

Introduction to Causal Machine learning

Igor Asanov

November 19, 2023



This handout summarizes the lectures slides. Please note that the handout is not very useful if you do not attend the class. The handout is also not a substitution for the book. The course is built around Martin Huber book: “Causal analysis: Impact evaluation and Causal Machine Learning with applications in R”.

Homepage: <http://www.igorasanov.com/teaching.html>

Literature:

- ! Huber, M., 2023. Causal analysis: Impact evaluation and Causal Machine Learning with applications in R. MIT Press.
- Efron, B. and Hastie, T., 2021. Computer age statistical inference, student edition: algorithms, evidence, and data science (Vol. 6). Cambridge University Press.
- Tibshirani, R.J. and Efron, B., 1993. An introduction to the bootstrap. Monographs on statistics and applied probability, 57(1).
- Course: Causal Inference with Machine Learning by V. Chernozhukov.

Software:

We will use **R** for most of the exercises: Here is the link on the homepage of **R**.

In the class I use **RStudio** as a front end and I would recommend you to install it too - it greatly simplifies workflow in **R**.

Contents

1	Introduction to Machine Learning	3
1.1	Motivation	3
1.1.1	Motivation: Spam Example	3
1.2	In and Out of Sample Predictions	4
1.3	Prediction Study Design	6
1.4	Random Forest	8
1.5	Casuality and Machine Learning?	8
1.6	Exercises	10
2	Recap: Statistics	11
2.1	Sample Variation	11
2.2	Hypothesis Testing	16
3	Casual Inference	17
3.1	Fundamental Problem of Casual Inference	17
3.2	Statistical solution	18
3.3	Exercises	25
4	Modern High Dimensional Linear Regression	26
5	Effect Heterogeneity	26
6	Identification under Selection on Observables	26

1 Introduction to Machine Learning

1.1 Motivation

- Econometrics, inferential statistics
 - $\bar{X} = 239$
 - Reject the the null-hypothesis that $\mu = 0$ and calculating the confidence intervals based on our estimations.
 - β is good estimator of relation between x and y
- However, what we often want is to predict certain event.
 - We want to predict that the letter is spam
 - What will person look next on Youtube
 - There will be a flu outbreak soon
 - Will be this teacher good based on his CV

1.1.1 Motivation: Spam Example

Dear Valentine,
you won this year.
Your money is waiting for you.
Just send us your bank account.
Best,
Valencia.
Dear|| Valentine||,
you|| won|| this|| year||.
Your|| money || is || waiting || for || you || .
Just|| send|| us|| your|| bank || account||.
Best||,
Valencia||.

Frequency of ||your||

$$f = 2/20 = 0.1$$

```
library(kernlab)

data(spam)

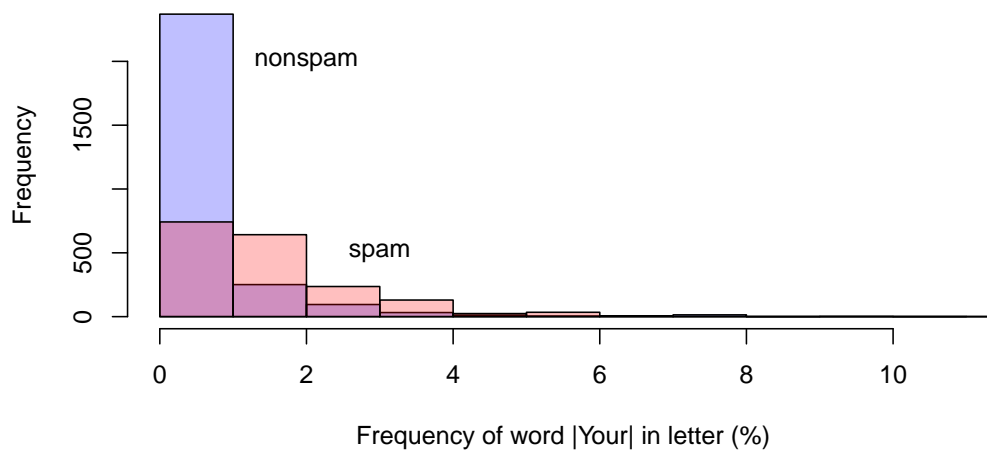
require(stargazer)

stargazer(head(spam[, c(1, 2, 21, 24, 58)]), summary = FALSE)
```

Table 1

	make	address	your	money	type
1	0	0.640	0.960	0	spam
2	0.210	0.280	1.590	0.430	spam
3	0.060	0	0.510	0.060	spam
4	0	0	0.310	0	spam
5	0	0	0.310	0	spam
6	0	0	0	0	spam

Histogram of word |your| based on type



```
prediction <- ifelse(spam[, "your"] > 1, "spam", "nonspam")
t <- table(prediction, spam[, "type"])/length(prediction)
require(xtable)
xtable(t * 100)
```

	nonspam	spam
nonspam	51.51	16.13
spam	9.08	23.28

Accuracy-correct predictions

$\approx 52+23=75\%$

1.2 In and Out of Sample Predictions

- In sample errors – the error rate on the same data set you used.
- Out of Sample error – the error rate you get on new data set.

- Key points:
 - We care about out of sample prediction
 - In sample error < out of sample error
 - In sample overfitting can occur
- Let's take a small sample from our data set.

```
set.seed(239)

require(dplyr)

s100 <- sample_n(spam, 100)

prediction <- ifelse(s100[, "your"] > 1, "spam", "nonspam")

t <- table(prediction, s100[, "type"])/length(prediction)
require(xtable)

xtable(t * 100)
```

