INTRODUCTION TO STATISTICS

LECTURER: PEDRO IZQUIERDO LEHMANN
pizquie1@jh.edu

LECTURE 1

TODAY

- WHAT IS STATISTICS?
- STATISTICAL MODEL
- SAMPLE
- ASSUMPTIONS
- FINITE SAMPLE V/S ASSYMPTOTIC
- OUTLINE OF 6 LECTURES.

WHAT IS STATISTICS?

- STATISTICS $\neq$ MATH.

  IT'S A DISCIPLINE THAT USES MATH TO BUILD MODELS OF REALITY. AND DRAW CONCLUSIONS BASED ON EXPERIENCE.

- MAIN ASSUMPTION: SOME PHENOMENON OF INTERST ITS GENERATED BY SOME PROBABILITY LAW. $\rightarrow f_0$

Ex 0: THROWING WEIGHTED COINS:

$$f_0 = \begin{cases} \text{HEADS} = 1 & \text{WITH PROB } p_0 \\ \text{TAILS} = 0 & \text{WITH PROB } 1-p_0 \end{cases}$$

$p_0 \in (0,1)$

WE DON'T KNOW f_0! BUT..

I) WE KNOW (WE ASSUME) SOME STRUCTURE OF f_0:

$$f_0 \in \mathcal{F}$$ ($\mathcal{F}$ IS A STRUCTURED SET OF DISTRIBUTIONS)

$\mathcal{F}$ IS CALLED A STATISTICAL MODEL.

EX: TOSSING COINS:

$$f_0 \in \mathcal{F} = \left\{ f = \begin{cases} 1 & \text{WITH PROB } p \\ 0 & \text{WITH PROB } 1-p \end{cases} \;\middle|\; p \in (0,1) \right\}$$

II) WE HAVE A SAMPLE OF OBSERVATION OF THE PHENOMENUM OF INTEREST:

$$x_1, \dots, x_1 \quad \text{(DATA POINTS)}$$

GIVEN I), THESE DATA POINTS ARE THE REALIZATION OF $\underbrace{\text{i.i.d}}_{}$ RANDOM VARIABLES

$$X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} f_0$$ THIS CAN BE RELAXED!

GOAL : CHARACTERIZE f_0 FROM

I) $f_0 \in \mathcal{F}$ (STATISTICAL MODEL)

II) $X_1, \dots, X_n \sim f_0$ (SAMPLES)

Ex: TOSSING COINS

- ESTIMATE p_0, e.g.

$$\hat{p}_0 = \frac{1}{n} \sum_{k=1}^{n} x_k$$

- CONFIDENCE INTERVALS

- HYPOTHESIS TESTING : CAN WE CONFIDENTLY SAY THAT $p_0 \notin S \subseteq [0,1]$.

STRUCTURE (ASSUMPTIONS)

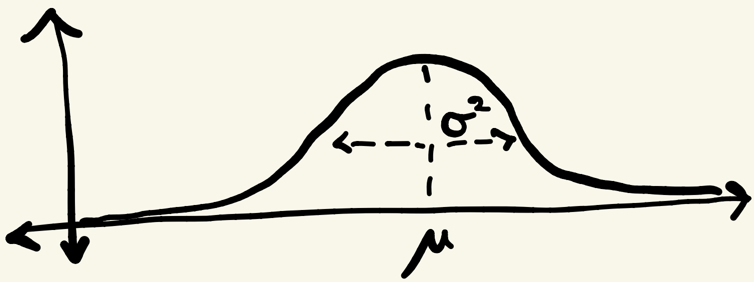
I) STATISTICAL MODEL

a) $\mathcal{F} = \{ f_\theta : \theta \in \Theta \subseteq \mathbb{R}^d \}$

PARAMETRIC STATISTICAL MODEL.

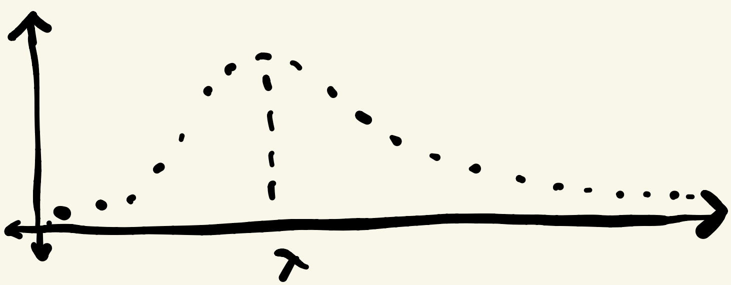
- POPULATION ATTRIBUTE (e.g. HEIGHT)

$$f_0 \in \mathcal{F} = \{ \mathcal{N}(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0 \}$$



- NUMBER OF GOALS IN A SOCCER MATCH

$$f_0 \in \mathcal{F} = \{ \text{POISSON}(\lambda) : \lambda \in (0, \infty) \}$$



b) NON PARAMETRIC MODELS

$\mathcal{F}$ IS SUCH THAT $\dim(\mathcal{F}) = \infty$.

Ex : SOME POPULATION ATTRIBUTE FOR WHICH WE HAVE LITTLE INTUITION,

$f_0 \in \mathcal{F}$ {ALL THE K-LIPSCHITZ CONTINUOUS DISTRIBUTIONS }

II) SAMPLES.

- $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} f_0$

  INDEPENDENT AND IDENTICALLY DISTRIBUTED

  STANDARD BUT STRONG!

- $X_1, \dots, X_n \sim f_0$ IDENTICALLY DISTRIBUTED AND EXCHANGABLE

  FOR ANY INDEX PERMUTATION $\sigma : [n] \to [n]$ (BIJECTIVE):

  $(X_{\sigma(1)}, \dots, X_{\sigma(n)})$ HAS THE SAME DISTRIBUTION

  INCREASINGLY POPULAR!

