Introduction to Statistical Inference

Floyd Bullard

SAMSI/CRSC Undergraduate Workshop at NCSU

23 May 2006

Introduction

Example 1

Example 2

Example 3

Example 4

Conclusion

Parametric models

Statistical inference means drawing conclusions based on data. There are a many contexts in which inference is desirable, and there are many approaches to performing inference.

Introduction

Example 1

Example 2

Example 3

Example 4

Conclusion

Parametric models

Statistical inference means drawing conclusions based on data. There are a many contexts in which inference is desirable, and there are many approaches to performing inference.

One important inferential context is *parametric models*. For example, if you have noisy (x, y) data that you think follow the pattern $y = \beta_0 + \beta_1 x + \textit{error}$, then you might want to estimate β_0, β_1, and the magnitude of the error.

Parametric models

Statistical inference means drawing conclusions based on data. There are a many contexts in which inference is desirable, and there are many approaches to performing inference.

One important inferential context is *parametric models*. For example, if you have noisy (x, y) data that you think follow the pattern $y = \beta_0 + \beta_1 x + \textit{error}$, then you might want to estimate β_0, β_1, and the magnitude of the error.

Throughout this week, we'll be examining parametric models. (More complex than this simple linear model of course.)

Likelihood ratios

There are numerous tools available for parameter estimation, and you'll be introduced to two or three of them this week. The one we'll look at this afternoon may be the most straightforward and easiest to understand: *likelihood ratios*.

Example 1

Suppose a large bag contains a million marbles, some fraction of which are red. Let's call the fraction of red marbles π. π is a constant, but its value is unknown to us. We want to estimate the value of π.

Introduction

Example 1

Example 2

Example 3

Example 4

Conclusion

Example 1 (continued)

Obviously we'd be just guessing if we didn't collect any data, so let's suppose we draw 3 marbles out at random and find that the first is white, the second is red, and the third is white.

Example 1 (continued)

Obviously we'd be just guessing if we didn't collect any data, so let's suppose we draw 3 marbles out at random and find that the first is white, the second is red, and the third is white.

Question: What would be the probability of that particular sequence, *WRW*, if π were equal to, say, 0.2?

Introduction

Example 1

Example 2

Example 3

Example 4

Conclusion

Example 1 (continued)

If $\pi = 0.2$, then the probability of drawing out the sequence *WRW* would be $0.8 \times 0.2 \times 0.8 = 0.128$.

Example 1 (continued)

If $\pi = 0.2$, then the probability of drawing out the sequence *WRW* would be $0.8 \times 0.2 \times 0.8 = 0.128$.

Question: What would be the probability of that particular sequence, *WRW*, if $\pi = 0.7$?

Example 1 (continued)

If $\pi = 0.7$, then the probability of drawing out the sequence *WRW* would be $0.3 \times 0.7 \times 0.3 = 0.063$.

Notice that $\pi = 0.7$ is less likely to have produced the observed sequence *WRW* that is $\pi = 0.2$.

Example 1 (continued)

If $\pi = 0.7$, then the probability of drawing out the sequence *WRW* would be $0.3 \times 0.7 \times 0.3 = 0.063$.

Notice that $\pi = 0.7$ is less likely to have produced the observed sequence *WRW* that is $\pi = 0.2$.

Question: Of all possible values of $\pi \in [0, 1]$, which one would have had the *greatest* probability of producing the sequence *WRW*?

Example 1 (continued)

Your gut feeling may be that $\pi = \frac{1}{3}$ is the candidate value of π that would have had the greatest probability of producing the sequence we observed, *WRW*. But can that be proven?

Example 1 (continued)

The probability of observing the sequence *WRW* for some unknown value of π is given by the equation

$$L(\pi) = (1-\pi)(\pi)(1-\pi) = \pi \cdot (1-\pi)^2.$$

Differentiating gives:

$$\begin{aligned} \frac{d}{d\pi} L(\pi) &= \pi \cdot 2(1-\pi)(-1) + (1-\pi)^2 \cdot 1 \\ &= 3\pi^2 - 4\pi + 1 \\ &= (3\pi - 1)(\pi - 1) \end{aligned}$$

Example 1 (continued)

The function $L(\pi)$ is called the *likelihood function*, and the value of π that maximizes $L(\pi)$ is called the *maximum likelihood estimate*, or *MLE*. In this case we did indeed have an MLE of $\frac{1}{3}$.

