Econ 506: Advanced Economic Statistics

Yuan Liao
Department of Economics
Rutgers University

May 3, 2020

Contents

1 Review of Probability and Distribution Theory

1.1 Set theory

- Set theory was founded by Georg Cantor (1874). He gave the following definition of a set:

  > By an "aggregate", we are to understand any collection into a Whole M of definite and separate objects of our intuition or of our thought. These objects are called the elements of M.

- Subset: $A \subset B$ iff: $\forall a \in A$, then $a \in B$.

  $A = B$ iff: $A \subset B$ and $B \subset A$.

- The Null Set Axiom: There exists a set, denoted by ; and called the empty set, which has no elements.

- Union: $A \cup B = \{a : a \in A \text{or } a \in B\}$

  Countable union: $\cup_{i=1}^{\infty} = \lim_n \cup_{i=1}^{n} A_i$

- Intersection: $A \cap B = \{a : a \in A \text{ and } a \in B\}$

  Countable intersection: $\cap_{i=1}^{\infty} = \lim_n \cap_{i=1}^{n} A_i$

- It is easy to prove the following laws using these definitions:

$$A \cup B = B \cup A (\text{same with intersection})$$
$$(A \cup B) \cup C = A \cup (B \cup C) (\text{same with intersection})$$
$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$
$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$

To illustrate the basic ideas of proving, we prove the last result.

Proof. $\forall a \in A \cap (B \cup C)$, then $a \in A$ and $a \in B \cup C$, so $a \in B$ or $a \in C$. If $a \in B$ then $a \in A \cap B$; if $a \in C$ then $a \in A \cap C$. Either case, $a \in (A \cap B) \cup (A \cap C)$. Thus left$\subset$ right.

Now $\forall a \in$ right, then either $a \in A \cap B$ or $a \in A \cap C$. In the first case, $a \in B$; in the second case, $a \in C$. Hence $a \in B \cup C$. In either case, $a \in A$. Hence $a \in$ left. Hence right $\subset$ left. □

- Difference: $A \backslash B = \{a \in A \text{ and} a \notin B\} = A \cap B^c$. symmetric difference $A \Delta B$

  Hausdorff distance:

$$\max\{\sup_{a \in A} d(a, B), \sup_{b \in B}, d(b, A)\}$$

- Complement: $A^c = \{a : a \notin A\}$.
- De Morgan's Law: $(A \cup B)^c = A^c \cap B^c$; $(A \cap B)^c = A^c \cup B^c$

1.2 Probability space

1.2.1 Sigma algebra

- Let Ω be a set (intuitively, it collects all the possible outcomes). A $\sigma-$ algebra defined on Ω, denoted by $\mathcal{F}$, is a class of subsets of Ω, such that:

  1. The elements of $\mathcal{F}$ are subsets of Ω.
  2. $\emptyset \in \mathcal{F}$
  3. $\Omega \in \mathcal{F}$
  4. It is closed under complementation.
  5. It is closed under countable intersections and unions.

  Intuitively, if we want to consider random events, then we want to also consider all possible unions, intersections, or all kinds of operations of these random events. This motivates sigma algebra.

- We usually denote the sigma algebra by $\mathcal{F}$, and the space as $(\Omega, \mathcal{F})$. This pair is called measurable space.
- Example: $\Omega = \{1...6\}$. Then $\mathcal{F} = \{\emptyset, \Omega\}$ is a sigma algebra. If in addition, we want to add $\{1\}$. Then what is the smallest sigma algebra in this case containing $\{1\}$?

$$\mathcal{F} = \{\emptyset, \Omega, \{1\}, \{2-6\}\}$$

- Let $\Omega \subset \mathbf{R}^d$. Borel sigma on Ω is a collection of all sets that can be formed from open sets through countable union, countable intersection, and complements. It can be proved to be the smallest σ-algebra containing all open subsets of Ω.

1.2.2 Probability measure

- A function $P : \mathcal{F} \to [0, 1]$ is called a probability measure if:

  1. $\forall A \in \mathcal{F}$, $P(A) \geq 0$
  2. $P(\emptyset) = 0$
  3. $P(\Omega) = 1$
  4. For all countable pairwise disjoint collections $\{A_1, A_2....\}$, each $A_i \in \mathcal{F}$, we have
  $$P(\cup_i A_i) = \sum_i P(A_i)$$

- It has the following properties:

  1. $P(A) = 1 - P(A^c)$
  2. $$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
  $$P(A \cup B \cup C) = P(A)+P(B)+P(C)-P(A \cap B)-P(A \cap C)-P(B \cap C)+P(A \cap B \cap C)$$
  $$P(\cup A_i) = \sum_i P(A_i) - \sum_{ij} P(A_i \cap A_j) + \sum_{ijk} P(A_i \cap A_j \cap A_k) - ... + ...$$
  3. If $\{A_n\}$ or $\{B_n\}$ is monotone, meaning that $A_n \subset A_{n+1}$, or $B_{n+1} \subset B_n$, then
  $$\lim_n P(A_n) = P(\lim_n A_n), \lim_n P(B_n) = P(\lim_n B_n)$$
  We just need to figure out $\lim_n A_n$ or $\lim_n B_n$. In fact, A_n becomes larger, so $\lim_n A_n = \cup A_n$; B_n becomes smaller, so $\lim_n B_n = \cap B_n$

1.3 random variable

1.3.1 definition

We now look at probabilities of sets that are generated due to some random events. These random events are generated by some random thing, called "random variable".

- Consider $(\Omega, \mathcal{F})$, this forms all the possible outcomes of some random event. Also consider $(A, \mathcal{B})$, where $A \in \mathbf{R}$, and $\mathcal{B}$ is its borel sigma algebra. The latter quantifies the random event.

- A random variable is a function $X : (\Omega, \mathcal{F}) \to (A, \mathcal{B})$, so that

$$\forall B \in \mathcal{B}, \quad \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}$$

  Intuitively, this means for any quantified value, there must be a possibility to achieve it.

- Example: When tossing a coin, we obtain H and T. $\Omega = \{H, T\}$. We can quantify it to be $\{0, 1\}$. Then this quantification is called a random variable: $X = 0$ or 1. Here we can define

$$X(\omega) = \begin{cases} 1 & \omega = H \\ 0 & \omega = T \end{cases}$$

  For example, we can take $\mathcal{F} = \{\emptyset, \{H\}, \{H, T\}, \{T\}\}$, and $\mathcal{B} = \{\emptyset, \{0, 1\}, \{0\}, \{1\}\}$ if we take $B = \{0, 1\} \in \mathcal{B}$, then what is

$$\{\omega \in \Omega : X(\omega) \in \{0, 1\}\}?$$

  in fact,

$$\{\omega \in \Omega : X(\omega) \in \{0, 1\}\} = \{\omega \in \Omega : X(\omega) = 0 \text{ or } 1\} = \{H, T\} \in \mathcal{F}.$$

- Homework Question: there are three colored balls: B, R, W. We define $\Omega = \{B, R, W\}$. We randomly pick up one, and define a function:

$$X(\omega) = \begin{cases} 1 & \omega = B \\ 2 & \omega = R \\ 3 & \omega = W \end{cases}$$

  and define $\{\omega \in \Omega : X(\omega) \in \emptyset\} = \emptyset$.

  We also define $A = \{1, 2, 3\}$, and $\mathcal{B} =$ borel σ algebra generated by $\{1, 2, 3\}$. Then

  1. call $\mathcal{F}_2$ as the full sigma algebra that contains all subsets,

is $X : (\Omega, \mathcal{F}_2) \to (A, \mathcal{B})$ a r.v ? prove or disprove

2. what is the smallest sigma algebra containing $\{R\}$,call it $\mathcal{F}_1$?
3. is $X : (\Omega, \mathcal{F}_1) \to (A, \mathcal{B})$ a r.v ? prove or disprove

Proof. 1.

$$\{\omega \in \Omega : X(\omega) \in \emptyset\} = \emptyset \in \mathcal{F}$$

$$\{\omega \in \Omega : X(\omega) \in \{1,2,3\}\} = \{B, R, W\} \in \mathcal{F}$$

$$\{\omega \in \Omega : X(\omega) \in \{2\}\} = \{R\} \in \mathcal{F}$$

$$\{\omega \in \Omega : X(\omega) \in \{1,3\}\} = \{BW\} \in \mathcal{F}$$

and also all other sets...... after checking all them, so yes it is sigma r.v.

2. $\mathcal{F}_1 = \{\emptyset, \{B, R, W\}, \{R\}, \{B, W\}\}$
3.

$$\{\omega \in \Omega : X(\omega) \in \{3\}\} = \{W\} \notin \mathcal{F}_1$$

so it is not.

This example says, the domain of the r.v. has to be big enough. □

- Sums and products of r.v's are still r.v.

1.3.2 CDF

Let X be a random variable.

- The cumulative distribution function of X is a function $F : \mathbb{R} \to [0, 1]$ defined as

$$F(x) = P(X \leq x)$$

- It satisfies:

1. nondecreasing: say P(no more than 3) $\leq$ P(no more than 4) . It is possible they are the same, if $X(\omega) \in \{1, 2, 5\}$.
2. $F(-\infty) = 0$, $F(+\infty) = 1$
3. right continuity: $F(x_0^+) = F(x_0)$. This means,

$$\lim_{x \geq x_0, x \to x_0} P(X \leq x) = P(X \leq x_0)$$

- But there can be a jump on the left side. Example: $X(\omega) \in \{1, 2, 5\}$, each with chance 1/3. Then $P(X \leq 2) = 2/3, \quad P(X \leq 2 + \epsilon) = P(X = 1 or 2) = 2/3$ But $P(X \leq 2 - \epsilon) = P(X = 1) = 1/3$.

- discrete r.v.: which is r.v. whose outcome possibility set is either finite or countable.

- If X is discrete, then $P(X = x)$ is called the probability mass function of X.

- A continuous r.v. is a r.v. such that $P(X = x) = 0$, $\forall x$.

- Note that a r.v. can be neither continuous nor discrete (mixed). e.g., takes value $\{1, 2, 3\} \cup [4, 5]$.

1.3.3 PDF

- The probability density function of a random variable X is a function f, so that
$$F(x) = \int_{-\infty}^{x} f(t)dt.$$

- If f is also continuous, then $F'(x) = f(x)$.
$$\int_{-\infty}^{+\infty} f(t)dt = \lim_{x \to \infty} \int_{-\infty}^{x} f(t)dt = \lim_{x \to \infty} F(x) = 1.$$

- density function usually exists for continuous functions.

- {x: f(x)>0} is called support function, usually used to characterize the possible values for continuous r.v.

1.3.4 Expectation and Variance

- It is difficult to formally define the expectation without a solid probability foundation. But getting into too much details is boring. We just give a rough definition

- If continuous,
$$EX = \int x f(x) dx$$

If discrete,

$$EX = \sum_x xP(X = x)$$

- $Var(X) = E(X - EX)^2$.

1.4 Joint distributions

1.4.1 Joint and marginal distributions

- For two r.v., the joint distribution function is

  $$F_{X,Y}(x, y) = P(X \leq x, Y \leq y).$$

  The marginal distribution is just $P(X \leq x)$ and $P(Y \leq y)$.

- Joint density: $f_{X,Y}$ satisfies

  $$F_{X,Y}(x, y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(t, s)dtds \quad \forall x, y$$

- We have

  $$f_X(x) = \int f_{X,Y}(x, y)dy$$

  Proof. Let $g(x) = \int f_{X,Y}(x, y)dy$. Simply show

  $$F(x) := \int_{\infty}^{x} g(t)dt, \forall x.$$

  The left is $F(x) = P(X \leq x) = \lim_{y\to\infty} P(X \leq x, Y \leq y) = \lim_{y\to\infty} \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(t, s)dsdt$.
  The right is $\int_{\infty}^{x} g(t)dt = \int_{\infty}^{x}\int f_{X,Y}(t, y)dydt = \int_{\infty}^{x}\int f_{X,Y}(t, s)dsdt = \int_{\infty}^{x} \lim_{y\to\infty}\int^{y} f_{X,Y}(t, s)dsdt$
  Compare the two sides, we have to show $\int^x$ and $\lim_y$ can be interchanged:

  $$\lim_y \int^x h(y, t)dt = \int^x \lim_y h(y, t)dt, \quad h(y, t) = \int_{-\infty}^{y} f_{X,Y}(t, s)ds$$

  This is given by the Dominated Convergence Theorem, because

  $$|h(y, t)| \leq g(t)$$

and $g(t)$ is integrable: $\int g(x)dx = \iint f_{X,Y}(x,y)dydx = F_{X,Y}(+\infty,+\infty) = 1$.

$\square$

- $\partial^2 F(x,y) = f(x,y)$.
- Joint expectation is $E\mathbf{X} = (EX_1, ..., EX_n)^T$.
- Covariance is $Cov(X,Y) = E(X-EX)(Y-EY)$.
- Covariance matrix is $Cov(\mathbf{X}) = (Cov(X_i,X_j)) = E\mathbf{X}\mathbf{X}^T - (E\mathbf{X})(E\mathbf{X})^T$. Why is covariance matrix useful? It can be shown that

  $$Var(a^T\mathbf{X}) = E(a^T\mathbf{X})^2 - (Ea^T\mathbf{X})^2 = a^T E\mathbf{X}\mathbf{X}^T a - a^T E\mathbf{X}E\mathbf{X}^T a = a^T Cov(\mathbf{X})a.$$

1.4.2 Application: Markowitz's modern portfolio theory

- Say on t, I have \$1, I spend it to N stocks, spending each $w_1, ..., w_N$.
- I got amount w_1/p_{1t},..., w_N/p_{Nt}.
- On $t+1$, I sell them, I got income $\sum_i w_i \frac{p_{i,t+1}}{p_{it}}$. My cost was $1 = \sum_i$, hence return is

  $$r_t = \sum_i w_i(\frac{p_{i,t+1}}{p_{it}} - 1) = \sum_i w_i \frac{p_{i,t+1} - p_{it}}{p_{it}}$$

- $R_{it} = \frac{p_{i,t+1} - p_{it}}{p_{it}}$ is called "return" for each stock, and $r_t = \sum_i w_i R_{it} = \mathbf{w}'\mathbf{R}_t$ is called the return of the portfolio $(w_1...w_N)$.
- Question: how to choose $\mathbf{w}$? The expected return is $\mathbf{w}'E\mathbf{R}_t$. The "risk" is defined as $Var(r_t) = Var(\mathbf{w}'\mathbf{R}_t) = \mathbf{w}_t' Cov(\mathbf{R}_t)\mathbf{w}_t$.
- A portfolio that gives maximum return for a given risk, or minimum risk for given return is an **efficient portfolio**. From the portfolios that have the same return, the investor will prefer the portfolio with lower risk
- Markowitz 1956 says:

  $$\min \mathbf{w}'Cov(\mathbf{R}_t)\mathbf{w}, s.t. \mathbf{w}'E\mathbf{R}_t = \mu, \quad \mathbf{w}'\mathbf{1} = 1.$$

  The solution $\mathbf{w}^* \propto Cov(\mathbf{R}_t)^{-1}\mathbf{b}(\mu)$.

- So the optimal risk $\mathbf{w}^{*'}Cov\mathbf{w}^*$ is a function of μ. This yields a Efficient Frontier, Risk v.s. Return. Every portfolio on it is efficient portfolio. Every portfolio below the curve is inefficient because under the same risk, the inefficient portfolio yields smaller return.

  For instance, when $N = 3$; I generated 50 possible μ from [0.03, 0.08]; I plot the generated μ with the corresponding optimal $\sqrt{\mathbf{w}^{*'}Cov\mathbf{w}^*}$.

- To estimate $E\mathbf{R}_t, Cov(\mathbf{R}_t)$, use sample average and sample variance. In matlab, suppose we have three sequences, then write X be 3 by T, the mean vector and covariance matrix are:

  mean(X,2);

  cov(X');

  Using my code, after getting bmu, bSigma, simply run

  portfolio(bmu,bSigma);

- As an example, I used weekly data from Sept 16, 2015-Sept 14, 2016 of Samsung, Apple and MS. The mean is [0.0072, 0.0074, 0.0007], the covariance is [13,-1,5;-1,14,2;5,2,15]*1e-4

  run portfolio(bmu,bSigma);

1.4.3 Conditional distributions

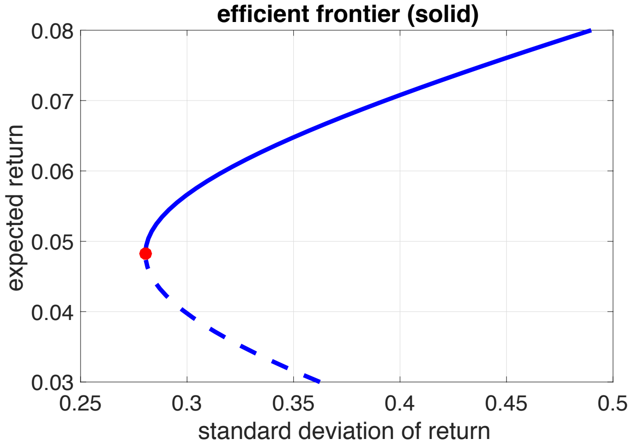
- The conditional probability is simply

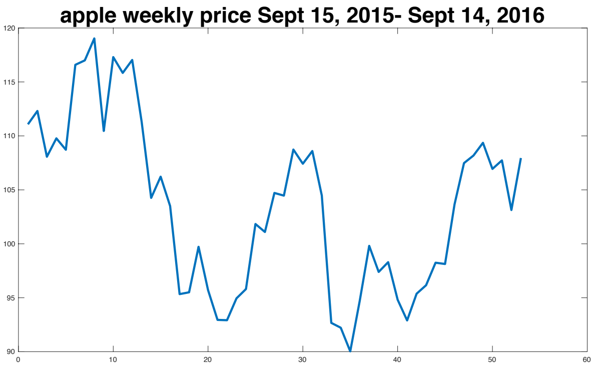
$$P(A|B) = \frac{P(AB)}{P(B)}$$

- If A, B are independent, then $P(A|B) = P(A)$.

  Example 1.1. Sensitive question survey: estimate the proportion of students who ever cheated in exams. ask "have you cheated? "

  Each respondent then flips the coin without showing it to the interviewer. If the coin lands heads, then the respondent answers "yes" to the question, regardless cheated or not. If the coin lands tails, then the respondent answers truthfully to the question.


efficient frontier (solid)
0.08
0.07
0.06
0.05
0.04
0.03
expected return
0.25
0.3
0.35
0.4
0.45
0.5
standard deviation of return


apple weekly price Sept 15, 2015- Sept 14, 2016
120
115
110
105
100
95
90
0
10
20
30
40
50
60

Define events A=(cheated),

$$P(yes) = P(yes, H) + P(yes, T) = \underbrace{P(yes|H)}_{=1} P(H) + P(yes|T)P(T)$$

$$= 0.5 + P(A)0.5$$

Hence

$$P(A) = 2P(yes) - 1$$

The key is

$$P(yes|T) \underbrace{=}_{telltruth} P(A|T) \underbrace{=}_{independence} P(A).$$

Getting T is independent of cheated. If we divide students into two groups, getting H or getting T. The rate of cheated should be the same (there is no point that those getting T have higher chance of cheating).

- Other examples: Monty Hall game.

  google "math, ucsd, month hall"

- Conditional density is

$$f_{X|Y}(x|y) = \frac{f_{XY}(x, y)}{f_Y(y)}$$

1.4.4 Conditional mean

- Conditional mean. It is difficult to define the conditional mean rigorously. Heuristically,

$$E(X|Y) = \int x f_{X|Y}(x|y)dx$$

- $Var(X|Y) = E[(X - E(X|Y))^2|Y]$

- Important laws

$$E(f(Y)X|Y) = f(Y)E(X|Y) : \text{treat } f(Y) \text{ constant}$$

$$E(E(X|Y)) = EX$$

$$Var(X) = EVar(X|Y) + VarE(X|Y)$$

To prove $E(E(X|Y)) = EX$, note that

$$E(X|Y = y)f(y) = \int xf(x, y)dx \Rightarrow \int E(X|Y = y)f(y)dy = \int xf(x)dx$$

Also, if A is a matrix, X is a vector

$$Var(AX) = AVar(X)A^T.$$

- Important result in econometrics: $\forall g$,

  $$E(Y - g(X))^2 \geq E(Y - g^*(X))^2$$

  where $g^*(X) = E(Y|X)$.

  Proof.

  $$E(Y-g(X))^2 = E(Y-g^*+g^*-g(X))^2 = E(Y-g^*)^2+E(g^*-g)^2+2E(Y-g^*)(g^*-g)$$

  To prove the third term is zero, note

  $$E(Y - g^*)(g^* - g) = EE((Y - g^*)(g^* - g)|X) = E(g^* - g)E(Y - g^*|X) = 0.$$

  □

- This means among all functions of X, $E(Y|X)$ is closest to Y. Hence $E(Y|X)$ is often called the projection of Y on X.

  Define

  $$e := Y - E(Y|X), \Rightarrow E(e|X) = 0$$

  meaning that the projection of e on X is zero; meaning that e is orthogonal to X. Hence in the following decomposition:

  $$Y = g(X) + e$$

  if $g(X) = E(Y|X)$, then $E(e|X) = 0$.

- Reversely, if we have the decomposition $Y = g(X)+e$ and know that $E(e|X) = 0$, then $g(X) = E(Y|X)$, by taking conditional mean on both sides.

- Hence we proved: in the following decomposition:

$$Y = g(X) + e$$

$g(X) = E(Y|X)$ iff $E(e|X) = 0$, known as "mean independence".

- In econometric applications,
  - $g(X)$ is often known as the effect of X on Y, which is the central object to study in empirical microeconomics.
  - We have shown that if $E(e|X) = 0$, then such an effect can be "identified" as $E(Y|X)$, meaning that it is uniquely determined if the joint distribution f_{XY} is known.
  - Such a condition, $E(e|X) = 0$, is known as "exogeneity", meaning that e is not determined by X.
  - In this case, the effect is also the closest to Y among all functions of X.
  - What about $Var(e|X)$? It is not zero, and can be calculated: because $Ee|X = 0$, we know $g(X) = E(Y|X)$, thus $e = Y - E(Y|X)$, thus

    $$Var(e|X) = (Ee^2|X) - 0 = E((Y - E(Y|X))^2|X) = Var(Y|X)$$

    Often it depends on X. This is often known as "heteroskedasticity", meaning that the conditional variance of e depends on X.
  - If we divide the population into subpopulations according to X, this means the variance is different in different populations, which is quite reasonable.
  - Reversely, called "homoskedasticity", simple but not reasonable.

1.4.5 independence

- Intuitive definition:

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$$

P(tomorrow rains, I win the lottery)= P(tomorrow rains) P(I will the lottery)

- Formal definition is more complicated.

1. Independence of events: events are mutually independent, if
$$P(\cap A_i) = \prod_i P(E_i)$$
2. rv' s X_i are independent, if for any $a_1 ... a_n$, the events $\{X_i \leq a_i\}$ are independnet.
3. More formal definition: sigma algebras $\mathcal{F}_1 ... \mathcal{F}_n$ are independent: if for any $A_1 \in \mathcal{F}_1, ..., A_n \in \mathcal{F}_n$, these events $A_1 ... A_n$ are independent.
4. We say rv' s X_i are independent, if the sigma algebra's generated by $\{X_i \leq a\}$ (for all a), are independent.

