STA442: Methods of Applied Statistics
University of Toronto
Fall 2022

Course introduction

Instructor: Jun Young Park

Note: This course material is for students' reference purposes only, and the reproduction of this material in any form is *not* allowed (prohibited).

What you've learned so far

Based on the course prerequisites, you learned

- **STA237**: An introduction to probability using simulation and mathematical frameworks, with emphasis on the probability needed for more advanced study in statistical practice. Topics covered include probability spaces, random variables, discrete and continuous probability distributions, probability mass, density, and distribution functions, expectation and variance, independence, conditional probability, the law of large numbers, the central limit theorem, sampling distributions. Computer simulation will be taught and used extensively for calculations and to guide the theoretical development.
- **STA238**: An introduction to statistical inference and practice. Statistical models and parameters, estimators of parameters and their statistical properties, methods of estimation, confidence intervals, hypothesis testing, likelihood function, the linear model. Use of statistical computation for data analysis and simulation.
- **STA302**: Introduction to data analysis with a focus on regression. Initial Examination of data. Correlation. Simple and multiple regression models using least squares. Inference for regression parameters, confidence and prediction intervals. Diagnostics and remedial measures. Interactions and dummy variables. Variable selection. Least squares estimation and inference for non-linear regression.
- **STA303**: Analysis of variance for one-and two-way layouts, logistic regression, loglinear models, longitudinal data, introduction to time series.

These are already more than sufficient knowledge that an undergraduate statistics student should know.

Expectations for this course

- Most of the students in this course will be graduating soon. Some of you will get a job in industry, others will pursue graduate studies, etc.
- Regardless, it is likely that people who'll be working with you expect you to be "experts" in statistics, because you took many statistics courses during your degree program.
- Most of the research questions you'll be working with will be pretty straightforward and can be answered using prerequisites for this course.

To work with data in independent and responsible manner, it is critical that you know what's *really* going on with the statistical models you use.

Some examples

- “What is the difference between confidence and probability?”
- “I tested for correlation between two variables and the p-value is 0.5. So I conclude that these variables are independent.”
- “I have collected three subjects’ data to test the efficacy of a drug. I am not sure if it is sufficient.”
- “I developed a new hypothesis testing method. I will be rejecting every null hypothesis regardless of the data I have.”
- “Do we *really* need to use paired t test? Why don’t we just use independent two-sample t test? I learned it in my intro stats course.”
- “I first tested for the gender effect in salary after using multiple linear regression with age and gender. It didn’t provide me a p value less than 0.05. I dropped age term and fitted the model again, and it’s now 0.03. Great!”
- “I tested if a gene is associated with breast cancer using t test. Out of 20000 genes, there were 800 genes with p-value less than 0.05.”

Course theme: revisiting statistical methods

If what you can do is (only) fitting a (generalized/mixed) linear model to some data and get a p-value using a software, your understanding is very limited.

The **purpose of this course** is

- revisit key statistical concepts that we thought we knew,
- see “what happens” if we do things right or wrong,
- learn alternative methods when classical methods fail,
- learn relationship between (seemingly different) methods,

and, importantly,

- deduce the best practice of statistics.

Course theme: revisiting statistical methods

Compared to STA442 offered in the past/taught by other instructors:

- We'll not learn Bayesian modeling (using INLA) and survival analysis. There are lots of theories there and I don't think the topics can be learned in a few weeks at the undergraduate level. Interested students can take
  - STA475 (F): Survival Analysis
- We'll not learn time series or spatial methods. I believe it can be covered in other courses.
  - STA457 (F/S): Time Series Analysis
  - STA465(S): Theory and Methods for Complex Spatial Data
- However, we'll learn how these methods can be *interpreted* in terms of correlated data analysis.
- We will focus more on computer intensive methods using `R`.

