1

PROBABILITY SPACES

THE IDEA

We begin prob with the vague notion of a random experiment. A repeatable process that does not have a determinate outcome.

We begin to abstract this by formalising the:

- Sample Space: Ω set of all outcomes
- Outcomes: An element $\omega \in \Omega$
- Events: Subsets $A \subset \Omega$ such that we can define probability on it.

But note that In the general case we cannot use ARBITRARY subsets as events.

SETS

we denote the union of disjoint sets $D \cup E = D + E$

The symetric difference of a set is: $A \Delta B = A \setminus B + B \setminus A$
$= (A \cup B) \setminus (A \cap B)$

INDICATORS: For an arbitrary set A the indicator function $\mathbb{1}_A(\omega) = \begin{cases} 1, & \omega \in A \\ 0, & \omega \notin A \end{cases}$

For two sets A & B: $\mathbb{1}_{A^c} = 1 - \mathbb{1}_A$, $\mathbb{1}_{A \cup B} = \max\{\mathbb{1}_A, \mathbb{1}_B\}$

$\mathbb{1}_{A \cap B} = \mathbb{1}_A \mathbb{1}_B$, $\mathbb{1}_{A \Delta B} = |\mathbb{1}_A - \mathbb{1}_B|$

EVENTS & ALGEBRAS

We need that our events are closed under cirtain operations so that when we manipulate events we still have events.

A family $\mathcal{F}$ of subsets of Ω is a σ-Algebra if it satisfies:

A.1) $\Omega \in \mathcal{F}$
A.2) $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$
A.3) $A_1, A_2, \dots \in \mathcal{F} \Rightarrow \bigcup_{n \geq 1} A_n \in \mathcal{F}$

Note that from De'Morgans Law this also implies closed under countable intersections.

Thus we will now call the elements of an appropriate σ-Algebra generated on subsets of Ω Events

T: For two σ-Algebras $\mathcal{F}_1$ & $\mathcal{F}_2$ on a common sample space Ω, then $\mathcal{F}_1 \cap \mathcal{F}_2$ is also a σ-Algebra.

T: $\mathcal{F}_n$, $n \in \mathbb{N}$, σ-Algebras on a common sample space $\Rightarrow \bigcap_{n \geq 1} \mathcal{F}_n$ is a σ-Algebra.

Given some Ω how do we create a σ-Algebra?

σ ALGERA GENERATED:

① For a single $A \subset \Omega$, $\sigma(A) = \{\emptyset, A, A^c, \Omega\}$

② For $\mathcal{G} = \{A_1, \dots, A_n\}$ a finite partition of Ω

$\sigma(\mathcal{G}) = \{\sum_{i \in I} A_i : I \subset \{1, \dots, n\}\}$

For an arbitrary collection of sets, consider that we can form a partition by taking all possible intersections.

T: For any family of subsets of Ω, $\mathcal{G}$, $\exists$ a unique σ Algebra, $\sigma(\mathcal{G})$ s.t. $\mathcal{G} \subset \sigma(\mathcal{G})$ & $[\mathcal{H}$ a σ Algebra on Ω & $\mathcal{G} \subset \mathcal{H} \Rightarrow \sigma(\mathcal{G}) \subset \mathcal{H}]$

BOREL SET: This is the cannonical σ Alg generated used on $\mathbb{R}$.

$$\mathcal{B}(\mathbb{R}) = \sigma\{(a, b] \mid a, b \in \mathbb{R}, a < b\}$$

$$\mathcal{B}(\mathbb{R}^m) = \sigma\{\prod_{i=1}^{m} (a_i, b_i] \mid a_i, b_i \in \mathbb{R}, a_i < b_i\}$$

A_n occur i.o. slide 11.

PROBABILITY SPACE

A set, Ω, paired with a σ Alg generated on its subsets, $\mathcal{F}$, is called a measurable space $(\Omega, \mathcal{F})$

A probability on $(\Omega, \mathcal{F})$ is a function $\mathbb{P}: \mathcal{F} \to \mathbb{R}$ satisfying

P.1) $P(A) \geq 0$, $A \in \mathcal{F}$
P.2) $P(\Omega) = 1$
P.3) For any pairwise disjoint $A_1, A_2, \dots \in \mathcal{F}$ $\quad P(\bigcup_{j \geq 1} A_j) = \sum_{j \geq 1} P(A_j)$

The tripple $(\Omega, \mathcal{F}, \mathbb{P})$ is called a prob space.

Ex. Degenerate Dist at a fixed $\omega \in \Omega$

$$\varepsilon_\omega(A) = \mathbb{1}_A(\omega)$$

E. Counting measure on $\mathbb{N}$

$$\mu(B) = \sum_{n \geq 1} \varepsilon_n(B), \quad B \in P(\mathbb{N})$$

From these axioms we can further deduce these properties of the probability measure:

T: $P(\emptyset) = 0$, T: $P(\bigcup_{n=1}^{m} A_j) = \sum_{n=1}^{m} P(A_j)$ For pairwise disjoint $A_1, \dots, A_m$

T: $P(A^c) = 1 - P(A)$

T: $A \subset B \Rightarrow P(B \setminus A) = P(B) - P(A) \Rightarrow P(A) \leq P(B)$

T: $P(A \cup B) = P(A) + P(B) - P(B \cap A)$

T: Booles Ineq: $P(\bigcup_{j \geq 1} A_j) \leq \sum_{j \geq 1} P(A_j)$

T: Borel-Cantelli: $\sum_{n \geq 1} P(A_n) < \infty \Rightarrow P(A_n \text{ i.o.}) = 0$

CONTINUITY PROPERTIES:

The infinite case of the countable additivity property of our probability is responsible for important continuity properties of the $\mathbb{P}$.

