

# SOSC 5340: Overview of Statistical Inference and Prediction

Han Zhang

# Outline

Logistics

Probability

Statistics

Estimation

Inference

Prediction

Summary

## Instructors

**Instructor:** ZHANG, Han

- Office: 2379
- Email: zhangh@ust.hk
- Office Hour: Mon 10-11AM
  - In-office
  - Zoom: <https://hkust.zoom.us/j/6522716568>

**Teaching Assistant:** Li, Jingchen

- Office: 3001
- Email: jlieg@connect.ust.hk
- Office Hour: TBD

# Self Introduction

## Components

- Second course in SOSC's statistics sequence for research graduate students (after SOSC 5090);
- Three core goals of social sciences:
  1. Description: describing one variable
  2. Prediction: correlation between two social phenomena.
  3. Explanation: are the correlation causal?
- Three set of knowledge/skills
  1. Statistical estimation and inference
  2. Applied regression modeling
  3. Causal inference (second half of the semester)

## Grading

Attendance	10%
Assignments	30%
Presentation of a published research (15 min)	10%
Presentation of your final paper (20 min)	15%
Write-up of your final paper	35%

## Assignments

- Homework assignment: short coding homework to make sure that you know how to run models we covered in the lectures.
  - 3-4 times
- Our TA will hold tutorial sections to teach you
  - how to run these models before assignments.
  - discuss solutions of previous assignment
  - 3-4 times

## Presentation of an academic article

- As a researcher, you will need to present your own research at academic conferences.
- This exercise prepares you with relevant skills
- A list of academic publications will be distributed later
- These publications apply what we have learned in the class to real social science problems, or discuss methodological pitfalls in current applied research
- You are required to select on article, and present it to the entire class (15 minutes)



## Final Paper

- As a researcher, you will need to apply what you have learnt to a real social science problem, and write an academic article.
- It is very important to write and present your own work.
- You need to
  - Present your own final paper to the class (15%)
  - Write it down (35%)
- Treat this as a real paper that has the potential to be published at academic journals/presented at academic conferences.

## Materials

- Some textbooks that inspired the slides
  - Scott Cunningham. *Causal Inference: the Mixtape*. Yale University, 2021. My weekly schedule is most similar to this book. It can be freely viewed at <https://mixtape.scunning.com/index.html>
  - Aronow, Peter M., and Benjamin T. Miller. *Foundations of Agnostic Statistics*. Cambridge University Press, 2019. (more mathematical; mostly used for the first half of the class).
  - Joshua D. Angrist and Jorn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricists Companion*. Princeton University Press, 2009. (more applied; mostly used for the second half of the class).
  - Hansen, Bruce. *Econometrics*, 2020. Free at the author's website <https://www.ssc.wisc.edu/~bhansen/econometrics/>

## Coding

- We will use R for lectures and tutorials
- If you prefer Stata, that is okay
- Scott Cunningham. *Causal Inference: the Mixtape* has both R and Stata codes.

## Three types of social science questions

1. Description: describing one variable
2. Prediction: correlation between two social phenomena.
3. Explanation: are the correlation causal?
  - Today's lecture focuses on the first two
  - How do we use statistics to do description and prediction

## Random variables

- Random variable: abstraction of some concept we care about.
- Examples:
  - define random variable  $X$  as gender; it can take several values from *male*, *female*, *transgender*, ...
  - define random variable  $X$  as height; it can take numeric values.

## Probability distribution

- Descriptive problem: measure the probability distribution of a random variable
- Probability density function (PDF):  $f(x)$ 
  - How likely does random variable  $X$  take a particular value  $x$
  - $f(x) = P(X = x)$
- Cumulative distribution function (CDF):  $F(X) = P(X \leq x)$ 
  - The probability that a random variable  $X$  takes a value equal to or less than  $x$ ?

## Joint and Conditional Probability

- Predictive/correlation problem can be addressed by the joint/conditional probabilities of random variables
- Joint probability density function:  $f(X = x, Y = y)$ 
  - Probability that  $X$  takes value  $x$  and  $Y$  takes value  $y$
- Conditional probability density function
  - Probability that  $Y$  takes value  $y$ , given that  $X$  takes value  $x$ .