- $(X_1, \dots, X_n) \sim f_0 \rightarrow$ VECTOR VALUED DISTRIBUTION

  WEAK BUT NOT VERY VERSATILE!

PAY ATTENTION TO OUR ASSUMPTIONS!

- ASSUMPTIONS DON'T HOLD:
  → MATHEMATICAL MODEL DOES NOT REPRESENT OUR PHENOMENOM OF INTEREST
  → CONCLUSION ARE DEVOIDED FROM MEANING.
- ASSUMPTION HOLDS APPROXIMATELY
  → THE ABOVE ISSUE IS JUST APPROXIMATELY PROBLEMATIC
- WE NEED ASSUMPTIONS TO DO MATH.

TRADE OFF!

FINITE SAMPLE VS ASSYMPTOTIC

- IDEALLY WE WANT FINITE SAMPLE RESULTS:

  → $n < \infty$.

  (ANY REAL SAMPLE IS FINITE!)

  Ex: HOEFFDING BOUND: IF

  $X_1, \dots, X_n \overset{i.i.d}{\sim} f_0$, $0 \leq X_1 \leq 1$; THEN

$$\forall \varepsilon: \mathbb{P}\left(\left|\frac{1}{n}\sum_{k=1}^{n} X_k - \mathbb{E}(X_1)\right| \geq \varepsilon\right) \leq e^{-2n\varepsilon^2}$$

- WE MIGHT NOT HAVE SUFFICIENT STRUCTURE.
- MATH CAN BE TOO HARD/INTRACTABLE.

- WE CAN STUDY ASSYMPTOTICS!

  → $n = \infty$, $n \longrightarrow \infty$

  (NO REAL SAMPLE IS INFINITE)

  Ex (WEAK) LAW OF LARGE NUMBERS

$$\forall \varepsilon: \mathbb{P}\left(\left|\frac{1}{n}\sum_{k=1}^{n} X_k - \mathbb{E}(X_1)\right| \geq \varepsilon\right) \xrightarrow{n \to \infty} 0$$

LIFE (MATH) IS EASIER!

OUTLINE FOR NEXT LECTURES

1) WHAT'S STATISTICS?
2) STATISTICS & ESTIMATORS
3) CONFIDENCE INTERVALS
4) HYPOTHESIS TESTING
5) FINITE SAMPLE & ASSYMPTOTIC PARAMETRIC METHODS
6) LINEAR REGRESSION

TOPICS WE ARE NOT INTRODUCING BUT YOU MIGHT LIKE

- NON PARAMETRIC STATS
- HIGH DIMENSIONAL STATS
- BAYESIAN STATS
- STATISTICAL LEARNING

☺

LECTURE 2

LAST TIME

- WHAT IS STATISTICS?
- SAMPLE
- STATISTICAL MODEL
- ASSUMPTIONS
- FINITE SAMPLE V/S ASYMPTOTICS

TODAY

- STATISTICS & ESTIMATORS
- CONSISTENCY
- BIAS & VARIANCE
- EFFICIENCY
- SUFFICIENCY.

STATISTICS & ESTIMATOR

I) DENOTE $\underbrace{\vec{x}=(x_1,\ldots,x_n)}_{\text{SAMPLE}} \in (\mathbb{R}^d)^n$ AS OUR DATA

II) ASSUME OUR SAMPLE IS THE REALIZATION OF

$$\vec{X}=(X_1,\ldots,X_n)\sim \underline{f_0}\in \underbrace{\mathcal{F}}_{\text{STATISTICAL MODEL.}}$$

- WE CALL **STATISTIC** OR **SAMPLE STATISTIC** TO ANY QUANTITY THAT CAN BE COMPUTED AS A FUNCTION OF THE SAMPLE

$$T(\vec{x}) \text{ FOR SOME } T:(\mathbb{R}^d)^n \longrightarrow \mathbb{R}^m$$

- FROM II), THE STATISTIC IS A REALIZATION OF THE RANDOM VECTOR

$$T(\vec{X})$$ ALSO CALL THIS A STATISTIC!

- IF OUR STATISTICAL MODEL IS PARAMETRIC

$$\mathcal{F} = \{ f_\theta : \theta \in \Theta \}, \quad \Theta \subseteq \mathbb{R}^\ell$$

THEN A STATISTIC THAT ESTIMATES THE PARAMETER θ IS CALLED AN ESTIMATOR OR PARAMETER ESTIMATOR, DENOTED $\hat{\theta}$.

Ex: $X_1, \dots, X_n \overset{i.i.d}{\sim} f_{\theta_0} \in \mathcal{F} = \{ f_\theta : \theta \in \Theta \}$

$$\theta_0 = \mathbb{E}(X_1)$$

ESTIMATORS FOR THE MEAN θ_0 ARE

$$\hat{\theta}_M = \frac{1}{n} \sum_{i=1}^{n} X_i$$ 1ST-MOMENT ESTIMATOR

$$\hat{\theta}_{MLE} = \underset{\theta \in \Theta}{\operatorname{argmax}}\ \mathcal{L}_n(\theta; \vec{X})$$

i.i.d.

$$\left[\prod_{i=1}^{n} f_\theta(X_i) \quad \text{LIKELIHOOD FUNCTION} \right]$$

MAXIMUM LIKELIHOOD ESTIMATOR

CONSISTENCY

- WE SAY AN ESTIMATOR $\hat{\theta}$ OF A PARAMETER θ IS **CONSISTENT** IF

$$\hat{\theta} \xrightarrow{n\to\infty} \theta \quad \text{IN PROBABILITY.}$$

$$\leftrightarrow \lim_{n\to\infty} \mathbb{P}(|\hat{\theta}-\theta| > \varepsilon) = 0$$

- CONSISTENCY IS THE "LEAST" THAT WE CAN EXPECT FROM AN ESTIMATOR!