Evaluations

- **Attendance** (5%)
- **Class participation**: Up to 3% of the final grade can be topped up based on the participation. It is highly encouraged to ask more questions and add your thoughts on the course material.
- **Homework** (40%): There will be **four** homework sets, requiring a solid understanding of theories and computing. Homework is a **group work** consisting of 3-4 students, and students submit their responses as a team.
- **Midterm exam** (30%): In person exam is scheduled to be held in November –. It is to be done individually in person and evaluates technical skills learned in class.
- **Final project** (25%): Students will be asked to read and summarize a research paper and conduct simulation studies as needed.
- **Peer evaluations**: Students will also be asked to rate other members' contributions to the work.

Assignments

There are 4 assignments in total

- Due dates (tentative): Oct 4, Oct 25, Nove 8, Dec 6.
- There are team of 3-4 students assigned randomly for each assignment.
- You'll be asked to complete peer evaluations based on the contributions.
  - Grades for peer evaluations will be disclosed after all assignments. I don't disclose peer evaluations for each assignment.
- You must not share answers codes with students other than your teammates. Violation of this policy is an academic offence and will be investigated and reported.
- All codes should be written in `R`. (No `Python`, `SAS`, `SPSS`, etc.)
- Up to 20% of the total marks can be deducted based on neatness, clarity, conciseness of the answers.
  - Most answers are straightforward. Don't write several paragraphs describing everything.
  - Check grammars and make it 'readable' by using margins, spacings, etc.
  - Make your codes optimized and commented.
  - LaTeXor R markdown is recommended, but not required.

Midterm exam

- It is scheduled early November (tentative)
- It evaluates
  - Technical skills covered in class.
  - (Pseudo) `R` codes.
  - Interpreting data analyses.

Final project

- 25% of your final grade.
- 4-5 pages report, to be completed individually.
- You'll be assigned a research article related to course materials. You'll be asked to
  - summarize the background, statistical methods, and results
  - list the limitations and potential improvements.
  - conduct simulation studies and data analysis, if needed.
- Articles will be randomly assigned out of 5-6 articles.
- More information wil be posted on Quercus.

Tips

- Up to 3% of the final grade can be topped up based on the participation.
- There are no silly questions.
  - If you didn't understand course material clearly, likely many classmates didn't understand either.
  - I like students who ask questions and are treating course materials critically.
  - There are many students enrolled in this course, and the classroom is large. It is encouraged to sit close.
- For those who need a reference letter from me, I will prioritize students whose class participation and exam grade are superb.

Piazza

We will use Piazza for discussing any questions on course materials, other than homework or exam questions:

https://piazza.com/utoronto.ca/fall2022/sta442

Office hours

Instructor

- Jun Young Park: Tuesdays 9-10AM, Thursdays 9-10AM

Teaching Assistants

- Ying Zhou: Fridays 3-5PM, until October 4
- Sigeng Shen: Fridays 3-5PM, until November 1
- Jiayue Huang: Fridays 3-5PM, until December 6

Notes:

- Do not dig for answers to the homework problems during the office hours. TAs and I will not answer questions such as
  - I don't know anything but I need hint.
  - Is this response correct?
- During the office hours, you need to first show your understandings of course materials to get hints.

Questions?

Monte Carlo methods (simulations)

A simple example

We know that the area of a circle with radius 1 is π. Proving it actually is not that easy if you haven't learned calculus. Suppose, instead, that we know the formula for the area of a square (much easier!). From here,

- We may draw a random dot within this square.

  R codes

  ```
  ## (x,y) is the coordinate of a random dot
  x = runif(1, min = -1, max = 1) #1 random number from Unif(-1,1)
  y = runif(1, min = -1, max = 1)
  ```

- While A is unknown, the probability that a dot will lie within a circle is $A/4$.
- We may repeat drawing dots M times, say 20000 times, and count the proportion of dots that lied within the circle.

$$\frac{\text{Area of a circle}}{\text{Area of a square}} = \frac{A}{4} \approx \frac{M_0}{M} \Rightarrow A \approx 4 \times \frac{M_0}{M}$$

- We can simply increase the number of dots to get a closer estimate of the area.

