Daily definition #1: Inferential statistics with one sample t-test

LSTP- Statistics & Coding

2022-04-06

Daily definition: inferential statistics

- Inferential statistics is a major branch of statistics.
- Making an inference is drawing conclusions about a population from a representative/random sample taken from that population.
- Inferential statistics allow us to understand how our results may differ if we take another sample from the same population and repeat the study.
- Understanding the degree of uncertainty in our results allows us to take this uncertainty into account when drawing conclusions.

Example

We can use inferential statistics to see if a drug can move from clinical trial to market.
How did they know we could give the Covid vaccine to the entire population?

Illustration 1/2

We generated a population of size 100,000 with a mean of 25 and a standard deviation of 1 and drew a random sample of size 100.

```
set.seed(350)
#Population of size 100 000 with mean 25
#and standard deviation 1
y <- rnorm(n=100000, mean = 25, sd = 1)

#Draw a random sample of size 100
x <- sample(y, size = 100, replace = FALSE)
mean(x)
```

```
## [1] 25.06182
```

```
x <- sample(y, size = 100, replace = FALSE)
mean(x)
```

```
## [1] 25.16545
```

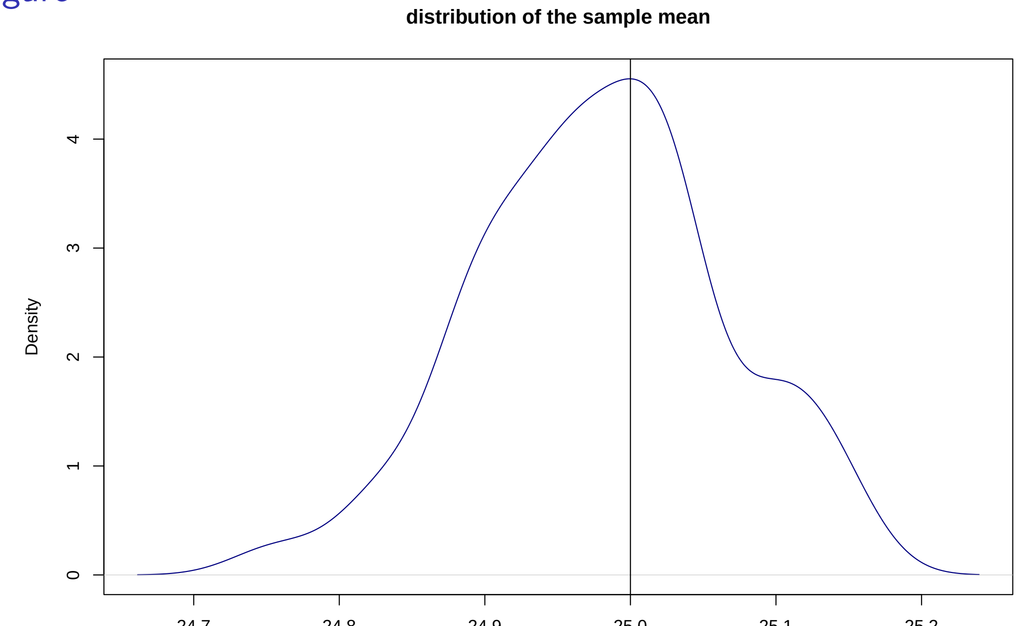
Illustration 2/2

If we repeat the random sampling 100 times:

```r
sample100 <- as.numeric(sapply(1:100,
           function(i) mean(sample(y, size = 100))))

#Figure
plot(density(sample100),
     main = "distribution of the sample mean",
     col = "navyblue")
    abline(v=25)
```

Figure


distribution of the sample mean
Density
0
1
2
3
4
24.7
24.8
24.9
25.0
25.1
25.2

Draw one sample of size 100

```r
set.seed(250)
x <- sample(y, size = 100, replace = FALSE)
mean(x)
```

```
## [1] 25.03077
```

One sample t-test to compare an observed mean to a theortical mean

The p-value = 0.756 is greater than the significance level of 0.05 implying that the sample mean is not significantly different from the population mean of 25.

```r
t.test(x, mu = 25, alternative = "two.sided")
```

```
##
##  One Sample t-test
##
## data:  x
## t = 0.31161, df = 99, p-value = 0.756
## alternative hypothesis: true mean is not equal to 25
## 95 percent confidence interval:
##  24.83485 25.22668
## sample estimates:
## mean of x
##  25.03077
```

Daily definition #2: Statistical Bias

LSTP- Statistics & Coding

2022-04-07

Daily definition: Statistical Bias

- Statistical bias can be defined as the difference between the true parameter and the mean of the sampling distribution for the estimator of the parameter.
- In other words, statistical bias is the difference between the truth and the average representation of the truth.
- For example, we expect the sample mean $\bar{X}$ to be an unbiased estimator of the population mean μ ; thus $Bias(\bar{X}) = E(\bar{X}) - \mu$ is expected to be equal to 0.

