the maximum likelihood estimate for μ is equal to the data value

What's the maximum likelihood estimate for μ?

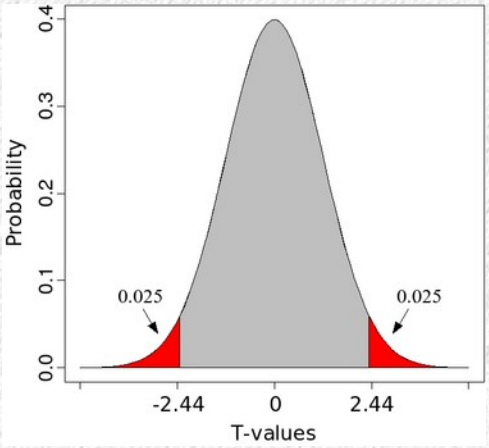


Single data value equal to 0

Probability and likelihood for a sample of data points

Probability

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



Use a sampling distribution, such as the t

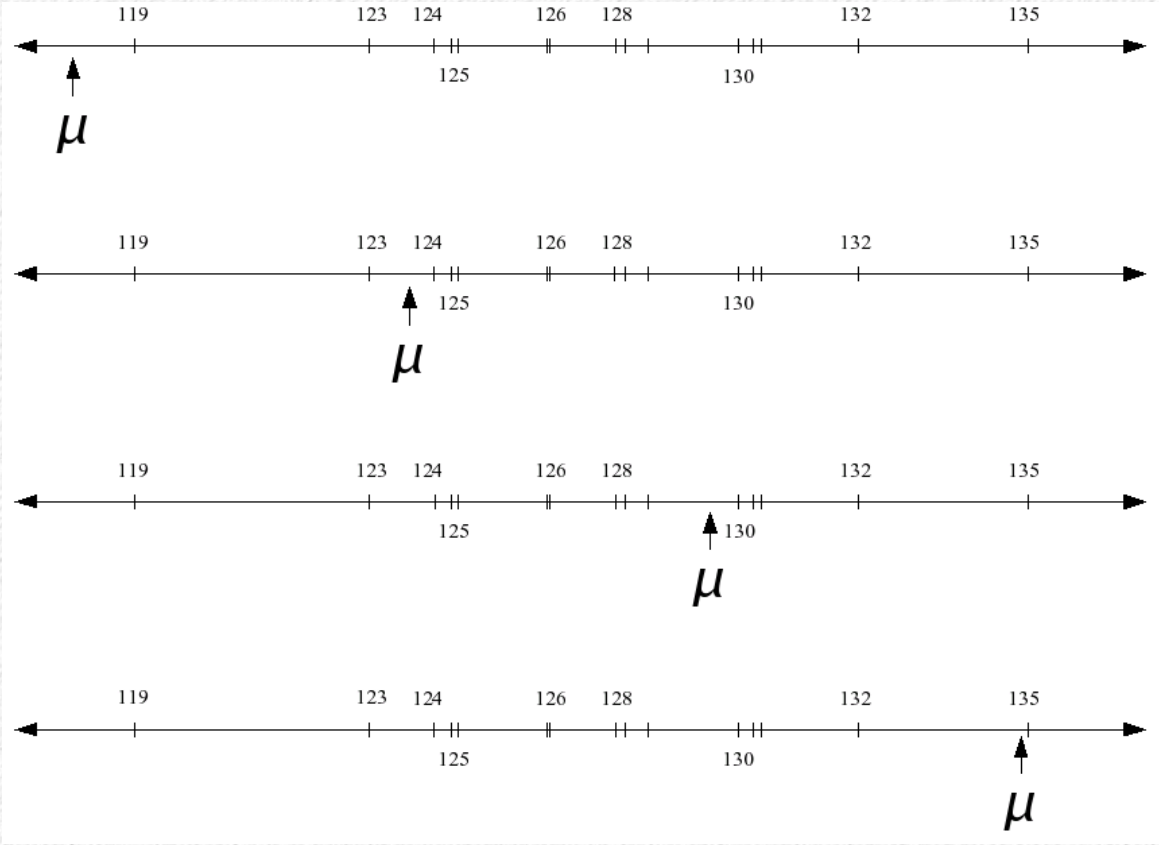
Likelihood