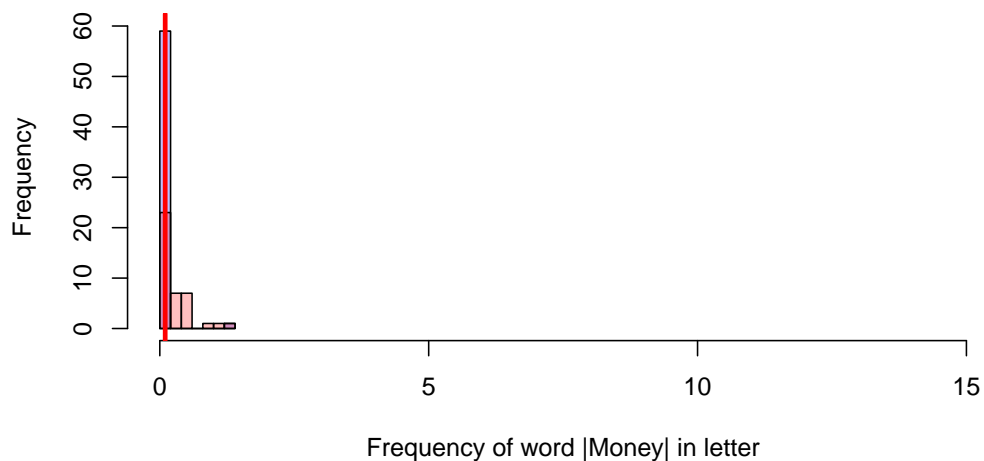
	nospam	spam
nospam	48.00	14.00
spam	12.00	26.00

Accuracy-correct predictions

$\approx 48 + 26 = 74\%$

- What about other variables?

Histogram of word |Money| based on type



```
prediction<-ifelse(s100[,"your"]>1,"spam",
                 ifelse(s100[,"money"]>0,
                        "spam","nonspam"))

t<-table(prediction, s100[,"type"])/length(prediction)
require(xtable)

xtable(t*100)
```

	nonspam	spam
nonspam	48.00	9.00
spam	12.00	31.00

Accuracy-correct predictions

$\approx 48 + 31 = 79\%$

- But what about out applying this rule to the whole dataset?

```
prediction<-ifelse(spam[,"your"]>1,"spam",
                 ifelse(spam[,"money"]>0.2, "spam","nonspam"))

t<-table(prediction, spam[,"type"])/length(prediction)
require(xtable)

xtable(t*100)
```

	nonspam	spam
nonspam	50.88	13.98
spam	9.72	25.43

Accuracy-correct predictions

$\approx 51 + 25 = 76\%$

1.3 Prediction Study Design

- Define error rate e.g. Accuracy
- Split data into:
 - Training, Test, Validation data set
- Pick features (variables) and model on training set
 - Use Cross-validation
- Apply to test data set (once)
- Apply on validation data set
- Sensitivity – $TP / (TP+FN)$

Table 2: Error Types

		Predicted	
		Spam	Not spam
Actual	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

- Specificity – $TN / (FP+TN)$
- Positive Predictive Value – $TP / (TP+FP)$
- Negative Predictive Value – $TN / (FN+TN)$
- Accuracy – $(TP+TN) / (TP+FN+TN+FP)$
- We estimate the model on the training set e.g. random 60% of sample
- We test the model on test data set and refine it e.g. calculate sensitivity of the model
- We apply model model on validation data set.

Can we do this multiple times?

1. Use training set
 2. Split it again into training/test set
 3. Build a model on training set
 4. Evaluate on test set
 5. Repeat from 1 to 4 n times.
 6. Estimate average errors e.g. Mean Squared Error = $\frac{\sum (y_i - \hat{y}_i)^2}{n}$
- It allows:
 - Pick up features(variables)
 - Select model
 - Parameters of Prediction function

How to split training/test set?

- Random sampling without replacement.
 - Alternative: Random sampling without replacement, bootstrap. (needs correction)
- K -fold cross validation.
 - Devide data set in K parts. (Needs to carefully choose K)

How to build a model and select variables?

- OLS
- Lasso Regression
- Trees
- Random Forest
- Neural networks
- ...

1.4 Random Forest

- Recall Spam example
 - We can base our prediction on word `||your||`
 - Then we can use word `||money||`
 - ...
 - We build a decision tree till the groups are very small

```
prediction <- ifelse(s100[, "your"] > 1, "spam", "nonspam")
```

```
prediction<-ifelse(spam[, "your"]>1, "spam",  
  
ifelse(spam[, "money">0.2, "spam", "nonspam"))
```

- We find the variable that best separate our outcome
- Divide the data on into two groups based on this variable
- Within each split, find the variable that separates outcomes the best
- Continue until the groups are too small

Why not to do this multiple times?

- Bootstrap sample
- At each split randomly select n variables
- Grow multiple trees and vote
- Pros: Extremely high accuracy
- Cons: Computationally intensive, interpretability, overfit

1.5 Casuality and Machine Learning?

- ML focus on prediction out of “black-box”
- → Causality is not necessary there!
 - Prediction of crime recidivism: Ethnicity or Income? see “Weapon of math destruction” by Cathy O’Neil
 - Google flu prediction algorithm: Sensitivity to words, time variance. See ”The Parable of Google Flu ...”

I run out of examples. Let’s ask ChatGPT

Bring examples when machine learning fails to establish casual relationships with ciation, please

Certainly, here are examples illustrating when machine learning fails to establish causal relationships, along with citations:

- 1. Ice Cream Sales and Drowning Incidents:**
 - Example: Machine learning might reveal a correlation between the increase in ice cream sales and the number of drowning incidents. However, it cannot establish a causal relationship, as both events are influenced by the temperature during summertime.
 - Citation: Phillips, N. (1985). The relationship between drowning and the sale of ice cream and iced drinks in New Zealand. *The Medical Journal of Australia*, 143(11-12), 610-612.
- 2. Crime Rates and Police Car Numbers:**
 - Example: Machine learning might find a correlation between the number of police cars patrolling an area and crime rates. However, it cannot confirm that adding more police cars directly reduces crime; other factors, such as community engagement, may play a role.
 - Citation: Klick, J., & Tabarrok, A. (2005). Using terror alert levels to estimate the effect of police on crime. *Journal of Law and Economics*, 48(1), 267-279.
- 3. Obesity and Diet Soda Consumption:**
 - Example: Machine learning could reveal a correlation between diet soda consumption and obesity. However, it cannot establish causality, as people who are already obese might consume diet soda to manage their weight.
 - Citation: Fowler, S. P., Williams, K., Resendez, R. G., Hunt, K. J., Hazuda, H. P., & Stern, M. P. (2008). Fueling the obesity epidemic? Artificially sweetened beverage use and long-term weight gain. *Obesity*, 16(8), 1894-1900.
- 4. Educational Attainment and Income:**
 - Example: Machine learning might discover a strong correlation between educational attainment and higher income. However, it cannot determine if more education directly causes higher income or if factors like networking and job opportunities are equally significant.
 - Citation: Goldin, C., & Katz, L. F. (2008). *The Race between Education and Technology*. Harvard University Press.
- 5. Social Media Usage and Loneliness:**
 - Example: Machine learning might uncover a connection between heavy social media use and feelings of loneliness. However, it cannot establish whether social media use causes loneliness, or if lonely individuals are more drawn to social media for connection.
 - Citation: Primack, B. A., Shensa, A., Sidani, J. E., Whaite, E. O., Lin, L. Y., Rosen, D., ... & Miller, E. (2017). Social media use and perceived social isolation among young adults in the U.S. *PLoS ONE*, 12(8), e0182145.

