DATA252/DATA551: Modeling and Simulation

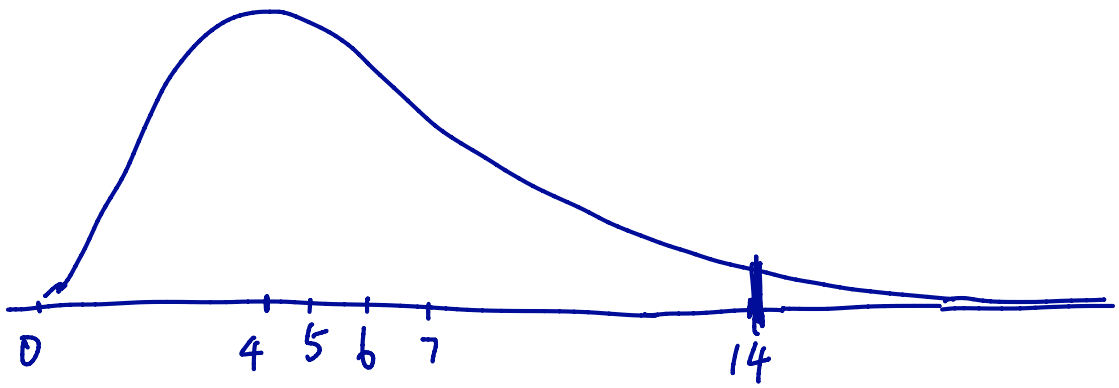
Lecture 6: Intro to model-based inference: Estimating the incubation period of COVID-19[1]

March 16, 2020

[1] Materials based on *The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application*, Lauer and Grantz, et. al., published at Annals.org on March 10, 2020. Data and code available at https://github.com/HopkinsIDD/ncov_incubation.

Set-up

Our goal is to estimate the incubation period of COVID-19, or, more precisely, the **distribution** of the incubation period.



- Why is this important?
- What properties of this distribution are especially of interest?

location : mean, median

spread : SD, IQR, Range ? → max incubation period ? 99% percentile ?

skewness → right skewed ?

Data

On Moodle, **nCoV_simple.csv** contains the incubation periods of 50 patients. Download this file, read in the data to R and convert it to a numeric vector of length 50.

```
ncov_simple=read.csv(file.choose())
ncov_simple=ncov_simple$days
ncov_simple
```

```
##  [1]  9.06  9.50  4.00  0.31  2.50  5.00 11.00  4.00  1.
## [12]  2.79  2.79  5.50  4.00  5.50  0.50  4.00  3.50  0.
## [23]  3.00  3.50  4.50  3.50  4.00  8.00  4.15  1.50  2.
## [34]  2.85  2.50  6.65  3.00  4.00  2.50  2.50  2.00  6.
## [45]  5.50  6.00  2.99  5.00  5.50  6.50
```

Statistical inference

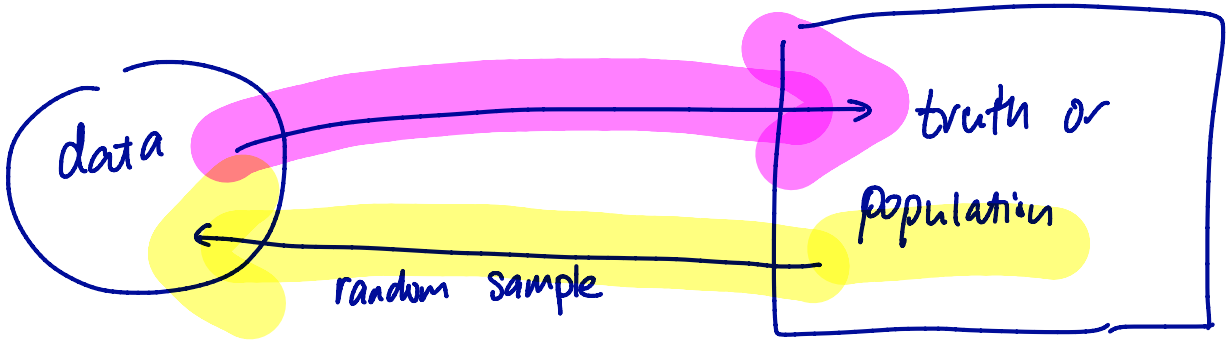
Print some **summary statistics** of this dataset:

```
summary(ncov_simple)
```