- if X and Y are independent, then $F_{XY}(x, y) = F_X(x)F_Y(y)$. The same applies to densities. Proved by taking partial derivatives.

- Also, $P(A|B) = P(A)$.

- Independence is much stronger than mean independence.

- If $X_1 ... X_n$ are mutually independent, then $Var(\sum_i X_i) = \sum_i Var(X_i)$

- Find the variance of OLS using the formula for conditional variance. $\hat{\beta} = (\frac{1}{n}\sum_i x_i x_i^T)^{-1}\frac{1}{n}\sum_i x_i y_i$ Suppose ϵ and X are independent.

  So $\hat{\beta} - \beta = (\frac{1}{n}\sum_i x_i x_i^T)^{-1}\frac{1}{n}\sum_i x_i \epsilon_i$. Now $E\hat{\beta} - \beta|X = 0$.
$$\text{var}(\hat{\beta}|X) = (\frac{1}{n}\sum_i x_i x_i^T)^{-1}\text{var}(\frac{1}{n}\sum_i x_i \epsilon_i|X)(\frac{1}{n}\sum_i x_i x_i^T)^{-1}$$
$$= (\frac{1}{n}\sum_i x_i x_i^T)^{-1}\frac{1}{n^2}\sum_i x_i \text{var}(\epsilon_i|X) x_i^T(\frac{1}{n}\sum_i x_i x_i^T)^{-1}$$
$$= (\frac{1}{n}\sum_i x_i x_i^T)^{-1}\frac{1}{n^2}\sum_i x_i x_i^T(\frac{1}{n}\sum_i x_i x_i^T)^{-1}\sigma_e^2 = (\frac{1}{n}\sum_i x_i x_i^T)^{-1}\sigma_e^2$$
  So $\text{var}(\hat{\beta}) = E(\frac{1}{n}\sum_i x_i x_i^T)^{-1}\sigma_e^2$.

1.4.6 Transformations of distributions

- if I know f_X for X, what about any invertible function $Y = g(X)$'s pdf ? By invertible function, I mean $g(g^{-1}(x)) = x$.

In fact suppose g^{-1} is increasing, $P(g(X) \leq y) = P(X \leq g^{-1}(y))$ whose derivative is $f_Y(y) = f_X(g^{-1}(y))\frac{dg^{-1}(y)}{dy}$. If g^{-1} is decreasing, do it similarly but use $|\frac{dg^{-1}(y)}{dy}|$.

- Example: $Y = X^2$. My suggestion is just do it. if $y \leq 0$, density is zero. If $y > 0$,

$$P(Y \leq y) = P(X^2 \leq y) = P(0 \leq X \leq \sqrt{y}) + P(-\sqrt{y} \leq X < 0)$$

$$= F_X(\sqrt{y}) - F_X(0) + F_X(0) - F_X(-\sqrt{y})$$

Differentiate on both sides,

$$f_Y(y) = f_X(\sqrt{y})0.5y^{-1/2} + f_X(-\sqrt{y})(0.5y^{-1/2}).$$

- Two dim,

Suppose X, Y and U, V are one-one mapping. So the two events $(X, Y) \in S$ and $(U, V) \in T$ are equivalent. Then

$$P((U,V) \in T) = P((X,Y) \in S) = \int_S f_{XY}(x,y)d(x,y) = \int_T f(x(u,v), y(u,v))d(x(uv), y(uv))$$

$$= \int_T f_{XY}(x(u,v), y(u,v))|\frac{\partial(x,y)}{\partial(u,v)}|d(u,v)$$

where $|\frac{\partial(x,y)}{\partial(u,v)}|$ means the absolute value of the determinant of the Jacobian

$$\begin{pmatrix} \partial_u x & \partial_v x \\ \partial_u y & \partial_v y \end{pmatrix}$$

This is true for any set T, thus

$$f_{UV} = f_{XY}(x(u,v), y(u,v))|\frac{\partial(x,y)}{\partial(u,v)}|$$

- If we are just interested in one r.v. $U(x,y)$, find $V(x,y)$ so that (XY) and (UV) are one-to-one, then integrate out V.

- Example: Find the pdf of X+Y. Let $V = Y, U = X + Y$. Then

  $$X = U - V, \quad Y = V. \quad \begin{pmatrix} \partial_u x & \partial_v x \\ \partial_u y & \partial_v y \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}$$

  whose determinant is 1. Hence $f_{UV} = f_{XY}(u - v, v)$. Thus

  $$f_U(u) = \int f_{UV}(u, v)dv = \int f_{XY}(u - v, v)dv.$$

  but need to be careful about the range where we integrate.

  Suppose $f_{XY} = x + y$ if $x, y \in (0, 1)$, and is zero otherwise. Then the range of (u, v) is $v \in (0, 1)$, and $u - v \in (0, 1)$, meaning that $v < u < v + 1$.

  We now want different range for u to define the support of f_U. Note that $u - 1 < v < u$. And overall $u \in (0, 2)$.

  Hence when $0 < u < 1$, then $0 < v < u$;

  when $1 \leq u < 2$, then $u - 1 < v < 1$.

  So

  $$f_U(u) = \int f_{UV}(u, v)dv = \int f_{XY}(u-v, v)dv = \begin{cases} \int_0^u udv = u^2 & u \in (0, 1) \\ \int_{u-1}^1 udv = -u^2 + 2u & u \in [1, 2) \end{cases}$$

1.5 Commonly used distributions

1.5.1 bernoulli an Binomial

- It the distribution of a random variable, which takes the 1 with success probability of p and the value 0 with failure probability of $1 - p$. The pmf is

  $$P(X = x) = p^x (1 - p)^{1-x}, x = 0, 1$$

- The study of employment rate or success rate, or proportion of something, can be understood as bernoulli. You randomly pick up a person, either employed or not, probability is p or 1-p.

- Binomial (n,p) is the total number of success from n independent Bernoulli (p) experiments. You randomly pick up n people, X are employed.

$$P(X=x)=\binom{n}{x}p^x(1-p)^{n-x}, x=0,1....,n$$

- sum of two independent binomials: $B(n_1,p)+B(n_2,p)=B(n_1+n_2,p)$.

1.5.2 multinomial (multiple choice model)

It is a generalization of binomial.

- An experiment is conducted independently n times. Each time it has k possibilities: $c_1,...,c_k$, with prob $p_1...p_k$, so $\sum_{i=1}^k p_i=1$.

  Suppose the experiment is conducted independently n times, and you find there are X_1 times c_1; X_2 times c_2, ..., X_k times c_k, so $\sum_{i=1}^k X_i=n$. Then

$$P(X_1=n_1,...,X_k=n_k)=\frac{n!}{n_1!...n_k!}p_1^{n_1}...p_k^{n_k}$$

- A better notation is to replace n_i with x_i.
- So the calculate this distribution, we need to specify n and $p_1....p_k$.

  Example 1.2. multiple choice model. Suppose an unemployed individual has k training programs to choose from: $c_1...c_k$. The choice of each program has a probability (you can imagine a population of unemployed people, they choose different programs. The proportion of each program is p).

  So this is like randomly pick up a person, and observe this person to be in program c_i.

  How to determine these p_i ? Each program gives this individual a utility U_j, so we have $U_1,...,U_k$. The utility is random, and has decomposition

$$utility=systematic+random$$

  So for each $i\leq k$,

$$U_i=\eta_i+\epsilon_i$$

  Here η is not individual specific, non random, called "expected utility". But ϵ is random. So different people have different $(U_1,...,U_k)$. Each person's choice is based on maximizing her individual utility: for each $i \leq k$,

$p_i = P(\text{choose } c_i) = P(\max(U_1...U_k) = U_i) = P(\max(\eta_1+\epsilon_1, ..., \eta_k+\epsilon_k) = \eta_i+\epsilon_i)$

here only ϵ is random. But suppose $\epsilon_1...\epsilon_k$ are independent, and from the same distribution, given by (Type I extreme value distribution, max of infinitely many independent exponential distributions)

$$f(\epsilon) = \exp(-\epsilon - \exp(-\epsilon))$$

then

$$p_i = \frac{\exp(\eta_i)}{\exp(\eta_1) + ... + \exp(\eta_k)}$$

In the special case $k = 2$, we come back to binomial. Then

$$p_1 = \frac{\exp(\eta_1)}{\exp(\eta_1) + \exp(\eta_2)}, \quad p_2 = 1 - p_1$$

This model is used frequently in multinomial models, used for labor economics.

- Example where the extreme value distribution does not work well: suppose $\eta = (\log 2, 0, 0)$. Then $\exp(\eta) = (2, 1, 1)$. Hence $p = (2/4, 1/4, 1/4)$. Now suppose $\eta = (\log 2, 0)$. Then $\exp(\eta) = (2, 1)$. Hence $p = (2/3, 1/3)$. This means, given the same expected utility, if the number of choices is different, then the proportion probabilities can be different.

  For instance, suppose you have choices train, blue bus, red bus. Suppose the expected utility is the same for buses, so people who choose bus are indifference between blue and red. The utility is $(\log 2, 0, 0)$. Then we see $p = (2/4, 1/4, 1/4)$, half take train, people who take bus splits 1-1.

  Now suppose the service of blue bus is discontinued. Then people who took it should switch to red bus, and utility does not change. We should have $p = (2/4, 2/4)$. But with the utility $(\log 2, 0)$, in fact $p = (2/3, 1/3)$.

1.5.3 poisson

- $$P(X = x) = \frac{\lambda^k e^{-\lambda}}{k!}$$

- It describes the probability of a given number of independent events occurring in a unit of time. For instance, it models the number of customers arriving at a counter or call centre, or the number of cars arriving at a traffic light, or the number of losses/claims occurring in a given period of time (insurance).

- λ is the averaged number of occurring.
- $P(poisson = x; \lambda) = \lim_n P(binomial(n, \lambda/n) = x)$. here x is the number of success of n independent experiments (enter the store or not)
- $Poisson(\lambda_1) + Poisson(\lambda_2) = Poisson(\lambda_1 + \lambda_2)$

1.5.4 Normal distribution, chi square, T, F (CAPM)

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$$

- it can be proved $\int f(x) = 1$.
- more general have two components μ, σ^2. Standardization

$$\frac{X - \mu}{\sigma} \sim N(0, 1)$$

- Linear combinations of normal is still normal.
- multivariate normal:

$$f(x) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp(-0.5(x-\mu)'\Sigma^{-1}(x-\mu))$$

- if $X_1 ... X_n$ are independent standard normal, then $\sum X_i$ is chi square (n).

$$f(x) = \frac{1}{\Gamma(n/2) 2^{n/2}} x^{n/2-1} e^{-x/2}, \quad x > 0$$

- **T distribution**

  If $Z \sim N(0,1)$, $V \sim chi^2(n)$, and $Z \perp V$, then

$$T = \frac{Z}{\sqrt{V/n}}$$

  is T with degrees of freedom n.

- It is due to William Gosset's 1908 paper under the pseudonym "student". This distribution is important due to the following result by student: if X1..Xn are iid $N(\mu, \sigma^2)$, then $\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim T(n-1)$

  Here is a proof: first, student's theorem says: if $X_1 ... X_n$ are iid $N(\mu, \sigma^2)$, let $\bar{X}$ be the mean, and $S^2 = \frac{1}{n-1}\sum_i (X_i - \bar{X})^2$ be the sample variance. Then $\bar{X}$ and S^2 are independent, and $(n-1)S^2/\sigma^2 \sim \chi^2(n-1)$.

  Now

  $$\frac{\bar{X}-\mu}{S/\sqrt{n}} = \frac{(\bar{X}-\mu)/(\sigma/\sqrt{n})}{\sqrt{(n-1)S^2/((n-1)\sigma^2)}} = \frac{N(0,1)}{\sqrt{\chi^2(n-1)/(n-1)}} = T(n-1).$$

- As n goes to infinity, T is close to standard normal.

  $$f_{T,n}(x) \propto (1 + \frac{x^2}{n})^{-(n+1)/2}$$

- It is used a lot in statistics

- **F distribution**

  Suppose $X \sim \chi^2(n)$, $Y \sim \chi^2(m)$, $X \perp Y$. Then

  $$\frac{X/n}{Y/m} \sim F_{n,m}$$

- Why is F so useful in statistics?

  $$Z = \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \sim N(0,1), \quad Z^2 \sim \chi^2(1)$$

  $$Y := (n-1)S^2/\sigma^2 \sim \chi^2(n-1)$$

  $Y \perp Z$, so

  $$F_{1,n-1} = \frac{Z^2}{Y/(n-1)} = \frac{n(\bar{X}-\mu)^2/\sigma^2}{S^2/\sigma^2} = \frac{n(\bar{X}-\mu)^2}{S^2}$$

  if the true mean is indeed μ, then the above is $F_{1,n-1}$. Useful in testing problems.

- Application:

Example 1.3. testing CAPM of a single stock (can be extended to more stocks)

$$y_t = \alpha + bf_t + e_t.$$

$Ey = \alpha + bEf$. Test $H_0 : \alpha = 0$. We are paid only for the risk we take. Estimated α using OLS. Under normality and H_0,

$$\frac{\sqrt{T}(\hat{\alpha} - 0)}{\sqrt{1 + (\frac{\bar{f}}{S_f})^2}} \sim N(0, \sigma^2)$$

Intuitively, call

$$\bar{X} = \frac{\hat{\alpha}}{\sqrt{1 + (\frac{\bar{f}}{S_f})^2}}, \quad \mu = 0$$

and call $S^2 = \hat{\sigma}^2$ using the estimated σ. Then it can be proved that

$$F_{1,T-2} \sim \frac{T(\bar{X})^2}{S^2} = \frac{T\hat{\alpha}^2/\hat{\sigma}^2}{1 + (\frac{\bar{f}}{S_f})^2}$$

we lose two df because one is to estimate σ, the other is to estimate e_t. This is GRS test.

- Uniform distribution.
  $f(x) = 1\{a < x < b\}\frac{1}{b-a}$.
  Property: $F_X(X)$ is Uniform(0,1)

1.5.5 Exponential and Pareto distributions

- **Exponential:**

$$\lambda e^{-\lambda x}.$$

- Often used in survival analysis: survival function:

$$S(t) = P(X > t) = 1 - CDF = 1 - \exp(-\lambda t)$$

- It is the only continuous distribution that has the memoryless property:

$$P(X > t + s | X > t) = P(X > s)$$

This means, the probability that you can live for another ten years, given that you are 70 years old, is the same regardless of how old you are.

- **Pareto:** CDF $F(x) = 1 - (\frac{x_m}{x})^\alpha$, $x \geq x_m$.
- Pareto originally used this distribution to describe the allocation of wealth among individuals It increases to 1 very fast first. It means most of individuals in this population have small x. So a larger portion of the wealth is owned by a smaller percentage of the people.
- Pareto principle or the "80-20 rule" which says that 20% of the population controls 80% of the wealth. This rule is determined by α.
- If X is Pareto, then $\log(X/x_m)$ is exponential with $\lambda = \alpha$. Inversely, if Y is exponential, then $x_m e^Y$ is Pareto. Intuitively, Pareto CDF grows not as fast as exponential (Pareto grows polynomially), so Pareto has to be very large $Pareto = \exp(exponential)$, in order to catchup the growth of exponential.

1.6 Important inequalities

Markov $P(|X| > x) \leq E|X|/x$, $\forall x$. This implies $P(|X| > x) < EX^2 < x^2$. We shall use this to prove converges in probability. Intuitively, it means if the moments are small, so will be X.

Chebychev : if g, r nonnegative. $P(g(X) > r) \leq Eg(X)/r$, This implies

$$P(|X - \mu| > t\sigma) \leq \frac{\sigma^2}{t^2\sigma^2} = t^{-2}$$

This implies WLLN.

It also implise $P(|N(0,1)| > t) < \sqrt{\frac{2}{\pi}} t^{-1} \exp(-t^2/2)$.

Proof:

$$P(N(0,1) > t) = \frac{1}{\sqrt{2\pi}} \int_t e^{-x^2/2} dx \leq \frac{1}{\sqrt{2\pi}} \int_t \frac{x}{t} e^{-x^2/2} dx = \frac{1}{\sqrt{2\pi}} \frac{\exp(-t^2/2)}{t}.$$

Now $P(|Z| > t) = 2P(Z > t)$. Similarly, a lower bound is $\sqrt{\frac{2}{\pi}} \frac{t}{t^2+1} e^{-t^2/2}$. So this gives the approximation order of the CDF of standard normal, we see it decays exponentially fast.

Cauchy Schwarz $(EXY)^2 \leq EX^2 EY^2$.

Applications: $|EX| \leq E|X| \leq (EX^2)^{1/2}$. This is why $Var = EX^2 - (EX)^2 \geq 0$.

Holder if $1/p + 1/q = 1$, $p, q \geq 1$, then

$$E|XY| \leq (E|X|^p)^{1/p} (E|Y|^q)^{1/q}$$

Jensen If g is concave, then

$$Eg(X) \leq g(EX)$$

when g is strictly concave, the equality holds iff X is constant

Proof. $\int g(x)dF \leq g(\int xdF)$ for concave g. □

The inequality flips when g is convex.

Applications:

Example 1.4 (Kullback-Leibler divergence)**.** the distance of two densities p, q is measured by Kullback-Leibler divergence: it is not symmetric. So if we have p as the truth, and would like to know how close our guess q to p.

$$D(p||q) = \int p \log(\frac{p}{q}) dx$$

Because log is concave,

$$-D(p||q) = \int p \log(\frac{q}{p}) dx = E(\log(\frac{q(X)}{p(X)})) \leq \log(E\frac{q(X)}{p(X)}) = \log \int p\frac{q}{p} dx = \log 1 = 0.$$

Hence $D(p||q) \geq 0$ for all q. And is zero if $q = p$.

It measures "the distance to p"

Example 1.5 (Risk Aversion)**.** A utility function satisfies:

(1) U is strictly increasing

(2) the first derivative is strictly decreasing. So U(x) is concave.

Then it leads to Risk aversion

$$U(E(X)) \geq EU(X)$$

This means that the investor would prefer to keep the expected payoff EX rather than the expected utility of the investments. (the utility of expected payoff is higher than the expected utilities)

For instance, suppose you decide whether to take a bet, the payoff is X, with $EX = 0$. If you do not take the bet, your utility is $U(0)$. But if you take the bet, in the long run, your expected utility is $EU(X)$. Note that

$$U(0) = U(EX) \geq EU(X)$$

So you will decline any bet whose expected payoff is zero.

2 Large Sample Theory: elementary

2.1 Convergence in Probability

- in prob: $\forall\epsilon$, $P(|X_n - X| > \epsilon) \to 0$.
- equivalently write $X_n \to^P X$.
- Properties: $X_n \to^P X$, $Y_t \to^P Y$, then $X_n \pm Y_n, X_nY_n, aX_n$ all converges to their obvious counterparts. Here these things can be also vectors or matrices.
- $o_P(1)$: if $X_n \to^P 0$;

  $O_P(1)$ if $\forall\epsilon > 0, \exists C > 0$, $P(|X| > C) < \epsilon$.
- stronger convergence: converges almost surely: $P(X_n \to X) = 1$. All properties for convergence in probability carry over.

2.2 WLLN and consistency

- Suppose X1... Xn are independent. And all their variances are bounded by a constant C.
- Then $Var(\bar{X}) = \frac{1}{n^2}\sum_i Var(X_i) < C/n$, implying, for any $\epsilon > 0$, by Markov,

$$P(|\bar{X} - \frac{1}{n}\sum_i EX_i| > \epsilon) \leq \frac{Var(\bar{X})}{\epsilon^2} \leq \frac{C}{\epsilon^2 n} \to 0.$$

This implies $\bar{X} - \frac{1}{n}\sum_i EX_i \to^P 0$. In particular, if their expectations are the same, then $\bar{X} \to^P EX$. This is known as the weak law of large number.

- Also, by Markov, any sequence $X_n \to^P X$ as long as $E(X_n - X)^2 \to 0$.
- Consistency: Given the data X1...Xn, a statistic is a function $T(X1...Xn)$. We say it is consistent for estimating θ if

$$T \to^P \theta.$$

  The above result shows, usually we can directly check whether $E(T - \theta)^2 \to 0$ by finding $Var(T)$ and bias of T.

2.3 convergence in distribution

- Now we know $\bar{X} - EX \to^P 0$, how fast does it go ? For any sequence $a_n \to \infty$ but slower than $\sqrt{n}$, we have, since $n/a_n^2 \to \infty$.

$$P(a_n|\bar{X} - \frac{1}{n}\sum_i EX_i| > \epsilon) \leq \frac{Var(\bar{X})}{\epsilon^2} \leq \frac{C}{\epsilon^2 n/a_n^2} \to 0.$$

  Thus

$$a_n|\bar{X} - EX| \to^P 0$$

- But when $a_n = \sqrt{n}$, it is no longer the case. In fact $\sqrt{n}|\bar{X} - EX|$ is a random variable that does not vanish, even if n is large. The distribution is normal, this is called CLT.
- To understand CLT, we start with convergence in distribution.
- Definition: $X_n \to^d X$, if for any x at which F_X is continuous,

$$F_n(x) \to F(x)$$

  This is true for any X as long as its CDF is F. So usually we say $X_n \to^d F$ instead.
- Why we only require continuous points ? Because if we require convergence everywhere, that definition requires too much, and not very useful then. For instance, it is likely $\lim F_n$ exists but not right cts, and thus not a CDF. so if we require $\lim F_n = F$ everywhere, then no CDF exists. This will rule out applications to these kinds of F_n.

- For example, suppose F_n is some CDF, and such that $F_n(x) = \Phi(\sqrt{n}x)$. Then

  $$\lim F_n = \begin{cases} 0 & x < 0 \\ 1/2 & x = 0 \\ 1 & x > 0 \end{cases}$$

  $\lim F_n$ is not right continuous, so is not a CDF. But we should not define "convergence distribution" conservatively because then X_n does not converge.

  If we only require converges at continuous points, then

  $$F_n \to^d F = \begin{cases} 0 & x < 0 \\ 1 & x \geq 0 \end{cases}$$

- Slutsky's theorem: If $X_n \to^d X$, $Y_n \to^P Y$, then $X_n Y_n, X_n + Y_n, Y_n^{-1} X_n$ converges in distribution.

- Continuous mapping theorem:

  (1) in prob: if $X_n \to^P c$ a constant c, and g is continuous at c, then $g(X_n) \to^P g(c)$.

  (2) in dist:

  if $X_n \to^d X$, and g is continuous up to a zero prob measure, meaning that the set on which g is discontinuous, D_g, satisfies $P(X \in D_g) = 0$, then

  $$g(X_n) \to^d g(X)$$

2.4 CLT

- Suppose X1... Xn are iid with bounded variance, then var$^{-1/2}\sqrt{n}(\bar{X} - mean) \to^d N(0, 1)$.

- The point is, regardless of the distribution of X, its mean is normal asymptotically. This makes lots of inferences based on random samples possible.