Example

- Let's say the true salary of an entry level statistician working at Google is 100k.
- If someone takes several samples of entry level statisticians working at Google, for example 30 samples, to estimate the mean salary, the estimated value should be very close to the true value if not equal.
- Multiple reasons can cause the absolute bias to be higher than expected.
- For example, using the wrong approach to choose the sample or measurement errors during data collection or even during the design of the questionnaire by using the wrong definitions.

Illustration

- Consider a population of junior statisticians working at Google.
- The population size is 10,000, the average salary is 100,000 and the standard deviation is 1.

```
set.seed(550)
#Population of entry level statisticians working at
#Google ; size 10,000, mean salary 100k
#and standard deviation 1.
y <- rnorm(n=10000, mean = 100000, sd = 1)
```

Illustrations

Draw a random sample of size 100 from the population:

```
#Draw a random sample of size 100 from the population
x <- sample(y, size = 100, replace = FALSE)
mean(x)
```

```
## [1] 99999.91
```

Let's repeat the random sampling 1000 times with a sample size of 100:

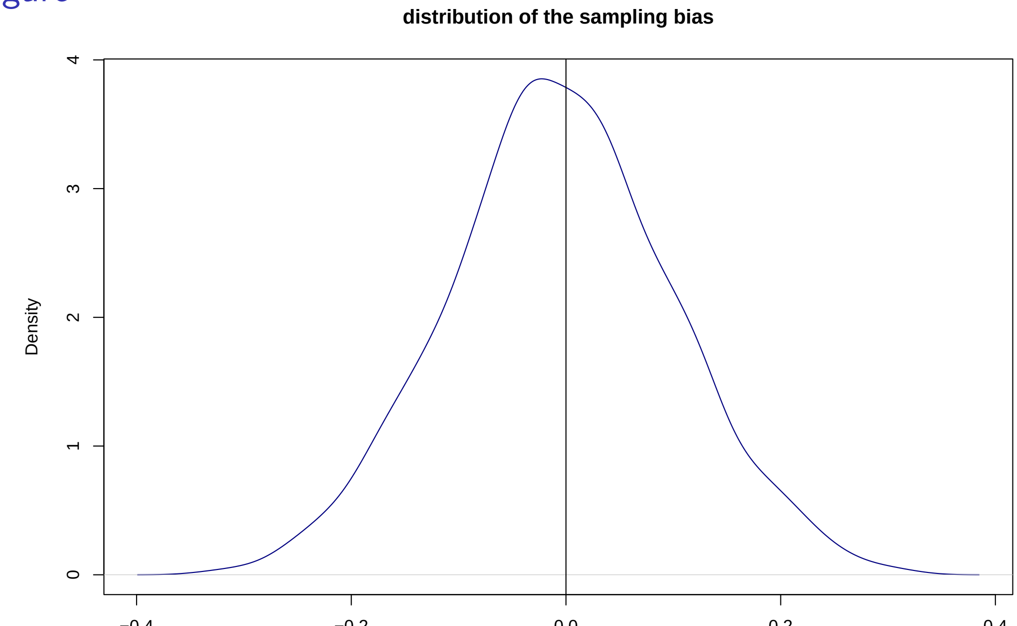
```
#repeat the random sampling with sample size 100
bias100 <- as.numeric(sapply(1:1000, function(i)
100000-mean(sample(y, size = 100))))
mean(bias100) #bias close to 0
```

```
## [1] -0.005503009
```

Figure

```r
plot(density(bias100),
main = "distribution of the sampling bias",
col = "navyblue",
xlab="difference between the true mean
salary and the estimates")
abline(v=0)
```