$\underset{n \to \infty}{A_n \uparrow A} \iff A_1 \subset A_2 \subset \cdots$ & $\bigcup_{n \geq 1} A_n = A$

$\underset{n \to \infty}{A_n \downarrow A} \iff A_1 \supset A_2 \supset \cdots$ & $\bigcap_{n \geq 1} A_n = A$

T: A function $\mathbb{P}: \mathcal{F} \to \mathbb{R}$ satisfies P.1, P.2 & has finite additivity. THEN (the following are $\iff$)

$\mathbb{P}$ has property P.3

$\iff [A_n \uparrow A \Rightarrow \mathbb{P}(A_n) \uparrow \mathbb{P}(A)] \iff [A_n \downarrow \emptyset \Rightarrow \mathbb{P}(A_n) \downarrow 0]$

$\iff [A_n \downarrow A \Rightarrow \mathbb{P}(A_n) \downarrow \mathbb{P}(A)]$

$\boxed{2}$

PROBABILITIES ON $\mathbb{R}$.

Probabilities on $\mathbb{R}$ are defined on the measureable space $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

BUT $\mathcal{B}(\mathbb{R})$ is HUGE! How do we specify the probability on all events in $\mathcal{B}(\mathbb{R})$ for a P?

DISTRIBUTION FUNCTION:

It turns out we can entirely specify a P by its distribution function.

Distribution Function (D.F) of a probability P on $\mathbb{R}$ is the function $F_P : \mathbb{R} \to \mathbb{R}$, $t \longmapsto P(-\infty, t]$

T: For any probability P on $\mathbb{R}$ its D.F. F satisfies all of...

D.1) Non Decreasing: $s < t \Rightarrow F(s) \leq F(t)$

D.2) Right continuous: $F(t) = F(t+)$

D.3) $\lim_{t \to -\infty} F(t) = 0$ & $\lim_{t \to \infty} F(t) = 1$

T: $F_P = F_{P'} \iff P = P'$ (The probability & its distribution function entirely determine one another.)

T: For any $F : \mathbb{R} \to \mathbb{R}$ satisfying D.1-3 there corresponds EXACTLY ONE Probability P on $\mathcal{B}(\mathbb{R})$.

CLASSIFYING $\mathbb{P}$ on $\mathbb{R}$

P is Discrete on $\mathbb{R}$

$\iff (\exists C \subset \mathbb{R})(C$ countable & $P(C) = 1)$

T: $\iff \exists \{t_i\}_{i \geq 1} \subset \mathbb{R}$ & $\exists \{p_i > 0\}_{i \geq 1}$ such that

$\sum p_i = 1$ & $P = \sum p_i \varepsilon_t$

$\iff \exists \{t_i\}_{i \geq 1} \subset \mathbb{R}$ & $\exists \{p_i > 0\}_{i \geq 1}$ such that

$\sum_i p_i = 1$ & $F_P(t) = \sum_i p_i \mathbb{1}(t_i \leq t)$

P is absolutely continuous (A.C) on $\mathbb{R}$ if $\exists f_P : \mathbb{R} \to \mathbb{R}$ such that $F_P(t) = \int_{-\infty}^{t} f_P(x)\,dx$.

So A.C. distributions are the ones with densities. Note that $f_P = F_P'$ almost everywhere. We must be careful about points of discontinuity.

T: Any function $f : \mathbb{R} \to \mathbb{R}$ that is

① $f \geq 0$ ② Integrable ③ $\int f_P\,dx = 1$

specifies a probability on $\mathbb{R}$.

A mixed distribution P is one such that for some $p \in (0,1)$ $P = pP_d + (1-p)P_a$ where P_d is discrete & P_a is A.C.

A singular distributions are continuous but not A.C., they Dont have a density however no single point has a positive probability.

T: Any probability on $\mathbb{R}$ has a unique representation of the form

$$P = \underbrace{\alpha_d P_d}_{\text{Discrete}} + \underbrace{\alpha_a P_a}_{\text{A.C}} + \underbrace{\alpha_s P_s}_{\text{Singular.}}, \quad \alpha_i \geq 0, \sum_i \alpha_i = 1$$

3

RANDOM VARIABLES

We naively think of a R.V. as a function of an outcome of a Random experiment that captures some information about the experiment. We also however want to be able to calculate the probabillty of our R.V. mapping to a certain value (of set of values)

Thus we define a Random Variable (R.V) as a function $X : \Omega \longrightarrow \mathbb{R}$ such that $\forall B \in \mathcal{B}(\mathbb{R}) \quad X^{-1}(B) \in \mathcal{F}$.

$$X^{-1}(B) = \{\omega \in \Omega \mid X(\omega) \in B\}$$

T: For an arbitrary family of subsets of $\mathbb{R}$ $\{B_\alpha \mid \alpha \in I\}$

- $B_\alpha \subset B_\beta \Rightarrow X^{-1}(B_\alpha) \subset X^{-1}(B_\beta)$
- $\bigcup_{\alpha\in I} X^{-1}(B_\alpha) = X^{-1}(\bigcup_{\alpha\in I} B_\alpha)$ (similarly for $\cap$)
- $B_\alpha \cap B_\beta = \emptyset \Rightarrow X^{-1}(B_\alpha) \cap X^{-1}(B_\beta)$
- $X^{-1}(B^c_\alpha) = [X^{-1}(B_\alpha)]^c$

Ex. Random indicators: For any event A, $\mathbb{1}_A$ is a R.V.