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

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

Using likelihoods for estimation with a sample of data

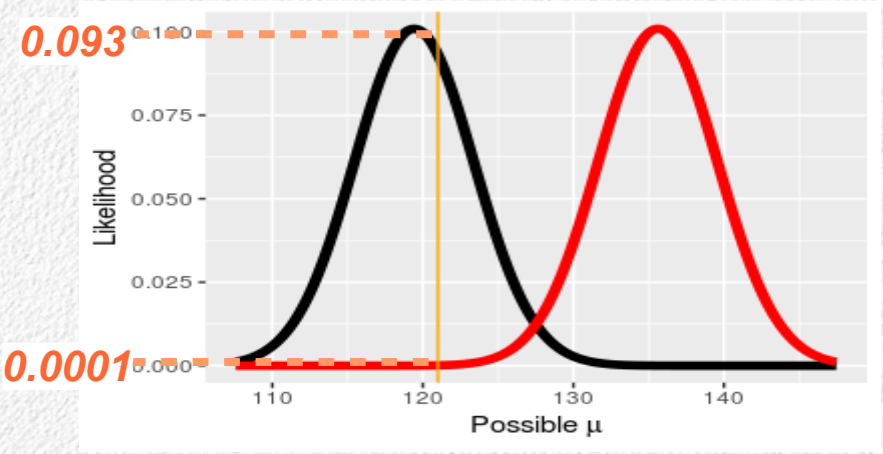
	The Data
	123.67
	126.90
	130.78
	125.30
	124.86
	126.96
	135.61
	119.42
	128.74
	132.53
	130.36
	128.31
	130.63
	128.13
	125.17



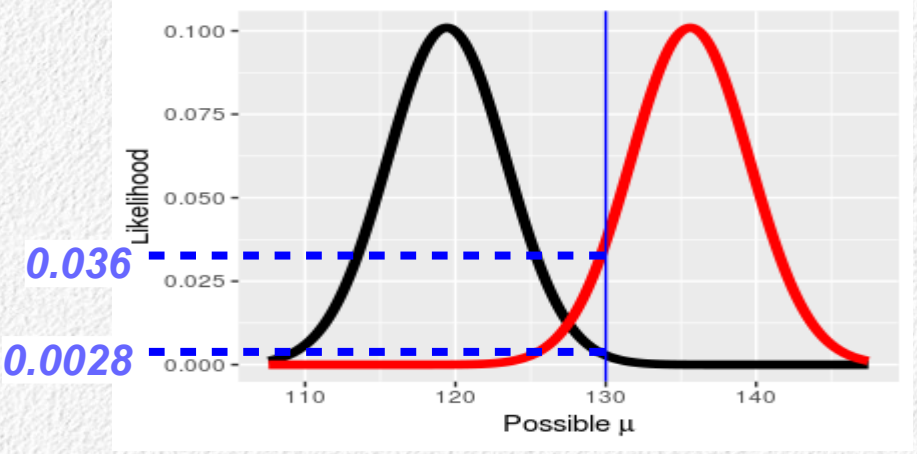
Infinite number of possible values of μ – which is best?