Regenerate

- We intuitively feel that
 - Machine learning predictions in the cases above are kind of funny
 - Machine learning predictions may have dangerous (policy) implications e.g. less cars to the police?
- To understand how to avoid this, we will have to understand causal relationship.

1.6 Exercises

1. **In and Out of Sample Predictions I:** What is in sample errors?
2. **In and Out of Sample Predictions II:** What is out of Sample error?
3. **Prediction study design I:** Describe simple prediction study design
4. **Prediction study design II:** Describe Cross-Validation
5. **Install R and Rstudio.**

Download R from <https://www.r-project.org/> and install it.

Download RStudio <https://www.rstudio.com/> and install it.

6. **Simple linear model: Crime rate.**

```
# Install and call the package 'Ecdat'.
install.packages("Ecdat")
require(Ecdat)

# Use the data 'Crime'
data(Crime)
`?`(Crime)

# Plot the crime rate with respect to number of policemen
# per capita.
plot(crmrte ~ polpc, ylab = "Crime Rate", xlab = "Number of Policemen per capita",
     pch = 18, data = Crime)

# Estimate simple linear relation between crime rate and
# number of policemen per capita.
lmcrime <- lm(crmrte ~ polpc, data = Crime)
summary(lmcrime)

# Plot the crime rate with respect to number of policemen
# per capita.
plot(crmrte ~ polpc, ylab = "Crime Rate", xlab = "Number of Policemen per capita",
     pch = 18, data = Crime)
abline(lmcrime, col = "red")
```

How do you interpret this result? What are the policy implications?

7. **Let's use Machine Learning to predict Crime rate**

```
require(Ecdat)

# Install and call the package 'randomForest' and
# 'datasets'
install.packages("randomForest")
install.packages("datasets")
library(randomForest)
library(datasets)

set.seed(239) # set seed to reproduce results

ind <- sample(2, nrow(Crime), replace = TRUE, prob = c(0.5, 0.5))
train <- Crime[ind == 1, ] # training data set
test <- Crime[ind == 2, ] # testing data set
nrow(train)
nrow(test)

rf1cr <- randomForest(crmrte ~ ., ntree = 50, data = train, importance = TRUE)
rf1cr
```

```

varImpPlot(rf1cr, main = "Variable Importance")

lm1 <- lm(crmrte ~ density + polpc + smsa + prbarr + pctmin +
          wfed, data = train)

summary(lm1)

# Out of Sample

predictedcrmrate <- predict(lm1, test)

compare <- cbind(test, predictedcrmrate)

plot(compare$crmrate, compare$predictedcrmrate, xlab = "Actual Crime Rate",
      ylab = "Predicated Crime Rate")
hist(compare$crmrate - compare$predictedcrmrate)
sum((compare$crmrate - compare$predictedcrmrate)^2)/length(compare$crmrate)

```

8. Your Turn

- Call the package `Ecdat` and use the data `DoctorAUS`. What is the data set about?
- What is the influence of number of doctor visits (`doctorco`) on the number of illness in past 2 weeks (`illness`)? Visualize this relation.
- Estimate a simple linear relation between the number of doctor visits and the number of illness in past 2 weeks. Visualize regressions output.
- Make training and testing dataset based on `DoctorAUS`.
- Use Random Forest `randomForest` from package `randomForest` to build regression predicting the number of doctor visits on your training set.
- “Test” your model on the testing set. Plot Actual number of doctor visits against predicted one.
- Calculate mean square error (MSE) for your testing set.

2 Recap: Statistics

2.1 Sample Variation

- Population – the set of all entities which could theoretically be observed.
- Sample – a part of the population that we observe.
- Typically we observe A but want to know something about B .

Example

We know working hours Y of person x in year t , but we want to know it for the same (other) person next year (time) $t + 1$.

- We can do it, use A to *infer* something about B , if we assume that they come from the same population.
- We can look at the first observation and make conclusion.
- Perhaps, it is not best idea.
- It is better to draw a random sample from the same population.
 - ⇒ Values are identically and independently distributed .i.i.d.

Law of Large Numbers

If Y_1, \dots, Y_n i.i.d. and $0 < \sigma_Y^2 < \infty$ and , then \bar{Y} is a consistent estimator of Y .