Figure



distribution of the sampling bias
Density
0
1
2
3
4
−0.4
−0.2
0.0
0.2
0.4

Daily Definition #3

Descriptive statistics by hand and with R

Author: LSTP-Statistics & Coding

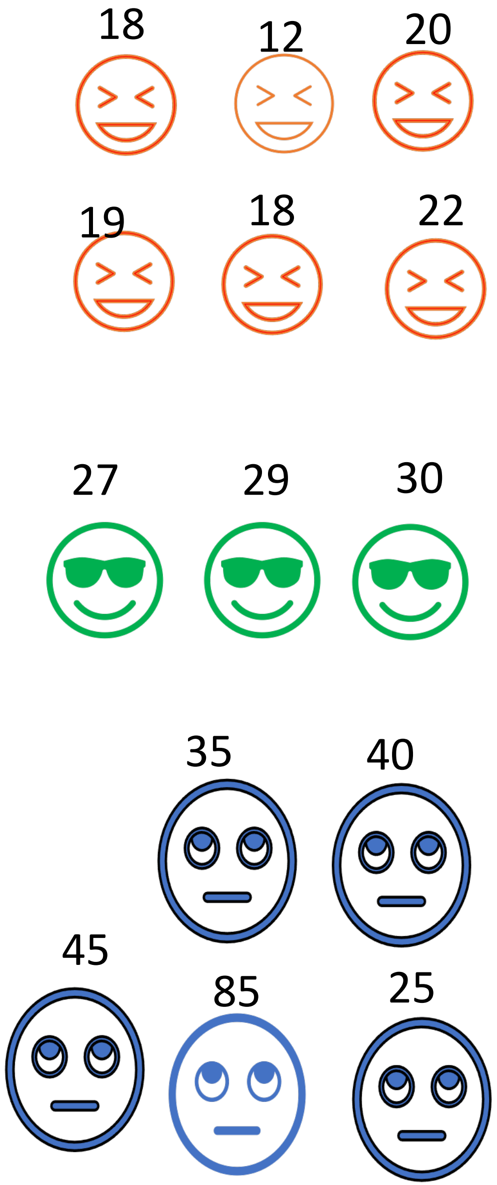
2022-04-08

Definition

Descriptive statistics are useful to understand a dataset before diving into complex analysis. Description of a dataset with basic statistics like the mean, the median or the variance, is an important step in data exploration.

Random Sample

Data Collection



Variables

- **Age:** quantitative variable
- **Color:** categorical variable, nominal scale (no order).
- **Level of satisfaction**: categorical variable, ordinal scale.

Organized dataset

ID	Age	Color	Level of satisfaction
1	18	Orange	Very satisfied
2	12	Orange	Very satisfied
3	20	Orange	Very satisfied
4	24	Orange	Very satisfied
5	28	Orange	Very satisfied
6	22	Orange	Very satisfied
7	27	Green	Satisfied
8	27	Green	Satisfied
9	33	Green	Satisfied
10	35	Blue	Neutral
11	40	Blue	Neutral
12	45	Blue	Neutral
13	85	Blue	Neutral
14	25	Blue	Neutral

Basic statistics	18 12 20 24 18 22	27 29 30	35 40 45 85 25
Frequency	6	3	5
Mean	$\frac{18+12+20+24+18+22}{6} = 19$	$\frac{27+27+33}{3} = 29$	$\frac{35+40+45+85+25}{5} = 46$
Median	12, 18,18, 24,20,22 median = $\frac{18+24}{2} = 21$	27, 27, 33 Median = 27	25,35,40,45,85 Median=40
Mode	12, 18,18, 24,20,22 Mode = 18	27,27,33 Mode = 27	No mode
Min	12	27	25
Max	24	33	85
Range	24-12 = 12	33-27 = 6	85-25 = 60

Basic statistics

	Variance	Standard-deviation
18 12 20 24 18 22	$\frac{(18-19)^2+(12-19)^2+(20-19)^2+(24-19)^2+(18-19)^2+(22-19)\text{^}2}{6} = 14.33$	$\sqrt{14.33} = 3.79$
27 29 30	$\frac{(27-29)^2+(27-29)^2+(33-27)\text{^}2}{3} = 8$	$\sqrt{8} = 2.83$
35 40 45 85 25	$\frac{(35-46)^2+(40-46)^2+(45-46)^2+(85-46)^2+(25-46)\text{^}2}{5} = 424$	$\sqrt{424} = 20.59$

Summary

Basic statistics	😆	😎	🙄	R code
Frequency	6	3	5	table(x)
Mean	19	29	46	mean(x)
Median	21	27	40	median(x)
Mode	18	27	No mode	Use table(x)
Min	12	27	25	min(x)
Max	24	33	85	max(x
Range	12	6	60	diff(range(x))
Variance	14.33	8	**424 (high variance compare to other groups)**	var(x) ; note that in R var(x) divide the numerator by (n-1) instead of n.
Standard-deviation	3.79	2.83	20.59	sd(x) ; sqrt(var(x))

Some notes with a focus on the blue and neutral group

The mean is sensitive to outliers, the median and the mode are not.

Variance and standard deviation are useful to understand how your data points are spread around the mean. In this example, the group with blue color and neutral level of satisfaction has the highest variance ; it is already an indication of potential outliers.

The average age is 46 but only one individual has more than 45 (presence of an outlier = 85).