Two possible values of μ

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



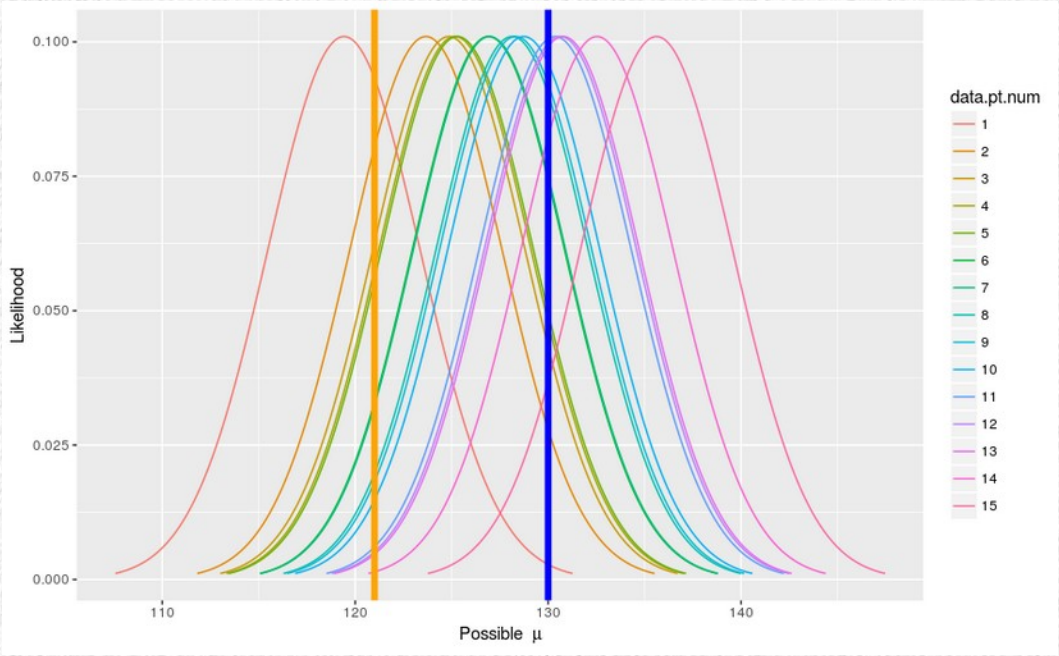
$$L(130 \mid x = 119.42, 135.61, \sigma = 3.95)$$



Likelihood of 121 is the product of likelihoods in orange: 0.093 x 0.0001 = 0.0000093

Likelihood of 130 is the product of likelihoods in blue: 0.0028 x 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			

$$L(121|data) = 3.4\times10^{-28}$$

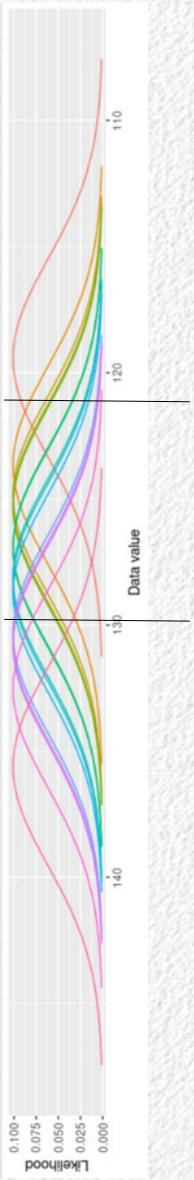
$$L(130|data) = 1.12\times10^{-19}$$

Which is bigger?

Individual likelihoods for two possible means

Product of these is the L(μ | data) for each

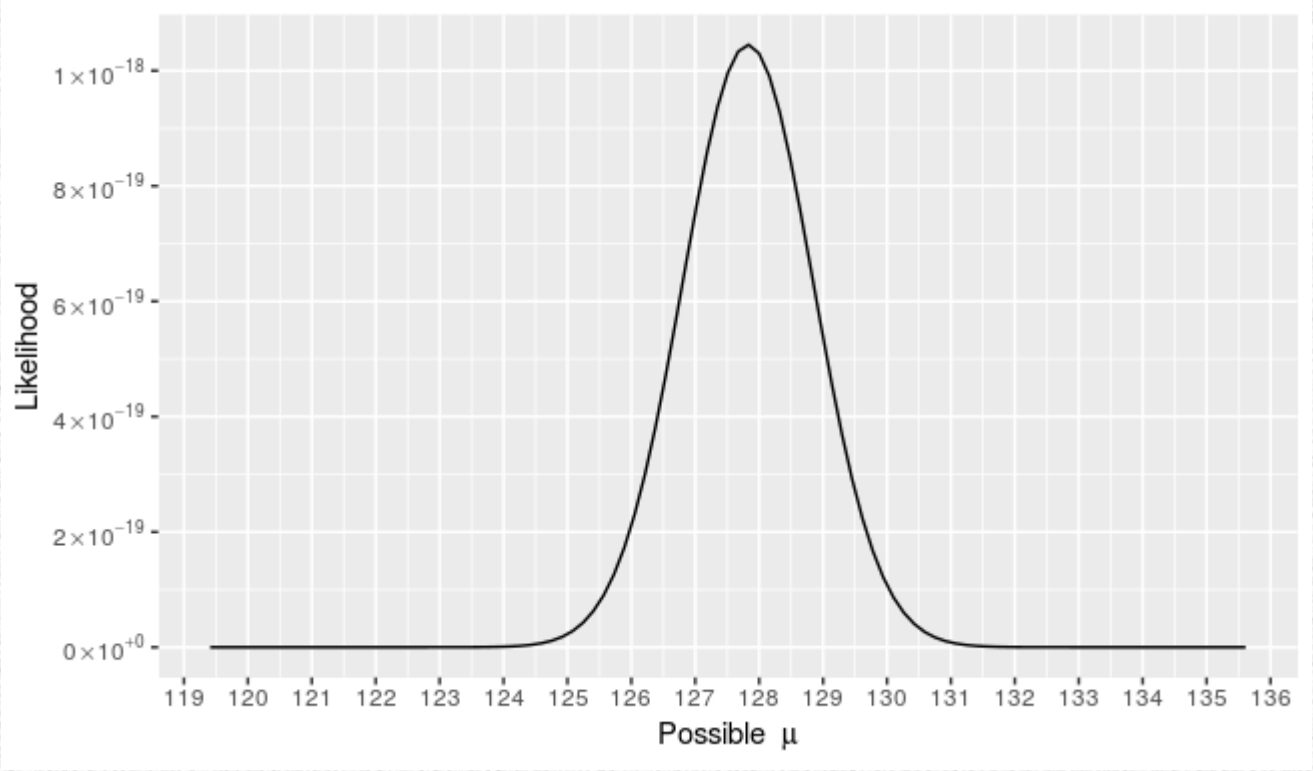
A range of possible μ and their likelihoods



