Fall 2021

Homework 14

Maximum Likelihood Estimates (MLE), Nonparametric Statistical Procedures, Bayesian Statistics Principles

Name:	Ravleen Kaur
Class ID:	44
School ID:	1214783
Course:	Statistics for Data Science
Course ID:	DTSC-620-W01
Date:	12/19/2021

Notation Summary

• We denote the random variables (r.v.'s) arising from a random sample as subscripted uppercase letters:

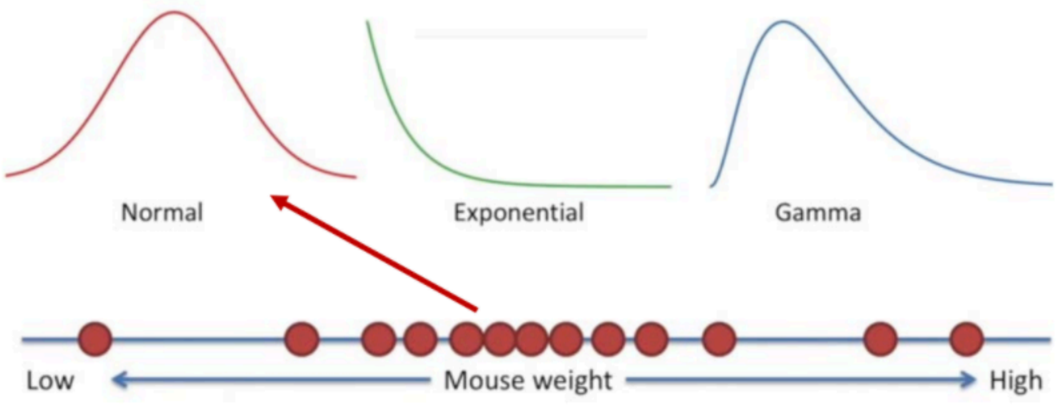
• The corresponding observed values of a specific random sample are then denoted as subscripted lowercase letters labeling r.v.'s that have some values/data/constants/literals:

- Example: Values/data/constants/literals

Data Model

• The reason one wants to fit a model/description such as distribution/pdf to given data is it to apply existing mathematical & algorithmic tools and make it easier to work with the data (of the model matching type).

- Typical r.v. data models are pdf's:



Probability & Likelihood

- Pdf p(x) model is probability bound, meaning that AUC must be equal 1.
- Likelihood model L(x) is similar data model to probability p(x) data model without AUC constraint.

Likelihood Maximal Location

- Both pdf p(x) and L(x) have maximal values at the same location x.
- Searching for the L(x) maximum location may be easier than searching for the pdf p(x) maximum location x= μ.

Maximum Likelihood Estimation (MLE)

- Using L(x)=ln [p(x)] to search for Maximally Likely Estimate (MLE) of the normal pdf p(x) mean value is simpler than using p(x) itself.

Maximum Likelihood Estimation (MLE)

- Without altering the location x=μ of the maximal value of p(x), function L(x) = ln [p(x)] transforms exponential

elements to summation elements which are easier to handle.