Simple R.V.: $\sum_{i=1}^{n} a_i \mathbb{1}_{A_i}$, $a_i \in \mathbb{R}$, $A_i \in \mathcal{F}$, $i \leq n < \infty$

T: For a R.V X, $\sigma(X) = \{X^{-1}(B) \mid B \in \mathcal{B}(\mathbb{R})\}$ is a σ Alg.

DISTRIBUTIONS OF R.V.

The distribution of a R.V X on $(\Omega, \mathcal{F}, \mathbb{P})$ is defined as $P_X(B) = \mathbb{P}(X \in B)$, $P_X : \mathcal{B}(\mathbb{R}) \longrightarrow \mathbb{R}$. This P_X is in fact a probability on $\mathbb{R}$.

Thus the distribution function of X is
$$F_X(t) = P_X((-\infty, t]) = \mathbb{P}(X \leq t)$$

X is discrete/A.C/singular if F_X is D/A.C/S. same for R.Vec

The survival function of X is $S_X(t) = 1 - F_X(t)$

Two R.V. X & Y are identically distributed $\Leftrightarrow P_X = P_Y$

FUNCTIONS OF R.V:

T: X a R.V, g increasing $(g'' \geq 0)$ & continuous on $\mathbb{R}$ $\Longrightarrow Y = g(X)$ is a R.V with D.F $F_Y(t) = F_X(g^{-1}(t))$

T: X an A.C R.V, g continuously diff on open set U such that $P(X \in U) = 1$.

$\Longrightarrow Y = g(X)$ is A.C RV. $\quad f_Y(t) = \left|\frac{d}{dt} g^{-1}(t)\right| f_X(g^{-1}(t))$

Note that we must also have that g is invertable & that inverse is differentable.

For a D.F F the quantile function is
$Q(x) = \inf\{t : F(t) \geq x\}$, $x \in (0,1)$

T: $U \sim U[0,1] \Longrightarrow X = Q(U) \sim F$

RANDOM VECTORS

A Random Vector (R.vec) $X = (X_1, \ldots, X_n) : \Omega \longrightarrow \mathbb{R}^n$ is a function such that $X^{-1}(B) \in \mathcal{F} \quad \forall B \in \mathcal{B}(\mathbb{R})$

T: $X = (X_1, \ldots, X_n)$ R.vec $\Longleftrightarrow (\forall i)(X_i$ is a R.V$)$

T: $X = (X_1, \ldots, X_n)$ a R.vec & g measurable $g : \mathbb{R}^n \to \mathbb{R}^n$ $\Longrightarrow g(X)$ is a R.vec

g measurable $\Leftrightarrow \forall B \in \mathcal{B}(\mathbb{R})$ $g^{-1}(B) \in \mathcal{B}(\mathbb{R})$

We have the D.F. in the multivariate case as
$F_X(t_1, \ldots, t_n) = \mathbb{P}(X_1 \leq t, \ldots, X_d \leq t_d)$, $(t_1, \ldots, t_n) \in \mathbb{R}^n$.

T: An A.C dist has a density f_X satisfying
$$F_X(t_1, \ldots, t_n) = \int_{-\infty}^{t_1} \cdots \int_{-\infty}^{t_n} f_X(x_1, \ldots, x_n)\, dx_1 \cdots dx_n$$

T: $(X_1, \ldots, X_n)$ is discrete $\Longleftrightarrow (\forall i)(X_i$ is discrete$)$

T: $(X_1, \ldots, X_n)$ is A.C. $\Longrightarrow (\forall i)(X_i$ is A.C$)$ &

Integrate out all but the desired variable.

$$f_{X_i}(x) = \int \cdots \int f_X(s_1, \ldots, x, \ldots, s_n)\, ds_1 \cdots ds_n$$

T: X a R.vec, $g : \mathbb{R}^n \to \mathbb{R}^n$ has a smooth inverse

X A.C $\Longrightarrow Y = g(X)$ is A.C R.V such that
$$f_Y(t) = |\det[g^{-1}(t)]|\, f_X(g^{-1}(t))$$

$\det(h(t)) = J(h(t)) = \det\left[\frac{dh_i}{dt_i}\right]$ is the jacobian.

INDEPENDENCE

A collection of R.V. $X_1, \ldots, X_n$ are independent if
$(\forall B_1, \ldots, B_n \in \mathcal{B}(\mathbb{R}))(P(X_1 \in B_1, \ldots, X_n \in B_n) = \prod_{i=1}^{n} P(X_i \in B_i))$

This is a very general definition of independence however it is not convenient as a test. For showing some R.Vs are independent we will often use the following.