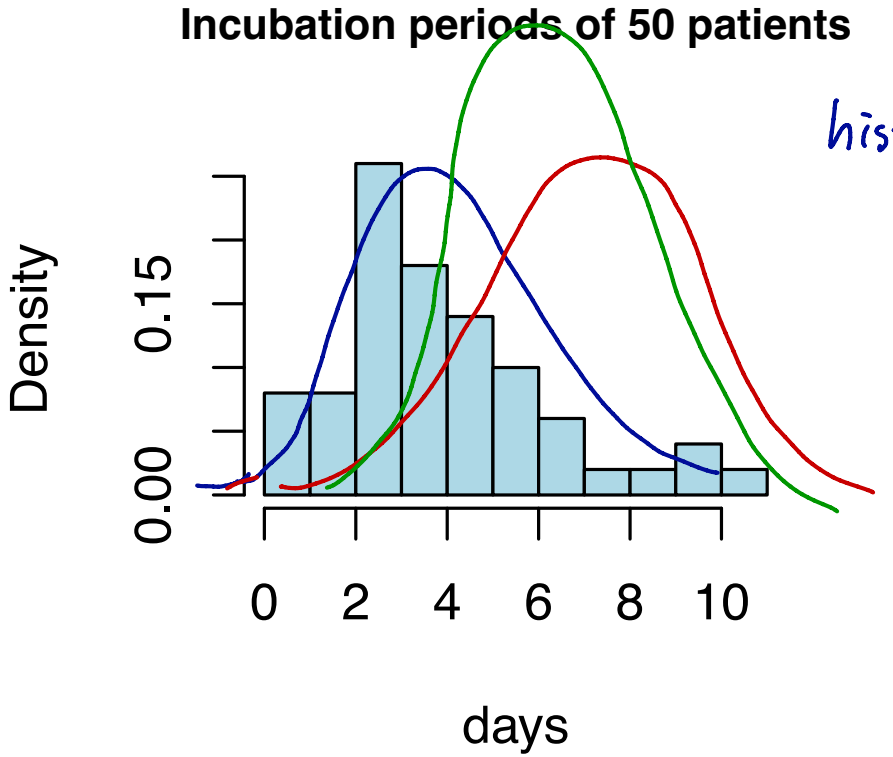
based on 50 patients

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.310   2.562   4.000   4.108   5.375  11.000
```

- Does this mean the true median of incubation period is 4 days? No.
  Does this mean the true maximum of the incubation period is 11 days? No.
- What is statistical inference?



Inference on the true distribution

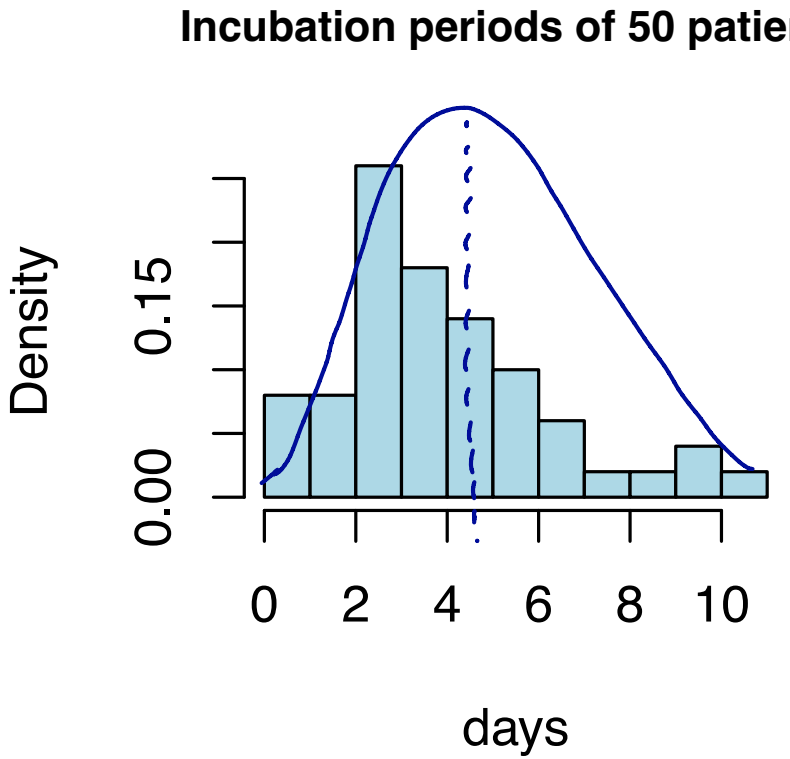


hist(ncov_simple)

- Step 1: Choose a model (say, normal)
- Step 2: Estimate parameters of the model[2] (For normal: mean, SD)

[2] Sometimes people estimate key parameters (like mean, median, SD) without assuming a model. For instance, we can use the sample mean to estimate the population mean.

Parameter Estimation



location

What is your "best estimates" of the population mean and standard deviation?

Best estimate for mean: 4.11 → sample mean
mean(ncov_simple)

SD: 2.34

Sample SD
sd(ncov_simple)

Parameter Estimation

Use R to perform parameter estimation