Central Limit Theorem

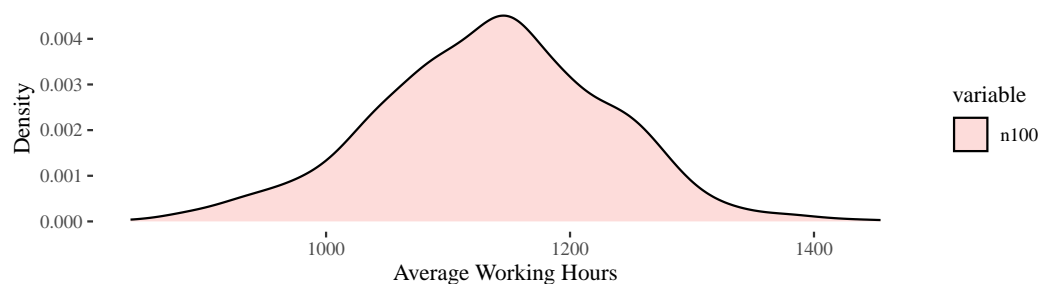
If Y_1, \dots, Y_n i.i.d. and $0 < \sigma_Y^2 < \infty$ and n is large, then the distribution of \bar{Y} approximates a normal distribution

```
library(Ecdat)
data(Workinghours)
# Make sample of 'hours'
s1 <- sample(Workinghours$hours, 100, replace = TRUE)
str(s1)

## int [1:100] 735 0 0 1312 1928 1348 2783 920 911 0 ...

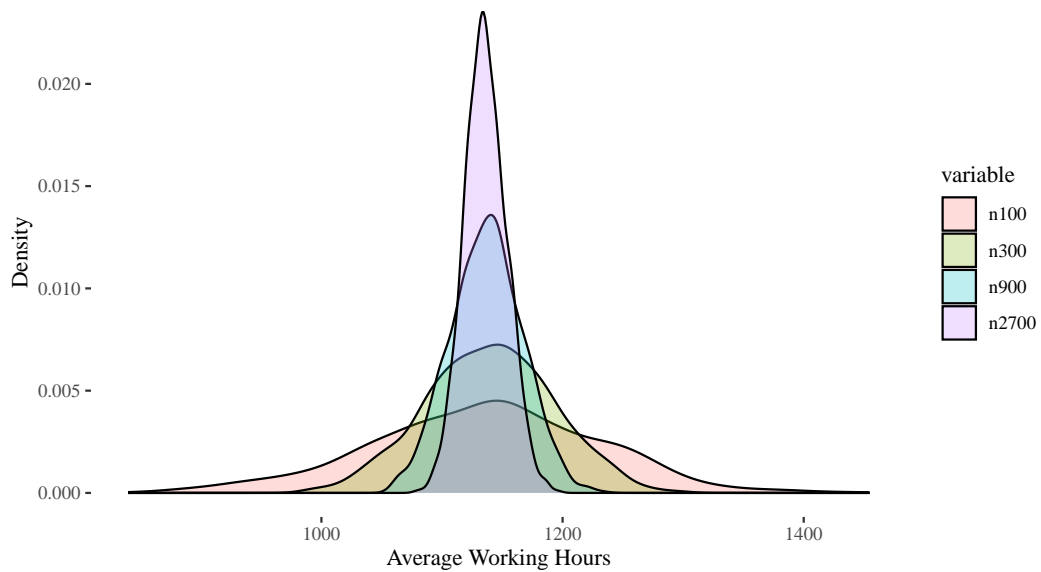
# Make 1000 samples of 'hours' of size 100
n100 <- replicate(1000, mean(sample(Workinghours$hours, 100,
  replace = TRUE)))
```

```
require(ggplot2)
require(ggthemes)
require(reshape2)
ggplot(melt(data.frame(n100)), aes(x = value, fill = variable)) +
  geom_density(alpha = 0.25) + theme_tufte() + xlab("Average Working Hours") +
  ylab("Density")
```



```
n300 <- replicate(1000, mean(sample(Workinghours$hours, 300,
  replace = TRUE)))
n900 <- replicate(1000, mean(sample(Workinghours$hours, 900,
  replace = TRUE)))
n2700 <- replicate(1000, mean(sample(Workinghours$hours, 2700,
  replace = TRUE)))
```

```
ggplot(melt(data.frame(n100, n300, n900, n2700)), aes(x = value,
  fill = variable)) + geom_density(alpha = 0.25) + theme_tufte() +
  xlab("Average Working Hours") + ylab("Density")
```



Standard Deviation

$$sd = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$$

```
s100 <- sample(Workinghours$hours, 100, replace = TRUE)
s300 <- sample(Workinghours$hours, 300, replace = TRUE)
s900 <- sample(Workinghours$hours, 900, replace = TRUE)
s2700 <- sample(Workinghours$hours, 2700, replace = TRUE)
sqrt(sum((s100 - mean(s100))^2)/99)
```

```
## [1] 841.4578
```

```
sd(s100)
```

```
## [1] 841.4578
```

```
sd(s300)
```

```
## [1] 888.9162
```

```
sd(s900)
```

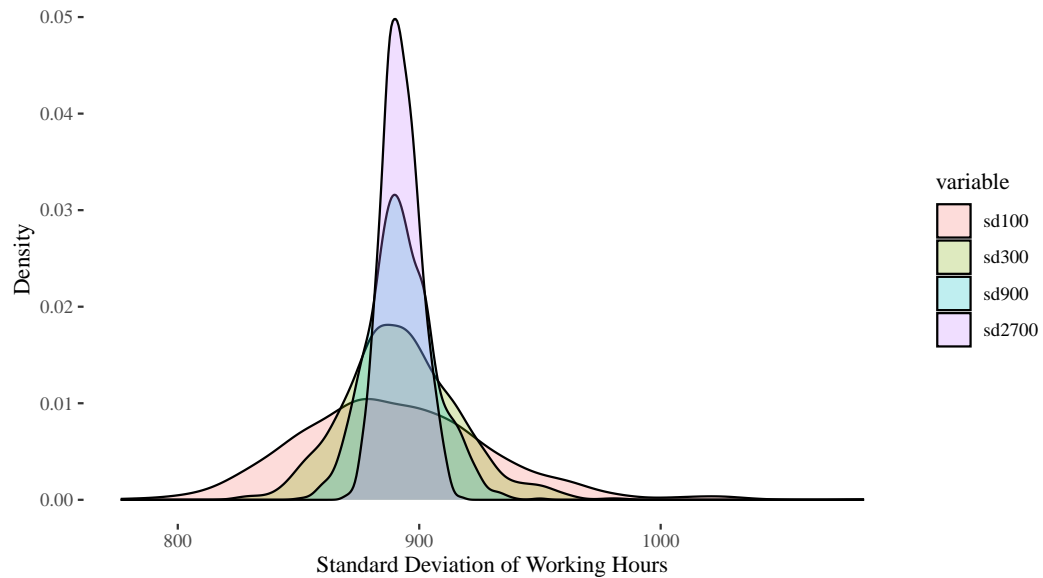
```
## [1] 874.7947
```

```
sd(s2700)
```

```
## [1] 896.1815
```

```
sd100 <- replicate(1000, sd(sample(Workinghours$hours, 100, replace = TRUE)))
sd300 <- replicate(1000, sd(sample(Workinghours$hours, 300, replace = TRUE)))
sd900 <- replicate(1000, sd(sample(Workinghours$hours, 900, replace = TRUE)))
sd2700 <- replicate(1000, sd(sample(Workinghours$hours, 2700,
replace = TRUE)))
```

```
ggplot(melt(data.frame(sd100, sd300, sd900, sd2700)), aes(x = value,
  fill = variable)) + geom_density(alpha = 0.25) + theme_tufte() +
  xlab("Standard Deviation of Working Hours") + ylab("Density")
```