T: $X_1, \ldots, X_n$ R.V independent $\Longleftrightarrow$ $\forall t_1, \ldots, t_n \in \mathbb{R}$
$F_{X_1, \ldots, X_n}(t_1, \ldots, t_n) = \prod_{i=1}^{n} F_{X_i}(t_i)$

T: Discrete $X_1, \ldots, X_n$ R.V independent $\Longleftrightarrow$ $\forall t_1, \ldots, t_n \in \mathbb{R}$
$P(X_1 = t_1, \ldots, X_n = t_n) = \prod_{i=1}^{n} P(X_i = t_i)$

T: AC R.V. $X_1, \ldots, X_n$ indepencent $\Longleftrightarrow$ $\forall t_1, \ldots, t_n \in \mathbb{R}$
$f_{X_1, \ldots, X_n}(t_1, \ldots, t_n) = \prod_{i=1}^{n} f_{X_i}(t_i)$

T: g_i measurable functions, $X_1, \ldots, X_n$ independent
$\Longrightarrow Y_j = g_j(X_j)$ are also independent.

Events $A_1, \ldots, A_n$ are independent

$\Longleftrightarrow \mathbb{1}_{A_1}, \ldots, \mathbb{1}_{A_n}$ are independent as R.V.

T: $\Longleftrightarrow (\forall I \subset \{1, \ldots, n\})(P(\bigcap_{i\in I} A_i) = \prod_{i\in I} P(A_i))$.

T: $\Longleftrightarrow A^c_1, \ldots, A^c_n$ are independent

4

EXPECTATIONS

DEFINING EXPECTATIONS

In second year we talked about $E(X)$ for A.C & discrete as simply $\int x f(x)\,dx$. This however is just a computational tool, not an informative definition. It is also not general enough, what about singular or mixed distributions? We also want our def to align with frequentist intuition

Indicator

For $\mathbb{1}_A$, $A \in \mathcal{F}$ we define $E(X) = P(A)$

simple R.V.

For $X = \sum_{i=1}^{n} a_i \mathbb{1}_{A_i}$ we define $E(X) = \sum_{i=1}^{n} a_i P(A_i)$

Now note that any non-negative R.V can be approximated by an increasing sequence of simple R.V.s $\{X_n\}_{n\geq 1}$ in the following way

$$\forall \omega \in \Omega \quad X_n(\omega) \underset{n\to\infty}{\uparrow} X(\omega) . \quad ☆$$

We will use our current definition & this approximation to define

$X \geq 0$ Arbitrary, $E(X) = \lim_{n\to\infty} E(X_n)$ Where $\{X_n\}_{n\geq 1}$ is a sequence of simple R.V.s as in ☆

T: This definition is consistent. Different sequences will give the same expectation.

Let $X^+ = \max\{X, 0\}$ & $X^- = -\min\{X, 0\}$

& Note that for an arbitrary $X = X^+ - X^-$

A R.V. is integrable if $E(|X|) < \infty \iff X \in L^1$

If X is integrable $E(X) = E(X^+) - E(X^-)$

The expectation of a R.V. X over an event A is $E(X ; A) = E(X \mathbb{1}_A)$.

UNDERSTANDING EXPECTATION

T: Expectation as a function is

- Monotone : $X \leq Y$ & $E(Y) < \infty \Rightarrow E(X) \leq E(Y)$
- Linear: $\forall a, b \in \mathbb{R}$, $X, Y \in L^1 \Rightarrow E(aX + bY) = aE(X) + bE(Y)$

We can denote E using Lebesgue integrals

$$E(X) = \int_\Omega X(\omega)\,\mathbb{P}(d\omega) = \int_\Omega X(\omega)\,d\mathbb{P}(\omega) = \int_\Omega X\,d\mathbb{P}$$

Moving probability spaces from $(\Omega, \mathcal{F}, \mathbb{P})$ (general) to $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), P_X)$ allows us to shift

T: $E(X) = \int_\Omega X(\omega)\,d\mathbb{P}(\omega) = \int_{\mathbb{R}} x\,dP_X(x)$ when $E(X)$ is defined.

distribution func of P_X.

Notation: $\int g(x)\,dP_X(x)$ often denoted $\int g(x)\,dF_X(x)$

T: If F is A.C with density $f = F'$ (a.e) and both f & g are piecewise continuous then

$$\int g(x)\,dF(x) = \int_{-\infty}^{\infty} g(x) f(x)\,dx$$

Lebesgue Riemann

USING EXPECTATIONS

T: $Y = g(X) \sim P_Y$, $X \sim P_X \Rightarrow E(Y) = \int g(x)\,dP_X(x)$

T: $X \geq 0$, $E(X) = \int_0^\infty (1 - F_X(x))\,dx$

T: $Y = g(X)$ for nice $g \Rightarrow E(Y) = \int g(x)\,dF_X(x)$

commonly $E(g(X)) = \int g(x) f_X(x)\,dx$ for AC.

$\quad\quad\quad\quad — \!/\!/ — \quad = \sum_{t_i \in C_X} g(t_i) P(X = t_i)$.