## Probability (exercise)

- Two treatments for kidney stones

Kidney Stone	Treatment A		Treatment B	
	cured	patient	cured	patient
Small	81	87	234	270
Large	192	263	55	80
Total	273	350	289	350

- $X$  is whether the patient is cured (1) or not (0)
- What is the conditional probability of being cured, given treatment A and B?
  - $P(X = 1 | \text{treatment} = A)$
  - $P(X = 1 | \text{treatment} = B)$
- $P(X = 1 | \text{treatment} = A) = 273/350 = 78\%$
- $P(X = 1 | \text{treatment} = B) = 289/350 = 83\%$
- treatment B is more effective in the entire population



## Probability (exercise)

Kidney Stone	Treatment A		Treatment B	
	cured	patient	cured	patient
Small	81	87	234	270
Large	192	263	55	80
Total	273	350	289	350

- What is the conditional probability of being cured, conditional on treatment status and stone size?
- Small Kidney Stone:
  - $P(X = 1 | treatment = A, size = small) = 81/87 = 93\%$
  - $P(X = 1 | treatment = B, size = small) = 234/270 = 87\%$
- Large Kidney Stone:
  - $P(X = 1 | treatment = A, size = large) = 192/263 = 73\%$
  - $P(X = 1 | treatment = B, size = large) = 55/80 = 69\%$
- B is more effective in the entire population, but A is more effect for both patients with small and large kidney stones.
- This is known as the **Simpson's Paradox**. Why?

## Expected Value

- Expectation (expected value)  $E(X)$ :
  - The **average** value of a random variable  $X$
- Categorical variable's expectation:
  - Let  $X$  be a random variable with a finite number of finite outcomes  $x_1, x_2, \dots, x_k$  occurring with probabilities  $p_1, p_2, \dots, p_k$
  - $E(X)$  is the weighted average of  $X$ , with probability as weights
  - $E[X] = x_1 p_1 + x_2 p_2 + \dots + x_k p_k = \sum_{i=1}^k x_i p_i$
- Continuous variable's expectation

$$E(X) = \int x \cdot f(x) dx$$

## Expected Value (exercise)

- What is the  $E(X)$  of the random variable  $X$ ?

X	P(X)
0	0.8
1	0.1
2	0.06
3	0.03
4	0.04

## Expected Value

- Useful formula of expected values
  1. Linearity of expectation:
$$E(aX + bY + c) = aE(X) + bE(Y) + c$$
  2. Constant's expectation is constant:  $E(c) = c$

## Conditional Expectation

- Conditional expectation  $E(Y|X = x)$ :
  - What is the average value of a random variable  $Y$ , when we already know that random variable  $X$  takes a **fixed** value  $x$
- Useful formula 3: Law of Iterated Expectation (Law of Total Expectation)

$$E[Y] = E[E[Y|X]] = \begin{cases} \sum_x E[Y|X = x]P(X = x) & \text{discrete } X \\ \int_{-\infty}^{\infty} E[Y|X = x]f(x)dx & \text{continuous } X \end{cases} \quad (1)$$

- Basically, this theorem says that:
  - if we have knowledge about one variable  $X$  ( $P(X)$ )
  - and how  $X$  relates to  $Y$  (through  $P(Y|X)$ )
  - we can calculate the average of another variable  $Y$ .

## Variance

- The variance measures the dispersion or the “spread” of a probability distribution.
- The variance of a random variable  $X$ , denoted  $V(X)$ , is the expected value of the square of the deviation of  $X$  from its mean:
- $$V(X) = E[(X - E(X))^2]$$
- Standard deviation:  $\sigma = \sqrt{V(X)}$

# Variance

## Definition (Alternative Formula for Variance)

$$V(X) = E[X^2] - E[X]^2$$

Proof.

$$V(X) = E[(X - E(X))^2] \tag{2}$$

$$= E[X^2 - 2XE(X) + E(X)^2] \tag{3}$$

$$= E(X^2) - 2E[XE(X)] + E[E(X)^2] \tag{4}$$

$$= E(X^2) - 2E(X)E(X) + E(X)^2 \tag{5}$$

$$= E(X^2) - E(X)^2 \tag{6}$$



# Probability and Statistics

- Probability is defined on population
- Probability of population is often very hard to obtain;
  - We need to have information of every unit in the population
  - But it's often unrealistic
- Statistics (or sample statistics) is an quantity computed from samples