Ex : 1ST MOMENT ESTIMATOR FOR MEAN μ

- $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$

i) $\mathbb{E}(\hat{\mu}) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) \overset{\text{i.i.d.}}{=} \mu$

ASSUME THIS IS FINITE

ii) $\mathrm{Var}(\hat{\mu}) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) \overset{\text{i.i.d}}{=} \frac{1}{n}\mathrm{Var}(X_1)$

iii) $\mathbb{P}(|\hat{\mu}-\mu| > \varepsilon) \overset{\text{MARKOV}}{\leq} \mathbb{E}[|\hat{\mu}-\mu|^2]/\varepsilon^2$

$$\overset{i)}{=} \frac{\mathrm{Var}(\hat{\mu})}{\varepsilon^2} \overset{ii)}{=} \frac{\mathrm{Var}(X_1)}{n\varepsilon^2} \xrightarrow{n\to\infty} 0$$

CONSISTENT!

OBS : THIS IS A PROOF FOR THE WEAK LAW OF LARGE NUMBERS!

BIAS & VARIANCE

- FOR AN ESTIMATOR $\hat{\theta}$ OF A PARAMETER θ WE DEFINE ITS BIAS AS

$$\text{Bias}(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$$

- WE WANT $\text{Bias}(\hat{\theta}) = 0$ UNBIASED ESTIMATOR

Ex: 1ST MOMENT ESTIMATOR FOR MEAN μ

$$\mathbb{E}(\hat{\mu}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(X_i) \overset{\text{i.i.d}}{=} \mathbb{E}(X_1) = \mu$$

UNBIASED!

Ex: NAIVE ESTIMATOR FOR VARIANCE σ^2

$$\hat{\sigma}_N^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2$$

$$\mathbb{E}(\hat{\sigma}_N^2) = \underbrace{\left(1 - \frac{1}{n}\right)}_{\frac{n-1}{n}} \sigma^2$$

BIASED!

Ex: UNBIASED ESTIMATOR FOR THE VARIANCE σ^2

$$\hat{\sigma}^2 = \left[\frac{n}{n-1}\right]\hat{\sigma}_N^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu})^2$$

$$\mathbb{E}(\hat{\sigma}^2) = \left[\frac{n}{n-1}\right]\mathbb{E}(\hat{\sigma}_N^2) = \cancelto{1}{\left[\frac{n}{n-1}\right]\left[\frac{n-1}{n}\right]}\sigma^2$$

UNBIASED!

BUT $\mathrm{Var}(\hat{\sigma}^2) = \mathrm{Var}\left(\left[\frac{n}{n-1}\right]\hat{\sigma}_N^2\right)$

$$= \underbrace{\left[\frac{n}{n-1}\right]^2}_{>1} \mathrm{Var}(\hat{\sigma}_N^2)$$

$$> \mathrm{Var}(\hat{\sigma}_N^2)$$

- BIAS/VARIANCE TRADE OFF.

$$0 < \|\hat{\theta} - \theta\|_2 := \mathbb{E}\left[(\hat{\theta} - \theta)^2\right]$$ MEAN SQUARED ERROR

$$= \mathbb{E}\left[(\hat{\theta} - \mathbb{E}(\hat{\theta}) + \mathbb{E}(\hat{\theta}) - \theta)^2\right]$$

$$= \mathbb{E}\left[(\hat{\theta} - \mathbb{E}(\hat{\theta}))^2\right] + (\mathbb{E}(\hat{\theta}) - \theta)^2$$

$(a+b)^2 = a^2 + b^2 + 2ab$

$$+ \cancelto{0}{2\,\mathbb{E}\left[(\hat{\theta} - \mathbb{E}(\hat{\theta}))(\mathbb{E}(\hat{\theta}) - \theta)\right]}$$

$$= \mathrm{Var}(\hat{\theta}) + \mathrm{Bias}(\hat{\theta})^2$$

TRADE OFF!

LECTURE 3

LAST TIME

- STATISTICS, ESTIMATORS
- CONSISTENCY
- UNBIASEDNESS
- BIAS / VARIANCE TRADE OFF

TODAY

- ESTIMATORS AGAIN
- EFFICENCY
- SUFFICIENCY
- CONFIDENCE INTERVALS

EFFICIENCY

- WE KNOW UNBIASED ESTIMATORS WITH ZERO VARIANCE DON'T EXIST.
- HOW SMALL/GOOD CAN BE THEIR VARIANCE ?

CRAMER RAO LOWER BOUND

- ANY UNBIASED ESTIMATOR $\hat{\theta}$ OF θ, (UNDER SOME REGULARITY CONDITIONS) :

$$\mathrm{Var}(\hat{\theta}) \geq I(\theta_0)^{-1}$$

FISHER'S INFORMATION

$$I(\theta_0) = \mathbb{E}\left[\frac{\partial}{\partial\theta} \log \mathcal{L}_n(\theta_0, \vec{X})^2\right]$$

θ_0 : TRUE PARAMETER OF STATISTICAL MODEL $\mathcal{F}$

- WE SAY $\hat{\theta}$ IS EFFICIENT OR THAT IS THE MINIMUM VARIANCE UNBIASED ESTIMATOR (M.V.U.) IFF:

$$\text{Bias}(\hat{\theta}) = 0 \quad \text{AND} \quad \text{Var}(\hat{\theta}) = I(\theta)^{-1}$$

Ex: $\hat{\theta}_{MLE} = \underset{\theta \in \Theta}{\arg\max}\, \mathcal{L}_n(\theta, \vec{x})$

IS ASSYMPTOTICALLY UNBIASED AND EFFICIENT (UNDER SOME REGULARITY CONDITIONS):

$$\sqrt{n}\left(\hat{\theta}_{MLE} - \theta\right) \xrightarrow{n \to \infty} N(0, I(\theta)^{-1})$$

CONVERGENCE IN DISTRIBUTION

UNBIASED

EFFICIENT

→ AS LONG AS WE HAVE ENOUGH SAMPLES WE SHOULD USE THE M.L.E.

