AMS 553.430 - Introduction to Statistics

Lectures by Dwijavanti P Athreya
Notes by Kaushik Srinivasan

Johns Hopkins University
Fall 2018

Introduction

Math 553.430 is one of the most important courses that is required/recommended for the engineering-based majors at Johns Hopkins University.

These notes are being live-TeXed, through I edot for Typos and add diagrams requiring the Ti*k*Z package separately. I am using Texpad on Mac OS X.

I would like to thank Zev Chonoles from The University of Chicago and Max Wang from Harvard University for providing me with the inspiration to start live-TeXing my notes. They also provided me the starting template for this, which can be found on their personal websites.

Please email any corrections or suggestions to ksriniv4@jhu.edu.

Lecture 0 (2018-08-30)

Introduction to Probability (553.420) Review

Part 1 - Counting

1. Multiplication rule (Basic Counting Principle)
2. Combinations/Permutations
   - Sampling with or without replacement. $\Rightarrow$ Inclusion-Exclusion Principle

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \qquad {}^nP_k = \frac{n!}{(n-k)!}$$

3. Birthday Problem
4. Matching Problem (inclusion-exclusion principle)
   - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
   - $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$
   - etc...
5. n balls going into m boxes (all are distinguishable)

   Example. n balls numbered $1, 2, \cdots, n$. n boxes labelled $1, 2, \cdots, n$. Distribute the balls into the boxes, one in each box. M_i = ball i is in box i
6. Multinomial Coefficients e.g. assign A, B, C, D, to different students $\rightarrow$ anagram problem – n distinct objects into r distinct groups

$$\frac{n!}{n_1!n_2!n_3!\dots n_r!} = \binom{n}{n_1, n_2, n_3, \dots, n_r}$$

7. Pairing Problem

$$2n \text{ people, paired up} \begin{cases} \text{ordered: } \binom{2n}{2,2,\cdots,2} & \text{e.g. different courts for players} \\ \\ \text{unordered: } \frac{\binom{2n}{2,2,\cdots,2}}{n!} \end{cases}$$

8. Partition of integers $\longrightarrow n$: sum of integer, r: number of partitions

$$\binom{n+r-1}{r-1} = \binom{n+r-1}{n}$$

Basics of Probability

Axioms

1 $0 \leq P(A) \leq 1, \forall A$

2 $P(\Omega) = 1 \rightarrow$ where Ω is the sample space

3 Countable additivity

- if $A_1, \cdots, A_n$ are mutually exclusive, then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = P(A_1) + P(A_2) + \cdots = \sum_{i=1}^{\infty} P(A_i)$$

$\Rightarrow P(A) = 1 - P(A^c)$
$P(A) = \dfrac{|A|}{|\Omega|}$

Conditional Probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Law of Total Probability

$$P(A) = \sum_j P(A|B_j)P(B_j) = \sum P(A \cap B_j) \qquad \underbrace{\bigcup_j B_j = \Omega}_{\text{partition of } \Omega}$$

Bayes Rule

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_j P(A|B_j)P(B_j)} \qquad \underbrace{\bigcup_j B_j = \Omega}_{\text{partition of } \Omega}$$

Independent events

If we have events $A_1, A_2, \cdots, A_n$, then

$$P(A_1 \cap A_2 \cap A_3 \cdots A_n) = P(A_1) \cdot P(A_2) \cdot P(A_3) \cdot \cdots P(A_n)$$

Introduction to Discrete and Continuous Random Variables

Random Variable - a real valued function defined on the sample space of an experiment $X : \Omega \rightarrow \mathbb{R}, \forall \omega \in \Omega, X(\omega) \in \mathbb{R}$

Function	Discrete	Continuous
Probability Function	PMF: $P(X=x)$	PDF: $f_x(x)$
Probability Distribution	$\sum_x P(X=x)=1$	$\int_x f_x(x)dx=1$
Expectation	$E[X]=\sum_x xP(X=x)$	E[X] $=\int_x xf(x)dx$
Variance	$Var[X]=E[X^2]-(E[X])^2$	$Var[X]=E[X^2]-(E[X])^2$

Law of the Unconscious Statistician (LOTUS)

1-dim $\quad E[g(x)]=\sum_x g(x)P(X=x) \Big/ E[g(x)]=\int_x g(x)f(x)dx$

2-dim $\quad E[g(X,Y)]=\sum_y\sum_x g(x,y)P(X=x,Y=y) \Big/ E[g(X,Y)]=\int_y\int_x g(x,y)f(x,y)dxdy$

Discrete Distributions

1. Bernoulli(p)
2. Binomial(n,p)
3. Poisson (λ)
4. Geometric(p)
5. Negative Binomial(n,p)
6. Hypergeometric(N,M,n)

Bernoulli Distribution

X is a random variable with Bernoulli(p) distribution

$$X \sim Bernoulli(p)$$
$$P(X=x)=\begin{cases} p & x=1 \\ 1-p & x=0 \end{cases}$$

Binomial Distribution

A sum of i.i.d. (identical, independent distribution) Bernoulli(p) R.V.

$$X \sim Binomial(n,p)$$
$$\text{Support} : x \in \{0,1,\cdots n\}$$
$$n : \text{sample size} \qquad p : \text{probability of success}$$
$$P(X=k)=\binom{n}{k}p^k(1-p)^{(n-k)}$$
$$E[X]=np \qquad\qquad Var(X)=np(1-p)$$

- Approximation methods $\Rightarrow$

- if n is large, p very small and $np < 10$. $\Rightarrow$ use Normal $(np, np(1-p))$
- $p \approx \frac{1}{2} \Rightarrow$ Use Poisson $(\lambda = np)$

- Mode:
  - if $(n+1)p$ integer, mode = (n+1)p or (n+1)p - 1.
  - if $(n+1)p \notin \mathbb{Z}$ mode is $\lfloor (n+1)p \rfloor$
  - **Proof:** consider $\dfrac{P(X=x)}{P(X=x-1)}$ going below 1.

Poisson Distribution

$$X \sim Poisson(\lambda)$$
$$x \in \{0, 1, \cdots\}$$
$$\lambda : \text{parameter}$$
$$P(X=x) = \frac{e^{-\lambda}\lambda^x}{x!}$$
$$E[X] = \lambda \qquad\qquad Var(X) = \lambda$$

- Approximations
  - if n is large $\Longrightarrow$ Normal(λ, λ)
- Sums of Poisson
  Let $X \sim Po(\lambda) \quad Y \sim Po(\mu) \quad \Longrightarrow \quad X + Y \sim Po(\mu + \lambda)$

Negative Binomial

$$X \backsim NB(r, p)$$
$$\text{Support} : x = \{r, r+1, \ldots\}$$
$$r = \text{the rth success}$$
$$p = \text{probability of success}$$
$$P(X=k) = \binom{k+r-1}{k} \cdot (1-p)^r \cdot p^k$$

A sum of i.i.d Geometric(p) R.V.
■ a^{th} head before b^{th} tail

Example. A coin has probability p to land on a head, $q = 1 - p$ to land on a tail.

$P[5^{th}$tail occurs before the 10^{th} head]?

$$\begin{cases} = P[\text{5th tail occurs before or on the 14th flip}] \\ = P[\text{Neg Binomial}(5, q) = 5, 6, 7, \cdots, 14] \\ = \sum_{x=5}^{14} \binom{x-1}{4} q^5 p^{x-5} \end{cases} \quad \text{(or)} \quad \begin{cases} = P[\text{at least 5 tails in 14 flips}] \\ = P[binom(14, q) = 5, 6, 7, \cdots, 14] \\ = \sum_{x=5}^{14} \binom{14}{x} q^x p^{14-x} \end{cases}$$

Geometric Distribution

$$X \sim Geometric(p)$$
$$\text{Support} : x \in \{1, 2, \cdots\}$$
$$p : \text{probability of success}$$
$$P(X = r) = (1-p)^{(r-1)} \cdot p$$
$$\text{prob for 1st success on } r\text{th trial}$$
$$E[X] = \frac{1}{p} \qquad Var(X) = \frac{1-p}{p^2}$$

Example. ■ Coupon Question

Variation A: N different types of coupons $\rightarrow P($ get a specific type$) = \frac{1}{N}$

Question: E[draws to get 10 different coupons]?

Answer:

$$X = X_1 + X_2 + \cdots + X_{10} \qquad X_i = \#\text{ draws to get the ith distinct coupon type}$$

$\boxed{X_i \backsim Geo(p_i)}$ $\quad p_i$: prob to get a new coupon $\leftarrow$ success, given that we have $i-1$ types of coupons

$$\text{Hence, } E[X_1] = 1$$
$$E[X_2] = \frac{1}{p_2} = \frac{1}{\frac{N-1}{N}} = \frac{N}{N-1}$$
$$E[X_3] = \frac{1}{p_3} = \frac{1}{\frac{N-2}{N}} = \frac{N}{N-2}$$
$$\vdots$$
$$E[X_{10}] = \frac{1}{p_{10}} = \frac{1}{\frac{N-9}{N}} = \frac{N}{N-9}$$

So, $$E[X] = E[X_1] + E[X_2] + \cdots + E[X_{10}] = E[\sum_{i=1}^{10} X_i] = 1 + \frac{N}{N-1} + \frac{N}{N-2} + \cdots + \frac{N}{N-9}$$

Variation B: Same setting, now you draw 10 times.

Question: E[# different types of coupons]?

Answer:

$$X = I_1 + I_2 + \cdots + I_N$$
$$I_i \begin{cases} 1 & \text{if we have this type of coupon} \\ 0 & \text{o/w} \end{cases}$$

$$
\begin{aligned}
E[I_i] &= P(\text{we draw coupon i in 10 draws}) \\
&= 1 - P(\text{we don't have coupon i}) \qquad \text{we use binomial distribution where } 1 - P(N=0) \\
&= 1 - \left(\frac{N-1}{N}\right)^{10}
\end{aligned}
$$

$$E[X] = E[\sum_{i=1}^{N} I_i] = NE[I_i] = \boxed{N\left[1 - \left(\frac{N-1}{N}\right)^{10}\right]}$$

Hypergeometric Distribution

$$X \sim Hyp(N, M, n)$$

$$N \in \{0, 1, 2, \ldots\} \quad M \in \{0, 1, \ldots, N\} \quad n \in \{0, 1, \ldots, N\}$$

$$\text{Support} : k \in \{\max(0, n + M - N), min(n, M)\}$$

N is the population size $\quad K$ is the no. of success states in the population

n is the no. of draws (i.e. quantity drawn in each trial)

k is the no. of observed successes

$$P(X = k) = \frac{\binom{M}{k}\binom{N-M}{M-k-1}}{\binom{N}{n}}$$

Continuous Distributions

Uniform Distribution

$$X \sim Unif(a, b)$$

$$f_X(x) = \begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0 & o/w \end{cases}$$

$$E[X] = \frac{a+b}{2} \qquad Var(X) = \frac{(b-a)^2}{12}$$

Normal Distribution

$$X \backsim N(\mu, \sigma^2) \Rightarrow Z = \frac{X-\mu}{\sigma} \backsim N(0,1) \text{ with CDF } P(Z \leq z) = \Phi(z)$$
$$\Phi(-x) = 1 - \Phi(x)$$
$$\text{Support:} \quad x \in (-\infty, \infty)$$
$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$
$$E[X] = \mu \qquad\qquad Var(X) = \sigma^2$$

- Sums and differences of Normal R.V.

$$\begin{aligned} X_1 &\sim \mathcal{N}(\mu, \sigma^2) & X_2 &\sim \mathcal{N}(\mu, \sigma^2) \\ Y_1 &= X_1 + X_2 & Y_2 &= X_1 - X_2 \\ &\underbrace{Y_1 \sim \mathcal{N}(2\mu, 2\sigma^2)}_{\text{has } \mu} & &\underbrace{Y_2 \sim \mathcal{N}(0, 2\sigma^2)}_{\text{doesn't have } \mu} \end{aligned}$$

  - The sum and difference of Normal R.V. are Normal R.V.
  - Any Linear Combination of Independent Normal R.V. is a Normal R.V.

- Dependence
  - $Y_2 = X_1 - X_2$ density does not depend on μ. But density of $X_1 + X_2$ does.
  - Key idea is used in Data Reduction

Exponential distribution

$$X \sim Exp(\lambda)$$
$$\text{Support:} \quad x \in [0, \infty)$$
$$f_X(x) = \lambda e^{-\lambda x}$$
$$E[X] = \frac{1}{\lambda} \qquad\qquad Var(X) = \frac{1}{\lambda^2}$$

Lack of memory property: $P(X \geq s + t | X \geq t) = P(X \geq s)$

- M = min of $exp(\lambda)$ and $exp(\mu) \Rightarrow M \backsim exp(\lambda + \mu)$
- M = min of $X_1, X_2, \cdots, X_n$, where $X_i \backsim_{\text{i.i.d.}} exp(\lambda) \Rightarrow exp(n\lambda)$

Gamma Distribution

$$X \sim Gamma(\alpha, \beta)$$
$$\text{Support:} \quad x \in [0, \infty)$$
$$F_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$
$$E[X] = \frac{\alpha}{\beta} \qquad Var(X) = \frac{\alpha}{\beta^2}$$
$$\textbf{Gamma Function:} \quad \Gamma(z) = (z-1)! = \int_0^\infty x^{z-1} e^{-x} dx$$
$$\Gamma(n) = (n-1)!$$
$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$$

- Sums of Gamma
  - $Gamma(s, \lambda) \underset{\text{ind}}{+} Gamma(s, \lambda) = Gamma(s + t, \lambda)$

Beta Distribution

$$X \sim Beta(\alpha, \beta)$$
$$\text{Support:} \quad x \in [0, 1]$$
$$f_X(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}$$
$$E[X] = \frac{\alpha}{\alpha + \beta} \qquad Var(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

- **Gamma to Beta**

$$X \sim Gamma(\alpha_1, \beta) \qquad Y \sim Gamma(\alpha_2, \beta)$$
$$\text{Then transformation} \quad U = \frac{X}{X+Y} \sim Beta(\alpha_1, \alpha_2) \qquad (\text{Use } X = UV,\ Y = V - UV)$$

Chi-Square

$$\textbf{Chi-Square:} \ \chi_n^2 \text{is Chi-square with degrees of Freedom } n$$
$$\chi_n^2 = Z_1^2 + Z_2^2 + \cdots + Z_n^2 \qquad \text{where } Z_i \backsim \text{standard normal.} Z_i \backsim Gamma\left(\frac{1}{2}, \frac{1}{2}\right)$$
$$\Rightarrow \chi_n^2 = n \text{ i.i.d. } Z_i \backsim Gamma\left(\frac{1}{2}, \frac{1}{2}\right)$$

$$= Gamma\left(\frac{n}{2}, \frac{1}{2}\right)$$

CDF in General

- $F_x(t) = P(X \leq t)$

$$= \sum_{x \leq t} P(X = x) \qquad \text{discrete}$$
$$= \int_{-\infty}^{t} f(x)dx \qquad \text{continuous}$$

- **Discrete:** "Left open, right closed" $\Rightarrow$ if you flip the sign (from $<$ to $\leq$) in the left, you flip the sign of a (from a to a^-)
  - $P(a < x \leq b) = F(b) - F(a)$
  - $P(a \leq x \leq b) = F(b) - F(a^-)$
  - $P(a < x < b) = F(b^-) - F(a)$
  - $P(a \leq x < b) = F(b^-) - F(a^-)$
- **Continuous:** (because a point doesn't have a mass)

$$P(a \leq x \leq B) = \int_a^b f(x)dx = F(b) - F(a)$$

Integration by Recognition

$$1 = \int_{-\infty}^{\infty} \frac{e^{-\frac{x^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}} dx \qquad \sigma\sqrt{2\pi} = \int_{-\infty}^{\infty} e^{-\frac{x^2}{2\sigma^2}} \qquad \text{(normal dist.)}$$

Joint Distribution

Discrete

$$P_{X,Y}(x,y) = P(X = x, Y = y)$$
$$\text{Indep} \Rightarrow P_X(x)P_Y(y)$$

Continuous

$$F_{X,Y}(x,y) = f_X(x)f_Y(y)$$
$$= \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y)$$

- **Marginal Density/PMF**:

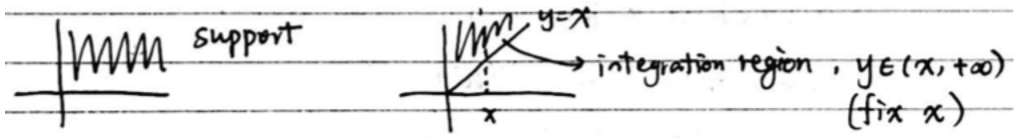
Continuous: $f_X(x) = \int_x f_{X,Y}(x,y)dy \quad and \quad f_Y(y) = \int_y f_{X,Y}(x,y)dx$

** the bounds for y in the integration can depend on x, and vice versa*

Discrete: $P_X(x) = \sum_y P(X = x, Y = y) \quad and \quad P_Y(y) = \sum_x P(X = x, Y = y)$

- Use joint pdf to compute probability

$$\text{e.g. } P(X<Y) = \int_0^\infty \int_x^\infty f(x,y)\,dy\,dx \qquad \text{assume } x>0, y>0$$



- **Independence:** If X, Y are independent, then

$$\textbf{Continuous:} \qquad f(x,y) = f_X(x) f_Y(y)$$
$$\textbf{Discrete:} \qquad P(X=x, Y=y) = P(X=x)P(Y=y)$$

- **Convolution:** assume X, Y are independent

$$\textbf{Discrete:} \qquad P_{X+Y}(a) = \sum_y P_X(a-y)P_Y(y) = \sum_x P_X(x)P_Y(a-x)$$
$$\textbf{Continuous:} \qquad f_{X+Y}(a) = \int_y f_X(a-y) f_Y(y)\,dy = \int_y f_X(x) f_Y(a-x)\,dx$$

MGF: we can use this $\quad M_{X+Y}(t) = M_X(t)M_Y(t) \longrightarrow$ then identify dist of X+Y from mgf

- **Density Transformation:**

$$\begin{tikzcd}
F_{X,Y} \arrow[r, "\text{diff}"] & f_{X,Y} \arrow[d, "\text{substitute}"] \arrow[dr, dotted, "\text{algebra}"] & \\
 & F_{g(X,Y)} \arrow[d, "\frac{d}{dt}"] & f_{g(X,Y)} \arrow[dl, dotted, bend left, "\text{integrate}"] \\
 & f_{g(X,Y)} &
\end{tikzcd}$$

X_1 & X_2 are indep r.v. $\Rightarrow$ want to find density of $\frac{X_1}{X_2}$

$Y_1 = \frac{X_1}{X_2} \qquad$ choose Y_2 to make work easier

$$\begin{tikzcd}
f_{X_1} \arrow[dr] & & f_{X_2} \arrow[dl] \\
 & f_{X_1,X_2} \arrow[d, "\text{dens. transf.}"] & \\
 & f_{\frac{X_1}{X_2},???} \arrow[d] & \\
 & f_{\frac{X_1}{X_2}} &
\end{tikzcd}$$

Density Transformation

For density transformation **e.g.** finding pdf of $U = X + Y$

– Convolution

– MGF

– Jacobian

– CDF Transformation

- Use CDF: Computer $P(Y \leq y) = P(g(x) = y)$
- **1-dim:** If Y is monotonically increasing or decreasing: $Y = g(x)$ $\boxed{f_Y(y) = f_X(x(y)) \cdot |(x^{-1})'(y)|}$
- **2-dim:** Joint Density:

$$(X,Y) \to (U,V) \qquad U = h_1(X,Y) \qquad V = h_2(X,Y)$$
$$f_{U,V}(u,v) = f_{X,Y}(x(u,v), y(u,v)) \cdot |J|$$
$$\text{where} \quad J = \begin{vmatrix} \dfrac{\partial x}{\partial u} & \dfrac{\partial x}{\partial v} \\ \\ \dfrac{\partial y}{\partial u} & \dfrac{\partial y}{\partial v} \end{vmatrix} \quad \text{determinant}$$

- if $Z = X + Y$ (2-dim $\to$ 1-dim) use CDF. Compute $P(Z \leq z) = P(X + Y \leq z)$. Integrate $f(x,y)$ over this region.

Sterling's Formula

$$n! \approx \sqrt{2\pi n} \cdot \left(\frac{n}{e}\right)^n$$

This is only really useful when n is large, when factorials are represented as ratios.