- Other types of CLT for independent data (non iid), see wikipedia

2.4.1 Proof

- characteristic functions

$$\varphi_X(t) = Ee^{itX} : \quad i = \text{imaginary unit}$$

Chara. function exists for ANY distribution. Two distributions are identical iff their chara. functions are the same.

- For standard normal,

$$\varphi_Z(t) = \exp(-\frac{t^2}{2})$$

- To prove CLT, we just need to prove

$$\varphi_{Z_n}(t) \to \exp(-\frac{t^2}{2}), \quad \forall t$$

where $Z_n = \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}$.

Proof. : let $Z_j = (X_j - \mu)/\sigma$. Then $Z_n = \sqrt{n}\bar{Z}$.

$$\varphi_{Z_n}(t) = Ee^{itZ_n} = E\exp(it\sqrt{n}\bar{Z}) = E\exp(\sum_j it\frac{1}{\sqrt{n}}Z_j) = E\prod_j \exp(it\frac{1}{\sqrt{n}}Z_j)$$

$$\underbrace{=}_{indept} \prod_j E\exp(it\frac{1}{\sqrt{n}}Z_j) = \prod_j \varphi_{Z_j}(\frac{t}{\sqrt{n}}) \underbrace{=}_{identical} \varphi_Z^n(\frac{t}{\sqrt{n}}).$$

Now

$$\varphi_Z(t) = \varphi_Z(0) + t\varphi_Z'(0) + \frac{t^2}{2}\varphi_Z''(0) + b(t)$$

where $b(t) = \frac{t^3}{6}\varphi_Z'''(ct)$, for some $c \in (0,1)$ depending on t.
Now $\varphi_Z(0) = 1$; $\varphi_Z'(t) = Ee^{itZ}iZ$; $\varphi_Z''(t) = -EZ^2e^{itZ}$. So $\varphi_Z'(0) = EiZ = 0$; $\varphi_Z''(0) = -EZ^2 = -1$. So

$$\varphi_Z(t) = 1 - \frac{t^2}{2} + b(t)$$

Thus

$$\varphi_Z(\frac{t}{\sqrt{n}}) = 1 - \frac{t^2}{2n} + b(\frac{t}{\sqrt{n}})$$

so

$$\varphi_{Z_n}(t) = (1 - \frac{t^2}{2n} + b(\frac{t}{\sqrt{n}}))^n, \quad b(\frac{t}{\sqrt{n}}) = \frac{1}{6}(\frac{t}{\sqrt{n}})^3 \varphi_Z'''(\tilde{t})$$

Now a key step is to argue that

$$(1 - \frac{t^2}{2n} + b(\frac{t}{\sqrt{n}}))^n \approx (1 - \frac{t^2}{2n})^n,$$

this is understood intuitively here. Then

$$(1 - \frac{t^2}{2n})^n \to e^{-t^2/2}$$

2.5 delta-method

- It is an application of the CMT. Suppose g is twice differentiable, and $\sqrt{n}(\bar{X} - \mu) \to^d N(0, \sigma^2)$, then

  $$\sqrt{n}(g(\bar{X}) - g(\mu)) \to^d N(0, g'(\mu)^2 \sigma^2)$$

- Proof:

  $$g(\bar{X}) - g(\mu) = g'(W)(\bar{X} - \mu)$$

  where Z lies on the segment joining $\bar{X}, \mu$. Thus, $\bar{X} \to^P \mu$ implies $W \to^P \mu$. Thus continuous mapping theorem says $g'(W) \to^P g'(\mu)$. Thus Slutsky theorem says $\sqrt{n} g'(W)(\bar{X} - \mu) \to^d Z$, where $Z =^d N(0, \sigma^2) g'(\mu) =^d N(0, g'(\mu)^2 \sigma^2)$.

- multidimensional: suppose X is multidimensional and $\sqrt{n}(\bar{X} - \mu) \to^d N(0, \Sigma)$. Then

  $$\sqrt{n}(g(\bar{X}) - g(\mu)) \to^d N(0, \nabla g(\mu)' \Sigma \nabla g(\mu).)$$

- Application: ratio estimators:

  Example 2.1. We would like to estimate the proportion of unemployed people within female $P(unemployed|female)$. We randomly pick up n people, among whom m are female , and z of these female are unemployed. Our estimator is z/m. What about its variance ?

  $$X = 1\{female\}, \quad Y = 1\{female, unemployed\}$$

$$z = \sum_{i=1}^{n} Y_i, \quad m = \sum_{i=1}^{n} X_i$$

so $z/m = \bar{Y}/\bar{X}$. We are looking for the variance of $\bar{Y}/\bar{X}$.

Then (Y, X)'s mean is (μ_y, μ_x). Covariance is

$$\Sigma = \begin{pmatrix} \mu_y(1-\mu_y) & \sigma_{xy} \\ \sigma_{xy} & \mu_x(1-\mu_x) \end{pmatrix}$$

where

$$\sigma_{xy} = E(XY) - \mu_x\mu_y = P(XY=1) - \mu_x\mu_y = P(Y=1) - \mu_x\mu_y = (1-\mu_x)\mu_y$$

Let $g(y,x) = y/x$. Then $\nabla g = (1/x, -yx^{-2})$. Hence variance is

$$\nabla g(\mu)'\Sigma\nabla g(\mu) = (\frac{1}{\mu_x}, -\frac{\mu_y}{\mu_x^2}) \begin{pmatrix} \mu_y(1-\mu_y) & (1-\mu_x)\mu_y \\ (1-\mu_x)\mu_y & \mu_x(1-\mu_x) \end{pmatrix} \begin{pmatrix} \mu_x^{-1} \\ -\mu_y/\mu_x^2 \end{pmatrix}$$

$$= (\mu_x^{-1}\mu_y(1-\mu_y) - \mu_y^2(1-\mu_x)/\mu_x^2, 0) \begin{pmatrix} \mu_x^{-1} \\ -\mu_y/\mu_x^2 \end{pmatrix} = \mu_x^{-2}\mu_y(1-\mu_y) - \mu_y^2(1-\mu_x)/\mu_x^3$$

$$= \frac{(\mu_x - \mu_y)\mu_y}{\mu_x^3}$$

To estimate: $\mu_x \approx m/n, \mu_y \approx z/n$. Hence estimate variance to be

$$\frac{1}{n}\frac{(\mu_x - \mu_y)\mu_y}{\mu_x^3} \approx \frac{(m-z)z}{m^3}$$

- As an example, suppose $P(unemployed) = 0.1$; and unemployed is independent of sex. I calculate z/m. I repeat this 100 times, and get the empirical variance of these 100 estimated z/m. Compare it with the theoretical

  $$\frac{1}{n}\frac{(\mu_x - \mu_y)\mu_y}{\mu_x^3} = \frac{1}{n}\frac{(0.5 - 0.1 \times 0.5)0.1 \times 0.5}{0.5^3} = \frac{0.18}{n}$$

  see figure 1. Matlab code:

  ```
  or=100;                  % number of replications for a fixed n
  sample=[200:50:1000];
  ```

```
for i=1:length(sample)
n=sample(i);
female=binornd(1,0.5,n,or);                % 1 if female
unemploy=binornd(1,0.1,n,or);                % 1 if unemployed
Z=female.*unemploy;                % female and unemployed
m=sum(female,1);                % number of females
z=sum(Z,1);                % number of female and unemployed
est=z./m;
esvar(i)=var(est);
th(i)=0.18/n;
end;
plot(sample,esvar,sample, th);
```

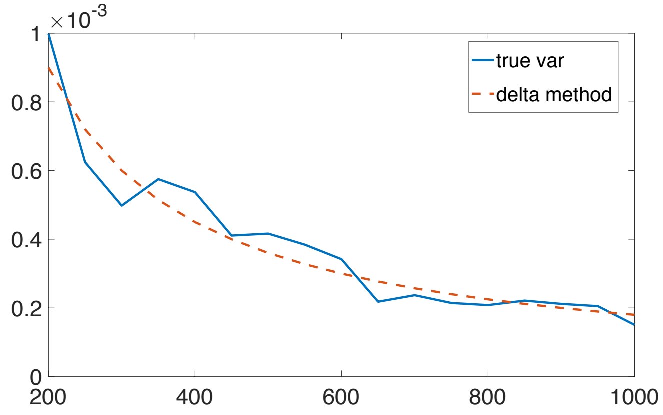


Figure 1: var of unemployment: as sample size increases

2.6 Example for OLS

- Consistency of OLS.

$$\hat{\beta} = (\frac{1}{n}\sum_i x_i x_i^T)^{-1}\frac{1}{n}\sum_i x_i y_i$$

then

$$\hat{\beta} - \beta = (\frac{1}{n}\sum_i x_i x_i^T)^{-1}\frac{1}{n}\sum_i x_i \epsilon_i$$

Now by WLLN, $\frac{1}{n}\sum_i x_i x_i^T \to^P Exx^T$. $\frac{1}{n}\sum_i x_i\epsilon_i \to^P 0$. Thus Hence by the continuous mapping, we get the consistency of $\hat{\beta}$.

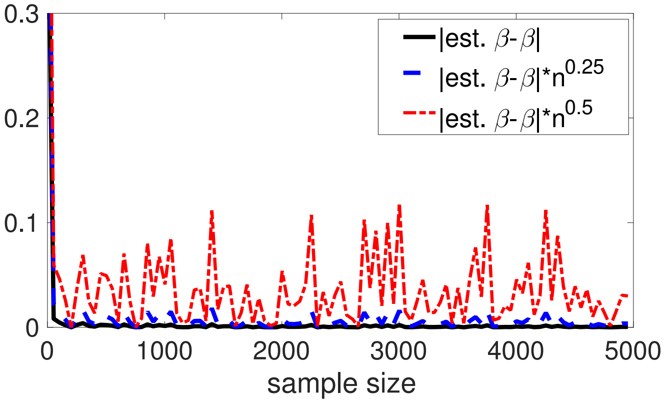
- Normality: if $Var(x_i\epsilon_i) < C$, then $\sqrt{n}\frac{1}{n}\sum_i x_i\epsilon_i \to^d N(0, Var(x_i\epsilon_i))$ Hence by Slutsky's theorem

  $$\sqrt{n}(\hat{\beta}-\beta) \to^d (Exx^T)^{-1}\times N(0, Var(x_i\epsilon_i)) =^d N(0, (Exx^T)^{-1}Var(x\epsilon)(Exx^T)^{-1})$$

- In the example below, we estimate $y = a + x\beta + e$, $e \sim N(0,1)$, $a = 1, \beta = 3$. We plot, as a function of n, the average of 500 replications of

  $$n^c|\hat{\beta} - \beta|$$

  for $c = 0, 0.25, 0.5$.



3 Large Sample Theory: Empirical process

Consider $\theta \in \Theta$ as a parameter space
Consider a function $f(X,\theta)$ and iid data

$$\sup_{\theta\in\Theta}|\frac{1}{n}\sum_i[f(X_i,\theta) - Ef(X_i,\theta)]|$$

More generally, let $\mathcal{F}$ as a (nonrandom) function family, such as

$$\{f(.,\theta) : \theta \in \Theta\}$$

we are interested in

$$\sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_i [f(X_i) - Ef(X_i)]|$$

3.1 ULLN

the first question is if

$$\sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_i [f(X_i) - Ef(X_i)]|$$

converge. This is known as ULLN

Here we prove the convergence of

$$E\sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_i [f(X_i) - Ef(X_i)]|$$

3.1.1 subGaussian

Def

$$P(|X| \geq t) \leq 2\exp(-ct^2/K^2)$$

up to constants, the smallest possible K is the subGaussian norm, defined as

$$\|X\|_\psi = \inf\{t > 0 : E\exp(X^2/t^2) \leq 2\}$$

Lemma 1

Suppose: and $|W_{id}| \leq M_n$ and the sequence is independent, then Then

$$E\max_{d=1...N}\left|\frac{1}{n}\sum_i W_{id}\right| \leq C\sqrt{\frac{\log N}{n}}$$

Lemma 2

if $W_{1d}..W_{nd}$ are independent zero mean subGaussian, then $\frac{1}{\sqrt{n}}\sum_i W_{id}$ is also subGaussian

$$E\max_{d=1...N}|\frac{1}{n}\sum_i W_{id}| \leq C\sqrt{\frac{\log N}{n}}\max_{d=1...N}(\frac{1}{n}\sum_i \|W_{id}\|_\psi^2)^{1/2}$$

- **Proof of Lemma 1**

  First,
  $$E\exp(2tW_{id}) \leq \exp(Ct^2)$$
  for any $t > 0$, by Jensen
  $$\exp(t2E\max_{d=1...N}\left|\sum_i W_{id}\right|) \leq E\exp(2t\max_{d=1...N}\left|\sum_i W_{id}\right|) \leq E\sum_{d=1...N}\exp(2t\left|\sum_i W_{id}\right|)$$
  $$= NE\exp(2t\left|\sum_i W_{id}\right|) \leq NE\exp(2t\sum_i W_{id}) + NE\exp(-2t\sum_i W_{id})$$
  $$\leq N(E\exp(2tW_{id}))^n + N(E\exp(-2tW_{id}))^n \leq 2N\exp(Cnt^2)$$
  So
  $$2E\max_{d=1...N}\left|\sum_i W_{id}\right| \leq \frac{\log 2 + \log N}{t} + Cnt$$
  The optimal $t \sim \sqrt{\frac{\log N}{n}}$.

- **Proof of Lemma 2** : it is from the following two more general lemmas:

  Lemma 3 Suppose X_i is subGaussian but the sequence does not have to be independent then
  $$E\max_{i=1...N}|X_i| \leq C\sqrt{\log N}\max_i \|X_i\|_\psi$$

  Lemma 4 if $W_1..Wn$ are independent zero mean subGaussian, then $\frac{1}{\sqrt{n}}\sum_i W_i$ is also subGaussian
  $$\|\frac{1}{\sqrt{n}}\sum_i W_i\|_\psi^2 \leq C\frac{1}{n}\sum_i \|W_i\|_\psi^2$$

  The usual CS inequality would make $\|\frac{1}{\sqrt{n}}\sum_i W_i\|^2 \leq \sum_i \|W_i\|^2$. So this new bound is much sharper than CS.

3.1.2 symmetrization

Theorem 3.1. *Let ϵ_i be Radamacher sequence: $P(\epsilon = \pm 1) = 1/2$ independent of $X1...Xn$ Then for ANY $\mathcal{F}$ and ANY sequence $X1...Xn$*

$$E \sup_{f \in \mathcal{F}} |\frac{1}{n} \sum_i [f(X_i) - Ef(X_i)]| \leq 2E \sup_f \left| \frac{1}{n} \sum_i \epsilon_i f(X_i) \right|$$

Proof. Let Yi be an identically distributed copy of Xi, and $(Y1...Yn)$ is independent of $(X1...Xn)$ So

$$Ef(X_i) = Ef(Y_i) = Ef(Y_i)|X1...Xn$$

Also,

$$f(X_i) = Ef(X_i)|X1...Xn$$

So

$$|\frac{1}{n} \sum_i [f(X_i) - Ef(X_i)]| = |E \left(\frac{1}{n} \sum_i f(X_i) - f(Y_i) \middle| X_1 ... X_n \right) |$$

$$\leq E \sup_f \left| \frac{1}{n} \sum_i f(X_i) - f(Y_i) \right| \Big| X_1 ... X_n$$

So

$$E \sup_{f \in \mathcal{F}} |\frac{1}{n} \sum_i [f(X_i) - Ef(X_i)]| \leq E \sup_f \left| \frac{1}{n} \sum_i f(X_i) - f(Y_i) \right|$$

Now le ϵ be Radamacher: $P(\epsilon = \pm 1) = 1/2$.
then if $W =^d -W$, and $W \perp \epsilon$, then $W =^d \epsilon W$

$$f(X_i) - f(Y_i) =^d f(Y_i) - f(X_i) =^d \epsilon_i (f(X_i) - f(Y_i))$$

So

$$E \sup_{f \in \mathcal{F}} |\frac{1}{n} \sum_i [f(X_i) - Ef(X_i)]| \leq E \sup_f \left| \frac{1}{n} \sum_i \epsilon_i (f(X_i) - f(Y_i)) \right|$$

$$\leq 2E \sup_f \left| \frac{1}{n} \sum_i \epsilon_i f(X_i) \right|$$

□

3.1.3 ϵ-cover

- ϵ-cover is the smallest number of ϵ- balls to cover $\mathcal{F}$

  $$N(\mathcal{F}, \|.\|, \epsilon) = \min\{N : \text{there are N balls radius } \epsilon : B_1(\epsilon), ..., B_N(\epsilon), \mathcal{F} \subset \cup_{i=1}^N B_i(\epsilon)\}$$

  here the radius is defined using $\|.\|$.
  If $\mathcal{F}$ is compact in the $\|.\|$-space, then immediately N is finite.

- The idea is to divide the complex set $\mathcal{F}$ into a collection of small balls.
  For example

  $$\sup_f H(f) \leq \sup_{f, g_i \text{ is the ball center closet to f}} \|H(f) - H(g_i)\| + \max_{i \leq N} \|H(g_i)\| \leq \epsilon C + \max_{i \leq N} \|H(g_i)\|$$

  The problem becomes $\max_{i \leq N} \|H(g_i)\|$. N is not so big. So the problem becomes not so complex.

- Come back to the proof

  $$2E \sup_f \left| \frac{1}{n} \sum_i e_i f(X_i) \right| \leq \underbrace{2E \sup_f \left| \frac{1}{n} \sum_i e_i f(X_i) 1\{|f(X_i)| > M_n\} \right|}_{I}$$
  $$+ \underbrace{2E \sup_f \left| \frac{1}{n} \sum_i e_i f(X_i) 1\{|f(X_i)| \leq M_n\} \right|}_{II}$$

- I is bounded by

  $$I \leq 2E \left| \sup_f |f(X_i) 1\{\sup_f |f(X_i)| > M_n\} \right|$$

  which $= o(1)$ if $E \sup_f |f(X)|^2 < \infty$ as $M_n \to \infty$.

- II is bounded using covering number. Define

  $$\mathcal{F}_{M_n} = \{f 1\{|f(.)| < M\} : f \in \mathcal{F}\}$$

Let $\cup_{d=1}^{N} B_d(\epsilon)$ be ϵ-cover of $\mathcal{F}_{M_n}$, here

$$N = N(\mathcal{F}_{M_n}, \|.\|, \epsilon)$$

Entropy number:

$$\log N(\mathcal{F}_{M_n}, \|.\|, \epsilon) = o(n)$$

Let g_d be the center of those balls.

$$II = 2E \sup_{g\in\mathcal{F}_{M_n}} \left| \frac{1}{n} \sum_i e_i g(X_i) \right| \leq 2E \sup_{g\in\cup_{d=1}^{N} B_d(\epsilon)} \left| \frac{1}{n} \sum_i e_i g(X_i) \right|$$

$$= 2E \max_{d=1\ldots N} \sup_{g\in B_d(\epsilon)} \left| \frac{1}{n} \sum_i e_i g(X_i) \right|$$

$$\leq 2E \max_{d=1\ldots N} \sup_{g\in B_d(\epsilon)} \left| \frac{1}{n} \sum_i e_i g(X_i) - e_i g_d(X_i) \right| + 2E \max_{d=1\ldots N} \left| \frac{1}{n} \sum_i e_i g_d(X_i) \right|$$

$$\leq 2\epsilon + 2E \max_{d=1\ldots N} \left| \frac{1}{n} \sum_i e_i g_d(X_i) \right|$$

Now g_d is bounded, and Radamacher is bounded.

- To bound

$$2E \max_{d=1\ldots N} \left| \frac{1}{n} \sum_i e_i g_d(X_i) \right|$$

let $W_{id} := e_i g_d(X_i)$. (we know $|g(.)| < M_n$), So $|W_{id}| \leq M_n$.

$$2E \max_{d=1\ldots N} \left| \frac{1}{n} \sum_i W_{id} \right| \leq C\sqrt{\frac{\log N}{n}}$$

- Put together

$$E \sup_{f\in\mathcal{F}} |\frac{1}{n} \sum_i [f(X_i) - Ef(X_i)]| \leq o(1) + 2\epsilon + C\sqrt{\frac{\log N}{n}} = o(1)$$

if

$$\log N(\mathcal{F}_{M_n}, \|.\|, \epsilon) = o(n)$$

3.2 Convergence rate

The ULLN does NOT provide rate of convergence. This section aims to prove

$$E\sup_{f\in\mathcal{F}}|\frac{1}{n}\sum_i f(X_i)-Ef(X_i)|=O(\frac{1}{\sqrt{n}})$$

The first step is still the symmetrization:

$$E\sup_{f\in\mathcal{F}}|\frac{1}{n}\sum_i f(X_i)-Ef(X_i)|\leq 2E\sup_{f\in\mathcal{F}}|\frac{1}{n}\sum_i f(X_i)\epsilon_i|:=\frac{1}{\sqrt{n}}E\sup_{f\in\mathcal{F}}|Z(f)|$$

where ϵ_i is iid Radamacher sequence and $Z(f):=\frac{1}{\sqrt{n}}\sum_i 2f(X_i)\epsilon_i$.

Bounding $E\sup_{f\in\mathcal{F}}|Z(f)|$ requires a finer argument, known as "chaining".

3.2.1 Chaining

overview (1) Recall that ϵ-cover is a cover on $\mathcal{F}$ so that

$$\forall f\in\mathcal{F}\Rightarrow\exists\pi(f)\in\epsilon\text{-cover},\quad \|f-\pi(f)\|<\epsilon$$

The number of such balls is $N(\mathcal{F},\|.\|,\epsilon)$.

Call such cover (or net) be T.

We now make this ϵ-cover finer and finer, by decreasing ϵ.

(2) Fix f_0,

$$E\sup_{f\in\mathcal{F}}|Z(f)|\leq E\sup_{f\in\mathcal{F}}|Z(f)-Z(f_0)|+E|Z(f_0)|$$

The first term is the key.

The main idea is still to reduce the number of "effective elements " in $\mathcal{F}$, transforming $f\in\mathcal{F}$ to a smaller set using ϵ-cover.

step 1: build nets start with a crude net $T_0=\{f_0\}$, which is just a point, then suppose $\sup_{\mathcal{F}}\|f\|<0.5\bar{C}$. We have

$$\sup_f\|f_0-f\|<\bar{C}$$

thus we set

$$T_0=\{f_0\},\quad \pi_0(f)=f_0,\quad \epsilon_0=\bar{C}$$

Now let the net finer , there are $\pi_1(f) \in T_1$, so that

$$\epsilon_1 = \epsilon_0 2^{-1}, \quad \|\pi_1(f) - f\| < \epsilon_1$$

continuously make the net finer and finer:
there are $\pi_k(f) \in T_k$ so that

$$\epsilon_k = \epsilon_0 2^{-k}, \quad \|\pi_k(f) - f\| < \epsilon_k$$

step 2: chaining Now consider a walk

$$f_0 \to f$$

from the fixed f_0 to an arbitrary $f \in \mathcal{F}$: along the chain:

$$f_0 = \pi_0(f) \to \pi_1(f) \to \pi_2(f) \to .... \to f$$

Then

$$Z(f) - Z(f_0) = \sum_{k=0}^{\infty} Z(\pi_{k+1}(f)) - Z(\pi_k(f))$$

$$E \sup_{f \in \mathcal{F}} |Z(f) - Z(f_0)| \leq E \sum_{k=0}^{\infty} \sup_{f \in \mathcal{F}} |Z(\pi_{k+1}(f)) - Z(\pi_k(f))|$$

Now $\pi_{k+1}(f)$ and $\pi_k(f)$ are close:

$$\|\pi_{k+1}(f) - \pi_k(f)\| \leq \|\pi_{k+1}(f) - f\| + \|\pi_k(f) - f\| \leq \epsilon_{k+1} + \epsilon_k \leq 2\epsilon_k.$$

Although it is still $\sup\{f \in \mathcal{F}\}$, this in fact is a much smaller sup, because it equivalent to:

$$\sup_{f \in \mathcal{F}} |Z(\pi_{k+1}(f)) - Z(\pi_k(f))|$$
$$\leq \max_{g_1, g_2 \in A_k} |Z(g_1) - Z(g_2)|$$

where

$$A_k = \{(g_1, g_2) : \|g_1 - g_2\| \leq 2\epsilon_k, g_1 \in T_{k+1}, g_2 \in T_k\}$$

The number of elements in T_k is bounded by the ϵ_k-cover number.