## I.I.D. random variables

- Example:  $X$  is height and we want its probability distribution of all HK residents
- We can collect every HKer's height; high cost and population changes
- Or we can sample a HKer and record his/her height  $X_1$ , and then sample another HKer get height  $X_2$ .
- This process continues 100 times, we get  $(X_1, X_2, \dots, X_{100})$
- $(X_1, X_2, \dots, X_{100})$  are independent and identically distributed (I.I.D.)
  - **independent**: our  $i$  th draw does not depend on the  $j$  th draw;
    - in math:  $P(X_i, X_j) = P(X_i)P(X_j)$
  - **identically distributed**: they all come from the same probability distribution: HKer's height.
    - They are not coming from a different distribution, say, heights of desks

## I.I.D. random variables (exercise)

- When the independent assumption may be violated?
  - e.g., samples are not random, but HKUST students.
- When the identically distributed assumption may be violated?
  - e.g., population changes during the sampling process.

## Sample Mean of I.I.D. random variables

- Let  $X_1, \dots, X_n$  be i.i.d. random samples of random variable  $X$
- We do not know  $E(X)$  and we want to estimate it using samples

### Definition (Sample Mean)

The sample mean  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$

Theorem (The Expected Value of the Sample Mean is the Population Mean)

$$E(\bar{X}) = E(X)$$

- Implication:
  - Expectation of the sample mean equals population mean
  - We cannot directly obtain population mean
  - But **mean of sample mean is something easier to obtain**

## Sample Mean of I.I.D. random variables

The Expected Value of the Sample Mean is the Population Mean.

$$E(\bar{X}) = E\left(\frac{1}{n}(X_1 + \cdots + X_n)\right) \quad (7)$$

$$= \frac{1}{n}E(X_1 + \cdots + X_n) \quad (8)$$

$$= \frac{1}{n}[E(X_1) + \cdots + E(X_n)] \quad (9)$$

$$= \frac{1}{n}[E(X) + \cdots E(X)] \quad (10)$$

$$= E(X) \quad (11)$$



## Mean of sample mean

- Mean of sample mean  $\rightarrow$  suggests that you need to take all possible surveys, calculate the mean from each survey, and take the averages
- Basically it's the average of an average
- In real life, we only have one survey (i.e., one  $\bar{X}$ ), so we still cannot calculate  $E(\bar{X})$ , which means that we also do not know the value of  $E(X)$
- How is the **one sample mean**  $\bar{X}$  associates with the population mean  $E(X)$ ?

## Law of Large Numbers

- Let  $X_1, \dots, X_n$  be i.i.d. random samples of random variable  $X$

Theorem (Weak Law of Large Numbers, Jacob Bernoulli, 1713)

*The sample mean  $\bar{X}$  **converges in probability** to the population mean  $E(X)$ , as  $n \rightarrow \infty$ .*

- Convergence in probability:
  - If  $a$  and  $b$  convergence in probability, it is very likely that their difference will be very small.
  - $\lim_{n \rightarrow \infty} P(|a - b| \leq \epsilon) = 1$ , for all  $\epsilon > 0$ .
- Implication of the **Weak Law of Large Numbers**:
  - We can use one sample mean to **estimate** population mean

## Variance of the Sample Mean

- Let  $X_1, \dots, X_n$  be i.i.d. random samples of random variable  $X$ , with finite variance  $V(X)$

### Theorem (Sampling Variance of the Sample Mean)

*The variance of the sample mean is  $V(\bar{X}) = \frac{V(X)}{n}$*

- That is, variance of the sample mean decreases, as  $n$  increases.

## Estimator

- Estimation: use (sample) statistics to infer population quantities
- Let  $X_1, \dots, X_n$  be i.i.d. random samples of random variable  $X$
- We care about some population quantity of interest  $\theta$  (e.g., mean, variance, median, etc)

### Definition (Estimate and Estimator)

Estimator of a **population** quantity  $\theta$  is a function of the **samples**,  $\hat{\theta} = h(X_1, \dots, X_n)$ ;  $\hat{\theta}$  is the **estimate** of  $\theta$ .

- In a nutshell, **statistics uses estimator to provide estimate of population quantity**
- Example: an estimator of population mean  $E(X)$  is sample mean  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$



## Estimators

- There are usually many different estimators of the same quantity.
  - E.g.,  $X_1$  is also an estimator of  $E(X)$ . But intuitively it is not as good as the sample mean  $\bar{X}$
  - How can we say one estimator is better than the other?
  - What properties should good estimators have?