T: X_1 & X_2 independent & $g_i(X_i) = L^1$

$$\Rightarrow E(g_1(X_1)\, g_2(X_2)) = E(g_1(X_1))\, E(g_2(X_2))$$

INEQUALITIES:

T: Jensens: $X \in L^1$, g convex $\Rightarrow g[EX] \leq E[g(X)]$

T: Lyapunov : For $0 < r \leq s$ $\left(E\left[|X|^r\right]\right)^{\frac{1}{r}} \leq \left(E\left[|X|^s\right]\right)^{\frac{1}{s}}$

T: Chebyshev : g positive nondecreasing on $\mathbb{R} \to \mathbb{R}$

$\Rightarrow$ For any X $\quad P(X \geq a) \leq \dfrac{E(g(X))}{g(a)}$

T: Cauchy: $E|XY| \leq \sqrt{E(X^2)E(Y^2)}$

MOMENTS:

The k^{th} moment of X is $E(X^k)$

The k^{th} central moment of X is $E[(X - E(X))^k]$

The 2^{nd} central moment is $V(X) = E(X^2) - [E(X)]^2$

The mixed moments of X & Y are $E(X^n Y^m)$.

For $X, Y \in L^2$ $\quad Cov(X, Y) = E[(X - E(X))(Y - E(Y))]$

$$= E[XY] - E(X)E(Y)$$

$$Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{V(X)V(Y)}}$$

T: $V(X + Y) = V(X) + V(Y) + 2\,cov(X, Y)$

T: $|Corr(X, Y)| = 1 \iff P(Y = aX + b) = 1$ for $a \neq 0, b \in \mathbb{R}$

MULTI-DIMENSION:

When considering R.Vectors $\underbrace{X = (X_1, \dots, X_d)}_{\text{A row vector}}$ $d \geq 3$

Covariance becomes a matrix.-

$$C_X^2 = [Cov(X_i, X_j)] = E[(X - E(X))^T (X - E(X))]$$

$C_X^2(i, i) = V(X_i)$, C_X^2 is symetric, C_X^2 is semi-positive definite.

$C_X^2(i, j) = C_X^2(j, i)$ $\quad \forall x \in \mathbb{R}^d$ $\; x C_X^2 x^T \geq 0$

T: If $X = (X_1, \dots, X_n)$ has iid components $X_i \sim N(0, 1)$

& $Y = \mu + XA$, $\mu \in \mathbb{R}^m$, $A \in \mathbb{R}^{n \times m}$

Row

$\Rightarrow Y \sim MVN(\mu, A^T A)$, For $Y \sim MVN(\mu, C_Y^2)$ it will have density...

$$f_Y(y) = \frac{1}{\sqrt{(2\pi)^m \det(C_Y^2)}} \exp\left[-\frac{1}{2}(y - \mu)\left[C_Y^2\right]^{-1}(y - \mu)^T\right]$$

5

CONDITIONAL EXPECTATIONS

DEFINING CE

When we dont know anything about a R.V. X our best guess is E(X). What about if we do know something about X but not its value, say the outcome of a related R.V.

Suppose what we knew about the outcome of a random experiment is that an event A occured. We define CE in this context as...

The Conditional Expectation (CE) of R.V. X given event A is $E(X|A) = \frac{E(X\mathbb{1}_A)}{P(A)}$.

Next consider if we have a partition of $\{A_1, \dots, A_n\}$ of the sample space Ω. If all we know is which of these events occurred then we have simpley the value of a simple R.V. $Y = \sum_{i=1}^{n} y_i \mathbb{1}_{A_i}$. Say Y takes value y_i.

Our best guess for X is then $E(X|A_i)$.

since $A_i = \{Y = y_i\}$ & $\hat{X} = E(X|A_i) = h(y)$

let $E(X|Y) = h(Y)$. ← SIMPLE R.V.

T: Let $X \in L^1$ & Y be R.V on common prob space

$\implies \exists \hat{X}$ a R.V satisfying

$\hat{X}$ is a R.V on $\sigma(Y)$

CE.1) $\hat{X}$ flat on atoms of $\sigma(Y)$

CE.2) $E(\hat{X} : A) = E(X : A) \quad \forall A \in \sigma(Y)$

$E(X:A) = E(X\mathbb{1}_A)$.

That is unique up to values on sets of zero probability.

We call this unique R.V $E(X|Y)$.

Note if $\mathcal{F} \subset \sigma(Y)$ & we replace CE.1) with the same condition on $\mathcal{F}$ the theorem is still true, we call this CE of X given σ Alg $\mathcal{F}$ $E(X|\mathcal{F})$.

PROPERTIES OF CE

T: φ is 1-1 function (injective) $\implies E(X|Y) = E(X|\varphi(Y))$

T: Linearity: $(\forall a, b \in \mathbb{R})(E(aX + bZ|Y) = aE(X|Y) + bE(Z|Y))$

T: Monotone: $X \leq Z$ a.s. $\implies E(X|Y) \leq E(Z|Y)$ a.s.

T: $Z = g(Y) \implies E(ZX|Y) = Z\,E(X|Y)$

T: X & Y independent $\implies E(X|Y) = E(X)$

T: Double $\mathbb{E}$: $E[E(X|Y_1, Y_2)|Y_1] = E(X|Y_1)$

In particular $E[E(X|Y)] = E(X)$.

OTHER CONDITIONALS

Conditional probabillties are defined for an event A & R.V Y $P(A|Y) = E(\mathbb{1}_A|Y)$.

Conditional distributions are non-trivial however it can be proved that conditional distributions $P_{X|Y}(B|y) = P(X \in B|Y) = E(\mathbb{1}(X \in B)|Y)$ always exist.

When (X,Y) is A.C we can use conditional densities $f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$

Where $f_Y(y) = \int f_{X|Y}(x,y)\,dx$.

Thus $E(X|Y) = \int x f_{X|Y}(x|Y)\,dx$.