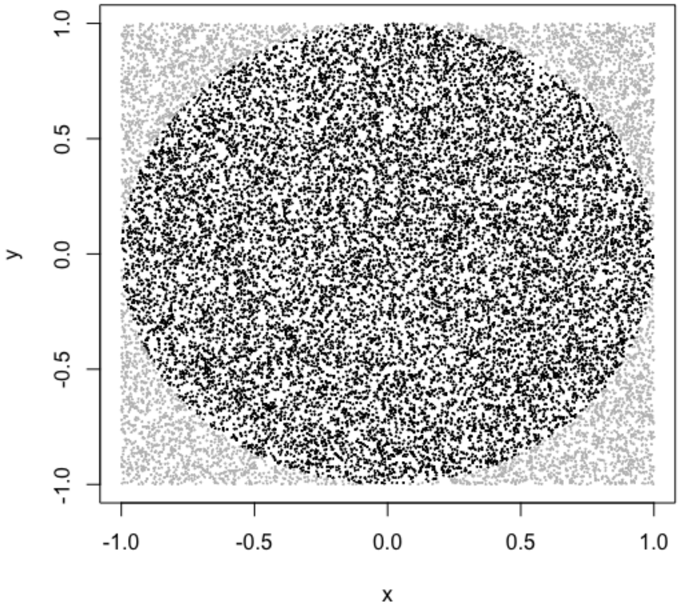
A simple example

R codes

```
set.seed(442)
M = 20000 #number of random dots
x = runif(m, min = -1, max = 1) #n.sim random numbers from Unif(-1,1)
y = runif(m, min = -1, max = 1)

M0=0
for (i in 1:m){
  if (x[i]^2+y[i]^2<1){
    M0=M0+1
    }
  }

4*M0/M


##Or, simply,
4*mean(x^2+y^2<1)
```

A simple example

In this implementation, we got an estimate of the area of the unit circle by 3.114. Not the same, but quite similar.



R codes

```
index=which(x^2+y^2<1)
plot(x,y,type="n")
points(x[index],y[index], cex=0.1)
points(x[-index],y[-index], cex=0.1, col="gray")
```

A simple example

How do we quantify how close our estimates are?

If we define $X = 1$ if X (a random dot) lies in a circle and 0 otherwise, then we have

$$X \overset{i.i.d}{\sim} \text{Bernoulli}(A/4)$$

which, by central limit theorem,

$$\hat{p} \equiv \overline{X} \sim \mathcal{N}\left(p, \frac{p \times (1-p)}{M}\right)$$

Therefore, with $\hat{p} = M_0/M$, we can construct a 95% confidence interval by

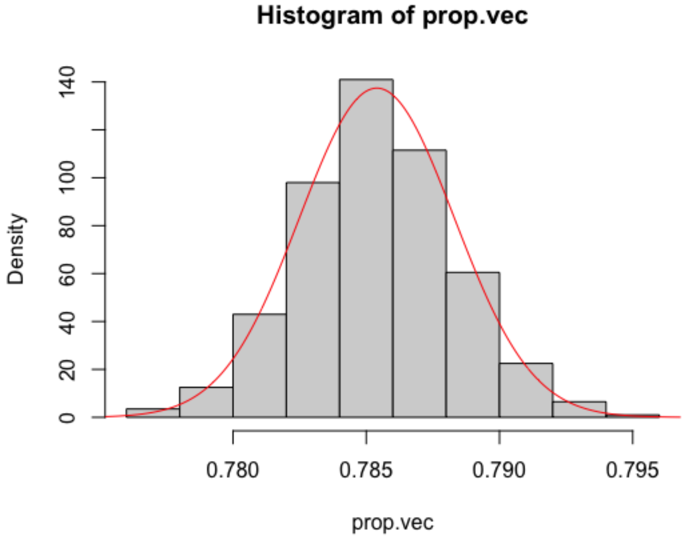
$$\hat{p} \pm 1.96 \cdot \sqrt{\frac{p \times (1-p)}{M}}$$

which is approximated by

$$\hat{p} \pm 1.96 \cdot \sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{M}}$$