So

$$E\sup_{f\in\mathcal{F}}|Z(f)-Z(f_0)|\le E\sum_{k=0}^{\infty}\max_{g_1,g_2\in A_k}|Z(g_1)-Z(g_2)|$$

where

$$|A_k|\le|T_{k+1}||T_k|\le N(\mathcal{F},\|.\|,\epsilon_k)^2$$

step 3: subGaussian recall

$$Z(f):=\frac{1}{\sqrt{n}}\sum_i W_i(f)$$

where

$$W_i(f)=2f(X_i)\epsilon_i$$

$$E\sup_{f\in\mathcal{F}}|Z(f)-Z(f_0)|\le E\sum_{k=0}^{\infty}\max_{g_1,g_2\in A_k}|\frac{1}{\sqrt{n}}\sum_i W_i(g_1)-W_i(g_2)|$$

Suppose (1) $W_i(f)$ is subGaussian, then

$$\frac{1}{\sqrt{n}}\sum_i W_i(g_1)-W_i(g_2)$$

is also subGaussian.

Suppose (2) we can interchange E with the infinite sum, then by subGaussian properties,

$$E\sup_{f\in\mathcal{F}}|Z(f)-Z(f_0)|\le\sum_{k=0}^{\infty}E\max_{g_1,g_2\in A_k}|\frac{1}{\sqrt{n}}\sum_i W_i(g_1)-W_i(g_2)|$$

$$\le\sum_{k=0}^{\infty}C\sqrt{\log N(\mathcal{F},\|.\|,\epsilon_k)}\max_{g_1,g_2\in A_k}(\frac{1}{n}\sum_i\|W_i(g_1)-W_i(g_2)\|_{\psi}^2)^{1/2}$$

step 4: covering number

$$\le\sum_{k=0}^{\infty}C\sqrt{\log N(\mathcal{F},\|.\|,\epsilon_k)}\max_{g_1,g_2\in A_k}(\frac{1}{n}\sum_i\|g_1(X_i)-g_2(X_i)\|_{\psi}^2)^{1/2}$$

$$\leq \sum_{k=0}^{\infty} C\sqrt{\log N(\mathcal{F}, \|.\|, \epsilon_k)} \max_{g_1, g_2 \in A_k} \sup_x \|g_1(x) - g_2(x)\|$$

$$\leq \sum_{k=0}^{\infty} C\epsilon_k \sqrt{\log N(\mathcal{F}, \|.\|_\infty, \epsilon_k)}$$

so we set $\|.\| = \|.\|_\infty = \sup_x \|.\|$

$\epsilon_k - \epsilon_{k+1} = \epsilon_k - \frac{1}{2}\epsilon_k = \frac{1}{2}\epsilon_k$

$$\epsilon_k = 2\int_{\epsilon_{k+1}}^{\epsilon_k} dt$$

when $t < \epsilon_k$, the cover is finer, needs more balls , so

$$N(\mathcal{F}, \|.\|_\infty, \epsilon_k) \leq N(\mathcal{F}, \|.\|_\infty, t)$$

so

$$E\sup_{f\in\mathcal{F}} |Z(f) - Z(f_0)| \leq \sum_{k=0}^{\infty} C \int_{\epsilon_{k+1}}^{\epsilon_k} \sqrt{\log N(\mathcal{F}, \|.\|_\infty, \epsilon_k)} dt$$

$$\leq C \sum_{k=0}^{\infty} \int_{\epsilon_{k+1}}^{\epsilon_k} \sqrt{\log N(\mathcal{F}, \|.\|_\infty, t)} dt$$

$$= C \int_{\bar{C}}^{\infty} \sqrt{\log N(\mathcal{F}, \|.\|_\infty, t)} dt$$

But if $\mathcal{F}$ is bounded in $\|.\|_\infty$ then the above can be sharper:

$$\leq C \int_{\bar{C}}^{bound} \sqrt{\log N(\mathcal{F}, \|.\|_\infty, t)} dt$$

Together,

$$E\sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_i f(X_i) - Ef(X_i)| = O(\frac{1}{\sqrt{n}}) + O(\frac{1}{\sqrt{n}}) \int_{\bar{C}}^{bound} \sqrt{\log N(\mathcal{F}, \|.\|_\infty, t)} dt$$

3.2.2 Dudley's inequality

We have in fact proved Dudley's inequality:

Let X_t be a mean zero process on a metric space with sub-Gaussian increments:

$$\|X_t - X_s\|_\psi \leq K\|t-s\|$$

Then

$$E \sup_{t\in T} X_t \leq CK \int_0^\infty \sqrt{\log N(T, \|.\|, \epsilon)} d\epsilon$$

3.2.3 Stochastic equicontinuity

The probability bound for

$$\sup_{\|t-s\|} \|X_t - X_s\|$$

can be proved similarly using the chaining argument. So the chaining is not only applied to expectations.

3.3 Peeling device

3.3.1 General argument of peeling device

Making the main target of interest "not too big or too small".

The goal: prove

$$P(\forall A, X_n(A) \in \Omega^c) \to 1$$

where $X_n(.)$ is some random function, and Ω^c is some event of interest.

Proof. Aim to show:

$$P(\exists A, X_n(A) \in \Omega) \to 0.$$

Let $\|A\|_n$ be some easy and related norm. Consider "skins":

$$B_j := \{A : 2^j \leq \|A\|_n \leq 2^{j+1}\}$$

elements in B_j are "not too big or too small". make sure first that the interesting $\|A\|_n \geq 2$. Then

$$P(\exists A, X_n(A) \in \Omega) = P(\cup_{j=1}^\infty \{\exists A \in B_j, X_n(A) \in \Omega\}) \leq \sum_{j=1}^\infty P(\exists A \in B_j, X_n(A) \in \Omega)$$

Now suppose we can find random function Y_n and constant c, so that

$$\{\exists A \in B_j, X_n(A) \in \Omega\} \subset \{\exists A \in B_j, |Y_n(A)| \geq c\}$$

(make sure they are nearly the same, this "subset" is as dense as possible)

Then

$$P(\exists A, X_n(A) \in \Omega) \leq \sum_{j=1}^{\infty} P(\sup_{A \in B_j} |Y_n(A)| \geq c)$$

Now proceed in either of two ways

- Method 1: probably crude

$$\sum_{j=1}^{\infty} P(\sup_{A \in B_j} |Y_n(A)| \geq c) \leq \sum_{j=1}^{\infty} \frac{E \sup_{A \in B_j} |Y_n(A)|}{c} \to 0$$

- Method 2:

  Show for some $\delta_n \to 0$, and $\epsilon_{nj} \to 0$

$$P(\sup_{A \in B_j} |Y_n(A)| \geq E \sup_{A \in B_j} |Y_n(A)| + \delta_{n,j}) \leq \epsilon_{nj} \quad (*)$$

  So

$$\sum_{j=1}^{\infty} P(\sup_{A \in B_j} |Y_n(A)| \geq c) \leq \sum_{j=1}^{\infty} \epsilon_{nj} + \sum_{j=1}^{\infty} \underbrace{P(c - E \sup_{A \in B_j} |Y_n(A)| \leq \delta_{n,j})}_{0}$$

  Here c is chosen to be much larger than $E \sup_{A \in B_j} |Y_n(A)|$ so that the second term is zero. Also, $\sum_{j=1}^{\infty} \epsilon_{nj} \to 0$. Together

$$P(\exists A, X_n(A) \in \Omega) \to 0$$

- (*) often follows from **Massart inequality**:

  Let

$$Z := \sup_{f \in \mathcal{F}} |\frac{1}{n} \sum_i f(X_i) - Ef(X_i)|$$

  Then:

(i)

$$P(Z \geq EZ + t) \leq \exp(-nt^2/8)$$

provided that almost surely, $|f(X_i)| \leq c_i$ and $\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_i c_i^2 \leq 1$,

(ii)

$$P(Z \geq EZ + \frac{tK}{3} + \sqrt{2t}\sqrt{1 + 2KEZ}) \leq \exp(-nt)$$

provided that $|f(X_i)| \leq K$ almost surely and $\frac{1}{n} \sum_i \sup_{f \in \mathcal{F}} Ef^2(X_i) \leq 1$.

- Finally, to bound $E \sup_{A \in B_j} |Y_n(A)|$ OR EZ, often use symmetrization

$$E \sup_{f \in \mathcal{F}} |\frac{1}{n} \sum_i f(X_i) - Ef(X_i)| \leq 2E \sup_{f \in \mathcal{F}} |\frac{1}{n} \sum_i \epsilon_i f(X_i)|$$

where ϵ_i is Radamacher. Then apply contraction

3.3.2 Nonsmooth M-estimation

- The problem:

Let $l(X_i, \theta)$ be the individual loss

$$M_n(\theta) = \frac{1}{n} \sum_i l(X_i, \theta)$$

$$M(\theta) = EM_n(\theta)$$

Suppose

$$\theta_0 = \arg\min M(\theta)$$

The goal is show the rate of

$$\hat{\theta} = \arg\min M_n(\theta)$$

withOUT assuming the smoothness of M_n, M.

- Assumptions

(1) Let

$$\mathbb{G}_n(\theta) = \sqrt{n}(M_n(\theta) - M(\theta))$$

Intuitively, $\mathbb{G}_n = O_P(1)$. We now **assume the maximal inequality** of $\mathbb{G}_n$:

$$E \sup_{\|\theta-\theta_0\|\le\delta} |\mathbb{G}_n(\theta) - \mathbb{G}_n(\theta_0)| \le c\delta^a$$

for some $a > 0$. This is kind of stochastic continuity, whose proof requires chaining arguments, which we do not pursue.

(2) In addition, we assume the consistency of $\hat{\theta}$.

(3) local curvature

$$\inf_{\|\theta-\theta_0\|<C} \frac{M(\theta) - M(\theta_0)}{\|\theta-\theta_0\|^\kappa} > d$$

The constant κ quantifies the local curvature of the loss function.

(4) Need:

$$a < \kappa$$

- result:
$$\|\hat{\theta} - \theta_0\| = O_P(n^{-\frac{1}{2(\kappa-a)}})$$

  smaller $\kappa - a$ leads to faster rate

  (1) usual case, $a = 1$, $\kappa = 2$; the rate is $\sqrt{n}$.

  (2) jump case, $\kappa = 1$, $a = 1/2$, the rate is n

  (3) max score case. $\kappa - a = 3/2$

- The goal is to relate $\|\hat{\theta} - \theta_0\| >$ ... with $\|\mathbb{G}(\theta_0) - \mathbb{G}(\hat{\theta})\| >$ ..., so that we can use the maximal inequality. Such a connection can be built using the local curvature property of M:

$$\mathbb{G}(\theta_0) - \mathbb{G}(\widehat{\theta}) \ge d\sqrt{n}\|\widehat{\theta} - \theta_0\|^\kappa \quad (*)$$

  Then the max inequality heuristically implies

$$\|\theta_0 - \hat{\theta}\|^a \ge Cn^{1/2}\|\theta_0 - \hat{\theta}\|^\kappa, \quad \kappa > a$$

  This gives the rate

  But strictly speaking, this argument requires $\|\hat{\theta} - \theta_0\|$ is “not too small, not too big”, so its lower and upper bound is basically the same.

- This then requires “peeling device”:

$$2^j < r_n\|\hat{\theta} - \theta_0\| < 2^{j+1}$$

Let $r_n \to \infty$ be the rate of convergence:

$$r_n\|\hat{\theta} - \theta_0\| = O_P(1)$$

The goal is to show, for any $\epsilon > 0$, there is $t > 0$ so that

$$P(r_n\|\hat{\theta} - \theta_0\| > 2^t) < \epsilon$$

The consistency ensures $\|\hat{\theta} - \theta\| < C$ (not very rigorous). So $r_n\|\hat{\theta} - \theta_0\| < r_n C$ The LHS is bounded by

$$P(r_n\|\hat{\theta} - \theta_0\| > 2^t) = \sum_{j>t, 2^{j+1} \le r_n C} P(2^{j+1} > r_n\|\hat{\theta} - \theta_0\| > 2^j)$$

We bound the RHS, to find such t.
If (*) is true then $2^{j+1} > r_n\|\hat{\theta} - \theta_0\| > 2^j$ implies

$$\sup_{2^{j+1} > r_n\|\theta - \theta_0\| > 2^j} |\mathbb{G}(\theta_0) - \mathbb{G}(\theta)| \ge \mathbb{G}(\theta_0) - \mathbb{G}(\widehat{\theta}) \ge d\sqrt{n}(\frac{2^j}{r_n})^\kappa$$

So by the maximal inequality

$$P(2^{j+1} > r_n\|\hat{\theta} - \theta_0\| > 2^j) \le P(\sup_{2^{j+1} > r_n\|\theta - \theta_0\| > 2^j} |\mathbb{G}(\theta_0) - \mathbb{G}(\theta)| \ge d\sqrt{n}(\frac{2^j}{r_n})^\kappa)$$

$$\le E \sup_{2^{j+1} > r_n\|\theta - \theta_0\| > 2^j} |\mathbb{G}(\theta_0) - \mathbb{G}(\theta)| \frac{1}{d\sqrt{n}}(\frac{r_n}{2^j})^\kappa$$

$$\le C\frac{1}{\sqrt{n}}(\frac{2^{j+1}}{r_n})^a(\frac{r_n}{2^j})^\kappa \le C\frac{r_n^{\kappa-a}}{\sqrt{n}}2^{-(\kappa-a)j}$$

So

$$P(r_n\|\hat{\theta} - \theta_0\| > 2^t) = \sum_{j>t, 2^{j+1} \le r_n C} P(2^{j+1} > r_n\|\hat{\theta} - \theta_0\| > 2^j)$$

$$\leq C\frac{r_n^{\kappa-a}}{\sqrt{n}}\sum_{j>t,2^{j+1}\leq r_nC}2^{-(\kappa-a)j}$$

Take t as a large enough constant, we can make the sum be less than ϵ, then

$$P(r_n\|\hat{\theta}-\theta_0\|>2^t)\leq \epsilon\frac{r_n^{\kappa-a}}{\sqrt{n}}<\epsilon$$

So the rate of convergence is

$$r_n\sim n^{-\frac{1}{2}\frac{1}{\kappa-a}}$$

- It remains to show (*), we use local curvature

$$\mathbb{G}_n(\theta_0)=\sqrt{n}[M_n(\theta_0)-M(\theta_0)]\geq\sqrt{n}[M_n(\hat{\theta})-M(\theta_0)]$$
$$=\sqrt{n}[M_n(\hat{\theta})-M(\hat{\theta})+M(\hat{\theta})-M(\theta_0)]\geq\mathbb{G}_n(\hat{\theta})+d\|\hat{\theta}-\theta_0\|^{\kappa}\sqrt{n}$$

3.4 P-Donsker

Def: if

$$\{\mathbb{G}_nf:f\in\mathcal{F}\}\Rightarrow\{\mathbb{G}f:f\in\mathcal{F}\}$$

where $\mathbb{G}$ is a Gaussian process , then $\mathcal{F}$ is called P-donsker

Example:

$$\frac{1}{\sqrt{n}}\sum_i\epsilon_i1\{X_i<t\}$$

Then

$$\{\frac{1}{\sqrt{n}}\sum_i\epsilon_i1\{X_i<t\}:t\in[0,1]\}$$

is P-donsker

3.4.1 Tightness

A stochastic process is denoted by

$$\{X(f):f\in\mathcal{F}\}$$

an easier notation than

$$\{f(X) : f \in \mathcal{F}\}$$

- Tightness:

  An extension of $O_P(1)$ to general metric space:

  For any $\epsilon > 0$, there is a compact set D so that

  $$P(X(.) \in D) > 1 - \epsilon$$

- asymptotically tight

  $$P(X_n(.) \in D^\delta) > 1 - \epsilon$$

  for all δ, ϵ, D compact.

3.4.2 weak convergence of processes

- weak convergence of stoch processes:

  An extension of convergence in distr to stocha process

  $$\{X_n(f) : f \in \mathcal{F}\} \Rightarrow \{X(f) : f \in \mathcal{F}\}$$

  iff for all bounded and continuous operators T,

  $$ET(X_n) \to ET(X)$$

- Theorem:

  $$\{X_n(f) : f \in \mathcal{F}\} \Rightarrow \{X(f) : f \in \mathcal{F}\}$$

  if and only if:

  (1) finite dimensional weak convergence:

  any M and $f_1, ..., f_M$,

  $$((X_n(f_1), ..., X_n(f_M)) \to^d ((X(f_1), ..., X(f_M))$$

  (2) $X_n(.)$ is asymptotically tight, or stoch. equicontinuous

  $$P(\sup_{|f_1 - f_2| \le \delta} |X_n(f_1) - X_n(f_2)| < \epsilon) \to 1$$

$\forall \epsilon, \exists \delta$

A tight and S equi are kind of equivalent, by Th 1.5.7 of VW

3.4.3 Proof of P-Donsker

- Theorem: Suppose $\mathcal{F}$ has an envelop function F and

$$\int_0^\infty \sup_Q \sqrt{\log N(\mathcal{F}, L_2(Q), \|F\|_{L_2(Q)}\epsilon)} d\epsilon < \infty$$

Then $\mathcal{F}$ is P-Donsker

Proof. The goal is to prove asym tightness

$$P(\sup_f \|\mathbb{G}_n f\| > M) < \epsilon$$

In fact,

$$P(\sup_f \|\mathbb{G}_n f\| > M) \leq \frac{1}{M} E \sup_f |\frac{1}{\sqrt{n}} \sum_i f(X_i) - E f_1(X_i)|$$

using the symmetrization

$$\leq \frac{2}{M} E \sup_f |\frac{1}{\sqrt{n}} \sum_i \epsilon_i f(X_i)|$$

covered by ϵ'-balls, with centers g_j

$$\leq \frac{2}{M} E \sup_{f, g_j} |\frac{1}{\sqrt{n}} \sum_i \epsilon_i (f(X_i) - g_j(X_i))| + \frac{2}{M} E \max_j |\frac{1}{\sqrt{n}} \sum_i \epsilon_i g_j(X_i)|$$

The first is further bounded by some kind of "continuity"

The second is further bounded by ϵ-cover number

3.4.4 VC dimensions

Remains to verify

$$\int_0^\infty \sup_Q \sqrt{\log N(\mathcal{F}, L_2(Q), \|F\|_{L_2(Q)}\epsilon)} d\epsilon < \infty$$

- chapter 2.6 of VW Many classes $\mathcal{F}$ satisfy this, known as VC class, whose VC dimension is finite.

  mononone, lipschiz of one parameter

- As for the ULLN, we may need

  $$\int_0^{bound} \sqrt{\log N(\mathcal{F}, \|.\|_\infty, t)} dt < \infty$$

  For Lip function family,

  $$N(\mathcal{F}, \|.\|_\infty, \epsilon) \leq (\frac{C}{\epsilon})^{C/\epsilon}$$

  and $\int_0^1 \sqrt{\frac{C}{t} \log \frac{C}{t} dt} < \infty$

3.5 Functional delta method

3.5.1 Motivation

- The goal is to estimate quantiles. Let F be a CDF

  $$\phi(F, p) = F^{-1}(p), \quad p \in (0, 1)$$

  More generally

  $$\phi(F, p) = \inf\{t : F(t) \geq p\}$$

- To estimate, let

  $$F_n(x) = \frac{1}{n} \sum_i 1\{X_i \leq x\}$$

  The sample quantile is simply

  $$\phi(F_n, p) = \inf\{t : F_n(t) \geq p\}$$

- The problem is to show the convergence of $\phi(F_n, p) - \phi(F, p)$, where regular delta method does not apply because ϕ is not smooth.

3.5.2 Hadamard differentiability

Def, roughly speaking,

(1) ϕ is **Hadamard-differentiable** at θ, if there is a continuous linear function T so that

$$\frac{\phi(\theta + t_n h_n) - \phi(\theta)}{t_n} \to T(h)$$

for all $t_n \to 0$ and $h_n \to h \in$ the domain of ϕ.

(2) ϕ is **Hadamard-Directional differentiable** at θ, if the same holds, but

(*) replace all $t_n \to 0$, to : only require all $t_n \to 0^+$

(*) h does not have to be linear.

We write

$$\phi'_\theta(h) := T(h)$$

- So HDD is weaker

  Note that $\phi(x) = |x|$ is only HDD at zero, bu not H-D at zero.

  (can check easily)

- The above definition is inaccurate. Formal definition can be found easily online.

  For example, we can restrict the condition on h_n to only consider continuous and bounded sequences.

- Check the H-D for

$$\frac{1}{g(x)}$$

Theorem: functional delta method (Fang and Santos)

Suppose $X_n(.)$ is some stochastic process,

$$r_n(X_n - \theta) \Rightarrow X(.)$$

where X is tight.

Suppose $\phi(.)$ is H-DD at θ, then

$$r_n(\phi(X_n) - \phi(\theta)) \Rightarrow \phi'_\theta(X)$$

3.5.3 Estimating quantiles

- First verify the H-D of

  $$\phi(F, .) = \inf\{t : F(t) \geq .\}$$

  where $F(t)$, as a CDF, is a process. Suppose F has a density f bounded away from zero, and that F has bounded support.

  $$\phi(F, p) = F^{-1}(p).$$

  Need to check

  $$\frac{\phi(F + t_n h_n, p) - \phi(F, p)}{t_n}$$

  write $\xi_n = \phi(F + t_n h_n, p)$ and $\xi = \phi(F, p)$.