6

SUFFICIENT STATISTICS

THE MODEL

For observed data we make the assumptions that the underlying RE is given by $(\Omega, \mathcal{F}, \mathbb{P}_\theta)$. $\mathbb{P}_\theta$ is a probability depending on parameter $\theta \in \Theta \subset \mathbb{R}^d$ whos value we dont know.
We observe a Random vector $X(\omega) = X \in \mathbb{R}^n$
P_θ is the distribution on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ induced by X & $\mathbb{P}_\theta$.

STATISTICS

A measurable function $S(X)$ is a statistic.
Statistics are also thus R.vec.

Estimators of a parameter θ are statistics with codomain Θ.

SUFFICIENCY

A statistic S is sufficient (SS) for a parameter θ if the conditional distribution $P_\theta(X \in B \mid S)$, $B \in \mathcal{B}(\mathbb{R})$ doesnt depend on θ.

T: For φ 1-1 function (Injection) & S a SS for θ
$\Rightarrow$ $\varphi(S)$ is also SS for θ.

How can we find & show statistics are sufficient?
It is most convenient to use densities.
Importantly only A.C. distributions have densities, however A.C. is relative to some measure!
So discrete distributions are A.C. relative to the counting measure.

T: Suppose all P_θ are A.C. with respect to some measure μ, with densities $f_\theta(x) = \frac{dP_\theta}{d\mu}(x)$.

S is a SS for θ $\iff$ $\exists \psi(s,\theta)\ \exists h(x)$
$$f_\theta(x) = \psi(S(x), \theta) h(x)$$
Neyman-Fisher

Note that $S(x)$ may be a R.Vec

NOTE: For $X = (x_1, \ldots, x_n)$ iid $\Rightarrow f_\theta(x) = \prod^n f_\theta(x_i)$

T: If T is a statistic and $S = \varphi(T)$ for some φ is a SS for θ $\Longrightarrow T$ is SS for θ.

$\theta^*(X) = \hat{\theta}^* = \underset{\theta \in \Theta}{\operatorname{argmax}} f_\theta(X)$ is the maximum likelyhood estimator (MLE) of θ from X.

T: S a SS for $\theta \Longrightarrow \hat{\theta}^*$ is a function of S only.

BIAS

How can we compare estimators?
We know that there is not a perfect estimator for anything other than the degenerate distribution. ie. $\nexists \theta^*$ such that $E(\theta^* - \theta)^2$ is minimised $\forall \theta$.

We need to ask for less, so we compare estimators within certain classes of estimators.

$\theta^*_0 \in K$, a class of estimators for θ, is efficient in K $\iff \forall \theta^* \in K$, $E_\theta(\theta^*_0 - \theta)^2 \leq E_\theta(\theta^* - \theta)^2$ $\forall \theta \in \Theta$.

A common class is the class of estimators with bias $b(\theta)$. $K_b = \{\theta^* \mid E_\theta(\theta^*) = \theta + b(\theta), \forall \theta \in \Theta\}$.

K_0 is the class of unbiased estimators.

T: An estimator efficient in K_b is unique up to values on a set of 0 probability.

T: Rao-Blackwell: $\theta^* \in K_b$, S a SS for θ
$\Longrightarrow$ $\theta^*_s = E_\theta(\theta^* \mid S)$ has properties

- θ^*_s is a function of S only
- $\theta^*_s \in K_b$
- $E_\theta(\theta^*_s - \theta)^2 \leq E_\theta(\theta^* - \theta)^2$, $\forall \theta \in \Theta$.

For $\theta \in \mathbb{R}^d$ we can measure the performance of an estimator using $E_\theta(\theta^* - \theta, a) \leftarrow$ for $a \in \mathbb{R}^d$ where $(\cdot,\cdot)$ is the scalar product. Dispersion.
We prefer an estimator if its dispersion is lower $\forall a$.

T: MV R-B: Same except the last condition now...

- $E_\theta(\theta^*_S - \theta, a)^2 \leq E_\theta(\theta^* - \theta, a)^2$, $\forall \theta \in \Theta$

7

CONVERGENCE OF RANDOM VAR:

We know what it formally means for a sequence of numbers to converge to a point.
What might it mean for a sequence of functions, specifically R.V.s to converge to a single function.
We have several interesting & useful notions.

$X_n \xrightarrow[n\to\infty]{a.s.} X \iff (\exists A \subset \Omega)(P(A)=1 \;\&\; \forall \omega \in A \;\; X_n(\omega) \xrightarrow{n\to\infty} X(\omega))$

Pointwise convergence on set of prob 1.

$X_n \xrightarrow[n\to\infty]{P} X \iff (\forall \varepsilon > 0)(P(|X_n - X| > \varepsilon) \xrightarrow[n\to\infty]{} 0)$

$X_n \xrightarrow[n\to\infty]{L^2} X \iff (X_n, X \in L^2)(\;E(X_n - X)^2 \xrightarrow[n\to\infty]{} 0)$

$X_n \xrightarrow[n\to\infty]{L^1} X \iff (X_n, X \in L^1)(\;E|X_n - X| \xrightarrow[n\to\infty]{} 0)$

$X_n \xrightarrow[n\to\infty]{d} X \iff \lim_{n\to\infty} F_{X_n}(t) = F_X(t)$ at all continuity points of F_X.

