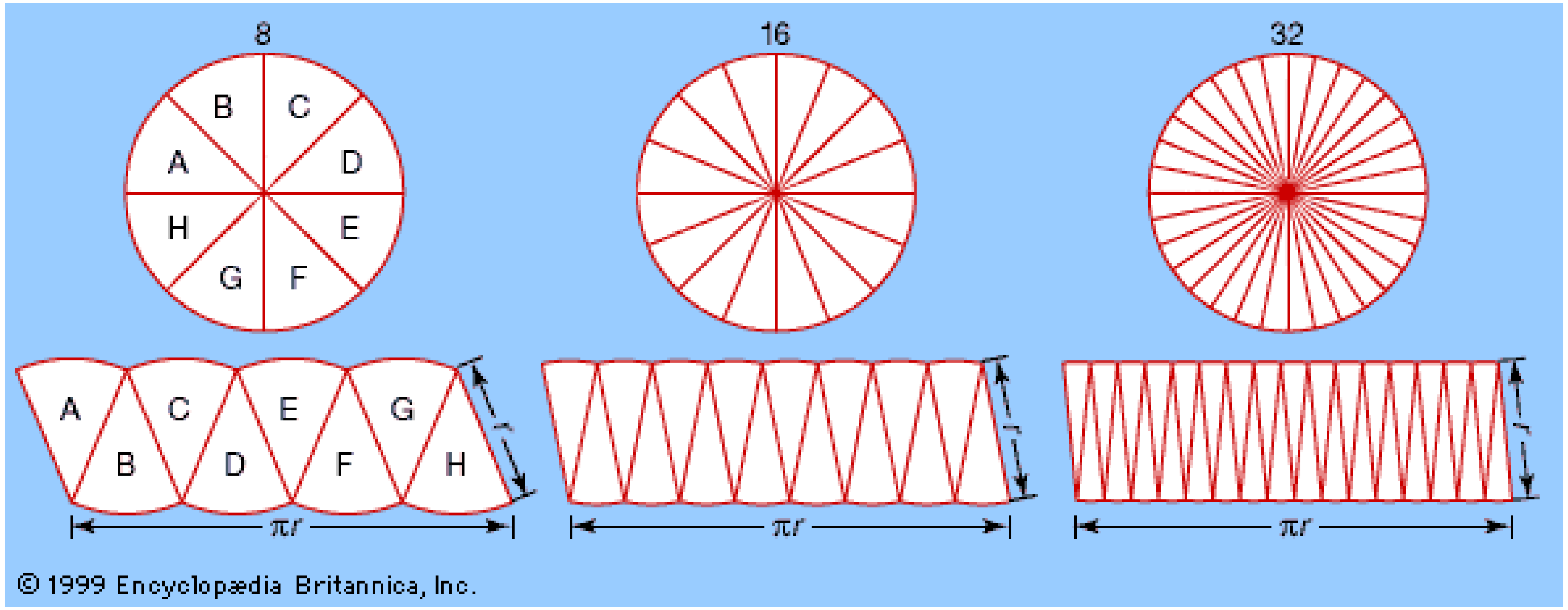
Measure Theory

Measuring Area of a Circle



Algebra

Let Ω be a set: for example $\mathbf{R}$. We want to describe the families of subsets on which we can make sense of measures.

Definition (**Algebra**). *A collection $\mathcal{A}$ of subsets of a set Ω is an* **algebra** *if*

- $\emptyset \in \mathcal{A}$ *and* $\Omega \in \mathcal{A}$
- *If* $A \in \mathcal{A}$ *then* $A^c = \Omega \setminus A \in \mathcal{A}$
- *If* $A, B \in \mathcal{A}$ *then* $A \cup B \in \mathcal{A}$

Note that an algebra is also closed under finite intersections because

$$A \cap B = (A^c \cup B^c)^c \, .$$

σ - Algebra

Definition (**σ-algebra**). *A collection $\mathcal{F}$ of subsets of a set Ω is a* **σ-algebra** *if*

- $\emptyset \in \mathcal{F}$ *and* $\Omega \in \mathcal{F}$
- *If* $A \in \mathcal{F}$ *then* $\Omega \backslash A \in \mathcal{F}$
- *If* $A_1, A_2, A_3, \ldots \in \mathcal{F}$ *then* $\bigcup A_i \in \mathcal{F}$

If $\mathcal{F}$ is a σ-algebra and (A_i) is a countable sequence of members of $\mathcal{F}$ then the intersection $\bigcap A_i$ is also a member.

