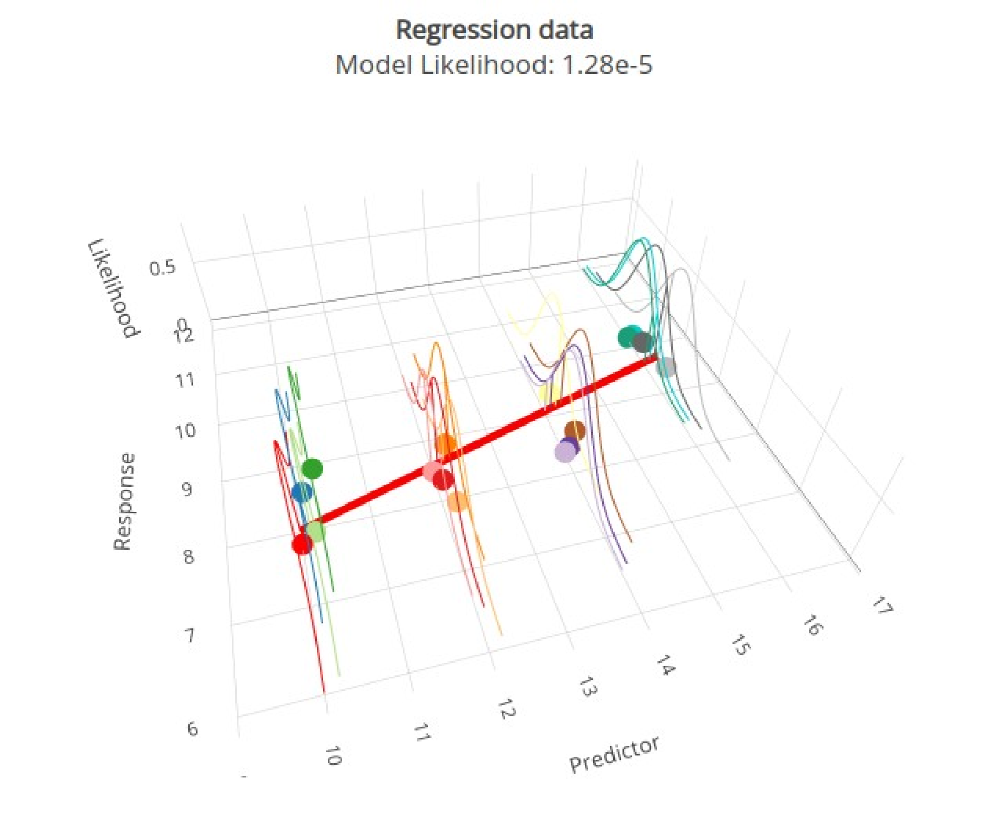
Curve fitting with likelihoods



Likelihood

- Likelihood is a general approach to statistics that can be used for:
  - Estimating parameters, building confidence intervals
  - Testing hypotheses
  - Comparing hypotheses against one another
- Likelihood appears to be similar to probability, but has a very different interpretation

Definition of likelihood

- Likelihood is a measure of support for a particular estimated value given a set of data
- We need to specify a particular statistical distribution of deviations from the estimated value (such as normal, binomial, etc.) to use the approach
- We then use the formula for the assumed distribution to calculate the likelihood

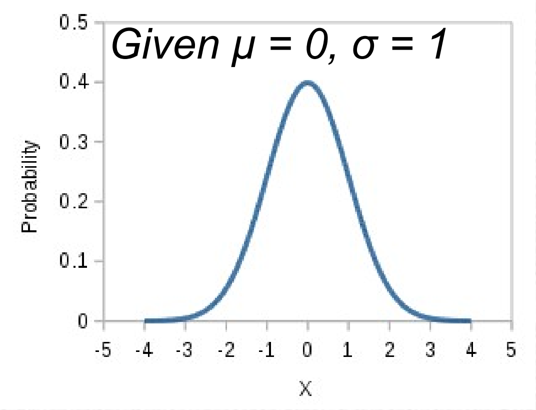
Likelihood and probability

- Probabilities treat parameters (μ, σ) as known, and calculate the chances of observing data values given the parameters → $p(x_i \mid \mu, \sigma)$
- Likelihoods invert this – they treat the data as known, and ask how likely a set of parameters is given the known data → $L(\mu, \sigma \mid x_i)$