The median age is 40, meaning 50% of individuals of the blue and neutral group have more than 40 years or you can say 50% of individuals have less than 40 years ; in this case, the median makes more sense.

The range is also a measure of variability ; helpful to be aware of potential outliers.

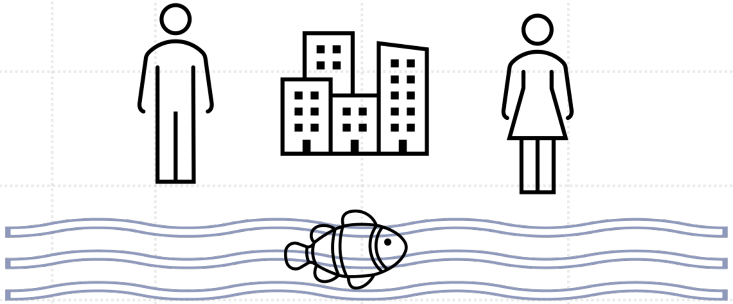
Daily Definition #4
Frequentist vs Bayesian: a Tale of Two Different Views

Author: LSTP-Statistics & Coding

2022-04-09

Question: what is the probability of having a good catch (yes/no) of fish per day?

Frequentist city

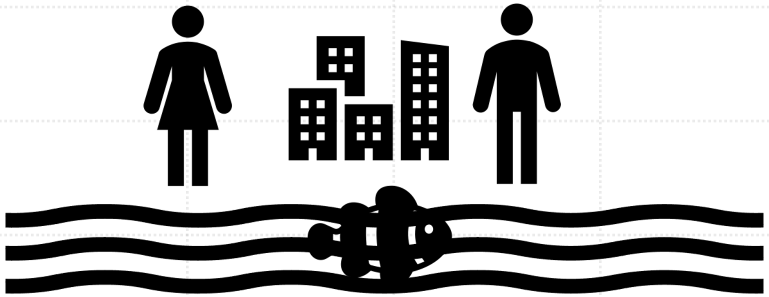


Observed data for 365 days:
225 out of 365:

$$\frac{225}{365} \approx 0.62$$

Answer: 0.62

Bayesian city



Observed data for 365 days:
225 out of 365

$$\longrightarrow \frac{225}{365} \approx 0.62$$

The wizard has an "*a priori*":
Not enough! I see a catastrophe coming ; we will lose 50% of our fish each day.

Update: $$\frac{225}{365} * 0.5 \approx 0.31$$

Answer: 0.31

Frequentists point of view:

- **Interested in the relative frequency of the event (point estimates)** ; in other words, the number of times the event occurs divided by the total number of events occurring. In our example the event of interest is having a good catch and the possibilities per day are {yes, no}. For 365 days we can have:

  yes, yes, no, yes, …., no

Then we calculate the number of “yes” among the total number of events.

$$\hat{\theta} = \frac{225}{365}$$

The sample proportion $\hat{\theta}$ corresponds to the maximum likelihood estimator.

- Parameters are fixed and data are random ; we write p($data$ | θ).

Bayesians point of view:

- Not happy with point estimates.

- Express **uncertainty** about the probability of having a good catch $\boldsymbol{\theta}$**.** Thus, they update the information provides by the data with *a prior probability* **p(*θ*)** using Bayes’ theorem:

$$\mathbf{p}(\theta\ |\mathbf{data}) = \frac{\boldsymbol{p(data|\theta)} * \boldsymbol{p(\theta)}}{\boldsymbol{p(data)}} \xrightarrow{\textbf{called}} \text{posterior probability}$$

p(data) is a normalizing constant or scaling factor ; $\boldsymbol{p(data|\theta)}$ is the data-generating process.

- We often reduce the Bayes’ formula to this:

$$\mathbf{p}(\theta\ |\mathbf{data}) \propto \boldsymbol{p(data|\theta)}\ * \boldsymbol{p(\theta)}$$

- Data are fixed and parameters random ; we write p(θ |$data$).

Note: in statistics everything after the symbol | (under the condition) is considered as fixed.

Et voilà!

Thanks for reading!

Daily definition #5: Bayesian statistics - Impact of the prior distribution

LSTP- Statistics & Coding

2022-04-10

Daily definition #5: Impact of the prior distribution

- Consider a Beta-binomial model with $\theta \sim Beta(\alpha, \beta)$, $X \sim Binomial(n, \theta)$.
- The prior: $p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$.
- The likelihood: $p(x|\theta) \propto \theta^{x}(1-\theta)^{n-x}$.
- The posterior: $p(\theta|x) \propto p(x|\theta)p(\theta) - - - > p(\theta|x) \propto \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}$.
- Expectation: $E(\theta|x) = \dfrac{\alpha + x}{\alpha + \beta + n}$.

