

Statistics 235A / Math 235A

Probability & Measure

UC Davis, Fall 2016

Instructor: Ethan Anderes

Let's start the class with a paradox.

Suppose X, Y are two independent positive random variables (each with a density).

Question:

What is the distribution of X given the event $\{X=Y\}$?

Two possible solutions:

(1) Take $X|_{X=Y}$ to mean $X|_{Z=0}$ where $Z = X-Y$.

By "change of variables" $X|_{Z=0}$ has density

$$f_{X|Z=0}(x) = \frac{f_{X,Y}(x, 0)}{f_Z(0)} \propto f_{X,Y}(x, x-0) \underbrace{\left| \frac{d(x-y)}{d(x-y)} \right|}_{=1}$$

$$\text{since } \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} X \\ Y-x \end{pmatrix}$$

(2) Take $X|_{X=Y}$ to mean $X|_{W=1}$ where $W = \frac{Y}{X}$.

Again by "change of variables" $X|_{W=1}$ has density

$$f_{X|W=1}(x) = \frac{f_{X,W}(x, 1)}{f_W(1)} = f_{X,Y}(x, x \cdot 1) \underbrace{\left| \frac{d(x-y)}{d(x-y)} \right|}_{= \left| \frac{1}{W} x \right|}$$

$$\text{since } \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} X \\ XW \end{pmatrix}.$$

Paradox:

(1) says $X|_{X=Y}$ has density $\propto f_{X,Y}(x, x)$

(2) says $X|_{X=Y}$ has density $\propto x f_{X,Y}(x, x)$

why are they different? which one is correct?

(1)

Course Content

(2)

- Rigorous measure theoretic underpinnings of Probability Theory:
 - measure & integration
 - σ -fields
 - conditional probability.
- Advanced topics using the general theory:
 - Martingales
 - Markov chains
 - Convergence of prob measures
 - Borel cantelli lemmas
 - LIL

Remark: Although I rarely (never) do research in measure theory, I find it extremely useful to know about it and not be afraid of it.

Course Material

- I have a collection of notes (still a work in progress) which you can find on my github page "github.com/EthanAnderes/ProbabilityAndMeasureNotes"
- I'll also put p.d.f copies of my lecture notes on github
- Book Ref: "Prob & Measure" by Billingsley.
- "Probability, Theory & Examples" by Durrett.

Course grading

- weekly HWs. (80% of total grade)
- Exams (20% of total grade)
- I'll announce each one with at least 1 week notice

Borel's Normal number theorem

(3)

In modern terms this thm is basically the SLLN for coin flips.

Thm: If X_1, X_2, \dots are indep R.V.s s.t.

$$X_i = \begin{cases} 1 & \text{w.p. } \frac{1}{2} \\ 0 & \text{w.p. } \frac{1}{2} \end{cases}$$

then $P\left(\lim_{n \rightarrow \infty} \frac{X_1 + \dots + X_n}{n} = \frac{1}{2}\right) = 1$.

Here is how one would prove it in modern times.

$$\text{Let } S_n = 2^n \left[\frac{X_1 + \dots + X_n}{n} - \frac{1}{2} \right]$$

$$= \sum_{i=1}^n (2X_i - 1) \quad \text{rademacher coin flip} \\ = R_i = \pm 1,$$

Note $E(R_i) = 0$, $\text{var}(R_i) = E(R_i^2) - 0 = 1$

$$\begin{aligned} P\left(\left|\frac{S_n}{n}\right| \geq \varepsilon\right) &= 2P\left(\frac{S_n}{n} \geq \varepsilon\right) \\ &= 2P(S_n \geq n\varepsilon) \\ &= 2P(e^{\varepsilon S_n} \geq e^{n\varepsilon}) \\ &\quad \text{want to apply Markov on the MGF of } S_n \\ &\leq 2 \frac{E(e^{\varepsilon S_n})}{e^{n\varepsilon}} \quad \text{since it is compactly supported} \\ &= 2e^{-n\varepsilon^2} \left[E(e^{\varepsilon R_i}) \right]^n \quad \because \text{very smooth at } \varepsilon=0 \end{aligned}$$

where

$$\begin{aligned} E(e^{\varepsilon R_i}) &= \frac{1}{2} (e^{-\varepsilon} + e^{\varepsilon}) \\ &\leq e^{\varepsilon^2/2}, \quad \text{use Taylor series} \end{aligned}$$