Probability and likelihood – single observation

Probability

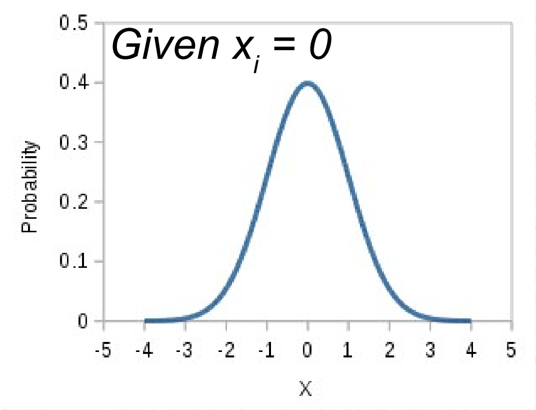
$$p(x_i \mid \mu, \sigma)$$



Use a probability distribution to represent a "random variable"

Likelihood

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



Likelihood of possible values of the mean given observed x

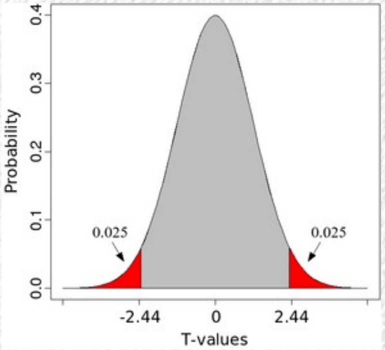
Some nice features of likelihoods

- Likelihoods can be combined
  - Data can be added as it becomes available, such as adding 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 obseration
- Parameter estimates and confidence intervals
  - “Maximum likelihood estimates”
  - Even when analytical formulas aren't available

Probability and likelihood – a sample of data points

Probability

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



Use a "sampling distribution", such as the t

Likelihood

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

Likelihood of a sample is the product of likelihoods of data points

Likelihood functions

- Derived from probability distributions
- Used to model differences between hypothetical values and observed data (residuals)
- Example: Normally distributed deviations

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

Likelihood of parameters given a single data point - the normal probability distribution

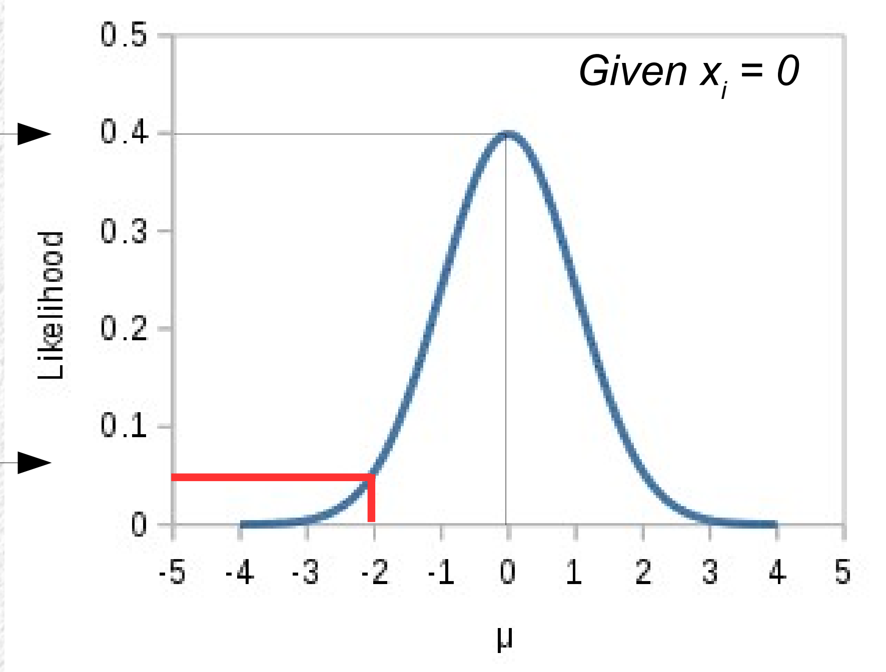
$$L(\mu \mid x_{i \ldots n}) = \prod \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left[\frac{x_i-\mu}{\sigma}\right]^2}$$

