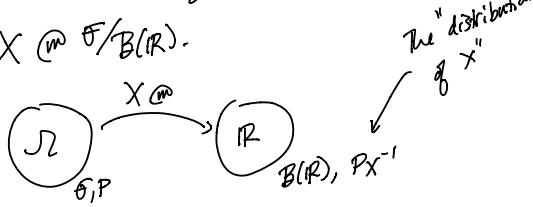


## Lecture 11: Random variable densities and expected values

Recall that random variable is a map  $X: \Omega \rightarrow \mathbb{R}$   
along with a probability measure  $(\Omega, \mathcal{F}, P)$   
s.t.  $X \in \mathcal{B}(\mathbb{R})$ .



Def: The expected value of  $X$ , denoted  $E(X)$ , is defined as

$$E(X) = \int_{\Omega} x dP = \int_{\mathbb{R}} x^+ dP - \int_{\mathbb{R}} x^- dP$$

when it is defined, i.e. when  $X \in \mathcal{Q}(\Omega, \mathcal{F}, P)$ .

You should think of  $X$  as a placeholder for a random number  $X(w)$  obtained by choosing  $w \in \Omega$  at random according to  $P$ . Then  $E(X)$  is essentially what you "expect"  $X$  to be.

e.g. For  $\theta \in [0, 1]$  the r.v.  $X$  has a Bernoulli  $\theta$  distribution, denoted

$$X \sim \text{Ber}(\theta), \text{ if}$$

$$P(X=1) = \theta$$

$$P(X=0) = 1 - \theta.$$

Let  $A = \{w : X(w) = 1\}$  so that

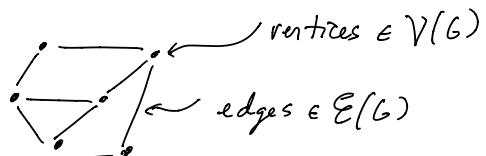
$$X(w) = I_A(w) \quad P\text{-a.e.}$$

$$\begin{aligned} \therefore E(X) &= E(I_A(w)) \text{ by "a.e. useful"} \\ &= P(A) \text{ by def.} \\ &= \theta \quad \text{since } A = \{X=1\}. \end{aligned}$$

Note: It is always the case that (1)

$$E(I_A) = P(A) \text{ whenever } A \in \mathcal{F}.$$

e.g. This example uses expected value & probability to prove a "non-probabilistic" statement about graphs  $G$ :



These types of proofs were made famous by Erdős.

Claim: Every graph  $G$  has a bipartite subgraph  $H$  for which  $\#E(H) \geq \frac{1}{2} \#E(G)$ .

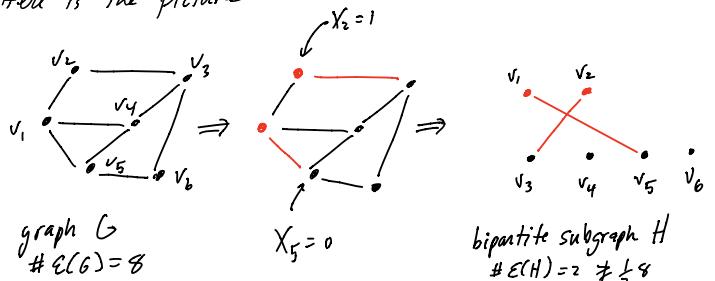
Proof:

Suppose  $G$  has  $n$  vertices labeled  $v_1, v_2, \dots, v_n$ . Let  $X_1, X_2, \dots, X_n$  be  $n$  independent  $\text{Ber}(\frac{1}{2})$  r.v.s defined on some probability space  $(\Omega, \mathcal{F}, P)$ . Define the subgraph  $H$  as follows

$$V(H) := V(G)$$

$$E(H) := \left\{ v_i v_j \in E(G) : (X_i, X_j) = (1, 0) \text{ or } (X_i, X_j) = (0, 1) \right\}$$

Here is the picture



Notice that  $\#E(H)$  is a r.v. in  $\mathcal{Q}_S(\Omega, \mathcal{F})$ .

Let  $\Sigma = \{(i, j) : i > j \text{ and } v_i v_j \in E(G)\}$  index all edges  $E(G)$  so that

Now

$$\begin{aligned} E(\#\mathcal{E}(H)) &= E\left(\sum_{(i,j) \in \mathcal{X}} I_{\{(X_i, X_j) = (1,0)\}} + I_{\{(X_i, X_j) = (0,1)\}}\right) \\ &\stackrel{\text{B.g. 3}}{=} \sum_{(i,j) \in \mathcal{X}} P(X_i=1, X_j=0) + P(X_i=0, X_j=1) \\ &= \sum_{(i,j) \in \mathcal{X}} \left(\frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2}\right) \\ &= \frac{1}{2} \underbrace{\sum_{(i,j) \in \mathcal{X}} 1}_{\#\mathcal{E}(G)} \quad (*) \end{aligned}$$

Now if  $\#\mathcal{E}(\tilde{H}) < \frac{1}{2} \#\mathcal{E}(G)$  for all bipartite subgraphs  $\tilde{H}$  of  $G$  then

$$E(\#\mathcal{E}(H)) < \frac{1}{2} \#\mathcal{E}(G)$$

$\curvearrowleft$  easy to see since  
 $\#\mathcal{E}(H)$  is a simple r.v.

This contradicts  $(*)$  so that we must have  
 $\exists$  a bipartite subgraph  $\tilde{H}$  s.t.

$$\#\mathcal{E}(\tilde{H}) \geq \frac{1}{2} \#\mathcal{E}(G). \quad \underline{\text{QED}}$$

Let's look at a couple fundamental properties of expected value.

### Theorem (Jensen's inequality)

If  $\varphi: \mathbb{R} \rightarrow \mathbb{R}$  is a convex function and  $X \in L_1(\Omega, \mathcal{F}, P)$  then  $\varphi(X)$  is quasi-integrable and

$$\varphi(E(X)) \leq E(\varphi(X)).$$

Proof:

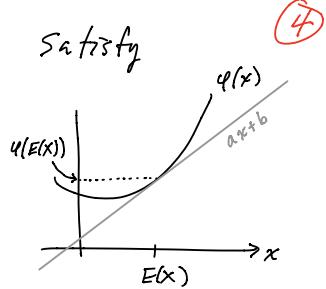
Let  $x \mapsto ax+b$  be a supporting line of  $\varphi$  passing through the point  $(\underbrace{E(X)}, \varphi(\underbrace{E(X)})$ )

$\curvearrowleft$  finite.

(3)

In particular let  $a, b \in \mathbb{R}$  satisfy

$$\begin{cases} ax+b \leq \varphi(x) \quad \forall x \in \mathbb{R} \\ a E(X)+b = \varphi(E(X)) \end{cases}$$



$$\therefore aX+b \leq \varphi(X) \quad (*)$$

Notice that  $aX+b$  is integrable since  $X$  is integrable.

Also  $\varphi$  is convex & mapping into  $\mathbb{R}$   $\Rightarrow \varphi$  is continuous

$$\Rightarrow \varphi \text{ is C}^1$$

$\Rightarrow \varphi(X)$  is a r.v.  
and in  $L^1$  by  $(*)$ .

$$\therefore \underbrace{a E(X)+b}_{\parallel} \leq E(\varphi(X)) \quad \text{by B.g. 3}$$

$\varphi(E(X)) \quad \underline{\text{QED}}$

e.g. Let's use Jensen's inequality to get an understanding of why type of maximal fluctuation we might expect to see if we stare at a large number of random variables

### Theorem (What to expect of the max)

Let  $X_1, X_2, \dots, X_n$  be r.v.s in  $L_1(\Omega, \mathcal{F}, P)$ .

Suppose  $\exists \sigma > 0$  s.t.

$$E(e^{tx_i}) \leq \exp\left(\frac{t^2\sigma^2}{2}\right), \quad \forall t > 0, \forall i \in n$$

$$\text{then } E(\max_{1 \leq i \leq n} X_i) \leq \sigma \sqrt{2 \log n}.$$

Proof:

$$\begin{aligned} \exp(t E(\max_{i \leq n} X_i)) &\stackrel{\text{Tausen}}{\leq} E(\exp(t \max_{i \leq n} X_i)) \\ &= E\left(\max_{i \leq n} \exp(t X_i)\right) \\ &\quad \text{since } e^{tx} \uparrow \text{ in } x \\ &\leq E\left(\sum_{i \leq n} \exp(t X_i)\right) \\ &\quad \text{since these are } \geq 0 \\ &\stackrel{\text{Big 3}}{=} \sum_{i \leq n} E(\exp(t X_i)) \\ &\leq n \exp\left(\frac{t^2 \sigma^2}{2}\right) \text{ by assumption.} \end{aligned}$$

Taking log gives

$$E(\max_{i \leq n} X_i) \leq \underbrace{\frac{\log n}{t} + \frac{t \sigma^2}{2}}_{\text{Now choose a good } t} \quad \forall t > 0.$$

Notice

$$\begin{aligned} \frac{d}{dt} \left( \frac{\log n}{t} + \frac{t \sigma^2}{2} \right) &= -\frac{\log n}{t^2} + \frac{\sigma^2}{2} = 0 \\ \Updownarrow \\ t &= \sqrt{\frac{2 \log n}{\sigma^2}} \end{aligned}$$

Plugging this into our inequality gives

$$\begin{aligned} E(\max_{i \leq n} X_i) &\leq \frac{\sigma}{\sqrt{2 \log n}} \log n + \frac{\sqrt{2 \log n}}{\sigma} \frac{\sigma^2}{2} \\ &= \frac{\sigma}{\sqrt{2}} \sqrt{\log n} + \frac{\sigma}{\sqrt{2}} \sqrt{\log n} \\ &= \sigma \sqrt{2 \log n} \quad \underline{\text{QED}} \end{aligned}$$

(5)

Theorem (expected value factors on indep r.v.s)

Suppose  $X$  and  $Y$  are (possibly extended) independent r.v.s on  $(\Omega, \mathcal{F}, P)$ . If  $X \geq 0$  &  $Y \geq 0$  or  $X, Y \in L_1(\Omega, \mathcal{F}, P)$  then  $XY \in L_1(\Omega, \mathcal{F}, P)$  and  $E(XY) = E(X)E(Y)$ .

Proof:

Case 1: Suppose  $X, Y \in \mathcal{M}_S(\Omega, \mathcal{F}, P)$  so that

$$X = \sum_{i=1}^n a_i I_{A_i} \quad Y = \sum_{j=1}^m b_j I_{B_j}$$

where  $a_1, \dots, a_n$  are distinct and  $A_1, \dots, A_n$  are a disjoint measurable partition of  $\Omega$  (and same for  $b_j$ 's &  $B_j$ 's).