T: $\iff \forall f$ continuous & bounded $E[f(X_n)] \xrightarrow[n\to\infty]{} E[f(X)]$

RELATIONS BETWEEN CONVERGENCE

T: $\xrightarrow{a.s.} \;\Rightarrow\; \xrightarrow{P} \;\Rightarrow\; \xrightarrow{d}$

$\xrightarrow{L^2} \;\Rightarrow\; \xrightarrow{P}$; $(X_n, X \in L^2)(\xrightarrow{P} \;\Rightarrow\; \xrightarrow{L^2})$

TRANSFORMATIONS

T: $g: \mathbb{R} \to \mathbb{R}$ continuous

- $X_n \xrightarrow{a.s.} X \implies g(X_n) \xrightarrow{a.s.} g(X)$
- $X_n \xrightarrow{P} X \implies g(X_n) \xrightarrow{P} g(X)$
- $X_n \xrightarrow{d} X \implies g(X_n) \xrightarrow{d} g(X)$

CONVERGENCE THMS:

T: Monotone Convergence: $X_n \geq 0$ R.Vs on a common probability space, $X_n \uparrow X \implies E(X_n) \uparrow E(X)$

T: Fatous Lemma: $X_n \geq 0 \implies E(\liminf_{n\to\infty} X_n) \leq \liminf_{n\to\infty} E(X_n)$

T: Dominated Convergence: $(\forall n)(|X_n| \leq Y$ a.s. $\&\; E(Y) < \infty)$

$X_n \xrightarrow[n\to\infty]{a.s.} X \implies \lim_{n\to\infty} E(X_n) = E(X)$

$\{X_n\}$ is iid sequence of $B(p)$ R.Vs

$S_n = \sum_{i=1}^{n} X_i$

T: Weak LLN: $\frac{S_n}{n} \xrightarrow[n\to\infty]{P} p$

T: Strong LLN: $\frac{S_n}{n} \xrightarrow[n\to\infty]{a.s.} p$.

8

Characteristic Functions

For any R.V. X its characteristic function (ChF) is $\varphi_X : \mathbb{R} \to \mathbb{C}$, $\varphi_X(t) = \mathbb{E}(e^{itX})$

Because of this definition the chF always exists, and is finite.

T: $|\varphi_X(t)| \leq 1$ T: $\varphi_X(0) = 1$

T: $Y = aX + b$, $a, b \in \mathbb{R} \Rightarrow \varphi_Y(t) = e^{itb} \varphi_X(at)$.

T: $\overline{\varphi_X(t)} = \varphi_X(-t) = \varphi_{-X}(t)$

T: ChF is real valued $\Longleftrightarrow$ X is symetric $\Longleftrightarrow X \overset{d}{=} -X$.

T: Any ChF is uniformly continuous.

T: X & Y independent $\Longrightarrow \varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t)$

T: $k \in \mathbb{N}$, $E|X|^k < \infty \Longrightarrow \varphi_X(t)$ is k times cont' differentiable & $\mathbb{E}(X^k) = i^{-k} \frac{d^k}{dt^k} \varphi_X(t) \Big|_{t=0}$

T: Inversion: $\int |\varphi_X(t)| dt < \infty \Rightarrow X$ has continuous density

$$f_X(x) = \frac{1}{2\pi} \int e^{-itx} \varphi_X(t) \, dt \, .$$

T: ChF uniquely specify the distribution.

T: $\int |t^k \varphi_X(t)| dt < \infty \Rightarrow X$ has k times diff' continuous density

T: $X_n \xrightarrow{d} X \Longleftrightarrow (\forall t \in \mathbb{R})(\varphi_{X_n}(t) \longrightarrow \varphi_X(t))$

T: $(\forall t \in \mathbb{R})(\varphi_{X_n}(t) \longrightarrow \varphi(t))$ where φ_{X_n} are ChF & $\varphi_X(t)$ is continuous at $0 \Rightarrow \varphi_X$ is chF of some R.V X & $X_n \xrightarrow{d} X$.

Clearly then the chF contains a lot of information about the distribution. This is why we use them because they are compact & have plentiful info.

APPLICATIONS TO STATS

Because of the property that the chF of a sum is the product of chF, they are convenient for proofs in statistics about sums.

T: WLLN: $X_1, X_2, \ldots$ iid $E|X_1| < \infty \Longrightarrow \frac{\sum_{i=1}^{n} X_i}{n} \xrightarrow[n \to \infty]{P} E(X_1)$

T: CLT: Further $E(X_1^2) < \infty$ & $V(X_1) = \sigma^2 > 0 \Longrightarrow \frac{\sum_{i=1}^{n} X_i - nE(X_1)}{\sigma\sqrt{n}} \xrightarrow[n \to \infty]{d} N(0,1)$

We can relax the condition of iid in the above to a more general condition. One example is for $E(X_i) = 0$ (rescale if necissary). Lyapunov Condition: $B_n^2 = V(\sum_{i=1}^{n} X_i)$, then we need $B_n^{-3} \sum_{i=1}^{n} E|X_i|^3 \xrightarrow[n \to \infty]{} 0$

T: Poisson LT: $X_{n1}, \ldots, X_{nn}$ independent R.V.