```
library('MASS') #need to install first, if not installed
fitdistr(ncov_simple,'normal')
```

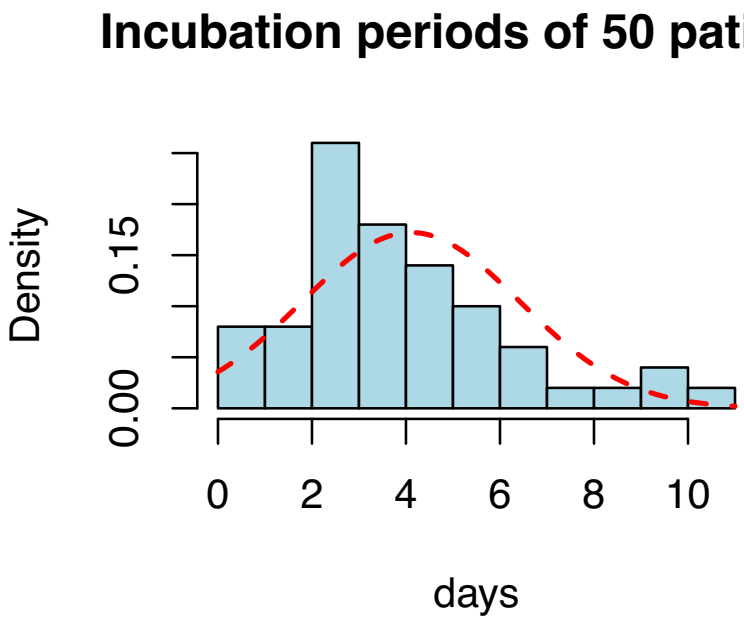
```
##       mean          sd
##  4.1076000   2.3166377
##  (0.3276220) (0.2316638)
```

Sample mean

$\propto 2.34$

Add estimated density to histogram

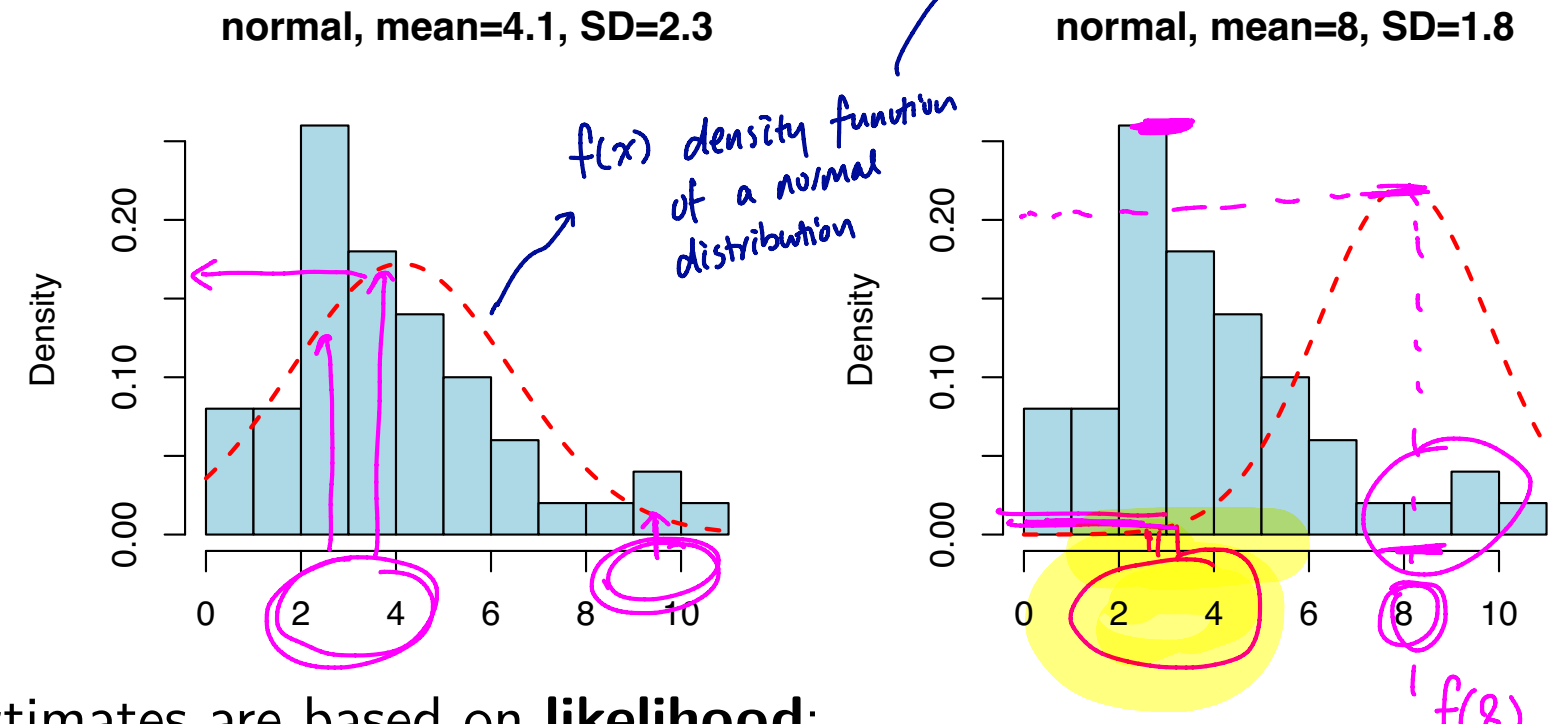
```
hist(ncov_simple,main='Incubation periods of 50 patients',
     col='lightblue',freq=F,xlab='days',nclass=10)
curve(dnorm(x,mean=4.1076,sd=2.3166),
     add=T,col='red',lwd=2,lty='dashed')
```



Maximum likelihood estimators (MLE)

$$f(x) = \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{2.3} e^{-\frac{1}{2 \cdot 2.3^2}(x-4.1)^2}$$

How are the parameters estimated?



Estimates are based on **likelihood**:

Likelihood (given a specific model) $\equiv f(x_1) \cdot f(x_2) \cdots f(x_{50})$

x_1: incubation period of 1st patient

x_{50}: incubation period of 50th patient

Q: which graph will produce a **larger** likelihood?

Left → Larger likelihood → fit data better

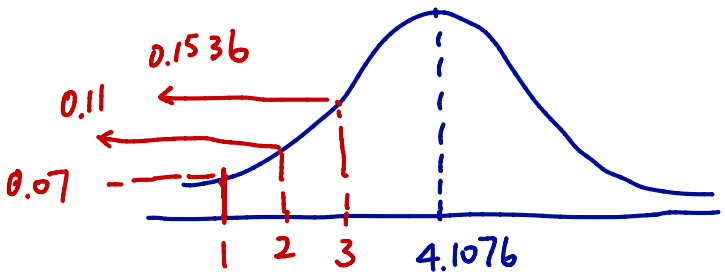
Calculate (log) likelihood in R

the bigger, the better

Recall, `dnorm` evaluates the density function of a normal distribution