Example

- If $\alpha = \beta = 1 - -- >$ non-informative prior $- - - >$ posterior mainly influenced by the likelihood meaning same shape as the likelihood.
- If $\alpha > 1$ et $\beta > 1 - -- >$ Posterior distribution influenced by both the likelihood and the prior distribution.

Illustration

Set values of parameters

```r
alpha=c(1,2,5,1,2,7)
beta=c(1,2,5,1,2,5)
n=c(3,3,3,15,15,15)
x=c(0,0,0,3,3,3)
theta<-seq(0,1,0.001)
```

Plot densities and to estimate the theoretical and empirical expectation (1/3)

```
plotden<-function(theta,x,n,alpha,beta)
  {
  #Initialization --> vector of theoretical expectation
  Esp.bayes_theo<-c()
  #Initialization --> vector of empirical expectation
  Esp.bayes.emp<-c()
  par(mfrow=c(2,3),cex=0.5)
```

Plot densities and to estimate the theoretical and empirical expectation (2/3)

```
for(i in 1:length(alpha)){
  set.seed(1971101811)
  #prior distribution
  prior<-dbeta(theta, alpha[i], beta[i])
   #likelihood
  likelihood<-dbinom(x[i],n[i],theta)
  #normalized likelihood
  likelihood.norm<-likelihood/(sum(likelihood)*0.001)
  alpha.star<-alpha[i]+x[i] #update
  beta.star<-beta[i]+n[i]-x[i] #update
  posterior<-(dbeta(theta, alpha.star,beta.star)) #posterior
```

Plot densities and to estimate the theoretical and empirical expectation (3/3)

```r
    eq <- "=";
    neq <- "n =";
    xeq <- "x=";
    plot(theta, prior, type="l",col="navyblue",
         main = bquote(alpha ~ .(eq) ~ .(alpha[i]) ~
      beta ~ .(eq) ~ .(beta[i]) ~.(neq) ~
        .(n[i]) ~ .(xeq) ~ .(x[i])),ltw=2,
         ylab="Distributions", xlab= expression(theta), ylim = range(c(pri

    lines(theta,likelihood.norm,type="l", col="purple", lty=2, ltw=2)
    lines(theta,posterior,type="l", col="red",lty=3,ltw=2)
    op <- par(cex = 0.4)
    legend("topright",
           c("Prior","Likelihood.norm","Posterior"),
           col=c("navyblue","purple","red"),
           text.col = "black",bty="n",lwd=c(2,2,2),lty=c(4,2,3))
```

Estimation of expectations

```
    Esp.bayes_theo[i]<-
  round(((alpha[i]+x[i])/(alpha[i]+beta[i]+n[i])),2)
    n.samples <- 10000
    Esp.bayes.emp[i]<-
      round(mean(rbeta(n.samples, alpha.star, beta.star)),2)
  }
 return(list(Esp.bayes_theo = Esp.bayes_theo,
             Esp.bayes_emp = Esp.bayes.emp))
    }

Res<-plotden(theta,x,n,alpha,beta)
```

Figure (1/2)

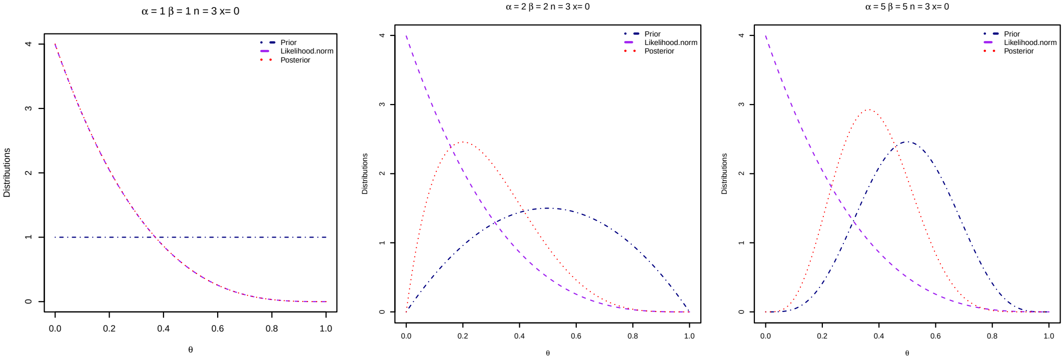
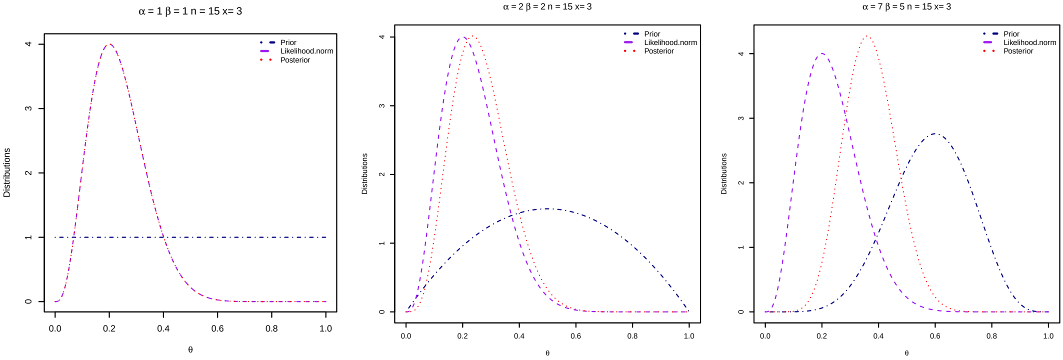