A simple example

If we repeat this 1000 times and get estimates of $\pi/4$, the distribution looks normal with mean $\pi/4$.



```
set.seed(442)

n.sim=1000   #Number of repeats
M = 20000
prop.vec=rep(NA,n.sim)
for (j in 1:n.sim){
  x = runif(m, min = -1, max = 1)
  y = runif(m, min = -1, max = 1)

  M0=0
  for (i in 1:m){
    if (x[i]^2+y[i]^2<1){
      M0=M0+1
    }
  }
  prop.vec[j]=M0/M
}

hist(prop.vec, prob=T)
s=seq(0, 1, length.out = 10000)
mu = pi/4; sigma= sqrt(pi/4*(1-pi/4)/M))
lines(s, dnorm(s, mean = mu, sd = sigma, col="red"))
```

Simulation studies

Simulation studies allow us to do statistical experiments and validate our approaches.

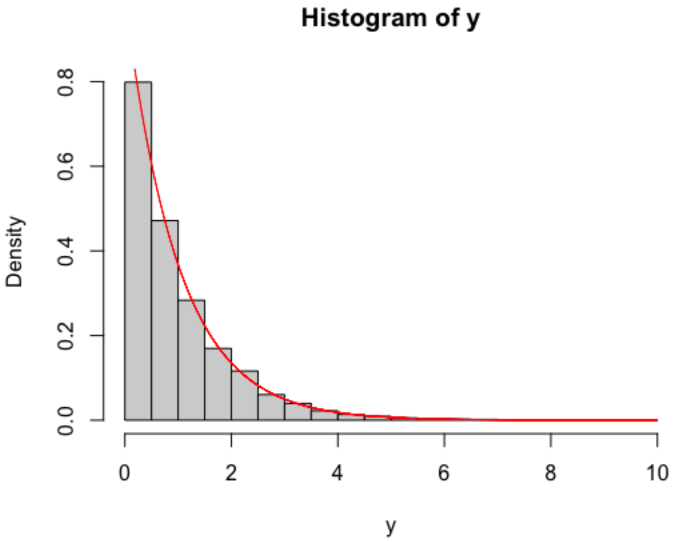
- Example 1: It is known that if $X \sim$ Unif$(0, 1)$, then $-\log(X) \sim$ Exp(1).

```
set.seed(442)                                #For reproducibility
x=runif(10000, min = 0, max = 1)             #Generate random samples from Unif(0,1)
y=-log(x)                                    #Transform variables
hist(y, probability = T)
s=seq(0, 10, length.out=10000)
lines(s, dexp(s, rate = 1), col="red")  #Compare it with the "theoretical" pdf of the exponential
```

Simulation studies

Simulation studies allow us to do 'statistical experiments' and validate our approaches.

- Example 1: It is known that if $X \sim$ Unif$(0, 1)$, then $-\log(X) \sim$ Exp(1).



Simulation studies

Simulation studies allow us to do ‘statistical experiments’ and validate our approaches.

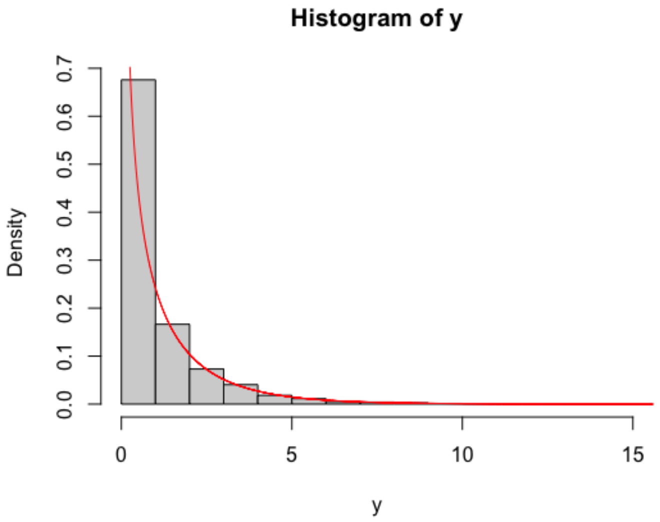
- Example 2: It is known that if $X \sim \mathcal{N}(0, 1^2)$, then $X^2 \sim \chi^2_1$.

```
set.seed(442)                       #For reproducibility
x=rnorm(10000, mean = 0, sd = 1)    #Generate random samples from N(0,1^2)
y=x^2                               #Transform the variable
hist(y, probability = T)
s=seq(0, 20, length.out=10000)
lines(s, dchisq(s, df = 1), col="red")  #Compare it with "theoretical" pdf of the Chi-square
```

Simulation studies

Simulation studies allow us to do ‘statistical experiments’ and validate our approaches.