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

Now let $\varepsilon_n \rightarrow 0$ slow enough that $n\varepsilon_n^2 \rightarrow \infty$ & $\sum_{n=1}^{\infty} e^{-n\varepsilon_n^2/2} < \infty$

$$\therefore \sum_{n=1}^{\infty} P\left(\left|\frac{S_n}{n}\right| \geq \varepsilon_n\right) < \infty$$

$\therefore P\left(\left|\frac{S_n}{n}\right| < \varepsilon_n \text{ for all but finite } n\right) = 1$
by the first Borel-Cantelli lemma

$$\therefore P\left(\lim_{n \rightarrow \infty} \frac{S_n}{n} = 0\right) = 1$$

$$\therefore P\left(\lim_{n \rightarrow \infty} \frac{X_1 + \dots + X_n}{n} = \frac{1}{2}\right) = 1.$$

QED.

But if your Borel in 1909 you...

- don't know what a prob measure P is
- don't know the meaning of "length(A)" for non-standard sets A.C.R.

- don't even know if there exists an infinite sequence of R.V.s X_1, X_2, \dots

$$\text{s.t. } X_i = \begin{cases} 0 & \text{w.p. } \frac{1}{2} \\ 1 & \text{w.p. } \frac{1}{2} \end{cases}$$

- don't know Markov's thm or first Borel-Cantelli lemma.

(Side Note: Borel's original proof in 1909 had a mistake which was fixed in 1914 by Hausdorff).

Let's prove this from first principles to...

- Motivate the definition of a probability space.

- show the existence of an infinite sequence of independent coin flip random variables (which will be used again later in the course).

- In case you time traveled back to 1909 you could write a proof others would understand.

- Show what one can do with simple finite additivity.

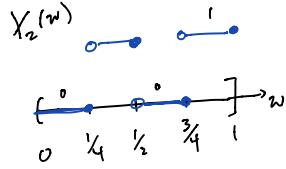
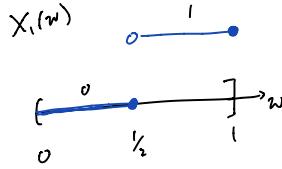
(4)

For $k=1, 2, \dots$ let $X_k: [0,1] \rightarrow \{0,1\}$ (5)

be the function:

$X_k(w) = \underbrace{\text{k-th binary digit of } w}_{\text{non-terminating, i.e. use } 0.011111\dots \text{ instead of } 0.1000\dots}$

e.g.



$$\text{e.g. } X_2(3/4) = X_2(0.101111\dots) = 0$$

Let $N = \left\{ w \in [0,1] : \frac{1}{n} \sum_{k=1}^n X_k(w) \rightarrow \frac{1}{2} \right\}$
= Normal numbers in $[0,1]$

$A = N^c$
= Abnormal numbers in $[0,1]$

Borel's question: is $\text{length}(N) = 1$?
is $\text{length}(A) = 0$?
what does "length" mean?

Dof: $(\mathcal{R}, \mathcal{F}, P)$ is a finitely additive probability (FAP) model if

1. \mathcal{R} is a set
2. \mathcal{F} is a collection of subsets of \mathcal{R} s.t.
 - 2a. $\mathcal{R} \in \mathcal{F}$
 - 2b. $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$
 - 2c. $A, B \in \mathcal{F} \Rightarrow A \cup B \in \mathcal{F}$

i.e. \mathcal{F} is a field

3. $P: \mathcal{F} \rightarrow [0,1]$ s.t.

$$3a: P(\mathcal{R}) = 1$$

$$3b: P(A \cup B) = P(A) + P(B)$$

disjoint

members of \mathcal{F} .

In our case:

$$\mathcal{R} = [0,1]$$

\mathcal{F} = Finite disjoint unions of intervals $[a, b] \subset \mathcal{R}$ denote $\mathcal{B}_{[0,1]}$

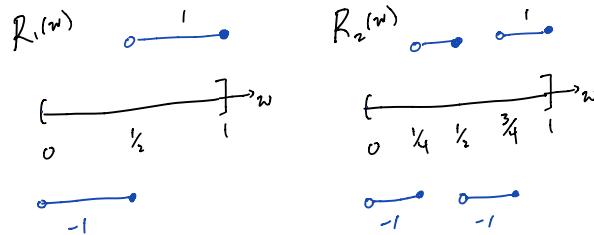
$$P\left(\bigcup_{i=1}^n [a_i, b_i]\right) = \sum_{i=1}^n (b_i - a_i).$$

disjoint

forms a FAP.