Conditional distribution

Discrete $$P_{X|Y=y}(x|y) = \frac{P_{X,Y}(x,y)}{P_Y(y)} = \frac{P(X = x, Y = y)}{P(Y = y)}$$
$$\Rightarrow \sum_y P_{X,Y}(x,y) = \sum_y P_{X|Y=y}(x|y) \cdot P_Y(y)$$

Continuous $$f_{X|Y=y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$
$$\Rightarrow f_X(x) = \int_y f(x,y)dy = \int_y f_{X|Y=y}(x|y) \cdot f_Y(y)dy$$
$$F_{X|Y}(x|y) = \int_{-\infty}^{x} f_{X|Y}(x|y)dx$$

$$\begin{tikzcd}
f_{X|Y} \arrow[d] & f_Y \arrow[dl, bend left, "x"] \\
f_{X,Y} \arrow[d, "\text{integrate}"] \arrow[dr] & \\
f_X \arrow[r, "\text{divide}"] & f_{Y|X}
\end{tikzcd}
\qquad
\begin{tikzcd}
f_{X,Y} \arrow[d, "\text{integrate}"] \arrow[dr, bend left, "\text{divide}"] & \\
f_X \arrow[r, "\text{divide}"] & f_{X|Y} \arrow[d, "\text{integrate } dx"] \\
& F_{X|Y}
\end{tikzcd}$$

Conditional Expectation

$$
\begin{aligned}
&E[X|Y=y] = \sum_x xP(X=x|Y=y)\\
&E[X|Y=y] = \int_x xf(x|y)dx\\
&E[X|Y] : \text{compute } E[X|Y=y] \text{ first, replace } y \text{ with } Y
\end{aligned}
$$

- **Properties:**
  - $E[aU+bV|Y=y] = aE[U|Y=y]+bE[V|Y=y]$ $\boxed{LOTUS}$
  - If $g(Y)=X$ then $E[X|Y=y]=X$
  - If $\underline{X \text{ and } Y}$ are independent, then $E[X|Y=y]=E[X]$

Conditional Variance

$$\boxed{Var(X|Y) = E[(X-E[X|Y])^2]} \qquad \text{(conditional variance)}$$

$$\boxed{Var(X|Y) = E[X^2|Y]-(E[X|Y])^2} \qquad \text{(unconditional variance)}$$

Ordered Statistics

Consider $X_1, X_2, \cdots, X_n$ $\quad X_{(j)} = $ j-th smallest

$$
\begin{aligned}
F_{\max(X_i)}(t) &= P(\max X_i \le t) = P(X_1 \le t)\cdot P(X_2 \le t)\cdots P(X_n \le t)\\
&= [F_X(t)]^n \qquad \boxed{f_{\max X_i}(t) = nF(t)^{n-1}f_X(t)}\\
F_{\min(X_i)}(t) &= 1-P(\min x_i \ge t) = 1-P(X_1 \ge t)\cdot P(X_2 \ge t)\cdots P(X_n \ge t)\\
&= 1-[1-F_X(t)]^n \qquad \boxed{f_{\min X_i}(t) = n[1-F(t)]^{n-1}f_X(t)}
\end{aligned}
$$

General: j-th order statistic

$$f_{x(j)}(t) = \binom{n}{j-1, 1, n-j} F_X(t)^{j-1}\cdot f_X(t)\cdot [1-F_X(t)]^{n-j}$$

As Beta distribution: Let $U_1, U_2, \ldots, U_N \sim i.i.d.$ Uniform$(0,1)$ and let $1 \le j \le N$
$U_{(j)} =$ jth smallest in $U_{(1)}, U_{(2)}, \ldots, U_{(N)}$ (ordered statistics). Then,

$$
\begin{aligned}
U_{(j)} &\sim Beta(j, N-j+1)\\
E[U_{(j)}] &= \frac{j}{N+1}
\end{aligned}
$$

Expectation and Variance

Law of Total Expectation:

$$E[X] = E[E[X|Y]]$$

Law of Total Variance:

$$Var(X) = E[Var(X|Y)] + Var[E(X|Y)]$$

Expectation

1. linearity of expectation
2. How to compute
   (a) LOTUS or definition (use density to integrate)
   (b) MGF: $M^{(n)}(0) = E[X^n]$ or by recognition
   (c) $E[X^2] = Var[X] + E[X]^2$
   (d) Tail probability X is non-neg R.V. $(x > 0)$ then $E[X] = \sum\limits_{t=0}^{\infty} P(X \geq t)\ or = \int\limits_{0}^{\infty} P(X \geq t)dt$

Variance

1. $Var(X_1 + X_2 + \cdots + X_n) = \sum\limits_{i=1}^{n} Var(X_i) + \sum\limits_{i \neq j} Cov(X_i, X_j)$

   if X_i, X_j identical (not independent) $= nVar(X_i) + n(n-1)Cov(X_i, X_j) \qquad i \neq j$

   $$\boxed{\text{Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)}}$$

2. **Covariance:**

$$Cov(X, Y) = E[XY] - E[X]E[Y]$$
$$Cov(X, c) = 0 \qquad c\ is\ a\ constant$$
$$Cov(X + Y, Z) = Cov(X, Z) + Cov(Y, Z)$$
$$Cov(cX, dZ) = cd \cdot Cov(X, Z)$$
$$Cov(aX + b, cY + d) = ac \cdot Cov(X, Y) \qquad a, b, c, d \text{ are constants}$$
$$Cov(X, Y) = 0 \qquad \text{If } X \perp Y \text{ (independent)}$$

3. **Correlation Coefficient:**

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}} = \frac{Cov(X, Y)}{\sigma_x \sigma_y}$$

MGFs

Let X be a random variable. Then

$$M_X(t) = E[e^{tX}]$$

it can also be written as:

$$= E\left[\sum_{j=0}^{\infty} \frac{(tX)^j}{j!}\right]$$

$$= E\left[\sum_{j=0}^{\infty} \left(\frac{X^j}{j!} \cdot t^j\right)\right]$$

$$\boxed{M_X^{(n)}(0) = E[X^n]}$$

If X and Y are independent, then

$$\begin{aligned} M_{X+Y}(t) &= E[E^{(X+Y)t}] \\ &= E[e^{tX}]E[e^{tY}] \\ &= M_X(t)M_Y(t) \end{aligned}$$

Limit Theorems

Markov's Inequality

For any non-negative random variable X

$$P(X \geq a) \leq \frac{E(X)}{a} \qquad \text{(for } \underline{\text{any}}\ a > 0\text{)}$$

Proof. Let $X \geq 0$ a random variable and let $a > 0$. Define new random variable from X as Y_a

$$Y_a = \begin{cases} 0 & \text{if } X < a \\ a & \text{if } X \geq a \end{cases}$$

$$0 \leq Y_a \leq X \implies \underbrace{E[Y_a]}_{a \cdot P(X \geq a)} \leq E[X]$$

$$E[Y_a] = 0 \cdot P(Y_a < a) + a \cdot P(X \geq a)$$

$$E[Y_a] = a \cdot P(X \geq a) \leq E[X] \implies \boxed{P(X \geq a) \leq \frac{E(X)}{a}}$$

■

Chebyshev's Inequality

For any random variable Y with mean μ_y and variance σ_y^2

$$P(|Y - \mu)y| \geq c) \leq \frac{\sigma_y^2}{c^2} \qquad \text{(for } \underline{\text{any}}\ c > 0\text{)}$$

Proof.

$$P(|Y-\mu_y)| \geq c) = P(\underbrace{|Y-\mu_y)|^2}_{=X} \geq c^2)$$
$$P(|Y-\mu_y)|^2 \geq c^2) \leq \frac{E[|Y-\mu_y|^2]}{c^2} = \frac{\sigma_y^2}{c^2}$$

∎

This is the same as

- $P(|Y-\mu_y| \geq k\sigma_y) \leq \frac{1}{k^2}$
- $P(|Y-\mu_y| \leq k\sigma_y) \geq \underbrace{1-\frac{1}{k^2}}_{\text{very conservative}}$

Central Limit Theorem

$$\sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu_n, n\sigma_x^2)$$
$$\frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}\left(\mu_n, \frac{\sigma_x^2}{n}\right)$$

Weak Law of Large Numbers

If $X_1, X_2, \cdots$ are *i.i.d.* with a mean μ

$$\text{then} \qquad \lim_{n\to\infty} P\left(\left|\bar{X_n}-\mu\right| \geq \epsilon\right) = 0$$

Strong Law of Large Numbers

$$X \xrightarrow{p} \mu_X \qquad \text{as } n \to \infty$$
$$Pr\left(\lim_{n\to\infty} \bar{X_n} = \mu\right) = 1$$

Jensen's Inequality

If $p_1, \ldots, p_n$ are positive numbers and $\sum_{i=1}^{n} p_i = 1$, and f is a real continuous function that is convex, then

$$f\left(\sum_{i=1}^{n} p_i x_i\right) \leq \sum_{i=1}^{n} p_i f(x_i)$$

Conversely, if f is a concave function

$$f\left(\sum_{i=1}^{n} p_i x_i\right) \geq \sum_{i=1}^{n} p_i f(x_i)$$

Lecture 1 (2018-08-30)

Survey Sampling

We have a population of objects under study (people, animals, places, etc.). We will consider a single numerical measurement associated to object $i : x_i$

Example. $N = 5000, x_i =$ height of person i, Population size = N. We denote population measurements $\{x_1, x_2, \cdots, x_N\}$
Compute population quantities:

- population total $\tau = \sum_{i=1}^{N} x_i$
- population mean $\mu = \frac{\tau}{N} = \frac{\sum_{i=1}^{N} x_i}{N}$

Note: τ and μ are population parameters, their computation depends on all the population data.

Question. How to estimate τ and μ based on a sample of observation from this population?

Classical Answer: Choose a "random" sample of objects and associated measurements denoted $\{x_1, x_2, \cdots, x_n\}$. *Note:* capital X_i denote random variables.
Whiter "Random"? Two types of ways to sample:

- without replacement
- with replacement

Claim 1. If X_i are drawn without replacement, then the distribution of X_1 and X_2 are identical. Is this true? **In fact, it is** $\Rightarrow$ They are **NOT** independent but they are identically distributed.

$$\text{P(Ace in Pos 1)} = \text{P(Ace in Pos 2)} = \tfrac{4}{52}$$

Combinatorial Approach

"well-shuffled deck" $\leftrightarrow$ all 52! rearrangements of the card are equally likely. How many rearrangements have ace at pos 1? $4 \cdot 51!$

$$P(A_1) = \frac{4 \cdot 51!}{52!} = \frac{4}{52} = P(A_2) = P(A_{19}) = P(A_{36})$$

Question. If X_1 and X_2 are identically distributed, then how do they differ between corresponding draws with replacement?

Answer. Independence. We can have Random Variables that are identically distributed and not independent. Note if independent, $P(A_2|A_1) = P(A_2)$.

with replacement

$$P(A_1) = \frac{4}{52}, \quad P(A_2) = \frac{4}{52}$$
$$P(A_2|A_1) = \frac{4}{52}$$

without replacement

$$P(A_1) = \frac{4}{52}, \quad P(A_2) = \frac{4}{52}$$
$$P(A_2|A_1) = \frac{3}{51}$$

We can see from this that depending on sampling method, we gain or lose independence. In the finite population sampling method, we have $1, \ldots, N$ objects we care about.
Loss of Independence when choosing sampling method is important.

Lecture 2 (2018-09-05)

Finite Population sampling – without i without replacement. Mean/expected value and variance of $\bar{X}$
Suppose our population is given by $\{x_1, \ldots, x_N\} = \{1, 2, 2, 7, 8, 9\}$ where

$$N = 6, \quad x_1 = 1 \quad x_2 = 2 \quad x_3 = 2 \quad x_4 = 7 \quad x_5 = 8 \quad x_6 = 9$$

Could also describe it by counting.

Distinct Value	frequency
$\varphi_1 = 1$	$n_1 = 1$
$\varphi_2 = 2$	$n_2 = 2$
$\varphi_3 = 7$	$n_3 = 1$
$\varphi_4 = 8$	$n_4 = 1$
$\varphi_5 = 9$	$n_5 = 1$

Possible sample of size $n = 6$, where we sample without replacement

$$X_1 = 7 \quad X_2 = 2 \quad X_3 = 8 \quad X_4 = 9 \quad X_5 = 1 \quad X_6 = 2$$

Sample here is the same as population as n=N

Same thing with replacement

$$X_1 = 9 \quad X_2 = 9 \quad X_3 = 9 \quad X_4 = 9 \quad X_5 = 9 \quad X_6 = 9$$

Typically N is large and $n << N$
Recall population parameters

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} \qquad \tau = N\mu = \sum_{i=1}^{N} X_i$$

Next, σ^2 (population variance)

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \qquad (\sigma^2 \text{ is pop. variance})$$

Alternatively, we can also express σ^2 as

$$\begin{aligned}
\sigma^2 = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N} &= \frac{\sum_{i=1}^{N}(x_i^2 - 2\mu x_i + \mu^2)}{N} \\
&= \frac{\sum_{i=1}^{N} x_i^2}{N} - \frac{2\mu}{N}\underbrace{\sum_{i=1}^{N} x_i}_{\mu} + \frac{\cancel{N}\mu^2}{\cancel{N}} \\
&= \frac{\sum_{i=1}^{N} x_i^2}{N} - 2\mu^2 + \mu^2
\end{aligned}$$

$$= \underbrace{\left(\frac{1}{N}\sum_{i=1}^{N} x_i^2\right)}_{\text{2nd moment}} - \mu^2 = \mu^{(2)} - \mu^2$$

Define: $\mu^{(k)} = \frac{1}{N}\sum\limits_{i=1}^{N} x_i^k$

Sample Mean $\bar{X}$ as an estimator

A function of the sample data for the population μ.
Note: If the sample is random ($X_1, \ldots, X_n$ are R.Vs), then $\bar{X}$ is **random!**
Questions:

1. How is $\bar{X}$ distributed? - in theory, if we know 1, then we know the answers 2 & 3 too.
2. What is $E[\bar{X}]$?
3. What is $Var(\bar{X})$?

Let's address 2

Consider $E[\underbrace{X_1}_{\text{first draw}}]$ possible values for $X_1 = \{x_1, \ldots, x_N\}$

$$P(X_1 = x_k) = \frac{1}{\binom{N}{1}} = \frac{1}{N}$$

e.x. $\{\underbrace{1}_{x_1}, \underbrace{2}_{x_2}, \underbrace{2}_{x_3}, \underbrace{7}_{x_4}, \underbrace{7}_{x_5}, \underbrace{9}_{x_6}\}$ gives every separate entry a unique ticket even if they are the same

$$E[X_1] = \frac{1}{N}\sum_{k=1}^{N} x_k = \mu = E[X_2] \qquad \text{(b/c } X_1 \text{ \& } X_2 \text{ are identically dist.)}$$

In sampling without replacement X_i & X_j are still identically distributed, but they are not independent.
In sampling with replacement, X_i & X_j are *i.i.d.*
Note that whether or not $X_1, \cdots, X_n$ are independent,

$$E\left[\sum_{i=1}^{N} X_i\right] = \sum_{i=1}^{N} E[X_i]$$

Note: The sample mean is equal to expected population mean regardless of sampling with or without replacement.

$$E[\bar{X}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{N} E[X_i] = \frac{n\mu}{n} = \mu$$

Since $E[\bar{X}] = \mu$, we say $\bar{X}$ is an unbiased estimator for μ. **BUT** $\underbrace{\bar{X}}_{\text{R.V.}} \neq \overbrace{\mu}^{\text{constant}}$

Let's address 3

Sampling with replacement.

Theorem. *Sampling from finite population with replacement*

$$Var(\bar{X}) = \frac{\sigma^2}{n}$$

Proof. Here $X_1, \cdots, X_n$ are *i.i.d.*. In general, X_i's are R.V. and a_i's are constants

$$Var\big(\sum_i a_i X_i\big) = \sum_i \sum_j a_i a_j cov(X_i, X_j)$$

If $X_1, \cdots, X_N$ are independent, $\underset{i \neq j}{Cov(X_i, X_j)} = 0$! Hence

$$Var(\bar{X}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} Var\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \underbrace{Var(X_i)}_{\text{a constant}}$$

$$\boxed{Var(\bar{X}) = \frac{Var(X_i)}{n} = \frac{\sigma^2}{n}}$$

■

We need to compute $Var(X_i)$. Observe that $Var(X_i)$ are same for all: *Why?* because they are identical.
Also notice $\frac{Var(X_i)}{n}$ decreases with n.
Observe that for all finite n, $Var(\bar{X})$ is not 0 unless $Var(X_i) = 0$!
Note: $Var(X_i) = E[(X_i - E(X_i))^2] = E[(X_i - \mu^2)] = \frac{1}{N}\sum(x_i - \mu)^2 = \sigma^2$
<u>So</u> $Var(X_i) = 0$ **iff** all $X_i \equiv \mu$

Lemma. *bX is <u>consistent</u> for μ, i.e. $\forall \delta > 0$, the $P(|\bar{X} - \mu| > \delta) \longrightarrow 0$ as $n \to \infty$*

For this Lemma, we need to Prove Chebyshev's Inequality, which is

$$P(|Z - E(Z)| > \delta) \leq \frac{Var(Z)}{\delta^2}$$

Use this identity!

$$E[\bar{X}] = \mu, \qquad Var(\bar{X}) = \frac{\sigma^2}{n}$$

$$P(|\bar{X} - E(\bar{X})| > \delta) \leq \frac{Var(\bar{X})}{\delta^2} = \frac{\sigma^2}{n\delta^2} \to 0 \qquad \text{as } n \to \infty$$

Lecture 3 (2018-09-10)

Sampling without replacement

$Var(\bar{X}) =$ when sampling without replacement

Theorem. *Sampling from finite population without replacement*

$$Var(\bar{X}) = \frac{\sigma^2}{n}\left[\underbrace{\frac{N-n}{n-1}}_{FPN}\right] \qquad \text{(finite population correction)}$$

<u>Points to Note</u> - In sample without replacement,

- If $n = N$, $Var(\bar{X}) = 0$
- If $n = 1$, $Var(\bar{X}) = \frac{\sigma^2}{n} = \sigma^2$, same as with replacement
- Check: for $n > 1$, how does $\frac{N-n}{N-1}$ relate to 1? The $Var(\bar{X})$ is <u>always</u> less without replacement

Proof. Start

1

$$Var(\bar{X}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_i \sum_j Cov(X_i, X_j)$$

$$\left(\text{When sampling with replacement, } Cov(X_i, X_j) = 0 \text{ if } i \neq j\right)$$

In sampling without replacement, we cannot assert that $Cov(X_i, X_j) = 0$ and we'll compute it explicitly.

$$\text{Recall} \qquad Cov(X_i, X_j) = E[X_iX_j] - \underbrace{E[X_i]E[X_j]}_{\mu^2}$$

$$\mu^2 \leftarrow \text{as identical but not independent} \qquad = E[X_iX_j] - \mu^2$$

2 To calculate $E[X_iX_j]$, let us list distinct values in population

Example. $\{\underbrace{5}_{x_1}, \underbrace{5}_{x_2}, \underbrace{8}_{x_3}, \underbrace{11}_{x_4}, \underbrace{8}_{x_5}, \underbrace{17}_{x_6}, \underbrace{9}_{x_7}\}$ Let $n_l = \#$ of times ζ_l appears in population.