- Example 2: It is known that if $X \sim \mathcal{N}(0, 1^2)$, then $X^2 \sim \chi^2_{df=1}$.



"Data generating process"

Consider a simple linear regression:

$$y_i = \beta_0 + x_i\beta_1 + \epsilon_i \quad \text{where} \quad \epsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$$

In this "model", there are a number of assumptions.

- The covariate x_i is fixed.
- $E(y_i) = \beta_0 + x_i\beta_1$. The expected value of y is linearly associated with x_i.
- The noise ϵ_i is independent to x_i.
- The noises ϵ_i are independent to each other.
- The noises are distributed normally.

“Data generating process”

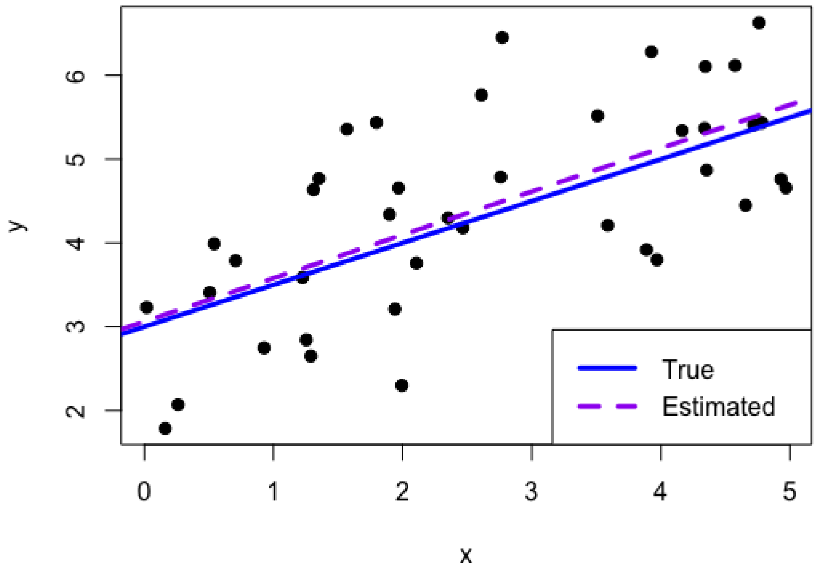
Consider a simple linear regression:

$$y_i = \beta_0 + x_i\beta_1 + \epsilon_i \quad \text{where} \quad \epsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$$

Based on this DGP, we can “generate” data using R

- Fix sample size n.
- Get x_i (or generate it and treat it as fixed)
- Fix $\beta_0, \beta_1, \sigma^2$ you want to simulate.
- Generate y_i.

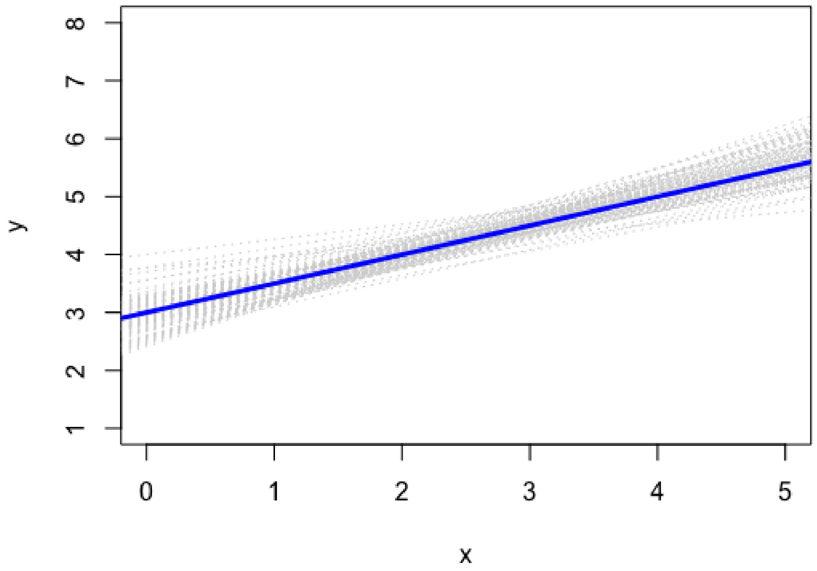
“Data generating process”