	Data values															
	119.42	123.67	124.86	125.17	125.3	126.9	126.96	128.13	128.31	128.74	130.36	130.63	130.78	132.53	135.61	Likelihood
Value for μ																
119	0.1004	0.0502	0.0336	0.0298	0.0283	0.0137	0.0133	0.0070	0.0063	0.0048	0.0016	0.0013	0.0012	0.0003	0.0000	5.81E-35
120	0.0999	0.0656	0.0474	0.0429	0.0411	0.0220	0.0214	0.0121	0.0110	0.0087	0.0032	0.0027	0.0024	0.0007	0.0000	1.74E-31
121	0.0932	0.0804	0.0627	0.0579	0.0558	0.0331	0.0324	0.0198	0.0182	0.0148	0.0061	0.0052	0.0047	0.0014	0.0001	1.99E-28
122	0.0816	0.0924	0.0777	0.0732	0.0712	0.0468	0.0459	0.0303	0.0282	0.0236	0.0108	0.0093	0.0085	0.0029	0.0003	8.69E-26
123	0.0670	0.0996	0.0904	0.0869	0.0852	0.0620	0.0611	0.0435	0.0409	0.0351	0.0178	0.0156	0.0145	0.0055	0.0006	1.45E-23
124	0.0516	0.1006	0.0986	0.0967	0.0957	0.0771	0.0763	0.0585	0.0557	0.0492	0.0276	0.0247	0.0231	0.0098	0.0013	9.28E-22
125	0.0372	0.0954	0.1009	0.1009	0.1007	0.0900	0.0893	0.0738	0.0711	0.0645	0.0402	0.0366	0.0346	0.0164	0.0027	2.27E-20
126	0.0252	0.0849	0.0969	0.0988	0.0994	0.0984	0.0981	0.0873	0.0851	0.0794	0.0549	0.0508	0.0486	0.0258	0.0052	2.12E-19
127	0.0160	0.0708	0.0872	0.0907	0.0921	0.1010	0.1010	0.0969	0.0956	0.0917	0.0703	0.0662	0.0639	0.0379	0.0094	7.58E-19
128	0.0095	0.0554	0.0736	0.0781	0.0800	0.0972	0.0976	0.1009	0.1007	0.0992	0.0845	0.0809	0.0788	0.0523	0.0158	1.04E-18
129	0.0053	0.0406	0.0583	0.0631	0.0651	0.0877	0.0884	0.0986	0.0995	0.1008	0.0952	0.0928	0.0912	0.0677	0.0249	5.41E-19
130	0.0028	0.0280	0.0433	0.0478	0.0498	0.0742	0.0751	0.0903	0.0922	0.0960	0.1006	0.0997	0.0990	0.0823	0.0368	1.08E-19
131	0.0014	0.0181	0.0302	0.0340	0.0357	0.0589	0.0599	0.0776	0.0801	0.0857	0.0997	0.1006	0.1008	0.0937	0.0511	8.25E-21
132	0.0006	0.0109	0.0197	0.0227	0.0240	0.0439	0.0447	0.0625	0.0653	0.0718	0.0927	0.0951	0.0963	0.1001	0.0665	2.41E-22
133	0.0003	0.0062	0.0121	0.0142	0.0151	0.0307	0.0314	0.0472	0.0499	0.0565	0.0808	0.0844	0.0862	0.1003	0.0812	2.69E-24
134	0.0001	0.0033	0.0069	0.0083	0.0089	0.0201	0.0206	0.0335	0.0358	0.0416	0.0661	0.0702	0.0724	0.0942	0.0929	1.15E-26
135	0.0000	0.0017	0.0037	0.0046	0.0050	0.0123	0.0127	0.0223	0.0241	0.0288	0.0507	0.0548	0.0571	0.0831	0.0998	1.88E-29