## Desirable Property: Unbiasedness

- For an estimator  $\hat{\theta}$ , its bias is defined as  $E(\hat{\theta}) - \theta$

### Definition (Unbiased Estimator)

An estimator  $\hat{\theta}$  of  $\theta$  is an unbiased estimator if  $E(\hat{\theta}) = \theta$  or bias is 0

- Question: sample mean  $\bar{X}$  is an unbiased estimator of population mean  $E(X)$ . Why?
- Answer: because the expectation of sample mean equals to population mean ( $E(\bar{X}) = E(X)$ )

[see proof here](#)

## Desirable property: Consistency

### Definition (Consistent Estimator)

An estimator  $\hat{\theta}$  is a consistent estimator if  $\hat{\theta}$  converges in probability to  $\theta$ , as  $n \rightarrow \infty$ .

- Question: sample mean is a consistent estimator of population mean. Why?
- Answer: because of the Law of Large Numbers.

## Estimation basics

- Let  $X_1, \dots, X_n$  be i.i.d. random samples of random variable  $X$  with variance  $V(X)$
- We see that sample mean  $\bar{X}$  is an **unbiased and consistent** estimator of population mean  $E(X)$
- It is tempting to extend this method to other population quantities, by:
  1. express some more complex population quantity as the combinations of estimable simple quantities (i.g.,  $E(X)$ )
  2. plug-in the sample estimator.
- This is called **plug-in** principle.

## Application: plug-in estimator for population variance

- We want to apply the **plug-in** principle to estimate population variance  $V(X)$ .
- Step 1: express  $V(X) = E[X^2] - E[X]^2$ 
  - we already know how to estimate  $E(X)$ :  $\bar{X}$
- Step 2: **plug-in**  $\bar{X}$  in place of  $E(X)$

### Definition (Plug-in Variance Estimator)

$$\hat{V}(X) = \overline{X^2} - \bar{X}^2$$

- Is this plug-in variance estimator a good estimator?
  - As we have learned, an good estimator should have two good properties
  - unbiased?
  - consistent?

## Plug-in estimator for population variance

- Unbiased estimator means  $E(\hat{\theta}) - \theta = 0$
- Our variance estimator of  $V(X)$  is  $\hat{V}(X) = \overline{X^2} - \bar{X}^2$
- Unbiasedness:

$$E(\hat{V}(X)) = E[\overline{X^2} - \bar{X}^2] = E[\overline{X^2}] - E[\bar{X}^2] \quad (12)$$

$$= E[X^2] - (E[X]^2 + V[\bar{X}]) \quad (13)$$

$$= (E[X^2] - E[X]^2) - \frac{V[X]}{n} \quad (14)$$

$$= V[X] - \frac{V[X]}{n} \quad (15)$$

$$= \frac{n-1}{n} V[X] \quad (16)$$

## Unbiased Estimator for Population Variance

- Plug-in population variance estimator  $\hat{V}(X) = \overline{X^2} - \bar{X}^2$  is **biased**
- Plug-in population variance estimator  $\hat{V}(X) = \overline{X^2} - \bar{X}^2$  is **consistent** (as  $n \rightarrow \infty$ ,  $\frac{n-1}{n}$  goes to 1)
- In general, **plug-in estimator is consistent, but may be biased** (advanced topic).

### Theorem (Unbiased Estimator of Population Variance)

$\hat{V}(X) = \frac{n}{n-1}(\overline{X^2} - \bar{X}^2)$  is an unbiased and consistent estimator of population variance  $V(X)$

## Inference vs Estimation

- **Estimation** is about getting the **(point) estimate**  $\hat{\theta}$  of quantity of interest  $\theta$ 
  - With unbiased estimator,  $\hat{\theta}$  on average equals to  $\theta$
  - With consistent estimator,  $\hat{\theta}$  converges to  $\theta$  with more and more sample data
  - But in reality, we have only one  $\hat{\theta}$
- **Inference** is about how certain we are about the estimate  $\hat{\theta}$ 
  - The first information is the variance of the estimate  $\hat{\theta}$ 
    - lower variance -> more certain



## Example: Variance of the sample mean

- What's the variance of sample mean (as an estimate)?
- Try **plug-in** estimator
  - Step 1: express the quantity of interest  $V(\bar{X}) = \frac{V(X)}{n}$
  - Step 2: plug-in (since we just shown how to estimate  $V(X)$  in the previous slide)

### Theorem (Estimator of the Sampling Variance of the Sample Mean)

$$\hat{V}(\bar{X}) = \frac{\hat{V}(X)}{n}$$

- Plug-in estimator is an unbiased and consistent estimator this time (proof omitted)
- $\sqrt{\hat{V}(\bar{X})}$  is called **standard error**.