Distinct Value	frequency
$\zeta_1 = 5$	$n_1 = 2$
$\zeta_2 = 8$	$n_2 = 2$
$\zeta_3 = 11$	$n_3 = 1$
$\zeta_4 = 17$	$n_4 = 1$
$\zeta_5 = 9$	$n_5 = 1$

$$P[X_i = 5] = \frac{2}{7} = \frac{n_1}{N} \qquad \text{(i draws identical)}$$

$$\Rightarrow P[X_i = \zeta_l] = \frac{n_l}{N}$$

$$n_1 + n_2 + \ldots + n_m = \sum_{j=1}^{m} n_j = N$$

$$E[X_i X_j] = \sum_{k=1}^{m} \sum_{l=1}^{m} \zeta_k \zeta_l \underbrace{P[X_i = \zeta_k, X_j = \zeta_l]}_{?}$$

$$P[X_i = \zeta_k, X_j = \zeta_l] = \underbrace{P[X_j = \zeta_l | X_i = \zeta_k]}_{3} \cdot \underbrace{P[X_i = \zeta_k]}_{=\frac{n_k}{N}}$$

3 Cases for Conditional probability

$$P[X_j = \zeta_l | X_i = \zeta_k] \overset{cases}{=} \begin{cases} \frac{n_l}{N_1} & l \neq k \to \text{numbers are diff.} \\ \frac{n_l - 1}{N-1} & l = k \to \text{numbers are same} \end{cases}$$

4 So we have

$$E[X_i X_j] = \sum_{k=1}^{m} \sum_{l=1}^{m} \zeta_k \zeta_l P[X_i = \zeta_k, X_j = \zeta_l]$$

$$E[X_i X_j] = \sum_{k=1}^{m} \sum_{l=1}^{m} \zeta_k \zeta_l P[X_j = \zeta_l | X_i = \zeta_k] \cdot P[X_i = \zeta_k]$$

$$= \sum_{k} \zeta_k P[X_i = \zeta_k] \zeta_k \left(\sum_{l} \zeta_l P[X_j = \zeta_l | X_i = \zeta_k] \right)$$

$$= \sum_{k} \zeta_k P[X_i = \zeta_k] \zeta_k \left(\sum_{l \neq k} \zeta_l P[X_j = \zeta_l | X_i = \zeta_k] + \zeta_k P[X_j = \zeta_k | X_i = \zeta_k] \right)$$

$$= \sum_{k} \zeta_k P[X_i = \zeta_k] \zeta_k \left(\underbrace{\sum_{l \neq k} \zeta_l \frac{n_l}{N-1}}_{5} + \zeta_k \frac{n_k - 1}{N-1} \right)$$

5 When $l \neq k$ and we want to remove all l terms

$$\sum_{l \neq k} \zeta_l \frac{n_l}{N-1} = \frac{1}{N-1} \sum_{l \neq k} \zeta_l n_l$$

$$\left(\sum_{l} \zeta_l n_l = \tau = n\mu \right) \quad \text{population total}$$

$$= \frac{1}{N-1} (\tau - \zeta_k n_k)$$

(6) Now Back

$$
\begin{aligned}
E[X_iX_j] &= \sum_k \zeta_k \frac{n_k}{N}\left(\frac{1}{N-1}(\tau - \zeta_k n_k) + \zeta_k \frac{n_k - 1}{N-1}\right) \\
&= \frac{1}{N(N-1)} \sum_k \zeta_k n_k \left[(\tau - \cancel{\zeta_k n_k}) + \cancel{\zeta_k n_k} - \zeta_k\right] \\
&= \frac{1}{N(N-1)} \sum_k \zeta_k n_k \left[\tau - \zeta_k\right] \\
&= \frac{1}{N(N-1)} \left(\sum_k \zeta_k n_k \tau - \sum_k \zeta_k^2 n_k\right) \\
&= \frac{1}{N(N-1)} \left[\tau^2 - \sum_k \zeta_k^2 n_k\right]
\end{aligned}
$$

(7) What is $\sum_k (\zeta_k)^2 \frac{n_k}{N}$? Second moment $E[X_i^2]$ $\quad E[X_i^2] = \sigma^2 + \mu^2$

$$
\begin{aligned}
E[X_i^2] = \sigma^2 + \mu^2 &\qquad \frac{\tau^2}{N} = N\mu^2 as \mu = \frac{\tau}{N} \\
E[X_iX_j] &\Longrightarrow \frac{1}{N-1}\left[N\mu^2 - (\sigma^2 + \mu^2)\right] \\
&= \frac{1}{N-1}[(N-1)\mu^2 - \sigma^2] = \mu^2 - \frac{\sigma^2}{N-1}
\end{aligned}
$$

$$
\begin{aligned}
\text{So} \quad Cov(X_i, X_j) &= \mu^2 - \frac{\sigma^2}{N-1} - \mu^2 \\
&= -\frac{\sigma^2}{N-1} \qquad (\text{Cov} < 0) \\
\text{So} \quad Cov(X_i, X_j) &= Var(X_i) = \sigma^2
\end{aligned}
$$

(8) Putting it all together

$$
\begin{aligned}
Var(\bar{X}) &= \frac{1}{n^2}\left(\sum_{i \neq j} Cov(X_i, X_j) + \sum_{i=1}^{n} Var(X_i)\right) \\
&= \frac{1}{n^2}\left(\sum_{i \neq j} -\frac{\sigma^2}{N-1} + n\sigma^2\right) \\
&= \frac{1}{n^2}\left(\frac{-n(n-1)\sigma^2}{N-1} + \frac{\sigma^2}{n}\right) \\
&= \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right) \\
&\boxed{= \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)}
\end{aligned}
$$

■

Lecture 4 (2018-09-12)

- Binary data- special case.
- Approximate distance of $\bar{X}$ when n is large but $n << N$
- Estimating population Variance
- Bivariate data

Recall that population is <u>dichotomous</u> or <u>binary</u> then $x_i = \begin{cases} 1 \\ 0 \end{cases}$

Moreover if we consider $x_i = 1$ as a "success" and $x_i = 0$ as a "failure", then

$$\mu = \frac{\sum_{i=1}^N X_i}{N} = \frac{\#\text{ of successess in population}}{\text{population size}} = p \qquad (\text{pop}^n \text{ proportion of success})$$

$$\text{Now,} \quad \sigma^2 = \underbrace{\frac{\sum_{i=1}^N X_i}{N}}_{\mu} - \mu^2 = p - p^2 = p(1-p) = pq$$

$$\mu \text{ as } 1 \Rightarrow 1^2 = 1 \qquad\qquad 0 \Rightarrow 0^2 = 0$$

Recall that if $Y \sim \text{Bernoulli}(p)$, $Y_i = \begin{cases} 1 & \text{w/ prob } p \\ 0 & \text{w/ prob } 1-p \end{cases}$

$$E[Y] = p$$
$$Var(Y) = p(1-p)$$

Last few weeks involved an analysis of $\bar{X}$, $E(\bar{X})$, $Var(\bar{X})$. Could also ask: How is $\bar{X}$ distributed if n is large.

Confidence Intervals - Sampling W.R.

If sampling **with replacement**, where $X_1, \ldots, X_n$ denotes sample, we know X_i's are *i.i.d.* Hence when n is large, by CLT $\bar{X}$ has an approximately normal distribution.

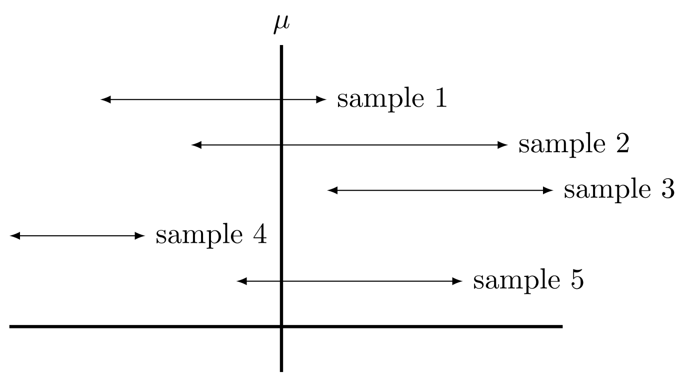
$$P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq x\right) \longrightarrow \Phi(x) \qquad \text{as } n \to \infty$$

When sampling with replacement, we can use this to obtain confidence intervals for μ: Let $\alpha \in (0,1)$ be given.

$$\text{Let } Z_\alpha \in \mathbb{R} \text{ such that } P(Z > Z_\alpha) = \alpha \text{ where } Z \sim N(0,1)$$

By the Central Limit Theorem, for n large (sampling w/replacement)

$$= P\left(-Z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq Z_{\alpha/2}\right)$$
$$= P\left(\underbrace{\bar{X} - Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}}_{\text{Random}} \leq \mu \leq \underbrace{\bar{X} + Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}}_{\text{Random}}\right)$$
$$Var(\bar{X}) = 0 \qquad \text{Never happens}$$



In repeated sampling, approx $(1-\alpha)$ of intervals contain μ, and (α) frac will not.

$$\text{We say } \boxed{\bar{X} - Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}} \text{ is } 100(1-\alpha)\% \text{ 2-sided confidence interval for } \mu$$

Problem: This interval involved σ which is unknown. Observe that if n is large, then $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ is still approx $N(0,1)$ in distribution where (no population parameters)

$$\boxed{s^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2} \qquad \text{(sample variance)}$$

So we obtain

$$\boxed{\bar{X} \pm Z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}} \qquad \text{as a } 100(1-\alpha) \text{ CI for } \mu$$

In the dichotomous case,

$$\bar{X} = \frac{\text{\# of the succession sample}}{\text{sample size}} = \hat{p}$$

$$100(1-\alpha)\% \text{ CI for} \qquad p : \hat{p} \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Confidence Intervals - Sampling W.o.R.

Recall now what happens when sampling **without replacement**

Here, $X_1, X_2, \ldots, X_n$ remain identically distributed, but not independent

We surmised, that if $n << N$, $X_i \& X_j$ have an *"approximate independence"*

Example 1. Let population consist of 1000 elements. In this case:

$$\text{blue } - 1 - 200, \qquad \text{red } - 2 - 300, \qquad \text{green } - 1 - 500$$

$$\left.\begin{array}{l} P(X_1 = 3) = \frac{1}{2} \\ P(X_2 = 3 | X_1 = 3) = \frac{499}{999} \end{array}\right\} \text{not independent, but have approximate independence.}$$

In short, $n << N$, each successive draw does not alter probabilities that much, precisely b/c removal is only of a sample # of population elements.
So if $n << N$, then even in sampling W.O.R, X_i's retain an approximate independence. Further if n is "large" and small relative to N, (note delicate point!) then $\bar{X}$ will still have an approx Normal distribution.

$$\frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)}} \sim N(0,1)$$

Observe σ^2 us still unknown. We'd like to consider estimators for σ^2

Estimator for variance W.o.R

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Try to understand $E[\hat{\sigma}^2]$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_1^2 - 2X_i\bar{X} + \bar{X}^2)$$
$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - 2\bar{X}\bar{X} + \bar{X}^2 = \frac{1}{n}\sum_{i=1}^{n}X_i^2 - \bar{X}^2$$
$$E[\hat{\sigma}^2] = \underbrace{E\left[\frac{1}{n}\sum_{i=1}^{n}X_i^2\right]}_{1} - \underbrace{E[\bar{X}^2]}_{2} \qquad \text{can get } E[\bar{X}^2] \text{ from } Var(\bar{X})$$

1 $\quad E\left[\frac{1}{n}\sum_{i=1}^{n}X_i^2\right] = \frac{1}{n}\sum_{i=1}^{n}E[X_i^2] = \sigma^2 + \mu^2$

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

$$E[\bar{X}^2] = \underbrace{Var(\bar{X})}_{\text{computed}} + \mu^2$$

Combining, we get:

$$E[\hat{\sigma}^2] = \sigma^2 + \mu^2 - (Var(\bar{X}) + \mu^2)$$
$$E[\hat{\sigma}^2] = \sigma^2 - \left[\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)\right]$$

The estimator is biased, but

$$E[\hat{\sigma}^2] = \sigma^2\left(\underbrace{1 - \frac{N-n}{(n)(N-1)}}_{\text{constant, } c}\right)$$
$$E[\hat{\sigma}^2] = C\sigma^2$$

and thus $\frac{\hat{\sigma}^2}{C}$ is an unbiased estimator.

Lecture 5 (2018-09-17)

- Approximation methods / Delta-methods
- Bivariate populations
- Ratio estimations

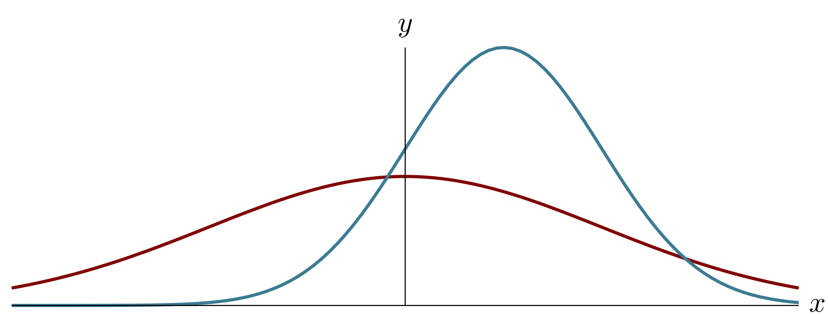
We calculated $E\underbrace{\left[\hat{\sigma}^2\right]}_{C\sigma^2}$ where $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ and you can use our computations to generate an unbiased estimator for population variance σ^2. Can also use his to calculate $E[s^2]$, where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

Bias-Variance Tradeoff

1. Unbiased estimators are useful: if T is an unbiased estimator for θ then $E[T] = \theta$.
2. However, if we wish to evaluate two estimatorsL- one biased and other unbiased, we may not universally want to choose the unbiased one always, we need to consider $\boxed{variance}$.

Why? Suppose that T is an estimator for θ.
The Mean Squared Error (MSE):

$$MSE = E[(T-\theta)^2] \xrightarrow{\text{exercised}} \underbrace{Var(T)}_{\text{Variance}} + \underbrace{(E(T)-\theta)^2}_{\text{Bias}}$$



We can see from the above plots that the red graph has an estimator θ closer to μ, but has a higher variance. However, estimator B has an unbiased estimator, but has a smaller variance. Depends on sampling analysis.

Bivariate population sampling

Suppose we have a population of N objects. On each object we have a pair of measurements: (x_i, y_i)
Note: When sampling from this population if object i is in sample, then both measurements in pair (x_i, y_i) are retained. In particular (x_i, y_i) appears exactly once in the population, and sample w/o repl, then you cannot retrieve measurement i later.

Parameters

$$\mu_X = \frac{1}{N}\sum_{i=1}^{N} X_i \qquad \tau_X = N\mu_X$$

$$\tau_Y = N\mu_Y$$

$$\sigma_Y^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \mu_Y)^2 \qquad \mu_Y = \frac{1}{N}\sum_{i=1}^{N} Y_i \qquad \sigma_X^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_X)^2$$

Covariance

$$\sigma_{XY}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_Y)$$

Suppose $\mu_X \neq 0$ Define $r = \frac{\mu_X}{\mu_Y}$ What is a reasonable estimator r?

Could consider $R = \frac{\bar{X}}{\bar{Y}}$

Now Suppose that μ_X. were known. Consider $\mu_X \cdot R = \frac{\mu_X}{\bar{X}}\bar{Y}$.
Plausible estimator for μ_Y. But why? we already have $\bar{Y}$, an unbiased estimator for μ_Y. We will see that $\mu_X \cdot R$, the so called **ratio estimate**, is

1. a biased estimate
2. can contribute in reduction in variance relative to $\bar{Y}$

So we will need to understand $E[R], Var(R)$ & approximations of $E[R]$ & $Var(R)$

Approximation Methods

Let X be a random variable with mean $= \mu_X$ and variance $= \sigma_X^2$. Let $Z = g(X)$, where $g : \mathbb{R} \to \mathbb{R}$, g a deterministic function of x.
Question: How to compute $E[Z]$?
Answer: If density of X is known, (call this f_X), then

$$E(Z) = \int_{\mathbb{R}} g(X) f_X(x) dx \qquad \text{involves an integral}$$

Cumbersome even if f_X is known; closed form solution to integral exists; not possible to get exact value even if f_X known, but no closed form solution; not even possible to write integral if f_X unknown. If g is linear, then it is OK e.g. $E[g(X)] = E[aX + b] = a\mu_X + b$

Taylor Expansions

Taylor expansion of g about μ_X (Why? Think Chebyshev!)

$$g(x) \approx g(\mu_X) + g'(\mu_X)(x - \mu_X) + \frac{g''(x)(x-\mu_X)^2}{2!} + \cdots + \text{higher order terms}$$

$$\boxed{g(X) \approx g(\mu_X) + g'(\mu_X)(X - \mu_X) + \frac{g''(X)(X-\mu_X)^2}{2!}}$$

$$E[Z] \approx E[g(\mu_X)] + E[g'(\mu_X)(X-\mu_X)] + E\left[\frac{g''(\mu_X)}{2!}(X-\mu_X)^2\right]$$

$$\approx g(\mu_X) + g'(\mu_X)\underbrace{\cancel{E[(X-\mu_X)]}}_{0} + \frac{g''(\mu_X)}{2!}E[(X-\mu_X)^2]$$

$$\boxed{E[Z] \approx g(\mu_X) + \frac{g''(\mu_X)}{2!}\sigma_X^2}$$

But $R = \frac{\bar{Y}}{\bar{X}}$, a function of two variables!

Consider $\quad g(x,y) : \mathbb{R}^2 \to \mathbb{R}$

Taylor expand g about (μ_x, μ_y)

1 Linear Approximation

$$g(x,y) \approx g(\mu_x,\mu_y) + \frac{\partial g}{\partial x}(\mu_x,\mu_y)\cdot(x-\mu_x) + \frac{\partial g}{\partial y}(\mu_x,\mu_y)\cdot(y-\mu_y)$$

2 Second order approximation

$$g(x,y) \approx g(\mu_x,\mu_y) + \frac{\partial g}{\partial x}(\mu_x,\mu_y)\cdot(x-\mu_x) + \frac{\partial g}{\partial y}(\mu_x,\mu_y)\cdot(y-\mu_y)$$
$$+\frac{1}{2}\frac{\partial^2 g}{\partial x^2}(\mu_x,\mu_y)\cdot(x-\mu_x)^2 + \frac{1}{2}\frac{\partial^2 g}{\partial y^2}(\mu_x,\mu_y)\cdot(y-\mu_y)^2 + \frac{\partial g}{\partial x \partial y}(\mu_x,\mu_y)\cdot(x-\mu_x)(y-\mu_y)$$

Evaluating $E[g(X,Y)]$

$$E[g(X,Y)] \approx g(\mu_x,\mu_y) + \frac{\partial g}{\partial x}(\mu_x,\mu_y)\cdot\underbrace{\cancel{E[(x-\mu_x)]}}_{0} + \frac{\partial g}{\partial y}(\mu_x,\mu_y)\cdot\underbrace{\cancel{E[(y-\mu_y)]}}_{0}$$
$$+\frac{1}{2}\frac{\partial^2 g}{\partial x^2}(\mu_x,\mu_y)\cdot E[(x-\mu_x)^2] + \frac{1}{2}\frac{\partial^2 g}{\partial y^2}(\mu_x,\mu_y)\cdot E[(y-\mu_y)^2] + \frac{\partial g}{\partial x \partial y}(\mu_x,\mu_y)\cdot E[(x-\mu_x)(y-\mu_y)]$$

When the dust settles,

$$\boxed{E[g(X,Y)] \approx g(\mu_x,\mu_y) + \frac{1}{2}\frac{\partial^2 g}{\partial x^2}(\mu_x,\mu_y)\cdot\sigma_X^2 + +\frac{1}{2}\frac{\partial^2 g}{\partial y^2}(\mu_x,\mu_y)\cdot\sigma_Y^2 + \frac{\partial g}{\partial x \partial y}(\mu_x,\mu_y)\cdot Cov(X,Y)}$$

Lecture 6 (2018-09-19)

- Approximation methods, Δ-methods
- Ratio estimations
- Parametric Estimation

Let X be a r.v. mean μ_X and variance σ_X^2. Let g be a deterministic function $g : \mathbb{R} \to \mathbb{R}$.
Let $Z = g(X)$ How to approximate $E[g(X)] = g(Z)$? We could do

$$E[Z] \approx g(\mu_X) + \frac{1}{2} g''(\mu_X) \cdot Var(X)$$

Whether or not this approximation is accurate depends on contribution to higher order terms.
If $Z = g(X,Y)$, then $E[Z]$ is

$$E[Z] \approx g(\mu_x, \mu_y) + \frac{1}{2}\frac{\partial^2 g}{\partial x^2}(\mu_x,\mu_y) \cdot \sigma_X^2 + + \frac{1}{2}\frac{\partial^2 g}{\partial y^2}(\mu_x,\mu_y) \cdot \sigma_Y^2 + \frac{\partial g}{\partial x \partial y}(\mu_x,\mu_y) \cdot \sigma_{XY}$$

Goal: Understand $E[R]$, $Var(R)$ where $R = \frac{\bar{Y}}{\bar{X}}$ and we are sampling W.o.R from a finite bivariate population

---

Let's consider what happens when $g(X,Y) = \frac{Y}{X}$

$$\frac{\partial g}{\partial x} = \frac{-y}{x^2} \to \frac{\partial^2 g}{\partial x^2} = \frac{2y}{x^3} \qquad \frac{\partial g}{\partial y} = \frac{1}{x} \to \frac{\partial^2 g}{\partial y^2} = 0 \qquad \frac{\partial^2 g}{\partial x \partial y} = -\frac{1}{x^2}$$

Here we will look at $g(\bar{X},\bar{Y}) = \frac{\bar{X}}{\bar{Y}}$ $\qquad E[\bar{X}] = \mu_x$ and $E[\bar{Y}] = \mu_y$

$$E[g(\bar{X},\bar{Y})] = E\left[\frac{\bar{X}}{\bar{Y}}\right] \approx \frac{\mu_y}{\mu_x} + \frac{1}{2}\left(\frac{2\mu_y}{(\mu_x)^3}\right)\sigma_{\bar{X}}^2 + 0 - \frac{1}{\mu_x^2}\sigma_{\bar{X}\bar{Y}}$$

Do we think $\mu_x R$ is unbiased for μ_y **Answer:** <u>No</u>, it is not unbiased b/c look at approximation

What about variance?