Example

Example. *For any set* Ω *the collection of all subsets is a* σ*-algebra.*

Example. *For any set* Ω *and any subset* A*, the family*

$$\{\emptyset, A, A^c, \Omega\}$$

is a σ*-algebra.*

Generated σ - Algebra

Lemma (**Intersection of σ-algebras**). *If $(\mathcal{F}_\alpha)_\alpha$ is an arbitrary collection of σ-algebras then*

$$\bigcap_\alpha \mathcal{F}_\alpha$$

is a σ-algebra.

Definition (**Generated σ-algebra**). *For any family $\mathcal{U}$ of subsets of a set Ω there is a smallest σ-algebra on Ω including $\mathcal{U}$, known as the σ-algebra* **generated by** $\mathcal{U}$ *and denoted $\sigma(\mathcal{U})$.*

Open Set

- An open set o is a member of a given collection of subsets $\mathcal{O}$ of a given set Ω,
- The collection $\mathcal{O}$ has the property of containing
  - every union of its members,
  - every finite intersection of its members,
  - the empty set,
  - and the whole set itself.

Borel σ - Algebra

Consider the set

$$\Sigma = \{\mathcal{A} : \mathcal{A} \text{ sigma-algebra containing all open subsets of } \mathbb{R}\}.$$

(For example, $\mathcal{P}(\mathbb{R})$ is one of the elements of Σ.) Then the intersection of all such sigma-algebras

$$\mathcal{B} = \bigcap_{\mathcal{A} \in \Sigma} \mathcal{A}$$

is the smallest σ-algebra containing all open subsets of $\mathbb{R}$, and it's called the **Borel σ-algebra**.

Any subset of the Borel σ – Algebra is a Borel set.

Measures

Definition (Measure). *Let $\mathcal{F}$ be a σ-algebra on a set Ω. A function $\mu : \mathcal{F} \to [0, \infty]$ is called a* **measure** *if*

- $\mu(\emptyset) = 0$
- *If $A_1, A_2, A_3, \ldots \in \mathcal{F}$ are disjoint then*

$$\mu\left(\bigcup A_i\right) = \sum \mu(A_i).$$

The last property is known as "countable additivity".

Example

Example (Exercise). *For any set* Ω *we may consider the* σ*-algebra of all subsets and define the measure of a set* A *to be the number of points in* A*.*

This is called **counting measure.**

Example. *For* $\Omega = \{H, T\}$ *and the* σ*-algebra of all subsets we may define the measures of the sets individually as* $\mu(\{H\}) = 1/2$, $\mu(\{T\}) = 1/2$ *(and* $\mu(\{H, T\}) = 1$*).*

Measure Space

- Measurable Space – $(\Omega, \mathcal{F})$
- Measure Space - $(\Omega, \mathcal{F}, \mu)$

Probability Measure

- A *measure* P is a probability measure over $(\Omega, \mathcal{F})$
  - If $P(\Omega) = 1$

- $(\Omega, \mathcal{F}, P)$ is a probability space

- Elements of $\mathcal{F}$ are called *events*

Random Variable

- A random variable provides us with a numerical value, depending on the outcome of an experiment (denoted by X(ω))
  - a function from the sample space to the real numbers

A function $X : \Omega \to \mathbb{R}$ *is a* **random variable** *if the set* $\{\omega \mid X(\omega) \leq c\}$ *is* $\mathcal{F}$*-measurable for every* $c \in \mathbb{R}$.