Note: $N, A \notin \mathcal{F}$ so it doesn't make sense to ask if $P(N) = 1$ & $P(A) = 0$.

As before let $R_k(w) = 2X_k(w) - 1$



$$\text{and set } S_n(w) = \sum_{k=1}^n R_k(w)$$

Even though $N, A \notin \mathcal{F}$ we do have

$$\left\{ \frac{S_n}{n} \geq \varepsilon \right\} := \left\{ w \in [0,1] : \left| \frac{S_n(w)}{n} \right| \geq \varepsilon \right\} \in \mathcal{F}$$

Thm ($W \& L$): $\forall \varepsilon > 0$

$$P\left(\frac{S_n}{n} \geq \varepsilon\right) \rightarrow 0$$

as $n \rightarrow \infty$.

Proof: Note that R_1, R_2, \dots are orthogonal functions in $L^2([0,1])$

$$\text{i.e. } \langle R_i, R_j \rangle = \int_0^1 R_i(w) R_j(w) dw = \delta_{ij}$$

$$\begin{aligned}
 n &= \int_0^1 \left(\sum_{i=1}^n R_i(w) \right)^2 dw \\
 &= \int_0^1 S_n^2(w) dw \\
 &\geq \int_{|S_n/n| \geq \varepsilon} S_n^2(w) dw \\
 &\quad \text{on this set of } w \text{ we have } S_n^2 \geq n^2 \varepsilon^2 \\
 &\geq n^2 \varepsilon^2 \int_{|S_n/n| \geq \varepsilon} 1 dw \\
 &= n^2 \varepsilon^2 P\left(|S_n/n| \geq \varepsilon\right) \\
 &\quad \text{defined since this is simply length}\left(\left\{|S_n/n| \geq \varepsilon\right\}\right). \\
 \therefore P\left(\left|\frac{S_n}{n}\right| \geq \varepsilon\right) &\leq \frac{1}{n \varepsilon^2} \rightarrow 0 \text{ as } n \rightarrow \infty. \quad \text{QED-} \\
 &\quad \text{basically chebyshev}
 \end{aligned}$$

One of your exercises will show
one can get better bounds:

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

From first principles.

Now we just need to do a
Borel-Cantelli argument by hand.

Thm (SLN, Borel's normal number thm)

For all $\varepsilon > 0$, $\exists F_1, F_2, \dots \in \mathcal{F}$ s.t.

$$A \subset \bigcup_{k=1}^{\infty} F_k$$

$$\sum_{k=1}^{\infty} P(F_k) \leq \varepsilon.$$

i.e. A is negligible.

the combined "length" of the F_k 's $\leq \varepsilon$.

Proof:

For any $\varepsilon_n \rightarrow 0$ we have

$$\left\{ \left| \frac{S_n}{n} \right| < \varepsilon_n \text{ for } n \right\} \subset \left\{ \lim_{n \rightarrow \infty} \frac{S_n}{n} = 0 \right\} = N$$

$$\therefore A = N^c = \left\{ \left| \frac{S_n}{n} \right| \geq \varepsilon_n \text{ for infinitely many } n \right\}$$

$$\subset \bigcup_{n=m}^{\infty} \left\{ \left| \frac{S_n}{n} \right| \geq \varepsilon_n \right\}, \text{ for }$$

This is the cover
for some appropriately
chosen m & ε_n

$$\begin{aligned}
 \text{where } \sum_{n=m}^{\infty} P\left(\left|\frac{S_n}{n}\right| \geq \varepsilon_n\right) \\
 \leq \sum_{n=m}^{\infty} 2e^{-n\varepsilon_n^2/2}
 \end{aligned}$$

Can be made to be \leq any given ε
by choosing $\varepsilon_n \rightarrow 0$ so that $e^{-n\varepsilon_n^2/2}$
is summable, then choosing m large
enough. QED.