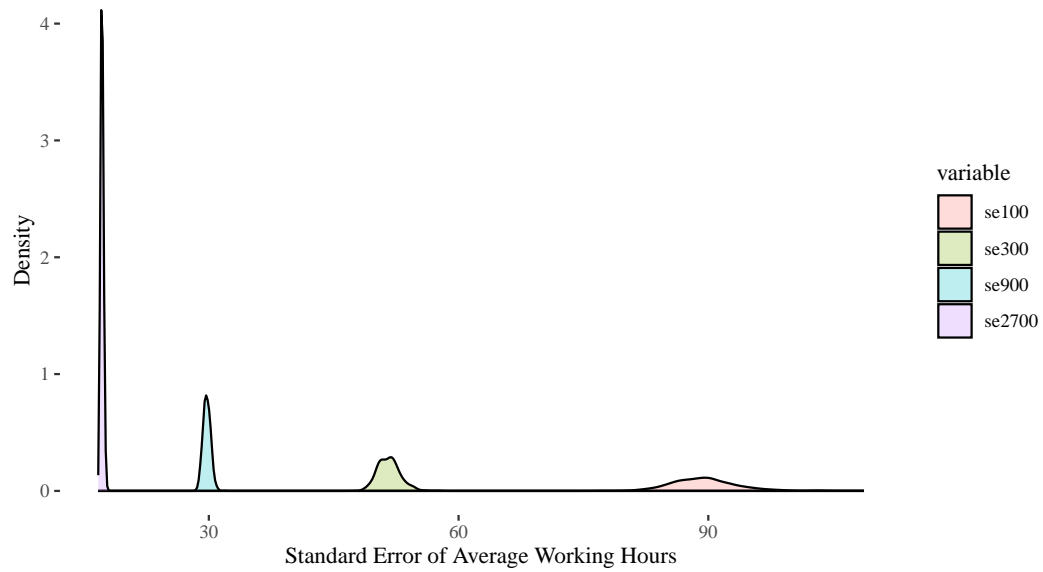
Standard Error of the Mean

$$se = \frac{sd}{\sqrt{n-1}}$$

```
sd(s100)/sqrt(length(s100) - 1)
## [1] 84.56969
sd(s300)/sqrt(length(s300) - 1)
## [1] 51.40735
sd(s900)/sqrt(length(s900) - 1)
## [1] 29.17604
sd(s2700)/sqrt(length(s2700) - 1)
## [1] 17.25022
```

```
se100 <- replicate(1000, sd(sample(Workinghours$hours, 100, replace = TRUE))/sqrt(99))
se300 <- replicate(1000, sd(sample(Workinghours$hours, 300, replace = TRUE))/sqrt(299))
se900 <- replicate(1000, sd(sample(Workinghours$hours, 900, replace = TRUE))/sqrt(899))
se2700 <- replicate(1000, sd(sample(Workinghours$hours, 2700,
  replace = TRUE))/sqrt(2699))
```

```
ggplot(melt(data.frame(se100, se300, se900, se2700)), aes(x = value,
  fill = variable)) + geom_density(alpha = 0.25) + theme_tufte() +
  xlab("Standard Error of Average Working Hours") + ylab("Density")
```



Central Limit Theorem

If Y_1, \dots, Y_n i.i.d. and $0 < \sigma_Y^2 < \infty$ and n is large, then the distribution of \bar{Y} approximates a normal distribution

Confidence intervals

Upper 95% limit = $\bar{x} + 1.96 \cdot se$

Lower 95% limit = $\bar{x} - 1.96 \cdot se$

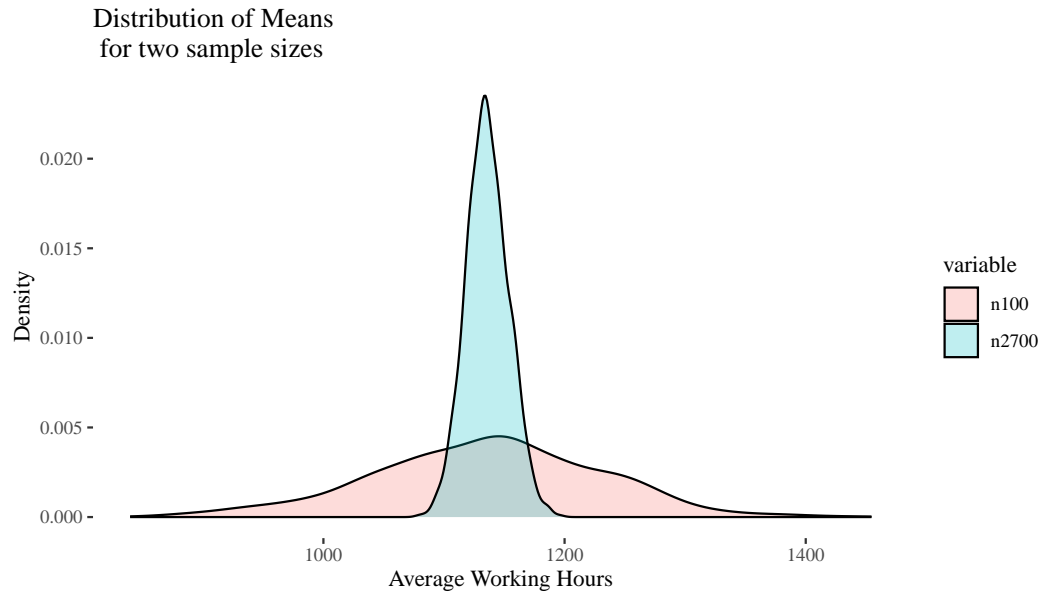
```

mean(s100) - 1.96 * sd(s100)/sqrt(length(s100) - 1)
## [1] 1251.433
mean(s100) + 1.96 * sd(s100)/sqrt(length(s100) - 1)
## [1] 1582.947
mean(s2700) - 1.96 * sd(s2700)/sqrt(length(s2700) - 1)
## [1] 1084.962
mean(s2700) + 1.96 * sd(s2700)/sqrt(length(s2700) - 1)
## [1] 1152.583

```


Table 3: Error Types

		Null Hypothesis, H_0	
		Valid	Invalid
Judgement of H_0	Reject	False Positive (Type Error I)	Correct Inference
	Fail to Reject	Correct Inference	False Negative (Type Error II)



Note: The exact confidence interval is slightly different since we have to correct for the fact that we sample from larger sample.