- First check

  $$(F + t_n h_n)(\xi_n) = p = F(\xi)$$

  So this implies

  $$F(\xi_n) + O(t_n) = F(\xi)$$

  This implies

  $$\xi_n - \xi = O(t_n)$$

  so

  $$F(\xi_n) - F(\xi) = -t_n h_n(\xi_n) = -t_n h_n(\xi) + O(t_n^2)$$

  LHS is $f(\xi)(\xi_n - \xi) + O(t_n^2)$ so

  $$\frac{\xi_n - \xi}{t_n} = -\frac{h_n(\xi)}{f(\xi)} + O(t_n) \to -\frac{h(\xi)}{f(\xi)}$$

- So

  $$\frac{\phi(F + t_n h_n, p) - \phi(F, p)}{t_n} \to -\frac{h(\phi(F, p))}{f(\phi(F, p))} := \phi'_F(h)$$

- So apply F-delta

  $$\sqrt{n}(\phi(F_n, p) - \phi(F, p)) \Rightarrow -\frac{h^*(F^{-1}(p))}{f(F^{-1}(p))}$$

where

$$h^* = \text{ the weak limit of empirical CDF-CDF}$$

$$\sqrt{n}(\frac{1}{n}\sum_i 1\{X_i \leq .\} - F(.)) \Rightarrow h^*(.)$$

- The above proof requires showing $\{\mathbb{G}_n 1\{X \leq t\} : t\}$ is P-Donsker. On a bounded support, this follows from

$$\int_0^\infty \sup_Q \sqrt{\log N(\mathcal{F}, L_2(Q), \|F\|_{L_2(Q)}\epsilon)} d\epsilon < \infty$$

  where for class of functions of x, indexed by t:

$$\mathcal{F} = \{1\{x \leq t\} : t\}$$

  This class is monotonically decreasing , with bounded envelop.

3.5.4 Examples

- Example: then the $|x|$ at zero can satisfy it:

$$\theta = 0, \quad \phi(x) = |x|$$

  for all $t_n \to 0^+$

$$\frac{\phi(\theta + t_n h_n) - \phi(\theta)}{t_n} = \frac{t_n |h_n|}{t_n} = |h_n| \to |h| := \phi_0'(h)$$

  So from $\sqrt{n}(\bar{X} - \theta) \to^d Z$, we have

$$\sqrt{n}(|\bar{X}| - \theta) \to^d \phi_0'(Z) = |Z|$$

  Of course, this result also follows directly from continuous mapping theorem:

$$|\sqrt{n}\bar{X}| \to^d |Z|$$

- Example: it is also easy to check

$$\phi(\theta) = \theta_1 \vee \theta_2, \quad \phi'_\theta(h) = h_1 \vee h_2.$$

(HW)

4 Mathematical Statistic Theory

4.1 Parametrization and Identification

4.1.1 Parametrization

- We are given an iid data $X_1 ... X_n$. We are interested in the distribution of X, called P. Many relationships or economic effects can be characterized as functionals of P.
- Assume $P \in \mathcal{P}$, a family of distributions. Here $\mathcal{P}$ is called "model'.
- For example, suppose $X_i \sim N(\mu, \sigma^2)$, then $\mathcal{P} = \{N(\mu, \sigma^2) : \mu \in \mathbf{R}, \sigma^2 > 0\}$
- In general, we assume $\mathcal{P}$ is partially known: known up to a parameter θ, and write $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. These models are called "parametric" if θ is finite dimensional, instead of functions. If $\mathcal{P}$ is completely unspecified, it is called "nonparametric". In between, it is "semiparametric".

4.1.2 Identification

- There are infinitely many parametrization for a model. For instance, $N(\mu, \sigma^2)$ is the same as $N(\mu, \sigma_1^2 + \mu^2)$. Here $\sigma_1 \neq \sigma_2$, but they are the same model. In general, the problem is: there exist $\theta_1 \neq \theta_2$, but $P_{\theta_1} = P_{\theta_2}$.
- θ is identified (statistician call "identifiable"; econometricians call "identified"; some called "point identified"), if $P_{\theta_1} = P_{\theta_2}$ implies $\theta_1 = \theta_2$.
- Sometimes we are not interested in θ, instead , we are interested in the identification of $q(\theta)$. Then q is identified if $P_{\theta_1} = P_{\theta_2} \Rightarrow q(\theta_1) = q(\theta_2)$, even if $\theta_1 \neq \theta_2$ possibly.
- Example:

$$X_i = \mu + \epsilon_i$$

Assume $\epsilon_i \sim N(\alpha, \sigma^2)$. Then $P_{\mu,\alpha,\sigma^2} = N(\mu+\alpha, \sigma^2)$. But $P_{\mu,\alpha,\sigma^2} = P_{\mu-3,\alpha+3,\sigma^2}$; thus (μ, α, σ) and $(\mu - 3, \alpha + 3, \sigma^2)$ give the same model. To achieve identification, we make further assumptions, just like $\alpha = 0$. Note that this kind of assumption is not unique. We just choose the simple one.

- Example: OLS

  Example 4.1. Consider

  $$Y_i = x_i^T \beta + \epsilon_i, \quad \epsilon \sim N(0, \sigma^2).$$

  - Is β identified? No without further assumptions. If $\dim(x) > 1$, we can always find $v \neq 0$ so that $Ex_i^T v = 0$. Then

    $$Y_i = x_i^T \underbrace{(\beta - v)}_{\tilde{\beta}} + \underbrace{x_i^T v + \epsilon_i}_{\tilde{\epsilon}}, \quad E\tilde{\epsilon}_i = 0$$

    Then $(\tilde{\beta}, \sigma^2)$ gives the same model. But $\beta \neq \tilde{\beta}$,
  - But if we only care about $q(\beta) = Ex^T\beta$, then it is identified as $EY_i = Ex^T\beta = Ex^T\tilde{\beta}$.
  - To identify β, often we further assume **exogeneity**

    $$Ex_i\epsilon_i = 0. \quad rank(Exx^T) = \dim(x)$$

    This implies

    $$Ex_iY_i = Ex_ix_i^T\beta + 0 \Rightarrow \beta = (Exx^T)^{-1}ExY.$$

    Hence β is identified because it is uniquely determined.
  - If $Ex\epsilon = 0$, but the rank condition does not hold, still unidentified. We can always find $v \neq 0$, but satisfy:
    (i) $Exx^Tv = 0$ and (ii) $Ex_i^Tv = 0$.
    Then

    $$Y_i = x_i^T \underbrace{(\beta - v)}_{\tilde{\beta}} + \underbrace{x_i^T v + \epsilon_i}_{\tilde{\epsilon}}$$

    Then $E\tilde{\epsilon} = 0$ and $E\tilde{\epsilon}x = 0$
  - So both exogeneity and rank conditions are needed. In economic applications, the rank condition is often easy to satisfy, but not the exogeneity condition.

-

    Example 4.2 (probit model on binary choice)**.** Recall

$$y_i = \begin{cases} 1 & x_i^T\beta - \epsilon_i > 0 \\ 0 & x_i^T\beta - \epsilon_i < 0 \end{cases}$$

    Now you can always define $\tilde{\beta} = \beta/\sigma$ and $\tilde{\epsilon} = \epsilon/\sigma$.

$$y_i = \begin{cases} 1 & x_i^T\tilde{\beta} - \tilde{\epsilon}_i > 0 \\ 0 & x_i^T\tilde{\beta} - \tilde{\epsilon}_i < 0 \end{cases}$$

    so β and $\tilde{\beta}$ cannot be distinguished.

    A probit model assumes $\epsilon_i \sim N(0, \sigma^2)$.

    Then

$$P(Y = 1|x) = \int_{-\infty}^{x^T\beta} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{t^2}{2\sigma^2})dt = \int_{-\infty}^{x^T\beta} \frac{1}{\sqrt{2\pi}} \exp(-\frac{t^2}{2\sigma^2})d(\frac{t}{\sigma})$$

    let $t/\sigma = z$. then $t < x^T\beta$ implies $z < x^T\beta/\sigma$. then

$$P(Y = 1|x) = \int^{x^T\beta/\sigma} \frac{1}{\sqrt{2\pi}} \exp(-0.5z^2)dz = \Phi(x^T\beta/\sigma)$$

    So $(\beta/\sigma, 1)$ is not identified from (β, σ). Usually assume $\sigma = 1$.

- Mathematical statistics: Suppose identification is given, how to do stat analysis?

    Econometrics: (1) under what conditions is the identification satisfied?

    (2) what if the identification does not satisfy, how to do inference ?

4.2 Sufficiency and Completeness

4.2.1 Sufficiency

- A **statistic** is any function of the data and denoted by $T(X)$ or T.

    Sufficiency is used to reduce the data with statistics without loss of information.

- knowing T is as good as knowing X.
- Precisely: T is called sufficient for θ if the conditional distribution $X|T$ does not involve θ.

  Explanation: the distribution of X depends on θ (e.g, mean=θ), but once T is known, the distribution of X does not contain further information of θ. So the information of θ is stored in "the T part of X".
- Example: toss coin n times and get X1...Xn.

  $$P(X1...Xn) = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$$

  Homework: prove $X|\sum x_i$ does not depend on θ.
- Why do we want sufficient statistics ? An estimator has

  $$MSE = bias^2 + variance$$

  An estimator is unbiased if bias=0;

  Rao-Blackwell theorem: Suppose W is an unbiased estimator for θ; T is a sufficient statistic for θ. Then $E(W|T)$ is unbiased and $Var(E(W|T)) \leq Var(W)$. Hence conditioning on a sufficient statistic will improve, so we should try to consider functions of sufficient statistics. However, this is useful only when we know the conditional distributions.

  Proof

  $$Var(W) = Var(E(W|T)) + EVar(W|T) \geq Var(E(W|T)).$$

  Also note that $W|T$ no longer depends on θ due to the sufficiency.
- Finding sufficient statistics: **General theorem (factorization)**: T is sufficient for θ iff there exists a function $g(t,\theta)$ and a function h so that

  $$f(x|\theta) = g(T(x),\theta)h(x)$$
- Example: independent Poisson:

  $$p(x,\theta) = \prod_i \frac{\theta^{x_i}e^{-\theta}}{x_i!} = \frac{\theta^{\sum x_i}e^{-n\theta}}{\prod x_i!}$$

so $\sum x_i$ is sufficient

- There are many sufficient statistics. T is called **minimum sufficient statistic** if for any other sufficient statistic S, T is a function of S. By function, we mean if $S(x) = S(y)$ then must $T(x) = T(y)$.

  Intuitively: if you know S, then you must know T. So T does not have redundant information compared to S.

- Theorem: let $f(x|\theta)$ be the pdf. Suppose for two sample points x, y, the ratio $f(y|\theta)/f(x|\theta)$ is a constant function of θ iff $W(x) = W(y)$. Then W is a minimum sufficient statistic for θ.

- Example: consider $N(\mu, \sigma^2)$.

$$\frac{f(y|\mu,\sigma^2)}{f(x|\mu,\sigma^2)} = \frac{\exp(n(\bar{y}-\mu)^2 + (n-1)s_y^2)/(2\sigma^2)}{...}$$

$$= \exp(-n((\bar{y})^2 - (\bar{x})^2) + 2n\mu(\bar{y}-\bar{x}) - (n-1)(s_y^2 - s_x^2)/(2\sigma^2))$$

  it does not depend on μ, σ^2 iff $\bar{y} = \bar{x}$ and $s_y^2 = s_x^2$. Hence $(\bar{x}, s_x^2)$ is mim sufficient.

4.2.2 Completeness

- A distribution F is complete if $E_F g(X) = 0$ implies $g(X) = 0$ a.s.

- The book is about complete statistics: T is complete if for any g, it should satisfy:

  $E_\theta g(X) = 0$ for all θ implies $P_\theta(g(X) = 0) = 1$ for all θ.

- The complete statistics is often used to combine with the sufficiency to describe minimum sufficiency, as well as good estimations.

- But in econometrics, we focus more on the distribution's completeness, because this understanding is helpful to understand identification. The distribution completeness implies:

$$E_\theta g_1(X) = E_\theta g_2(X) \Rightarrow g_1 = g_2$$

  Econometric models often describe moment conditions. So if the unknown effect satisfies $E_\theta g_1(X)$ is fixed, then this implies g_1 is identified as long as X's distribution is complete.

Example 4.3 (Nonparametric IV regression)**.** Consider an effect of schooling X on the income.

$$Y = g(X) + U$$

X is endogenous, and suppose there is IV W: proxy to college. Then

$$E(Y|W) = E(g|W)$$

Whether this relation identifies g depends on if $g_1 \neq g_2$ but $E(g_1|w) = E(g_2|w) = E(y|w)$.

If $X|W$ is complete, then $E(g_1 - g_2|W) = 0 \Rightarrow g_1 = g_2$. So g is identified.

4.3 MLE and Efficiency

4.3.1 Likelihood

- The probability of observing what we have observed (intuitively)

$$L(\theta) = \prod f(X_i, \theta)$$

- Importantly, the true θ_0 maximizes expected log likelihood

$$E_{\theta_0}(\log f(X;\theta_0)) > E_{\theta_0}(\log f(X,\theta)), \forall \theta \neq 0$$

Proof: by Jenson,

$$E_{\theta_0}(\log f(X,\theta)) - E_{\theta_0}(\log f(X;\theta_0)) = E_{\theta_0} \log \frac{f(\theta)}{f(\theta_0)} \leq \log E_{\theta_0} \frac{f(\theta)}{f(\theta_0)} = \log \int \frac{f(\theta)}{f(\theta_0)} f(\theta_0) dx = 0$$

equality holds only if $\theta = \theta_0$

- This also implies

$$E_{\theta_0} \nabla_\theta \log f(\theta_0) = 0$$

4.3.2 MLE

let

$$l(\theta) = \frac{1}{n} \sum_i \log f(X_i, \theta) = \frac{1}{n} \sum_i g(X_i, \theta)$$

as the log likelihood

- MLE is defined as $\max l(\theta)$

$$\nabla l(\hat{\theta}) = \frac{1}{n}\sum_i \nabla \log f(X_i, \hat{\theta}) = 0$$

- If $\hat{\theta}$ is MLE, then $h(\hat{\theta})$ is the MLE of $h(\theta)$
- MLE is consistent
- Normality:

$$\sqrt{n}(\hat{\theta} - \theta_0) \to^d N(0, (Var(\nabla g(X, \theta_0)))^{-1})$$

  Proof:

$$0 = \nabla l(\hat{\theta}) = \nabla l(\theta_0) + \nabla^2 l(\bar{\theta})(\hat{\theta} - \theta_0)$$

  hence $\sqrt{n}(\hat{\theta}-\theta_0) = -\sqrt{n}\nabla^2 l(\bar{\theta})^{-1}\nabla l(\theta_0) = -\sqrt{n}\nabla^2 l(\bar{\theta})^{-1}\frac{1}{n}\sum_i \nabla g(X_i, \theta_0)$ Note that $E\nabla g(X_i, \theta_0) = E\nabla \log f(X, \theta_0) = 0$ so

$$-\sqrt{n}\frac{1}{n}\sum_i \nabla g(X_i, \theta_0) \to^d N(0, Var(\nabla g(X, \theta_0)))$$

  Also, if twice differentiable,

$$\nabla^2 l(\bar{\theta}) = \frac{1}{n}\sum_i \nabla^2 g(X_i, \bar{\theta}) \approx \frac{1}{n}\sum_i \nabla^2 g(X_i, \theta_0) \to^P E\nabla^2 g(X_i, \theta_0)$$

  Hence by slutsky,

$$\sqrt{n}(\hat{\theta}-\theta_0) \to^d N(0, \Sigma), \quad \Sigma = (E\nabla^2 g(X_i, \theta_0))^{-1} Var(\nabla g(X, \theta_0))(E\nabla^2 g(X_i, \theta_0))^{-1}$$

- We now simplify Σ:

$$Var(\nabla g(X, \theta_0)) = -E\nabla^2 g(X_i, \theta_0) := I(\theta_0) \quad \text{Information matrix}$$

  (find the proof of this identity on wikipedia) Given this,

$$\Sigma = (Var(\nabla g(X, \theta_0)))^{-1}$$

  The larger the information, the smaller variance.

4.3.3 MLE example of multiple choice model

$$U_{ij} = X_i^T \theta_j + e_{ij}, \quad i \leq n, \quad j \leq K$$

The i th person's random utility if chooses program j. But we do not observe U_{ij}. We only observe each person's choice.

Data: X_i and Y_{ij}, where $Y_{ij} = 1$(person i chooses program j)

Goal: The goal is to estimate the marginal effect of x_i^T on the utility: θ_j.

Modeling: assume e_{ij} is iid $\exp(-x - \exp(-x))$. Then

$$p_{ij} = P(Y_{ij} = 1|x_i) = P(\max(U_{i1}...U_{iK}) = U_{ij}|x_i) = \frac{\exp(x_i^T \theta_j)}{\sum_{k=1}^K \exp(x_i^T \theta_k)}.$$

Likelihood: assume X is discrete. Conditional on x_i, the probablity of person i's choice is:

$$\prod_{j=1}^K p_{ij}^{Y_{ij}}$$

example: K=3. This is $p(choose1)^{Y_{i1}} p(choose2)^{Y_{i2}} p(choose3)^{Y_{i3}}$ If choose 2, then $Y_{i1} = Y_{i3} = 0$, and $Y_{i2} = 1$. The above becomes $p(choose2)$.

Now unconditionally, it is

$$\prod_{j=1}^K p_{ij}^{Y_{ij}} P(X_i = x_i)$$

Assume independence, then the likelihood function is:

$$\prod_{i=1}^n \prod_{j=1}^K p_{ij}^{Y_{ij}} P(X_i = x_i)$$

Log-likelihood:

$$l(\theta_1...\theta_K) = \sum_i \sum_j Y_{ij} \log p_{ij} + K \sum_i P(X_i = x_i)$$

$$\propto \sum_i \sum_j Y_{ij} \log p_{ij} = \sum_i \sum_j Y_{ij} \log(\frac{\exp(x_i^T \theta_j)}{\sum_{k=1}^K \exp(x_i^T \theta_k)})$$

$$= \sum_i \sum_j Y_{ij}[X_i^T \theta_j - \log \sum_{k=1}^{K} \exp(X_i^T \theta_k)]$$

4.3.4 Efficiency

- Given the information we have, how well can we do?
- biaseness: $ET - g(\theta)$. unbiasedness means bias=0.

  $$MSE = bias^2 + Var$$

  Among unbiased estimators, smaller variance wins.
- Many estimators are asymptotically unbiased, meaning that

  $$\sqrt{n}(\hat{\theta} - \theta) \to^d N(0, V)$$

  V is some variance. So the $var(\hat{\theta}) \approx \frac{1}{n}V$. The goal is to find the smallest V.
- Cramer-Rao: Suppose X1...Xn are iid from $f(x, \theta_0)$. Suppose we want to estimate $g(\theta_0)$ using an unbiased estimator T. Then

  $$Var(T) \geq \frac{g'(\theta_0)^2}{nI(\theta_0)}$$

  Heuristic proof: Let $\mathbf{x} = (x_1...x_n)$ Note that

  $$\frac{d \log \prod_i f(x_i, \theta)}{d\theta} = \frac{\frac{d}{d\theta} \prod_i f(x_i, \theta)}{\prod_i f(x_i, \theta)}$$

  This means

  $$\frac{d}{d\theta} \prod_i f(x_i, \theta) = \underbrace{\frac{d \log \prod_i f(x_i, \theta)}{d\theta}}_{Z(\theta)} \times \prod_i f(x_i, \theta)$$

  Now integrate

  $$0 = \frac{d}{d\theta} \int \prod_i f(x_i, \theta) = \frac{d}{d\theta} 1 = \underbrace{\int \frac{d}{d\theta} \prod_i f(x_i, \theta) = \int Z(\theta) \times \prod_i f(x_i, \theta)}_{above} = E_\theta Z(\theta)$$

Variance:

$$Var(Z(\theta)) = Var(\frac{d\log\prod_i f(x_i,\theta)}{d\theta}) = Var(\sum_i \frac{d\log f(x_i,\theta)}{d\theta}) = nVar(\frac{d\log f(x_i,\theta)}{d\theta})$$

$$= nI(\theta)$$

So

$$g'(\theta) = \frac{d}{d\theta}E_\theta T = \int T(\mathbf{x})\frac{d}{d\theta}\prod_i f(x_i,\theta)d\mathbf{x} = \int T(\mathbf{x})Z(\theta)\prod_i f(x_i,\theta)d\mathbf{x} = E_\theta TZ(\theta)$$

$$= E_\theta TE_\theta Z(\theta) + Cov_\theta(T, Z(\theta)) = Cov_\theta(T, Z(\theta))$$

So

$$|g'(\theta)|^2 \leq Var(T)Var(Z(\theta)) = Var(T)nI(\theta)$$

- if to estimate θ, then $g' = 1$. So the best variance is $(nI)^{-1}$. This is the MLE's variance. So MLE is the best.
- An unbiased estimator is efficient if its variance is Cramer-Rao lower bound. It is asymptotically efficient if asymptotic variance is C-R bound.
- Efficient estimators are often useful to achieve smaller confidence intervals.
- In econometrics, however, efficiency is not always easy to get, because likelihood functions are not often easy to get. Instead, we only rely on some moment conditions or regression models. In these cases, **semi-parametric efficiency** is useful, meaning that it's the smallest variance for estimators (under a class of distributions satisifying the given moment conditions).

5 Computations

5.1 Newton Raphson

- Given a log-likelihood function $l(\theta)$, how do we maximize it ? In smooth cases, numerically maximizing l is the same as numerically solving $\nabla l = 0$.

  Newton-Raphson: $\nabla l(\hat{\theta}) = 0$. Then $0 \approx \nabla l(\theta_0) + \nabla^2(\theta_0)(\hat{\theta} - \theta_0)$.

  $$\hat{\theta} = \theta_0 - \nabla^2 l(\theta_0)^{-1}\nabla l(\theta_0)$$

We can start by replacing θ_0 with an initial estimate.

Newton is very sensitive to initial.

Example 5.1 (probit model on binary choice)**.** A person decides to work if the utility is positive.

$$y_i = \begin{cases} 1 & x_i^T\beta - \epsilon_i > 0 \\ 0 & x_i^T\beta - \epsilon_i < 0 \end{cases}$$

Then $P(y_i = 1|x_i) = P(\epsilon_i < x_i^T\beta|x_i) = \Phi(x_i^T\beta)$. In this model, the goal is β.