Note that  $A_i$  is indep of  $B_j$  since

$$\begin{aligned} A_i &= \{X = a_i\} \in \sigma\langle X \rangle \quad \text{indep } \sigma\text{-fields} \\ B_j &= \{Y = b_j\} \in \sigma\langle Y \rangle \end{aligned}$$

$$\therefore E(XY) = E\left(\sum_{i,j} a_i b_j I_{A_i \cap B_j}\right)$$

$$= \sum_{i,j} a_i b_j \underbrace{P(A_i \cap B_j)}_{P(A_i)P(B_j)}$$

$$\begin{aligned} \text{Note that } a_i \text{ or } b_j \text{ could} \\ \text{be } \infty \text{ but } \\ \mathcal{M}_S \text{ allows} \\ I_{A_i \cap B_j} \text{ to} \\ \text{Big 3 (1) [a]} \text{ apply} \\ &= \left(\sum_i a_i P(A_i)\right) \left(\sum_j b_j P(B_j)\right) \\ &= E(X)E(Y). \end{aligned}$$

Case 2: Suppose  $X, Y \in \mathcal{M}(\Omega, \mathcal{F})$ .

Notice that we also have that

$$X \in \mathcal{M}(\Omega, \sigma\langle X \rangle) \quad Y \in \mathcal{M}(\Omega, \sigma\langle Y \rangle).$$

(6)

Therefore the structure theorem implies there exists  $X_n \in \mathcal{N}_s(\Omega, \mathcal{F}, P)$  and  $Y_n \in \mathcal{N}_s(\Omega, \mathcal{F}, P)$  such that

$$X_n \uparrow X \text{ and } Y_n \uparrow Y.$$

$$\therefore E(X_n Y_n) = E(X_n) E(Y_n) \text{ by case 1.}$$

$$\begin{aligned} \therefore E(XY) &= E(\lim_n^{\uparrow} X_n Y_n) \\ &= \lim_n^{\uparrow} E(X_n Y_n) \text{ by little 3} \\ &= \lim_n^{\uparrow} E(X_n) E(Y_n) \\ &= E(X) E(Y). \end{aligned}$$

Case 3: Suppose  $X, Y \in L_1(\Omega, \mathcal{F}, P)$ .

Notice that

$$\begin{aligned} (XY)^+ &= X^+ Y^+ + X^- Y^- \\ (XY)^- &= X^+ Y^- + X^- Y^+ \end{aligned}$$

all finite

Therefore

$$\begin{aligned} E(XY)^+ &= E(X^+) E(Y^+) + E(X^-) E(Y^-) < \infty \\ E(XY)^- &= E(X^+) E(Y^-) + E(X^-) E(Y^+) < \infty \end{aligned}$$

by little 3, case 2 and the fact that  $\sigma\langle X^+ \rangle \subset \sigma\langle X \rangle, \dots \text{ & } \sigma\langle Y^- \rangle \subset \sigma\langle Y \rangle$ .

(Notice I'm implicitly using " $X @ \text{w.r.t. } \sigma\langle Y \rangle$ " & "subclasses".)

$\therefore XY \in L_1(\Omega, \mathcal{F}, P)$  and

$$\begin{aligned} E(XY) &= E(XY)^+ - E(XY)^- \\ &= E(X^+) E(Y^+) + E(X^-) E(Y^-) \\ &\quad - E(X^+) E(Y^-) - E(X^-) E(Y^+) \\ &= E(X^+) E(Y) - E(X^-) E(Y) \\ &= E(XY). \end{aligned}$$

QED

(7)

Notice this fully generalizes to situations like this: suppose  $X_1, X_2, \dots$  are independent  $L_1(\Omega, \mathcal{F}, P)$  r.v.s then

$$E(X_1 X_2 \cdots X_n) = E(X_1) E(X_2) \cdots E(X_n)$$

and

$$\begin{aligned} E(g(X_1, X_2, \dots) h(X_2, X_4, \dots)) \\ = E g(X_1, X_3, \dots) E h(X_2, X_4, \dots) \end{aligned}$$

if  $g, h \in \mathcal{N}(\mathbb{R}^\infty, \mathcal{B}(\mathbb{R}^\infty))$ . To show these just use the previous Thm, ANOVA & " $X @ \text{w.r.t. } \sigma\langle Y \rangle$  Thm".

Def: If  $X$  and  $Y$  are two r.v.s on  $(\Omega, \mathcal{F}, P)$

then set

$$\text{var}(X) = \text{"the variance of } X":= E(X - E(X))^2$$

$$\text{sd}(X) = \text{"the standard deviation of } X":= \sqrt{\text{var}(X)}$$

$$\text{cov}(X, Y) = \text{"the covariance b/w } X \text{ & } Y":= E[(X - E(X))(Y - E(Y))]$$

when they are defined.

Note:  $\text{var}(X)$ ,  $\text{sd}(X)$  &  $\text{cov}(X, Y)$  may not be defined if the expectations which define them do not exist.

Remark:  $\text{sd}(X)$  measures how spread out  $X$  is on  $\mathbb{R}$  &  $\text{cov}(X, Y)$  measure how  $X$  &  $Y$  co-vary together.

(9)

Later we will see that  $sd(X)$  and  $\text{cov}(X, Y)$  are essentially the functional analysis equivalent to  $L_2$  norm &  $L_2$  inner product. Here is a hint as to why.

### Theorem (Hölder)

Let  $X$  and  $Y$  be two R.V.s on  $(\Omega, \mathcal{F}, P)$ .

If  $p, q > 0$  s.t.  $\frac{1}{p} + \frac{1}{q} = 1$ , then

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

even if  $X$  and  $Y$  are not quasi-integrable.

If  $XY \in \mathcal{Q}(\Omega, \mathcal{F}, P)$  then

$$|E(XY)| \leq (E|X|^p)^{\frac{1}{p}} (E|Y|^q)^{\frac{1}{q}}. \quad (\ast\ast)$$

Proof:

$(\ast)$  is trivially true if any one of the following is true:

$$E|X|^p = \infty, E|X|^p = 0, E|X|^q = \infty, E|X|^q = 0.$$

since "a.e. useful facts"

So suppose  $E|X|^p, E|Y|^q \in (0, \infty)$ . implies  $|X|^q \stackrel{\text{a.e.}}{=} 0$  so that  $|XY| \stackrel{\text{a.e.}}{=} 0$ .

Define

$$Z := \frac{X}{(E|X|^p)^{\frac{1}{p}}} \quad \& \quad W := \frac{Y}{(E|Y|^q)^{\frac{1}{q}}}$$

Now we simply show

$$E|ZW| \leq 1.$$

Use Young's inequality

$$a^{w_1} b^{w_2} \leq w_1 a + w_2 b \quad (\ast\ast\ast)$$

when  $a, b \geq 0$  &  $w_1, w_2 > 0$  s.t.  $w_1 + w_2 = 1$

Young's inequality follows since log is concave so that

$$w_1 \log a + w_2 \log b \leq \log(w_1 a + w_2 b).$$

(10)

Now  $E|ZW| = E((|Z|^p)^{\frac{1}{p}} (|W|^q)^{\frac{1}{q}})$   
has the form  $(***)$  when  
 $a = (Z(W))^p$ ,  $b = (W(W))^q$ ,  
 $w_1 = \frac{1}{p}$  &  $w_2 = \frac{1}{q}$ .

$$\stackrel{(***)}{\leq} \underbrace{\frac{1}{p} E|Z|^p}_{=1} + \underbrace{\frac{1}{q} E|W|^q}_{=1}, \text{ using Big 3 (2) [1]}$$

$$= 1.$$

If  $XY \in \mathcal{Q}(\Omega, \mathcal{F}, P)$  then

$$|E(XY)| \leq E|XY|$$

by corollary to Big 3.

QED.

### Corollary:

If  $X$  and  $Y$  are two r.v.s on  $(\Omega, \mathcal{F}, P)$  s.t.  $E(X^2) < \infty$  &  $E(Y^2) < \infty$  then  $\text{cov}(X, Y)$ ,  $sd(X)$  &  $sd(Y)$  are well defined and

$$|\text{cov}(X, Y)| \leq sd(X) sd(Y)$$

Proof: By Hölder  $E|X| \leq \sqrt{E|X|^2} < \infty$  so that  $X, Y \in L_1(\Omega, \mathcal{F}, P)$  and  $E(X) < \infty, E(Y) < \infty$ .

Let  $\tilde{X} = X - E(X)$ ,  $\tilde{Y} = Y - E(Y)$ . Also by Hölder

$$\text{we have } E|\tilde{X}\tilde{Y}| \leq \sqrt{E(\tilde{X}^2)} \sqrt{E(\tilde{Y}^2)} < \infty$$

$$\therefore \tilde{X}\tilde{Y} \in \mathcal{Q} \text{ & } |\text{cov}(X, Y)| \leq \sqrt{E(\tilde{X}^2)} \sqrt{E(\tilde{Y}^2)}$$

$$= sd(X) sd(Y) \quad \text{QED}$$

### Corollary:

(11)

If  $X$  and  $Y$  are two independent r.v.s s.t.  $E(X^2) < \infty$  &  $E(Y^2) < \infty$  then  $\text{cov}(X, Y)$ ,  $E(X)$  and  $E(Y)$  are well defined, finite and

$$\begin{aligned}\text{cov}(X, Y) &= E[(X - EX)(Y - EY)] \\ &= \underbrace{E(X - EX)}_{=0} \underbrace{E(Y - EY)}_{=0}\end{aligned}$$

Once we prove the following theorem we will discuss how it gives an interesting relation btwn Lebesgue integration and Riemann integration.

### Theorem:

Suppose  $X$  is a r.v. on  $(\Omega, \mathcal{F}, P)$  s.t.

$X \geq 0$  P-a.e.. Then

$$\begin{aligned}E(X) &= \stackrel{(*)}{=} \int_0^\infty P(X > t) dt \\ &\stackrel{(**)}{=} \int_0^\infty P(X \geq t) dt \quad \text{Lebesgue integral}\end{aligned}$$

### Proof:

First notice that  $t \mapsto P(X > t)$  &  $t \mapsto P(X \geq t)$  are both monotonic and therefore  $\mathcal{B}(\mathbb{R})/\mathcal{B}(\mathbb{R})$ .

Clearly these two functions are in

$\mathcal{L}^1([0, \infty], \mathcal{B}([0, \infty]), \mathbb{P})$  so the integrals in  $(*)$  and  $(**)$  are well defined.

To show  $(*)$  use the "1-2 argument".