Introduction

Example 1

Example 2

Example 3

Example 4

Conclusion

Example 1 (continued)

The MLE may be the "best guess" for π, at least based on the maximum likelihood criterion, but surely there are other values of π that are also plausible. How should be find them?

Example 1 (continued)

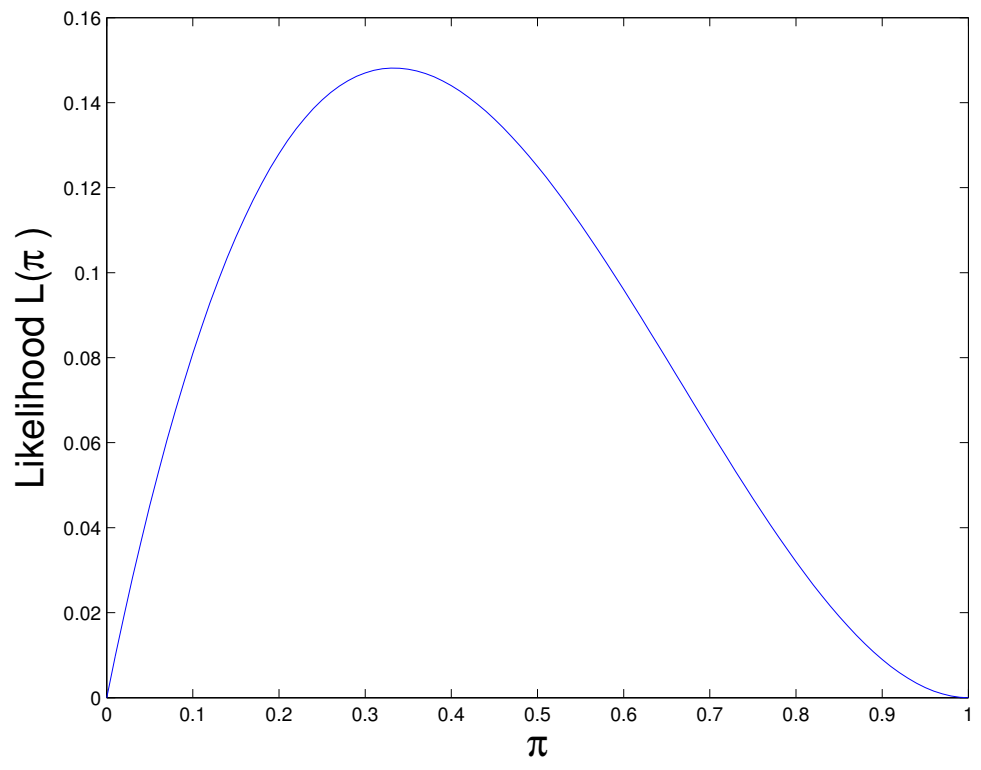

Figure: The likelihood function $L(\pi)$ plotted against π. What values of π are plausible, given the observation *WRW*?

Example 1 (continued)

Here is the MATLAB code that generated the graph on the previous slide:

```
p = [0:0.01:1];
L = p.*((1-p).^2);

plot(p,L)
xlabel('\pi')
ylabel('Likelihood L(\pi )')
```

Example 2

Okay. Now suppose that you again have a bag with a million marbles, and again you want to estimate the proportion of reds, π. This time you drew 50 marbles out at random and observed 28 reds and 22 whites.

Come up with the MLE for π and also use `MATLAB` to give a range of other plausible values for π.