$P(Y = y|x) = \Phi(x^T\beta)^y(1 - \Phi(x^T\beta))^{1-y}$. Full likelihood:

$$\prod_i P(Y_i, X_i) = \prod P(Y_i = y_i|x_i)f(x_i) \propto \prod P(Y_i = y_i|x_i) = \prod_i \Phi(x_i^T\beta)^{y_i}(1-\Phi(x_i^T\beta))^{1-y_i}$$

log-likelihood:

$$l(\beta) = \sum_i y_i \log\Phi(x_i^T\beta) + \sum_i (1-y_i)\log(1-\Phi(x_i^T\beta))$$

$$\nabla l(\beta) = \sum_i y_i \frac{\phi(x_i^T\beta)}{\Phi(x_i^T\beta)} X_i + \sum_i (1-y_i)\frac{-\phi(x_i^T\beta)}{1-\Phi(x_i^T\beta)} X_i$$

$$\nabla^2(\beta) = -\sum_i y_i X_i \frac{x_i^T\beta\phi\Phi + \phi^2}{\Phi^2} X_i^T + \sum_i (1-y_i)\frac{x_i^T\beta\phi(1-\Phi) - \phi^2}{(1-\Phi)^2} X_i X_i^T$$

Numerical example

n=5. $(X, Y) = (3,1)(2,1)(3.3,1)(1.5,0)(1.2,0)$ Suppose no intercept. matlab code:

runprob(1,5). Here 1 is initial. 5 is number of iterations.

example

logisitic:

$$\nabla Q = -\sum_i y_i x_i + \sum_i z_i x_i, \quad z_i = \frac{\exp(x_i\beta)}{1+\exp(x_i\beta)}$$

$$\nabla^2 Q = \sum_i x_i x_i^T z_i(1-z_i)$$

5.2 Coordinate descent

$$\min L(\theta_1 ... \theta_p)$$

Sometimes F is not directly differentiable. We coordinately min it

$$\min_{\theta_k} L(\theta_{-k}, \theta_k), \quad k = 1...p$$

Convergence
Suppose we can write

$$L(\theta_1 ... \theta_p) = F(\theta_1 ... \theta_p) + \sum_j h_j(\beta_j)$$

F is differentiable and convex
h part is separable as above
Then guaranteed to converge to the global minimizer

5.3 Gradient Descent

5.3.1 unconstraint gradient descent

$$\min L(\beta)$$

Note that the loss function

$$L(\beta) \approx L(\beta^t) + \nabla L(\beta^t)(\beta - \beta^t) + \frac{1}{2}(\beta^t - \beta)^T \nabla^2 L(h^*)(\beta^t - \beta)$$

- recall N-R iteration: use β^t in place of h^*, leading to:

  $$\beta^{t+1} = \beta^t - \nabla^2 L(\beta^t)^{-1} \nabla L(\beta^t)$$

- The gradient descent update: replace $\nabla^2 L(h^*)$ by either $1/s^t I$ or D_t, where D_t is a diagonal matrix.

  Then
  $$\beta^{t+1} = \beta^t - s^t \nabla L(\beta^t)$$

  or

$$\beta^{t+1} = \beta^t - D_t^{-1}\nabla L(\beta^t)$$

5.3.2 majorization and convergence

- Let
$$F_t(\beta) = L(\beta^t) + \nabla L(\beta^t)(\beta - \beta^t) + \frac{1}{2s_t}(\beta^t - \beta)^T(\beta^t - \beta)$$
suppose s_t is such that $F_t(\beta) \geq L(\beta)$. Also $F_t(\beta^t) = L(\beta^t)$

  then
$$L(\beta^{t+1}) \quad \underbrace{\leq}_{majorization} \quad F_t(\beta^{t+1}) \underbrace{\leq}_{\min F_t} F_t(\beta^t) = L(\beta^t)$$

  So always decreases

- Convergence analysis

  Proof. For convex function $L(\beta)$, and Also by KKT, we have three conditions:
$$(1) L(\beta) \geq L(\beta^t) + \nabla L(\beta^t)^T(\beta - \beta^t)$$
$$(2) - L(\beta^{t+1}) \geq -F_t(\beta^{t+1})$$
$$(3) \beta^{t+1} = \beta^t - s^t \nabla L(\beta^t)$$

  (1)+(2)
$$L(\beta) - L(\beta^{t+1}) \geq \underbrace{L(\beta^t) + \nabla L(\beta^t)^T(\beta - \beta^t) - F_t(\beta^{t+1})}_{A}$$

  we calculate A using (3)
$$\begin{aligned} A &= \frac{1}{s}(\beta^t - \beta^{t+1})'(\beta - \beta^{t+1}) - \frac{1}{2s}\|\beta^{t+1} - \beta^t\|^2 \\ &= \frac{1}{2s}\|\beta^{t+1} - \beta\|^2 - \frac{1}{2s}\|\beta^t - \beta\|^2 \end{aligned}$$

So

$$L(\beta) - L(\beta^{T+1}) \geq L(\beta) - L(\beta^{t+1}) \geq \frac{1}{2s}\|\beta^{t+1} - \beta\|^2 - \frac{1}{2s}\|\beta^t - \beta\|^2$$

sum up from $t = 0, 1, ..., T$

$$(T+1)(L(\beta) - L(\beta^{T+1})) \geq \frac{1}{2s}\|\beta^{T+1} - \beta\|^2 - \frac{1}{2s}\|\beta^0 - \beta\|^2 \geq -\frac{1}{2s}\|\beta^0 - \beta\|^2$$

implies, for any β

$$L(\beta^{T+1}) \leq L(\beta) + \frac{1}{2s(T+1)}\|\beta^0 - \beta\|^2$$

now take β as the global minimum.

5.3.3 Proximal gradient descent

$$\min L(\beta) + h(\beta)$$

where h is convex but nondifferentiable

- approx

$$\min L(\beta^t) + \nabla L(\beta^t)(\beta^t - \beta) + \frac{1}{2s^t}(\beta^t - \beta)^T(\beta^t - \beta) + h(\beta)$$

which is

$$\min \frac{1}{2}\|\beta^t - s_t \nabla L(\beta^t) - \beta\|^2 + h(\beta)$$

hopefully this problem is easier to solve

5.4 lasso as an example

In modern statistical applications, Lasso is useful when regressors are many

$$Y = X\beta + e.$$

Then

$$\min \frac{1}{n}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

- coordinate descent

  Coordinately, each becomes a quadratic + L1 problem.

  $$\frac{1}{n}\sum_i (y_i - x_{i1}\theta - X_{i,-1}^T\beta_{-1})^2 = \frac{1}{n}\sum_i (z_i - x_{i1}\theta)^2 = (h-\theta)^2 + constant$$

  now the problem becomes

  $$\min_\theta \frac{1}{2}(h-\theta)^2 + \lambda|\theta|$$

  The solution is coordinate soft-thresholding.

- proximal gradient descent

  $$\min \frac{1}{n}\|\beta^t + s_t X'(T - X\beta^t) - \beta\|_2^2 + \lambda\|\beta\|_1$$

  The solution is vector soft-thresholding.

6 Statistical Inferences

6.1 Hypothesis Tests

6.1.1 Introduction

- Suppose the distribution of a random variable X depends θ. Suppose we think either $\theta \in \Theta_0$ or $\theta \in \Theta_1$. This is a partition of the parameter space. We usually label these hypotheses as

  $$H_0 : \theta \in \Theta_1, \quad H_1 : \theta \in \Theta_1.$$

  H0 is referred to as the null hypothesis, usually refers to a general statement or default position that there is no relationship between two measured phenomena, or no difference among groups.

  H1 is referred to as the alternative hypothesis, often referred to as the maintained hypothesis or the research hypothesis.

- Rejecting or disproving the null hypothesisand thus concluds that there are grounds for believing that there is a relationship between two phenomena (e.g.

that a potential treatment has a measurable effect)

- The decision rule of rejecting H0 or not is based on a sample X1... Xn.

  A test is to develop a decision rule, which partition the sampling space into a rejection region (critical region) and an acceptance region.

  $$\text{Reject } H_0 \Leftrightarrow (X_1...X_n) \in C$$

  Example: $C = \{\mathbf{x} : \bar{X} > 3\}$. The decision is $1\{\mathbf{x} \in C\}$

- Decision could be wrong, because data are random.

  Type I error is the incorrect rejection of a true null hypothesis "false positive"

  Type II error is the failure to reject a false null hypothesis " false negative"

- The ideal goal if to select a critical region which minimizes the probabilities of both Type I and Type II errors. However, this is not possible in general. For example,

  $$H_0 : \theta = 1, H_a : \theta > 1$$

  rule 1: reject if $\bar{X} > 3$

  rule 2: reject if $\bar{X} > 4$.

  Rule 2 rejects less often because its rejection region is smaller: $\{\mathbf{x} : \bar{X} > 4\} \subset \{\mathbf{x} : \bar{X} > 3\}$: reject under Rule 2 must lead to reject under Rule 1.

  Reject less often yields smaller type I, but larger Type II.

- Often, we consider Type I error to be the worse of these two types of errors. Then a solution is to

  (i) First make sure Type I error is well controlled

  (ii) Then, among those with controlled type I error, choose a test that minimizes the Type II error

- **Size** of a test: α:

  $$\alpha = \sup_{\theta \in \Theta_0} P_\theta(\mathbf{x} \in C) = P(reject|H_0) = P(TypeI)$$

  Power function

  $$\forall \theta \in \Theta_a : \quad \beta(\theta) = P_\theta(\mathbf{x} \in C) = P(reject|H_1) = 1 - P(typeII)$$

consistency: A test of size α is consistent if $P(reject|H_0) \to 1$.

6.1.2 Examples of Econometric Tests

Testing mean-variance efficiency

$$R_{it} = \alpha_i + \beta_i f_t + e_{it}$$

Then $ER_{it} = \alpha_i + \beta_i Ef_t$ Here β_i measures the sensitivity to the market. Note that β_i represents the risk of the investment, because

$$Var(R_{it}) = \beta_i^2 Var(f_t) + Var(e_{it})$$

Under the "mean-variance efficiency", no risk no return. Thus $ER_{it} = \beta_i Ef_t$. It means $\alpha_i = 0$.

$$H_0 : \alpha_i = 0 \forall i$$

Shape restriction test

$$earning = g(edu) + error$$

$$H_0 : g(.) monotone$$

Significance

$$Y = a + bX + Z^T\theta + error$$

Z is control; X is treatment.

$$H_0 : b = 0$$

Parametric tests

$$Y = g(X) + error$$

$$H_0 : g(x) = x^T\theta \text{ for some} \theta \in \Theta$$

Exogeneity

$$Y = x^T\theta + \epsilon$$

Here x is endogenous. We have an IV Z, and would like to test the validity of the IV:

$$H_0 : E(\epsilon Z) = 0$$

Under H_0, θ is identified if $dim(Z) \geq \dim(\theta) - 1$:

$$EYZ = EZx^T\theta, EY = Ex^T\theta.$$

number of effective equations is enough.

Heterogeneity test

$$GDPgrowthRate_{it} = \beta_i baseGDP_{it} + control + error$$

β_i: effect of base GDP.

$$H_0 : \beta_1 = ... = \beta_N$$

6.1.3 Test of normal mean

X1....Xn are iid $N(\theta, \sigma^2)$. For simplicity we assume σ^2 known.

$$H_0 : \theta = \theta_0 \quad H_a : \theta \neq 0$$

- We reject if $|\bar{X} - \theta_0| > m$.

$$\alpha = P_{\theta_0}(|\bar{X} - \theta_0| > m) = P_{\theta_0}(\frac{|\bar{X} - \theta_0|}{\sigma/\sqrt{n}} > \frac{m}{\sigma/\sqrt{n}}) = P(|Z| > \sqrt{n}m/\sigma)$$

  So $m = z_{\alpha/2}\sigma/\sqrt{n}$.

- Power: For any θ, let $\delta = \theta_0 - \theta$.

$$\beta(\theta) = P_\theta(|\bar{X} - \theta_0| > m) = P_\theta(\bar{X} > m + \theta_0, or \bar{X} < -m + \theta_0)$$

$$= P_\theta(\frac{\bar{X} - \theta}{\sigma/\sqrt{n}} > \frac{m + \theta_0 - \theta}{\sigma/\sqrt{n}}, or \frac{\bar{X} - \theta}{\sigma/\sqrt{n}} < \frac{-m + \theta_0 - \theta}{\sigma/\sqrt{n}})$$

$$= P(Z > \frac{m + \delta}{\sigma/\sqrt{n}}, or Z < \frac{-m + \delta}{\sigma/\sqrt{n}}) = P(Z > \frac{z\sigma + \sqrt{n}\delta}{\sigma}, or Z < \frac{-z\sigma + \sqrt{n}\delta}{\sigma})$$

$$= P(Z > z_{\alpha/2} + \sqrt{n}\delta\sigma^{-1}) + P(Z < -z_{\alpha/2} + \sqrt{n}\delta/\sigma)$$

  Now let $\delta = \delta_n$. Suppose $\sqrt{n}|\delta_n| \to \infty$.

  if $\delta > 0$, then

$$\beta(\theta) \to 0 + P(Z < +\infty) = 1$$

  if $\delta < 0$, then

$$\beta(\theta) \to P(Z > -\infty) + 0 = 1$$

  Hence the test is consistent against a fixed alternative, and more generally, against alternatives as long as the "gap" is less than $\sqrt{n}$.

- Local alternative. If $\sqrt{n}|\delta_n| = o(1)$, then

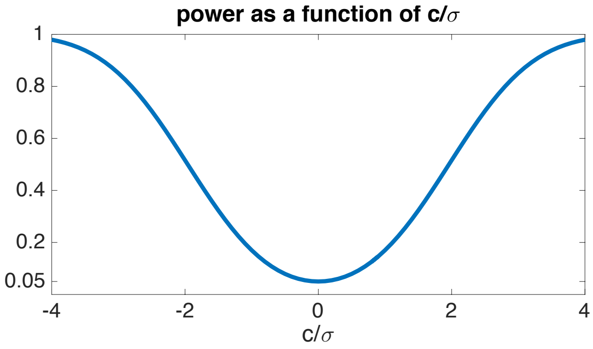
$$\beta(\theta) \to P(Z > z_{\alpha/2}) + P(Z < -z_{\alpha/2}) = \alpha.$$

  Hence the reject probability is as low as size even if $\delta_n \neq 0$. Too close to null to detect.

- If $\theta = \theta_0 - \frac{c}{\sqrt{n}}$, "Pitman drift". Then $\sqrt{n}\delta_n = c$.

$$\beta(\theta) = P(Z > z_{\alpha/2} + c\sigma^{-1}) + P(Z < -z_{\alpha/2} + c/\sigma) \in (\alpha, 1)$$

  So to compare with different tests, we sometimes look for tests that give the best power under Pitman alternatives.



Matlab code: c=[-4:0.1:4]; beta=1-normcdf(1.96+c)+normcdf(-1.96+c); plot(c,beta);

- When σ is unknown but its consistent estimator is available, then it is still the same , due to the Slutsky's lemma.

6.1.4 LRT

A useful test is likelihood ratio test.

$$LR = \frac{\sup_{\Theta_0} L(\theta, \mathbf{x})}{\sup_{\Theta} L(\theta, \mathbf{x})} \leq 1.$$

Reject if LR is small. So critical region:

$$C = \{\mathbf{x} : LR > c\} \text{ for some } c : \quad P_{\Theta_0}(LR > c) = \alpha.$$

- Intuitively, if LR is close to one, that means $L(\theta)$ is approximately maximized in Θ_0. Also, by regularity conditions, L should be maximized at the truth$\approx$ MLE, so this means the truth should be nearly Θ_0. Hence large LR shows no evidence against Θ_0.
- Use $\sup_\Theta$ or $\sup_{\Theta_a}$ in the denominator?
  - equivalent if MLE is obtained in Θ_a. For instance, $H_0 : \theta = 0; H_a : \theta > 0$ for normal mean test. And $\bar{X} > 0$.
  - In general, using $\sup_\Theta$ is easier to get distributions if MLE is not obtained in Θ_a.
  - Example: $X \sim N(\theta, 1)$. $H_0 : \theta = \theta_1$ vs $H_a : \theta = \theta_2$. Here $\bar{X}$ can be neither.

    if use $\sup_\Theta$:

    $$LR = \frac{\exp(-\sum_i (X_i - \theta_1)^2/2)}{\exp(-\sum_i (X_i - \bar{X})^2/2)} = \exp(-\frac{1}{2}\sum_i (X_i - \theta_1)^2 + \frac{1}{2}\sum_i (X_i - \bar{X})^2)$$

    So Reject if $-\frac{1}{n}\sum_i (X_i - \theta_1)^2 + \frac{1}{n}\sum_i (X_i - \bar{X})^2$ is small. Note that $\frac{1}{n}\sum_i (X_i - \theta_1)^2 = \frac{1}{n}\sum_i (X_i - \bar{X})^2 + (\bar{X} - \theta_1)^2$. So reject if

    $$(\bar{X} - \theta_1)^2 > c^2, \quad c = z_{\alpha/2}/\sqrt{n}$$

    power:

    $$\beta(\theta_2) = P_{\theta_2}(|\bar{X} - \theta_1| > c) = P_{\theta_2}(Z > z_{\alpha/2} - \sqrt{n}\delta) + P(Z < -\sqrt{n}\delta - z_{\alpha/2})$$

if use $\sup_{\Theta_a}$: suppose $\theta_2 - \theta_1 = \delta > 0$.

$$LR = \frac{\exp(-\sum_i (X_i - \theta_1)^2/2)}{\exp(-\sum_i (X_i - \theta_2)^2/2)} = \exp(-\frac{1}{2}\sum_i (X_i - \theta_1)^2 + \frac{1}{2}\sum_i (X_i - \theta_2)^2)$$

$$-\frac{1}{n}\sum_i (X_i - \theta_1)^2 + \frac{1}{n}\sum_i (X_i - \theta_2)^2 = (\bar{X} - \theta_2)^2 - (\bar{X} - \theta_1)^2 = -2\delta\bar{X} + C$$

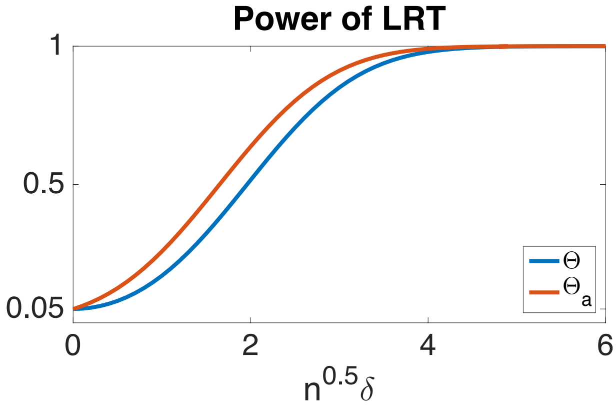
because $\delta > 0$, reject if

$$\bar{X} - \theta_1 > c \quad , c = z_\alpha/\sqrt{n}$$

power:

$$\beta(\theta_2) = P_{\theta_2}(\bar{X} - \theta_1 - \theta_2 > c - \theta_2) = P(Z > z_\alpha - \sqrt{n}\delta)$$

which power is larger ? Plot both powers as function of $\sqrt{n}\delta > 0$. Neyman-Pearson lemma shows that $\sup_{\Theta_a}$ is the most powerful test.



Matlab code: d=[0:0.1:6]; beta1=1-normcdf(1.96-d)+normcdf(-1.96-d); beta2=1-normcdf(1.65-d); plot(d,beta1, d, beta2, 'LineWidth', 4); legend('Theta', 'Theta a')

- If $\sup_\Theta L(\theta, \mathbf{x}) = L(MLE, \mathbf{x})$ and $\Theta_0 = \{\theta_0\}$ then

  $$-2\log LR|H_0 \to^d \chi^2_{dim(\theta)}$$

  Why ?

  $$2\log LR = 2l(\theta_0) - 2l(\hat{\theta})$$

  Note that $\nabla l(\hat{\theta}) = 0$. so (note that $l = \sum_i$)

  $$l(\theta_0) = l(\hat{\theta}) + \frac{1}{2}(\theta_0 - \hat{\theta})' \nabla^2 l(\bar{\theta})(\theta_0 - \hat{\theta})$$

  so

  $$2\log LR = (\theta_0 - \hat{\theta})' \nabla^2 l(\bar{\theta})(\theta_0 - \hat{\theta})$$

  Recall that

  $$\sqrt{n}(\hat{\theta} - \theta_0) \to^d N(0, I^{-1}), \quad I = Var(\nabla l_1(\theta_0)) = -E\nabla^2 l_1(\theta_0)$$

  Also, $\frac{1}{n}\nabla^2 l(\bar{\theta}) \approx \frac{1}{n}\nabla^2 l(\theta_0) \to^P E\nabla^2 l_1(\theta_0) = -I$ So

  $$-2\log LR \approx (\theta_0 - \hat{\theta})' I (\theta_0 - \hat{\theta}) n = \|\sqrt{n}(\hat{\theta} - \theta_0)' I^{1/2}\|_2^2$$

  $$= \|\sqrt{n}(\hat{\theta} - \theta_0)' Var^{-1/2}\|_2^2 \to^d \|Z(0, Identity_p)\|_2^2 = \chi^2_p$$

6.1.5 Neyman-Pearson Lemma

- Consider a simple test: $H_0 : \theta = \theta_0$, $H_a : \theta = \theta_1$. Let $\mathcal{C}$ be a class of tests such that

  $$P_{\theta_0}(\delta = 1) = \alpha. \tag{6.1}$$

- A test $\tilde{\delta} \in \mathcal{C}$ is called Uniformly Most Powerful test in class $\mathcal{C}$, if for any $\tilde{\delta} \in \mathcal{C}$,

  $$P_{\theta_1}(\delta = 1) \geq P_{\theta_1}(\tilde{\delta} = 1). \tag{6.2}$$

- N-P Lemma:

  $$\delta = 1\{f(\mathbf{x}; \theta_1) > kf(\mathbf{x}; \theta_0)\} \tag{6.3}$$

  If $k \geq 0$ is such that $\delta \in \mathcal{C}$, then δ is UMP in $\mathcal{C}$. Converse, ant UMP test should satisfy (6.3).

- The proof is on my Apple notes. Also see using this proof method to prove the optimality of Bayes' rule.
- **Summary**
  1. If
  $$H_0 : \theta = \theta_0 \quad H_a : \theta \in \Theta_a$$
  then
  $$\frac{f(\theta_0)}{\sup_{\Theta_a} f(\theta)}, \quad \frac{f(\theta_0)}{\sup_{\Theta} f(\theta)}$$
  are both LR test, as long as $\sup_{\Theta_a} f(\theta) = \sup_{\Theta} f(\theta) = f(\hat{\theta}_{MLE})$. And LR test is χ^2
  2. If $\sup_{\Theta_a} f(\theta) \neq f(\hat{\theta}_{MLE})$, then
  $$\frac{f(\theta_0)}{\sup_{\Theta_a} f(\theta)}$$
  is NOT χ^2, and not LR test.
  3. If $\Theta_a = \{\theta_1\}$ a singleton, then
  $$\frac{f(\theta_0)}{\sup_{\Theta_a} f(\theta)}$$
  is NOT χ^2, not LR test. But it has the largest power. In contrast, LR test does not have largest power
  4. If $H_a : \theta \neq \theta_0$, then it seems difficult to show LR test has the largest power either. In stead, using a similar techique, I can show a simpler statement: Suppose the set $\{\mathbf{x} : LR(\mathbf{x})\text{test rejects}\}$ has a finite Lebesgue measure. Then for any test δ, for any $\epsilon > 0$, there is $\theta(\epsilon) \neq \theta_0$, so that
  $$P_{\theta(\epsilon)}(\delta = 1) \leq P_{\theta(\epsilon)}(LR(\mathbf{x})\text{test rejects}) + \epsilon$$

6.1.6 Wald (test of Exogeneity)

- First estimate θ using $\hat{\theta}$, obtain its variance. Suppose
$$V^{-1/2}\sqrt{n}(\hat{\theta} - \theta) \to^d N(0, Id_p)$$
here Id means identity matrix.