SUFFICIENCY

- $\vec{X} = (X_1, \ldots, X_n) \sim f_{\theta_0} \in \mathcal{F} = \{f_\theta : \theta \in \Theta\}$

- IN SPIRIT, A STATISTIC $T(\vec{X})$ IS SUFFICIENT FOR A PARAMETER θ IF ITS USE INVOLVES NO LOSS OF INFORMATION ABOUT θ.

- MATHEMATICALLY, THIS TRANSLATES TO: FOR ANY $t \in \text{RANGE}(T)$:

$$\vec{X} \mid T(\vec{X})$$ DISTRIBUTION IS INVARIANT ON $\theta_0 \in \Theta$

Ex: $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} f_{p_0} \in \mathcal{F} = \{f_p : p \in (0,1)\}$,

$$f_p = \begin{cases} 1 & \text{WITH PROB } p \\ 0 & \text{WITH PROB } 1-p \end{cases}$$

$$T(\vec{X}) = \sum_{j=1}^{n} X_i$$ (HEADS COUNTER)

$$\mathbb{P}(\vec{X} = \vec{x} \mid T(\vec{X}) = t) = \frac{\mathbb{P}(\vec{X} = \vec{x}, T(\vec{X}) = t)}{\mathbb{P}(T(\vec{X}) = t)}$$

ASSUMING $T(\vec{x}) = t$ OTHERWISE THE PROB. IS ZERO.

$$= \frac{p_0^t (1-p_0)^{n-t}}{\binom{n}{t} p_0^t (1-p_0)^{n-t}} = \frac{1}{\binom{n}{t}}$$

$\to$ INVARIANT ON p_0!

SUFFICIENT!

MINIMAL SUFFICIENCY

- WE CAN HAVE SUFFICIENT STATISTICS THAT STORE UNNECESSARY INFORMATION ABOUT THE PARAMETER θ.

<u>Ex</u>: $T(\vec{X}) = \vec{X} \rightarrow [\vec{X} \mid T(\vec{X}) = \vec{x}] = \vec{x}$

$\rightarrow$ SUFFICIENT FOR ANY PARAMETER θ

STORES ALL THE INFORMATION OF THE SAMPLE!

- IN SPIRIT, A SUFFICIENT STATISTIC $T(\vec{X})$ FOR A PARAMETER θ IS MINIMAL SUFFICIENT IF IT STORES NO UNNECESSARY INFORMATION ABOUT θ.

- MATHEMATICALLY, THIS TRANSLATES TO

  $\forall$ SUFFICIENT STATISTIC $\tilde{T}(\vec{X})$

  $\exists$ A FUNCTION S SUCH THAT

$$S(\tilde{T}(\vec{X})) = T(\vec{X})$$

- LEHMANN-SCHEFFÉ CRITERION

CONFIDENCE INTERVALS

IN A PARAMETRIC STATISTICAL MODEL,

$$\vec{X} = (X_1, \dots, X_n) \sim f_{\theta_0} \in \mathcal{F},$$

$$\mathcal{F} = \{ f_\theta : \theta \in \Theta \}, \quad \Theta \subseteq \mathbb{R}^m$$

CONSIDER $\hat{\theta}$ FOR ESTIMATING θ, EVEN IF CONSISTENT, UNBIASED, EFFICIENT, ETC. STILL WE DON'T KNOW HOW PRECISE IS OUR ESTIMATION

SOLUTION: CONFIDENCE INTERVALS

WITH ADEQUATE STRUCTURE WE CAN FIND

$T_L(\vec{x})$, $T_U(\vec{x})$. S.T.

$$\mathbb{P}\left(T_L(\vec{x}) \leq \theta \leq T_U(\vec{x})\right) = \gamma \quad (\approx 1)$$

Ex: $X_1, \dots, X_n \overset{\text{i.i.d}}{\sim} f_{\theta_0} \in \mathcal{F}$, WHERE

$$\mathcal{F} = \{ \mathcal{N}(\mu, \sigma^2) ; \mu \in \mathbb{R}, \sigma^2 > 0 \}$$

THIS IS NOT AN STATISTIC!

FACT: $D = \dfrac{\sqrt{n}(\hat{\mu} - \mu_0)}{\hat{\sigma}^2} \sim t_{\underbrace{n-1}}$ t-STUDENT DISTRIBUTION

↳ DEGREES OF FREEDOM

→ FOR ANY γ (≈ 0) WE CAN FIND $p \in \mathbb{R}$ SUCH THAT

$$\mathbb{P}(-p \leq D \leq p) \leq \gamma$$

$$\rightarrow \mathbb{P}\Big(\underbrace{\hat{\mu} - \frac{p\hat{\sigma}^2}{\sqrt{n}}}_{T_L(\vec{x})} \leq \mu_0 \leq \underbrace{\hat{\mu} + \frac{p\hat{\sigma}^2}{\sqrt{n}}}_{T_U(\vec{x})} \Big) = \gamma$$

→ WITH PROB. γ: $\mu_0 \in [T_L(\vec{x}), T_U(\vec{x})]$

- γ LARGER → LARGER INTERVAL
- n LARGER → SMALLER INTERVAL!