Introduction

Example 1

Example 2

Example 3

Example 4

Conclusion

Example 2 (continued)

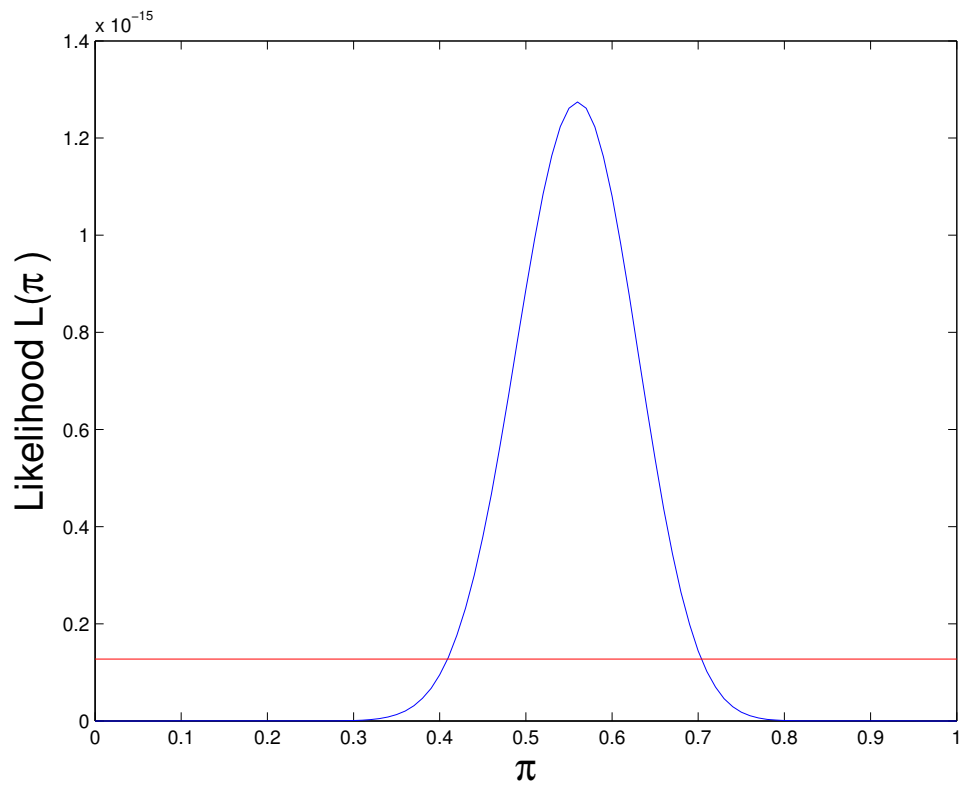

Figure: The red line is at 0.1 of the MLE's likelihood. Plausible values of π (by this criterion) are between 0.41 and 0.70.

A warning

The likelihood function L is *not* a probability density function, and it does *not* integrate to 1!

A comment

Notice that in the second example the scale of the likelihood function was much smaller than in the first example. (Why?)

Some of you likely know some combinatorics, and perhaps were inclined to include a binomial coefficient in the likelihood function:

$$L(\pi) = \binom{50}{28} \pi^{28}(1-\pi)^{22},$$

instead of

$$L(\pi) = \pi^{28}(1-\pi)^{22}.$$

Why might that matter? How does it change our inferential conclusions?

Example 3

Suppose now we have data that we will model as having come from a normal distribution with an unknown mean μ and an unknown standard deviation σ. For example, these five heights (in inches) of randomly selected MLB players:

Player	Position	Team	Height	Age
Pedro Lopez	2B	White Sox	73"	22
Boof Bonser	P	Twins	76"	24
Ken Ray	P	Braves	74"	31
Xavier Nady	RF	Mets	74"	27
Jeremy Guthrie	P	Indians	73"	27

Example 3 (continued)

Recall that the normal distribution is a continuous probability density function (pdf), so the probability of observing any number *exactly* is, technically, 0. But these players' heights are clearly rounded to the nearest inch. So the probability of observing a height of 73 inches when the actual height is rounded to the nearest inch is equal to the area under the normal curve over that span of heights that would round to 73 inches.