Then $\|V^{-1/2}\sqrt{n}(\hat{\theta}-\theta)\|_2^2 \to^d \chi_p^2$, by Continuous mapping theorem. The statistic is thus

$$n(\hat{\theta}-\theta_0)'V^{-1}(\hat{\theta}-\theta_0) \sim \chi_p^2$$

- If $\hat{\theta}$ is MLE, then

  $$V = I^{-1} = Var(\nabla l_1(\theta_0))$$

  Recall that -2 log LR $\approx \|\sqrt{n}(\hat{\theta}-\theta_0)'I^{1/2}\|_2^2$

  Hence Wald and LRT are asymptotically equivalent if $\hat{\theta}$ =MLE. So

  $$n(\hat{\theta}-\theta_0)^T I(\theta)(\hat{\theta}-\theta_0) \to^d \chi_p^2$$

  reject if

  $$n(\hat{\theta}-\theta_0)^T \hat{I}(\theta)(\hat{\theta}-\theta_0) > \chi_{p,0.05}^2$$

  where $\chi_{p,0.05}^2$ is the right tail percentile. You should use an estimated $I(\theta)$ as:

  $$\hat{I} = -\hat{E}\nabla^2 l_1(\theta_0) = -\frac{1}{n}\nabla^2 l(\hat{\theta}) := -\frac{1}{n}\sum_i \nabla^2 l(X_i, \hat{\theta})$$

- The good thing about Wald is that it is still available if the likelihood is not available.

  Example: test $EX = \theta = \theta_0$ or not. Without knowing the likelihood, we do not know whether $\hat{\theta} = \bar{X}$ is the MLE or not. Nevertheless, under H0,

  $$\sqrt{n}(\bar{X}-\theta_0) \to^d N(0,\sigma^2)$$

  So wald test is

  $$n(\bar{X}-\theta_0)^2/\sigma^2 \to^d \chi_1^2$$

  which is the same as Z test because it is one dimensional. We can estimate σ^2 by its sample variance.

  Power of Wald test

- Suppose θ_1 is the truth. MLE always estimates truth regardless of null true or false, so no matter null or alternative,

  $$Z := \sqrt{n}I(\hat{\theta})^{1/2}(\hat{\theta}-\theta_1) \to^d N(0, Id)$$

So

$$\frac{1}{\sqrt{n}}I(\hat{\theta})^{-1/2}Z+\theta_1=\hat{\theta}$$

Let $\delta=\theta_1-\theta_0$. Wald is

$$n(\hat{\theta}-\theta_0)^TI(\hat{\theta})(\hat{\theta}-\theta_0)=n(\frac{1}{\sqrt{n}}I(\hat{\theta})^{-1/2}Z+\theta_1-\theta_0)^TI(\hat{\theta})(\frac{1}{\sqrt{n}}I(\hat{\theta})^{-1/2}Z+\theta_1-\theta_0)$$

$$=(I(\hat{\theta})^{-1/2}Z+\sqrt{n}\delta)^TI(\hat{\theta})(\frac{1}{\sqrt{n}}I(\hat{\theta})^{-1/2}Z+\sqrt{n}\delta)$$

Let

$$I(\hat{\theta})^{1/2}(I(\hat{\theta})^{-1/2}Z+\sqrt{n}\delta)=W\approx^d N(\sqrt{n}I(\hat{\theta})^{1/2}\delta,Id)$$

So

$$Wald=W^TW\approx^d\chi_p^2(\lambda)$$

The noncentral chi square is defined as

$$\chi_2^p(\lambda)=\sum_{i=1}^p iidN(\mu_i,1),\quad\lambda=\sum_{i=1}^p\mu_i^2$$

In this context, $\lambda=\|\sqrt{n}I(\hat{\theta})^{1/2}\delta\|_2^2=n\delta^TI(\hat{\theta})\delta$.

- So we have proved: no matter null or alternataive,

$$Wald\to^d\chi_p^2(\lambda)$$

Under the null, $\delta=0$. The reject is

$$\beta(\theta)=P(\chi^2(\lambda)>\chi_{p,\alpha}^2).$$

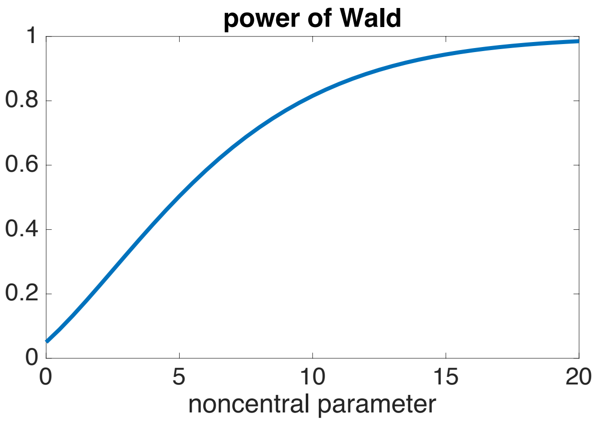
If $n\|\delta\|_2^2\to0$, accept the null; If $n\|\delta\|_2^2\to\infty$, power converges to one.

- More general estimator:

$$\beta(\theta)=P(\chi^2(\lambda)>\chi_{p,\alpha}^2).\quad\lambda=\delta^T\text{var}(\hat{\theta})^{-1}\delta$$

–Intuitively, the larger λ, the larger $\chi^2(\lambda)$, hence the larger power. This shows:

– for two estimators to construct two Wald tests, **if one has smaller variance, then its λ is larger, then its power is larger.**



Matlab code: p=2; d=[0:0.5:20]; cr=chi2inv(0.95,p); beta = ncx2cdf(cr,p,d,'upper'); plot(d,beta,'LineWidth',4);

– Comparing the power of two Wald test is the same as comparing the estimators' variances.

–So we should try to find the smallest variance estimator to construct Wald. It turns out, MLE implies the Wald with the best power.

Example 6.1 (Test of exogeneity)**.** Suppose

$$y = x^T\beta + u$$

and we have a set of valid IV's z so that $Ezu = 0$. We want to test whether x is exogenous:

$$H_0 : Eux = 0$$

so here $\theta = Eux$. Endogeneity arises, for instance, there is omitted variables in u.

To use Wald test, we first estimate θ, which equals

$$\hat{\theta} = \frac{1}{n}X^T\hat{U} = \frac{1}{n}X^T(Y - X\hat{\beta})$$

How to estimate β ? You can use OLS, but it is inconsistent under the alternative.

inconsistency of OLS under endogeneity

$$\hat{\beta}_{ols} = (X^T X)^{-1} X^T Y \to^P E(xx^T)^{-1} Exy = E(xx^T)^{-1} Ex(x^T\beta + u) = \beta + E(xx^T)^{-1} Exu$$

IV estimator

Note that $Euz = 0 \Rightarrow Ez(y - x^T\beta) = 0 \Rightarrow Ezy = Ezx^T\beta$. Assume Ezx^T is full ranked square matrix, then $\beta = (Ezx^T)^{-1}Ezy$, so

$$\hat{\beta}_{IV} = (Z^T X)^{-1} Z^T Y$$

Wald test

$$\hat{\theta} = \frac{1}{n} X^T [Y - X(Z^T X)^{-1} Z^T Y] = \frac{1}{n} X^T [I - X(Z^T X)^{-1} Z^T] Y = \frac{1}{n} X^T (I - P_{XZ}) Y.$$

let $P_{XZ} := X(Z^T X)^{-1} Z^T$. We have

$$Wald = n(\hat{\theta} - \theta_0)' V^{-1} (\hat{\theta} - \theta_0) = n\hat{\theta}^T Var(\hat{\theta})^{-1} \hat{\theta} \to^d \chi^2_p$$

Hausman's test (Hausman 78), also known as "Durbin-Wu-Hausman test"

They consider a different perspective: $\hat{\beta}_{iv} - \hat{\beta}_{ols}$. Reject if

$$n(\hat{\beta}_{iv} - \hat{\beta}_{ols})' \tilde{V}^{-1} (\hat{\beta}_{iv} - \hat{\beta}_{ols}) > C$$

under the null, they are close. But under the alternative, $\hat{\beta}_{iv} - \hat{\beta}_{ols} \to^P E(xx^T)^{-1} Exu = E(xx^T)^{-1}\theta$, which is nonzero. So essentially this is the same as Wald test based on $\hat{\theta}$.

To show it,

$$\hat{\beta}_{iv} - \hat{\beta}_{ols} = [(Z^T X)^{-1} Z^T - (X^T X)^{-1} X^T] Y = (X^T X)^{-1} [X^T X (Z^T X)^{-1} Z^T - X^T] Y$$

$$= -(X^T X)^{-1} X^T (I - P_{XZ}) Y = -(\frac{1}{n} X^T X)^{-1} \hat{\theta}$$

So Hausman's test is

$$n(\hat{\beta}_{iv} - \hat{\beta}_{ols})' \tilde{V}^{-1} (\hat{\beta}_{iv} - \hat{\beta}_{ols}) = \hat{\theta}' (\frac{1}{n} X^T X)^{-1} \tilde{V}^{-1} (\frac{1}{n} X^T X)^{-1} \hat{\theta} n$$

To make sure this is chi squared one, the thing in the middle should be also $Var(\hat{\theta})^{-1}$. So Hausman's test is Wald test.

6.1.7 Score test

Under the null,

$$E\nabla l_1(\theta_0) = 0$$

Hence reject if $\|\frac{1}{n}\sum_i \nabla l_1(X_i, \theta_0)\|_V^2$ is large.

$$0 = \nabla l(\hat{\theta}) = \nabla l(\theta_0) + \nabla^2 l(\bar{\theta})(\hat{\theta} - \theta_0)$$

Hence

$$\nabla l(\theta_0) \approx \nabla^2 l(\theta_0)(\hat{\theta} - \theta_0)$$

So in fact, we reject when

$$\nabla l(\theta_0)' V \nabla l(\theta) \approx (\hat{\theta} - \theta_0)' \nabla^2 l(\theta_0) V \nabla^2 l(\theta_0)(\hat{\theta} - \theta_0) > C$$

For MLE, we know

$$n(\hat{\theta} - \theta_0)^T I(\theta_0)(\hat{\theta} - \theta_0) \to^d \chi_p^2$$

hence we let $\nabla^2 l(\theta_0) V \nabla^2 l(\theta_0) = nI(\theta_0) \approx -\nabla^2 l(\theta_0)$, which is

$$V = -(\nabla^2 l(\theta_0))^{-1} \approx -(nE\nabla^2 l_1(\theta_0))^{-1} = \frac{1}{n} I(\theta_0)^{-1}$$

So we have

$$-\nabla l(\theta_0)^T (\nabla^2 l(\theta_0))^{-1} \nabla l(\theta_0) \approx n(\hat{\theta} - \theta_0)^T I(\theta_0)(\hat{\theta} - \theta_0) | H_0 \to^d \chi_p^2$$

we reject if

$$-\nabla l(\theta_0)^T (\nabla^2 l(\theta_0))^{-1} \nabla l(\theta_0) = \nabla l(\theta_0)^T I(\theta_0)^{-1} \nabla l(\theta_0)/n > \chi_{p,\alpha}^2$$

Example 6.2. Score test does not require likelihoods. Consider

$$y = x^T\theta + e, \quad Eze = 0$$

dim(z)$> dim(x)$. Then $\hat{\theta} = \arg\min Q(\theta)$. The score is $\nabla Q(\hat{\theta})$. Here

$$Q = [\frac{1}{n}\sum_i z_i(y_i - x_i^T\theta)]^T [\frac{1}{n}\sum_i z_i(y_i - x_i^T\theta)]$$

- **In summary**

  LRT $-2\log LR \to^d \chi_p^2$

Wald $n(\hat{\theta}-\theta_0)^T I(\hat{\theta})(\hat{\theta}-\theta_0) \to^d \chi^2_p$

Score $\nabla l(\theta_0)^T \frac{1}{n} I(\hat{\theta})^{-1} \nabla l(\theta_0) \to^d \chi^2_p$

reject if the left is bigger than $\chi^2_{p,\alpha}$

They are all asymptotically equivalent. **Should use** $I(\hat{\theta})$ because it is consistent even under alternative. Usually estimated by $I(\hat{\theta}) \approx -\frac{1}{n}\nabla^2 l(\hat{\theta})$

- Simulation.

$$H_0 : \theta = \theta_0; \quad H_a : \theta \neq \theta_0$$

$$X \sim iid \theta^{-1} \exp(-\theta^{-1}x)$$

LRT: $l(\theta) = -n\log\theta - n\bar{X}/\theta$, MLE=$\bar{X}$. So

$$-2\log LR = -2(-n\log\theta - n\bar{X}/\theta + n\log\bar{X} + n) = -2n(\log\frac{\bar{X}}{\theta_0} - \frac{\bar{X}}{\theta_0} + 1)$$

reject if $> \chi^2_1(0.05) = 3.8415$.

Wald: $I(\hat{\theta}) \approx -\frac{1}{n}\nabla^2 l(\hat{\theta}) = \frac{2\bar{X}}{\theta^3} - \frac{1}{\theta^2} = (\bar{X})^{-2}$. So

$$Wald1 = n(\bar{X}-\theta_0)^2(\bar{X})^{-2} = n(1-\frac{\theta_0}{\bar{X}})^2$$

reject if > 3.841

Score: $\nabla l = -n\theta^{-1} + n\frac{\bar{X}}{\theta^2}$, So

$$score1 = (\frac{n\bar{X}}{\theta_0^2} - \frac{n}{\theta_0})^2 \frac{\bar{X}^2}{n} = n(1-\frac{\theta_0}{\bar{X}})^2(\frac{\bar{X}}{\theta_0})^4$$

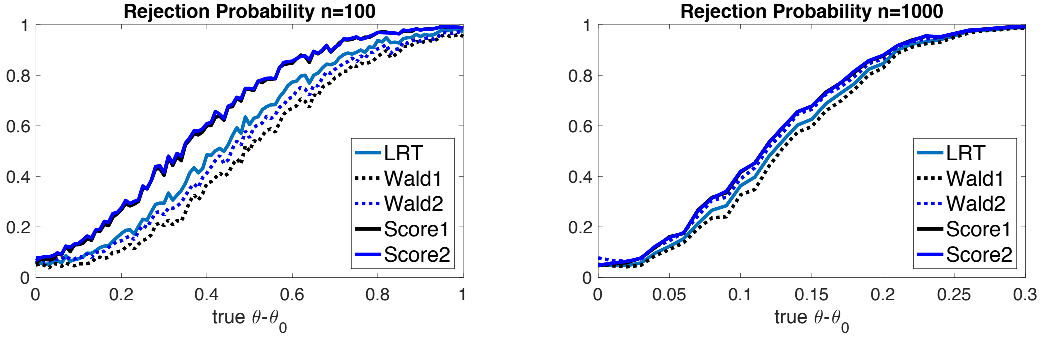
Here $delta = true\theta - \theta_0$. Also try n=10.

- You may also think about using $\hat{I} = \frac{1}{n}\sum_i(\nabla l_1(X_i,\hat{\theta}))^2$ as the estimated fisher information, which is $\hat{\sigma}^2/\bar{X}^4$, where $\hat{\sigma}^2 = \frac{1}{n}\sum_i(X_i-\bar{X})^2$, then

$$wald2 = n(\bar{X}-\theta_0)^2\hat{\sigma}^2/\bar{X}^4 = n(1-\frac{\theta_0}{\bar{X}})^2\frac{\hat{\sigma}^2}{\bar{X}^2}$$

$$score2 = (\frac{n\bar{X}}{\theta_0^2} - \frac{n}{\theta_0})^2\frac{\bar{X}^4}{n\hat{\sigma}^2} = n(1-\frac{\theta_0}{\bar{X}})^2(\frac{\bar{X}}{\theta_0})^4\frac{\bar{X}^2}{\hat{\sigma}^2}$$

In the simulation, $\theta_0 = 2$. 1000 replications.



Matlab code: n=100; delta=[0:0.01:1]; exptest(n,delta);

- Sometimes p may grow, then we use

$$\frac{\chi_p^2 - p}{\sqrt{2p}} \to N(0,1)$$

6.2 Confidence Regions

6.2.1 Introduction

$$P(\theta \in CR) = 1 - \alpha$$

Sometimes

$$P(\theta \in CR) \to 1 - \alpha$$

Sometimes use conservative CR

$$\lim P(\theta \in CR) \geq 1 - \alpha$$

In econometrics, usually

$$\sqrt{n}(\widehat{\theta} - \theta) \to^d N(0, \sigma^2)$$

then we get CR

$$\widehat{\theta} \pm z_{\alpha/2}\hat{\sigma}/\sqrt{n}$$

Figure 2: Proof of N-P Lemma, Optimality of Bayes rule, and general idea of optimality

App. of this idea : Pf of NP lemma $\quad \delta^* = 1_{f(\theta_1) \geq k f(\theta_0)}$

$\forall$ test δ. $P_{\theta_1}(\delta=1) = \int f(\theta_1) 1_{\delta=1} dx = \int_{f(\theta_1)\geq kf(\theta_0)} f(\theta_1) 1_{\delta=1} + \int_{f(\theta_1)<kf(\theta_0)} f(\theta_1) 1_{\delta=1}$

$= \cancel{\int_{f(\theta_1)\geq kf(\theta_0)} f(\theta_1) 1_{\delta=1,\delta^*=0}} + \int_{f(\theta_1)\geq kf(\theta_0)} f(\theta_1) 1_{\delta=1,\delta^*=1} + \int_{f(\theta_1)<kf(\theta_0)} f(\theta_1) 1_{\delta=1,\delta^*=1} + \int_{f(\theta_1)<kf(\theta_0)} f(\theta_1) 1_{\delta=1,\cancel{\delta^*=0}}$

$P_{\theta_1}(\delta^*=1) = \int_{f(\theta_1)\geq kf(\theta_0)} f(\theta_1) 1_{\delta^*=1,\delta=1} + \int_{f(\theta_1)\geq kf(\theta_0)} f(\theta_1) 1_{\cancel{\delta^*=1},\delta=0} + \int_{f(\theta_1)<kf(\theta_0)} f(\theta_1) 1_{\delta^*=1,\delta=1} + \cancel{\int_{f(\theta_1)<kf(\theta_0)} f(\theta_1) 1_{\delta^*=1,\delta=0}}$

$P_{\theta_1}(\delta=1) - P_{\theta_1}(\delta^*=1) = \int_{f(\theta_1)<kf(\theta_0)} f(\theta_1) 1_{\delta=1} - \int_{f(\theta_1)\geq kf(\theta_0)} f(\theta_1) 1_{\delta=0}$

$\leq k\int_{f(\theta_1)<kf(\theta_0)} f(\theta_0) 1_{\delta=1} - k\int_{f(\theta_1)\geq kf(\theta_0)} f(\theta_0) 1_{\delta=0}$

$= k\int f_0 1_{\delta=1} - \cancel{k\int_{f_1\geq kf_0} f_0 1_{\delta=1}} - k\int_{f_1\geq kf_0} f_0 + \cancel{k\int_{f_1\geq kf_0} f_0 1_{\delta=1}}$

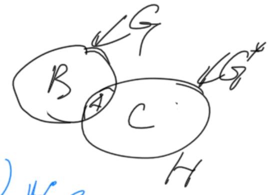
$= k\alpha - k\alpha = 0$

$$\boxed{\alpha = \int f_0 = \int_{f_1\geq kf_0} f_0 1_{\delta=1}}$$

main idea:

$\min_G F(G)$

To show $F(G) \geq F(G^*)$ $\forall G$:



$\forall G$, $F(G) = F_G(B) + F_G(A) + F_G(C) + F_G(H)$

$F(G^*) = F_{G^*}(B) + F_{G^*}(A) + F_{G^*}(C) + F_{G^*}(H)$

often ① Turns out: $F_G(A) = F_{G^*}(A)$ & $F_G(H) = F_{G^*}(H)$ $\quad \Rightarrow$

$\therefore F(G) - F(G^*) = F_G(B) - F_{G^*}(B) + F_G(C) - F_{G^*}(C)$

often ② $F_G(B) - F_{G^*}(B)$ can be combled to be $= M(\underset{=}{Y}) B$

for some y

$F_G(C) - F_{G^*}(C) = \tilde{M}(\tilde{y}) C$

Now use the definition & other knowns to show $M(Y)B + \hat{M}(\tilde{y})C \geq 0$

expamples of this proof procedure:

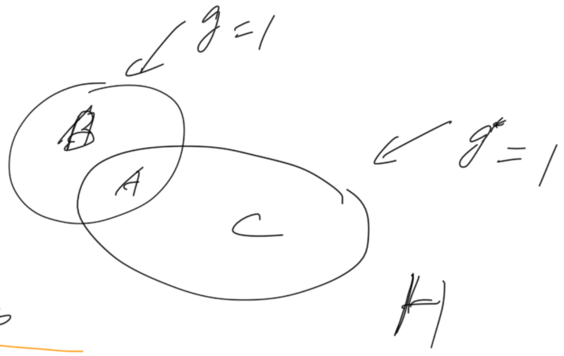
① Neyman-pearson lemma

② Optimality of Bayes rule

③ Optimality of Bayesian Set estimation (See "Bayesian copy.pdf" in my dropbox folder).

$$\therefore \min_g P(Y \neq g(x)) = g^* = \mathbb{1}\, P(Y=1|x) > 0.5$$

$$\because \forall g \qquad Y \neq g(x) =$$



$$P(Y \neq g|x) = P(Y=1|x)\mathbb{1}_{g=0} + P(Y=0|x)\mathbb{1}_{g=1}$$

$$= \underline{P(Y=1|x)\mathbb{1}_{g=0,\, g^*=1}} + \underline{P(Y=1|x)\mathbb{1}_{g=0,\, g^*=0}}$$

$$+ P(Y=0|x)\mathbb{1}_{g=1,\, g^*=0} + \underline{P(Y=0|x)\mathbb{1}_{g=1,\, g^*=1}}$$

$$P(Y \neq g^*|x) = P(Y=1|x)\mathbb{1}_{g^*=0,\, g=1} + \underline{P(Y=1|x)\mathbb{1}_{g^*=0,\, g=0}}$$

$$+ \underline{P(Y=0|x)\mathbb{1}_{g^*=1,\, g=1}} + \underline{P(Y=0|x)\mathbb{1}_{g^*=1,\, g=0}}$$

$$\underbrace{P(Y \neq g|x) - P(Y \neq g^*|x)}_{\geq 0} = \mathbb{1}_{g^*=1,\, g=0}\underbrace{[P(Y=1|x) - P(Y=0|x)]}_{>0} + \mathbb{1}_{g=1,\, g^*=0}\underbrace{[P(Y=0|x) - P(Y=1|x)]}_{>0}$$

6.2.2 shortest normal CI

Now we show this is the shortest CI. Suppose the interval is $[x, x+y]$. Then

$$\Phi(x+y) - \Phi(x) = 1 - \alpha$$

so length

$$y = f(x) = \Phi^{-1}(\Phi(x) + 1 - \alpha) - x$$

Use the formula

$$\frac{d}{dt}\Phi^{-1}(t) = \frac{1}{\Phi'(\Phi^{-1}(t))}$$

we have

$$\frac{d}{dx}\Phi^{-1}(\Phi(x)+1-\alpha) = \frac{d}{dt}\Phi^{-1}(t)|_{t=\Phi(x)+1-\alpha}\frac{d}{dx}\Phi(x) = \frac{\phi(x)}{\phi(\Phi^{-1}(t^*))}, \quad t^* = \Phi(x)+1-\alpha$$

$$f'(x) = \frac{\phi(x)}{\phi(m)} - 1, \quad m = \Phi^{-1}[\Phi(x) + 1 - \alpha]$$

setting $f' = 0$ yields $\phi(x) = \phi(m)$. So either $x = m$ or $x = -m$.