Step 1: Show  $(*)$  holds for  $X \in \mathcal{N}_s(\Omega, \mathcal{F})$  (12)

$\therefore X$  can be written in the form

$$X = \sum_{i=1}^n c_i I_{A_i} \quad \begin{matrix} \downarrow \\ \in [0, \infty] \end{matrix} \quad \text{measurable partition}$$

$$\text{so that } E(X) = \sum_{i=1}^n c_i P(A_i).$$

Now

$$\int_0^\infty P(X > t) dt = \int_0^\infty \sum_{i=1}^n P(\{X > t\} \cap A_i) dt$$

since  $A_i$ 's partition  $\Omega$

$$\begin{aligned}&\stackrel{\text{Big 3}}{=} \sum_{i=1}^n \int_0^\infty P(\{X > t\} \cap A_i) dt \\ &\quad \downarrow \text{on } A_i, X = c_i \\ &= \sum_{i=1}^n \int_0^\infty P(\{c_i > t\} \cap A_i) dt \\ &\quad \underbrace{\quad}_{\begin{cases} 0 & \text{if } c_i \leq t \\ P(A_i) & \text{o.w.} \end{cases}} \\ &= \sum_{i=1}^n c_i P(A_i) \\ &= E(X)\end{aligned}$$

Step 2:

If  $X \in \mathcal{N}(\Omega, \mathcal{F})$  then  $\exists X_n \in \mathcal{N}_s(\Omega, \mathcal{F})$  s.t.

$$X_n \uparrow X.$$

$$\therefore E(X) = E(\lim_n^{\uparrow} X_n)$$

$$= \lim_n^{\uparrow} E(X_n), \text{ little 3}$$

$$= \lim_n^{\uparrow} \int_0^\infty P(X_n > t) dt$$

$$= \int_0^\infty \lim_n^{\uparrow} P(X_n > t) dt, \text{ little 3}$$

$$= \int_0^\infty P(X > t) dt$$

by CFB since  $\{X_n > t\} \uparrow \{X > t\}$   
(but not  $\{X_n \geq t\} \uparrow \{X \geq t\}$ )

This gives  $(*)$ .

To show (\*\*) simply notice that (13)

$$P(X \geq t) = P(X > t) + \underbrace{P(X = t)}_{\text{This can only be non-zero for at most countably many } t\text{'s.}}$$

$$= P(X > t) \quad \mathbb{Z}^1\text{-a.e.}$$

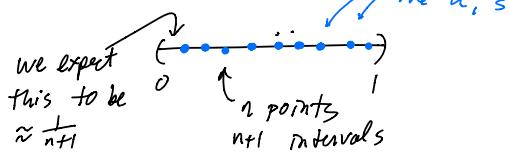
QED

The previous theorem can be useful for computing the expected value of the minimum of a sequence of r.v.

e.g. Let  $U_1, U_2, \dots, U_n$  be iid  $\text{Unif}(0,1)$  random variables. Then

$$\begin{aligned} E(\min_{1 \leq i \leq n} U_i) &= \int_0^\infty P(\underbrace{\min_{1 \leq i \leq n} U_i \geq t}_{\text{This event holds iff each } U_i \geq t}) dt \\ &= \int_0^\infty P(U_1 \geq t, \dots, U_n \geq t) dt \\ &= \int_0^\infty P(U_i \geq t)^n dt \quad \text{since } U_i\text{'s are indep and identically distributed} \\ &= \int_0^1 (1-t)^n dt \\ &= -\frac{(1-t)^{n+1}}{n+1} \Big|_0^1 = \frac{1}{n+1} \end{aligned}$$

In some sense we get the following picture



Using improper Riemann integration to define  $E(X)$ . (14)

To relate the formula  $E(X) = \int_0^\infty P(X > t) dt$  to Riemann integration first notice

Thm (CDF's have countably many jumps)

If  $X$  is a r.v. with cdf  $F(t) := P(X \leq t)$  then  $\{t \in \mathbb{R} : F \text{ is discontinuous at } t\}$  is countable

Proof:

Since  $F(t)$  is right continuous

\begin{aligned} F \text{ is discontinuous at } t &\iff F(t) \neq F(t-) \\ &\iff F(t) - F(t-) > 0 \\ &\iff P(X=t) > 0 \end{aligned}

But  $\{X=t : t \in \mathbb{R}\}$  is a countable collection of disjoint events  $\implies P(X=t) > 0$  for at most countably many  $t \in \mathbb{R}$  (by a Thm in Lecture 5).

QED.

This means that  $t \mapsto P(X > t) = P(X \leq t)$  is Riemann integrable on any bdd interval  $t \in [a, b]$  (since it is bdd and has countably many discontinuities)

Now when  $X \geq 0$

$$E(X) = \int_0^\infty P(X > t) dt$$

abstract integral  
 $\int_X(x) dP(x)$

this can be interpreted as a improper Riemann integral on  $[0, \infty]$ . ... this can be used to define  $E(X)$  with just improper Riemann integration!

## Probability inequalities involving expected value

(15)

It is often the case that computing bounds for expected value is easy. These inequalities then yield probability bounds

### Theorem (Markov's inequality)

If  $X$  is a r.v. on  $(\Omega, \mathcal{F}, P)$  s.t.  $X \geq 0$  P-a.e.

then  $P(X \geq \alpha) \leq \frac{E(X)}{\alpha}$  for all  $\alpha > 0$ .

Proof:  $X \geq 0$  P-a.e. implies that  
 $X \in \mathbb{Q}^+(\Omega, \mathcal{F}, P)$  and  $X I_{\{X \geq \alpha\}} \geq \alpha I_{\{X \geq \alpha\}}$

$$\begin{aligned} \therefore E(X) &\geq E(X I_{\{X \geq \alpha\}}) \quad \text{by Big 3(i)} \text{ and } X \geq 0 \text{ P-a.e.} \\ &\geq E(\alpha I_{\{X \geq \alpha\}}) \quad \text{by Big 3(ii)} \\ &= \alpha P(X \geq \alpha) \quad \text{definition of } \int_X dP \end{aligned}$$

QED.

### Corollary:

For any r.v.  $X$  on  $(\Omega, \mathcal{F}, P)$  and any  $\alpha > 0$ :

i)  $P(|X| \geq \alpha) \leq \frac{E(|X|^k)}{\alpha^k}$  Chernoff method

ii)  $P(X \geq \alpha) \leq \inf_{t > 0} \frac{E(e^{tX})}{e^{t\alpha}}$

iii) if  $E(X^2) < \infty$  then  $E(X) < \infty$  and Chebyshev's inequality

$$P(|X - EX| \geq \alpha) \leq \frac{\text{var}(X)}{\alpha^2}$$

Note: i) says that if  $E|X|^k < \infty$  for a large  $k > 0$  then  $P(|X| \geq \alpha)$  decays quickly as  $\alpha \rightarrow \infty$ . We used ii) in Lecture 1 when illustrating the "modern" way to prove Borel's Normal number theorem.

iii) shows why  $\text{sd}(X)$  controls the "spread" of  $X$ .

Lets use Chebyshev to show a famous theorem in analysis.

1.9

### Theorem (Weierstrass Approximation)

If  $f: [0,1] \rightarrow \mathbb{R}$  is continuous then for any  $\varepsilon > 0$  there exists a polynomial  $p(x)$  s.t.

$$\sup_{x \in [0,1]} |f(x) - p(x)| < \varepsilon.$$

Proof: Let  $U_1, U_2, \dots$  be iid r.v.s uniformly distributed on  $[0,1]$ . For each  $x \in [0,1]$

let  $F_x(\cdot)$  be the C.d.f. of a  $\text{Ber}(x)$  r.v. and set

$$S_n^x := F_x^{-1}(U_1) + \dots + F_x^{-1}(U_n)$$

independent  $\text{Ber}(x)$  r.v.s

A simple counting argument shows

$$S_n^x \sim \text{Bin}(n, x)$$

$$P(S_n^x = m) = \binom{n}{m} x^m (1-x)^{n-m}$$

for  $m = 0, 1, \dots, n$ . Moreover,

$$E\left(\frac{S_n^x}{n}\right) = x \quad \text{and} \quad \text{var}\left(\frac{S_n^x}{n}\right) = \frac{x(1-x)}{n}.$$

Also notice that

$$P_n(x) := E f\left(\frac{S_n^x}{n}\right) = \underbrace{\sum_{m=0}^n f\left(\frac{m}{n}\right) P(S_n^x = m)}_{\text{simple r.v.}} \underbrace{\text{Polynomial in } x \text{ of degree } n.}_{\text{of}}$$

Since  $f$  is uniformly continuous on  $[0,1]$  let

$$M := \sup_{x \in [0,1]} |f(x)| < \infty$$

$$S(\varepsilon) := \sup \{|f(x) - f(y)| : |x - y| < \varepsilon\}$$

and for a given  $\varepsilon > 0$  and  $x \in [0,1]$  let

$$A_{\varepsilon, x} := \left\{ w \in \mathbb{Z} : \left| \frac{S_n^x(w)}{n} - x \right| < \varepsilon \right\}$$

The definition of  $A_{\varepsilon,x}$  now implies (17)

$$w \in A_{\varepsilon,x} \Rightarrow \left| f\left(\frac{S_n^x(n)}{n}\right) - f(x) \right| \leq \delta(\varepsilon)$$

$$w \in A_{\varepsilon,x}^c \Rightarrow \left| f\left(\frac{S_n^x(n)}{n}\right) - f(x) \right| \leq 2M$$

so that

$$\begin{aligned} \left| f\left(\frac{S_n^x(n)}{n}\right) - f(x) \right| &= \left| f\left(\frac{S_n^x(n)}{n}\right) - f(x) \right| \left( I_{A_{\varepsilon,x}(n)} + I_{A_{\varepsilon,x}^c(n)} \right) \\ &\leq \delta(\varepsilon) I_{A_{\varepsilon,x}(n)} + 2M I_{A_{\varepsilon,x}^c(n)} \end{aligned}$$

To finish

$$\begin{aligned} |P_n(x) - f(x)| &= \left| E f\left(\frac{S_n^x}{n}\right) - f(x) \right| \\ &\leq E \left| f\left(\frac{S_n^x}{n}\right) - f(x) \right| \text{ by Jensen} \\ &\leq \delta(\varepsilon) P\left(\left|\frac{S_n^x}{n} - x\right| < \varepsilon\right) + 2M P\left(\left|\frac{S_n^x}{n} - x\right| \geq \varepsilon\right) \\ &\leq \delta(\varepsilon) + 2M \frac{\text{Var}\left(\frac{S_n^x}{n}\right)}{\varepsilon^2} \text{ by Chebyshev} \\ &= \delta(\varepsilon) + 2M \frac{x(1-x)}{n\varepsilon^2} \end{aligned}$$

Since  $x(1-x) \leq \frac{1}{4}$   $\forall x \in [0,1]$  we have

$$\sup_{x \in [0,1]} |P_n(x) - f(x)| \leq \delta(\varepsilon) + \frac{M}{2n\varepsilon^2} \quad (*)$$

By replacing  $\varepsilon$  with  $\frac{1}{n^{1/2}}$  (for example) and choosing  $n$  large enough I can make  $(*)$  as small as I want (since  $\delta(\varepsilon) \rightarrow 0$  as  $\varepsilon \rightarrow 0$ ).