## Central Limit Theorem

- Sample mean as an estimator has another good property: distribution of sample mean is always a normal distribution,
- Let  $X_1, \dots, X_n$  be i.i.d. random samples of random variable  $X$ , with finite  $E(X) = \mu$  and  $V(X) = \sigma^2 > 0$
- Standardized Sample Mean  $Z = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma}$ 
  - $E(Z) = 0$ ;  $V(Z) = \sigma(Z) = 1$ ; hence the name standardized sample mean.

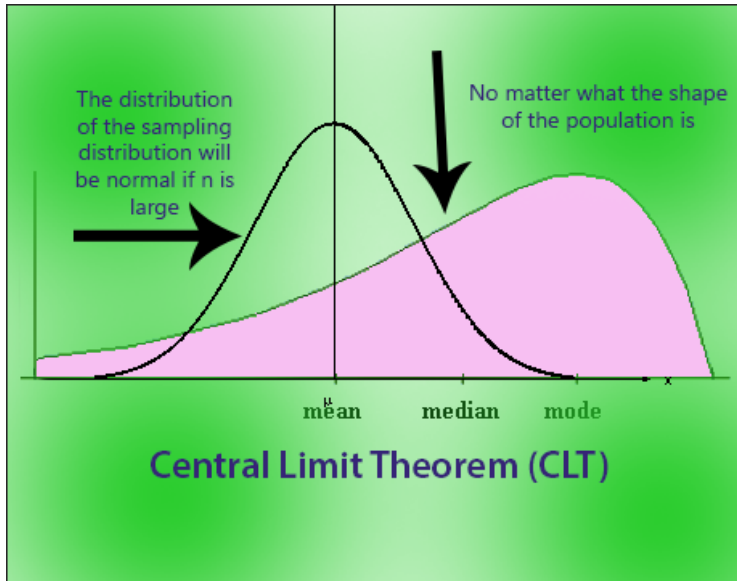
### Theorem (Central Limit Theorem)

- The *distribution of  $Z$*  converges to a standard normal distribution ( $Z \sim N(0, 1)$ ), as  $n \rightarrow \infty$ .
- Or equivalently,  $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ , as  $n \rightarrow \infty$

## Central Limit Theorem (CLT)

- Distribution of sample mean is always a normal distribution,
  - even when the population distribution **is not distributed normally**;
- $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$
- Implications:
  - $\bar{X} = \mu = E(X)$ ; therefore, CLT automatically implies the Law of Large Numbers
  - Variance of the sample mean equals the variance of this normal distribution
- Central limit theorem provides a general way for us to infer the uncertainty around our estimate of sample mean

## Central Limit Theorem



## Desirable Property: asymptotically normal

- Central Limit Theorem means that sampling distribution of the sample mean will tend to be approximately normal
- Some other estimators also have this good property:

### Definition (Asymptotic Normal Estimator)

An estimator  $\hat{\theta}$  is an asymptotically normal estimator, if  $\sqrt{n}(\hat{\theta} - \theta) \sim N(0, \phi^2)$  for finite  $\phi > 0$ , as  $n \rightarrow \infty$ .