Example 3 (continued)

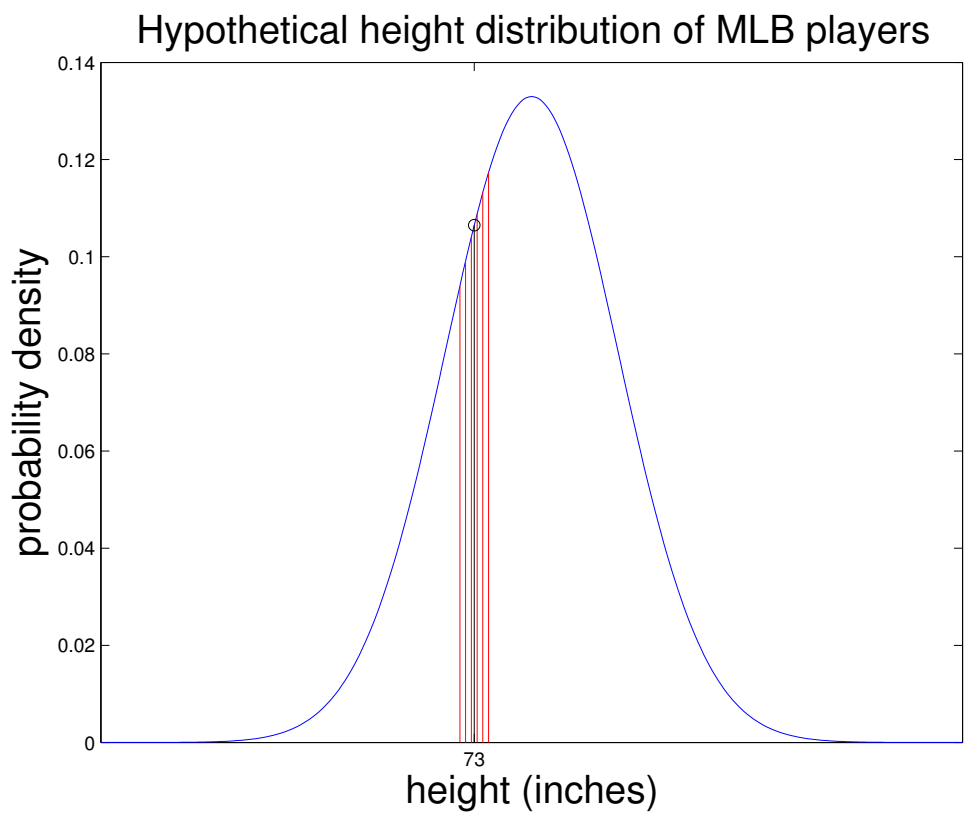

Figure: The probability that a player's height will be within a half inch of 73 inches is (roughly) *proportional to* the pdf at 73 inches.

Example 3 (continued)

So if $f(h)$ is a probability density funtion with mean μ and standard deviation σ, then the probability of observing the heights h_1, h_2, h_3, h_4, and h_5 is (approximately) *proportional to* $f(h_1) \cdot f(h_2) \cdot f(h_3) \cdot f(h_4) \cdot f(h_5)$.

Let's not forget what we're trying to do: estimate μ and σ! The likelihood function L is a function of both μ and σ, and it is proportional to the product of the five normal densities:

$$L(\mu, \sigma) \propto f(h_1) \cdot f(h_2) \cdot f(h_3) \cdot f(h_4) \cdot f(h_5),$$

where f is the normal probability density function with parameters μ and σ.

Example 3 (continued)

Happily, the normal probability density function is a built-in function in MATLAB:

```
normpdf(X, mu, sigma)
```

X can be a vector of values, and MATLAB will compute the normal pdf at each of them, returning a vector.