Let's return for a minute on general setting for approximations of moments of functions of random variables. Again $g(X,Y) = Z$

Let's write 1st order Taylor expansion for Z

$$Z \approx g(\mu_x,\mu_y) + \frac{\partial g}{\partial x}(\mu_x,\mu_y) \cdot (x - \mu_x) + \frac{\partial g}{\partial y}(\mu_x,\mu_y) \cdot (y - \mu_y)$$

So we find

$$\begin{gathered} Z \approx a + b(X - \mu_X) + c(Y - \mu_Y) \\ Var(Z) \approx b^2 Var(X) + c^2 Var(Y) + 2bc Cov(X,Y) \\ \approx \left[\underbrace{\frac{\partial g}{\partial x}}_{\text{b}}\right]^2 \sigma_X^2 + \left[\underbrace{\frac{\partial g}{\partial y}}_{\text{c}}\right]^2 \sigma_Y^2 + 2\left[\underbrace{\frac{\partial g}{\partial x}}_{\text{b}}\right]\left[\underbrace{\frac{\partial g}{\partial y}}_{\text{c}}\right]\sigma_{XY} \end{gathered}$$

We don't go further than linear as higher variance requires higher order moments e.g. $E[x^4] \leftarrow$ they don't matter.

$$Var(R) \approx \left[\frac{-\mu_y}{\mu_x^2}\right]^2 \sigma_{\bar{X}}^2 + \left[\frac{1}{\mu_x}\right]^2 \sigma_{\bar{Y}}^2 + 2\left[\frac{-\mu_y}{\mu_x^2}\right]\left[\frac{1}{\mu_x}\right]\sigma_{\bar{X}\bar{Y}} \qquad (\star)$$

Recall

$$\sigma_{\bar{X}}^2 = \frac{\sigma_x}{n}\left[\frac{N-n}{N-1}\right] \qquad \sigma_{\bar{Y}}^2 = \frac{\sigma_y}{n}\left[\frac{N-n}{N-1}\right]$$
$$\sigma_{\bar{X}\bar{Y}} = ? \quad \frac{\sigma_{xy}}{n}\left[\frac{N-n}{N-1}\right]$$

Recall

$$\sigma_{XY} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)$$
$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \implies \boxed{\sigma_{xy} = \rho\sigma_x\sigma_y}$$

Now $\star$ implies

$$\begin{aligned} Var(R) &\approx \frac{1}{n}\left[\frac{N-n}{N-1}\right]\left\{\frac{\mu_y^2}{\mu_x^4}\sigma_x^2 + \frac{1}{\mu_x^2}\sigma_y^2 - \frac{2\mu_y}{\mu_x^3}\sigma_{xy}\right\} \\ &\approx \frac{1}{n\mu_x^2}\left[\frac{N-n}{N-1}\right]\left\{\underbrace{\frac{\mu_y^2}{\mu_x^2}}_{r^2}\sigma_x^2 + \sigma_y^2 - 2\underbrace{\frac{\mu_y}{\mu_x^3}}_{r}\underbrace{\sigma_{xy}}_{\rho\sigma_x\sigma_y}\right\} \\ Var(R) &\approx \frac{1}{n\mu_x^2}\left[\frac{N-n}{N-1}\right]\left(r^2\sigma_x^2 + \sigma_y^2 - 2r\rho\sigma_x\sigma_y\right) \end{aligned}$$

Ratio Estimations

Ratio estimate for μ_Y is $\mu_X R \leftarrow$ useful if μ_X is known. We know from before that $E[\mu_X R] \neq \mu_Y$.

$$Var(\bar{Y}) = \frac{\sigma_y^2}{n}\left[\frac{N-n}{N-1}\right] \qquad E[\bar{Y}] = \mu$$

Ratio is useful if bias is small and variance reduction is significant (relative to $Var(\bar{Y})$).

Recall

$$\begin{aligned} E(R) &= \frac{\mu_x}{\mu_y} + \frac{1}{\cancel{2}}\frac{\cancel{2}\mu_y}{\mu_x^3}\cdot\frac{\sigma_y^2}{n}\left[\frac{N-n}{N-1}\right] - \frac{1}{\mu_X^2}\frac{\sigma_{xy}}{n}\left[\frac{N-n}{N-1}\right] \\ &\approx r + \frac{1}{n\mu_x^2}\left[\frac{N-n}{N-1}\right]\left(r\sigma_x^2 - \rho\sigma_x\sigma_y\right) \end{aligned}$$

Finally,

$$E[\mu_x R] \approx \mu_y + \frac{1}{\mu_y}\left(\frac{1}{n}\right)\left(\frac{N-n}{N-1}\right)(r\sigma_x^2 - \rho\sigma_x\sigma_y)$$

So is non-zero, but decaying in n.
Fact: For n large but small relative to N ($n << N$), R can be approx. using normal distribution.

Lecture 7 (2018-09-24)

- Properties of estimation
- Method of moments
- Maximum Likelihood
- Properties of estimators

Properties of Estimation

Let X_i, $1 \leq i \leq n$, be i.i.d. random variables with some cdf F_θ, where $\theta \subseteq \mathbb{R}^d$ is deterministic but potentially unknown vector.

We will often consider X_i's with a pdf or pmf f_θ as well.

Example.

1. X_i's are i.i.d. Bernoulli (p), p is unknown pmf: $P(X=1) = p$, $P(X=0) = 1-p$. How to estimate p if we observe $X_1, \ldots, X_n$?
2. X_i's are i.i.d. Poission(λ), $\lambda > 0$. How to estimate λ given $X_1, \ldots, X_n$?
3. X_i's are i.i.d. Exp(λ), $\lambda > 0$. How to estimate λ given $X_1, \ldots, X_n$?
4. X_i's are i.i.d. Uniform$[0, \theta]$. How to estimate θ given $X_1, \ldots, X_n$? - What if X_i's are i.i.d. Unif$[\alpha, \beta]$. How to estimate α, β then $X_1, \ldots, X_n$?
5. X_i's are i.i.d. Gamma$[\alpha, \beta]$. How to estimate (α, β) given $X_1, \ldots, X_n$? What if one of α, β is known?
6. Let $X_1, \ldots, X_n \in \mathbb{R}^d$ be i.i.d. multivariate normal with mean vector $\vec{\mu} \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{dxd}$. How to estimate μ_d and Σ from $X_1, \ldots, X_n$

In all cases, we are concerned with estimating the parameters associated to a cdf or density whose functional form we have specified and from which we have an i.i.d. sample.

If we did not specify the correctional form of F_θ, e.g. did not specify "normal", then the inference at F/t itself is classified as "non-parametric" inference.

According to laws of large numbers, both these lend credence to the idea that if we wish to estimate $\mu = E[X_i]$, $\bar{X}$ is a reasonable start.

Definition. The population moment is defined as

$$\begin{aligned} \mu^{(k)} &= \mathbb{E}[X_i^k] \\ &= \int x^k f(x) dx - \quad \text{given all data i.e. entire population.} \end{aligned}$$

Observe that $Y_i = X_i^k \sim$ i.i.d. and $\mathbb{E}[Y_i] = E[X_i^k]$

So laws of large numbers apply to $\bar{Y}$ and suggest:

$$\bar{Y} = \frac{\sum X_i^k}{n} = \frac{\sum Y_i}{n} = \text{kth sample moment of } X_i$$

is a reasonable estimate for $\mu^{(k)}$

Method of Moments estimators

Suppose we are interested in d parameters $\alpha_1, \ldots, \alpha_d$ (need not be population moments themselves)

Step 1 - This system related population moments to parameters $\alpha_1, \ldots, \alpha_d$

$$\begin{aligned} \mu^{(1)} &= g_1(\alpha_1, \ldots, \alpha_d) \\ &\vdots \\ \mu^{(d)} &= g_d(\alpha_1, \ldots, \alpha_d) \end{aligned}$$

Step 2 - Invert this to solve for α_1 in terms of $\mu^{(1)}, \ldots, \mu^{(d)}$

$$\begin{aligned} \alpha_1 &= h_1(\mu^{(1)}, \ldots, \mu^{(d)}) \\ &\vdots \\ \alpha_d &= h_d(\mu^{(1)}, \ldots, \mu^{(d)}) \end{aligned}$$

Step 3 - Now if h_1 functions are regular enough (continuous, differentiable, etc.). Then again by laws of large numbers, we can find $\alpha_1, \ldots, \alpha_d$

Example 1. Let X_i's be i.i.d. $\mathcal{N}(\mu, \sigma^2)$ Calculate MOM estimators for μ, σ^2

$$\alpha_1 = \mu = \mu^{(1)} = \bar{X} \qquad \alpha_2 = \sigma^2 = \mu^{(2)} - (\underbrace{\mu^{(1)}}_{\bar{X}})^2$$

$$\hat{\alpha_1}_{\text{MOM}} = \bar{X} \qquad \hat{\alpha_2}_{\text{MOM}} = \mu^{(2)} - (\underbrace{\mu^{(1)}}_{\bar{X}})^2$$

Example 2. Suppose X_i's are uniform $(0, \theta)$

$$\mu^{(1)} = \frac{\theta - 0}{2} \Rightarrow \theta = 2\mu^{(1)} \qquad \hat{\theta}_{\text{MOM}} = 2\bar{X}$$

Example 3. Suppose X_i's are $\exp(\lambda)$

$$\mu^{(1)} = \frac{1}{\lambda} = \bar{X} \qquad \hat{\lambda}_{\text{MOM}} = \frac{1}{\bar{X}}$$

Example 4. Let $X \sim \text{Gamma}(\alpha, \lambda)$

$$E[X] = \mu^{(1)} = \frac{\alpha}{\lambda} \qquad E[X^2] = \mu^{(2)} = \frac{\alpha(\alpha+1)}{\lambda^2} = \mu^{(1)^2} + \frac{\mu^{(1)}}{\lambda}$$

$$\hat{\alpha}_{\text{MLE}} = \frac{\hat{\mu}^{(1)}}{\hat{\mu}^{(2)} - \hat{\mu}^{(1)^2}} \qquad \hat{\lambda}_{\text{MLE}} = \frac{\hat{\mu^{(1)}}}{\hat{\mu^{(2)}} \hat{\mu^{(1)^2}}}$$

Lecture 8 (2018-09-26)

Maximum Likelihood Estimation

Suppose $X_1, \ldots, X_n$ are i.i.d. with common density $f(x|\theta)$ for some parameter θ or pmf $p(X|\theta)$

Note: functional form is assumed known, θ may not be. Recall joint density of $X_1, \ldots, X_n$ is $f(X_1, \ldots, X_n|\theta)$

$$\begin{aligned} f(X_1, \ldots, X_n|\theta) &= f_1(X_1|\theta) \cdot f_2(X_2|\theta) \cdots f_n(X_n|\theta) \\ &= \prod_{i=1}^{n} f(X_i|\theta) \end{aligned}$$

Note: $f(X_1, \ldots, X_n|\theta)$ has n arguments. **Do not drop the indices on the X_i's!!!!!**

The product/joint distribution in i.i.d. case is called the likelihood function (or joined likelihood).

Example 1. Let X_i's be i.i.d. Bernoulli(p). $0 \leq p \leq 1$

$$\begin{gathered} P(X_i = 1) = p \qquad P(X_i = 0) = 1 - p \\ P(X_i = x_i|p) = p^{x_i}(1-p)^{1-x_i} \qquad \text{for } x_i = 0 \text{ or } 1 \end{gathered}$$

Suppose we observe a collection of points $X_1, \ldots, X_n$ and suppose that $X_1 = x_1, \ldots, X_n = x_n$. What is the probability of observing string of values

$$\begin{aligned} P(X_1, \ldots, X_n|p) &= \prod_{i=1}^{n} P(X_i = x_i|p) \\ &= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \\ &= p^{\sum_i x_i}(1-p)^{n-\sum_i x_i} \end{aligned}$$

Central question: What values of the parameter makes the observed data maximally likely? i.e. what value of the parameter maximizing the likelihood.

For maximizing likelihood, can take the log-likelihood as it is also monotonically increasing.

$$\begin{aligned} l(\theta) = \log l(p) &= \log(p^{\sum_i x_i}(1-p)^{n-\sum_i x_i}) \\ &= \big(\sum_i x_i\big) \log p + \big(n - \sum_i x_i\big) \log(1-p) \end{aligned}$$

This is a sufficiently smooth function of p so can consider finding maxima via critical points.

$$\begin{gathered} \frac{\partial L}{\partial p} = \frac{\sum_{i=1}^{n} X_i}{p} - \frac{n - \sum_i x_i}{1-p} = 0 \qquad \text{solve for } p \\ \frac{n - \sum_i x_i}{1-p} = \frac{\sum_{i=1}^{n} X_i}{p} \implies \boxed{\hat{p}_{\text{MLE}} = \frac{\sum_i X_i}{n} = \bar{X}} \end{gathered}$$

We already Know:

1. $\bar{X}$ is unbiased
2. $\bar{X}$ is consistent
3. $\bar{X}$ is asymptotically normal
4. $\bar{X}$ has variance $\frac{\sigma^2}{n}$

We will stem asymptotic analogues of trace properties for MLEs more general.

Lecture 9 (2018-10-01)

- MLEs - normal, gamma, uniform
- Modes of convergence
- Slutsky's Theorem
- Asymptotic properties of MLEs

MLEs - Normal Distribution

Let $X_i, 1 \leq i \leq n$ be i.i.d. $\mathcal{N}(\mu, \sigma^2)$ The <u>likelihood</u>

$$\begin{aligned}
f(x_1, \leq, x_n|\mu, \sigma^2) &= \prod_{i=1}^{n} f(x_i|\mu, \sigma^2) \\
&= \prod_{i=1}^{n} \left[\frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x_i-\mu)^2}{2\sigma^2}\right)\right] \\
&= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right) \\
\underline{\text{The log-likelihood:}} \qquad &= -n\log(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2
\end{aligned}$$

Maximize this w.r.t. μ, σ. Here log-likelihood depends smoothly on parameters $\to$ can consider critical points as 1st step in maximization.

$$\frac{\partial l}{\partial \mu} = \frac{2}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0 \Longrightarrow n\mu = \sum_{i=1}^{n} x_i \Longrightarrow \boxed{\hat{\mu}_{\text{MLE}} = \bar{X}}$$

$$\frac{\partial l}{\partial \sigma} = \frac{-n\cancel{\sqrt{2\pi}}}{\sigma\cancel{\sqrt{2\pi}}} - \frac{-1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2 = 0 \Longrightarrow \frac{n}{\sigma} = \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2 \Longrightarrow \boxed{\hat{\sigma}^2_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2}$$

Need to make sure two partial derivatives vanish simultaneously

$$\mu = \bar{X}$$

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 \Longrightarrow \boxed{\hat{\sigma}_{\text{MLE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2}}$$

Capital X_i's because want a function of the random variables in our sample. $E[\bar{X}] = \mu$. So $\hat{\mu}$ MLE is <u>unbiased</u>. $\text{Var}(\hat{\mu}_{\text{MLE}}) = \frac{\sigma^2}{n}$

Question: Is $E[\hat{\sigma}_{\text{MLE}}] = \sigma$?

$$\text{Next,} \quad \hat{\mu}_{\text{MLE}} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

Support

Given a density function $f(x|\theta)$, we define the support of f to be

$$\operatorname{supp} f = \{x : f(x|\theta) > 0\}$$

Suppose Θ is the space (in $\mathbb{R}, \mathbb{R}^d$) to which θ belongs:

$$\text{If } X_i\text{'s are i.i.d.. Bernoulli}(p)\text{, then } \Theta = (0,1), \quad \operatorname{supp} f = \{0,1\}$$
$$\text{If } X_i\text{'s are i.i.d.. } \mathcal{N}(\mu, \sigma^2)\text{, then } \Theta = \{(a,b) : a \in \mathbb{R}, b > 0\}, \quad \operatorname{supp} f = \mathbb{R}$$

We say that the supp f is independent of θ if

$$\{x : f(x|\theta) > 0\} \quad \text{is the same set for all } \theta \in \Theta$$

MLEs - Uniform Distribution

Now let X_i be i.i.d. Unif$[0, \theta]$ $\quad\theta > 0\quad$ Here supp f is not independent of θ.

$$\operatorname{supp} f = \{x : f(x|\theta) > 0\} \qquad f(x|\theta) \begin{cases} 0 & x < 0 \\ \frac{1}{\theta} & 0 \leq x \leq \theta \\ 0 & x >> \theta \end{cases}$$

Joint Likelihood

$$f(x_1, x_2, \ldots, x_n|\theta) = \underbrace{\frac{1}{\theta} \cdot \frac{1}{\theta} \cdots \frac{1}{\theta}}_{\text{n times}} = \left(\frac{1}{\theta}\right)^n$$

$$\text{with indicator} \qquad f(x_1, x_2, \ldots, x_n|\theta) = \left(\frac{1}{\theta}\right)^n \left(I_{[0,\theta]}(x_1) \cdot I_{[0,\theta]}(x_2) \cdots I_{[0,\theta]}(x_n)\right)$$
$$= \left(\frac{1}{\theta}\right)^n I_{\min(x_i) \geq 0,\ \max(x_i) \leq \theta}$$

Note: $\left(\frac{1}{\theta}\right)^n$ is decreasing in θ

- So want to choose θ as small as possible
- So note lower bound on θ in terms of x_i's if likelihood is to remain positive.

$$\boxed{\hat{\theta}_{\text{MLE}} = \max_{i \in \{1, \ldots, n\}} (x_i)}$$

Modes of Convergence

Let X be $unif[0,1]$. Let $g_n(x) = nI_{[0,1/n]}(x)$
Let $Y_n = g_n(X)$

$$\text{If } X = 0, g_n(0) = n \quad \text{(grows unboundedly)}$$
$$\text{If } X = x \in (0,1],\ g_n(x) \text{ is eventually 0.}$$

If $X > 0, g_n(X) \to 0$. $X = a > 0$. if n large enough so $\frac{1}{n} < a$, then $g_n(a) = 0$
For all ω except $\omega = 0$, $Y_n(\omega) \to 0 : P(\{\omega = 0\}) = 0$
So we have set A. A $= \{\omega : \omega > 0\}$ with $P(A) = 1$, such that $\forall \omega \in A$, $Y_n(\omega) \to 0$. So $Y_n \to 0$ with probability 1.

Lecture 10 (2018-10-03)

- Asymptotic properties of MLEs

Look at $\log f(x|\theta) \to$ Next, compute $\frac{\partial}{\partial\theta}\log f(x|\theta)$. Suppose $X_1, \ldots, X_n \sim$ i.i.d. $f(x|\theta)$

We are often concerned w/maximizing $\log f(x|\theta)$ as a function of θ.