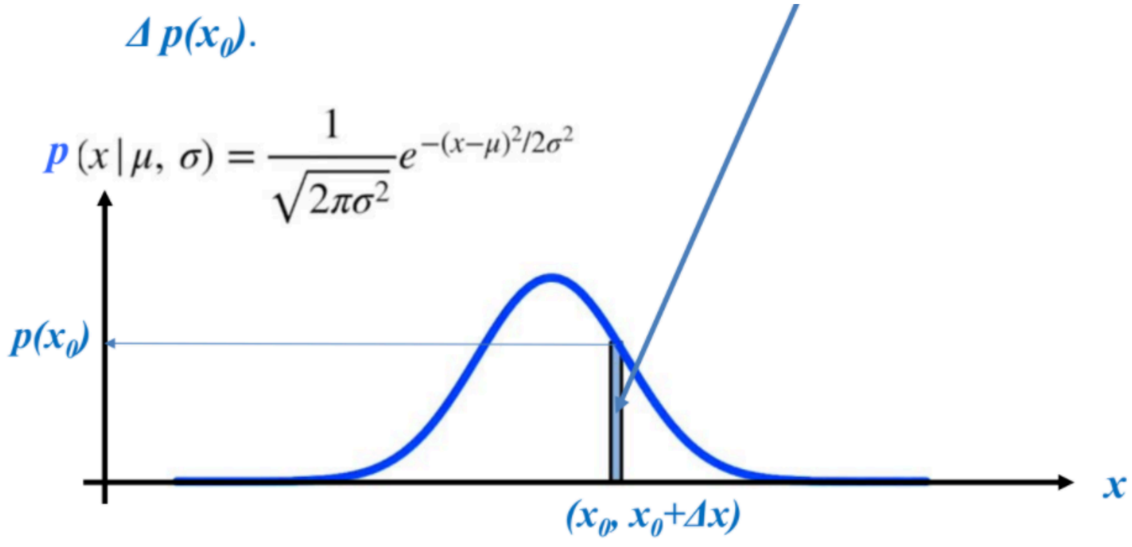
Likelihood & Probability

- Using pdf (a density function) one can find probability of x-

values in the small interval (x0,x0+Δx) as the AUC in that

interval which is approximately equal to the small rectangle

area:


$\Delta\, p(x_0).$
$p\,(x\,|\,\mu,\,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$
$p(x_0)$
x
$(x_0, x_0+\Delta x)$

Example: MLE of Normal Distribution

Parameters

• Suppose that we have observed the random sample X1, X2,

X3, ..., Xn, where Xi

~N(θ1,θ2), drawn from the stationary stochastic signal process of binary data transmission in noise over the period T of one data bit transmission:

What does stationary mean?

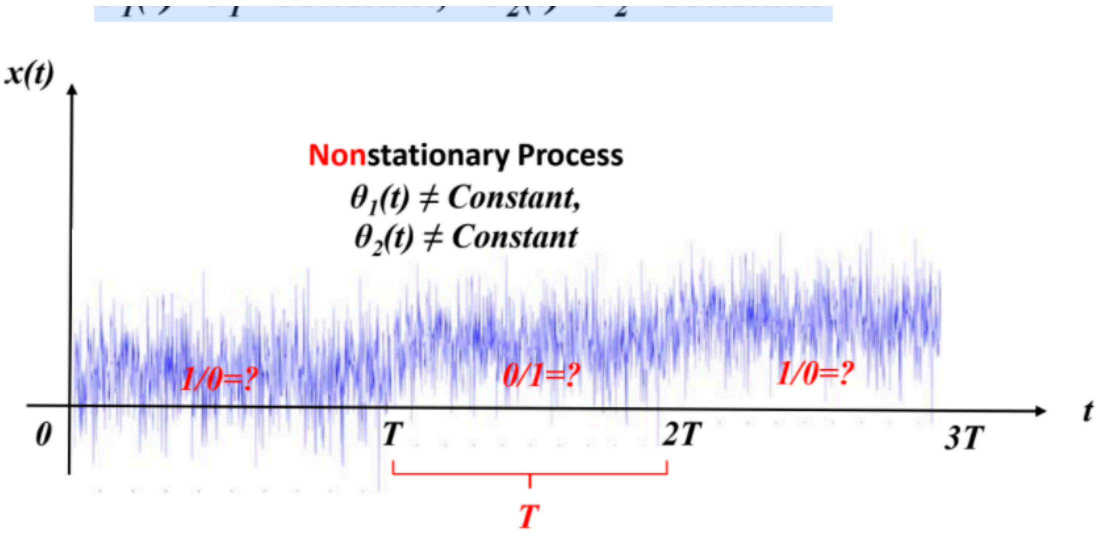
• Stationary process random sample (x1, x2, x3, ..., xn), over the

observation interval T, has model parameters (θ1,θ2) that do

not vary, remain constant:

θ1(t)=θ1=Constant, θ2(t)=θ2=Constant

Is MLE a random variable?



- Yes.

– It is produced of random variable sample values and as a

product is random too!

- However, less random.

