1]

Lectures 1 & 2: Probability Review

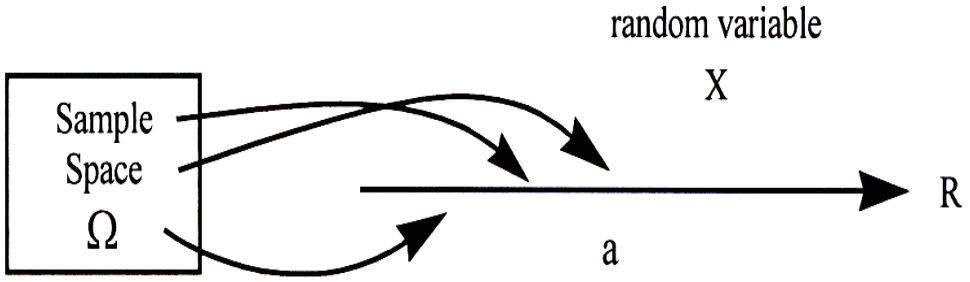
September 4, 2019

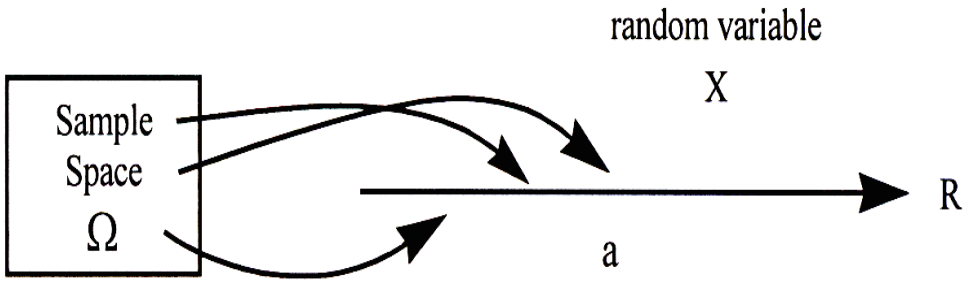
Random Variables

A **random variable** is a variable whose possible values are outcomes of a random phenomenon. For example, when tossing a fair coin, the side on which the coin lands, i.e., the outcome, is either heads or tails and is uncertain. However, since the coin must land on either heads or tails, the event of either happening has nonzero probability.

More formally, a random variable X is defined as a measurable function that maps the set of possible outcomes of a stochastic (unpredictable) process, Ω, to numerical quantities or a measurable space, R:

$$X : \Omega \to R \tag{1}$$





We say that "X is a random variable with sample space Ω that takes on values in the set R."

For example, suppose we perform an experiment where we toss a fair coin 2 times, and define the random variable to be the number of heads in the experiment. Then, $\Omega = \{HH, TT, HT, TH\}$ and the associated $R = \{2, 0, 1, 1\}$.

We also are interested in events, which are non empty subsets of Ω. For example, in the above experiment, an event A can be "1 or more tails", and its probability is the measure of how likely the event is when the experiment is conducted.

$$P(A) = P(TT \text{ or } TH \text{ or } HT) = P(TT) + P(TH) + P(HT) = \frac{3}{4}. \qquad (2)$$

$$P(A) = P(TT \text{ or } TH \text{ or } HT) = P(TT) + P(TH) + P(HT) = \frac{3}{4}. \tag{3}$$

The first equality is by definition of A, the second equality comes from the fact that each outcome in A is disjoint from every other outcome (outcomes cannot occur simultaneously), the final equality comes from the fact that each outcome is equally likely (because it is a fair coin and tosses are independent). More formally,

$$P(A) = \sum_{i=1}^{n} P(E_i), \tag{4}$$

where $E_1, E_2, ..., E_n$ are the outcomes in A, and if all outcomes are equally likely, then

$$P(A) = \frac{\text{number of outcomes in A}}{\text{number of outcomes in } \Omega}. \tag{5}$$

Necessarily, the value of probability is between 0 and 1 as the sample space, Ω is the whole possible set of outcomes.

Axioms

1. *Axiom 1:* The probability of an event is a non-negative real number:

$$P(A) \geq 0, \qquad \forall A \in \Omega. \tag{6}$$

2. *Axiom 2:* The probability of the entire sample space equals 1, i.e.,

$$P(\Omega) = 1. \tag{7}$$

3. *Axiom 3:* Any countable sequence of disjoint sets (synonymous with mutually exclusive events) $\{E_1, E_2, \ldots\}$ satisfies

$$P(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i) \tag{8}$$

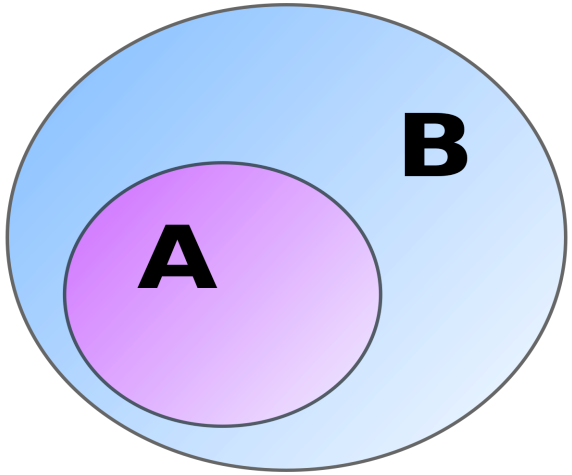
Consequences of Axioms

- *Complement of A:* The complement of an event A means $not(A)$ and is denoted as A^c. The probability of A^c means the probability of all the outcomes in sample space other than the ones in A. Thus,

$$P(A^c) = 1 - P(A) \tag{9}$$

  Given axiom 2 and the complement property, $P(\varnothing) = 1 - P(\Omega) = 0$.