Definition. Fisher Information: for a sample of size 1 from family $f(x|\theta)$. Denote $I(\theta)$, as follows

$$I(\theta) = \mathbb{E}\Bigg[\underbrace{\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2}_{\text{new r.v.}=y}\Bigg]$$

For a sample i.i.d. of size n,

$$\begin{aligned}\log f(x_1, x_2, \ldots, x_n|\theta) &= \log\prod_{i=1}^{n} f(x_i|\theta)\\ &= \sum_{i=1}^{n} f(x_i|\theta)\\ \text{So we find}\qquad \frac{\partial}{\partial\theta}\left(\sum_i \log f(x_i|\theta)\right) &= \sum_{i=1}^{n}\frac{\partial}{\log} f(x_i|\theta)\partial\theta\end{aligned}$$

Note that

$$E\left[\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right] \to \text{look the same for all } j \text{ by identical distribution}$$

So if we could

Lecture 11 (2018-10-10)

- MLEs - consistency
- Asymptotic normality

Question: Are MLEs always unbiased?
Answer: No,

$$\text{Consider} \qquad X_i \sim \text{ i.i.d. } \mathcal{N}(\mu, \sigma^2)$$
$$\text{MLE for } \sigma^2, \quad \hat{\sigma}^2 = \frac{1}{n}\sum(X_i - \bar{X})^2$$
$$E[S^2] = \sigma^2 \quad \text{where} \quad s^2 = \frac{\sum(X_i - \bar{X})}{n-1}$$
$$s^2 > \hat{\sigma}^2$$
$$E[\hat{\sigma}^2] < \hat{\sigma}^2$$
$$E[\hat{\sigma}^2] = \frac{n-1}{n}\sigma^2 \qquad \textbf{But} \qquad \mathbb{E}[\hat{\sigma}^2] \xrightarrow{n\to\infty} \sigma^2$$

So this estimator is asymptotically unbiased

$$\text{Bias}[\hat{\sigma}^2] = \left|\frac{n-1}{n}\sigma^2 - \sigma^2\right| \to 0 \quad \text{as} \quad n \to \infty$$

We will see arguments for why

1. MLEs are consistent
2. Asymptotically normal & asymptotically unbiased
3. Have a variance related to Fisher Information

Lecture 12 (2018-10-15)

NEED TO FINISH ATLEAST 8 LECTURES FROM BEFORE

- Modes of convergence; Slutsky's Theorem
- Asymptotic normality of MLEs
- Sufficiency
- Efficiency

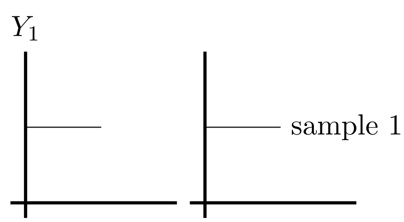
<u>4 Typical Modes of Convergence</u>

1. Convergence with probability 1
2. Convergence in probability
3. Convergence in L^P (expectation)
4. Convergence in distribution

$$Y_n = g_n(X) = n\mathbb{1}_{[0,\frac{1}{n})} \qquad X \sim unif[0,1]$$
$$Y_n \to y \quad \text{w.p. } 1 \quad \text{(away from zero) where } Y \equiv 0$$
$$g_n(X) = n^2\mathbb{1}_{[0,\frac{1}{n})}$$

1. Here $Y_n \to 0$ w.p. 1
2. $Y_n \to 0$ in probability
3. $E[|Y_n|] = n$ so $Y_n \to Y$ in Expectation or L^P for $p \geq 1$

<u>Exercise</u>: How can we construct a sequence Y_n s.t. $Y_n \to 0$ in probability but $Y_n \not\to 0$ w.p. 1?



For each $\omega \in (0,1)$ the Y_n's oscillate between 0 and 1, but the set of points at which Y_n is non-zero shrinks in probability.
Note: If $Y_n \to Y$ with probability 1, then $Y_n \to Y$ in probability, but converse is not necessarily true.

Theorem. *Slutsky's Theorem:*

1. *Suppose $X_n \to X$ in distribution ($X_n \xrightarrow{d} X$), $Y_n \to Y$ in probability. Then $X_n + Y_n \xrightarrow{d} X + Y$*
2. *If $X_n \xrightarrow{d} X$ and $Y_n \to c$ in probability: $X_n Y_n \xrightarrow{d} cX$*

<u>Why all this fuss?</u> Short answers: modes of convergence can be quire different!

Let's look at what happens to functions of random variables in particular:

Let $g : \mathbb{R} \to \mathbb{R}$ be smooth; and suppose $X_i \sim i.i.d. \quad f(x|\theta); \qquad \mu = \mathbb{E}[X_i]; \quad Var(X_i) = \sigma^2 < \infty$

So $\bar{X}$ is consistent for μ. Further, by CLT $\Rightarrow$

$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \underset{\text{approx}}{\sim} \mathcal{N}(0,1)$$
$$\frac{\sqrt{n}}{\sigma}(\bar{X} - \mu) \to \mathcal{N}(0,1)$$

How to understand approximatet/asymptotic behavior of $g(\bar{X})$? **Taylor expand** g about μ

$$g(x) \approx g(\mu) + g'(\mu)(x - \mu) + \frac{1}{2} g''(\mu)(x-\mu)^2$$

<u>Taylor's theorem with remainder:</u>

$$g(x) \approx g(\mu) + g'(\mu)(x - \mu) + \frac{g''(Z)}{2!}(x-\mu)^2$$

where Z is some point between x & μ

$$\Rightarrow g(\bar{X}) - g(\mu) = g'(\mu)(\bar{X} - \mu) + \frac{g''(Z)(\bar{X}-\mu)^2}{2!}$$
$$\sqrt{n}\big(g(\bar{X}) - g(\mu)\big) = \underbrace{\boxed{\sqrt{n} g'(\mu)(\bar{X} - \mu)}}_{\to \mathcal{N}(0,\text{some variance})} + \underbrace{\frac{\sqrt{n} g''(Z)(\bar{X}-\mu)^2}{2!}}_{?}$$

$$? = \sqrt{n} \underbrace{\frac{g''(Z)}{2!}}_{\substack{\text{suppose} \\ \text{we can bound} \\ \text{this piece}}} (\bar{X}-\mu)^2$$
$$\sqrt{n}(\bar{X} - \mu)^2 = \underbrace{\boxed{\sqrt{n}(\bar{X} - \mu)}}_{\substack{\text{converging} \\ \text{in distr} \\ \text{to normal}}} \underbrace{\boxed{(\bar{X} - \mu)}}_{\text{0 in prob.}}$$

So Slutsky's Theorem $\Rightarrow \sqrt{n}\big(g(\bar{X}) - g(\mu)\big) \to \mathcal{N}(0, \text{some variance})$

Recall our properties of MLE's from last week:

1 Consistency

2 Fisher information as a variance

3 Asymptotic normality: $\sqrt{nI(\theta_0)}\left(\hat{\theta}_{\text{MLE}} - \theta_0\right) \xrightarrow{d} \mathcal{N}(0,1)$

Let's look at $\ell(\theta) =$ log-likelihood

$$\text{MLE}: \quad 0 = \ell'(\hat{\theta})$$
$$\ell'(\theta) - \ell'(\theta_0) \approx \ell''(\theta_0)(\theta - \theta_0)$$

We conclude that for $\theta = \hat{\theta}$

$$\begin{aligned} \ell'(\hat{\theta}) &\approx \ell'(\theta_0) + \ell''(\theta_0)(\hat{\theta} - \theta_0) \\ \Rightarrow 0 &= \ell'(\theta_0) + \boxed{\ell'(\theta_0)}(\hat{\theta} - \theta_0) \end{aligned}$$

So if $\ell''(\theta_0) \neq 0$, we find

$$\boxed{(\hat{\theta} - \theta_0) \approx \frac{-\ell'(\theta_0)}{\ell''(\theta_0)}}$$

Now we can also write

$$\boxed{\sqrt{n}(\hat{\theta} - \theta_0) \approx -\frac{n^{-1/2}\ell'(\theta_0)}{n^{-1}\ell''(\theta_0)}}$$

$$\begin{aligned} \ell'(\theta) &= \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f(X_i|\theta_0) \\ &= \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log \Big(f(X_i|\theta_0) \Big) \Big|_{\theta=\theta_0} \\ \mathbb{E}[n^{-1/2}\ell'(\theta_0)] &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbb{E}\left[\frac{\partial}{\partial\theta} \log \Big(f(X_i|\theta_0) \Big) \Big|_{\theta=\theta_0} \right] = 0 \qquad \text{(by earlier result)} \\ Var\big(n^{-1/2}\ell'(\theta_0)\big) &= \frac{1}{n}\mathbb{E}\left[\left(\frac{\partial}{\partial\theta} \log \Big(f(X_i|\theta_0) \Big) \Big|_{\theta=\theta_0} \right)^2 \right] \end{aligned}$$

By independence of X_i's and Zero 1st moment of $\frac{\partial}{\partial\theta} \log f(X_i|\theta)\Big|_{\theta=\theta_0}$

$$\boxed{= I(\theta_0)}$$

<u>The denominator</u>:

$$\frac{1}{n}\ell''(\theta_0) = \frac{1}{n} \sum_{i=1}^{n} \underbrace{\left[\frac{\partial^2}{(\partial\theta)^2} \log f(X_i|\theta) \right]}_{Z_i} \Bigg|_{\theta=\theta_0}$$

Lecture 13 (2018-10-17)

- Asymptotic normality of MLEs (8.5)
- Efficiency & Sufficiency (8.7)
- Bayesian Estimation (8.6)

Suppose X_i are i.i.d. $f(x|\theta)$ where f satisfies regularity conditions 1) smoothness 2) $suppf$ is independent of θ)
Let $\hat{\theta}$ be MLE for θ suppose true value of θ is $\theta = \theta_0$. Then

$$\sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0) \xrightarrow[d]{n\to\infty} \mathcal{N}(0,1)$$

Note $Var(\hat{\theta})$ is asymptotically given by $\frac{1}{nI(\theta_0)}$

$$\sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0) = \frac{\hat{\theta} - \theta_0}{1/nI(\theta_0)}$$

Recall: (where $\ell(\theta)$ is log likelihood)

$$\sqrt{n}(\hat{\theta} - \theta_0) \approx \frac{n^{-1/2}\ell'(\theta_0)}{n^{-1}\ell''(\theta_0)}$$

Recall: last time we showed

$$Var(n^{1/2}\ell'(\theta_0)) = I(\theta_0)$$

Also the denominator is

$$\frac{1}{n}\ell''(\theta_0) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial^2}{(\partial\theta)^2}\log f(X_i|\theta)\right]\Bigg|_{\theta=\theta_0}$$

By LLN, this converges to

$$\mathbb{E}\left[\frac{\partial^2}{(\partial\theta)^2}\log f(X_i|\theta)\Bigg|_{\theta=\theta_0}\right] = +I(\theta_0)$$

So we've written

$$\sqrt{n}(\hat{\theta} - \theta_0) \approx \frac{W^{(n)}}{U^{(n)}}$$

We know that $U^{(n)} \to I(\theta_0)$ in probability But what is the numerator?

$$W^{(n)} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\underbrace{\left[\frac{\partial}{\partial\theta}\log f(X_i|\theta)\right]\Bigg|_{\theta=\theta_0}}_{Y_i}$$

Observe that Y_i's are ii, $E[Y_i] = 0; Var(Y_i) = I(\theta_0)$ So by CLT applied to $\frac{1}{\sqrt{n}}\sum\limits_{i=1}^{n} Y_i$, we find that

$$\frac{1}{\sqrt{nI(\theta_0)}}\sum Y_i \xrightarrow{d} \mathcal{N}(0,1)$$

So Slutsky's theorem $\Rightarrow \sqrt{n}(\hat{\theta}-\theta_0) \xrightarrow{d} \mathcal{N}(0,?)$ What is ?

So we've written

$$[\sqrt{n}(\hat{\theta}-\theta_0)]\sqrt{I(\theta_0)} \approx \frac{W^{(n)}}{U^{(n)}} \quad (\sqrt{I(\theta_0)})$$

Notice that $\dfrac{\sqrt{I(\theta_0)}}{U^{(n)}} \to \dfrac{1}{\sqrt{I(\theta_0)}}$ in probability

Note that $\dfrac{W^n}{\sqrt{I(\theta_0)}} = \dfrac{1}{\sqrt{I(\theta_0)}} \sum Y_i \longrightarrow \mathcal{N}(0,1)$

So what did we do?

1. First, we did a Taylor expansion (1st order) of log likelihood
2. We used that to write

$$\sqrt{n}(\hat{\theta}-\theta_0) \approx -\frac{n^{-1/2}\ell'(\theta_0)}{n^{-1}\ell''(\theta_0)}$$

Note: $$\sqrt{nI(\theta_0)}(\hat{\theta}-\theta_0) \approx \frac{\frac{1}{\sqrt{nI(\theta_0)}}\ell'(\theta_0)}{\boxed{\frac{-1}{I(\theta_0)} \cdot \frac{1}{n}\ell''(\theta_0)}}$$

3. We used Central Limit Theorem to conclude that

$$\frac{1}{\sqrt{n}}\ell'(\theta_0) \to \mathcal{N}(0, I(\theta_0))$$

4. By LLN, boxed piece converges in probability to $1/\sqrt{I(\theta_0)}$
5. By Slutsky's Theorem, $\sqrt{nI(\theta_0)}(\hat{\theta}-\theta_0) \xrightarrow[d]{n\to\infty} \mathcal{N}(0,1)$

Next: Surprising!

Suppose that $X_i \sim f(X_i|\theta)$ satisfying regularity conditions and let $T = r(X_1,\ldots,X_n)$ an estimator for θ Suppose that T is unbiased for θ. (T is not necessarily MLE or MOM...) Then

$$Var(T) \geq \frac{1}{nI(\theta)}$$

This is a remarkable <u>lower bound</u> on the variance of an unbiased estimator! An unbiased estimator T ($T = T_n = r(X_1,\ldots,X_n)$) Such that $Var(T_n) = \frac{1}{nI(\theta)}$ is said to be efficient

if $\dfrac{Var(T_n)}{1/nI(\theta_0)} \xrightarrow{n\to\infty} 1,$ then T_n is <u>asymptotically</u> efficient

<u>Relative Efficiency:</u> If we have two unbiased estimators $\hat{\theta}_1$ and $\hat{\theta}_2$, their relative efficiency is the ratio $\frac{Var(\hat{\theta}_1)}{Var(\hat{\theta}_2)}$

The asymptotic relative efficiency is the limit of this ratio as $n \to \infty$:

$$\lim_{n\to\infty} \frac{Var(\hat{\theta}_1)}{Var(\hat{\theta}_2)}$$

So far we've shown

1. MLEs are consistent
2. MLEs are asymptotically unbiased
3. MLEs are asymptotically normal
4. MLEs are asymptotically efficient

Sufficiency

Let $X_i \sim f(x|\theta)$. Suppose $T = r(X_1, \ldots, X_n)$ is a statistic (i.e. a function of $X_1, \ldots, X_n$) We say T is sufficient for θ if the conditional distribution of $X_1, \ldots, X_n$ given T is independent of θ

Theorem. *(Factorization)*
*A statistic T is sufficient for a parameter θ **iff** $f(x_1, \ldots, x_n|\theta) = g(T, \theta) \cdot h(X_1, \ldots, X_n)$*

Lecture 14 (2018-10-22)

Sufficiency: We say that a statistic T is sufficient for the parameter θ if the conditional distribution of the data $X_1, X_2, \ldots, X_n$ given T does not depend on θ.

Factorization Theorem: A statistic T is sufficient for a parameter θ ***iff*** the joint density can be factorized

$$f(x_1, \ldots, x_n|\theta) = g(T, \theta) \cdot h(X_1, \ldots, X_n)$$

Remark. Sufficient statistic need not be unique and many cases $h(x_1, \ldots, x_n) = 1$

Example 1. Let X_i be i.i.d. Bernoulli(p). Suppose $n = 3$. Let $T = X_1 + X_2 + X_3$.
Claim: T is sufficient for p. Let's look at an example

$$P(X_1 = 1, X_2 = 0, X_3 = 1|T = t) \begin{cases} 0 & \text{if } t \neq 2 \\ \frac{1}{\binom{3}{2}} & \text{if } t = 2 \end{cases} \quad \{t = 2\} = \frac{P(X_1 = 1, X_2 = 0, X_3 = 1|T = 2)}{P(T = 2)} = \frac{p^2 q}{\binom{3}{2} p^2 q}$$

Can also invoke Factorization:

$$\begin{aligned} p(x_1, \ldots, x_n|\theta) &= \theta^{\sum x_i} \cdot (1 - \theta)^{\sum x_i} \\ &= \underbrace{\Theta^T (1 - \Theta)^{n-T}}_{g(T,\theta)} \cdot \underbrace{1}_{h(x_1,\ldots,x_n)} \end{aligned}$$

Two Paradigms for Statistical Inference

1. **Frequentist:** parameters are unknown non-random variables.
   Goal: obtain estimate $T(X_1, \ldots, X_n)$ for this parameter and try to extract useful properties — consistency, asymptitic distributions, unbiasedness, minimum variance, ... Might want CIs for θ based on asymptotic distribution of T.

2. **Bayesian**: parameters are themselves random variables and these parameters have some probability distribution, $f_\lambda(\theta)$, this distribution might involve other parameters, called hyperparameters (often known).
   This distribution models uncertainty in your belief about θ. It is called a *prior.*

Next we have X_i i.i.d. $f(x|\theta)$. This is our data, and $f(x_1, \ldots, x_n|\theta)$ is our joined likelihood (common thread in both paradigms).

Goal: Use the observed data to recalculate conditional probabilities for θ given observed data i.e. to calculate a posterior distribution $f(\theta|x_1, \ldots, x_n)$

Then we use posterior distribution to extract information about θ include estimates (for θ):

1. posterior mean
2. posterior median
3. posterior mode

Example 2. Suppose X_i i.i.d. Bernoulli(p).

Suppose p satisfies a discrete prior:

$$p = \begin{cases} \frac{1}{4} & \text{w.p. } \frac{1}{3} \\ \frac{1}{2} & \text{w.p. } \frac{1}{3} \\ \frac{3}{4} & \text{w.p. } \frac{1}{3} \end{cases}$$

An example of continuous, "non-informative" prior:

$$p \sim \text{Unif}(0,1)$$

Given p, let $X_i \sim$ i.i.d Bernoulli(p) $\quad X_1 = 1, X_2 = 1, X_3 = 1, X_4 = 1$

Let's calculate posterior distribution of p:

$$P(p = p_0 | X_1 = 1, X_2 = 1, X_3 = 1, X_4 = 1) = \frac{P(p = p_0, X_1 = 1, X_2 = 1, X_3 = 1, X_4 = 1)}{\underbrace{P(X_1 = 1, X_2 = 1, X_3 = 1, X_4 = 1)}_{\text{function of the data}\to C(X_1,\ldots,X_n)}}$$

$$= \frac{P(X_1 = 1, X_2 = 1, X_3 = 1, X_4 = 1 | p = p_0) \cdot P(p = p_0)}{\sum\limits_{\text{all } a} P(X_1 = 1, X_2 = 1, X_3 = 1, X_4 = 1 | p = a) \cdot P(p = at)}$$

Now continue with the example

$$Pr(p = \frac{1}{4} | 1,1,1,1) = \frac{Pr(1,1,1,1 | p = \frac{1}{4}) \cdot Pr(p = \frac{1}{4})}{Pr(1,1,1,1)}$$
$$= \frac{(1/4)^4 \cdot (1/3)}{(1/4)^4 \cdot (1/3) + (1/2)^4 \cdot (1/3) + (3/4)^4 \cdot (1/3)} = u_1$$

$$Pr(p = \frac{1}{2} | 1,1,1,1) = \frac{Pr(1,1,1,1 | p = \frac{1}{2}) \cdot Pr(p = \frac{1}{2})}{Pr(1,1,1,1)}$$
$$= \frac{(1/2)^4 \cdot (1/3)}{(1/4)^4 \cdot (1/3) + (1/2)^4 \cdot (1/3) + (3/4)^4 \cdot (1/3)} = u_2$$

$$Pr(p = \frac{3}{4} | 1,1,1,1) = \frac{Pr(1,1,1,1 | p = \frac{3}{4}) \cdot Pr(p = \frac{3}{4})}{Pr(1,1,1,1)}$$
$$= \frac{(3/4)^4 \cdot (1/3)}{(1/4)^4 \cdot (1/3) + (1/2)^4 \cdot (1/3) + (3/4)^4 \cdot (1/3)} = u_3$$

$$p_{\text{post}} \begin{cases} 1/4 & u_1 \\ 1/2 & u_2 \\ 3/4 & u_3 \end{cases} \qquad \hat{p}_{\text{post}} = \frac{1}{4}u_1 + \frac{1}{2}u_2 + \frac{3}{4}u_3$$

Sometimes, we will find priors and posteriors and likelihoods such that prior and posterior belong to some family $\mathcal{F}$ and the likelihood belongs to $\mathcal{G}$. Here, we say $\mathcal{F}$, $\mathcal{G}$ are conjugate families of priors

Case when we have continuous distributions and want to obtain posterior densities:

$$f(\theta) = \text{prior density}$$
$$f(x_1, \ldots, x_n | \theta) = \text{ likelihood}$$

$$f(\theta|x_1,\ldots,x_n) = \frac{f(x_1,\ldots,x_n|\theta)\cdot f(\theta)}{\underbrace{\boxed{\int f(x_1,\ldots,x_n|\theta)\cdot f(\theta)d\theta}}_{\text{function of observed data}\to C(X_1,\ldots,X_n)}}$$

The denominator is a function of observed data i.e. it is a *normalizing constant in the posterior density*. Often we don't have to calculate it explicitly! **Note** that the posterior density depends on the data. It is however a density for θ. So often, we will want to manipulate the posterior density into a recognizable form as a function of θ with moments that might depend on the data.