2.2 Hypothesis Testing

1. Formulate Null-Hypothesis e.g. means is zero. $H_0 : E(Y) = \mu_{Y,0}$
2. Formulate alternative hypothesis:
 - $H_1 : E(Y) \neq \mu_{Y,0}$. (Two-sided test)
 - $H_1 : E(Y) > \mu_{Y,0}$. (One-sided test)
 - $H_1 : E(Y) < \mu_{Y,0}$. (One-sided test)

What can happen then?

- α , significance level – predefined probability of rejecting H_0 despite it being true, typically 5%.
- k – predefined probability of failing to reject H_0 when it is false.
- $(1-k)$, power – predefined probability of rejecting H_0 when it is false, typically 80%

P-Value

- **Level of significance** – Predefined probability of rejecting the null hypothesis, despite it being true.
- **P-value** – Probability of drawing a sample that is at least as averse to the null hypothesis as our data given that the null hypothesis is true.



Figure 1: The casual effect of treatment on outcomes.

Example

- H_0 : Coin is fair, $q_T = q_H = 0.5$
- H_1 : Heads are more likely, $q_H > 0.5$
- We observe 3 heads out of 4 throws.
- $p = \frac{4!}{3! \cdot 1!} \cdot (0.5)^3 \cdot 0.5 = 0.25$
- and more averse case $p = \frac{4!}{4! \cdot 0!} \cdot (0.5)^4 = 0.0625$
- p -value= 0.3125

```

binom.test(3, 4, p = 0.5, alternative = c("greater"))

##
## Exact binomial test
##
## data: 3 and 4
## number of successes = 3, number of trials = 4, p-value = 0.3125
## alternative hypothesis: true probability of success is greater than 0.5
## 95 percent confidence interval:
##  0.2486046 1.0000000
## sample estimates:
## probability of success
##                0.75
  
```

3 Casual Inference

3.1 Fundamental Problem of Casual Inference

Road Not Taken

“Two roads diverged in a yellow wood,
 And sorry **I could not travel both**
 And be one traveler, long I stood
 And looked down one as far as I could
 To where it bent in the undergrowth;

...

Two roads diverged in a wood, and
 I took the one less traveled by,
 And that has made all the difference.”

Robert Frost, 1920

- T - treatment e.g. educational program
- Y - outcome e.g. reading level
- $T_i = 1$ – if person i is under treatment
- $T_i = 0$ – if person i is not under treatment

- $Y_i(1)$ – (potential) outcome if person i is under treatment
- $Y_i(0)$ – (potential) outcome if person i is not under treatment

$$Y_i = Y_i(1)T_i + Y_i(0)(1 - T_i) \quad (1)$$

$$Y_i = Y_i(0) + \underbrace{(Y_i(1) - Y_i(0))}_{\text{Casual Effect}} T_i \quad (2)$$

Holland, (1986) Statistics and Causal Inference

Casual effect of $T_i = 1$ on unit i (relative to $T_i = 0$):

$$Y_i(1) - Y_i(0)$$

The problem: It is impossible to observe the value of $Y_i(1)$ and $Y_i(0)$ on the same unit, therefore, it is impossible to observe the effect of $T_i = 1$ on unit i .

Any solutions?

- “Scientific solution” - requires the unbelievably strong assumption of homogeneity of units, e.g., **identical** twins that are put under treatment and no treatment condition, and the **only** difference in outcomes between them is induced by treatment.
- Statistical solution – we can hope to estimate **averages** and the distribution of $Y(T)$ at the population level to compute **Average Treatment Effect (ATE)**:

$$\Delta = E[Y_i(1) - Y_i(0)] = E[Y_i(1)] - E[Y_i(0)] \quad (3)$$

3.2 Statistical solution

Knowing Average Treatment Effect (ATE) can be extremely useful as we can learn that treatment, on average, improved outcomes:

- Visits to doctors reduce the number of sick people in the population (unfortunately, not necessarily all will recover, but still good for society).
- Educational programs improve the student’s reading ability (not necessarily all will read equally, but still good for society).
- Training increased income on average (not necessary for all, but still welfare might improve).

Now, we can think about how to calculate the average treatment effect:

1. Observational data. We can use observational data on the population or use samples to infer the average treatment effect.
 2. Experimental data. We can treat randomly selected people and compare them to non-treated to infer the average treatment effect.
- Let’s think about calculating the average treatment effect if we have the population data at hand or sample.
 - For instance, we know if the person is sick and if this patient is treated by a doctor in the population.

- We can naïvely calculate the average of the outcome conditional on the **observed** “treatment” status:

$$E[Y_i|T = 1] = E[Y_i(1)|T = 1] \quad (4)$$

$$E[Y_i|T = 0] = E[Y_i(0)|T = 0] \quad (5)$$

- and subtract 4 from 5 to get **Average Predictive Effect (APE)**:

$$\pi = E[Y_i|T = 1] - E[Y_i|T = 0] = E[Y_i(1)|T = 1] - E[Y_i(0)|T = 0] \quad (6)$$

```
library(Ecdat)

data(DoctorAUS)
attach(DoctorAUS)

T <- ifelse(doctorco > 0, 1, 0) # 1 if treated by a doctor or specialist in the past 2 weeks
Y <- illness #number of illness in past 2 weeks

mean(Y[T == 1]) # average number of illnesses conditional on observed treatment by a doctor or specialist

[1] 2.125834

mean(Y[T == 0]) # average number of illnesses conditional on observed no treatment by the doctor

[1] 1.256218

mean(Y[T == 1]) - mean(Y[T == 0]) #The average predictive effect of doctor visits

[1] 0.8696158
```

→ The average predictive effect is positive ($\pi \approx 0.87 > 0$): Doctor or specialist visit is associated with more illness.

```
library(Ecdat)
data("VietNamI")
attach(VietNamI)

T <- ifelse(educ > 0, 1, 0) # 1 if a person got at least two years of higher education
Y <- illness #number of illnesses experienced in past 12 months

mean(Y[T == 1])

[1] 0.6150017

mean(Y[T == 0])

[1] 0.7094727

mean(Y[T == 1]) - mean(Y[T == 0])

[1] -0.09447091
```

→ The average predictive effect is negative ($\pi \approx -0.09 < 0$): Higher education is associated with less illness.