Examples

Example 1. Consider a sequence of five consecutive coin tosses. An appropriate sample space is $\Omega = \{0,1\}^n$, where "1" stands for heads and "0" for tails. Let $\mathcal{F}$ be the collection of all subsets of Ω, and suppose that a probability measure $\mathbb{P}$ has been assigned to $(\Omega, \mathcal{F})$. We are interested in the number of heads obtained in this experiment. This quantity can be described by the function $X : \Omega \to \mathbb{R}$, defined by

$$X(\omega_1, \ldots, \omega_n) = \omega_1 + \cdots + \omega_n.$$

Example 2. (**Indicator functions**) Suppose that $A \subset \Omega$, and let $I_A : \Omega \to \{0,1\}$ be the indicator function of that set; i.e., $I_A(\omega) = 1$, if $\omega \in A$, and $I_A(\omega) = 0$, otherwise. If $A \in \mathcal{F}$, then I_A is a random variable. But if $A \notin \mathcal{F}$, then I_A is not a random variable.

Example 3. (**A function of a random variable**) Suppose that X is a random variable, and let us define a function $Y : \Omega \to \mathbb{R}$ by letting $Y(\omega) = X^3(\omega)$, for every $\omega \in \Omega$, or $Y = X^3$ for short. We claim that Y is also a random variable. Indeed, for any $c \in \mathbb{R}$, the set $\{\omega \mid Y(\omega) \le c\}$ is the same as the set $\{\omega \mid X(\omega) \le c^{1/3}\}$, which is in $\mathcal{F}$, since X is a random variable.

Definition 2. **(The probability law of a random variable)** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $X : \Omega \to \mathbb{R}$ be a random variable.*

(a) *For every Borel subset B of the real line (i.e., $B \in \mathcal{B}$), we define $\mathbb{P}_X(B) = \mathbb{P}(X \in B)$.*

(b) *The resulting function $\mathbb{P}_X : \mathcal{B} \to [0, 1]$ is called the* **probability law** *of X.*

Proposition 1. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let X be a random variable. Then, the law $\mathbb{P}_X$ of X is a probability measure on $(\mathbb{R}, \mathcal{B})$.*

Cumulative Distribution Functions

Definition 4. (Cumulative distribution function) *Let X be a random variable. The function $F_X : \mathbb{R} \to [0, 1]$, defined by*

$$F_X(x) = \mathbb{P}(X \leq x),$$

is called the cumulative distribution function (**CDF**) *of X.*

Corollary 1. *There is a one-to-one correspondence between distribution functions F and probability measures $\mathbb{P}$ on $(\mathbb{R}, \mathcal{B})$.*

Functions of Random Variables

- Continuous functions of random variable are random variables

Another way that we can form a random variable is by taking the limit of a sequence of random variables. Let us first introduce some terminology. Let each f_n be a function from some set Ω into $\mathbb{R}$. Consider a new function $f = \inf_n f_n$ defined by $f(\omega) = \inf_n f_n(\omega)$, for every $\omega \in \Omega$. The functions $\sup_n f_n$, $\liminf_{n\to\infty} f_n$, and $\limsup_{n\to\infty} f_n$ are defined similarly. (Note that even if the f_n are everywhere finite, the above defined functions may turn out to be extended-valued.) If the limit $\lim_{n\to\infty} f_n(\omega)$ exists for every ω, we say that the sequence of functions $\{f_n\}$ **converges pointwise**, and define its **pointwise limit** to be the function f defined by $f(\omega) = \lim_{n\to\infty} f_n(\omega)$. For example, suppose that $\Omega = [0, 1]$ and that $f_n(\omega) = \omega^n$. Then, the pointwise limit $f = \lim_{n\to\infty} f_n$ exists, and satisfies $f(1) = 1$, and $f(\omega) = 0$ for $\omega \in [0, 1)$.