Lecture 15 (2018-10-24)

- Bayesian Estimation
- Sufficiency
- Likelihood Ratio Tests

We want to estimate a mean θ, for i.i.d. normal data. Suppose that the variance is known. We have a normal likelihood.

Consider a normal prior distribution for θ. Need to specify a prior mean & a prior variance.

Suppose we have a prior mean of θ_0 and a prior variance of σ^2_{pr}. Let's write all expressions in terms of precision $\xi = 1/\sigma^2$

$$\boxed{\text{Prior:} \qquad f(\theta) = \frac{(\xi_{\text{prior}})^{\frac{1}{2}}}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\xi_{\text{prior}}(\theta - \theta_0)^2 \right)}$$

Likelihood: Suppose that θ, mean, is unknown but σ^2, variance, is known; $\sigma^2 = \sigma_0^2 \longleftrightarrow \xi_0 = 1/\sigma_0^2$

$$\boxed{f(x|\theta, \xi_0) = \left(\frac{\xi_0}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2}\xi_{\text{prior}}(x - \theta_0)^2 \right)}$$

Note that ξ_{pr} is a measure of our uncertainty about θ

Question: Once we calculate the posterior distribution, we updated our "belief" about θ. In this new "belief" — i.e. this new posterior distribution, do we have more precision or less?

Let $X_1, \dots, X_n \sim$ i.i.d. $f(x|\theta, \xi_0)$. Calculate $f(\theta|x_1, \dots, x_n)$.

$$f(\theta|x_1, \dots, x_n) = \frac{f(x_1, \dots, x_n|\theta, \xi_0) \cdot f(\theta)}{\underbrace{\int_\theta f(x_1, \dots, x_n|\theta, \xi_0) \cdot f(\theta) d\theta}_{\substack{C(x_1, \dots, x_n) - \text{normalizing constant} \\ \theta \text{ has been integrated out}}}}$$

Likelihood: $$f(x_1, \dots, x_n|\theta, \xi_0) = \left(\frac{\xi_0}{2\pi}\right)^{\frac{n}{2}} \exp\left(-\frac{\xi_0}{2}\sum_{i=1}^n (x_i - \theta)^2 \right)$$

Product: $$f(x_1, \dots, x_n|\theta, \xi_0) \cdot f(\theta) = \underbrace{\left(\frac{\xi_0}{2\pi}\right)^{\frac{n}{2}} \left(\frac{\xi_{\text{pr}}}{2\pi}\right)^{\frac{1}{2}}}_{C} \exp\left(-\underbrace{\left[\frac{\xi_{\text{pr}}}{2}(\theta - \theta_0)^2 + \frac{\xi_0}{2}\sum_{i=1}^n (x_i - \theta)^2\right]}_{Q(\theta)} \right)$$

So the posterior is of the form $C\exp(-Q(\theta))$ where Q is a quadratic - hence, normal!

$Q(\theta)$ will depend on θ, $\underbrace{\theta_0, \{x_1, \dots, x_n\}}_{\text{known!}}$

<u>Objective:</u> Force, through rough sheer of algebra, $Q(\theta)$ into the form. **Why?** Because the form of the product tells us the posterior density belongs to normal family - we now want to figure out mean and precision. We are going to force just by algebra, where each terms are calculable.

$$\left[\frac{\xi_{\text{post}}}{2}(\theta - \theta_{\text{post}})^2\right]$$

We have

$$\begin{aligned}
&= \frac{\xi_{\text{pr}}}{2}(\theta - \theta_0)^2 + \frac{\xi_0}{2}\sum_i (X_i - \theta)^2 \qquad \text{in the exponent} \\
&= \frac{\xi_{\text{pr}}}{2}(\theta^2 - 2\theta\theta_0 + \theta_0^2) + \frac{\xi_0}{2}\sum_i (x_i^2 - 2x_i\theta + \theta^2) \\
&= \underbrace{\left[\frac{\xi_{\text{pr}} + n\xi_0}{2}\right]}_{a}\theta^2 - \underbrace{(\theta_0\xi_{\text{pr}} + n\bar{X}\xi_0)}_{b}\theta + \underbrace{\boxed{\frac{\theta_0^2\xi_{\text{pr}}}{2} + \frac{\xi_0\sum_i x_i}{2}}}_{c} \approx a\theta^2 + b\theta + c
\end{aligned}$$

How do we work with this?

$$\begin{aligned}
a\theta^2 + b\theta + c &= a\left(\theta^2 - \frac{b}{a}\theta + \frac{c}{a}\right) \\
&= a\left(\theta - \frac{2b}{2a}\theta + \left(\frac{b}{2a}\right)^2 + \frac{c}{a} - \left(\frac{b}{2a}\right)^2\right) \leftrightarrow \boxed{\exp\left(-a\left(\theta - \frac{b}{2a}\right)^2\right.} + \text{STUFF}\Bigg)
\end{aligned}$$

So we get Normal with mean μ and precision ξ: $C\exp(-Q(\theta))$

$$a = \frac{\xi_{\text{pr}} + n\xi_0}{2}$$

So posterior precision:

1. $\xi_{\text{pr}} + n\xi_0 > \xi_{\text{pr}}$
2. As $n \to \infty$, ξ_{pr} matters less

$$\text{Posterior mean:} \qquad \frac{b}{2a} = \frac{\theta_0\xi_{\text{pr}} + n\bar{x}\xi_0}{\xi_{\text{pr}} + n\xi_0} = \theta_{\text{post}} = \frac{\theta_0\xi_{\text{pr}}}{\xi_{\text{pr}} + n\xi_0} + \frac{n\bar{x}\xi_0}{\xi_{\text{pr}} + n\xi_0}$$

$$f_{\text{post}} \sim \mathcal{N}\left(\frac{b}{2a}, 2a\right) \qquad \text{where its } \mathcal{N}(\text{mean, precision}) \qquad \overset{\text{as } n \to \infty}{\theta_{\text{post}} \text{ looks like } \bar{X}!}$$

Sufficiency in this Context

if T is sufficient for θ,

$$f(x_1, \dots, x_n|\theta) = g(T, \theta)h(x_1, \dots, x_n)$$

So the posterior distribution is

$$\frac{f(x_1, \dots, x_n|\theta) \cdot f(\theta)}{\int_\theta f(x_1, \dots, x_n|\theta)f(\theta)d\theta} = \frac{g(T,\theta)h(x_1, \dots, x_n)f(\theta)}{\int_\theta g(T,\theta)h(x_1, \dots, x_n)f(\theta)d\theta} = \frac{g(T,\theta)\cancel{h(x_1, \dots, x_n)}f(\theta)}{\cancel{h(x_1, \dots, x_n)}\int_\theta g(T,\theta)f(\theta)d\theta}$$

Posterior density depends on data ONLY through <u>sufficient statistic</u>

Lecture 16 (2018-10-29)

Hypothesis Testing

A hypothesis is a conjecture about a population parameter. Recall that in parametric inference we often consider X_i i.i.d. $f(X|\theta)$, where θ is the parameter and $\theta \in \Theta =$ parameter space

Example. 1. $X_i \sim \text{Bernoulli}(p) \qquad p \in (0,1)$

2. $X_i \sim \mathcal{N}(\mu, \sigma^2), \quad \theta = (\mu, \sigma^2), \quad \Theta = \mathbb{R}_x(0, \infty)$

3. $X_i \sim \text{Unif}[0, \infty]$ and $\theta > 0$ so $\Theta = \mathbb{R}^+$

Hypothesis typically take the form (in the frequentist interpretation)

$$\theta = \theta_0$$
$$\text{or} \quad \theta \in \text{H}_0 \subset \Theta$$

We say that a hypothesis H is simple if it fully determines the distribution $f(x|\theta)$

Example. H: $\theta = 4$ in uniform case, then $f(x|\theta) = \frac{1}{4} I_{(0,4)}(x)$
H: $\theta > 3$ NOT SIMPLE

Any non simple hypothesis is called composite. Typically we want to evaluate a pair of competing conjecture.

We let these be denoted by H_0, the so called *NULL*, and H_1, the so called *ALTERNATE*.

In the frequentist framework, parameters are not random. Consider an especially simple starting point:

$$\text{H} = \text{H}_0 \bigcup \text{H}_a$$
$$H_0 : \theta \in \text{H}_0 \qquad \text{is simple}$$
$$H_a : \theta \in \text{H}_a \qquad \text{is simple}$$

Questions: Is the data we observe more likely under H_0 or under H_a

That is, what is the likelihood under H_0 and what is the likelihood under H_a and how do they compare?

$$H_0 : \theta = \theta_0$$
$$H_a : \theta = \theta_a$$

$$\textbf{Likelihood Ratio (LR)} : \quad \frac{f(x_1, \ldots, x_n|\theta_0)}{f(x_1, \ldots, x_n|\theta_a)}$$

If LR is large, suggests observed data more likely under H_0 so LR gives us a decision rule - $T(X_1, \ldots, X_n)$ - where T is binary either reject H_0 or fail to reject H_0.

Decision Rule may be in correct for a given string of data you might fail to reject H_0 when H_a is true or reject H_0 when H_0 is true

Types of Error

1. **Type I Error**: Reject H_0 when H_0 is true
2. **Type II Error**: Fail to reject H_0 when H_0 is false

It can be challenging to simultaneously control both.

$$\alpha = P(\text{Type I Error})$$
$$\beta = P(\text{Type II Error})$$

Instead, we will set a tolerance for the probability of Type I Error, α, called the significance level of the test, and we will look for the decision rule that satisfies this tolerance and also minimizes the probability of Type II Error.

Note: Decision rule to always accept H_0 has no Type I Error, but might have high probability of Type II Error.

There are many possible decision rules $T(X_1, \ldots, X_n)$. How to both control Type I error and Type II error? $\rightarrow$ look at likelihood

Supposed we say P(Type I Error) $\leq \alpha$. This will help us determine a rejection region: a set of values of data for which H_0 is rejected.

$$\text{Let} \quad d(X_1, \ldots, X_n) = \begin{cases} 0 & \text{if we do not reject } H_0 \\ 1 & \text{if we reject } H_0 \end{cases}$$

$$\textbf{LRT:} \quad \frac{f(X_1, \ldots, X_n|\theta_0)}{f(X_1, \ldots, X_n|\theta_n)} = g(X_1, \ldots, X_n; \theta_0, \theta_n)$$

We want to reject H_0 for observed data in which

$$\frac{f(X_1, \ldots, X_n|\theta_0)}{f(X_1, \ldots, X_n|\theta_n)} \quad \text{is small}$$

i.e. we want to choose a constant c s.t.

$$P\left(\frac{f(X_1, \ldots, X_n|\theta_0)}{f(X_1, \ldots, X_n|\theta_n)} \leq c \middle| H_0\right) \leq \alpha$$

Observe that the LRT depends on data and the specific non-random values $\theta_0 + \theta_a$. But to determine the critical value c, we only need to know the distribution of the data under H_0.

Example. X_i i.i.d. Bernoulli(p)

$$\left.\begin{aligned} H_0: &\quad \theta = p = p_0 \\ H_a: &\quad \theta = p = p_a \end{aligned}\right\} p_0 > p_a$$

$$\textbf{LRT} \quad \frac{p_0^{\sum X_i}(1-p_0)^{\sum(1-X_i)}}{p_a^{\sum X_i}(1-p_a)^{\sum(1-X_i)}} \qquad \text{where } n = \text{sample size}$$

Rejecting for small values of LRT i.e. when LRT $= c.$ is equivalent to rejecting when $\ln(LRT) = \ln(c) = d$

$$\text{Taking logs we get} \qquad \ln\left(p_0^{\sum X_i}(1-p_0)^{\sum(1-X_i)}\right) - \ln\left(p_a^{\sum X_i}(1-p_a)^{\sum(1-X_i)}\right)$$
$$= (\ln p_0 - \ln p_a)\sum_i X_i + [\ln(1-p_0) - \ln(1-p_a)]\left(\sum_i (1-X_i)\right)$$

Want this to be bounded from above in order to determine a critical region or rejection region

We expect to reject H_0 for small values on $\sum X_i$

$$\ln\left(\frac{p_0}{p_a}\right)\sum_i X_i + \ln\left(\frac{1-p_0}{1-p_a}\right)\sum_i (1-X_i) \leq d$$
$$\sum_i X_i\left[\ln\left(\frac{p_0}{p_a}\right) - \ln\left(\frac{1-p_0}{1-p_a}\right)\right] \leq d - n\ln\left(\frac{1-p_0}{1-p_a}\right)$$

So we reject if

$$\sum_i X_i \leq \underbrace{d - n\ln\left(\frac{1-p_0}{1-p_a}\right)}_{D}$$
$$\text{We want} \qquad P\left(\sum_i X_i \leq D|H_0\right) \leq \alpha$$

Suppose $\alpha = 0.05$. Note that under H_0 $\sum_i X_i \sim \text{Bin}(n, p_0)$. So can determine D such that $P(\sum X_i \leq D) \leq 0.05$

Now suppose that we have determined C for our rejection region observe that probability of **Type II Error** is given by

$$P(\text{LRT} > C|H_a)$$

Power

$$\text{Power} = 1 - P(\textbf{Type II Error})$$

Lecture 17 (2018-10-31)

- Neyman-Pearson
- Uniformly most powerfiul tests
- GLRTs

Neyman-Pearson Lemma

Theorem. *(The Neyman-Pearson)*
Let H_0, H_1 be simple, let $H_0 \bigcup H_a = H$. Suppose LRT rejects H_0 when $LR \leq C$ and that this test procedure hhas significance level α. Consider any other test with significance less than or equal to α. The power of this test is less than or equal to power of LRT.

Proof. Since H_0, H_a (or H_1) are both simple, let $f_0(x)$, $F_1(x)$ denote the respective densities under null and alternative. Any decision rule is of the form

$$d(x) = \begin{cases} 0 & \text{if } H_0 \text{ accepted} \\ 1 & \text{if } H_0 \text{ rejected} \end{cases}$$

$$\text{Note that} \qquad \mathbb{E}[d(\underline{X})] = P(d(\underline{X}) = 1)$$

$$\text{Note that significance level:} \qquad P(d(\underline{X}) = 1 | H_0) = \mathbb{E}[d(\underline{X})]$$

$$\textbf{Power:} \quad 1 - \beta = 1 - P(\text{Type II Error}) = P(d(\underline{X}) = 1 | H_1) = E_1(d(\underline{X}))$$

Now, let's consider the particular decision rule given by LRT

$$\textbf{Reject } H_0 \textbf{ if} \quad \frac{f_0(\underline{X})}{f_1(\underline{X})} < c \qquad c \text{ is chosen so that } P(\text{Type II Error}) = \alpha$$

$$E_0[d(\underline{X})] = \alpha \qquad \text{where } d(\underline{X}) \text{ is the LRT decision rule.}$$

Let d^* be any other decision rule with at most α as Type I error: $\mathbb{E}_0[d^*(X)] \leq \alpha$

It suffices to show:

$$\underbrace{\mathbb{E}_1[d^*(\underline{X})]}_{\text{power of } d^*} = \underbrace{\mathbb{E}_1[d^{(}\underline{X})]}_{\text{power of LRT}}$$

■

Key Inequality

$$d^*(\underline{x})[cf_1(\underline{x}) - f_0(\underline{x})] \leq \underbrace{d(\underline{x})}_{LRT}[cf_1(\underline{x}) - f_0(\underline{x})]$$

We reject LRT, i.e. $d(\underline{x}) = 1$, when

$$f_0(\underline{x}) < xf_1(\underline{x})$$
$$cf_1(\underline{x}) - f_0(\underline{x}) > 0$$

So if $\underline{x}$ is such that $d(\underline{x}) = 1$, then

$$cf_1(\underline{x}) - f_0(\underline{x}) > 0$$

$$\text{and} \qquad d^*(\underline{x})[cf_1(\underline{x}) - f_0(\underline{x})] \le cf_1(\underline{x}) - f_0(\underline{x})$$

If $\underline{x}$ is such that $d(\underline{x}) = 0$, then $d(\underline{x})[cf_1(\underline{x}) - f_0(\underline{x})] = 0$.
But also since $d(\underline{x}) = 0$, $cf_1(\underline{x}) - f_0(\underline{x}) \le 0$

Thus we now consider two options - either $d^*(\underline{x}) = 0$, in which case:

$$d^*(\underline{x})[cf_1(\underline{x}) - f_0(\underline{x})] = 0 \qquad \text{which leads to } 0 = 0$$

If $d(\underline{x}) = 0$ & $d^*(\underline{x}) = 1$, we have

$$\underbrace{d^*(\underline{x})}_{1}\underbrace{[cf_1(\underline{x}) - f_0(\underline{x})]}_{\text{non-positive}} \le 0 = \overbrace{d(\underline{x})}^{0}[cf_1(\underline{x}) - f_0(\underline{x})]$$

$$\text{so we find} \qquad cd^*(\underline{x})f_1(\underline{x}) - d^*(\underline{x})f_0(\underline{x}) \leqslant cd(\underline{x})f_1(\underline{x}) - d(\underline{x})f_0(x)$$

Let's note that, integrating over possible values $x_1, \ldots, x_n$ in the vector $\underline{x} = (x_1, \ldots, x_n)$

$$c\mathbb{E}_1(d^*(X)) - \mathbb{E}_0(d^*(\underline{X})) \le c\mathbb{E}_1(d(\underline{X})) - \mathbb{E}_0(d(\underline{X}))$$

$$\text{so note that} \qquad \underbrace{\mathbb{E}_0(d^*(\underline{X})) - \mathbb{E}_0(d(\underline{X}))}_{\text{-ve if } d^* \text{ has small TI error than d}} \geqslant c\Big(\mathbb{E}_1(d^*(X)) - \mathbb{E}_1(d(\underline{X}))\Big)$$

In which case

$$\mathbb{E}_1(d^*(X)) - \mathbb{E}_1(d(\underline{X})) < 0$$
$$\mathbb{E}_1(d^*(X)) < \mathbb{E}_1(d(\underline{X}))$$

Most powerful test

Example 1. X_i i.i.d. $\mathcal{N}(\mu, \sigma^2)$, suppose σ^2 is known

$$\text{Consider } H_0\colon \mu = \mu_0$$
$$H_a\colon \mu = \mu_a$$
$$\text{Folk wisdom: use } \bar{X} \text{ as T.S.}$$

$$LRT = \frac{f_0(\underline{X})}{f_1(\underline{X})} = \frac{\cancel{\left(\frac{1}{\sigma\sqrt{(2\pi)}}\right)^n} \exp\left\{\frac{-\sum_i (x_i - \mu_0)^2}{2\sigma^2}\right\}}{\cancel{\left(\frac{1}{\sigma\sqrt{(2\pi)}}\right)^n} \exp\left\{\frac{-\sum_i (x_i - \mu_a)^2}{2\sigma^2}\right\}}$$

$$\text{take logs :} \qquad \frac{-\sum_i (x_i - \mu_0)^2}{2\sigma^2} + \frac{-\sum_i (x_i - \mu_a)^2}{2\sigma^2} \le d$$

$$\text{Reject if} \qquad 2\bar{X}n\mu_0 - n\mu_0^2 - 2\bar{X}n\mu_a + \mu_a^2 \le d'$$
$$= 2n\bar{X}(\mu_0 - \mu_a) + n(\mu_a^2 - \mu_0^2) \le d'$$

Reject H_0 if

$$\bar{X}(\mu_0 - \mu_a) \le \frac{d' - n(\mu_a^2 - \mu_0^2)}{2n}$$

Since $\mu_a > \mu_0$, we find reject H_0 if

$$\bar{X} \geq \boxed{\frac{d' - n(\mu_a^2 - \mu_0^2)}{2n(\mu_0 - \mu_a)}} \qquad (\star)$$

i.e. we reject H_0 if $\bar{X}$ is sufficiently large. $\star$ looks complicated like it depends on μ_0, μ_a, etc.

$$P(\text{Reject } H_0 | H_0) = \alpha$$
$$P(\bar{X} > \star | H_0) = \alpha$$
we know from previous lectures that $\bar{X} \sim \mathcal{N}(\mu_0, \sigma^2/n)$
$$= P\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} > \frac{\star - \mu_0}{\sigma/\sqrt{n}} \middle| H_0\right) = \alpha$$
$$= P\left(Z > \frac{\star - \mu_0}{\sigma/\sqrt{n}}\right) = \alpha$$

So

$$\boxed{\star = Z_\alpha \frac{\sigma}{\sqrt{n}} + \mu_0}$$

So we reject if $\bar{X} > Z_\alpha \frac{\sigma}{\sqrt{n}} + \mu_0$ and we note that this rejection region is not dependent on explicit value of μ_a, as long as $\mu_a > \mu_0$

So note that the exact same test (reject H_0 if $\bar{X} > Z_\alpha \frac{\sigma}{\sqrt{n}} + \mu_0$) is most powerful for $H_0 : \mu = \mu_0$ vs $H_a : \mu = \mu_a$ for any choice of $\mu_a > \mu_0$.