Likelihood of parameters given all the data – the product of all the likelihoods given each single data point

Example of the normal likelihood function

Likelihood of
μ = 0 is 0.4

Likelihood of
μ = -2 is 0.05



Have a data value, x_i, equal to 0

The highest likelihood for a possible value for the mean given this one data point is the value of the data point itself

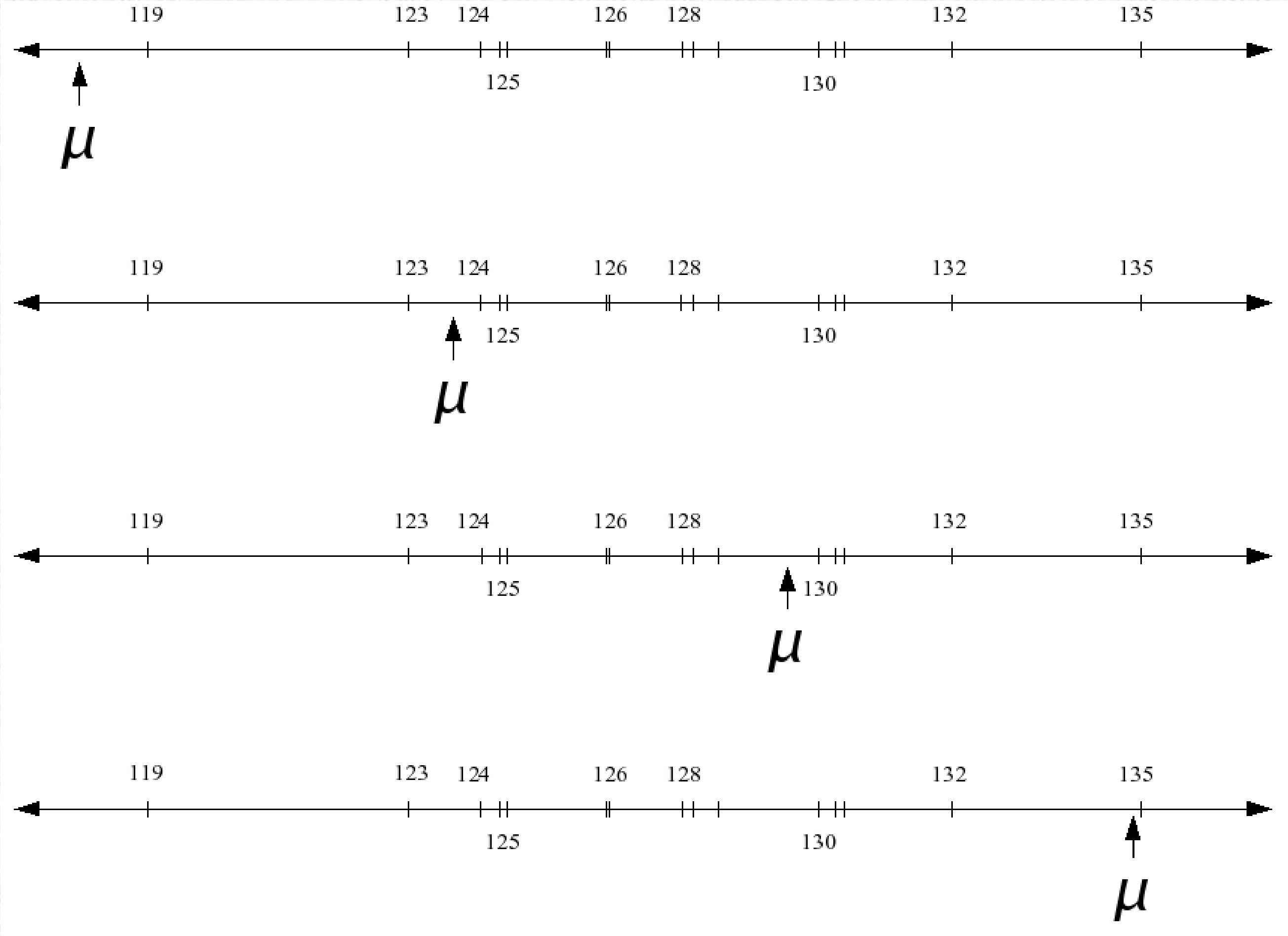
Other values for μ are possible, but have lower likelihood given the one data value we have