- Many estimators you will learn in this course is asymptotically normal
  - But not all estimators have this good property
- The good thing about asymptotically normal estimator is that we can obtain confidence interval easily

# Confidence Interval

## Definition (Confidence interval)

A  $\alpha$  confidence interval for quantity of interest  $\theta$  is an estimated interval that covers the true value of  $\theta$  with at least  $\alpha$  probability

- Example: in social sciences, we often uses  $\alpha = 95\%$  confidence interval that looks like  $[\theta_{min}, \theta_{max}]$ . The probability that the true  $\theta$  falls between  $[\theta_{min}, \theta_{max}]$  is at least 95%.
- Note 1: wide confidence intervals are valid, but not useful
  - e.g.,  $\theta$  is the average height of HKers;  $[0, 2.5]$  is a valid confidence interval but it is not very useful.
- Note 2: confidence interval does not need to be symmetric

# Normal Approximation-based Confidence Interval

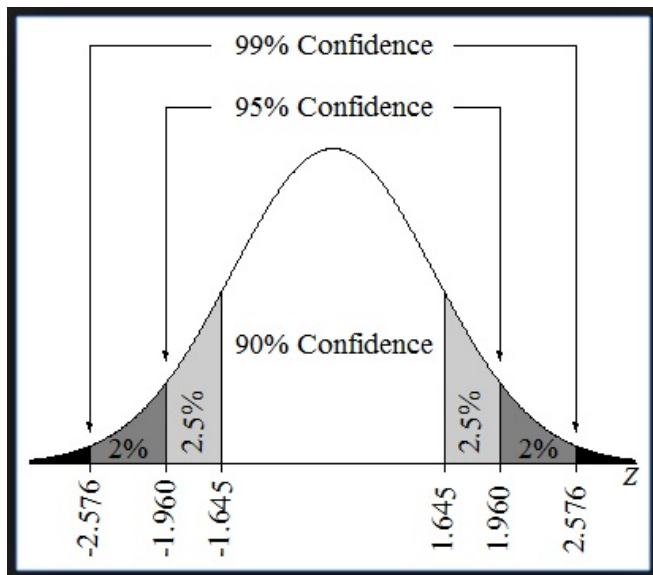
## Definition (Estimating Normal Approximation-based Confidence Interval)

A normal approximation-based confidence interval for  $\theta$  can be estimated by:

$$\left( \hat{\theta} - q_{\frac{1+\alpha}{2}} \sqrt{\hat{V}(\hat{\theta})}, \hat{\theta} + q_{\frac{1+\alpha}{2}} \sqrt{\hat{V}(\hat{\theta})} \right)$$

- $q$  is the quantile function of a standard normal distribution
  - $\alpha = 0.05$ ;  $q_{0.975} = 1.96$
  - $\alpha = 0.01$ ;  $q_{0.995} = 2.58$
- Normal Approximation-based Confidence Interval is valid for asymptotically normal estimators

## Illustration





## Steps to empirically estimate confidence interval

- Steps to estimate the Normal Approximation-based Confidence Interval for sample mean in a given sample
- Step 1: calculate sample mean  $\bar{X}$  and sampling variance of the sample mean  $\hat{V}(\bar{X})$
- Step 2: construct confidence interval as

$$\left( \bar{X} - q_{\frac{1+\alpha}{2}} \sqrt{\hat{V}(\bar{X})}, \bar{X} + q_{\frac{1+\alpha}{2}} \sqrt{\hat{V}(\bar{X})} \right)$$

- E.g., for 95% confidence interval

$$\left( \bar{X} - 1.96 \sqrt{\hat{V}(\bar{X})}, \bar{X} + 1.96 \sqrt{\hat{V}(\bar{X})} \right)$$

# Bootstrap

- Normal Approximation is not the only way to construct valid confidence intervals
  - it only works for asymptotic normal estimator
- The Bootstrap is more general method to construct confidence intervals; one of the most important modern statistical concept (Efron, 1979)
  - Drawback of Bootstrap: it's a data-driven method; slow; no analytical solutions.

## Bootstrap procedures

Assume we already have  $X_1, \dots, X_n$  be i.i.d. random samples of random variable  $X$ ). We are interested in estimating a  $\alpha$  confidence interval for a population quantity  $\theta$

1. Take a **with replacement** sample of size  $n$  from  $X_1, \dots, X_n$
2. Calculate the sample analog of  $\theta$
3. Repeat 1 and 2 for  $m$  times. We end up having  $m$  estimates of  $\theta$ ,  $(\hat{\theta}_1, \dots, \hat{\theta}_m)$
4. Take the  $\frac{1-\alpha}{2}$  and  $\frac{1+\alpha}{2}$  quantile of the values  $(\hat{\theta}_1, \dots, \hat{\theta}_m)$ . These two quantiles give us the bootstrap confidence intervals.