QED

(18)  
Chernoff's method is especially useful since  $t \mapsto E(e^{tX})$  is an important function called the moment generating function. More on that later.

Lets look at a special case which will give us the SLLN for iid r.v.s.

Theorem (Hoeffding's inequality)

Let  $X_1, X_2, \dots, X_n$  be iid r.v.s on  $(\Omega, \mathcal{F}, P)$ . If there exists finite real numbers  $a \leq b$  s.t.

$a \leq X_i \leq b$   
P-a.e.  $\forall i = 1, \dots, n$ , then

$$P\left(\left|\frac{S_n}{n} - \mu\right| \geq \varepsilon\right) \leq 2 \exp\left(-\frac{2n\varepsilon^2}{(b-a)^2}\right)$$

$\forall \varepsilon > 0$  where  $S_n := X_1 + \dots + X_n$  and  $\mu := E(X_1)$ .

Note: The iid assumption in Hoeffding can be relaxed and extended to martingales. We will cover this next quarter when we study dependence in random variables.

Note: The Hoeffding bound gives the exact same bound we derived by hand for our rademacher coin flip r.v.s  $R_1, R_2, \dots$  from lecture I:

$$P\left(\left|\frac{S_n}{n}\right| \geq \varepsilon\right) \leq 2e^{-n\varepsilon^2/2}$$

where  $\frac{S_n}{n} = \frac{R_1 + \dots + R_n}{n}$  and  $-1 \leq R_i \leq 1$ .

Lemme: If  $X$  is a r.v. on  $(\Omega, \mathcal{F}, P)$   
s.t.  $\exists$  finite  $a < b$  s.t.  $a \leq X \leq b$  &  $E(X) = 0$   
Then  $E(e^{tX}) \leq e^{t^2(b-a)^2/8} \quad \forall t \geq 0$ .

Proof:

Let  $w = \frac{x-a}{b-a}$  so that  $0 \leq w \leq 1$  and  
 $X = wb + (1-w)a$ . By convexity we have

$$e^{tX} \leq we^{tb} + (1-w)e^{ta}$$

$$\begin{aligned} \therefore E(e^{tX}) &\leq E(w)e^{tb} + E(1-w)e^{ta} \\ &= \frac{E(X)-a}{b-a}e^{tb} + \frac{b-E(X)}{b-a}e^{ta} \\ &= -\frac{a}{b-a}e^{tb} + \frac{b}{b-a}e^{ta} \\ &= (1-\theta)e^{-u\theta} + \theta e^{u(1-\theta)} \end{aligned}$$

$$\text{where } u = -(b-a)t \text{ & } \theta = \frac{b}{b-a}.$$

Notice the assumption  $E(X) = 0$  implies  
 $\theta \in [a, b]$  so that  $0 \leq \theta \leq 1$ .

If we can show

$$(1-\theta)e^{-u\theta} + \theta e^{u(1-\theta)} \leq e^{u^2/8} \quad (*)$$

✓  $u \in \mathbb{R}$  and  $\forall \theta \in [0, 1]$  we are done.

Taking log it is sufficient to show

$$\log((1-\theta)e^{-u\theta} + \theta e^{u(1-\theta)}) \leq u^2/8$$

||

$$\log(e^{-u\theta}[1-\theta + \theta e^u])$$

||

$$-u\theta + \log(1-\theta + \theta e^u)$$

||

$$K(u)$$

$$\text{Now } K'(u) = -\theta + \frac{\theta e^u}{1-\theta + \theta e^u} = -\theta + \frac{\theta}{\theta + (1-\theta)e^{-u}}$$

$$K''(u) = \frac{\theta(1-\theta)e^{-u}}{(\theta + (1-\theta)e^{-u})^2}$$

(19)

Now Taylor's thm gives

$$\begin{aligned} K(u) &= K(0) + uK'(0) + \frac{u^2}{2}K''(u^*) \quad u^* \in [0, u] \\ &= 0 + 0 + \frac{u^2}{2} \left( \underbrace{\frac{\theta}{\theta + (1-\theta)e^{-u^*}}}_{\in [0, 1]} \right) \left( 1 - \underbrace{\frac{\theta}{\theta + (1-\theta)e^{-u^*}}}_{\in [0, 1]} \right) \\ &\leq \frac{1}{4} \end{aligned}$$

$$\therefore K(u) \leq \frac{u^2}{8}$$

QED.

(20)

Proof of Hoeffding's inequality:

We can suppose w.l.g that  $E(X_i) = \mu = 0$ .

Now

$$\begin{aligned} P\left(\left|\frac{S_n}{n}\right| \geq \varepsilon\right) &= P\left(\left\{\frac{S_n}{n} \geq \varepsilon\right\} \cup \left\{-\frac{S_n}{n} \geq \varepsilon\right\}\right) \\ &= P\left(\frac{S_n}{n} \geq \varepsilon\right) + P\left(-\frac{S_n}{n} \geq \varepsilon\right) \end{aligned}$$

using Chernoff's method for any  $t > 0$

$$P\left(\frac{S_n}{n} \geq \varepsilon\right) \leq P\left(e^{tS_n} \geq e^{t\varepsilon n}\right)$$

$$\leq \frac{E(e^{tS_n})}{e^{tn\varepsilon}} \quad \text{by Markov Ineq.}$$

$$= e^{-tn\varepsilon} \prod_{i=1}^n E(e^{tX_i}) \quad \text{by indep}$$

$$\leq e^{t^2(b-a)^2/8} \quad \text{by lemma}$$

$$\leq e^{-tn\varepsilon} e^{nt^2(b-a)^2/8}$$

minimized at  $t = \frac{4\varepsilon}{(b-a)^2}$

$$\begin{aligned} \therefore P\left(\frac{S_n}{n} \geq \varepsilon\right) &\leq e^{-4n\varepsilon^2/(b-a)^2} e^{n4^2\varepsilon^2/8(b-a)^2} \\ &= e^{-2n\varepsilon^2/(b-a)^2} \end{aligned}$$

Similar arguments give the exact same upper bound for  $P\left(-\frac{S_n}{n} \geq \varepsilon\right)$ .

QED

Here is an example of the utility of Hoeffding.

e.g.

Theorem: (SLLN for bounded r.v.s)

Let  $X_1, X_2, \dots$  be iid r.v.s on  $(\Omega, \mathcal{F}, P)$  s.t.  $|X_i| \leq c$  for some finite  $c$ . Then

$$\frac{S_n}{n} \xrightarrow{n \rightarrow \infty} E(X_1) \text{ P-a.e.}$$

where  $S_n = X_1 + \dots + X_n$ .

Proof:

By Hoeffding's inequality

$$P\left(\left|\frac{S_n}{n} - E(X_1)\right| \geq \varepsilon\right) \leq 2 \exp\left(-\frac{2n\varepsilon^2}{(2c)^2}\right) \quad \forall \varepsilon > 0.$$

underbrace  
finitely summable over n  
for each  $\varepsilon > 0$ .

∴ The First Borel-Cantelli lemma applies so

$$\text{that } P\left(\bigcap_{\varepsilon \in \mathbb{Q}^+} \left\{\left|\frac{S_n}{n} - E(X_1)\right| \geq \varepsilon \text{ a.a.}\right\}\right) = 1$$

$$\therefore P\left(\frac{S_n}{n} \rightarrow E(X_1)\right) = 1.$$

QED.

Notice that this tells us we have the right definition for  $E(X) := \int_X dP$ , i.e.  $E(X)$  tells us the long run average of independent samples of  $X$  (with probability 1).

After we cover densities, later in this lecture, we will discuss

The Glivenko-Cantelli Theorem

which is an important application of the SLLN for bdd r.v.s.

(21)

In the Law of the iterated log we needed lower bounds to get tight control on the behavior of  $P(S_n \geq \alpha)$ . Let's look at one example of a lower bound.

(22)

Theorem: (Paley-Zygmund Ineq.)

If  $X$  is a non-negative r.v. s.t.  $E(X^2) < \infty$  then

$$(1-\alpha)^2 \frac{(EX)^2}{E(X^2)} \leq P(X \geq \alpha EX)$$

for all  $\alpha \in (0, 1)$ .

Proof: First notice that

$$X = X I_{\{X \leq \alpha EX\}} + X I_{\{X \geq \alpha EX\}}.$$

Taking expected values gives

$$E(X) \leq \alpha E(X) + E(X I_{\{X \geq \alpha EX\}})$$

$$\leq \alpha E(X) + \sqrt{E(X^2)} \sqrt{E(I_{\{X \geq \alpha EX\}}^2)}$$

by Hölder

$$= \alpha E(X) + \sqrt{E(X^2)} \sqrt{P(X \geq \alpha EX)}$$

Now since  $E(X), E(X^2)$  are both finite ( $E(X) < \infty$  by Hölder) we can shuffle terms around to get the result.

QED.

Note: Jensen's inequality shows  $(EX)^2 \leq E(X^2)$

$$\therefore \text{the lower bound } (1-\alpha)^2 \frac{(EX)^2}{E(X^2)} \leq 1$$

## Maximal Inequalities

(23)

We used two maximal inequalities in lecture 7 which were custom to our coinflip model. In this section we state general versions and use them later (after we get the cLT) to establish Kolmogorov's 3 series theorem.

### Theorem (Kolmogorov's Maximal inequality)

Let  $X_1, \dots, X_n$  be independent r.v.s s.t.  $E(X_k^2) < \infty$  and  $E(X_k) = 0$  f.p. If  $\alpha > 0$  then

$$P\left(\max_{1 \leq k \leq n} |S_k| \geq \alpha\right) \leq \frac{1}{\alpha^2} \underbrace{\text{var}(S_n)}_{= E(S_n^2)}$$

where  $S_n := X_1 + \dots + X_n$ .

Proof: The proof is exactly similar to the one we did for coin flips in lecture 7. Let's do it again using our theory of integration.