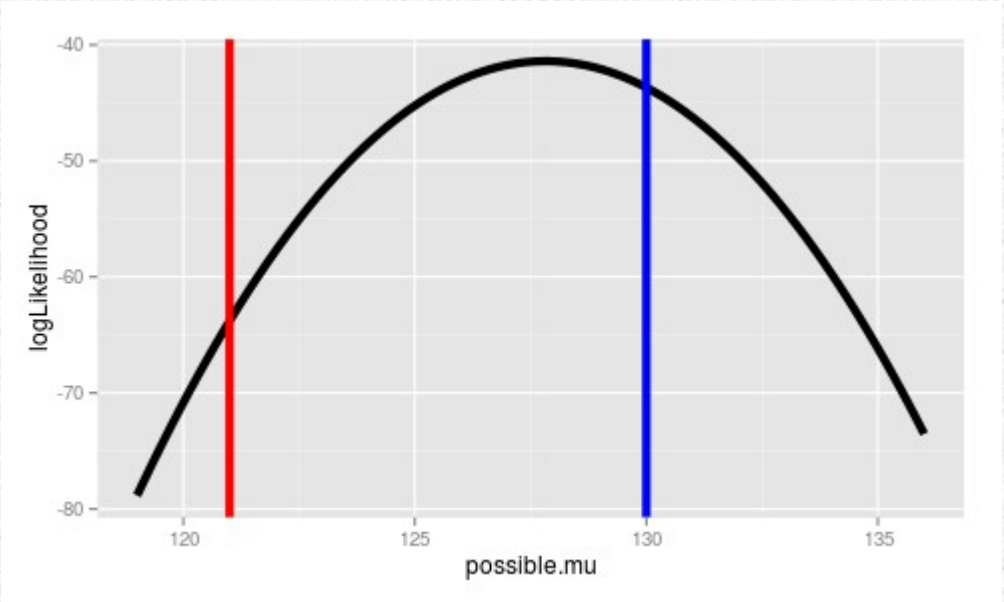
Calculate individual likelihoods, dataset likelihood for a range of possible values, pick the one with the highest likelihood as the maximum likelihood estimate

Likelihood function for data set



What's the maximum likelihood estimate for μ?