## Confidence Interval: Interpretations

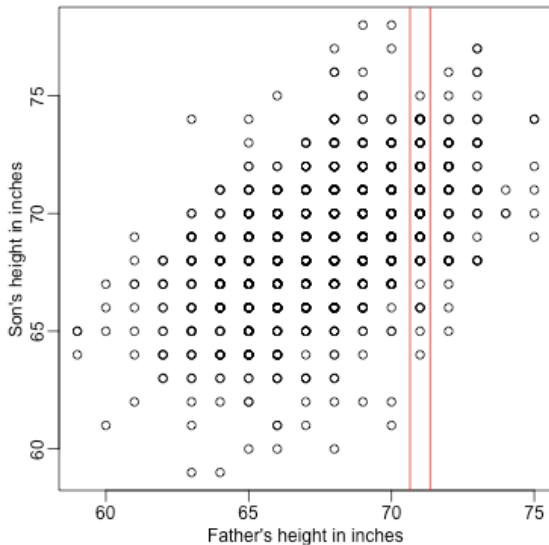
- Interpreting confidence intervals carefully
- a 95% confidence interval  $[\theta_{min}, \theta_{max}]$  contains the population quantity  $\theta$  with at least 95% probability.
- In reality, we are **estimating** confidence intervals  $[\theta_{min}, \theta_{max}]$ .
- How should we interpret 95% **estimated** confidence interval  $[\hat{\theta}_{min}, \hat{\theta}_{max}]$  then?
  - Through **repeated samples** (each time we sample  $n$  units), 95% of estimated confidence intervals would contain the population quantity  $\theta$
  - If we know  $[\theta_{min}, \theta_{max}]$ , we do not need repeated sampling
  - note that in reality we only have one sample (of  $n$  units)

## Prediction

- Now let us move on to two variable setting
- Given two variables  $Y$  and  $X$ , and we observed  $X$  takes the value  $x$ .
- A prediction of  $Y$  given  $X$  is a function  $g(X)$  that approximate  $Y$
- For instance,  $E(Y|X)$  is a prediction of  $Y$  given  $X$
- Again, there are tons of ways to predict  $Y$  given  $X$  (e.g., median of  $Y$  given  $X$ )

## Prediction (example)

- Predicting son's height with father's height



## Using Conditional Expectation as Prediction

- If  $g(X) = E(Y|X)$ , that is, we use the conditional expectation as the prediction
- The prediction error is  $\epsilon = Y - E(Y|X)$
- This prediction error has some good properties

Property 1:  $E(\epsilon) = 0$ .

$$E(\epsilon) = E[Y - E(Y|X)] \quad (17)$$

$$= E(Y) - E[E(Y|X)] \quad (18)$$

$$= E(Y) - E(Y), (\text{Law of Iterated Expectation}) \quad (19)$$

$$= 0 \quad (20)$$



## Conditional Expectation as Prediction (cont'd)

Property 2:  $E(\epsilon|X) = 0$ .

$$E(\epsilon|X) = E[Y - E(Y|X)|X] \quad (21)$$

$$= E(Y|X) - E[E(Y|X)|X] \quad (22)$$

$$= E(Y|X) - E(Y|X), (\text{Law of Iterated Expectation}) \quad (23)$$

$$= 0 \quad (24)$$





## Using Conditional Expectation as Prediction

- Property 2 means that on conditional on  $X$ , the mean of prediction error is 0
- This property is also called **mean independent** because  $E(\epsilon|X) = E(\epsilon) = 0$ 
  - That is, we only assume that **on average**  $X$  and error are independent
  - Recall **independence** means that  $P(\epsilon|X) = P(\epsilon)$

# Independent, mean independent, and uncorrelated

- Independent:  $P(XY) = P(X)P(Y)$
- Mean independent:  $E(Y|X) = E(Y)$
- Uncorrelated:  $E(XY) = E(X)E(Y)$
- In general, we have the following relationship (the reverse is not true):

$X, Y$  are **independent**  $\implies X, Y$  are **mean independent**  
 $\implies X, Y$  are **uncorrelated**.