So this is a uniformly most powerful test (UMP) for

$$H_0 : \mu = \mu_0 \text{ (simple } H_a)$$
$$\text{vs } H_0 : \mu > \mu_0 \text{ (comp. } H_a)$$

Lecture 18 (2018-11-05)

- Hypothesis tests/Confidence Intervals.
- Bayesian HTs.
- GLRTs and Wilks Theorem.
- Distributions based on normal.
- Midterm II next Wednesday

Confidence Intervals

Last time we considered $X_i \sim$ i.i.d. $\mathcal{N}(\mu, \sigma^2)$, σ^2 known

$$H_0 : \mu = \mu_0 \qquad H_a : \mu > \mu_0$$

We found that if we considered

$$\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \quad \text{as our test statistic with rejection region}$$

$$T > Z_\alpha, \qquad \underbrace{\text{i.e. reject } H_0 \text{ if } \bar{X} > Z_\alpha \frac{\sigma}{\sqrt{n}} + \mu_0}_{\text{test procedure is uniformly most powerful}}$$

Now, what if we had instead considered a two sided test?

$$H_0 : \mu = \mu_0 \qquad H_a : \mu \neq \mu_0$$

Here might consider rejecting H_0 if $|\bar{X} - \mu_0|$ is sufficiently large. i.e.

$$\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1) \text{ under } H_0$$

$$\text{So } \textbf{reject} \text{ if } \quad \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < -Z_{\alpha/2} \quad \text{or} \quad \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} > Z_{\alpha/2}$$

So we reject if

$$\bar{X} < -Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} + \mu_0$$

$$\bar{X} > Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} + \mu_0$$

So we accept H_0 if

$$\boxed{-Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} + \mu_0 < \bar{X} < Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} + \mu_0}$$

Suppose we want a random interval that contains the population parameter μ (whatever its value) with probability $1 - \alpha$ i.e. suppose we have some population parameter (in this case μ) whose value we'd like to estimate.

A $(100)(1-\alpha)$ % C.I. for μ is a random interval containing μ with specified probability $1-\alpha$

Note that a CI for μ a level α looks like

$$\left(\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

and we can say we accept $H_0 : \mu = \mu_0$ when $100(1-\alpha)\%$ CI given by $\left(\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$ contains μ_0

That is, we have a duality between CIs and HTs.

Theorem. *Suppose for every $\theta_0 \in \Theta$, $\exists$ a level α test of $H_0 : \theta = \theta_0$. Suppose $A(\theta_0) = \{\underline{X}$: decision rule is to accept $H_0\}$. Then let $C(X) = \{\theta \in \Theta : X \in A(\theta)\}$. Then $C(X)$ is a $100(1-\alpha)\%$ confidence region for θ.*

Conversely,

Theorem. *Suppose that $C(X)$ is a $100(1-\alpha)\%$ confidence region for θ. i.e.*

$$P(\theta_0 \in C(X)|\theta = \theta_0) = 1 - \alpha$$

for every θ_0. Then if we define

$$A(\theta_0) = \{\underline{X} : \theta_0 \in C(X)\}$$

this is an acceptance region for a level α test of $H_0 : \theta = \theta_0$

Bayesian Hypotheses Tests

Consider X_i's i.i.d. $f(x|\theta)$, suppose we now have a probability distribution over hypotheses: let H_0 and H_1 be two simple null and alternative hypotheses (respectively) and let $\pi_0 = P(H_0)$ and $\pi_1 = P(H_1)$. So in the Bayesian framework, we observe a vector of data and then update π_0 and π_1

$$\text{Compute} \quad P(H_1|X_1,\ldots,X_n) = \frac{P(X_1,\ldots,X_n|H_1)\pi_1}{P(X_1,\ldots,X_n|H_0)\pi_0 + P(X_1,\ldots,X_n|H_1)\pi_1}$$
$$P(H_0|X_1,\ldots,X_n) = \frac{P(X_1,\ldots,X_n|H_0)\pi_0}{P(X_1,\ldots,X_n|H_0)\pi_0 + P(X_1,\ldots,X_n|H_1)\pi_1}$$
$$\text{Decision Rule: } P(H_0|X_1,\ldots,X_n) > P(H_1|X_1,\ldots,X_n)$$

Observe that

$$\frac{P(H_0|X_1,\ldots,X_n)}{P(H_1|X_1,\ldots,X_n)} = \frac{P(X_1,\ldots,X_n|H_0)\pi_0}{P(X_1,\ldots,X_n|H_1)\pi_1}$$

Accept H_0 if $\star$ is greater than some constant.

So we are still comparing likelihoods, i.e. computing a L.R.

Detour now into Rice, Ch 6

Distributions derived from the normal distribution.

Suppose $\underline{X} \in \mathbb{R}^d$ has a multivariate normal distribution, so $\underline{X}$ has density

$$f_{\underline{X}}(\underline{t}) = C \exp\left(-\frac{1}{2}(\underline{t}-\underline{\mu})^T \Sigma^{-1} (\underline{t}-\underline{\mu}) \right) \quad \text{where} \quad \underline{t} = \begin{bmatrix} t_1 \\ \vdots \\ t_d \end{bmatrix} \quad \underline{\mu} = \begin{bmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_d] \end{bmatrix} \qquad \Sigma_{ij} = cov(X_i, X_j)$$

Facts (lemmas):

1. If $\underline{X} \sim$ jointly normal and $\sigma_{ij} = 0$, then X_i, X_j are independent. t
2. If $\underline{X}$ is normal and $O \in \theta(dxd)$, so $O^T O = I$ the OX is normal (invariance of normality under rotation).
3. If $X_1, \dots, X_d$ are separately normal and independent, then $\underline{X} = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}$ is jointly normal and Σ is diagonal.
4. if $\underline{X} = (X_1, \dots, X_n)$ where X_i's are i.i.d. normal then $\bar{X}$ and s^2 are independent.

Chi-squared

If $Z_1, \dots, Z_n$ are i.i.d. $\mathcal{N}(0,1)$, then

$$\sum_{i=1}^n Z_i^2 \sim \chi^2 \quad \text{n degrees of freedom}$$

Degrees of Freedom in a χ^2 correspond to number of independent squared normals in the sum.

Finally, if $X_1, \dots, X_n$ is i.i.d. $\mathcal{N}(0,1)$, then

$$(n-1)s^2 = (n-1)\left[\frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2\right] = \sum_{i=1}^n (X_i - \bar{X})^2 \sim \chi^2_{n-1}$$

Lecture 19 (2018-11-07)

Lemma. *In this case $\bar{X}$ is independent of $(X_1 - \bar{X}, X_1 - \bar{X}, \dots, X_n - \bar{X})$*

Lemma. *IF $X_i s$ are i.i.d $\mathcal{N}(\mu, \sigma^2)$ then $\bar{X}$ and s^2 are independent (follows immediately from earlier lemma).*

Lemma. *If $X_i \sim$ i.i.d $\mathcal{N}(\mu, \sigma^2)$ then $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2 \quad (n-1$ degrees of freedom)*

So if $X_i, \dots, X_n \sim$ i.i.d. $\mathcal{N}(\mu, \sigma^2)$

$$\text{We know} \quad \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$
$$\text{We also know} \quad \frac{s^2(n-1)}{\sigma^2} \sim \chi^2(n-1 \quad df)$$
$$\frac{s^2(n-1)}{\sigma^2} \quad \text{is independent of} \quad \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

So we find

$$\frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{s^2(n-1)}{\sigma^2}}} \sim t(n-1 \quad df)$$
$$= \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{s/\sigma} = \boxed{\frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t(n-1 \quad df)}$$

So we want to understand hypotheses about μ e.g.

$$H_0 : \mu = \mu_0, \quad \sigma^2 \text{ unknown (composite null)}$$
$$\text{vs } H_1 : \mu \neq \mu_0, \quad \sigma^2 \text{ unknown (composite altetrnate)}$$

But to build a framework for testing such hypotheses, we will consider generalized likelihood ratio test. Suppose we would like to test:

$$H_0 : \theta \in \Theta_0$$
$$H_a : \theta \in \Theta_a$$
$$\text{Suppose} \quad \Theta = \Theta_0 \bigcup \Theta_a$$

Consider l = likelihood: $l(X_1, \dots, X_n | \theta)$. Define:

$$\Lambda^* = \frac{\max_{\theta \in \Theta_0} l(X_1, \dots, X_n | \theta)}{\max_{\theta \in \Theta} l(X_1, \dots, X_n | \theta)}$$

Λ^* is called the generalized ratio and the test procedure in which we reject H_0 if $\Lambda^* \leq c$ is called GLRT or generalized likelihood ratio test.

In out class, Θ_0, and Θ_a will generally be "nice" subsets of Euclidean space, whose dimension is well defined and straight forward to calculate.

Example 1.

$$\Theta = \{(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}$$
$$\dim \Theta = 2, \text{ if } \Theta_0 = \{\theta : \mu = \mu_0, \sigma^2 = \sigma_0^2\}$$
$$\text{Note} \quad \dim \Theta_0 = 0$$

Suppose we consider $\mu > \mu_0$, $\sigma^2 = \sigma_0^2$

$$\text{Then if} \quad \Theta_0 = \{\theta : \mu > \mu_0, \sigma^2 = \sigma_0^2\}, \quad \dim \Theta_0 = 1$$
$$\text{If} \quad \Theta_0 = \{\theta : \mu > \mu_0, \sigma^2 > \sigma_0^2\}, \quad \dim \Theta_0 = 2$$

Theorem. *Under Certain regularity conditions, $-2 \log \Lambda^*$ has $n \to \infty$, an asymptotic distribution given by χ^2 $(\dim \Theta - \dim \Theta_0)$*

Example 2. Let $X_1, \ldots, X_n \sim$ i.i.d. $\mathcal{N}(\mu, \sigma^2)$. Consider

$$H_0 : \mu = \mu_0$$
$$H_a : \mu \leq \mu_0$$

Suppose σ^2 is known (all this happens when null is true)

$$\Theta = \{\mu \in \mathbb{R}\}, \quad \dim \Theta = 1 \text{ and } \dim \Theta_0 = 0$$
$$\Lambda^* = \frac{\max_{\theta \in \Theta_0} l(X_1, \ldots, X_n | \theta)}{\max_{\theta \in \Theta} l(X_1, \ldots, X_n | \theta)}$$
$$\Lambda^* = \frac{f(x_1, \ldots, x_n | \mu_0, \sigma^2)}{\max_{\mu} f(x_1, \ldots, x_n | \mu_0, \sigma^2)}$$
$$\vdots$$
$$\log \Lambda^* = \frac{-1}{2\sigma^2} \sum (X_i - \mu_0)^2 + \frac{1}{2\sigma^2} \sum (X_i - \bar{X})^2$$
$$\vdots$$
$$\begin{aligned}
&= \frac{2\mu_0}{2\sigma^2} n\bar{X} - \frac{n\mu_0^2}{2\sigma^2} - \frac{n\bar{X}^2}{2\sigma^2} \\
&= \frac{-n}{2\sigma^2} (\bar{X}^2 - 2\mu\bar{X} + \mu_0^2) \\
&= \frac{-n}{2\sigma^2} (\bar{X} - \mu_0)^2 \\
&\frac{-1}{2} \underbrace{\left(\frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \right)^2}_{\sim \mathcal{N}(0,1) \textbf{ under } H_0}
\end{aligned}$$

Hence Wilks theorem

$$-2 \log \Lambda^* = \left(\frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \right)^2 \sim \chi^2_{1df}$$

<u>Wilks</u> under regularity conditions $-2 \log \Lambda^*$ converges in distribution under H_0 to $\chi^2(\dim \Theta - \dim \Theta_0)$

Lecture 20 (2018-11-12)

- Distributions derived from the normal (χ^2, F, t)
- Analysis of Variance

Recall if X_i's are i.i.d. $\mathcal{N}(0,1)$, then

$$\sum_{i=1}^{m} X_i^2 \sim \chi^2(m \text{ df})$$

Fact if $U \sim \chi^2(n_1 \text{ df}), V \sim \chi^2(n_2 \text{ df})$ U, V are independent

$$U + V \sim \chi^2(n_1 + n_2 \text{ df})$$

Recall If $U \sim \mathcal{N}(0,1)$ and $V \sim \chi^2(v \text{ df})$ where U, V are independent, then

$$\frac{U}{\sqrt{V/\nu}} \sim t_{\nu \text{ df}}$$

We saw that if X_i are i.i.d. $\mathcal{N}(\mu, \sigma^2)$, then

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1 \text{ df}}$$

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_i (X_i - \bar{X})^2}$$

Finally, suppose $W_1 \sim \chi^2_{n_1 \text{ df}}$ and $W_2 \sim \chi^2_{n_2 \text{ df}}$ with W_1 and W_2 independent.

$$\frac{W_1/n_1}{W_2/n_2} \sim F(n_1, n_2)$$

Note that if $Y \sim F(n_1, n_2)$, then $\frac{1}{Y} \sim F(n_2, n_1)$

Suppose $X_1, \ldots, X_n$ are i.i.d. $\mathcal{N}(\mu, \sigma^2)$ (both parameters unknown)
How to test

$$H_0 : \mu = \mu_0 \qquad \sigma^2 > 0$$
$$H_a : \mu \neq \mu_0 \qquad \sigma^2 > 0$$
$$\Theta = \{(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0\}, \quad \Theta_0 = \{(\mu_0, \sigma^2), \sigma^2 > 0\}$$

So $\theta = (\mu, \sigma^2)$, we will consider

$$\frac{\max\limits_{\theta \in \Theta_0} l(X_1, \ldots, X_n|\theta)}{\max\limits_{\theta \in \Theta} l(X_1, \ldots, X_n|\theta)} = \frac{\cancel{\left(\frac{1}{\sigma\sqrt{(2\pi)}}\right)^n} \exp\left\{\frac{-\sum_i (x_i - \mu_0)^2}{2\sigma^2}\right\}}{\cancel{\left(\frac{1}{\sigma\sqrt{(2\pi)}}\right)^n} \exp\left\{\frac{-\sum_i (x_i - \mu)^2}{2\sigma^2}\right\}}$$

Denominator recall MLEs for μ, σ^2 in normalcase

$$\hat{\mu}_{\text{MLE}} = \bar{X}$$

$$\sigma^2_{\text{MLE}} = \frac{1}{n}\sum_i (X_i - \bar{X})^2$$

Numerator MLE for σ^2 when $\mu = \mu_0$: $\frac{1}{n}\sum_i (X_i - \mu_0)^2$

Numerator of GLRT

$$\left[\frac{1}{\sqrt{\frac{1}{n}\sum_i (X_i - \mu_0)^2}}\right]^n \exp\left(- \frac{\sum_i (X_i - \mu_0)^2}{\frac{2}{n}\sum_i (X_i - \mu_0)^2}\right) \cdot \left(\frac{1}{\sqrt{2\pi}}\right)^n$$

Denominator of GLRT

$$\left[\frac{1}{\sqrt{\frac{1}{n}\sum_i (X_i - \bar{X})^2}}\right]^n \exp\left(- \frac{\sum_i (X_i - \bar{X})^2}{\frac{2}{n}\sum_i (X_i - \bar{X})^2}\right) \cdot \left(\frac{1}{\sqrt{2\pi}}\right)^n$$

So GLRT looks like

$$\frac{\left(\sqrt{\frac{1}{n}\sum_i (X_i - \mu_0)^2}\right)^n}{\left(\sqrt{\frac{1}{n}\sum_i (X_i - \bar{X})^2}\right)^n}$$

We reject when GLR $\leq c$. Equivalent to rejecting H_0 when

$$\frac{\cancel{1/n}\sum \frac{(X_i - \bar{X})^2}{\sigma^2}}{\cancel{1/n}\sum \frac{(X_i - \mu_0)^2}{\sigma^2}} \qquad \text{is small}$$

Observe that under H_0,

$$\sum_{i=1}^n \frac{(X_i - \mu_0)^2}{\sigma^2} \sim \chi^2_{n \text{ df}}$$

Furthermore,

$$\sum_{i=1}^n \frac{(X_i - \bar{X})^2}{\sigma^2} = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1 \text{ df}}$$

But last week, we claimed that we would reject H_0 for large absolute values of $\frac{\bar{X}-\mu_0}{s/\sqrt{n}}$. Rejecting for large absolute values of this is same as rejecting for large values of

$$\left(\frac{\bar{X} - \mu_0}{s/\sqrt{n}}\right)^2 \sim (t_{n-1 \text{ df}})^2$$

"Fun" Fact:

$$U \sim t_{n-1 \text{ df}}, \qquad U = \frac{Z}{\sqrt{V/\nu}}$$

$$\text{then} \quad U^2 \sim F(1, \nu_1) \qquad U^2 = \frac{Z^2}{V/\nu_1} = \frac{Z^2/1}{V/\nu_1}$$

Suppose now, we have data from two normal populations.

$$X_1, \ldots, X_{n_1} \sim \mathcal{N}(\mu_1, \sigma_1^2)$$
$$Y_1, \ldots, Y_{n_2} \sim \mathcal{N}(\mu_2, \sigma_2^2)$$

Let's consider two cases

I Equal population variances: $\sigma_1^2 = \sigma_2^2$

II Unequal population variances: $\sigma_1^2 \neq \sigma_2^2$

Consider, in case I, testing

$$\begin{aligned} H_0 :& \mu_1 - \mu_2 = 0, \;\; \sigma^2 > 0 \qquad (\sigma^2 = \sigma_1^2 = \sigma_2^2) \\ H_A :& \mu_1 - \mu_2 \neq 0, \;\; \sigma^2 > 0 \end{aligned}$$

Intuition: We need an estimator for $\mu_1 : \bar{X}$ and $\mu_2 : \bar{Y}$ and σ^2 (pooled sample variance)

$$\boxed{s_p^2 = \frac{s_x^2(n_1-1) + s_y^2(n_2-1)}{n_1+n_2-2} = \frac{\sum_i (X_i - \bar{X})^2 + \sum_i (Y_i - \bar{Y})^2}{n_1+n_2-2}}$$

Claim 1. The GLRT for testing H_1 vs H_a in the equal variances case is equivalent to reject H_0 for large values of

$$\frac{(\bar{X} - \bar{Y} - 0)^2}{\left(\sqrt{s_p^2/(n_1+n_2)}\right)^2} = \left(\frac{\bar{X} - \bar{Y} - 0}{\sqrt{s_p^2(\frac{1}{n_1} + \frac{1}{n_2})}}\right)^2$$

Next, we will show that under H_0,

$$\frac{\bar{X} - \bar{Y} - 0}{\sqrt{s_p^2(\frac{1}{n_1} + \frac{1}{n_2})}} \sim t(n_1+n_2-2 \text{ df})$$

This will be complete as soon as we verify that :

I

$$\frac{s_p^2(n_1+n_2-2)}{\sigma^2} \sim \chi^2(n_1+n_2-2 \text{ df})$$

II $\bar{X} - \bar{Y}$ is independent of s_p^2

So if we want to test whether two normal populations with equal variances have equal means as well, we could use an F test and reject for large values.

Extent this idea to 3 or more populations:

$$\begin{aligned} \text{Population 1:} \quad & X_{11}, \dots, X_{1J} \\ \text{Population 2:} \quad & X_{21}, \dots, X_{2J} \\ \text{Population 3:} \quad & X_{31}, \dots, X_{3J} \\ X_{ij} \sim \mathcal{N}(\mu_i, \sigma^2) \quad & \text{all independent} \\ i\text{: which population,} \quad & j\text{: which element of sample} \end{aligned}$$

Want to test $H_0 = \mu_1 = \mu_2 = \mu_3$ vs H_a atleast two μ_i's differ

Punchline: We'll end up with an F-test.