Ln(likelihood) changes the scale, makes likelihoods additive



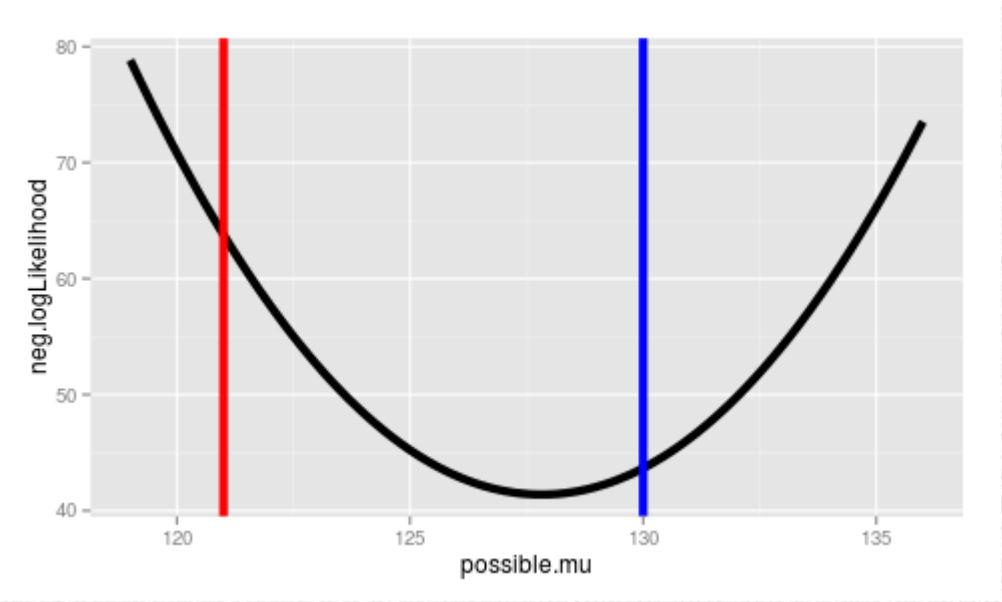
$$Likelihood = \prod \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left[\frac{x_i-\mu}{\sigma}\right]^2}$$

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

$$L(121|data) = 3.4\times10^{-28} \longrightarrow \ln(L(121|data)) = -63.79$$

$$L(130|data) = 1.12\times10^{-19} \longrightarrow \ln(L(130|data)) = -43.68$$

-Ln(likelihood) changes the direction



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

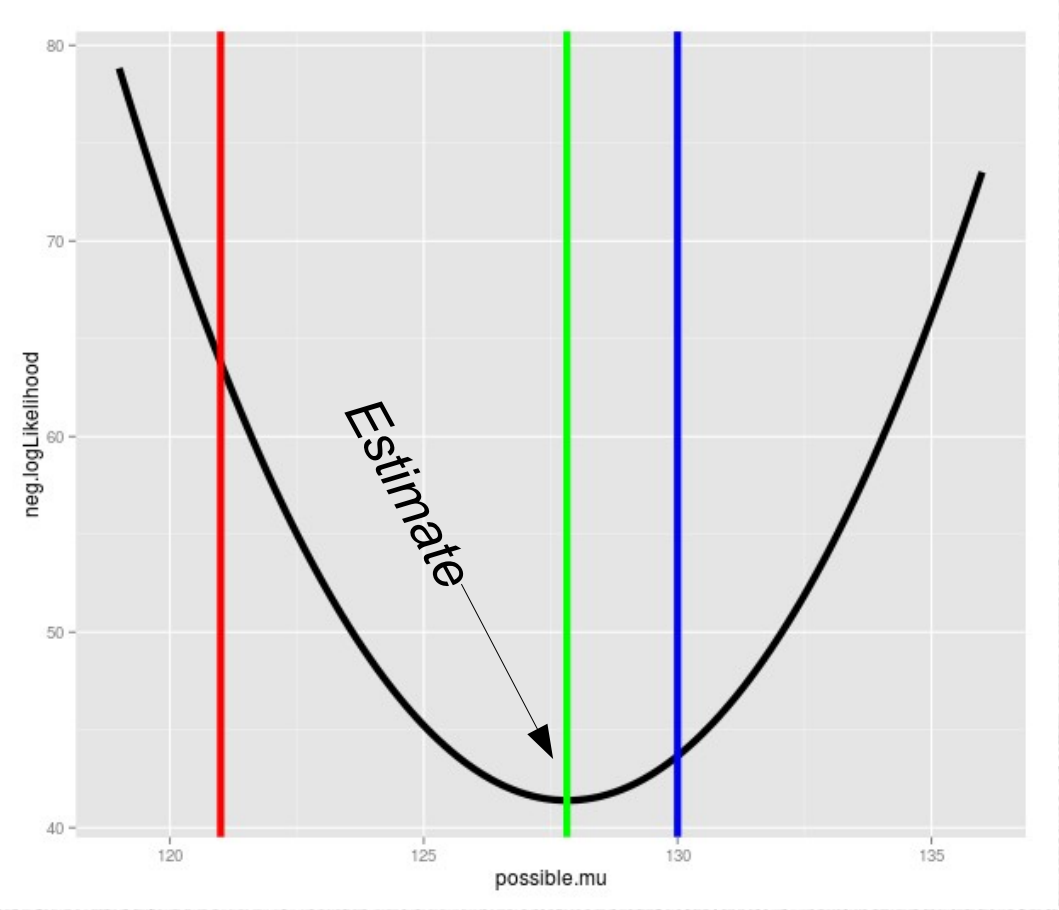
$$-\ln\left(L\left(130|data\right)\right)=43.68$$