```
test=c(1,2,3)
dnorm(test,mean=4.1076,sd=2.3166)
```

```
## [1] 0.07003351 0.11384808 0.15361038
```



(Log) Likelihood of the normal model:

estimates produced by R

```
log(prod(dnorm(ncov_simple,mean=4.1076,sd=2.3166)))
```

```
## [1] -112.9528
```

Try changing the parameters to any other number: does the (log) likelihood increase or decrease? decrease.

→ The estimates by R are called maximum likelihood estimates (MLE)

Some other models

prod(c(1,2,1,5)) ≡ 1*2*1*5

These distributions have the correct sample space of $[0, \infty)$ and allow more flexibility in shape (doesn't have to be symmetric)

- Lognormal (in R: `dlnorm`) → example
- Gamma (in R: `dgamma`) → exercise
- Weibull (in R: `dweibull`) → homework

Recall:

- Step 1: Choose a model
- Step 2: Estimate the parameters

Lognormal

- Estimate the parameters of the lognormal distribution

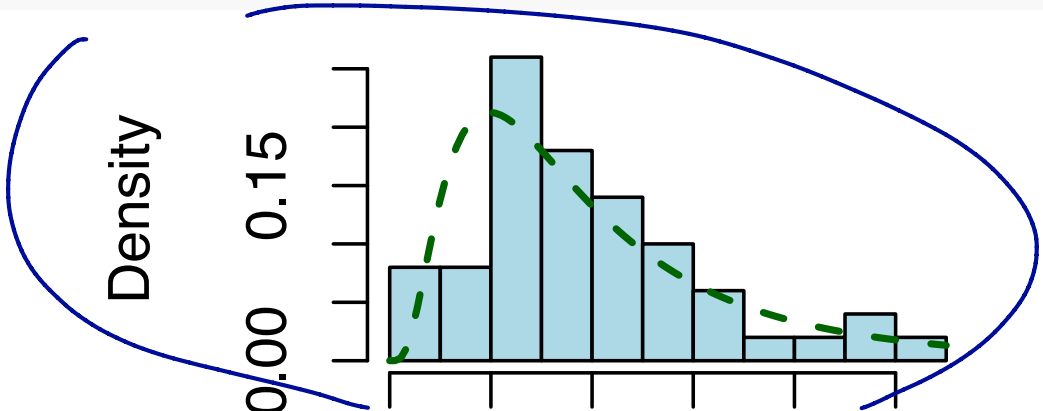
```
fitdistr(ncov_simple,'lognormal',lower=0.01)
```

```
##      meanlog        sdlog
##    1.20925432    0.73077736
##   (0.10334753) (0.07307774)
```

based on maximizing the likelihood

- Add the estimated density to histogram

```
hist(ncov_simple,main=NA,
     col='lightblue',freq=F,xlab='days',nclass=10)
curve(dlnorm(x,meanlog=1.21,sdlog=0.73),add=T,
      col='darkgreen',lwd=2,lty='dashed')
```



Lognormal

- Calculate the (log) likelihood