Lecture 21 (2018-11-26)

- Two sample testing
- ANOVA

Two-Sample Testing

$$X_1, X_2, \ldots, X_n \sim \mathcal{N}(\mu_1, \sigma_1^2)$$
$$Y_1, Y_2, \ldots, Y_m \sim \mathcal{N}(\mu_2, \sigma_2^2)$$
$$\text{Suppose } \sigma_1^2 = \sigma_1^2 = \sigma^2$$

But σ^2 unknown (common variance assumption). Consider

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_1 : \mu_1 - \mu_2 \neq 0$$

Can compute GLR (HW): hint

$$f(x_1, \ldots, x_n, y_1, \ldots, y_m | \mu_1, \mu_2, \sigma^2) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^{n+m} \exp\left(\frac{-\sum_{i=1}^{n}}{2\sigma^2}(x_i - \mu_1)^2\right) \cdot \left(-\frac{1}{2\sigma^2}\sum_{j=1}^{m}(y_j - \mu_2)^2\right)$$

In the homework, you will show that GLRT involves rejecting H_0 if

$$\left|\frac{\bar{X} - \bar{Y} - 0}{\sqrt{s_p^2\left(\frac{1}{n} + \frac{1}{m}\right)}}\right| \quad \text{is sufficiently large}$$

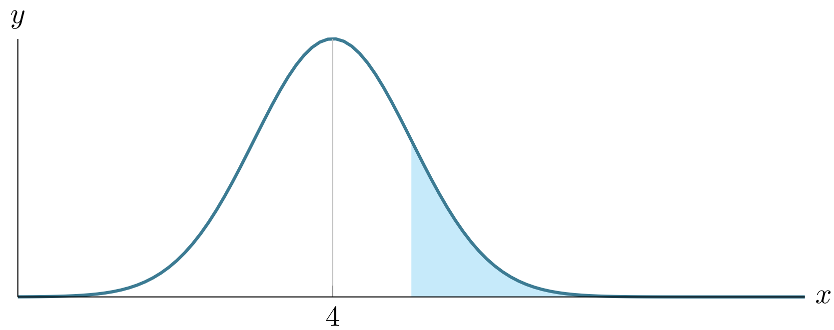
$$Var(\bar{X}) = \frac{\sigma^2}{n} \quad Var(\bar{Y}) = \frac{\sigma^2}{m} \qquad Var(\bar{X} - \bar{Y}) = \frac{\sigma^2}{n} + \frac{\sigma^2}{m} = \sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)$$

$$s_p^2 = \frac{S_X^2(n-1) + S_y^2(m-1)}{n+m-2} = \underbrace{\frac{S_X^2(n-1)}{(n-1+m-1)} + \frac{S_y^2(m-1)}{(n-1+m-1)}}_{\text{weighted average of sample variances}}$$

Under H_0,

$$\frac{\bar{X} - \bar{Y} - 0}{\sqrt{s_p^2\left(\frac{1}{n} + \frac{1}{m}\right)}} \sim t(n+m-2 \text{ df}) \qquad (\star)$$

Rejecting H_0 for large $ABSOLUTE$ values of $\star$ is equivalent to rejecting H_0 for large values of $(\star)^2$ i.e. rejecting H_0 when an F-distributed r.v. with 1 num df and $n+m-2$ denominator d.f. is large.



$\boxed{F_\alpha}$ Rejecting for large values of $\star^2 \Rightarrow$ upper tailed rejection region

Summary: To test equality of means for two independently sampled normal population with equal variances, we use an upper tailed F test, in essence, compares the variability in sample means $(\bar{X} - \bar{Y})$ against the "inherent" variability in the two populations (i.e. using s_p^2 as an estimate for σ^2)

Now consider comparing multiple population means

We consider Y_{ij} $= j^{th}$ data point/measurement from i^{th} sample (i^{th} treatment)

Suppose we have I populations/treatments, J measurements per treatment = common sample size

Assume:

$$Y_{ij} \sim \underbrace{\mu_i}_{\substack{\text{fixed}\\\text{deterministic}}} + \underbrace{\epsilon_{ij}}_{\substack{\text{random}\\\text{error}}}$$

$$\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \quad \underline{\text{all independent}} \quad \text{common variance: } \underline{\text{homoscedasticity}}$$

So Y_{ij} are all independent

$$Y_{ij} \sim \mathcal{N}(\mu_i, \sigma^2) \quad \text{for } 1 \le j \le J$$

$$\text{Consider } H_0 : \underbrace{\mu_1 = \mu_2 = \cdots = \mu_I}_{\text{all means equal}}$$

To make things a bit simpler, let's consider this parametrization:

$$Y_{ij} \sim \mu + \alpha_i + \epsilon_{ij}$$

$$\text{where: } \sum_{i=1}^{I} \alpha_i = 0 \quad \& \quad \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$$

Can always do this by setting $\boxed{\mu = \hat{\mu} = \frac{\sum \mu_i}{I}}$ $\quad \alpha_i = (\mu_i - \mu)$

$$\sum_{i=1}^{I} (\mu_i - \mu) = \sum \mu_i - I\mu = \sum \mu_i - I\bar{\mu} = 0$$

We will break up the variability in data into:

(1) A measure of variability across sample treatment means $\rightarrow SS_B$ ("sum of squares between")

(2) A measure of inherent variability (variability within each treatment) $\rightarrow SS_W$ ("sum of squares within")

$$\text{Total sum of squares} = \sum_i \sum_j (Y_{ij} - \bar{Y}_{\cdot\cdot})^2$$

$$\text{Key Identity} \quad \sum_i \sum_j (Y_{ij} - \bar{Y}_{\cdot\cdot})^2 = \sum_i \sum_j (Y_{ij} - \bar{Y}_{i\cdot} + \bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$$

Lecture 22 (2018-11-28)

- Sums of Squares decomposition
- ANOVA tables

One-way ANOVA

$$\textbf{Model:}\quad Y_{ij} \sim \mathcal{N}(\overbrace{\mu + \alpha_i}^{\mu_i}, \sigma^2)$$

$$\sum_{i=1}^{I} \alpha_i = 0, \quad Y_{ij}\text{'s are independent}$$

Wish to test H_0: μ_i all equal $\Leftrightarrow H_0 : \alpha_i \equiv 0 \ \forall i$

$$\begin{aligned} TSS = SSTot = SST &= \sum_i \sum_j (Y_{ij} - \bar{Y}_{\cdot\cdot})^2 \\ &= \sum_i \sum_j (Y_{ij} - \bar{Y}_{i\cdot} + \bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 \\ &= \underbrace{\sum_i \sum_j (Y_{ij} - \bar{Y}_{i\cdot})^2}_{SS_W} + \underbrace{\sum_i \sum_j (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2}_{SS_B} + 2 \sum_i \sum_j \cancelto{0}{(Y_{ij} - \bar{Y}_{i\cdot})}(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}) \end{aligned}$$

$$\bar{Y}_{i\cdot} = \frac{\sum_{j=1}^{J} Y_{ij}}{J} \qquad \text{Note: } J\bar{Y}_{i\cdot} = \sum_{j=1}^{J} Y_{ij}$$

$$\text{Hence} \quad \sum_{i=1}^{I} Y_{i\cdot} J = \sum_i \sum_j Y_{ij}$$

$$\frac{\sum_{i=1}^{I} Y_{i\cdot} \cancel{J}}{I\cancel{J}} = \frac{\sum_i \sum_j Y_{ij}}{IJ} = \bar{Y}_{\cdot\cdot}$$

$$\text{That is, } \bar{Y}_{\cdot\cdot} = \frac{\sum_{i=1}^{I} \bar{Y}_{i\cdot}}{I}$$

Note that if $\alpha_i \equiv 0$, i.e. if H_0 is true,

$$\bar{Y}_{i\cdot} \sim \mathcal{N}(\mu, \sigma^2/J) \text{ and } \bar{Y}_{i\cdot} \text{ r.vs are independent}$$

So let $W_i = \bar{Y}_{i\cdot}$ and

$$\bar{W} = \frac{1}{I} \sum_{i=1}^{I} \bar{Y}_{i\cdot} = \frac{1}{I} \sum_{i=1}^{I} W_i = \bar{Y}_{\cdot\cdot}$$

We are i.i.d under H_0, $W_i \sim \mathcal{N}(\mu, \sigma^2/J) \quad (I-1)S_W^2$

$$\text{So} \quad \frac{\sum_{i=1}^{I} (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2}{I-1} = \frac{1}{I-1} \sum_{i=1}^{I} (W_i - \bar{W})^2$$

$$\text{So we conclude} \quad \boxed{\frac{J\sum(W_i - \bar{W})^2}{\sigma^2} \sim \chi^2(I-1)}$$

$$\text{Note that } J\sum(W_i - \bar{W})^2 = \sum_i \sum_j (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$$

$$\text{Now, } \frac{SSB}{\sigma^2} \sim \chi^2(I-1) \quad \text{under } H_0$$

Next, what about

$$\sum_{i=1}^{I} \underbrace{\sum_{j=1}^{J} (Y_{ij} - \bar{Y}_{i\cdot})^2}_{\text{sample var } S_i^2(J-1)}$$

$$\text{Note:} \quad \frac{(J-1)S_i^2}{\sigma^2} \sim \chi^2(J-1)$$

So Since S_i^2's are independent,

$$\sum_{i=1}^{I} \frac{(J-1)S_i^2}{\sigma^2} \sim \chi^2(I(J-1))$$

Hence,

$$\frac{SS_W}{\sigma^2} \sim \chi^2(I(J-1))$$

Exercise (HW). Show that SS_W and SS_B are independent. Assuming independence, under H_0,

$$\frac{SS_B/(I-1)}{SS_W/I(J-1)} \sim F(I-1, I(J-1))$$

SS_B is sometimes called the $SSTr$, it encapsulates differences across sample treatment means $\bar{Y}_{i\cdot}$.

SS_W is a proxy for inherent variability error in model (i.e. σ^2). And SS_W is sometimes called "sums of squares for error" or SSE

ANOVA Table

$$\frac{TSS}{\sigma^2} = \frac{\sum\sum(Y_{ij} - \bar{Y}_{\cdot\cdot})^2}{\sigma^2} \sim \chi^2(IJ-1) \quad \textbf{under } h_0$$

Benefits of ANOVA F-test

(a) "Omnibus" test: can test equality of multiple means $\mu_1 = \mu_2 = \mu_3 \cdots = \mu_I$ vs atleast one pair differs at fixed level α

(b) Less burdensome than pairwise t-test

(c) Type I error is controlled

Drawback: F-test doesn't allow us to immediately identify which means are different (if we reject H_0, for example)

When comparing two populations, the t-test for equality of means, is equivalent to the F-test, because $t^2 = F$

Tuky's Test

Commonly used to identify which means might be different

Basic idea:

1. Order the sample treatment $\bar{Y}_{i\cdot}$
2. Calculate a threshold value
3. Check if $|\bar{Y}_{i_1\cdot} - \bar{Y}_{i_2\cdot}| > \omega$ or not
4. If yes, reject $H_0(i_1, i_2) : \mu_{i_1} = \mu_{i_2}$

The threshold value ω depends on the so called studentized range distribution.

This is distribution of the random variable

$$Q = \max_{i_1, i_2} \frac{|(\bar{Y}_{i_1\cdot} - \mu_{i_1}) - (\bar{Y}_{i_2\cdot} - \mu_{i_2})|}{S_p/\sqrt{J}}$$

$$S_p^2 = \text{ pooled sample variance} = MSE = \frac{SS_W}{I(J-1)}$$

Note that Q depends on two parameters:

1. how many treatments? (I)
2. how many dfs in MSE $(I(J-1))$ Need equal treatment sizes.

At level α, you can locate (in table in Rice)

$$q_\alpha(I, I(J-1)) \quad I \text{ is treatments and } I(J-1) \text{ is df}$$

So,

$$\omega = q_\alpha(I, I(J-1)) \frac{S_p}{\sqrt{J}}$$

and we reject H_0: $\mu_{i_1} = \mu_{i_2}$ if $|\bar{Y}_{i_1\cdot} - \bar{Y}_{i_2\cdot}| > \omega$

Lecture 23 (2018-12-03)

- One way ANOVA

$$Y_{ij} \sim \mu + \alpha_i + \epsilon_{ij}$$
$$\sum_{i=1}^{I} \alpha_i = 0, \quad \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$$

Two-way layout

Example 1. Drug type (A) and patient type (B)

$$Y_{ijk} \quad i: \text{ Factor A, level i}, \quad j: \text{ Factor B at level j}$$

"levels" of each factor correspond to a breakdown of possible populations of that factor

Example for Drug type: Chloroquine (1), Metloquine (2), antemisin (3)

Age: 0-2 (1), 2-6 (2), 6-16 (3), >6 (4)

Model:

$$Y_{ijk} = \underbrace{\mu + \alpha_i + \beta_j + \overbrace{\delta_{ij}}^{\text{interaction}}}_{\text{parameters}} + \underbrace{\epsilon_{ijk}}_{\text{random error}}$$
$$\sum_{i=1}^{I} \alpha_i = 0 \quad \sum_{j=1}^{J} \beta_i = 0 \quad \sum_{i=1}^{I} \delta_{ij} = \sum_{j=1}^{J} \delta_{ij} = 0$$

Three Hypotheses to test

- H_{0A} (main effects for Factor A): $\alpha_i \equiv 0 \quad \forall i$
- H_{0B} (main effects for Factor B): $\beta_j \equiv 0 \quad \forall j$
- H_{0AB} (main effects for Factor A and B): $\delta_{ij} \equiv 0 \quad \forall i, j$

- H_{1A} (main effects for Factor A): $\alpha_i \neq 0 \quad \forall i$
- H_{1B} (main effects for Factor B): $\beta_j \neq 0 \quad \forall j$
- H_{1AB} (main effects for Factor A and B): $\delta_{ij} \neq 0 \quad \forall i, j$

Sums of Squares decomposition

$$\sum\sum\sum (Y_{ijk} - \bar{Y}_{...})^2 = SS_{TOT}$$

Sums of Squares

$$SS_A = JK \sum_{i=1}^{I} (\bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot\cdot\cdot})^2$$

$$SS_B = IK \sum_{j=1}^{J} (\bar{Y}_{\cdot j\cdot} - \bar{Y}_{\cdot\cdot\cdot})^2$$

$$SS_{AB} = K \sum \sum (\bar{Y}_{ij\cdot} - \bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot j\cdot} + \bar{Y}_{\cdot\cdot\cdot})^2$$

$$SSE = \sum \sum \sum (Y_{ijk} - \bar{Y}_{ij\cdot})^2$$

We expect, under suitable standardization, that SS_A, SS_B, SS_{AB} and SSE should all be related to chi-squared distance.

Two Way ANOVA Table

Source	**SS**	**df**	Mean Squared	F
Factor A	SS_A	$I-1$	$MSA = \frac{SSA}{I-1}$	$\frac{MSA}{MSE} \sim F(I-1, IJ(K-1))$
Factor B	SS_B	$J-1$	$MSB = \frac{SSB}{J-1}$	$\frac{MSB}{MSE} \sim F(J-1, IJ(K-1))$
Interaction	SS_{AB}	$(I-1)(J-1)$	$MSAB = \frac{SS_{AB}}{(I-1)(J-1)}$	$\frac{MS_{AB}}{MSE} \sim F((I-1)(J-1), IJ(K-1))$
Error	SS_E	$IJ(K-1)$	$MSE = \frac{SSE}{IJ(K-1)}$	
Total	SS_{TOT}	$IJK-1$		

Lecture 24 (2018-12-05)

Last Lecture :(

- Simple Linear Regression

Simple Linear Regression

Model: Response variable y depends on predictor variable x, but also includes Random error.

$$y = \beta_0 + \beta_1 x + \epsilon$$
$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$

$$\textbf{Data:} \quad \{(x_i, y_i) : 1 \le i \le n\} y_i = \beta_0 + \beta_1 x + \epsilon_i \qquad \boxed{\epsilon_i \text{ are i.i.d. } \mathcal{N}(0, \sigma^2)}$$

Example 1. x_i = hours of study for student i
y_i = exam score
We need to estimate the parameters.

$$\beta_0, \beta_1 \text{ - (intercept, slope)}$$
$$\sigma^2 \text{ - (variance)}$$

Likelihood:

$$\left(\frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ \frac{-1}{2\sigma^2} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 \right\} \right)^2$$

Note that $\epsilon_i = y_i - (\beta_0 + \beta_1 x_i)$

<u>Exercise</u>: Calculate $\hat{\beta}_{0_{MLE}}, \hat{\beta}_{1_{MLE}}, \hat{\sigma^2}_{MLE}$

We typically think of x_i's as fixed given the y_i's as random can also consider (x_i, y_i) pairs arising from a bivariate normal distribution.

Now if we consider (x_i, y_i) as arising from a bivariate normal distribution, with correlation p, then conditional on x, <u>$y(x)$ has a normal distribution</u> with a mean that depends linearly on x (HW1, bivariate normal)

<u>Remark</u>: ANOVA is regression with categorical predictors. X_i determines which pop. To estimate $\beta_0, \beta_1, \sigma^2$

Sum of squared differences

$$\sum_{i=1}^{n} (y_i - (a + bx_i))^2 = \mathcal{L}(a, b)$$

Minimize $\mathcal{L}(a, b)$ over a, b. This leads to

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

$$\hat{b} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

Estimated slope

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \\
\hat{\beta}_1 &= \hat{y} - \hat{\beta}_1 \bar{x}
\end{aligned}$$

Estimated regression line: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

Actual mean of y: $y = \mathbb{E}[y] = \beta_0 + \beta_1 x$

For estimating σ^2 – note that if β_0, β_1 are known, σ^2 still remains to be estimated, but the MLE in this case can be expressed as a function of β_0, β_1

Total variability in data

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2 + 2 \sum (y_i - \bar{y})(y_i - \hat{y}_i)$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

$$\frac{\sum \hat{y}_i}{n} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 \bar{x} = \bar{y} \Rightarrow \boxed{\bar{\hat{y}} = \bar{y}}$$

$$\begin{aligned}
\hat{y}_i - \bar{y} &= \hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y} \\
&= \hat{\beta}_0 \left[\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \right] \bar{y} \\
&= \cancel{\bar{y}} = \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i - \cancel{\bar{y}} \\
&= \hat{\beta}_1 (x_i - \bar{x})
\end{aligned}$$

$$y_i - \bar{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

Now, if we wanted to estimate σ^2, and we know β_0, β_1, we might consider

$$\frac{\sum_{i=1}^{n} \overbrace{(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))}^{(\epsilon^2)}{}^2}{n}$$

So, since β_0, β_1 are unknown, can instead consider

$$\frac{\sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2}{n-2} = \hat{\sigma}^2$$

We have an estimate for β_1 and $b\hat{e}ta_1$. How is $b\hat{e}ta_1$ distributed?

$$\hat{\beta}_1 = \frac{\sum \overbrace{(x_i - \bar{x})}^{a_i}(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{a_i y_i - \boxed{\sum a_i \bar{y}}}{\sum (x_i - \bar{x})^2} = \frac{\sum a_i y_i}{\sum a_i^2}$$

$$\begin{aligned}\sum a_i \bar{y} &= \sum (x_i - \bar{x})\bar{y} \\ &= \bar{y}(\cancelto{0}{\sum x_i - n\bar{x}})\end{aligned}$$

So $\bar{\beta}_1$ is linear combination of normal random variables (y_i's)

$$Var(\hat{\beta}_1) = \frac{\sum (a_i)^2 \sigma^2}{(\sigma a_i^2)^2} = \frac{\sigma^2}{(\sum a_i^2)}$$

So that we can show (exercise) that $\mathbb{E}[\hat{\beta}_1] = \beta_1$, then we can use

$$\frac{\hat{\beta}_1 - 0}{\sqrt{\sigma^2 / \sum a_i^2}} \qquad \text{as a test statistic}$$

Instead we use $H_0 : \beta_1 = 0$ vs $H_a : \beta_1 \neq 0$

$$\frac{\hat{\beta}_1 - 0}{\sqrt{MSE / \sum (x_i - \bar{x})^2}} \sim t_{n-2}$$

Summary of Linear Regression

$$\textbf{Model} \qquad tty_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

$$\textbf{Estimates} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}; \quad \hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}; \quad \hat{\sigma^2} = s^2 = \frac{\sum (y_i - \hat{y}_i)^2}{n-2} = MSE$$

Is model useful? $\longrightarrow H_0 : \beta_1 \neq 0$ vs $\beta_1 \neq 0$