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 -logLikelihood 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

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

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

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

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

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

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

ML estimator for μ

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 → 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 →

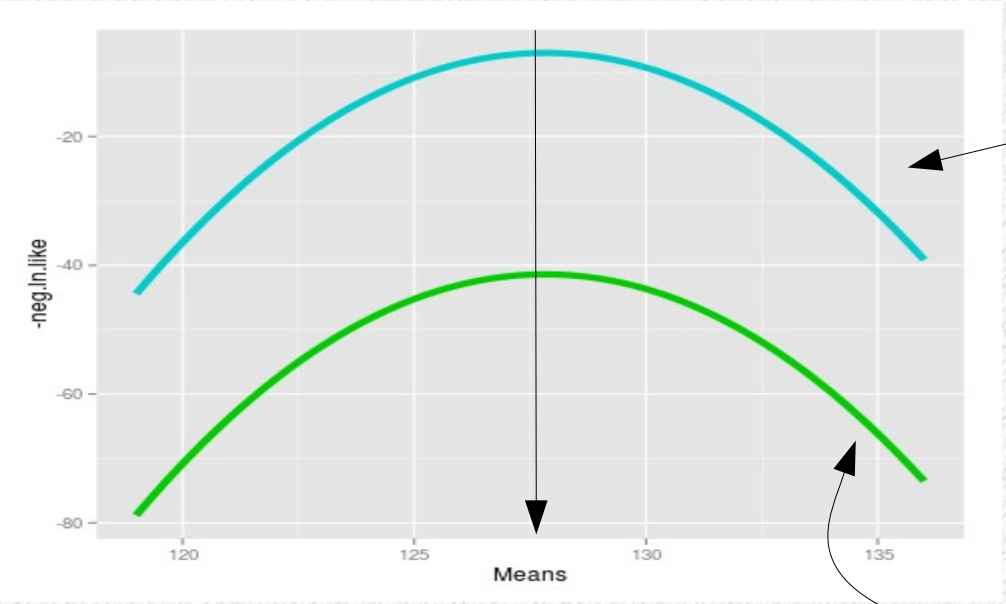
$$-0.5\,n\ln(2\pi)-0.5\,n\ln(\sigma^2)-\frac{1}{2\sigma^2}\sum(x_i-\mu)^2$$

and this →

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

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

- Remember, models are hypotheses
- We can calculate the likelihoods of models, which means we can calculate the likelihoods of hypotheses

  “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*

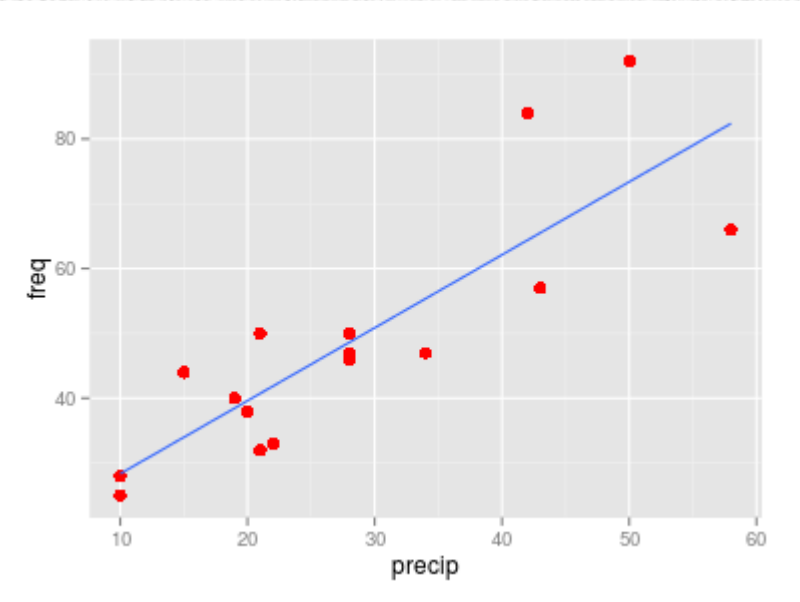
- Compare the support for different hypotheses by comparing likelihoods of different models