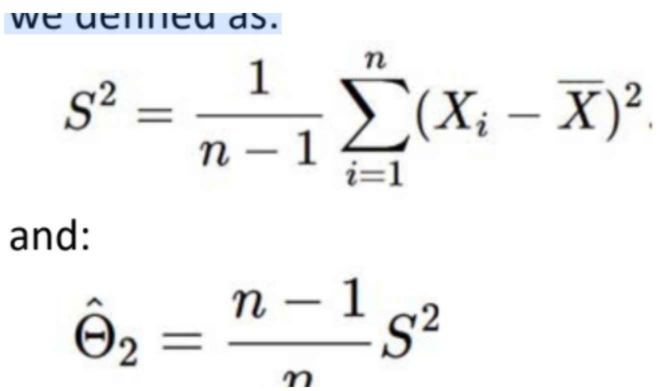
– Random−In/Random−Out!

Example: MLE Bias

- Note that Θ1 is the sample mean, X, and therefore it is an unbiased estimator of the parameter mean μ.

- MLE estimate Θ2 is very close to the sample variance which

we defined as:



we defined as:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

- and:

$$\hat{\Theta}_2 = \frac{n-1}{n}S^2$$

Did we use Probability Theory to compute MLE value?

- No!

– We used Calculus to find the maximum of the Likelihood function L(x) which is similar to the pdf p(x).

They are similar.

– Pdf p(x) is also a likelihood function L(x), but the opposite does not hold.

- Likelihood function L(x) is not a pdf. Levels of Measurement

- Measurements are always relative:

– Explicitly, or

– Implicitly.

- Which measure to take with different available data types (code)?

Nominal Level Data

- The nominal−level variables are organized into non−numeric NAMED LABELED categories that cannot be ranked (sorted) or compared quantitatively.

- Nominal–levels or categories of variables have no ordering and are – Mutually exclusive (i.e., each case object can only fit into ONLY one category) and

– Exhaustive (i.e., there is a category for each possible case).

Cannot be ordered!

Example: Nominal Level

• Shoes can be categorized based on

– Type (sports, casual, others),

– Gender (men, women, children, todler),

– Color (black, brown, others),

– Size (size−7, size−8, . ..???)

• These categories of shoes have no ordering (greater than, less than,

equal to), are mutually exclusive and exhaustive.

Ordinal Level Data

• In the ordinal level of measurement, the variables are still classified into categories, but these categories are ordered and there is no equivalent distance between the categories.

– The categories still must be mutually exclusive and exhaustive, but also have a logical order, can be ranked.

Example: Ordinal Level

• Class variable for a person can have values like:

– Upper class,

– Lower class,

– Middle class, etc.

• These values put a person into a particular category and there is also a defined implicit relative ordering between the classes like

– Upper−class > Middle−class > Lower−Class

• But there is no distance or boundaries between these classes,

Ordinal Level Data can be RANKED

- Class standing variable is measured at the ordinal level of measurement.
- The categories still must be mutually exclusive and exhaustive, but also have a logical−order/ semantic−order/implicit−order, can be

ranked.

Question: Exhaustive Ordinal Level

- What is the meaning of exhaustive?

Answer: Exhaustive Ordinal Level

- Wat is the meaning of exhaustive?
- Exhaustive means that all possible values/cases are/can−be listed/presented.

– No value/case is left unconsidered.

– The list is axhausted, complete.

Interval Level as Label Only

- In the interval level of measurement, the variables are still classified into ordered categories, but there is an equivalent distance between these categories.
- This allows for a direct comparison between categories such that the difference between any two sequential data points is exactly the same as the difference between any other two sequential data points.

– Interval neighboring values distance Δ is fixed.

Parametric Test Procedures

- Involve ratio rvv's and well assumed population parameters

– Example: Assumed normal distribution of the population with 2

parameters only mean and variance.