Define  $F_k := \{|S_1| < \alpha, \dots, |S_{k-1}| < \alpha, |S_k| \geq \alpha\}$ .

$$\begin{aligned} E(S_n^2) &= \int_{\Omega} S_n^2 dP \\ &\geq \int_{\Omega} S_n^2 \sum_{k=1}^n I_{F_k} dP \quad \text{since } F_k \text{'s are disjoint so } \sum_{k=1}^n I_{F_k} \leq 1. \\ &= \sum_{k=1}^n \int_{\Omega} S_n^2 I_{F_k} dP \quad S_n^2 = (S_n - S_k)^2 \\ &= \sum_{k=1}^n \int_{\Omega} [S_k^2 + 2S_k(S_n - S_k) + (S_n - S_k)^2] I_{F_k} dP \\ &\geq \sum_{k=1}^n \int_{\Omega} S_k^2 I_{F_k} dP + 2 \int_{\Omega} S_k I_{F_k} (S_n - S_k) dP \end{aligned}$$

Notice  $S_k \in \sigma(X_1, \dots, X_k)$  since  $S_k$  is a measurable function of  $X_1, \dots, X_k$ . Also  $I_{F_k} \in \sigma(X_1, \dots, X_k)$  since  $F_k \in \sigma(X_1, \dots, X_k)$ . Therefore  $S_k I_{F_k} \in \sigma(X_1, \dots, X_k)$  so that  $\sigma(S_k I_{F_k}) \subset \sigma(X_1, \dots, X_k)$

(24)

and similarly

$$\sigma(S_n - S_k) \subset \sigma(X_{k+1}, \dots, X_n)$$

Since the  $X_k$ 's are indep we have

$$\begin{aligned} E(S_k I_{F_k} (S_n - S_k)) &= E(S_k I_{F_k}) \underbrace{E(S_n - S_k)}_{=0} \\ &= 0 \\ \therefore E(S_n^2) &\geq \int_{\Omega} \sum_{k=1}^n S_k^2 I_{F_k} dP \quad \text{on } |S_k| \geq \alpha \text{ on } F_k \\ &\geq \alpha^2 \sum_{k=1}^n \underbrace{I_{F_k}}_{\substack{\text{on } |S_k| \geq \alpha \\ \text{on } F_k}} dP \quad \rightarrow I_{\{\max_{1 \leq k \leq n} |S_k| \geq \alpha\}} \\ &\geq \alpha^2 P\left(\max_{1 \leq k \leq n} |S_k| \geq \alpha\right) \end{aligned}$$

Q.E.D.

Here is Etemadi's neg which can be used without the moment assumptions on  $X_k$  in Kolmogorov's max neg.

### Theorem (Etemadi's maximal inequality)

Suppose  $X_1, X_2, \dots, X_n$  are independent r.v.s and  $\alpha > 0$ . Then

$$P\left(\max_{1 \leq k \leq n} |S_k| \geq 3\alpha\right) \leq 3 \max_{1 \leq k \leq n} P(|S_k| \geq \alpha)$$

Proof: The proof is similar to what we did for the coin flips.

Let  $F_k := \{|S_1| < 3\alpha, \dots, |S_{k-1}| < 3\alpha, |S_k| \geq 3\alpha\}$

Notice that the  $F_k$ 's are disjoint, (25)

$$\bigcup_{k=1}^n F_k = \left\{ \max_{1 \leq k \leq n} |S_k| \geq 3\alpha \right\} \text{ and}$$

$$\begin{aligned} w \in F_k \cap \{|S_n| < \alpha\} &\Rightarrow |S_k(w)| \geq 3\alpha \text{ and } |S_n(w)| < \alpha \\ &\stackrel{(*)}{\Rightarrow} |S_n(w) - S_k(w)| > 2\alpha \\ &\quad \text{and } w \in F_k \end{aligned}$$

Now

$$\begin{aligned} P\left(\max_{1 \leq k \leq n} |S_k| \geq 3\alpha\right) &= P\left(\max_{1 \leq k \leq n} |S_k| \geq 3\alpha, |S_n| < \alpha\right) \\ &\quad + P\left(\max_{1 \leq k \leq n} |S_k| \geq 3\alpha, |S_n| \geq \alpha\right) \\ &\leq P(|S_n| \geq \alpha) + \sum_{k=1}^{n-1} P(F_k \cap \{|S_n| < \alpha\}) \\ &\quad \text{drop } k=n \text{ term since } F_n \cap \{|S_n| < \alpha\} = \emptyset \\ &\leq P(|S_n| \geq \alpha) + \sum_{k=1}^{n-1} P(F_k \cap \{|S_n - S_k| > 2\alpha\}) \\ &\quad \text{by } (*) \\ &= P(|S_n| \geq \alpha) + \sum_{k=1}^{n-1} P(F_k) P(|S_n - S_k| > 2\alpha) \\ &\quad \text{since } F_k \subset \sigma(X_1, \dots, X_k) \text{ and } \{|S_n - S_k| > 2\alpha\} \subset \sigma(X_{k+1}, \dots, X_n) \\ &\leq P(|S_n| \geq \alpha) + \max_{1 \leq k \leq n} P(|S_n - S_k| > 2\alpha) \\ &\leq P(|S_n| \geq \alpha) + P(|S_n| \geq \alpha) \\ &\leq 3 \max_{1 \leq k \leq n} P(|S_k| \geq \alpha). \end{aligned}$$

QED.

## Characterizing and Constructing probability measures with densities (26)

In undergraduate probability we are taught that probability densities characterize "continuous r.v.s" & probability mass functions characterize "discrete r.v.s".

e.g.-

$$\text{if } P(X \in B) = \int_B e^{-x} I_{(0, \infty)}(x) dx$$

then this is the density

but if

$$P(X \in B) = \sum_{k \in B} \binom{n}{k} \left(\frac{1}{3}\right)^k \left(\frac{2}{3}\right)^{n-k}$$

then this is the probability mass function

It was annoying to me that you had to figure something out about  $X$ , i.e. continuous or discrete, before trying to compute probabilities. Our general integration theory unifies both cases and gives an accessible method for characterizing and constructing Prob measures.

For the rest of this section let  $(\Omega, \mathcal{F}, \mu)$  denote a measure space (unless stated otherwise).

Def: If  $f \in \mathcal{Q}(\Omega, \mathcal{F}, \mu)$  then the set function  $\int f d\mu : \mathcal{F} \rightarrow \bar{\mathbb{R}}$  mapping

$$A \in \mathcal{F} \mapsto \int_A f d\mu := \int_{\Omega} f I_A d\mu$$

is called the indefinite integral of  $f$  with respect to  $\mu$ .

Theorem ( $\int f d\mu$  is  $\sigma$ -additive)

If  $f \in \mathcal{Q}(\Omega, \mathcal{F}, \mu)$  then  $\int f d\mu$  is countably additive over disjoint  $\mathcal{F}$ -sets.

Proof: Let  $F_1, F_2, \dots$  be disjoint  $\mathcal{F}$ -sets.

Use the "2-3" argument.

Step 2: Suppose  $f \in \mathcal{N}(\Omega, \mathcal{F})$ .

$$\begin{aligned} \int_{V_{\mathcal{F}_k}} f d\mu &= \int_{\Omega} \sum_{n=1}^{\infty} f I_{F_n} d\mu \\ &\quad \text{since } F_n \text{'s disjoint} \\ &= \int_{\Omega} \limsup_n \sum_{k=1}^n f I_{F_k} d\mu \\ &\quad \text{since these are } \geq 0 \\ &\stackrel{\text{Big 3(2)}}{=} \limsup_n \int_{\Omega} \sum_{k=1}^n f I_{F_k} d\mu \\ &\stackrel{\text{Big 3(3)}}{=} \limsup_n \sum_{k=1}^n \int_{\Omega} f I_{F_k} d\mu \\ &= \sum_{k=1}^{\infty} \int_{F_k} f d\mu \end{aligned}$$

$\therefore$  the Thm holds over  $\mathcal{N}(\Omega, \mathcal{F})$ .

Step 3: Suppose  $f \in \mathcal{Q}(\Omega, \mathcal{F}, \mu)$ .

Certainly  $f I_{V_{\mathcal{F}_k}} \in \mathcal{Q}(\Omega, \mathcal{F}, \mu)$  since

$$(f I_{V_{\mathcal{F}_k}})^{\pm} = f^{\pm} I_{V_{\mathcal{F}_k}} \leq f^{\pm}$$

(27)

$$\begin{aligned} \therefore \int f d\mu &\stackrel{\text{def}}{=} \int_{V_{\mathcal{F}_k}} f^+ d\mu - \int_{V_{\mathcal{F}_k}} f^- d\mu \\ &= \sum_{k=1}^{\infty} \int_{F_k} f^+ d\mu - \sum_{k=1}^{\infty} \int_{F_k} f^- d\mu \\ &\quad \text{one of these is finite.} \\ &\quad \& \text{Both are defined} \\ &= \sum_{k=1}^{\infty} \left[ \int_{F_k} f^+ d\mu - \int_{F_k} f^- d\mu \right] \\ &\quad \text{by Big 3 for counting measure.} \\ &= \sum_{k=1}^{\infty} \int_{F_k} f d\mu. \end{aligned}$$

QED.

Corollary:

i) If  $f \in \mathcal{N}(\Omega, \mathcal{F})$  then

$\int f d\mu$  is a measure on  $(\Omega, \mathcal{F})$

ii) If  $f \in \mathcal{N}(\Omega, \mathcal{F})$  &  $\int f d\mu = 1$  then

$\int f d\mu$  is a probability measure on  $(\Omega, \mathcal{F})$ .

Definition:

For any two measures  $\mu, \nu$  on  $(\Omega, \mathcal{F})$  if  $\exists s \in \mathcal{N}(\Omega, \mathcal{F})$  s.t.

$$\nu(A) = \int_A s d\mu, \quad \forall A \in \mathcal{F}$$

then  $s$  is a density of  $\nu$  with respect to  $\mu$ .

Here is an important question:

Are densities unique?

The next example shows the answer must be

Not without assumptions

e.g. Let  $\Omega = \mathbb{R}$

$$\mathcal{F} = \{\emptyset, \Omega, (-\infty, \delta), [\delta, \infty)\}$$

$$\therefore \mathcal{L}'(A) = \int_A \frac{1}{\delta} d\mathcal{L}' = \int_A \frac{2}{\delta} d\mathcal{L}' \quad \forall A \in \mathcal{F}$$

these are both densities w.r.t  $\mathcal{L}'$  resulting in the same measure.

The following theorem gives sufficient conditions for uniqueness (as a corollary)

Thm: Suppose  $f, g \in Q(\Omega, \mathcal{F}, \mu)$ .

If  $f \in L_1(\Omega, \mathcal{F}, \mu)$  or  $g \in L_1(\Omega, \mathcal{F}, \mu)$  or  $\mu$  is  $\sigma$ -finite then

$$\int f d\mu \leq \int g d\mu \text{ on } \mathcal{F} \iff f \leq g \text{ } \mu\text{-a.e.}$$

Proof:

$\Leftarrow$ : Follows directly from Big 3 [1]

$\Rightarrow$ :

Case 1: Suppose  $f \in L_1(\Omega, \mathcal{F}, \mu)$  or  $g \in L_1(\Omega, \mathcal{F}, \mu)$ . We will show  $\mu(f > g) = 0$  by the "indicate what you want trick".

$$f I_{\{f > g\}} \geq g I_{\{f > g\}}$$

$$\Rightarrow \int f d\mu \geq \int g d\mu \text{ by Big 3 [1].}$$

$$\Rightarrow \int f d\mu = \int g d\mu \text{ since } \int f d\mu \leq \int g d\mu \text{ on } \mathcal{F}.$$

$$\Rightarrow \int (f-g) I_{\{f > g\}} d\mu = 0 \text{ by Big 3 [2] since } f \in L_1 \text{ or } g \in L_1.$$

$$\Rightarrow (f-g) I_{\{f > g\}} = 0 \text{ } \mu\text{-a.e. by a.e. useful facts}$$

$$\Rightarrow \mu(f > g) = 0$$

$$\hookrightarrow \text{since } f(\omega) > g(\omega) \Rightarrow (f(\omega) - g(\omega)) I_{\{f > g\}(\omega)} = 1$$

(2a)

Case 2: Suppose  $\mu$  is finite.