- Why do we observe it?
- We calculate this:

$$\pi = E[Y_i(1)|T = 1] - E[Y_i(0)|T = 0] \quad (7)$$

- Yet,

$$E[Y_i(1)|\mathbf{T}=\mathbf{1}] \neq E[Y_i(1)] \quad (8)$$

$$E[Y_i(0)|\mathbf{T}=\mathbf{0}] \neq E[Y_i(0)] \quad (9)$$

- Now, I can decide to assign people to the treatment and control groups randomly.
- Random assignment guarantees that treatment T is independent from potential outcome Y . $T \perp\!\!\!\perp Y$
- $T \perp\!\!\!\perp Y \Rightarrow E[Y_i(1)|\mathbf{T}=\mathbf{1}] = E[Y_i(1)] \Rightarrow$

$$\Delta = E[Y_i(1)] - E[Y_i(0)] = E[Y_i|T_i = 1] - E[Y_i|T_i = 0] \quad (10)$$

$$\Delta = \pi \quad (11)$$

STAR Experiment: Does reducing class size improve educational outcomes?

Over 7,000 students in 79 schools were randomly assigned to one of three interventions:

- regular class (22 to 25 students per teacher)
- small class (13 to 17 students per teacher),
- regular-with-aide class (22 to 25 students with a full-time teacher's aide)
- For simplicity, we will focus on regular classes and small class intervention

STAR Experiment: Does reducing class size improve educational outcomes?

0 regular class (22 to 25 students per teacher)

1 small class (13 to 17 students per teacher),

- Let's say we are interested in the reading abilities of students in grade 3
- Specifically, we want to know if reducing class improves reading.
- But before, let's look at how basic demographic characteristics of students are distributed across groups.

Table 4: Control Group: Regular Class

Statistic	N	Mean	St. Dev.
Girls	4,085	0.479	0.500
Caucasian	4,075	0.645	0.479
African-American	4,075	0.351	0.477
Other Ethnicity	4,075	0.004	0.064

Table 5: Treatment Group: Small Class

Statistic	N	Mean	St. Dev.
Girls	1,915	0.500	0.500
Caucasian	1,913	0.685	0.465
African-American	1,913	0.308	0.462
Other Ethnicity	1,913	0.006	0.079

```
library(AER)
data(STAR)

STAR3 <- subset(STAR, star3 != c(" regular+aide") & !is.na(STAR$read3) &
  !is.na(STAR$star3))

T <- ifelse(STAR3$star3 == "small", 1, 0) # 1 if student is in the small class treatment group

Y <- STAR3$read3 #total reading scaled score in 3rd grade.

mean(Y[T == 1])
```

[1] 621.0757

```
mean(Y[T == 0])
```

[1] 612.7576

```
mean(Y[T == 1]) - mean(Y[T == 0])
```

[1] 8.318068

Now, we can assess significance of this difference with t-test:

```
t.test(Y ~ T)

##
## Welch Two Sample t-test
##
## data: Y by T
## t = -7.7229, df = 3621.6, p-value = 1.461e-14
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
## 95 percent confidence interval:
## -10.429778 -6.206358
## sample estimates:
```

```
## mean in group 0 mean in group 1
##      612.7576      621.0757
```

Assumptions of Regression (OLS):

1. Strict exogeneity, $E(u_i|T_i = t) = 0$
2. (T_i, Y_i) are identically independently distributed (i.i.d.)
3. Large outliers are rare
4. $var(u|T = t)$ is constant
5. u is normally distributed, $u \sim \mathcal{N}(0, \sigma^2)$

Assumptions of Regression (OLS):

1. Strict exogeneity, $E(u_i|T_i = t) = 0$ ✓
 2. (T_i, Y_i) are identically independently distributed (i.i.d.)
 3. Large outliers are rare
 4. $var(u|T = t)$ is constant
 5. u is normally distributed, $u \sim \mathcal{N}(0, \sigma^2)$
- ✓ by randomization, n is large

Similarly with regression:

```
summary(lm(Y ~ T))

##
## Call:
## lm(formula = Y ~ T)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -113.758  -27.758   -1.076    26.242   162.242
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  612.7576     0.6004  1020.584 < 2e-16 ***
## T              8.3181     1.0628    7.827 5.87e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 38.37 on 5998 degrees of freedom
## Multiple R-squared:  0.01011, Adjusted R-squared:  0.009945
## F-statistic: 61.26 on 1 and 5998 DF, p-value: 5.869e-15
```

- Let's recall that we know the characteristics (W) of students e.g. gender. Can we include them?

Assumptions:

1. Strict exogeneity, $E(u_i|T_i = t) = 0$

2. (T_i, Y_i) are identically independently distributed (i.i.d.)
3. Large outliers are rare
4. $var(u|T = t)$ is constant
5. u is normally distributed, $u \sim \mathcal{N}(0, \sigma^2)$
6. No perfect multicollinearity ✓

✓ by randomization. Moreover, $T \perp\!\!\!\perp Y, W \Rightarrow$

$$E[Y_i(T)|W] = E[Y_i|T_i, W], T_i \in \{0, 1\} \Rightarrow$$

$$\Delta(W) = \pi(W)$$

Table 6: Outcomes of STAR project in Grade 3

	<i>Dependent variable:</i>			
	Y - Reading score			
	(1)	(2)	(3)	(4)
T - Small Class	8.318*** (1.063)	8.156*** (1.058)	7.982*** (1.062)	7.477*** (1.031)
Girl		7.567*** (0.987)	7.525*** (0.992)	7.598*** (0.962)
Teachers Experience (Years)			-0.013 (0.057)	0.099* (0.056)
Sub-urban School				20.294*** (1.434)
Rural School				24.702*** (1.287)
Urban				20.165*** (2.056)
Constant	612.758*** (0.600)	609.134*** (0.762)	609.500*** (1.113)	589.820*** (1.480)
Observations	6,000	6,000	5,951	5,951
R ²	0.010	0.020	0.019	0.078
Adjusted R ²	0.010	0.019	0.019	0.077

Note:

*p<0.1; **p<0.05; ***p<0.01

I. Shall we include these characteristics (W) – covariates?

- Let's note that
 - Standard errors in model 4 are slightly lower than in model 1.
 - R^2 or Adjusted R^2 is larger in the model 4 than in model 1.
- Covariates in randomized experiments is “denoising”, improving the precision of estimation of Average Treatment Effect.
- Yet, which one to include and which not, especially if I have a lot more than people e.g. $W \gg i$?