- Use data to learn about the population mean
- Require interval scale or ratio scale

– Whole numbers or fractions

• Example: Height in inches (72, 60.5, 54.7)

• Have stringent assumptions

– Example: Normal sample−based−statistic−distribution followed

by tests such as:

• z−test,

• t−test,

• F−test,

• ⍰2−test

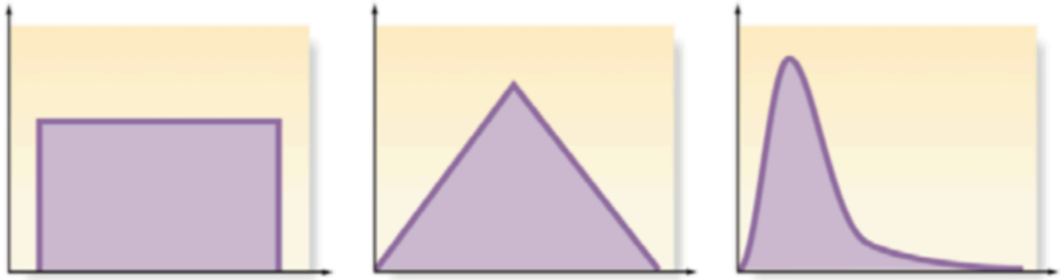
Parametric Statistics Inconvenience

• Parametric procedures with very small sample size, e.g., n<5, and very different sample rvv pdf from some symmetric bell−shaped pdf that resembles normal pdf, are unacceptable,

– Example: Even t−tests cannot be well applied.

• It needs roughly bell−shaped distribution. **Parametric Procedures & Outliers**

•



a. Flat distribution b. Peaked distribution c. Skewed distribution

Parametric Procedures are sensitive to outliers among rvv’s of

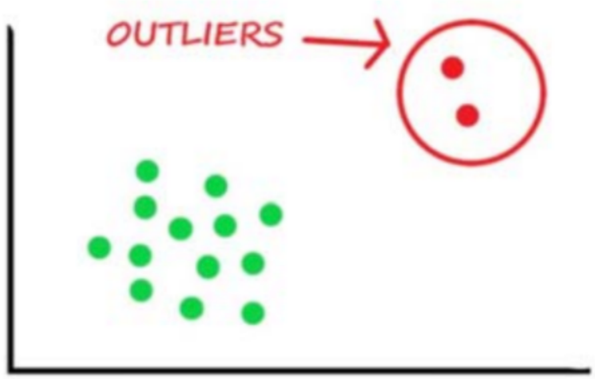
the sample.

– Outliers are rvv’s that show drastically different central

tendency (Have very different mean and/or variance)..

– Outliers can cause misleading situations:

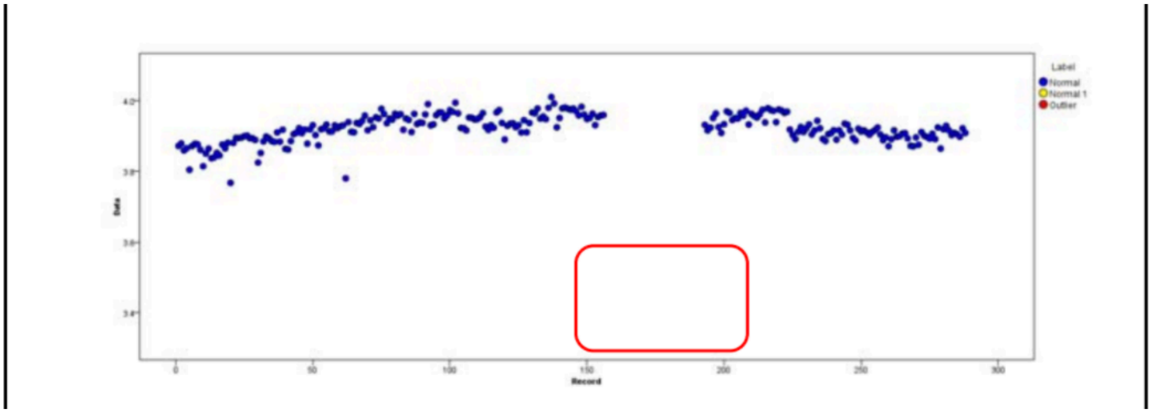
- Type−I or Type−II errors
- Change of the strength and direction of correlation.



Question: Parametric Procedures & Outliers

- When using data with outliers and parametric procedures, what is necessary to do?

Data must be cleansed of outliers before being processed using parametric procedures.