As such, we may compute the likelihood function at a particular μ and σ in MATLAB like this:

```
data = [73, 76, 74, 74, 73];
L(mu, sigma) = prod(normpdf(data, mu, sigma));
```

Example 3 (continued)

```
data = [73, 76, 74, 74, 73];

mu = [70:0.1:80];
sigma = [0:0.1:5];

L = zeros(length(mu), length(sigma));

for i = 1:length(mu)
  for j = 1:length(sigma)
    L(i,j) = prod(normpdf(data, mu(i), sigma(j)));
  end
end

surf(sigma, mu, L')
xlabel('sigma')
ylabel('mu')
```

Example 3 (continued)

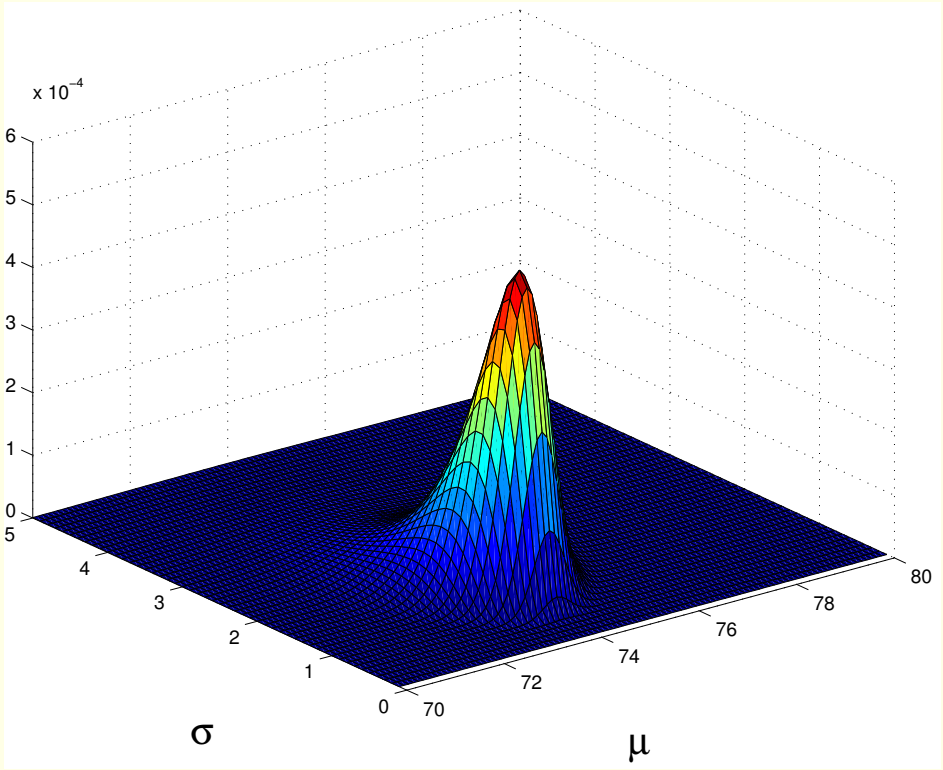

Figure: The likelihood function shows what values of the parameters μ and σ are most consistent with the observed data values.

Example 3 (continued)