$P(X_{n,j} = 1) = 1 - P(X_{n,j} = 0) = p_n$, $j = 1, \ldots, n$ & $np_n \longrightarrow \lambda \in (0, \infty)$

$$\Longrightarrow \sum_{i=1}^{n} X_{n,i} \xrightarrow{d} P(\lambda) \, .$$

FOR R.VECTORS

$X = (X_1, \ldots, X_d)$, $t = (t_1, \ldots, t_d) \in \mathbb{R}^d$ then

$\varphi_X : \mathbb{R}^d \longrightarrow \mathbb{C}$, $\varphi_X(t) = \mathbb{E}(e^{i(t,X)}) = \mathbb{E}(\exp[i \sum^d t_j X_j])$ ← Dot product

All key results carry over.

T: $Y = XA + b$, A a $d \times m$ matbix $b \in \mathbb{R}^m$

$\Longrightarrow \varphi_Y(s) = e^{i(s,b)} \varphi_X(sA^T)$

T: $\frac{\partial^{k_1+k_2}}{\partial t_1^{k_1} \partial t_2^{k_2}} \varphi_X(t) = i^{k_1+k_2} \mathbb{E}[X_1^{k_1} X_2^{k_2} e^{i(t,X)}]$ Ensuring we have appropriate finite moments

T: $(\forall b \in \mathbb{R}^d)(\varphi_{(b,X)}(t) = \varphi_X(tb))$ projection of X onto b.

T: WLLN & SLLN

T: CLT: $X_1, X_2, \ldots$ iid R-vect, $E\|X_1\|^2 < \infty$ & C_X^2 exists. ← covariance matrix. $\Longrightarrow \frac{\sum_{i=1}^{n} X_i - \mu n}{\sqrt{n}} \longrightarrow N(0, C_X^2)$

χ^2 testing...?

T: $Z \sim N(0, I_d)$, $b_1, \ldots, b_d$ orthonormal system

$\Longrightarrow Y = ((b_1, Z), \ldots, (b_d, Z)) \sim N(0, I_d)$

9

Distribution Free tests & MLE's

EMPIRICAL DF

For $X_1, \dots, X_n$ an iid sample we know that $S = (X_{(1)}, \dots, X_{(n)})$ is a ss for F, the DF of X_1. The same info is captured in the empirical distribution function

$$F_n^*(t) = \frac{1}{n} \sum_{j=1}^{n} 1\!\!1(X_j \leq t) = \frac{1}{n} \sum_{j=1}^{n} 1\!\!1(X_{(j)} \leq t)$$

The order stats are all the points of discontinuity of F_n^* but we can also find other statistics

$$\bar{X} = \int t \, dF_n^*(t) \quad , \quad s^2 = \frac{1}{n} \sum_{j=1}^{n} (x_j - \bar{x})^2 = \int t^2 dF_n^*(t) - \left(\int t \, dF_n^*(t) \right)^2$$

If there is a parameter $\theta = G(F)$ then we can have a good estimator given by $\theta^* = G(F_n^*)$

T: Glivenko-Cantelli : $X_1, X_2, \dots$ iid , DF F.

$$\Longrightarrow D_n = \sup_t | F_n^*(t) - F(t) | \xrightarrow{a.s.} 0$$

T: For $X_1, X_2, \dots$ iid DF F & $U_1, U_2, \dots \sim U[0,1]$

& $R_n^*(u) = \frac{1}{n} \sum_{i=1}^{n} 1\!\!1(U_j \leq u)$ uniform EDF.

$$\Longrightarrow D_n = \sup_t |F_n^*(t) - F(t)| = \sup_{u \in [0,1]} |R_n^*(u) - u| .$$

Independent of F.

Further $\sqrt{n} \left(R_n^*(u_1) - u_1, \dots, R_n^*(u_d) - u_d \right) \longrightarrow N(0, C^2(u))$

$$C^2(u) = \left[\min\{u_j, u_k\} (1 - \max\{u_j, u_k\}) \right]_{j,k=1,\dots,d}$$

We can use this for

Kolmogorov Test: $\lim_{n \to \infty} P(\sqrt{n} D_n \leq x) = 1 + 2 \sum_{k=1}^{\infty} (-1)^k e^{-2k^2 x^2}$

Mises-Smirnov ω^2-Test: $\omega_n^2 \overset{d}{=} \int_0^1 [\sqrt{n} (R_n^*(u) - u)]^2 du$

$$\lim_{n \to \infty} P(\omega_n^2 \leq x) = P\left(\int_0^1 V^2(u) \, du \leq x \right)$$

with $V(u) \sim N(0, u(1-u))$

MLE's

$X = (X_1, \dots, X_n)$ X_j have density $f_\theta(x)$ Then the MLE of θ is

$$\hat{\theta} = \operatorname*{argmax}_\theta f_\theta(X) = \operatorname*{argmax}_\theta \log(f_\theta(X))$$

T: Gibbs Inequality: f, g densities with respect to μ, on common space $\int f(x) \log(f(x)) \mu(dx) \geq \int f(x) \log(g(x)) \mu(dx)$ when both integrals are finite.

T: $\hat{\theta}_n \xrightarrow{P} \vartheta$ & $\sqrt{n} (\hat{\theta}_n - \vartheta) \xrightarrow{d} N\left(0, \frac{1}{I(\vartheta)}\right)$

True value

Where $$I(\vartheta) = \int \frac{[f_\vartheta'(x)]^2}{f_\vartheta(x)} \mu(dx)$$

T: $$E_\vartheta(\theta_n^* - \vartheta)^2 \geq \frac{1}{n I(\vartheta)}$$