Fundamental Nonparametric Concept

- Nonparametric methods do not need data cleansing.

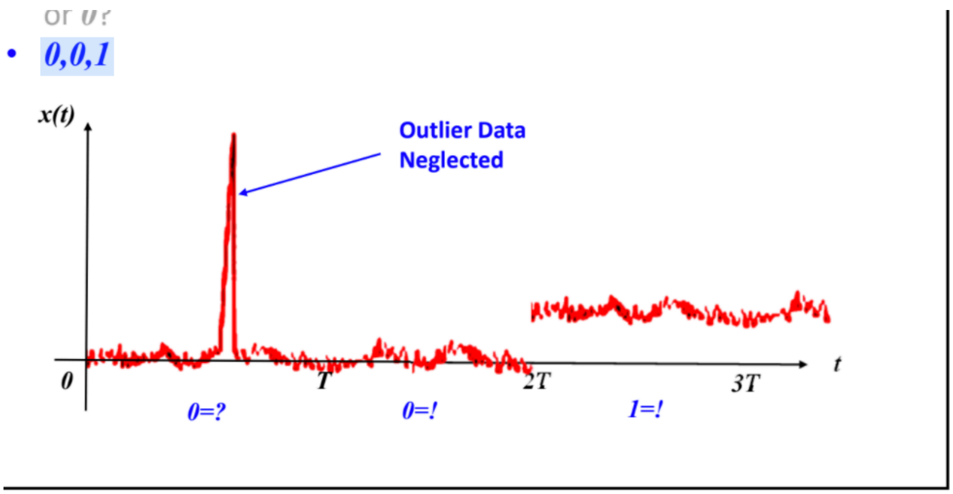
- Nonparametric procedures use rank/order of r.v.v. data instead of original r.v.v. data itself.
- Rough parameter of nonparametric procedures is median.

– The central tendency measure is median rather than the mean.

– Median is insensitive to outlier r.v.v.'s while mean, variance, covariance and correlation are very sensitive.

Question: Nonparametric Statistical Methods

- Assume we have random data sample n=10 times over T period of time.
- What does visual data inspection show as a result, data bits 1 or 0?



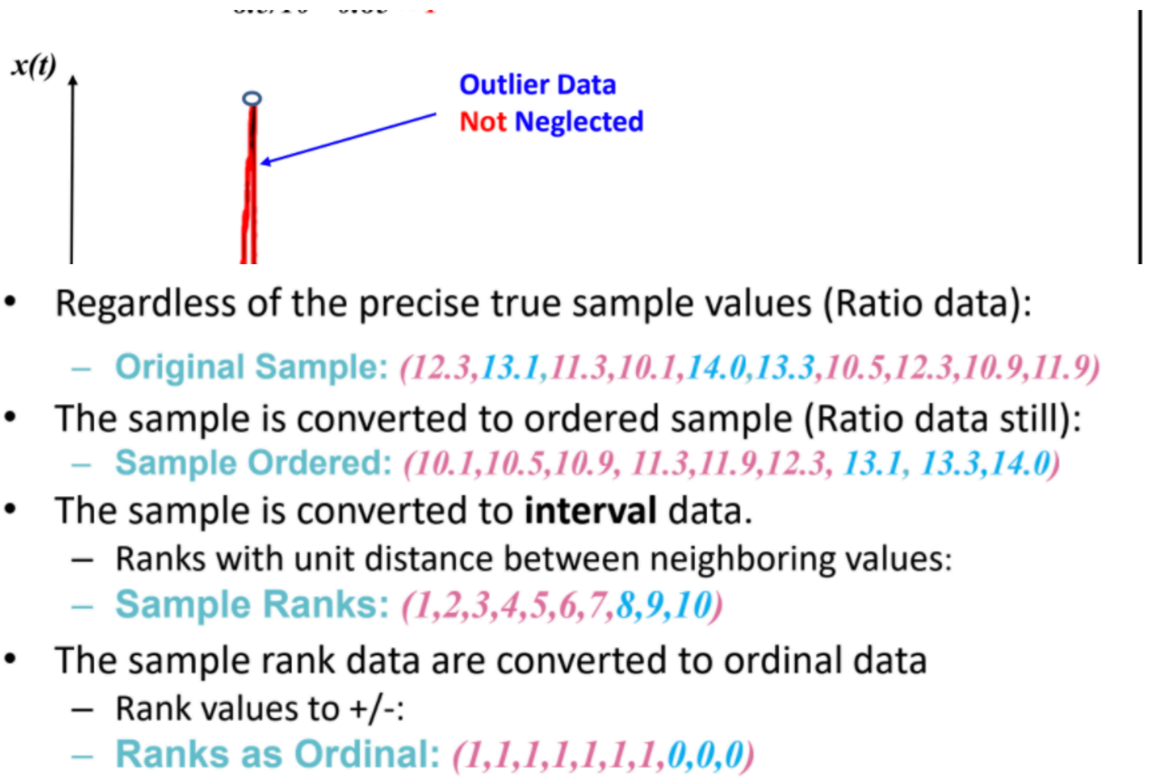
or 0?