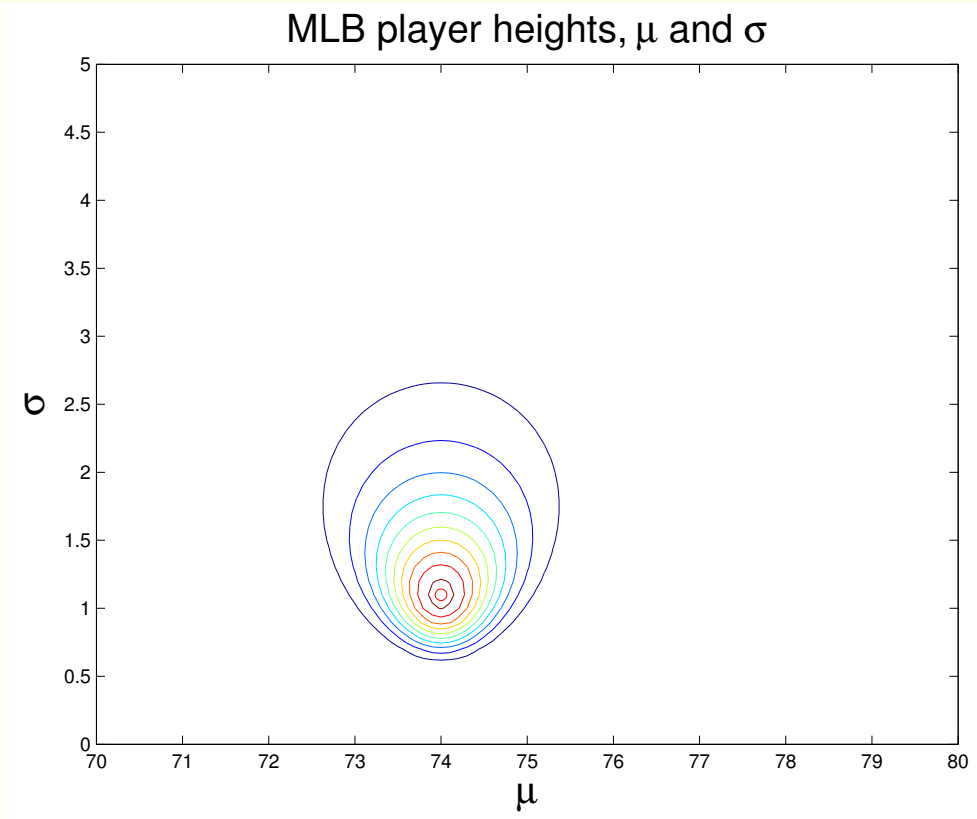

Figure: `contour(sigma, mu, L')`

Example 3 (continued)

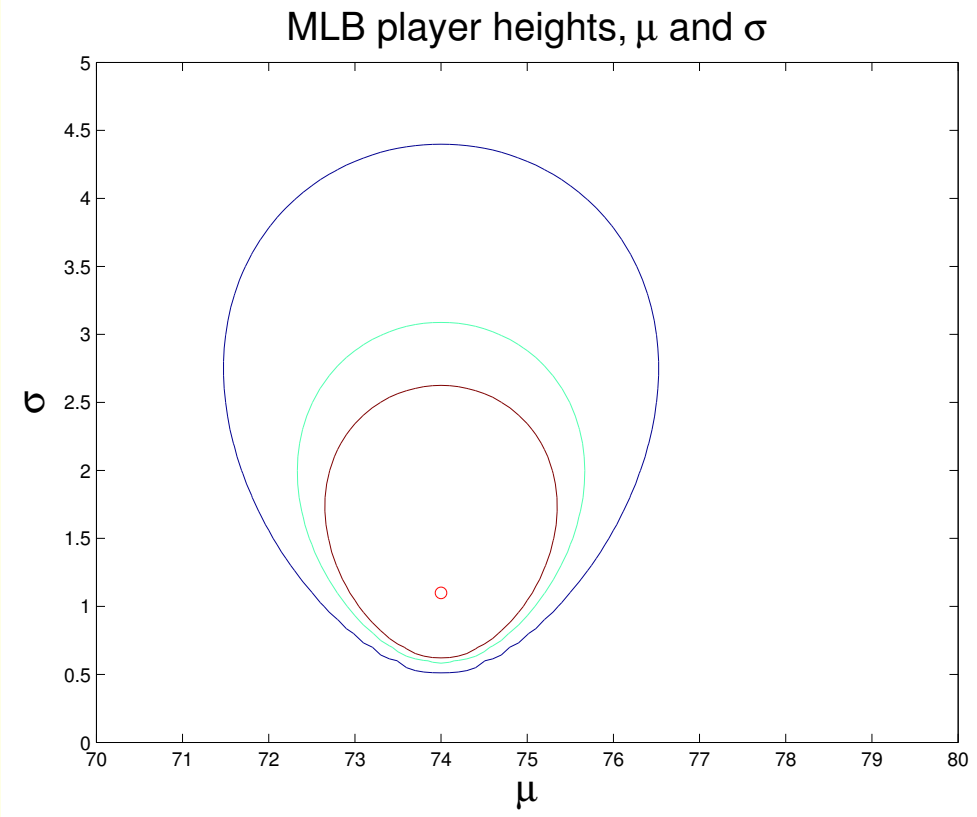

Figure: Level contours at 10%, 5%, and 1% of the maximum likelihood

Example 4 (a cautionary tale!)