R codes

```
set.seed(442)
n=40
beta0=3; beta1=0.5; sigma=1
x=runif(n, min = 0, max = 5)
epsilon=rnorm(n, mean = 0, sd = sigma)
y=beta0+x*beta1+epsilon
plot(x, y, pch=19)
abline(a=beta0, b=beta1, col="blue", lwd=3)
abline(lm(y˜x),col="purple", lty=2, lwd=3)
legend("bottomright",
c("True","Estimated"), col=c("blue", "purple"),
lty=c(1,2),lwd=c(3,3))
```

“Data generating process”



R codes

```
set.seed(442)
n.sim=100
n=40
beta0=3; beta1=0.5; sigma=1
plot(c(0,5),c(1,8), type="n",xlab="x", ylab="y")
for (sim in 1:n.sim){
  epsilon=rnorm(n, mean = 0, sd = sigma)
  y=beta0+x*beta1+epsilon
  abline(lm(y~x), lty=3, col="lightgrey")
}
abline(a=beta0, b=beta1, col="blue", lwd=3)
```

Application to statistical inference (we'll revisit it later)

Consider a one-sample t-test that $H_0 : \mu = 0$ and $H_1 : \mu \neq 0$.

Power is defined by the probability of rejecting null hypothesis H_0 when the alternative hypothesis H_1 is actually true.

$$\text{Power} = P(\text{reject } H_0 | H_0 \text{ is false})$$

- There is a closed form solution for power in one-sample t-test, but in general we don't have a closed-form solution for power.
- Power depends on a few factors, including
  - sample size (N)
  - effect size (i.e. true μ)
  - Type 1 error threshold (i.e. α such as 0.05)
- Analog: We may
  1. "simulate" data (random dots) from our population (square)
  2. compute a p-value
  3. repeat 1 and 2 many times
  4. compute the proportion of rejected cases (p-value less than α).

Application to statistical inference (we'll revisit it later)

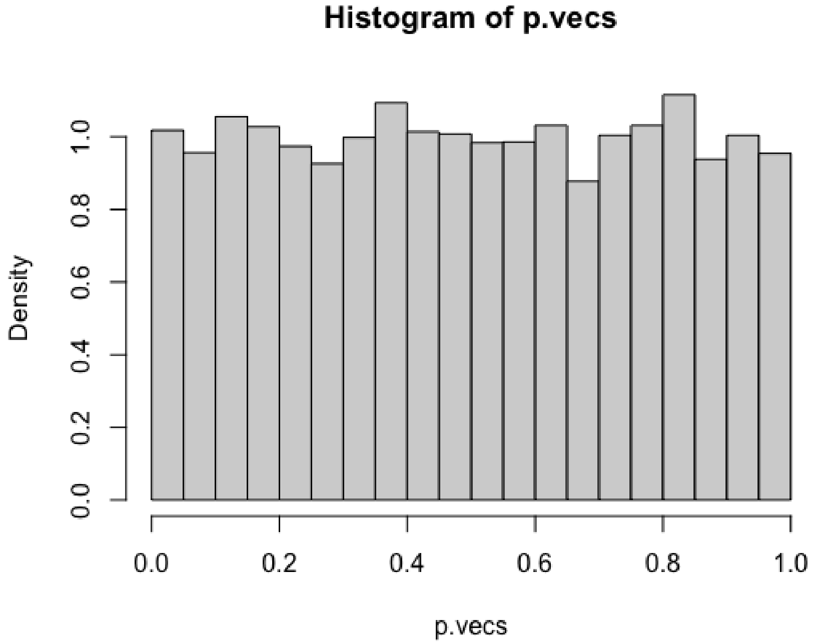
Q: What is the *distribution* of the p-value under the null hypothesis?