## Using Conditional Expectation as Prediction

- Property 3 says  $g(X)$  and error is uncorrelated; it can be derived from Property 2 (mean independence) and Property 1 ( $E(\epsilon) = 0$ )

Property 3:  $E(g(X)\epsilon) = 0$ , for any  $g(X)$ .

$$E[g(X)\epsilon] = E[g(X)(Y - E(Y|X))] \quad (25)$$

$$= E[g(X)Y - g(X)E(Y|X)] \quad (26)$$

$$= E[g(X)Y] - E[g(X)E(Y|X)] \quad (27)$$

$$= E[g(X)Y] - E[E(g(X)Y|X)], \text{ (} g(X) \text{ is a constant given } X \text{)} \quad (28)$$

$$= E[g(X)Y] - E[g(X)Y], \text{ (Law of Iterated Expectation)} \quad (29)$$

$$= 0 \quad (30)$$

## Evaluating Predictions

- We have seen that  $E(Y|X)$  is a good guess for  $Y$ :
  - Property 1: mean error is 0
  - Property 2: error and prediction  $g(X)$  is mean independent
  - Property 3: error and prediction  $g(X)$  is uncorrelated
- But mean error  $E(\epsilon) = E(Y - g(X))$  has one drawback:  
sensitive to the sign of error
- e.g.,  $Y = 0$ ; our guesses  $g(X)$  are -100, 100, -100, 100
  - Intuitively these guesses are not good
  - But  $E(\epsilon) = 0$

## MAE and MSE

- Mean Absolute Error (MAE):  $E[|Y - g(X)|]$
- **Mean Square Error (MSE)**:  $E[(Y - g(X))^2]$
- MSE make sure that you get penalized more, if the absolute error is large.
  - MSE is perhaps the most widely used error metric
- Both MAE and MSE  $\geq 0$ ; a good estimation thus should **minimize** MAE or MSE

## Conditional Expectation As the Best Predictor

- There are some even better properties of  $E(Y|X)$  that make it the **best** predictor, **given Mean Squared Error**

### Theorem (Conditional Expectation as the Best Predictor)

*Conditional Expectation Function  $E(Y|X)$  is the best predictor of  $Y$  because it minimizes Mean Squared Error*

- Proof sketch: we have two predictions for  $Y$ ,  $E(Y|X)$  and any other  $g(X) \neq E(Y|X)$
- We want to show that the MSE of any other  $g(X)$  is not smaller than the MSE of  $E(Y|X)$
- In math term:  $E[(Y - g(X))^2] \geq E[(Y - E(Y|X))^2]$
- Hint: use the conditional expectation error  $\epsilon = Y - E(Y|X)$

# Conditional Expectation As the Best Predictor

## Conditional Expectation as the Best Predictor.

$$E[(Y - g(X))^2] = E[(\epsilon + E(Y|X) - g(X))^2] \quad (31)$$

$$= E[\epsilon^2 + 2\epsilon(E(Y|X) - g(X)) + (E(Y|X) - g(X))^2] \quad (32)$$

$$= E[\epsilon^2] + 2E[\epsilon(E(Y|X) - g(X))] + E[(E(Y|X) - g(X))^2] \quad (33)$$

$$= E[\epsilon^2] + E[(E(Y|X) - g(X))^2], (\text{Property 3}) \quad (34)$$

$$\geq E[\epsilon^2] \quad (35)$$

$$= E[(Y - E(Y|X))^2] \quad (36)$$



## Conditional Expectation As the Best Predictor (implications)

- Note that this says that the conditional expectation gives **an upper bound** on how well we can make a guess of  $Y$  based on  $X$ , if we want to minimize MSE
  - If the conditional expectation itself is not a very good predictor, we can still make lots of errors
  - But in this case, other predictions can only be worse
- $E(Y|X)$  is the best predictor of  $Y$ , **if the criterion is to minimize MSE**
  - There are other criteria to evaluation predictions
- Our next half semester is devoted onto understanding how to estimate  $E(Y|X)$



## Today's summary

- Population/sample
- What is estimator; three good properties of estimator
- Inference; confidence interval; normal approximation vs. Bootstrap
- Conditional expectation is the best predictor in terms of minimizing MSE

## Today's readings

- Aronow, Peter M., and Benjamin T. Miller. *Foundations of Agnostic Statistics* . Cambridge University Press, 2019.  
(Chapter 2 - 4)
- Joshua D. Angrist and Jorn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricists Companion* . Princeton University Press, 2009.
  - Discuss Conditional expectation is the best predictor (Chapter 3.1)
  - Motivated differently from Aronow and Miller