Using likelihood for estimation

- Given some data, what is the best estimate for the mean of a population, μ?
- We use $\bar{x} = \frac{\sum x_i}{n}$, is that the best estimator?
- Maximum likelihood criterion: the parameter value with the highest likelihood given the data is the best estimate

Some 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?

Pick a likelihood function

- Need to specify a likelihood function to model deviations of estimates from data points
- We'll use the normal distribution
- We then find the value for μ with the highest likelihood across all the data
- Let’s try this out...

Log-likelihoods

- The log of the normal likelihood function is:

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

- Multiplicative terms are additive now
- Across multiple data values the likelihood function is:

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

- No effect on which value has the highest likelihood

Calculations – comparing likelihoods among possible means

	Data	Possible means	Likelihood	LogLikelihood
	119.42	119	5.81E-35	-78.8
	123.67	120	1.74E-31	-70.8
	124.86	121	1.99E-28	-63.8
	125.17	122	8.69E-26	-57.7
	125.3	123	1.45E-23	-52.6
	126.9	124	9.28E-22	-48.4
	126.96	125	2.27E-20	-45.2
	128.13	126	2.12E-19	-43.0
	128.31	127	7.58E-19	-41.7
	128.74	128	1.04E-18	-41.4
	130.36	129	5.41E-19	-42.1
	130.63	130	1.08E-19	-43.7
	130.78	131	8.25E-21	-46.2
	132.53	132	2.41E-22	-49.8
	135.61	133	2.69E-24	-54.3
		134	1.15E-26	-59.7
Mean	127.82	135	1.88E-29	-66.1
Std. Dev	3.95	136	1.17E-32	-73.5

Numerical solution – try different possible means, calculate logLikelihood for each

Pick the one with the lowest logLikelihood

Not an analytical solution! Only approximately correct (but often good enuf)

Minimum -LogLikelihood at the maximum likelihood

	Data		Possible means	Likelihood	LogLikelihood	-LogLikelihood
	119.42		119	5.81E-35	-78.8	78.8
	123.67		120	1.74E-31	-70.8	70.8
	124.86		121	1.99E-28	-63.8	63.8
	125.17		122	8.69E-26	-57.7	57.7
	125.3		123	1.45E-23	-52.6	52.6
	126.9		124	9.28E-22	-48.4	48.4
	126.96		125	2.27E-20	-45.2	45.2
	128.13		126	2.12E-19	-43.0	43.0
	128.31		127	7.58E-19	-41.7	41.7
	128.74		128	1.04E-18	-41.4	41.4
	130.36		129	5.41E-19	-42.1	42.1
	130.63		130	1.08E-19	-43.7	43.7
	130.78		131	8.25E-21	-46.2	46.2
	132.53		132	2.41E-22	-49.8	49.8
	135.61		133	2.69E-24	-54.3	54.3
			134	1.15E-26	-59.7	59.7
Mean	127.82		135	1.88E-29	-66.1	66.1
Std. Dev	3.95		136	1.17E-32	-73.5	73.5

Back to the app...