```
log(prod(dlnorm(ncov_simple,meanlog=1.21,sdlog=0.73)))
```

```
## [1] -115.7274
```

- *Based on likehihood*, which model better fits data? Lognormal or normal?

Log normal : -115.7274

normal : -112.95 → fits better

- In addition, using this model and estimated parameters, we can do many things...

lognormal (meanlog = 1.21, sdlog = 0.73)

  - What is the probability that the incubation period is shorter than 14 days? plnorm (14, meanlog = 1.21, sdlog = 0.73) 0.975
  - What is the probability that the incubation period is longer than 5 days? 1 - plnorm (5, meanlog = 1.21, sdlog = 0.73) 0.29
  - Simulate incubation periods of future patients:

rlnorm (1000, meanlog = 1.21, sdlog = 0.73)

Exercise: gamma

curve(dgamma(x, shape = 2.61, rate = 0.64), add = T)

shape = 2.6105974
rate = 0.6355530

1. Estimate the parameters of the gamma distribution.
2. Add the estimated density to the histogram.
3. Calculate the (log) likelihood of the gamma model. Based on likelihood, does the gamma model fit data better or worse than the normal and lognormal models?
4. Using the gamma model and estimated parameters, what is the probability that the incubation period is shorter than 14 days?

3. −110.5826 best fit (i.e, largest log likelihood)

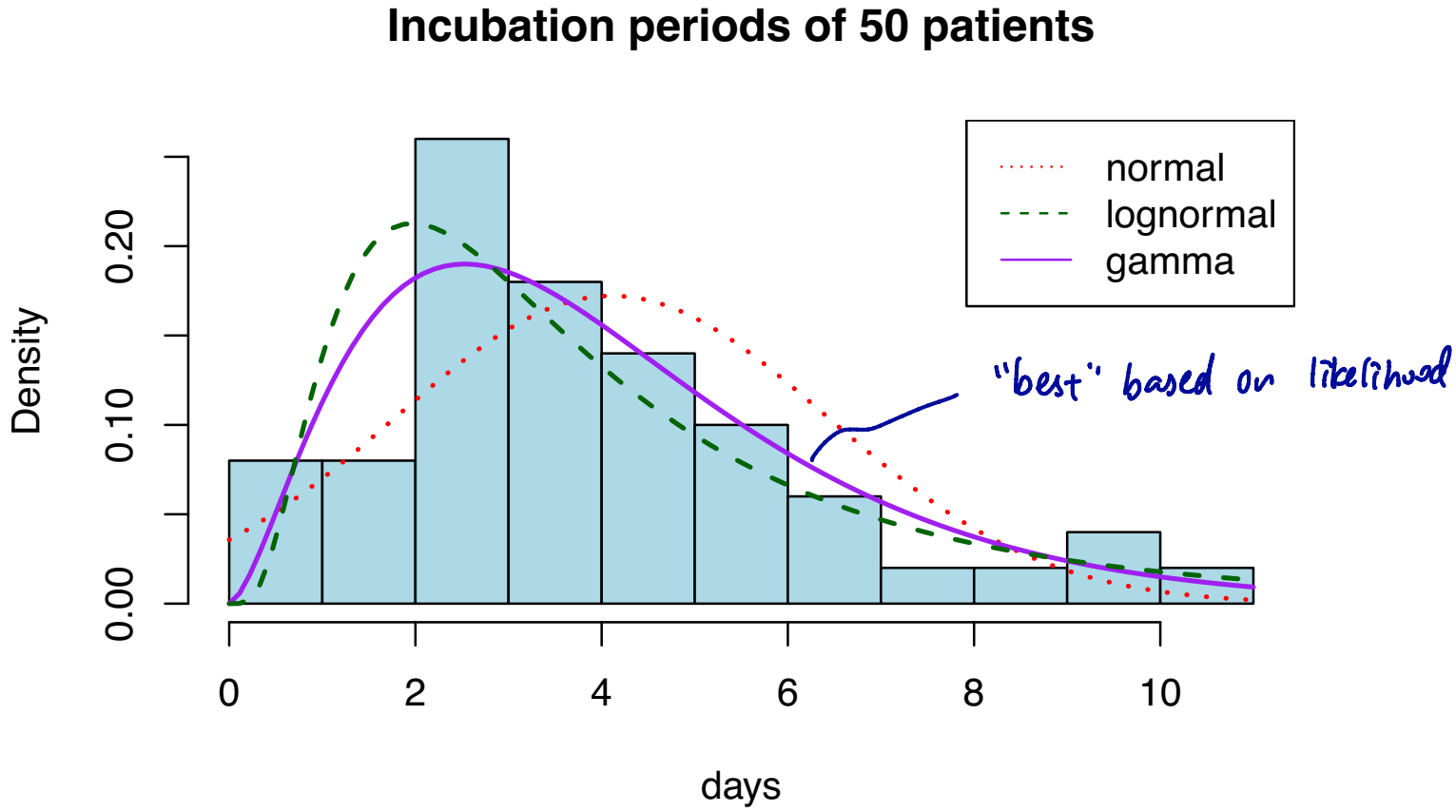
Log normal : −115.7274

normal : −112.45

w/out the log transformation, the actual likelihood 9.4×10^{-49}

4. pgamma(14, shape = 2.61, rate = 0.64) = 0.996

Discussion on model selection



- "Modeling is more of an art than a science"
- "All models are wrong; some models are useful"
- Some formal statistical procedures can be used: Compare likelihood, Chi-square test for model fitting
- People also consider: Context, assumptions, existing literature