- *0,0,1*

What kind of estimation is MLE that results in 1,0,1?

- Objective estimation!

– Objective and wrong?



- Regardless of the precise true sample values (Ratio data):
  - Original Sample: *(12.3,13.1,11.3,10.1,14.0,13.3,10.5,12.3,10.9,11.9)*
- The sample is converted to ordered sample (Ratio data still):
  - Sample Ordered: *(10.1,10.5,10.9, 11.3,11.9,12.3, 13.1, 13.3,14.0)*
- The sample is converted to **interval** data.
  - Ranks with unit distance between neighboring values:
  - Sample Ranks: *(1,2,3,4,5,6,7,8,9,10)*
- The sample rank data are converted to ordinal data
  - Rank values to +/-:
  - Ranks as Ordinal: *(1,1,1,1,1,1,1,0,0,0)*

Nonparametric Statistical Methods

• Nonparametric methods would not give such an importance to the outlier odd value which would cause wrong objective estimate.

– Nonparametric methods are ROBUST/insensitive to outlier data.

• Nonparametric methods use minimized−model approach.

– Minimal, i.e., no assumptions on the model.

• No CLT use.

•

Sign Test

• Tests one population median, ? (Greek eta)

• Corresponds to t−test for one mean

• Assumes population is continuous

– R.v.v’s can be float numbers (e.g., 3.14, 3.15, . . .).

• Small sample size n=10...20 test statistic:

• For large sample sizes n ≥ 30 normal approximation can be used within nonparametric procedure,

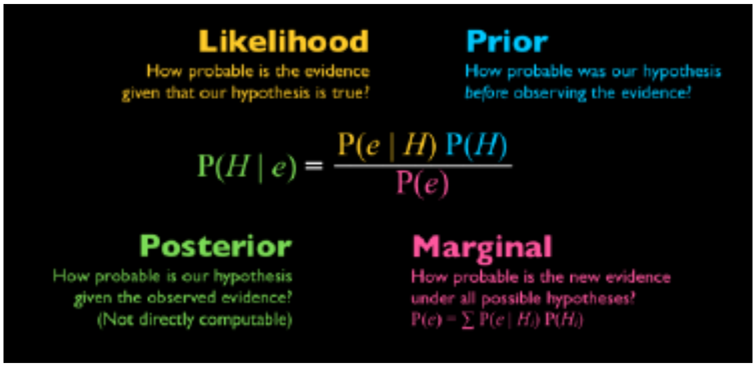
– Data are not considered as normal.

– Normal distribution N(0,1) is just used to make a decision.

P(H|E) = F(E) P(H)

The belief improvement function F(E) acting as likelihood function is processing new evidence supporting H.

F(E) + P(E|H)/P(E)



Likelihood
How probable is the evidence given that our hypothesis is true?