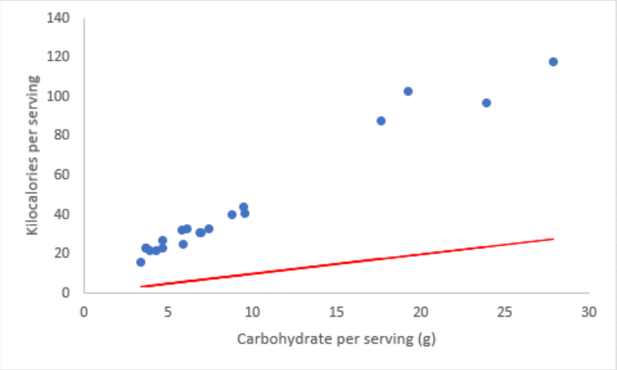
Curve fitting with maximum likelihood

- The predicted value from a curve is the average of y expected for a given value of x
- We can calculate the likelihood of parameter values given the residuals around the line that they produce
- Example: caloric content of foods as a function of carbohydrate content

Relationship between carbohydrates and kcal of foods

E21 =C21-D21

	A	B	C	D	E
1	food	Carbohydrate	Kilocalories	Predicted	Residuals
2	Cilantro	3.67	23	3.67	19.33
3	Oyster Mushrooms	6.09	33	6.09	26.91
4	Tomatillos	5.84	32	5.84	26.16
5	Yam	27.88	118	27.88	90.12
6	Carrots	9.58	41	9.58	31.42
7	Mustard Spinach	3.9	22	3.9	18.1
8	Pokeberry Shoots	3.7	23	3.7	19.3
9	Spaghetti Squash	6.91	31	6.91	24.09
10	Eggplant	5.88	25	5.88	19.12
11	Mustard Greens	4.67	27	4.67	22.33
12	Brown Mushrooms	4.3	22	4.3	17.7
13	Radishes	3.4	16	3.4	12.6
14	Maitake Mushrooms	6.97	31	6.97	24.03
15	Fava Beans	17.63	88	17.63	70.37
16	Chicory Greens	4.7	23	4.7	18.3
17	Hot Red Chili Pepper	8.81	40	8.81	31.19
18	Blackeye Pea Pods	9.5	44	9.5	34.5
19	Okra	7.45	33	7.45	25.55
20	Waterchestnuts	23.94	97	23.94	73.06
21	Fireweed Leaves	19.22	103	19.22	83.78
22					
23	Slope	1			
24	Intercept	0			
25					



Another app...

Finding the maximum likelihood with Solver

	A	B	C	D	E	F
1	food	Carbohydrate	Kilocalories	Predicted	Residuals	-logLikelihood
2	Cilantro	3.67	23	14.68	8.32	10.26489
3	Oyster Mushrooms	6.09	33	24.36	8.64	10.94329
4	Tomatillos	5.84	32	23.36	8.64	10.94329
5	Yam	27.88	118	111.52	6.48	6.860886
6	Carrots	9.58	41	38.32	2.68	2.509886
7	Mustard Spinach	3.9	22	15.6	6.4	6.732086
8	Pokeberry Shoots	3.7	23	14.8	8.2	10.01709
9	Spaghetti Squash	6.91	31	27.64	3.36	3.023286
10	Eggplant	5.88	25	23.52	1.48	1.885886
11	Mustard Greens	4.67	27	18.68	8.32	10.26489
12	Brown Mushrooms	4.3	22	17.2	4.8	4.492086
13	Radishes	3.4	16	13.6	2.4	2.332086
14	Maitake Mushrooms	6.97	31	27.88	3.12	2.828886
15	Fava Beans	17.63	88	70.52	17.48	39.80589
16	Chicory Greens	4.7	23	18.8	4.2	3.817086
17	Hot Red Chili Pepper	8.81	40	35.24	4.76	4.444286
18	Blackeye Pea Pods	9.5	44	38	6	6.112086
19	Okra	7.45	33	29.8	3.2	2.892086
20	Waterchestnuts	23.94	97	95.76	1.24	1.804286
21	Fireweed Leaves	19.22	103	76.88	26.12	86.89389
22						
23	Slope	4			Sum:	228.8681
24	Intercept	0				
25	StdDev	2				