(30)

First notice that

$$f \leq g \text{ on } \{f = \infty\} \cap \{g = \infty\} \quad (1)$$

$$f \leq g \text{ on } \{f = -\infty\} \cap \{g = \infty\} \quad (2)$$

$$f \leq g \text{ on } \{f = -\infty\} \cap \{g = -\infty\} \quad (3)$$

We also have

$$\mu(\{f = \infty\} \cap \{g = -\infty\}) = 0 \quad (4)$$

otherwise it would contradict the assumption

$$\int f d\mu \leq \int g d\mu. \text{ Now we just show}$$

$$f \leq g \text{ } \mu\text{-a.e. on } \{|f| < \infty\} \cup \{|g| < \infty\}. \quad (5)$$

Let  $A_n := \{|f| < n\}$ . Since  $\mu$  is a finite measure

$$f I_{A_n} \in L_1 \text{ & } g I_{A_n} \in Q.$$

Also

$$\int f I_{A_n} d\mu = \int_{A_n} f d\mu \stackrel{\text{By assumption}}{\leq} \int_{A_n} g d\mu = \int g I_{A_n} d\mu$$

on  $\mathcal{F}$ . Since  $f I_{A_n} \in L_1$ , case 1 implies

$$f I_{A_n} \leq g I_{A_n} \text{ } \mu\text{-a.e.}$$

$$\therefore f \leq g \text{ } \mu\text{-a.e. on } A_n := \{|f| < n\}$$

$$\therefore f \leq g \text{ } \mu\text{-a.e. on } \{|f| < \infty\} = \bigcup_{n=1}^{\infty} \{|f| < n\}.$$

A similar argument shows

$$f \leq g \text{ } \mu\text{-a.e. on } \{|g| < \infty\}.$$

Now the union of (1)-(5) gives

$$f \leq g \text{ } \mu\text{-a.e.}$$

Case 3: Suppose  $\mu$  is  $\sigma$ -finite.

Let  $F_k \in \mathcal{F}$  s.t.  $\mu(F_k) < \infty$  and  $\bigcup_{k=1}^{\infty} F_k = \Omega$ .

$$\therefore \mu(f > g) \leq \sum_{k=1}^{\infty} \underbrace{\mu(\{f > g\} \cap F_k)}_{= \mu_k(f > g) \text{ where } \mu_k(\cdot) := \mu(\cdot \cap F_k)} \quad (*)$$

$$\mu_k(\cdot) := \mu(\cdot \cap F_k)$$

Now case 2 applies to the finite measure  $\mu_\mu$  since (31)

$$\int f d\mu_\mu = \int f I_{F_\mu} d\mu \quad \text{by a "1-2-3" argument}$$

$$= \int f d\mu$$

$$\bullet_{\Omega F_\mu}$$

$$\text{by assumption} \quad \leq \int_{\Omega F_\mu} g d\mu = \int g d\mu_\mu$$

$\therefore$  case 2 implies  $\mu_\mu(f > g) = 0$ .

$\therefore \mu(f > g) = 0$  by (\*). QED

### Corollary (Uniqueness of densities)

Let  $f, g \in \mathcal{Q}(\Omega, \mathcal{F}, \mu)$ . If  $f$  or  $g$  is integrable or  $\mu$  is  $\sigma$ -finite then

$$\int f d\mu = \int g d\mu \text{ on } \mathcal{F} \Leftrightarrow f = g \mu\text{-a.e.}$$

Note: If  $P$  is a probability measure &  $\mu$  is a measure (over  $(\Omega, \mathcal{F})$ ) then

$$\text{i)} P(\cdot) = \int s d\mu \Rightarrow \text{both } \int s^+ d\mu < \infty \text{ and}$$

$$\int s^- d\mu < \infty.$$

$$\Rightarrow s \in L_1(\Omega, \mathcal{F}, \mu)$$

$$\Rightarrow s \text{ is unique } \mu\text{-a.e.}$$

$$\text{ii)} \mu(\cdot) = \int s dP \Rightarrow s \text{ is unique } P\text{-a.e.}$$

since  $P$  is  $\sigma$ -finite.

The next theorem shows how to compute (32)  
 $\int f d\nu$  when  $\nu$  has density  $s$  w.r.t.  $\mu$ .

### Theorem (Slap in the density: $d\nu = s d\mu$ )

Let  $\nu$  and  $\mu$  be measures on  $(\Omega, \mathcal{F})$  where  $\nu$  has density  $s$  w.r.t.  $\mu$ .

Then

$$f \in \mathcal{Q}^+(\Omega, \mathcal{F}, \nu) \Leftrightarrow fs \in \mathcal{Q}^+(\Omega, \mathcal{F}, \mu)$$

and either one implies

$$\int f d\nu = \int fs d\mu \quad \begin{matrix} \text{i.e. } d\nu = s d\mu \\ \text{i.e. } s = \frac{d\nu}{d\mu} \end{matrix}$$

Proof: Again use "1-2-3 argument".

Step 1: Suppose  $f \in \mathcal{N}_s(\Omega, \mathcal{F})$  so that  
 $f = \sum_{k=1}^n c_k I_{A_k}$  for  $A_k \in \mathcal{F}$  &  $0 \leq c_k \leq \infty$ .

since  $s \in \mathcal{N}(\Omega, \mathcal{F})$  (by def of densities) clearly

$$f \in \mathcal{Q}^-(\Omega, \mathcal{F}, \nu) \Leftrightarrow fs \in \mathcal{Q}^-(\Omega, \mathcal{F}, \mu).$$

$$\begin{aligned} \therefore \int f d\nu &= \sum_{k=1}^n c_k \nu(A_k) \\ &= \sum_{k=1}^n c_k \int_A s d\mu && \text{since } s \text{ is a density for } \nu \\ &&& \text{w.r.t. } \mu \\ &= \int \left( \sum_{k=1}^n c_k I_{A_k} \right) s d\mu && \text{by little 3} \\ &\quad \underbrace{\phantom{\sum_{k=1}^n c_k}}_f \end{aligned}$$

and this implies

$$f \in \mathcal{Q}^+(\Omega, \mathcal{F}, \nu) \Leftrightarrow fs \in \mathcal{Q}^+(\Omega, \mathcal{F}, \mu).$$

Step 2: Suppose  $f \in \mathcal{N}(\Omega, \mathcal{F})$ . Then  
the result follows similarly by little 3.

Step 3: From step 2

$$\int f^\pm d\nu = \int f^\pm s d\mu = \int (fs)^\pm d\mu$$

$$\therefore f \in \mathcal{Q}^\pm(\Omega, \mathcal{F}, \nu) \Leftrightarrow fs \in \mathcal{Q}^\pm(\Omega, \mathcal{F}, \mu) \quad \text{QED.}$$

### Notation:

(33)

Suppose  $v$  and  $\mu$  are measures on  $(\Omega, \mathcal{F})$ .

If  $v$  has a density w.r.t  $\mu$  I will denote it:

$$\frac{dv}{d\mu} \leftarrow \text{Non-negative } \mathbb{R} \text{ function mapping } \Omega \rightarrow \mathbb{R} \text{ s.t. } v(\omega) = \int \frac{dv}{d\mu} d\mu.$$

Moreover when I say  $\frac{dv}{d\mu}$  exists I mean there exists some density of  $v$  w.r.t.  $\mu$  (it will be unique  $\mu$ -a.e. if  $\frac{dv}{d\mu} \in L_1(\Omega, \mathcal{F}, \mu)$  or  $\mu$  is  $\sigma$ -finite).

### Theorem (Chain rule)

Suppose  $v, p, \mu$  are measures on  $(\Omega, \mathcal{F})$  with  $\mu$   $\sigma$ -finite. If  $\frac{dp}{dv}$  and  $\frac{dv}{d\mu}$  exists then  $\frac{dp}{d\mu}$  exists and satisfies

$$\frac{dp}{d\mu} = \frac{dp}{dv} \frac{dv}{d\mu} \quad \mu\text{-a.e.}$$

This Thm can be extended to the case  $\mu$  isn't  $\sigma$ -finite... then the RHS is a possibly non-unique density.

Proof:

$$\int \frac{dp}{dv} \frac{dv}{d\mu} d\mu = \int \frac{dp}{dv} dv \quad \text{by step in the density}$$

$$\begin{aligned} \uparrow \\ \in \eta(\Omega, \mathcal{F}) \subset \mathbb{G} &= \int dp \quad \text{by step in the density} \\ &= p(\omega) \quad \text{def} \end{aligned}$$

over  $\mathcal{F}$ .

Uniqueness follows since  $\mu$  is  $\sigma$ -finite.

$\square \in D$

Our final result is important for two reasons. First it will be used to significantly simplify the proof of the Radon-Nikodym theorem. Second it shows that measure theory (for  $\sigma$ -finite measures) can be considered a sub-theory of Probability!

(34)

### Theorem (Probabilist's world view)

i.e.  $\exists A \in \mathcal{F} : \mu(A) > 0$

If  $\mu$  is a non-trivial  $\sigma$ -finite measure on a measure space  $(\Omega, \mathcal{F})$ , then there exists a density  $s : \Omega \rightarrow (0, \infty)$  and a probability measure  $P$  on  $(\Omega, \mathcal{F})$  s.t.

$$d\mu = s dP$$

$$\text{i.e. } \mu(A) = \int_A s dP \quad \forall A \in \mathcal{F}.$$

Proof:

Let  $\Omega = \bigcup_{k=1}^{\infty} A_k$  where  $0 < \mu(A_k) < \infty$   
 disjoint  $A_k \in \mathcal{F}$   
 set  $A_p = F_p - (F_0 \cup \dots \cup F_{p-1}) \in \mathcal{F}$   
 where  $F_p = \bigcup_{k=1}^p A_k$  &  $\mu(F_p) < \infty$   
 to get this just  
 absorb  $A_j$  s.t.  
 $\mu(A_j) = 0$  into an  $A_k$   
 with positive measure.