LECTURE 4

LAST TIME

- EFFICIENCY
- SUFFICIENCY
- CONFIDENCE INTERVALS

NOW

- HYPOTHESIS TESTING
- ERRORS (TYPE I & II)
- LIKELIHOOD RATIO TEST

TESTING HYPOTHESIS

- MODERN SCIENCE ASSUMPTION (POPPER)
  HYPOTHESIS CAN NOT BE PROVEN TRUE
  BUT CAN BE FALSIABLE: WE CAN REJECT
  THEM IF EXPERIENCE DON'T MATCH IT.

- MANY POSSIBLE STATISTICAL APPROACHES
  - BAYESIAN
  - FREQUENTIST
    - FISHER
    - $\underbrace{\text{NEYMAN \& PEARSON}}_{\text{MOST WIDESPREAD AND FAMOUS}}$

NEYMAN & PERSON FREQUENTIST TEST

H_0 : NULL HYPOTHESIS (TO BE TESTED)

H_A : ALTERNATIVE HYPOTHESIS (H_0 NEGATION)

TO TEST H_0 :

0) COLECT A SAMPLE $\vec{x} = (x_1, \ldots, x_n)$

1) ASSUME A STATISTICAL MODEL $\mathcal{F}$ FOR $\vec{x}$

WE SHOULD BE EXPRESS H_0 IN TERMS OF A SUBSET $\mathcal{F}_0 \subseteq \mathcal{F}$

2) SELECT A STATISTIC $T_f(\vec{X})$, FOR ANY $f \in \mathcal{F}_0$

WE ASSUME WE KNOW f WHEN COMPUTING IT.

LARGER VALUES OF $T_f(\vec{x})$ SHOULD REFLECT EVIDENCE AGAINST H_0.

3) SELECT A SIGNIFICANCE LEVEL $\alpha \approx 0$

4) COMPUTE THE P-VALUE:

$$p(\vec{x}) = \underbrace{\sup_{f \in \mathcal{F}_0}}_{\text{BEST CASE OVER } H_0} \mathbb{P}_f\Big(\underbrace{T_f(\vec{X})}_{\text{RANDOM VARIABLE}} \geq \underbrace{T_f(\vec{x})}_{\text{CONCRETE OBSERVATION}}\Big)$$

UNDER H_0, THE PROBABILITY OF T BEING MORE DEVIATED THAN OUR OBSERVATION

$p(\vec{x}) \approx 0 \rightarrow$ OUR OBSERVATION IS DEVIATED UNDER H_0

$\rightarrow H_0$ IS UNLIKELY TO FIT OUR EXPERIENCE

5) REJECT H_0 IFF $p(\vec{x}) \leq \alpha$.

Ex: $X_1, \dots, X_n \overset{i.i.d}{\sim} f_{\bar\theta} \in \mathcal{F}$, $\bar\theta \to (\bar\mu, \bar\sigma^2)$

1) $\mathcal{F} = \{\mathcal{N}(\mu, \sigma^2) ; \mu \in \mathbb{R}, \sigma^2 > 0\}$

$H_0 : \bar\mu \leq c \leftrightarrow \mathcal{F}_0 = \{\mathcal{N}(\mu, \sigma^2) : \mu \leq c, \sigma^2 > 0\}$

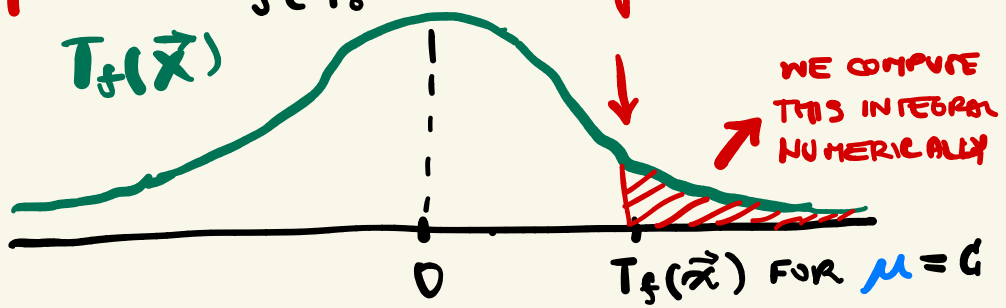
$H_a : \bar\mu > c$

2) FOR ANY $f \in \mathcal{F}_0$, WITH PARAMETERS (μ, σ^2), CONSIDER THE STATISTIC

$$T_f(\vec{x}) = \frac{\sqrt{n}(\hat\mu - \mu)}{\hat\sigma} \underset{\text{ASSUMING } \vec{X} \sim f}{\sim} t_{n-1}$$

3) CHOOSE $\alpha = 0.05 \approx 0$.

4) $p(\vec{x}) = \sup_{f \in \mathcal{F}_0} \underbrace{\mathbb{P}_f(T_f(\vec{X}) \geq T_f(\vec{x}))}$



FOR ANY OTHER $f \in \mathcal{F}_0$ WITH $\mu < c$, $T_f(\vec{x})$ IS TRANSLATED TO THE RIGHT IN THE ABOVE PLOT.

$\rightarrow$ THE AREA UNDER THE CURVE IS NOT LARGER!

$\rightarrow$ THE SUPREMUM IS ATTAINED FOR $\mu = c$.