Here are five randomly sampled MLB players' annual salaries:

Player	Position	Team	2006 Salary
Jeff Fassero	P	Giants	\$750 K
Brad Penny	P	Dodgers	\$5250 K
Chipper Jones	3B	Braves	\$12333 K
Jose Valverde	P	Diamondbacks	\$359 K
Alfredo Amezaga	SS	Marlins	\$340 K

Let's use the same technique we used with MLB players' heights to estimate the mean and standard deviation of players' salaries.

Example 4 (continued)

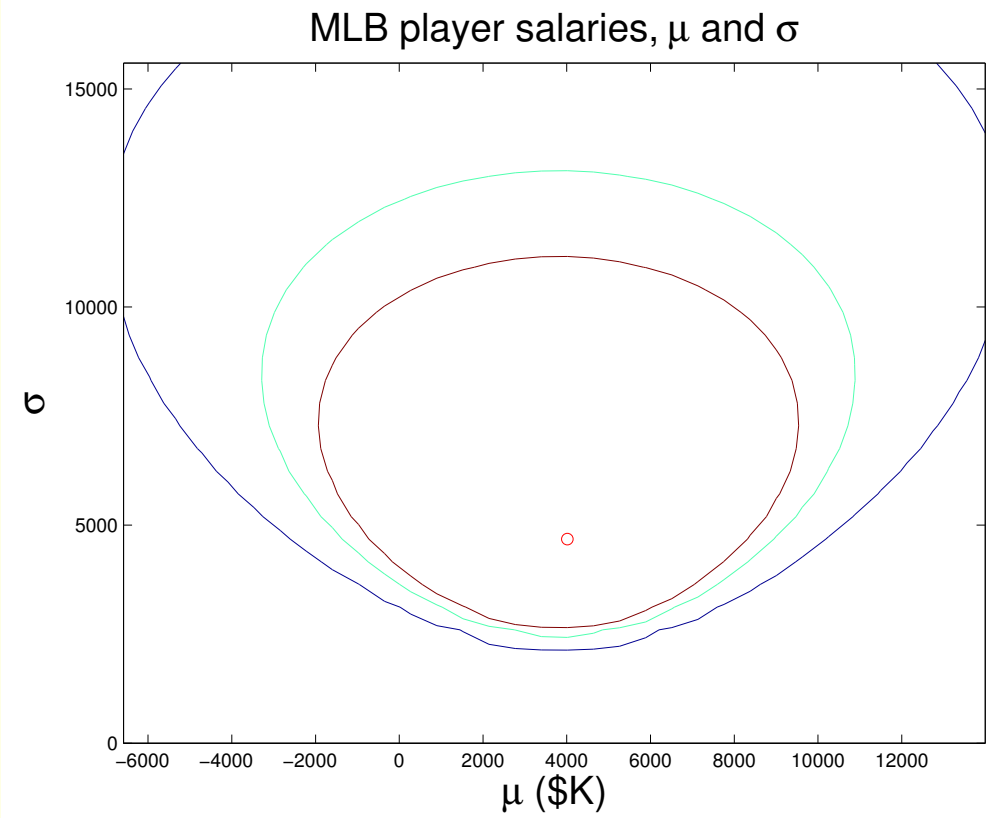

Figure: Level contours at 10%, 5%, and 1% of the maximum likelihood. What's wrong with this picture?

Example 4 (continued)

Moral: If the model isn't any good, then the inference won't be either.

Conclusion

Statistical inference means drawing conclusions based on data. One context for inference is the *parametric model*, in which data are supposed to come from a certain distribution family, the members of which are distinguished by differing parameter values. The normal distribution family is one example.

One tool of statistical inference is the *likelihood ratio*, in which a parameter value is considered "consistent with the data" if the ratio of its likelihood to the maximum likelihood is at least some threshold value, such as 10% or 1%. While more sophisticated inferential tools exist, this one may be the most staightforward and obvious.

Conclusion

Enjoy the week here at NCSU!

Feel free to ask any of us questions at any time!

Introduction

Example 1

Example 2

Example 3

Example 4

Conclusion