α = 1 β = 1 n = 3 x= 0
Distributions
θ
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
Prior
Likelihood.norm
Posterior
α = 2 β = 2 n = 3 x= 0
Distributions
θ
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
Prior
Likelihood.norm
Posterior
α = 5 β = 5 n = 3 x= 0
Distributions
θ
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
Prior
Likelihood.norm
Posterior

Figure (2/2)


α = 1 β = 1 n = 15 x= 3
α = 2 β = 2 n = 15 x= 3
α = 7 β = 5 n = 15 x= 3
Prior
Likelihood.norm
Posterior
Distributions
θ
0
1
2
3
4
0.0
0.2
0.4
0.6
0.8
1.0

Theoretical expectation for $\alpha = \beta = 1$, $n = 3$, et $x = 0$:

$$E(\theta|x) = \frac{1+0}{1+1+3} = \frac{1}{5} = 0.2$$

```
Res <- plotden(theta,x,n,alpha,beta)
#Theoretical expectation
Res$Esp.bayes_theo
```

```
## [1] 0.20 0.29 0.38 0.24 0.26 0.37
```

```
#Empirical expectation
Res$Esp.bayes_emp
```

```
## [1] 0.20 0.29 0.38 0.24 0.26 0.37
```

STATISTICS
&
CODING
Learn Through Practice

Daily Definition #6: Causal inference (simply explained)

And no, this is not to say that correlation does not imply causation (almost).

Author: LSTP-Statistics & Coding
2022-04-11

Causal inference can be defined as a branch of statistics useful to model the relation between a cause and an effect called outcome.

- In a more natural way, we can say that causal inference is part of our daily life.
- As human beings, we like to find causes for observed phenomenon.
- We might think that we are almost a random generator of reasons why what we observed happened.

Let us take an example where we model the dynamic between a boss and an employee.

Employee thinking:
bad work → Anger

Anxious Employee

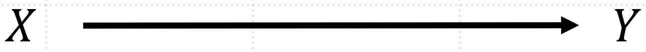
Boss thinking:
Bumped my car this morning → Anger

Angry boss

The employee thinks he did a bad job, which made his boss angry.

When the boss is angry because she had an accident with her car.

More formally, consider a cause X and an outcome Y. In causal inference, we use what we call **Directed Acyclic Graphs (DAGs)** or causal graphs to depict relation between variables.

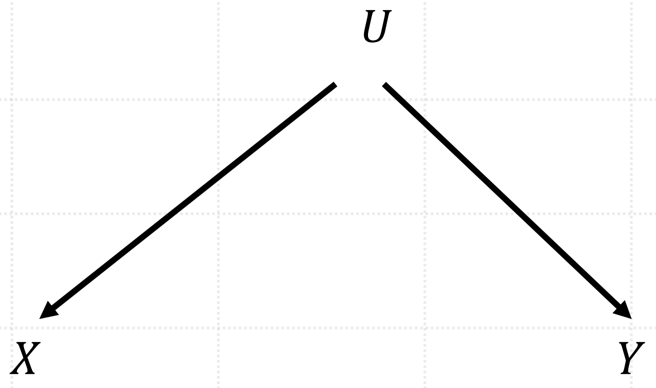


It is important in causal inference to clearly establish the order of relationship between variables.

Stating ***“eating sugar causes diabetes”*** is different from stating ***”diabetes causes eating sugar”.*** Both may work but they lead to two different research questions.

In causal inference, we seek to **quantify and isolate** the influence of X on Y. In other words, we want to make sure that X really causes Y **without artifacts ;** a major difference with the standard statistical modeling approach.