Choose any sequence  $w_1, w_2, \dots$  s.t.  $w_k > 0$  and

$$\sum_{k=1}^{\infty} w_k = 1. \quad \text{Now define}$$

$$s^* := \sum_{k=1}^{\infty} \frac{w_k}{\mu(A_k)} I_{A_k} \in \eta(\Omega, \mathcal{F})$$

$$P(\omega) := \int s^* d\mu$$

$$\text{Notice } P(\Omega) = \int_{\Omega} s^* d\mu$$

$$= \int_{\Omega} \lim_{n \rightarrow \infty} \sum_{k=1}^n \frac{w_k}{\mu(A_k)} I_{A_k} d\mu$$

$$\stackrel{\text{B13}^3}{=} \sum_{k=1}^{\infty} \int_{\Omega} \underbrace{\frac{w_k}{\mu(A_k)} I_{A_k}}_{=w_k} d\mu$$

$$= 1$$

By  $\sigma$ -additivity of  $\int s^* d\mu$  and  $s^* \in \mathcal{N}(\mathbb{R}, \mathcal{F})$ ,  $P(\cdot)$  is a probability measure and (35)

$$dP = s^* d\mu$$

To finish notice that  $0 < s^*(w) < \infty$ ,  $\forall w \in \mathbb{R}$ , and define  $s(w) := \frac{1}{s^*(w)}$  which maps into  $(0, \infty)$  and is  $\mathcal{B}(\mathbb{R})$  by closure.

Now  $\forall A \in \mathcal{F}$

$$\mu(A) = \int_A s s^* d\mu = \int_A s dP$$

by step in the density  
since  $dP = s^* d\mu$ .

$\square \in \mathcal{P}$

Note: Suppose  $s \in \mathcal{N}(\mathbb{R}, \mathcal{F})$  and  $\nu, \mu$  are two measures s.t.

$$d\nu = s d\mu$$

In an exercise you will show

- i)  $\nu$  is finite  $\iff s \in L_1(\mathbb{R}, \mathcal{F}, \mu)$
- ii)  $\nu$  is  $\sigma$ -finite  $\Rightarrow s < \infty$   $\mu$ -a.e.
- iii)  $s < \infty$   $\mu$ -a.e.  $\begin{cases} \text{and} \\ \mu \text{ is } \sigma\text{-finite} \end{cases} \Rightarrow \nu$  is  $\sigma$ -finite.

e.g. Let  $X$  be a r.v. which has a density w.r.t. Lebesgue measure. By this I mean

$$dP_X^{-1} = s dx \quad d\lambda'$$

for some  $s \in \mathcal{N}(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ . In particular

$$P(X \in B) = \int_B s(x) dx \quad \forall B \in \mathcal{B}(\mathbb{R}).$$

$\hookrightarrow$  unique  $\mathcal{I}'$ -a.e. since  $s \in L_1(\mathbb{R}, \mathcal{F}, \mathcal{I}')$  or  $\mathcal{I}'$  is  $\sigma$ -finite

BTW: some people write  $P(X \in dx) = s(x)dx$  (36) as synonymous for  $dP_X^{-1} = s dx$ . It sorta makes sense since by step in the density one has

$$P(X \in A) = \int_A P_X^{-1}(dx) = \int_A s(x) dx$$

$\downarrow P(X \in dx)$

Now suppose  $g: \mathbb{R} \xrightarrow{\text{(m)}} \mathbb{R}$ . If  $g(X)$  is quasi-integrable we have

$$\begin{aligned} \therefore E(g(X)) &= \int g(X) dP \\ &= \int g(x) dP_X^{-1}(x) \quad \text{by change of variables} \\ &= \int_{\mathbb{R}} g(x) s(x) dx \quad \text{by step in the density.} \\ &\qquad\qquad\qquad \text{This is what everybody sees in undergrad probability.} \end{aligned}$$

e.g.

Let  $X_1, X_2, \dots, X_n$  be indep. r.v. s.t.

$$X_i = \begin{cases} 1 & \text{with prob } \theta \\ 0 & \text{with prob } 1-\theta \end{cases}$$

where  $0 \leq \theta \leq 1$ . Set  $S_n = X_1 + \dots + X_n$ .

$S_n$  has a binomial distribution denoted

$$S_n \sim \text{Bin}(n, \theta).$$

Let  $s(k) = \begin{cases} \binom{n}{k} \theta^k (1-\theta)^{n-k} & \text{if } k = 0, \dots, n \\ 0 & \text{otherwise} \end{cases}$

$$\begin{aligned} \text{Now } P(S_n \in B) &= \int_B s(k) d\lambda(k) \\ &\qquad\qquad\qquad \text{Counting measure} \\ &\therefore dP_{S_n}^{-1} = s dk \end{aligned}$$

e.g. Probably the most important r.v., 37  
besides the coin flip  $\text{Ber}(\theta)$ , is the  
Gaussian r.v.. A r.v.  $X$  is said to  
be Gaussian, denoted  $X \sim N(\mu, \sigma^2)$ , if

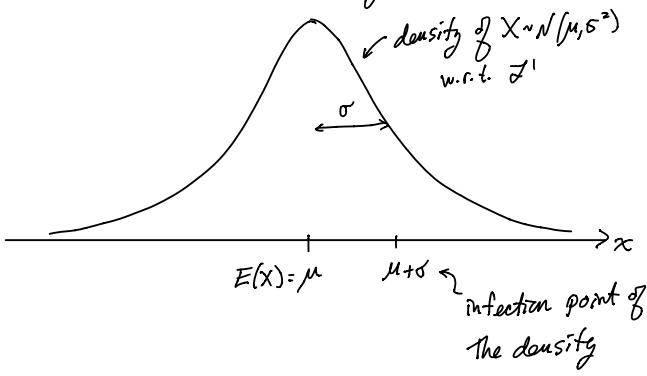
$$dP_{X^{-1}} = \frac{1}{\sqrt{\sigma^2 \pi}} \exp\left(-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2}\right) dx.$$

Notice that  $X \sim N(\mu, \sigma^2)$  implies

$$E(X) = \mu$$

$$\text{var}(X) = \sigma^2$$

so one has the following picture



If  $X_1, \dots, X_d$  are r.v.s defined on  $(\Omega, \mathcal{F}, P)$   
then the random vector  $X = (X_1, \dots, X_d)^T$   
is said to be jointly Gaussian, denoted  
 $X \sim N_d(\mu, \Sigma)$ , if

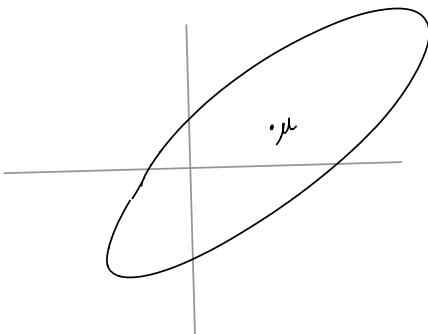
$$dP_{X^{-1}} = \frac{1}{\sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2} (x-\mu)^T \Sigma^{-1} (x-\mu)\right) dZ^d(x)$$

↑                      ↓  
 induced measure on  $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$        $d$ -dimensional vector       $d \times d$  positive definite matrix

in this case

$$\mu = \begin{pmatrix} E(X_1) \\ \vdots \\ E(X_d) \end{pmatrix} \text{ and } \text{cov}(X_i, X_j) = \Sigma_{ij}$$

The contours of  $X \sim N_2(\mu, \Sigma)$  look like 38



**Warning!** It is possible that both  $X_1$  and  $X_2$  are Gaussian r.v.s but  $(X_1, X_2)$  is not jointly Gaussian.

As an exercise check that if

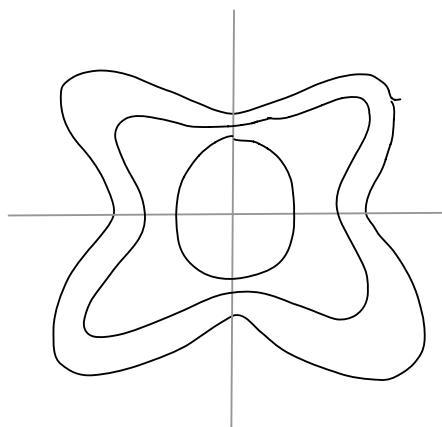
$$S_1(x_1, x_2) \text{ is the density of } N_2(0, 0), \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$S_2(x_1, x_2) \text{ is the density of } N_2(0, 0), \begin{pmatrix} 1 & -1/2 \\ -1/2 & 1 \end{pmatrix}$$

Then any random vector  $(X_1, X_2)$  with

$$\text{density } S(x_1, x_2) := \frac{S_1(x_1, x_2)}{2} + \frac{S_2(x_1, x_2)}{2}$$

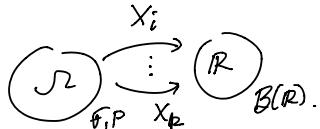
has Gaussian marginals, i.e.  $X_1 \sim N(0, 1)$   
and  $X_2 \sim N(0, 1)$ , but  $(X_1, X_2)$  are  
not jointly Gaussian since the contours  
of  $S(x_1, x_2)$  look like



## Glivenko-Cantelli

(39)

In statistics one often observes r.v.s  $X_1, X_2, \dots, X_n$  which are iid



Suppose each  $X_k$  has density  $\delta(x)$  w.r.t.  $\mathcal{F}$ , i.e.

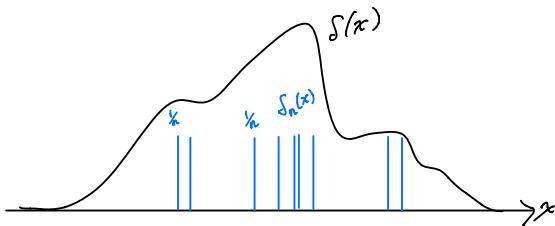
$$dP_{X_k}^{-1} = \delta(x) dx.$$

If  $\delta(x)$  is unknown it is natural to consider estimating it with the empirical measure based on  $X_1, X_2, \dots, X_n$  defined by

$$dP_n = \left[ \frac{1}{n} \sum_{k=1}^n I_{\{X_k\}}(x) \right] d\lambda$$

↑ ↑  
 empirical measure on  $(\Omega, \mathcal{F}, P)$       counting measure

Here is the picture:



For some events  $B \in \mathcal{B}(R)$ , such as intervals, one would expect

$$\frac{\#\{X_k \in B\}}{n} = \int_B \delta_n(x) d\lambda(x) \approx \int_B \delta(x) dx = P(X_1 \in B)$$

But for other  $B \in \mathcal{B}(R)$ , such as  $\{X_k : X_k = x\}$  one has

$$\frac{1}{n} = \int_B \delta_n(x) d\lambda(x) \neq \int_B \delta(x) dx = 0$$

So  $P_n$  doesn't uniformly estimate  $PX^{-1}$  over Borel events. (40)

If we instead compare the c.d.f.'s of the two measures  $\delta(x)dx$  and  $\delta_n(x)d\lambda(x)$  we get uniform approximation. This is the Glivenko-Cantelli theorem.