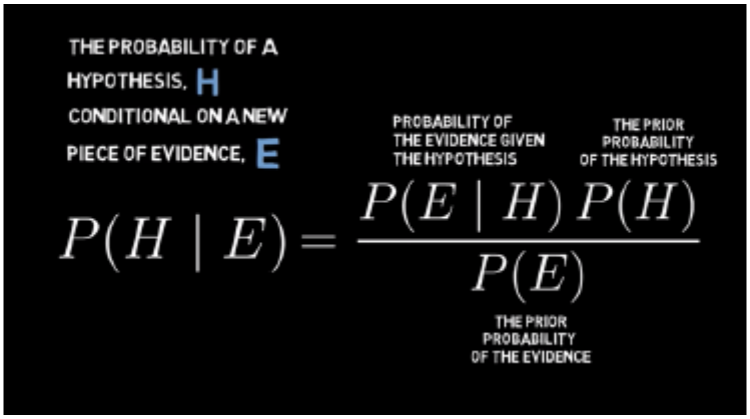
Prior
How probable was our hypothesis before observing the evidence?

$$P(H \mid e) = \frac{P(e \mid H)\, P(H)}{P(e)}$$

Posterior
How probable is our hypothesis given the observed evidence?
(Not directly computable)

Marginal
How probable is the new evidence under all possible hypotheses?
$P(e) = \sum P(e \mid H_i)\, P(H_i)$

Bayesian theorem probabilities may be replaced with pdf's, i.e., with the pdf estimates such as histograms obtained after observing large number of data samples (Statistics).



THE PROBABILITY OF A HYPOTHESIS, H CONDITIONAL ON A NEW PIECE OF EVIDENCE, E

PROBABILITY OF THE EVIDENCE GIVEN THE HYPOTHESIS

THE PRIOR PROBABILITY OF THE HYPOTHESIS

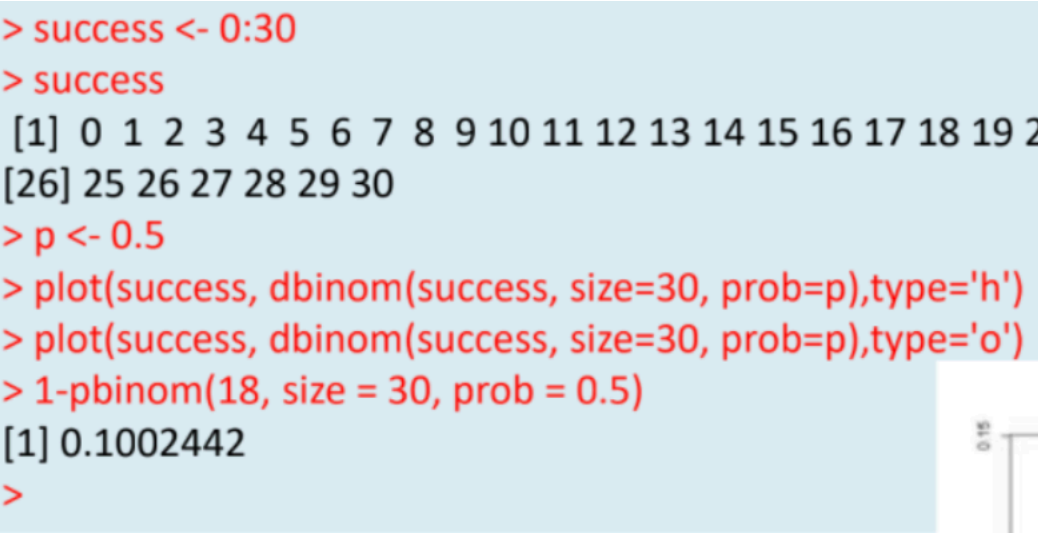
$$P(H \mid E) = \frac{P(E \mid H)\, P(H)}{P(E)}$$