For this experiment, we will

1. generate data with $N = 50$ from $\mathcal{N}(0, 1^2)$
2. apply one-sample t test for $H_0 : \mu = 0$ and extract a p-value
    - Note: The data was generated under $\mu = 0$, so we already know that the null hypothesis is true.
3. repeat 1 and 2 these many times.

Application to statistical inference (we'll revisit it later)

For one-sample t test, the distribution of p-values under the null hypothesis seems to be distributed Unif(0, 1). Is it just by an accident or is there any theoretical result? If there is a theory, can we generalize to any tests?



R codes

```
set.seed(442)
n.sim = 10000
p.vecs = numeric(n.sim)
for (i in 1:n.sim){
data = rnorm(n = 50, mean = 0, sd = 1)
    p.vecs[i] = t.test(data)$p.value
    }
hist(p.vecs, prob=T)
```

DGP vs Assumed model

All models are wrong but some are useful.
(one of my favorite phrases)

- (Almost) all statistical estimations and inferences are based on the assumption that

  "your assumed model is truly the data generating model".

- Unfortunately, it is not usually the case. No one really knows how the observed data has been generated.
- Does it mean that what we learned in college is useless?

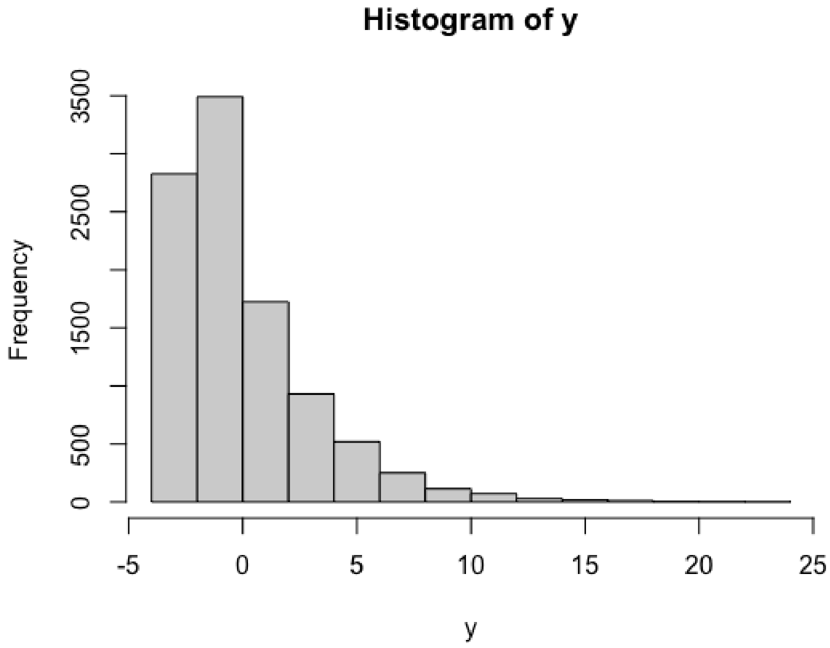
No, at least we hope that our assumed model is robust to misspecifications, and it is what we can evaluate through simulations.

Application to statistical inference (we'll revisit it later)

Simulation studies allow us to evaluate a model when its underlying assumption is misspecified.

One assumption in the linear regression is that the errors are normally distributed.
What happens if it is violated?

To evaluate this, we can generate data with non-normal errors that have zero means. One possible distribution is Exp$(\lambda) - 1/\lambda$, which is highly skewed.



R codes

```
set.seed(442)
lambda=1/3
x = rexp(10000, rate = lambda)
y = x-1/lambda
hist(y)
```