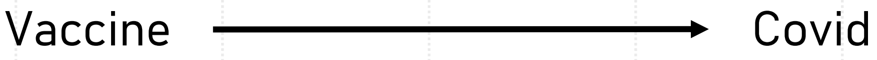
Indeed, we could have a third variable U that causes the observed association between X and Y.



If the variable U is not included in our model, we might conclude that X is correlated to Y which is **not true** ; U is called a **confounding variable.**

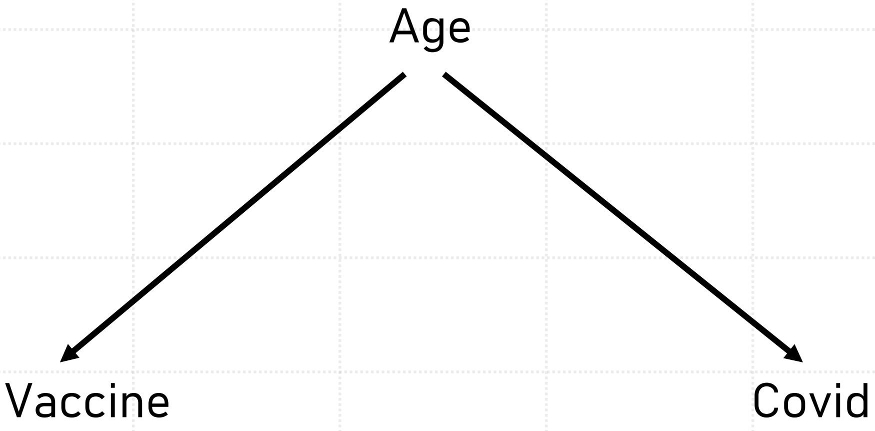
One very important assumption in causal inference is that we should not have **unmeasured confounders** (I know a big assumption).

Let us take another example where an individual received the covid vaccine and did not develop the disease. We might think that the vaccine caused the absence of disease but that may not be the case.



Vaccine → Covid

The true reason may be that the person is young and would not develop the disease even without a vaccine.



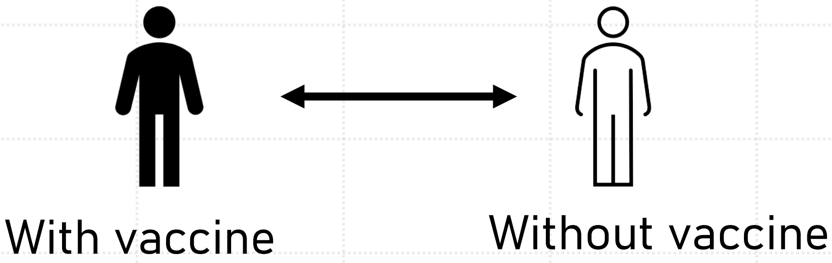
Age

Vaccine

Covid

However, how do we know? We cannot observe for the same individual a situation where the vaccine was taken and a situation where it was not.

In causal inference we have this great concept called **counterfactual**.
What would happen if the vaccine was not taken?
What is the **potential outcome**?
It is also called **“a missing-data problem”.**



If we had the power to know for a same individual what consequence taking or not the vaccine would provoke, we could measure the effect of the vaccine on the development or not of the disease.

Since we cannot measure individual causal effects, we always rely on **group comparisons**.
For example, we may compare a group called the **treatment group** to a group called the **control group**.

Key points about causal inference:

- Goal: model the relation between a cause and an outcome.
- Define the conceptual model using a DAG or causal graph.
- Ideally include all potential confounders (not very realistic)
- Emulate a counterfactual model by defining for example a treatment group and a control group to quantify the impact of a treatment on a disease.

Important: There are other assumptions to consider when building a causal model.