Solver Parameters

Set Objective: $F$23

To: Max / Min (selected) / Value Of: 0

By Changing Variable Cells: $B$23:$B$25

Subject to the Constraints:

Add | Change | Delete | Reset All | Load/Save

Make Unconstrained Variables Non-Negative

Select a Solving Method: GRG Nonlinear | Options

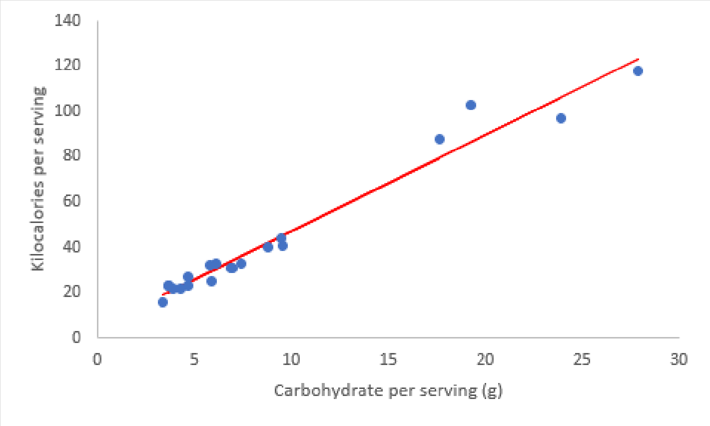
Solving Method

Select the GRG Nonlinear engine for Solver Problems that are smooth nonlinear. Select the LP Simplex engine for linear Solver Problems, and select the Evolutionary engine for Solver problems that are non-smooth.

Help | Solve | Close

Solver's solution

	A	B	C	D	E	F
1	food	Carbohydrate	Kilocalories	Predicted	Residuals	-logLikelihood
2	Cilantro	3.67	23	20.05309	2.946906	2.756245
3	Oyster Mushrooms	6.09	33	30.35379	2.646215	2.727571
4	Tomatillos	5.84	32	29.28966	2.710336	2.733427
5	Yam	27.88	118	123.1026	-5.10257	3.052087
6	Carrots	9.58	41	45.20891	-4.20891	2.910214
7	Mustard Spinach	3.9	22	21.03209	0.967914	2.624156
8	Pokeberry Shoots	3.7	23	20.18079	2.819211	2.743691
9	Spaghetti Squash	6.91	31	33.8441	-2.8441	2.746095
10	Eggplant	5.88	25	29.45992	-4.45992	2.947313
11	Mustard Greens	4.67	27	24.30958	2.690422	2.731593
12	Brown Mushrooms	4.3	22	22.73468	-0.73468	2.617385
13	Radishes	3.4	16	18.90384	-2.90384	2.751949
14	Maitake Mushrooms	6.97	31	34.09949	-3.09949	2.771974
15	Fava Beans	17.63	88	79.47361	8.52639	3.847669
16	Chicory Greens	4.7	23	24.43727	-1.43727	2.643403
17	Hot Red Chili Pepper	8.81	40	41.93142	-1.93142	2.671784
18	Blackeye Pea Pods	9.5	44	44.8684	-0.8684	2.62104
19	Okra	7.45	33	36.1426	-3.1426	2.776563
20	Waterchestnuts	23.94	97	106.332	-9.33202	4.092965
21	Fireweed Leaves	19.22	103	86.24142	16.75858	7.39653
22						
23	Slope	4.256483913			Sum:	62.16365
24	Intercept	4.431798256				
25	StdDev	5.415385965				
26						



How do they compare to analytical values?

	Solver	Analytical
Slope	4.256484	4.256491
Intercept	4.431798	4.431768
StdDev	5.415386	5.556068

Slope and intercept – very close
Standard deviation of residuals under-estimated

Today…

- We will use maximum likelihood to find the best-fit line for the photosynthesis data we used previously