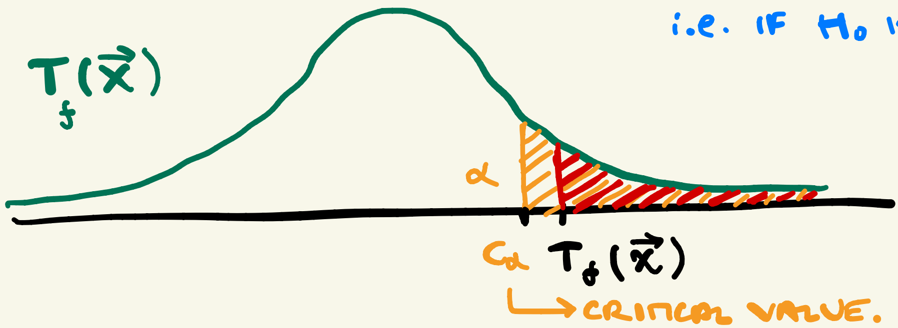
5) REJECT H_0 IFF $p(\vec{x}) \leq \alpha = 0.05$.

TESTING ERRORS

TYPE I: REJECTING H_0 WHEN IT IS TRUE.

$$\mathbb{P}(\text{TYPE I ERROR}) \leq \alpha$$

"=" IF $\mathcal{F}_0 = \{f_0\}$
i.e. IF H_0 IS SIMPLE



$$p(\vec{x}) \leq \alpha \longleftrightarrow T_f(\vec{x}) \geq c_\alpha$$

- BY CONSTRUCTION, TYPE I ERROR IS LOW!

→ THE TEST IS "PLAYING FAIR" WITH H_0.

TYPE II: NOT REJECTING H_0 WHEN IT IS FALSE.

$$\mathbb{P}(\text{TYPE II ERROR}) =: \beta$$

"POWER" OF TEST $:= 1 - \beta$

- β QUANTIFIES HOW "SEVERE" IS THE TEST WITH H_0.
- β IS NOT SET BY THE SCIENTIST
  IT IS A PROPERTY OF ANY TEST WITH FIXED α.

LIKELIHOOD RATIO TEST

- CONSIDER A STATISTICAL MODEL $\mathcal{F}$ AND A NULL HYPOTHESIS $\mathcal{F}_0 \subseteq \mathcal{F}$.

- THE LIKELIHOOD RATIO TEST IS DEFINED BY ITS STATISTIC

$$T(\vec{x}) = -2\ln\left[\frac{\sup\{f(\vec{x}) : f \in \mathcal{F}_0\}}{\sup\{f(\vec{x}) : f \in \mathcal{F}\}}\right]$$

- **NEYMAN PEARSON LEMMA**

IF $\mathcal{F} = \{f_0, f_A\}$, $\mathcal{F}_0 = \{f_0\}$

(SIMPLE NULL AND SIMPLE ALTERNATIVE)

THEN, FOR ANY FIXED SIGNIFICANCE α, THE LIKELIHOOD RATIO TEST IS THE MOST POWERFUL TEST WITH SIGNIFICANCE α.

LECTURE 5

LAST TIME

- TESTING (NEYMAN-PEARSON)

TODAY

- MORE TESTING
- FINITE SAMPLE TEST
- ASSYMPTOTIC TESTS
- LIKELIHOOD RATIO TEST

FINITE SAMPLE TESTS

- WITH ADEQUATE STRUCTURE, WE KNOW EXACTLY THE DISTRIBUTION OF THE TEST STATISTIC $T(\vec{x})$

→ THEN WE CAN EXACTLY* COMPUTE THE p-VALUE OF TEST FOR ANY FINITE n.

*UP TO NUMERICAL ERRORS

Ex: ONE SAMPLE TESTING

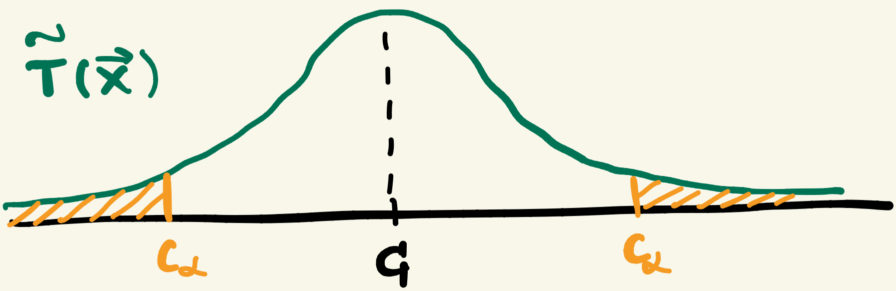
- $X_1, \dots, X_n \overset{i.i.d}{\sim} f_{\bar\theta} \in \mathcal{F}_X$, WHERE

  $\mathcal{F}_X = \{ N(\mu, \sigma^2);\ \mu \in \mathbb{R},\ \sigma^2 > 0 \}$

- $H_0: \bar\mu = c$, $H_A: \bar\mu \neq c$

- $T(\vec{x}) = |\tilde{T}(\vec{x})|$, $\tilde{T}(\vec{x}) = \dfrac{\overbrace{\sqrt{n}\,(\hat\mu - c)}^{\sim t_{n-1},\ \text{B.C. } \mathcal{F}_X}}{\hat\sigma}$

- FOR ANY α, $c_\alpha = t_{n-1}(\alpha/2)$.



- REJECT IFF $T(\vec{x}) \geq c_\alpha$

Ex: TWO SAMPLES TESTING.

- 1ST SAMPLE: $X_1, \dots, X_n \overset{i.i.d}{\sim} f_{\theta_X} \in \mathcal{F}_N$,

  2nd SAMPLE: $Y_1, \dots, Y_n \overset{i.i.d}{\sim} f_{\theta_Y} \in \mathcal{F}_N$

  (TWO SAMPLES, EACH FROM A DIFFERENT DISTRIBUTION OF THE SAME STATISTICAL MODEL. ASSUME SAME VARIANCE)
  $\sigma_X = \sigma_Y$

- $H_0: \mu_X = \mu_Y$, $H_A: \mu_X \neq \mu_Y$.