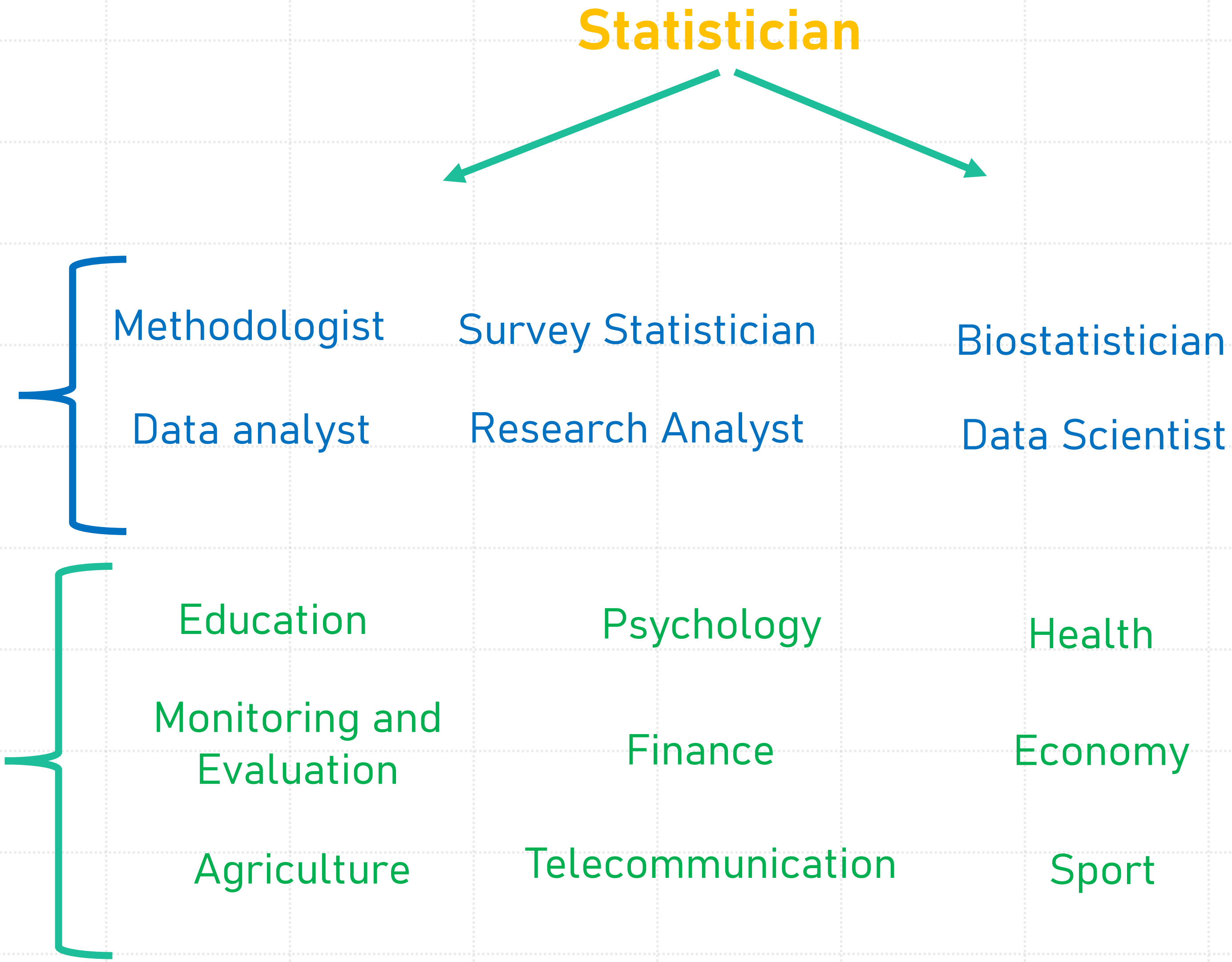
Et voilà!

Thanks for reading!

Daily Definition #7/#7 final: Why should you hire a statistician

Author: LSTP-Statistics & Coding
2022-04-12

Shapeshifters, statisticians can perform several types of work.



Statistician

Methodologist

Survey Statistician

Biostatistician

Data analyst

Research Analyst

Data Scientist

Education

Psychology

Health

Monitoring and Evaluation

Finance

Economy

Agriculture

Telecommunication

Sport

Wherever there is data to collect or analyze, a statistician can find his place.

Definition: A statistician knows how to design a survey, collect data, analyze a wide variety of data, and has computer skills to solve problems.

Statisticians typically do the following:

- Decide what data are needed to answer specific questions or problems
- Apply mathematical theories and techniques to solve practical problems in business, engineering, sciences, and other fields
- Design surveys, experiments, or opinion polls to collect data
- Develop mathematical or statistical models to analyze data
- Interpret data and communicate analyses to technical and nontechnical audiences
- Use statistical software to analyze data and create visualizations to aid decision making in business

Source: https://www.bls.gov/ooh/math/mathematicians-and-statisticians.htm

Sample job description for a methodologist

Key responsibilities of this role include:

• Work collaboratively to develop research plans, study designs, and analysis plans

• Use intermediate to advanced R skills to perform multivariable statistical analysis

• Conduct exploratory and statistical analyses of the data, primarily using R

• Produce meaningful reports and graphical presentations of data

• Document project work according to standards of practice guidelines

• Participate in writing research papers and reports

• Take part in team initiatives to develop an analytical knowledge base, standards of practice Basic

Requirements:

• MSc in Biostatistics, Epidemiology, Health Sciences, or related field

Is something missing?

Let us know in the comments.

Thanks for reading!