Example: temperature and gene frequencies in butterflies

- Gene frequencies are variable among populations, suspect this is due to climatic conditions
- Collect data on three variables
  - Maximum temp
  - Minimum temp
  - Precipitation
- From these variables we might hypothesize:
  - Precipitation affects gene frequencies
  - Since precipitation + temperature is the difference between rain and snow, the combination of precipitation and temperature affects gene frequencies
- Which of these is better supported by the data?

Likelihood of a model given the data

- Start with precip-only hypothesis
  - Linear model is that freq ~ precip
- Will use the normal likelihood function = assume the residuals are normally distributed
  - We can check this (model criticism)
- Can calculate likelihoods of residuals
  - Likelihood of each residual multiplied together is the likelihood of the model



Modeled as a linear relationship with:

Intercept = 17.0956

Slope = 1.1258

Model log-likelihood

- Log-likelihood of the model given each residual is calculated using normal likelihood function
  - n = 16 data values
  - $\hat{y}$ = predicted allele freq.
    = 17.0956 + 1.1258 (precip)
  - y_i are the data values
  - Residuals are $y_i - \hat{y}$
  - σ^2 = SSE/n = sum of squared residuals divided by sample size
- 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

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

Residuals

What do we do with a model log-likelihood?

- Likelihoods only tell us anything in comparison with other likelihoods
  - A log-likelihood of -59.088 doesn't mean anything in particular
- We can specify another hypothesis (i.e. another linear model) to explain variation in frequency, and compare its likelihood to the model with only precipitation as a predictor
- For example, we may want to know if temperature and precipitation combined are important – we need to:
  - Fit a model with precipitation only (done!)
  - Fit a model with precipitation + max temperature + min temperature
  - Calculate the likelihoods of the models and compare them
- This is easy to do in R

```
Response: freq
          Df Sum Sq Mean Sq F value    Pr(>F)
precip     1 3722.1  3722.1   34.48 4.061e-05 ***
Residuals 14 1511.3   107.9
```

First hypothesis (H1)

```
'log Lik.' -59.08808 (df=3)
```

Log-likelihoods for butterfly models

```
Response: freq
          Df Sum Sq Mean Sq F value    Pr(>F)
precip     1 3722.1  3722.1 125.900 1.018e-07 ***
max.temp   1 1118.6  1118.6  37.837 4.938e-05 ***
min.temp   1   37.9    37.9   1.282    0.2797
Residuals 12  354.8    29.6
```

Second hypothesis (H2)

```
'log Lik.' -47.49411 (df=5)
```

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

Which hypothesis is better supported by the data?

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

$$Support = \ln\left(\frac{L(H1)}{L(H2)}\right) = \ln\left(L(H1)\right) - \ln\left(L(H2)\right)$$

- The bigger the difference in the logs of the model likelihoods the bigger the difference in support
  - Positive difference = H1 better supported than H2

Likelihood ratio test comparing the two models

- We can use support ratios as the basis of a hypothesis test, called 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 the difference in residual df for the two models
  - Residual d.f. for the model with precip and both temps (H_2) is 14
  - Precip only (H_1) is 12
  - Difference is 2

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 support is significant

*LR tests only valid for **nested** models (i.e. one model has a subset of the predictors in the other)*

Response data has to be the same

- no missing values

- both un-transformed or both transformed, but no mixing

Model	log Likelihood	-2 log Likelihood	Chi-square	df	p
Precip	-59.09	118.18	23.2	2	9.17E-06
Precip + Min temp + Max temp	-47.49	94.98			

Some nice features of likelihoods

- Likelihoods can be combined
  - Data can be added as it becomes available, likelihoods updated
  - Can even 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 a single 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

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!