→ Here, Machine Learning tools will help us. We will study this in the first part of “Modern High Dimensional Linear Regression”

II. Treatment Heterogeneity.

- Let's note that
 - Treatment (small class) increases the reading score by 8 points on average.
 - * Average treatment effect (ATE) ≈ 8 .
 - It is comparable to the gender gap in reading: Girls get 8 points more than boys.
 - * Average predictive effect (APE) ≈ 8 .
- Nice, could it be that the treatment helps to close the reading gender gap by being particularly helpful to boys?
- We can check it by estimating the interaction effect of treatment on boys (B):
- $Y = \beta_0 + \beta_T T + \beta_B B + \beta_{T \cdot B} T \cdot B + u_i$
- But could it be that while it is particularly useful for boys in rural areas
- $Y = \beta_0 + \beta_T T + \beta_B B + \beta_R R + \beta_{T \cdot B} T \cdot B + \beta_{R \cdot B} R \cdot B + \beta_{T \cdot B \cdot R} T \cdot B \cdot R + u_i$
- What about the income level of the parents, and the education of the mother, low achieving girls in grade 3?
- ...
- We can continue this exercise till infinity (or sample size).
- Increasing the number of hypotheses till infinity (or sample size) and shrinking our analysis to a subgroup of one person i .
- What can we do about it?
- We can explore heterogeneity in the systematic agnostic way with the help of machine learning.
- We will study this in the second part of "Modern High-Dimensional Linear Regression"
- Fundamental problem of Casual Inference: We can not observe the same person under treatment and not in the same time
- Statistical solution: Look at the groups of people and compare averages to calculate Average Treatment Effect (Δ).
 - Observational. Average Predictive Effect (π), yet, $\pi \neq \Delta$
 - Experimental. Average Treatment Effect (Δ) and $\pi = \Delta$
- We collect a lot of characteristics (W) about people (i) so that $W \gg i$
 - Here machine learning can help us to learn more.

3.3 Exercises

1. **Sample Variation I:** What is population?
2. **Sample Variation II:** What is sample?
3. **Sample Variation III:** What is the difference between standard deviation and standard error of the mean?
4. **Sample Variation IV:** Here are two tables of error types:

Table 7: Inferential Statistics: Error Types

		Null Hypothesis, H_0	
		Valid	Invalid
Judgement of H_0	Reject	False Positive	Correct Inference
	Fail to Reject	Correct Inference	False Negative

Table 8: Prediction Study: Error Types

		Predicted Value, \hat{Y}	
		Positive	Negative
Actual Value, Y	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

- What is the key differences between them? Do they contradict each other?
- When is it useful to use one approach, and when the other?
- Give examples to use each way of thinking for economic problems.

5. Hypothesis Testing I:

- The average interview callback rate is 8% e.g., Bertrand, M. and Mullainathan, S., 2004.
- You send ten resumes and applications and get one callback.
- How likely are you to observe this outcome given that your resume is not worse than average?
- How likely are you to observe this outcome or worse, given that your resume is not worse than average?
- Calculate the p-value by formula and with simple `binom.test()`.
- Formula of binomial probability: $p = \frac{n!}{n!x!}(q)^x q^{n-x}$, where q is the probability of success on a single trial, x - number of success, n - number of trials.

6. Hypothesis Testing II:

- Call the package `AER` and use the data `ResumeNames`. What is it about?
- Calculate the average rate of callback.
- Calculate the rate of callback for applicants with afro-american sounding names. Use `subset()` function.
- Test with `binom.test()` that applicant with afro-american sounding names is less likely to receive the callback than the average applicant.
- Test with `binom.test()` that applicant with afro-american sounding names is less or more likely to receive the callback than the average applicant.

- Turn `call` variable to binary variable that takes 1 if "yes" with `ifelse()` function.
- Use regression (`lm()`) to test if an afro-american applicant is equally likely to receive the call as a caucasian applicant.

1. **Causal Inference I:** What is the fundamental problem of causal inference?

2. **Causal Inference II:** What are the two solutions proposed?

3. **Causal Inference III:** What is an average predictive effect (APE)?

4. **Causal Inference IV:** What is the average treatment effect (ATE)?

5. **Causal Inference V:** When APE and ATE is equal?

6. **Statistical Solution I:**

- Call the package `AER` and use the data `ResumeNames`.
- Check the variables. For which variables you can assume that ATE is equal to APE in predicting callback?
- Turn `call` variable to binary variable that takes 1 if "yes" with `ifelse()` function.
- Use regression (`lm()`) to test how ethnicity, gender, and industry is associated with callback. Interpret your results.

7. **Statistical Solution II: What predicts callback? or How to make a fictitious CV to get a callback?;**)

- Call the package `AER` and use the data `ResumeNames`.
- Split the dataset in training and testing set.
- Build a model on your training set with regression.
- Use function `predict(...type="response")` to predict callback on your testing set.
- Calculate your accuracy rate.
- Calculate your mean square error.
- Find a way to improve the prediction.
- Hint 1: Use `ifelse()` function to turn probabilities into binary outcomes.
- Hint 2: You can find your threshold level based on predictions.
- Hint 3: You can re-run the code many times.
- Hint 4: Consider using `randomForest` for variable selection once you setup the model. Re-run the code many times.

4 Modern High Dimensional Linear Regression

5 Effect Heterogeneity

6 Identification under Selection on Observables