Expectation of Random Variables

$$\mathbb{E}[X] = \int X(\omega)\,\mathrm{d}\mathbb{P}(\omega),$$

$$\mathbb{E}[f(X,Y)] = \int f(X(\omega),Y(\omega))\,\mathrm{d}\mathbb{P}(\omega).$$

$$\mathbf{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]$$

Convergence of Sequence of RVs

A sequence $X_1, X_2, \ldots$ of real-valued random variables, with cumulative distribution functions $F_1, F_2, \ldots$, is said to **converge in distribution**, or **converge weakly**, or **converge in law** to a random variable X with cumulative distribution function F if

$$\lim_{n\to\infty} F_n(x) = F(x),$$

for every number $x \in \mathbb{R}$ at which F is continuous.

Convergence in Probability

A sequence $\{X_n\}$ of random variables **converges in probability** towards the random variable X if for all $\varepsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}\big(|X_n - X| > \varepsilon\big) = 0.$$

More explicitly, let $P_n(\varepsilon)$ be the probability that X_n is outside the ball of radius ε centered at X. Then X_n is said to converge in probability to X if for any $\varepsilon > 0$ and any $\delta > 0$ there exists a number N (which may depend on ε and δ) such that for all $n \geq N$, $P_n(\varepsilon) < \delta$ (the definition of limit).

Almost Sure Convergence

To say that the sequence X_n converges **almost surely** or **almost everywhere** or **with probability 1** or **strongly** towards X means that

$$\mathbb{P}\left(\lim_{n\to\infty} X_n = X\right) = 1.$$

This means that the values of X_n approach the value of X, in the sense that events for which X_n does not converge to X have probability 0 (see *Almost surely*). Using the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and the concept of the random variable as a function from Ω to **R**, this is equivalent to the statement

$$\mathbb{P}\left(\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\right) = 1.$$

Pointwise Convergence

To say that the sequence of random variables (X_n) defined over the same probability space (i.e., a random process) converges **surely** or **everywhere** or **pointwise** towards X means

$$\forall \omega \in \Omega\colon \ \lim_{n\to\infty} X_n(\omega) = X(\omega),$$

where Ω is the sample space of the underlying probability space over which the random variables are defined.

Law of Large Numbers

i.i.d. (independent, identically distributed) random vars

$X_1, X_2, X_3, \ldots$

X_i has $\mu = E[X_i] < \infty$ and $\sigma^2 = Var[X_i]$

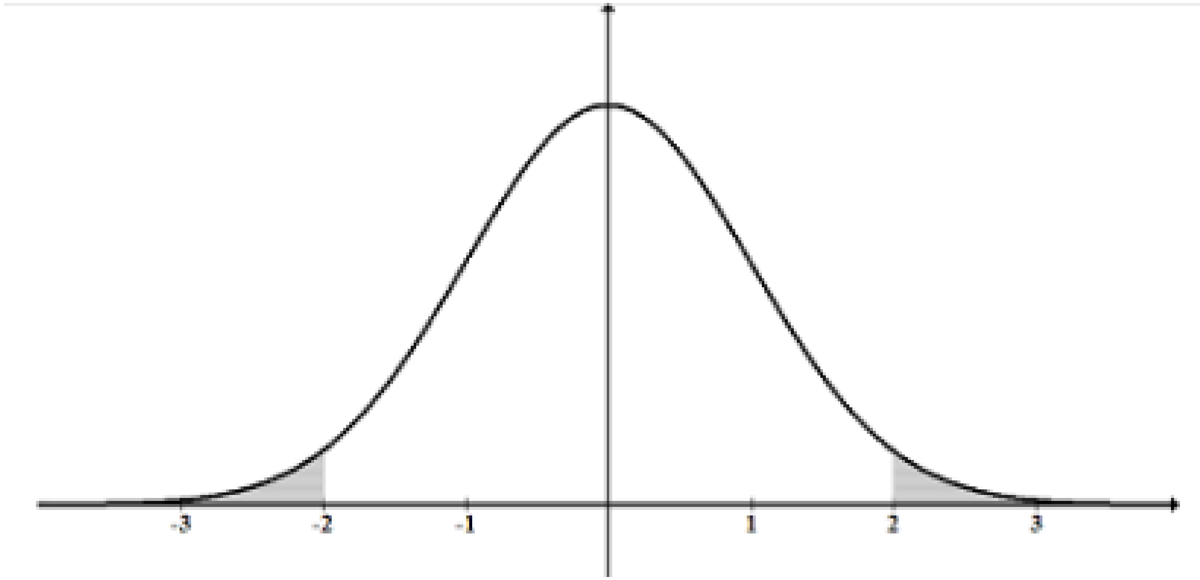
Consider the *sample mean:* $$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