- $T(\vec{X}, \vec{Y}) = \left| \underbrace{\dfrac{\hat{\mu}_X - \hat{\mu}_Y}{\sqrt{\hat{\sigma}_p^2 \left(\frac{1}{n} + \frac{1}{m}\right)}}}_{\sim t_{n+m-2}} \right|$, WHERE

  BECAUSE OF $\mathcal{F}_N$

$$\hat{\sigma}_p^2 = \frac{(n-1)\hat{\sigma}_X^2 + (m-1)\hat{\sigma}_Y^2}{n+m-2}$$

(POOLED t-TEST FOR EQUAL VARIANCES)

- FOR ANY α, $c_\alpha = t_{n-1}(\alpha/2)$

- REJECT IFF $T(\vec{x}, \vec{y}) \geq c_\alpha$

"LARGE SAMPLE" TEST
(ASSYMPTOTIC RESULT)

- ASSUME $X_1, \ldots, X_n \overset{i.i.d}{\sim} f_0 \in \mathcal{F}$ ↗ ANYTHING
- WE DON'T KNOW THE DISTRIBUTION OF ANY STATISTIC WE CAN THINK OF.
- WE CAN CONSIDER THE STATISTIC

$$T_f(\vec{x}) = \frac{\sqrt{n}\,(\hat{\mu} - \mu)}{\sigma} \xrightarrow[\text{CONVERGENCE IN DISTRIBUTION, BY C.L.T.}]{n \to \infty} \mathcal{N}(0,1)$$

WHERE $\mu = \mathbb{E}_f(X_1)$, $\sigma = \sqrt{\mathrm{Var}_f(X_1)}$

- WE KNOW THE ASSYMPTOTIC DISTRIBUTION OF $T_f(\vec{x})$.
- OTHER DISTRIBUTIONS THAT ARISE FROM ASSYMPTOTICS: χ^2, RATIOS OF χ^2s.

Ex: ONE SAMPLE TESTING OF PROPORTIONS.

- $X_1, \dots, X_n \overset{i.i.d}{\sim} f_{\bar{p}} \in \mathcal{F}_p$, WHERE

$$\mathcal{F}_p = \{f_p : p \in (0,1)\}, \quad f_p = \begin{cases} 1 & \text{w.p. } p \\ 0 & \text{w.p. } 1-p \end{cases}$$

- $H_0 : \bar{p} = p_0$, $\quad H_A : \bar{p} \neq p_0$

- $T_f(\vec{x}) = |\tilde{T}_f(\vec{x})|$

$$\tilde{T}_f(\vec{x}) = \frac{\sqrt{n}(\hat{p} - p_0)}{\sqrt{p_0(1-p_0)}} \overset{n\to\infty}{\sim} \mathcal{N}(0,1)$$

$(\hat{p} = \hat{\mu})$

- FOR ANY α, $c_\alpha = z(\alpha/2)$

- REJECT IFF $T_f(\vec{x}) \geq c_\alpha$

Ex: TWO SAMPLE PROPORTIONS.

- $X_1, \ldots, X_n \overset{i.i.d.}{\sim} f_{p_X} \in \mathcal{F}_P$
  $Y_1, \ldots, Y_n \overset{i.i.d.}{\sim} f_{p_Y} \in \mathcal{F}_P$

- $H_0: p_X = p_Y$, $H_A: p_X \neq p_Y$

- $T(\vec{x}, \vec{y}) = |\tilde{T}(\vec{x}, \vec{y})|$ (WALD STATISTIC WITH POOLING)

$$\tilde{T}(\vec{x}, \vec{y}) = \frac{\hat{p}_X - \hat{p}_Y}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n} + \frac{1}{m}\right)}} \overset{n \to \infty}{\sim} N(0,1)$$

- FOR ANY α, $c_\alpha = z(\alpha/2)$

- REJECT IFF $T(\vec{x}, \vec{y}) \geq c_\alpha$

(NUMERICAL METHODS)

LECTURE 6

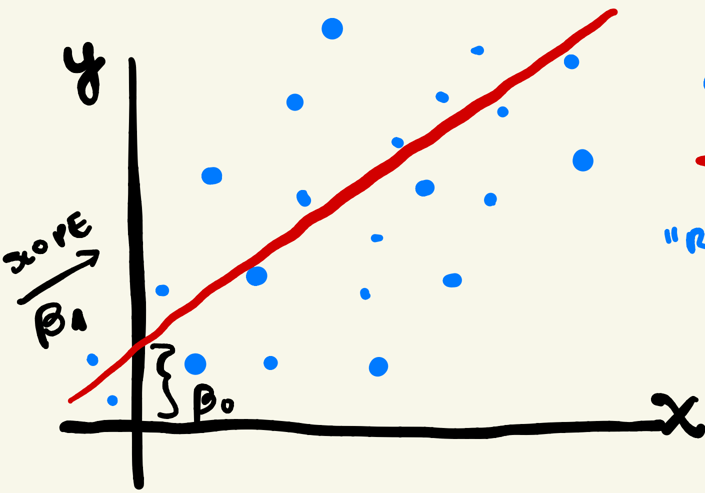
LAST TIME

- NEYMAN-PEARSON TESTING
- FINITE SAMPLE
- ASSYMPTOTICS

NOW

- LINEAR REGRESSION
- MODEL
- ESTIMATION
- FURTHER TOPICS

WHAT IS LINEAR REGRESSION?



- DATA
- LINEAR REGRESSION

"RESPONSE" "PREDICTOR"

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon$$

"INTERCEPT"

GEOMETRICALLY: "<u>BEST</u>" LINEAR APROXIMATION TO OUR DATA

STATISCALLY?