THE PRIOR PROBABILITY OF THE EVIDENCE

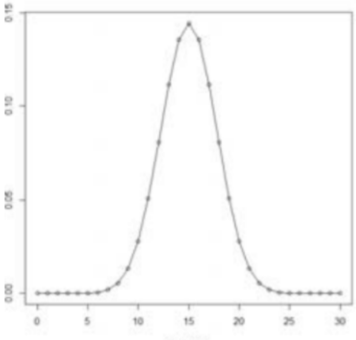
R-Session Graphs

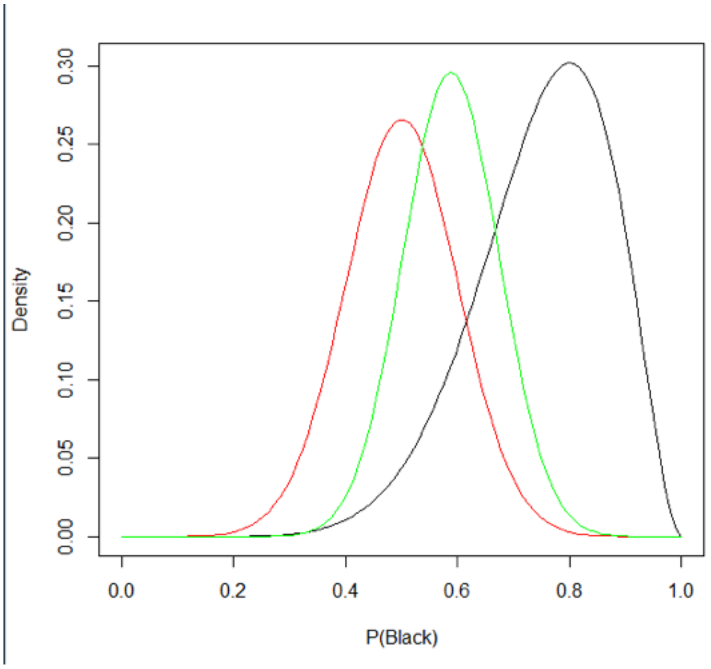
Exact p−value is 0.1002442 which

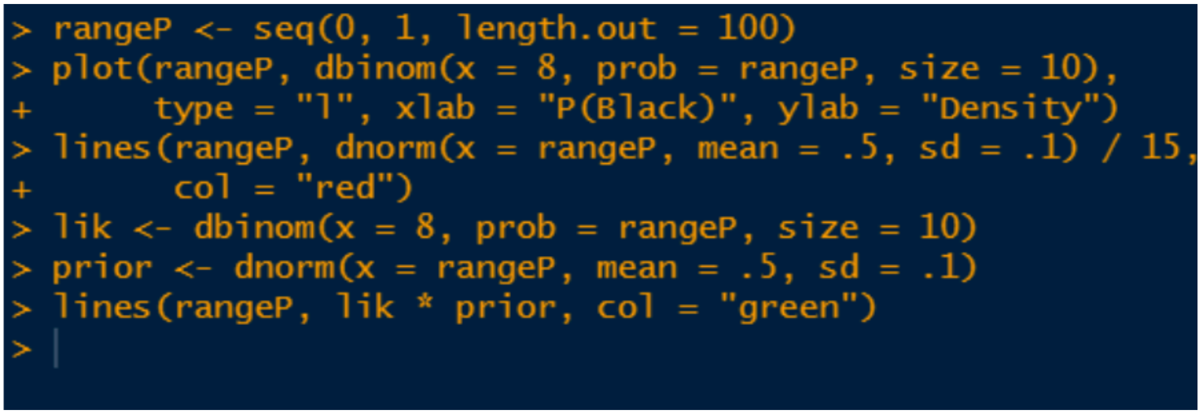
is larger than α=0.05.



```
> success <- 0:30
> success
 [1]  0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 2
[26] 25 26 27 28 29 30
> p <- 0.5
> plot(success, dbinom(success, size=30, prob=p),type='h')
> plot(success, dbinom(success, size=30, prob=p),type='o')
> 1-pbinom(18, size = 30, prob = 0.5)
[1] 0.1002442
>
```







```
> rangeP <- seq(0, 1, length.out = 100)
> plot(rangeP, dbinom(x = 8, prob = rangeP, size = 10),
+      type = "l", xlab = "P(Black)", ylab = "Density")
> lines(rangeP, dnorm(x = rangeP, mean = .5, sd = .1) / 15,
+       col = "red")
> lik <- dbinom(x = 8, prob = rangeP, size = 10)
> prior <- dnorm(x = rangeP, mean = .5, sd = .1)
> lines(rangeP, lik * prior, col = "green")
> 
```