If $x = m$, then $\Phi^{-1}[\Phi(x) + 1 - \alpha] = x$. So $\Phi(x) + 1 - \alpha = \Phi(x)$, impossible when $\alpha < 1$.

So $x = -m$, and then $\Phi(-x) = \Phi(x) + 1 - \alpha$. Also note that $1 - \Phi(-x) = \Phi(x)$ for all x. Thus

$$1 - \Phi(x) = \Phi(x) + 1 - \alpha, \Rightarrow \Phi(x) = \alpha/2$$

Note: the proof only uses the facts that Φ is differentiable and ϕ is symmetric.

6.2.3 Duality of CR and Tests

Consider $H_0 : \theta = \theta_0$.

- Intuitively, for those θ_0 that are accepted, they should be closer to the truth. So

  $$C(\mathbf{x}) = \{\theta_0 : \textit{accepted}\}$$

  is a CR.

  Proof: Fix θ_0, let $A(\theta_0)$ be acceptance region, that is

  $$\theta_0 \in C(\mathbf{x}) \Leftrightarrow \mathbf{x} \in A(\theta_0)$$

By definition,

$$P_{\theta_0}\{\theta_0 \in C(\mathbf{x})\} = P_{\theta_0}\{\mathbf{x} \in A(\theta_0)\} = 1 - \alpha$$

- Conversely, Suppose we have a CR $C(\mathbf{x})$, then

  $$A(\theta_0) = \{\mathbf{x} : \theta_0 \in C(\mathbf{x})\}$$

  is the acceptance region.

  Proof:
  $$P_{\theta_0}(\mathbf{x} \notin A(\theta_0)) = P_{\theta_0}(\theta_0 \notin C(\mathbf{x})) = \alpha.$$

- Example: $N(\theta, \sigma^2)$.

  Method 1:

  Reject θ_0 if $|\bar{X} - \theta_0| > 1.96\sigma/\sqrt{n}$. So accepte if $\theta_0 \in \bar{X} \pm 1.96\sigma/\sqrt{n}$; the latter is its CR.
  $$[\bar{X} - 1.96\sigma/\sqrt{n}, \bar{X} + 1.96\sigma/\sqrt{n}]$$

  Method 2:

  You can also consider One sided test. Reject if $\bar{X} - \theta_0 > 1.65\sigma/\sqrt{n}$. So accept if $\theta_0 > \bar{X} - 1.65\sigma/\sqrt{n}$; the latter is its CR.

  $$(\bar{X} - 1.65\sigma/\sqrt{n}, +\infty)$$

  Of course Method 1 gives smaller CR.

- Usually inverting one-sided test gives one-sided interval; two-sided gives two-sided interval. But not always.

- **Inverting Wald test**

  $$accept \Leftrightarrow n(\hat{\theta} - \theta_0)^T V(\hat{\theta} - \theta_0) < \chi^2_{d,\alpha}$$

  So $P(\theta_0 : n(\hat{\theta} - \theta_0)^T V(\hat{\theta} - \theta_0) < \chi^2_{d,\alpha}) \to 1 - \alpha$. In the one dim case, it is

  $$n|\hat{\theta} - \theta_0|^2/\hat{\sigma}^2 < \chi^2_{1,\alpha}$$

  equivalent to $\sqrt{n}|\hat{\theta} - \theta_0|/\hat{\sigma} < z_{\alpha/2}$, $\theta_0 \in \hat{\theta} \pm \frac{z\hat{\sigma}}{\sqrt{n}}$

Inverting LR test

$$accept \Leftrightarrow -2\log LR(\theta_0) < \chi^2_{d,\alpha}$$

so CI is $\{\theta_0 : -2\log LR(\theta_0) < \chi^2_{d,\alpha}\}$. In the one dim case, it is still hard to see, but we consider the normal case.

$$-2\log LR(\theta_0) = \frac{1}{\sigma^2}[\sum_i (X_i - \theta_0)^2 - \sum_i (X_i - \bar{X})^2] < \chi^2_{1,\alpha}$$

Now this can be simplified.

$$\frac{1}{\sigma^2}[n(\bar{X} - \theta_0)^2] < \chi^2_{1,\alpha} \Leftrightarrow \theta_0 \in \bar{X} \pm \frac{z\hat{\sigma}}{\sqrt{n}}$$

6.2.4 CR under partial identification

The duality allows us to construct CI without $\hat{\theta}$. Consider a missing data example. $M = 1$ means Y is missing. $\theta = P(Y = 1)$.

$$\theta = P(Y = 1|M = 1)P(M = 1) + P(Y = 1|M = 0)P(M = 0)$$

we have n in total, d not missing, k of them $Y = 1$.

classical method

forget about missing part: $P(Y = 1|M = 1) = P(Y = 1|M = 0)$. Then

$$\theta = P(Y = 1|M = 0), \quad \hat{\theta} = \frac{\sum_i 1\{notmissing, Y = 1\}}{\sum_i 1\{notmissing\}} = \frac{k}{d}$$

We use delta-method to obtain

$$se(\hat{\theta}) = \sqrt{\frac{[P(M = 0) - P(M = 0, Y = 1)]P(M = 0, Y = 1)}{nP(M = 0)^3}} = \sqrt{\frac{[d - k]k}{d^3}}$$

$$\theta \in \hat{\theta} \pm 1.96 se(\hat{\theta}) = \frac{k}{d} \pm 1.96\sqrt{\frac{[d - k]k}{d^3}}$$

which does not require to know n. This makes sense because you only care how many are not missing.

partial identification approach

We have $0 < P(Y = 1|M = 1) < 1$. So

$$P(Y = 1|M = 0)P(M = 0) \leq \theta \leq P(Y = 1|M = 0)P(M = 0) + P(M = 1)$$

which is

$$P(Y = 1, M = 0) \leq \theta \leq P(Y = 1, M = 0) + P(M = 1).$$

There is no way to identify θ. We say it is **partially identified**.

Consider a test, for a given t (known),

$$H_0 : P(Y = 1, M = 0) \leq t \leq P(Y = 1, M = 0) + P(M = 1)$$

$$P(Y = 1, M = 0) \approx \frac{k}{n}, \quad P(M = 1) \approx \frac{n-d}{n}$$

So we accept, for some $z > 0$,

$$\frac{k}{n} - z \leq t \leq \frac{k+n-d}{n} + z$$

Let $X = 1\{M = 0, Y = 1\}$. So $k/n = \bar{X}$, and $d/n = 1 - \bar{M}$, so accept when

$$\bar{X} - z \leq t \leq \bar{X} + \bar{M} + z$$

$$P(\bar{X} + \bar{M} + z \geq t) = P(\bar{X} + \bar{M} + z \geq t, t \geq \bar{X} - z) + P(\bar{X} + \bar{M} + z \geq t, t \leq \bar{X} - z)$$

$$= P(\bar{X} + \bar{M} + z \geq t, t \geq \bar{X} - z) + P(t \leq \bar{X} - z)$$

So

$$1 - \alpha \leq P(\bar{X} - z \leq t \leq \bar{X} + \bar{M} + z) = P(\bar{X} + \bar{M} + z \geq t) - P(t \leq \bar{X} - z)$$

$$= P(\bar{X} + \bar{M} - (\mu_x + \mu_m) \geq t - (\mu_x + \mu_m) - z) - P(t - \mu_x + z \leq \bar{X} - \mu_x)$$

$$= P(N(0,1) \geq \frac{\sqrt{n}(t - (\mu_x + \mu_m) - z)}{\sqrt{\text{var}(X + M)}}) - P(\frac{\sqrt{n}(t - \mu_x + z)}{\sqrt{\text{var}(X)}} \leq N(0,1))$$

$$= P(N(0,1) \geq \frac{\sqrt{n}(a - z)}{\sqrt{\text{var}(X + M)}}) - P(\frac{\sqrt{n}(b + z)}{\sqrt{\text{var}(X)}} \leq N(0,1))$$

Intuitively, you need this probability to be larger than $1 - \alpha$ even in the worse case. In $P(\bar{X} + \bar{M} + z \geq t)$, the worst case is given by larger t. In $P(t < \bar{X} - z)$, the worst case is by smaller t.

let $a = t - (\mu_x + \mu_m)$, $b = t - \mu_x$. Under

$$H_0 : \mu_x \leq t \leq \mu_x + \mu_m$$

$a \leq 0$ and $b \geq 0$. So

$$\geq P(N(0,1) \geq \frac{\sqrt{n}(-z)}{\sqrt{\mathrm{var}(X+M)}}) - P(\frac{\sqrt{n}z}{\sqrt{\mathrm{var}(X)}} \leq N(0,1))$$

$$= 1 - \Phi[\frac{\sqrt{n}(-z)}{\sqrt{\mathrm{var}(X+M)}}] - [1 - \Phi(\frac{\sqrt{n}z}{\sqrt{\mathrm{var}(X)}})] = \Phi(\frac{\sqrt{n}z}{\sqrt{\mathrm{var}(X)}}) - \Phi[\frac{\sqrt{n}(-z)}{\sqrt{\mathrm{var}(X+M)}}]$$

$$= \Phi(\frac{\sqrt{n}z}{\sqrt{\hat{\mathrm{var}}(X)}}) - \Phi[\frac{\sqrt{n}(-z)}{\sqrt{\hat{\mathrm{var}}(X+M)}}] = 1 - \alpha$$

Figure out z from the last equation. $\mathrm{var}(X) = \mu_x(1-\mu_x) \approx \frac{k}{n}(1-\frac{k}{n})$,

$\mathrm{var}(X+M) = E(X+M+2XM) - (\mu_x+\mu_m)^2 = \mu_x+\mu_m-(\mu_x+\mu_m)^2 = (\mu_x+\mu_m)(1-\mu_x-\mu_m)$

$$\approx \frac{(n-d+k)(d-k)}{n^2}$$

So

$$\Phi(\frac{n\sqrt{n}z}{\sqrt{k(n-k)}}) - \Phi(\frac{-n\sqrt{n}z}{\sqrt{(n-d+k)(d-k)}}) = 1 - \alpha$$

So we accept when

$$\bar{X} - z \leq t \leq \bar{X} + \bar{M} + z$$

this implies when $\mu_x \leq t \leq \mu_x + \mu_m$,

$$P(\bar{X} - z \leq t \leq \bar{X} + \bar{M} + z) \geq 1 - \alpha$$

Now the true value θ satisfies $\bar{X} - z \leq t \leq \bar{X} + \bar{M} + z$, thus

$$P(\bar{X} - z \leq \theta \leq \bar{X} + \bar{M} + z) \geq 1 - \alpha$$

$$\frac{k}{n} - z \leq \theta \leq \frac{k+n-d}{n} + z$$

This is the confidence interval for θ, allowing missing. This one depends on n.

In the special case when $P(M=0)=1$, no missing. $d=n$

$$\Phi(\frac{n\sqrt{n}z}{\sqrt{k(n-k)}})-\Phi(\frac{-n\sqrt{n}z}{\sqrt{k(n-k)}})=1-\alpha, \Rightarrow \frac{n\sqrt{n}z}{\sqrt{k(n-k)}}=z_{\alpha/2}$$

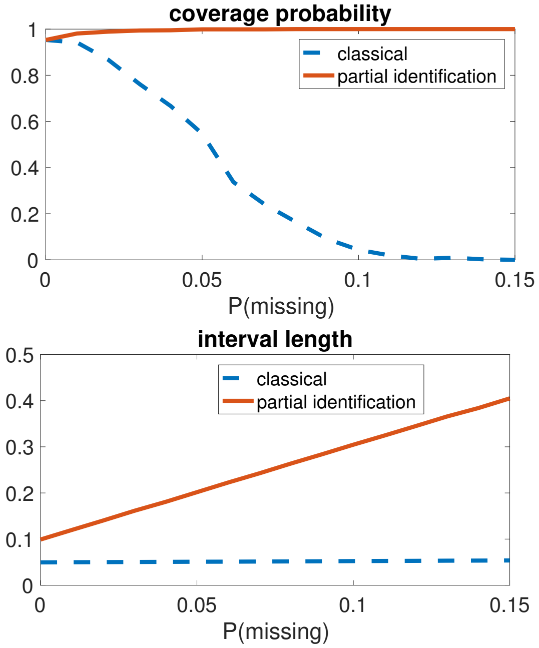
so

$$\frac{k}{n}\pm z_{\alpha/2}\frac{\sqrt{k(n-k)}}{\sqrt{n^3}}$$

Recall the classical one is $\frac{k}{d}\pm 1.96\sqrt{\frac{[d-k]k}{d^3}}$, same.

In the simulation below, we set $P(Y=1|M=1)=0.7$ and $P(Y=1|M=0)=0.2$. We plot the coverage probability as $P(M=1)=P(missing)$ increases. We also compare the length of the interval.

In the simulation, $n=1000$. 1000 replications.



6.2.5 More examples about partial identification

- English Auction

  Two assumptions:

  (1) bidders do not bid more than they want to pay

(2) Bidders do not allow an opponent to win at a price they are willing to beat.

$$y = x^T\beta + e, \quad E(e|x0$$

y: max value a bidder is willing to pay; unobservable

(y_1, y_2): bidders' final bid; winning bid. observable.

Δ : minimum increment: observable.

$$y_1 \leq y \leq y_2 + \Delta$$

$$E(y_1|x) \leq x^T\beta \leq E(y_2|x) + \Delta$$

β is not identified.

- censored data

$$y = x^T\beta + e$$

only observe $Z = \min(y, C)$.

y: number of years a patient live after treatment

C: number of years tracking the patient (e.g., C=10)

$$P(Z > x^T\beta|x) = P(Z > x^T\beta, y > x^T\beta|x) = P(e > 0, C > x^T\beta|x)$$

if we do not observe C,

$$\leq P(e > 0|x) = 0.5$$

β is not identified, but satisfies $P(Z > x^T\beta|x) \leq 0.5$.

- There is a huge literature on partial identification inference.

http://www.public.iastate.edu/~vardeman/stat543/stat543.html
http://www.math.hawaii.edu/~grw/Classes/2013-2014/2014Spring/Math472_1/Solutions06.pdf

7 Minimaxity

7.1 Shannon's information theory

- Entropy X

$$H(X) = -E_X \log p(X) = -\int p(x) \log p(x) dx$$

which represents the chaos of the distr of X.

- Conditional entropy $X|Y$:

$$H(X|Y) = -\int p(x|y) \log p(x|y) dx$$

- Joint entropy decomposition:

$$H(X,Y) = EH(X|Y) + H(Y)$$

$EH(X|Y)$: after knowing Y, the remaining chaos of X.

- Information

$$I(X,Y) = \int p(y)p(x|y) \log \frac{p(x|y)}{p(x)} dxdy = E_Y KL(p(x|y)||p(x))$$

7.2 Fano's inequality

For any Markov chain

$$V \to X \to \hat{V}$$

where V is uniform finite rv with number of possibilities $|V| \geq 2$

then

$$P(\hat{V} \neq V) \geq 1 - \frac{I(V,X) + \log 2}{\log |V|}$$

(1) The term $\log(|V|-1)$ arises from

$$H(V|V \neq V, \hat{V}) = -\sum_{v_i} P(V = v_i | V \neq V, \hat{V}) \log P(V = v_i | V \neq V, \hat{V})$$

$$\leq \log(|V|-1)$$

because given $\hat{V} \neq V$, V can take at most $|V|-1$ possible values. The largest entry of such is $\log(|V|-1)$

(2) $\log 2$ is the max entry for $1\{V \neq \hat{V}\}$

(3) Intuition: Because of the Markov, learning about V using $\hat{V}$, only relies on X.

So the success of learning depends on how much information X has regarding V.

If $I(V, X)$ is good, then the learning error $P(\hat{V} \neq V)$ can be "possibly" (lower bound) very small

7.3 Minimax risk

Let X be data, let $\hat{\theta}$ be estimator. Let $L(\|\hat{\theta} - \theta\|)$ be loss

Define minimax risk

$$M_L = \inf_{\hat{\theta}} \sup_{\theta} E_\theta L(\|\hat{\theta} - \theta\|)$$

Here E_θ treats θ as the truth.

- max means: among the worst scenario of the truth

  mini means: the best estimator.

  So minimax: in the most difficult case of the truth, the best we can do.

- 2-delta packing

  Consider a finite parameter space $\Theta = \{\theta_1 ... \theta_M\}$, each $\|\theta\| = \delta$.

  called 2-delta packing if

  $$\inf_{\theta_i \neq \theta_j, \in \Theta} \|\theta_i - \theta_j\| \geq 2\delta$$

- canonical estimating problem:

  (1) Uniformly choose $\theta \in \Theta$

  (2) draw data X from the distribution P_θ; latter is completely determined by θ

  (3) Let P_Θ denote the joint distribution over (θ, X)

  $$P_\Theta = P(X|\theta)P(\theta) = P_\theta \times Uniform(\Theta)$$

- Theorem: for the finite 2-delta packing Θ,

  $$M_L \geq L(\delta) \inf_{\hat{\theta}} P_\Theta(\hat{\theta} \neq \theta)$$

  Here Θ has to be the finite packing for the original parameter space, meaning that

  $$\Theta \subset \text{original parameter space for } \theta \text{ we care about}$$

- Theorem: combined with Fano's lemma

  $$M_L \geq L(\delta)\left(1 - \frac{I(\theta, X) + \log 2}{\log|\Theta|}\right)$$

  where θ is Uniform from a 2-delta packing finite space Θ, and X is drawn from P_θ

  In the above, M_L does not depend on the choice of packing, but $I(\theta, X)$ does.

7.4 Packing number

- we also need a lower bound for the packing $\log|\Theta|$. To make the result meaningful, we wish the lower bound as large as possible.
- Packing number for a set Ω: the largest possible elements in this set, mutually δ-distance away
- For a unit ball in the d-dim Eu-space, the delta-packing number is at least

  $$(\frac{C}{\delta})^d$$

  So in practice, we can take parameter θ for packing as

  $$\theta_j = v_j\delta$$

  where v_j is in the unit ball. Use a 1/2-packing for the unit ball: $\{v_1, ..., v_M\}$, then $|M| \geq (\frac{C}{\delta})^d$. Let

  $$\Theta = \{v_j\delta : j = 1...M\}$$

  making sure $v_j\delta$ is inside the original parameter space. This then becomes the packing for the original parameter space

  $$\|\theta_i - \theta_j\| \geq \delta\|v_i - v_j\| \geq \frac{\delta}{2}$$

7.5 Normal models

The key is to find a good δ, and a finite packing Θ.

Usually $I(\theta, X)$ is increasing in δ. Larger delta makes θ's more separated from each other, easier to detect using data, so more information.

7.5.1 Normal mean model

Suppose $X \sim N(\theta, \sigma^2)$.

- Consider a packing Θ with number of elements at least 2^d, where $d = \dim(\theta)$.

$$\delta \geq \|\theta_1 - \theta_2\| \geq \frac{\delta}{2}$$

  for each pair $\theta \in \Theta$. (need $0.5 \leq \|v_1 - v_2\| \leq 1$)

  here

$$|M| \geq (\frac{C}{\delta})^d \geq 2^d$$

  make sure $C/\delta \geq 2$.

- Then

$$\inf_{\hat{\theta}} \sup_{\theta} E_\theta \|\hat{\theta} - \theta\|^2 \geq (\frac{\delta}{4})^2 \left(1 - \frac{I(\theta, X) + \log 2}{\log |V|}\right)$$
$$\geq \frac{\delta^2}{16} \left(1 - \frac{I(\theta, X) + \log 2}{d \log 2}\right)$$

- To calculate the information, now for two normal distributions , the KL is

$$KL(N(\theta_1, \Sigma) || N(\theta_2, \Sigma)) = \frac{1}{2}(\theta_1 - \theta_2)^T \Sigma^{-1} (\theta_1 - \theta_2)$$

  So for any pair θ_1, θ_2 in the packing, the likelihood (from n observations) $P_{\theta_1}, P_{\theta_2}$ satisfy

$$KL(P_{\theta_1} || P_{\theta_2}) = \frac{n}{2\sigma^2} \|\theta_1 - \theta_2\|^2$$

  In addition, we note that $X \sim P_\theta$, and $\theta \sim Uniform(\Theta)$; Also note that

$$I(\theta, X) = E_{\theta \sim Uni(\Theta)} KL(P_\theta || p(X))$$

  where $p(X)$ is marginal on X. Combining all these and concavity of log, one can show

$$I(\theta, X) \leq average(KL(P_{\theta_1} || P_{\theta_2})) \leq \frac{n}{2\sigma^2} \delta^2$$

- So

$$\inf_{\hat{\theta}} \sup_{\theta} E_\theta \|\hat{\theta} - \theta\|^2 \geq \frac{\delta^2}{16} \left(1 - \frac{\frac{n}{2\sigma^2}\delta^2 + \log 2}{d \log 2}\right)$$

Now take

$$\delta^2 = d\sigma^2 \log 2/(2n)$$

$$1 - \frac{\frac{n}{2\sigma^2}\delta^2 + \log 2}{d \log 2} = 1 - \frac{1}{d} - \frac{1}{4} \geq \frac{1}{4}$$

So

$$\inf_{\hat{\theta}} \sup_{\theta} E_\theta \|\hat{\theta} - \theta\|^2 \geq \frac{1}{180} \frac{d\sigma^2}{n}$$

7.5.2 linear regression

Suppose $Y \sim N(X\theta, \sigma^2)$.
Here data is Y, we condition on X. Here all is n dimensional

- Still use same packing:

  $$\|v_j\| \leq 1, \quad \theta_j = \delta v_j, \quad 0.5 \leq \|v_1 - v_2\| \leq 1$$

  make sure δ is not too large. packing number

  $$\log |M| \geq \log(\frac{C}{\delta})^d \geq d \log 2$$

  $$0.5\delta \leq \|\theta_1 - \theta_2\| \leq \delta$$

- $$\begin{aligned} I(\theta, Y) &\leq average(KL(N(X\theta_1, \sigma^2)||N(X\theta_2, \sigma^2))) \\ &\leq \frac{1}{2\sigma^2}\|X(\theta_1 - \theta_2)\|^2 \leq \frac{1}{2\sigma^2}\lambda_{\max}(X^T X)\delta^2 \end{aligned}$$

  take

  $$\frac{1}{2\sigma^2}\lambda_{\max}(X^T X)\delta^2 = \frac{d \log 2}{4}$$

- So suppose $d \geq 2$,

  $$\begin{aligned} \inf_{\hat{\theta}} \sup_{\theta} E_\theta \|\hat{\theta} - \theta\|^2 &\geq (\frac{\delta}{4})^2 \left(1 - \frac{\frac{1}{2\sigma^2}\lambda_{\max}(X^T X)\delta^2 + \log 2}{d \log 2}\right) \\ &= (\frac{\delta}{4})^2 (0.75 - 1/d) = C \frac{d\sigma^2}{\lambda_{\max}(X^T X)} \end{aligned}$$