STATISTICAL MODEL

INDICES

- $x^1, \dots, x^n \in \mathbb{R}^d$ (PREDICTOR)
- $y_1, \dots, y_n \in \mathbb{R}$ (RESPONSE)

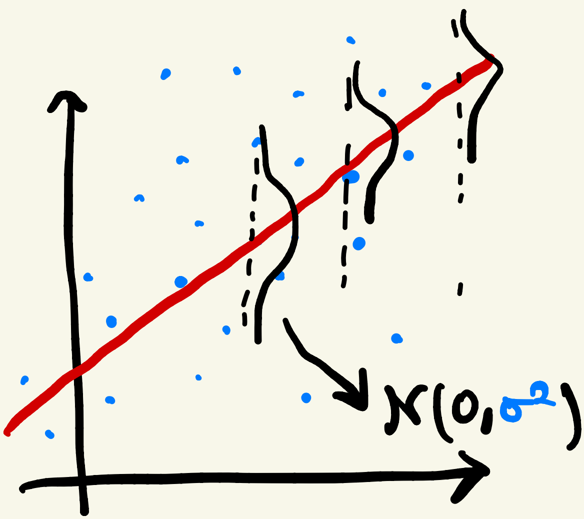
MAIN ASSUMPTION

$$y_i = \beta_0 + \sum_{k=1}^{d} x^i_k \beta_k + \varepsilon_i \quad \forall i \in \{1, \dots, n\}$$

WHERE $\varepsilon_1, \dots, \varepsilon_n \overset{i.i.d.}{\sim} N(0, \sigma^2) \rightarrow$ THIS CAN BE RELAXED!

STATISTICAL MODEL COMPACT FORM

- $\vec{Y} \sim f_\theta \in \mathcal{F} = \{ f_\theta : \theta = (\beta, \sigma^2) \in \mathbb{R}^{d+1} \times \mathbb{R}_{++} \}$
- $f_\theta = X\beta + \varepsilon$
- $\varepsilon \sim N(0, \sigma^2 I_d)$
- $X = \underset{\in\, \mathbb{R}^{n \times (d+1)}}{\begin{bmatrix} 1 & x^{1T} \\ \vdots & \vdots \\ 1 & x^{nT} \end{bmatrix}}$



PARAMETER ESTIMATION

- WE FOCUS ON THE β PARAMETER. CONSIDER

$$T(X,y) = \underset{\beta \in \mathbb{R}^d}{\operatorname{argmin}} \frac{1}{2} \| X\beta - y \|_2^2$$

OPTIMAL LEAST SQUARES ESTIMATOR (O.L.S)

LET $h(\beta) = \frac{1}{2} \| X\beta - y \|_2^2$

$$\nabla h(\beta) = \frac{2}{2} X^T (X\beta - y)$$

OPTIMALITY CONDITIONS FOR CONVEX FUCTIONS GIVE

$$\nabla h(\hat{\beta}) = 0 \leftarrow \hat{\beta} = (X^T X)^{-1} X^T y$$

OBS: $X^T X$ IS INVERTIBLE IFF THE COLUMNS OF X ARE LINEARLY INDEPENDENT.

IF THEY AREN'T, REPLACE INVERSE WITH PSEUDOINVERSE. (MATRIX ANALYSIS)

$$\boxed{\hat{\beta} = (X^T X)^{\dagger} X^T y}$$

LINEAR ESTIMATOR W.R.T y!

LEAST SQUARES IS BLUE

BEST ✓
LINEAR ✓
UNBIASED ✓
ESTIMATOR

- $\mathbb{E}(\hat{\beta}) = \mathbb{E}((X^T X)^{-1} X^T \vec{Y})$

↳ SOURCE OF RANDOMNESS

$$= \mathbb{E}((X^T X)^{-1} X^T (X\beta + \varepsilon))$$

$$= \mathbb{E}((X^T X)^{-1} (X^T X)\beta) + \mathbb{E}((X^T X)^{-1} X^T \varepsilon)$$

LINEARITY OF EXPECTATION

$$= \mathbb{E}(\beta) + (X^T X)^{-1} X^T \cancelto{0}{\mathbb{E}(\varepsilon)}$$

$$= \beta \longrightarrow \text{UNBIASED ESTIMATOR}$$

GAUSS MARKOV THEOREM

ANY ESTIMATOR $\tilde{\beta}$ THAT IS A LINEAR FUNCTION OF y AND IS UNBIASED HAS LARGER VARIANCE THAN $\hat{\beta}$:

$$\underbrace{\mathbb{E}(\tilde{\beta}\tilde{\beta}^T)}_{\text{COVARIANCE MATRIX}} \succeq \mathbb{E}(\hat{\beta}\hat{\beta}^T) = I_d \sigma^2$$

SUMMARY

WE HAVE SEEN

- SAMPLES & STATISTICAL MODEL
- STATISTICS & ESTIMATORS
- CONFIDENCE INTERVALS & TESTING
- LINEAR REGRESSION

MORE OF THIS IN AMS :

- MATHEMATICAL STATISTICS (FOUNDATIONS) [FALL]
- STATISTICAL THEORY I (DISCUSSION) [FALL]
- STATISTICAL THEORY II (ASYMPTOTICS) [SPRING]

WHAT WE DIDN'T COVER

- BAYESIAN STATISTICS [FALL] / SPRING
- STATISTICAL LEARNING
  - STATISTICAL PATTERN RECOGNITION [SPRING]
  - MACHINE LEARNING I [FALL]
  - MACHINE LEARNING II [SPRING]
- NON PARAMETRIC STATISTICS
  - APPLIED STATISTICS AND DATA ANALYSIS II [S]
- HIGH DIMENSIONAL STATISTICS/PROBABILITY
  - MATHEMATICS OF DATA SCIENCE [FALL]
  - HIGH DIM. PROB. & APPROXIMATION [SPRING]