As setup for the theorem let

$$F(t) := \int_{(-\infty, t]} \delta(x) dx = P(X_1 \leq t)$$

$$F_n(t) := \int_{(-\infty, t]} \delta_n(x) d\lambda(x) = \frac{1}{n} \sum_{k=1}^n \underbrace{I_{(-\infty, t]}(X_k)}_{\substack{\# \text{ of } k's \text{ b/w } 1 \text{ to } n \\ \text{s.t. } X_k \leq t}}$$

$F_n$  is called the empirical c.d.f.

## Theorem (Glivenko-Cantelli)

If  $X_1, X_2, \dots$  are iid r.v.s on  $(\Omega, \mathcal{F}, P)$ .

Then

$$\sup_{t \in \mathbb{R}} |F(t) - F_n(t)| \xrightarrow{n \rightarrow \infty} 0 \quad P\text{-a.e.}$$

Proof:

Notice that  $I_{(-\infty, t]}(X_1), \dots, I_{(-\infty, t]}(X_n)$  are iid bdd r.v.s. Therefore the SLLN for bdd r.v.s applies and gives

$$\underbrace{\frac{1}{n} \sum_{k=1}^n I_{(-\infty, t]}(X_k)}_{= F_n(t)} \xrightarrow{n \rightarrow \infty} \underbrace{E I_{(-\infty, t]}(X_1)}_{= P(X_1 \leq t) = F(t)} \quad P\text{-a.e.}$$

for each fixed  $t \in \mathbb{R}$ . Similarly

$$\underbrace{\frac{1}{n} \sum_{k=1}^n I_{(-\infty, t)}(X_k)}_{= F_n(t-)} \xrightarrow{n \rightarrow \infty} \underbrace{E I_{(-\infty, t)}(X_1)}_{= P(X_1 < t) = F(t-)} \quad P\text{-a.e.}$$

For each  $t \in \mathbb{R}$  let  $\mathcal{D}_t^\circ \subset \mathcal{D}$  be the measurable event s.t.

- $P(\mathcal{D}_t^\circ) = 1$
- $F_n(t) \rightarrow F(t)$  &  $F_n(t^-) \rightarrow F(t^-)$  everywhere on  $\mathcal{D}_t^\circ$

Now fix  $\varepsilon > 0$ .

Choose  $m$  points (depending on  $\varepsilon$ ) s.t.

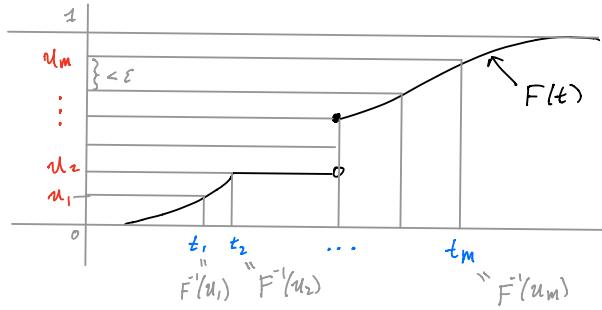
$$0 < u_1 < u_2 < \dots < u_m < 1$$

↑ ↑ ... ↑ →  
with spacing less than  $\varepsilon$

and define

$$t_p = F^{-1}(u_p) := \inf \{t : F(t) \geq u_p\}.$$

The reason for this choice is to get more  $t_p$ 's where  $F(t)$  changes quickly. Picture:



Recall the "off sandwich lemma"

$$F(F^{-1}(u)^-) \leq u \leq F(F^{-1}(u)) \quad \forall u \in (0,1).$$

Therefore

$$(*) \quad k=1, \dots, m \Rightarrow F(t_p^-) \leq u_k$$

$$(**) \quad k=2, \dots, m+1 \Rightarrow u_{k-1} \leq F(t_{k-1}^-)$$

(41)

Let  $\mathcal{D}_\varepsilon := \bigcap_{k=1}^m \mathcal{D}_{t_p}^\circ$  so that  $P(\mathcal{D}_\varepsilon) = 1$  and

$$w \in \mathcal{D}_\varepsilon \Rightarrow \begin{cases} F_n(t_p) \rightarrow F(t_p) & \text{as } n \rightarrow \infty \\ F_n(t_p^-) \rightarrow F(t_p^-) & \text{as } n \rightarrow \infty \\ \forall p = 1, \dots, m \end{cases}$$

$\Rightarrow \exists$  a finite  $N = N(w, \varepsilon)$  s.t.

$$(i) \quad \sup_{1 \leq p \leq m} |F_n(t_p) - F(t_p)| \leq \varepsilon$$

$$(ii) \quad \sup_{1 \leq p \leq m} |F_n(t_p^-) - F(t_p^-)| \leq \varepsilon$$

$\forall n \geq N$ .



e.g. take  $N$  to be the max over  $N_p$ 's and  $M_p$ 's s.t.

$$n \geq N_p \Rightarrow |F_n(t_p) - F(t_p)| \leq \varepsilon$$

$$n \geq M_p \Rightarrow |F_n(t_p^-) - F(t_p^-)| \leq \varepsilon$$

To finish we show

$$w \in \mathcal{D}_\varepsilon \Rightarrow \sup_{t \in \mathbb{R}} |F_n(t) - F(t)| \leq 2\varepsilon, \quad \forall n \geq N(w, \varepsilon)$$

This is sufficient since  $P(\mathcal{D}_\varepsilon) = 1$  hence

$$P\left(\bigcap_{\varepsilon \in \mathbb{Q}^+} \mathcal{D}_\varepsilon\right) = 1$$

$\sup_{t \in \mathbb{R}} |F_n(t) - F(t)| \rightarrow 0$  for these  $w$ 's

(42)

Case 1:  $t \in (t_{k-1}, t_k)$  &  $n \geq N$ .

(43)

$$\begin{aligned} \therefore F_n(t) &\leq F_n(t_{k-1}^-) \text{ since } P_n([-n, t]) \leq P_n([-n, t_k]) \\ &\leq F(t_{k-1}^-) + \varepsilon \text{ by (i)} \\ &\leq u_k + \varepsilon \text{ by (*)} \\ &= u_{k-1} + \underbrace{(u_k - u_{k-1})}_{\leq \varepsilon} + \varepsilon \\ &\leq F(t_{k-1}^-) + 2\varepsilon \text{ by (**)} \\ &\leq F(t) + 2\varepsilon \text{ by monotonicity} \end{aligned}$$

Also  $F(t) \leq F(t_{k-1}^-)$

$$\begin{aligned} &\vdots \quad \leftarrow \text{same as above without extra } \varepsilon \\ &\leq F(t_{k-1}^-) + \varepsilon \\ &\leq F_n(t_{k-1}^-) + 2\varepsilon \text{ by (i)} \\ &\leq F_n(t) + 2\varepsilon \text{ by monotonicity.} \end{aligned}$$

$$\therefore \forall \varepsilon \in \mathbb{R}_\varepsilon \Rightarrow \sup_{t \in (t_{k-1}, t_k)} |F_n(t) - F(t)| \leq 2\varepsilon, \forall n \geq N$$

Case 2:  $t < t_1$  &  $n \geq N$ .

Similar to case 1 using the fact that  $0 \leq F(t_1^-) \leq F(t_1) \leq \varepsilon$  by (\*).

$$\therefore \forall \varepsilon \in \mathbb{R}_\varepsilon \Rightarrow \sup_{t < t_1} |F_n(t) - F(t)| \leq 2\varepsilon, \forall n \geq N$$

Case 3:  $t > t_m$  &  $n \geq N$ .

Similar to case 2 using  $1 - F(t_m) \leq \varepsilon$  by (\*\*)

$$\therefore \forall \varepsilon \in \mathbb{R}_\varepsilon \Rightarrow \sup_{t > t_m} |F_n(t) - F(t)| \leq 2\varepsilon, \forall n \geq N$$

Finally by cases 1, 2, 3, (i), (ii) we have

$$\forall \varepsilon \in \mathbb{R}_\varepsilon \Rightarrow \sup_{t \in \mathbb{R}} |F_n(t) - F(t)| \leq 2\varepsilon, \forall n \geq N$$

As was to be shown.

QED

Scheffé's Theorem

(44)

The last theorem we will cover in this lecture shows that pointwise convergence of probability densities implies uniform convergence of probabilities over the  $\sigma$ -field.

Theorem (Scheffé)

Let  $P_n$  &  $P$  be two probability measures on  $(\Omega, \mathcal{F})$  which have densities  $f_n$  &  $f$  w.r.t some measure  $\mu$  on  $(\Omega, \mathcal{F})$ .

If  $f_n \rightarrow f$   $\mu$ -a.e. as  $n \rightarrow \infty$  then

$$\|P_n - P\|_{TV} := \sup_{A \in \mathcal{F}} |P_n(A) - P(A)| \stackrel{(i)}{\leq} \int_{\Omega} |f_n - f| d\mu \stackrel{(ii)}{\rightarrow} 0$$

called the total variation norm as  $n \rightarrow \infty$ .

Proof:

To show (i) notice that

$$\begin{aligned} |P_n(A) - P(A)| &= \left| \int_A f_n d\mu - \int_A f d\mu \right| \\ &= \left| \int_A (f_n - f) d\mu \right| \quad \text{by Bg 3 since } f_n, f \in L_1(\Omega, \mathcal{F}, \mu) \\ &\leq \int_A |f_n - f| d\mu \quad \text{"a.e.-useful facts"} \\ &\leq \int_{\Omega} |f_n - f| d\mu \end{aligned}$$

To prove (ii) set  $\Delta_n := f - f_n$ . Now

$$\begin{aligned} \int_{\Omega} |\Delta_n| d\mu &= \int_{\Delta_n > 0} \Delta_n d\mu - \int_{\Delta_n < 0} \Delta_n d\mu \\ &= 2 \int_{\Delta_n > 0} \Delta_n d\mu \quad \text{since } \int_{\Delta_n < 0} \Delta_n d\mu = \int_{\Omega} f_n d\mu = \int_{\Omega} f d\mu = 0 \\ &= 2 \int_{\Delta_n^+} \Delta_n^+ d\mu \quad \text{by DCT since } \lim_n \Delta_n^+ = (\lim_n \Delta_n)^+ = 0 \text{ } \mu\text{-a.e.} \\ &\rightarrow 0 \quad \text{and } \Delta_n^+ = (f - f_n)^+ \leq f^+ \epsilon L_1(\Omega, \mathcal{F}, \mu) \\ &\quad \uparrow \text{monotonicity of } (\cdot)^+ \end{aligned}$$

We will use this theorem in the  
lecture on convergence in distribution for a nice  
proof of a result by Schoenberg & von Neumann  
which states that functions of the form

$$K(h, g) := \int_0^\infty \exp(-t^2 \|h-g\|_{\mathcal{H}}^2) dP(t)$$

are positive definite over all  $h, g \in \mathcal{H}$  where  
 $\mathcal{H}$  is a Hilbert space (Ref: Steerneman &  
van Perlo-ten Kleij).