The Weak Law of Large Numbers:

For any $\varepsilon > 0$, as $n \to \infty$

$$\Pr(|\overline{X} - \mu| > \epsilon) \longrightarrow 0.$$

Heavy Tails



The normal curve has the most data points in the center, with fewer data points the further from the average.

Concentration Inequalities

- Markov Inequality

Let $X \geq 0$ be a **non-negative** random variable (discrete or continuous), and let $k > 0$. Then:

$$\mathbb{P}(X \geq k) \leq \frac{\mathbb{E}[X]}{k}$$

Equivalently (plugging in $k\mathbb{E}[X]$ for k above):

$$\mathbb{P}(X \geq k\mathbb{E}[X]) \leq \frac{1}{k}$$

Chebyshev Inequality

Let X be any random variable with expected value $\mu = \mathbb{E}[X]$ and finite variance $\mathsf{Var}(X)$. Then, for any real number $\alpha > 0$:

$$\mathbb{P}(|X - \mu| \geq \alpha) \leq \frac{\mathsf{Var}(X)}{\alpha^2}$$

Equivalently (plugging in $k\sigma$ for α above, where $\sigma = \sqrt{\mathsf{Var}(X)}$):

$$\mathbb{P}(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

Chernoff Bound (for Binomial distrib.)

Let $X \sim \text{Bin}(n, p)$ and let $\mu = \mathbb{E}[X]$. For any $0 < \delta < 1$:

Upper tail bound:

$$\mathbb{P}\left(X \geq (1+\delta)\mu\right) \leq \exp\left(-\frac{\delta^2 \mu}{3}\right)$$

Lower tail bound:

$$\mathbb{P}\left(X \leq (1-\delta)\mu\right) \leq \exp\left(-\frac{\delta^2 \mu}{2}\right)$$

where $\exp(x) = e^x$.

Hoeffding Bound

Let $Z_1, Z_2, ..., Zn$ *be independent bounded random variables such that* $Z_i \in [a_i, b_i]$ *with probability 1. Let* $S_n = \sum_{i=1}^{n} Z_i$*. Then for any* $t > 0$*, we have*

$$P(|S_n - E[S_n]| \geq t) \leq 2e^{-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}}$$

Uniform Law of Large Numbers

Let $\mathcal{F}$ be a collection of functions $f : \mathcal{X} \to \mathbb{R}$. Then $\mathcal{F}$ satisfies a *ULLN* (for a distribution P) if

$$\|P_n - P\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |P_n f - Pf| \xrightarrow{p} 0.$$

Empirical Distribution Function

Let $X_1, \ldots, X_n$ be a collection of i.i.d. random variables with cdf F_X. Then the *empirical distribution function* will be denoted $F_n(x)$, and defined for $x \in \mathbb{R}$ by

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{[X_i,\infty)}(x)$$

where $I_A(\omega)$ is the indicator function for set A.

Glivenko Cantelli Theorem

Let $X_1, \ldots, X_n$ be a collection of i.i.d. random variables with cdf F_X, and let $F_n(x)$ denote the empirical distribution function. Then, as $n \longrightarrow \infty$,

$$P\left[\sup_{x\in\mathbb{R}} |F_n(x) - F_X(x)| \longrightarrow 0\right] = 1$$

or equivalently

$$P\left[\lim_{n\longrightarrow\infty} \sup_{x\in\mathbb{R}} |F_n(x) - F_X(x)| = 0\right] = 1.$$

that is, the convergence is ***uniform in x.***