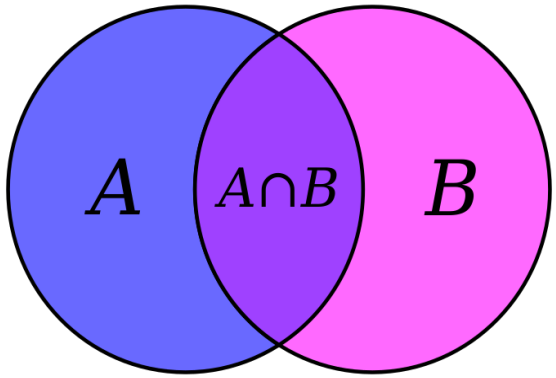
- *Monotonicity:* if $A \subseteq B$ then $P(A) \leq P(B)$.



- *Sum rule:* The probability that A or B will happen is the sum of the probabilities that A will happen and that B will happen, minus the probability that both A and B will happen.

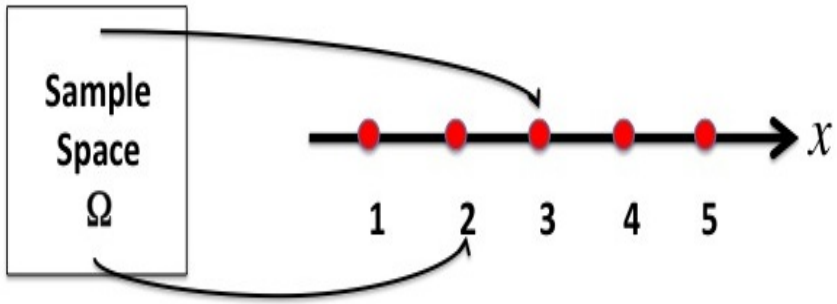
$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \tag{10}$$

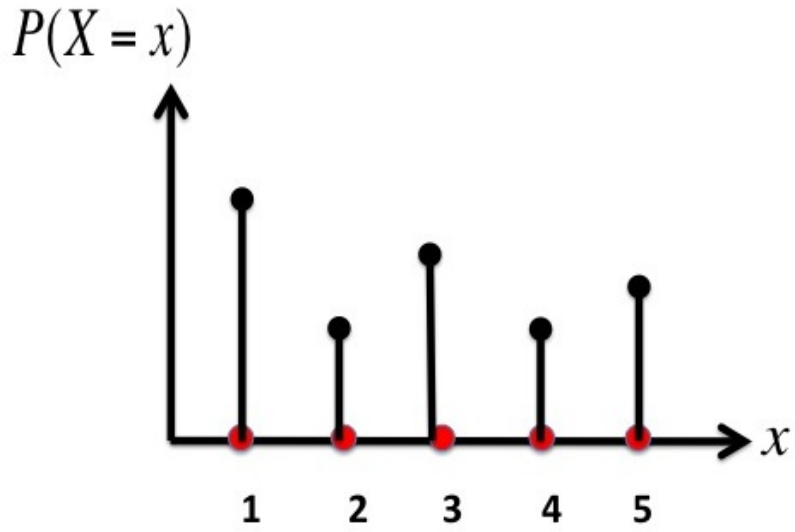
- *Independent Events*: Any two events are independent of each other if one has zero effect on the other i.e. the occurrence of one event doe not affect the occurrence of the other. If A and B are two independent events then, $P(A \cap B) = P(A)P(B)$.



Discrete Random Variables

When the image (or range) of X is finite or countably infinite, the random variable is called a *discrete random variable* and its distribution can be described by a *probability mass function (PMF)* which assigns a probability to each value in the image of X.





Bernoulli Processes

A Bernoulli process consists of a finite or infinite sequence of Bernoulli trials, which have the following properties:

- the trials are independent of each other;
- there are only two possible outcomes for each trial, arbitrarily labeled "success" (1) or "failure" (0);
- the probability of success is the same for each trial.

We denote the Bernoulli random variable on trial i as $X_i \in \{0, 1\}$, where $P(X_i = 1) = p_i$ for $i = 1, 2, 3, ...$ Consequently, $P(X_i = 0) = 1 - p_i$. Formally, we say the the Bernoulli random variable has the following PMF:

$$P(X_i = x) = \begin{cases} p & x = 1 \\ 1 - p & x = 0 \end{cases} \tag{11}$$

Expected Value and Variance of a PMF

The **expected value** or **mean** of a random variable is intuitively the average value of infinitely many repetitions of the experiment it represents. Formally, the expected value of a discrete random variable is the probability-weighted average of all possible values:

$$E(x) = \sum_x xP(X = x). \tag{12}$$

The **variance** of a random variable is the square of the average deviation from the expected value:

$$var(x) \equiv \sigma_x^2 = \sum_x (X - E(x))^2 P(X = x). \tag{13}$$

Mean and Variance of the Bernoulli Process

$$E(x) = P(X = 0) \cdot 0 + P(X = 1) \cdot 1 = (1 - p) \cdot 0 + p \cdot 1 = p. \qquad (14)$$

$$\sigma_x^2 = \sum_x (x - E(x))^2 p(X = x) = (0-p)^2 \cdot (1-p) + (1-p)^2 \cdot p = p(1-p). \quad (15)$$

Bernuolli Random Variable

$$E(x) = p, \qquad \sigma_x^2 = p(1 - p)$$

Sequences of Bernoulli Trials

Flipping a biased coin 2 times with $P(\text{heads}) = p$ is a Bernoulli process where $p_i = p$ for $i = 1, 2$.

In fact, if the sequence $\{X_1, X_2, ...\}$ are independent and identically distributed than, $p_i = p$ and we call this a sequence of *i.i.d. Bernoulli trials.*

Binomial Processes

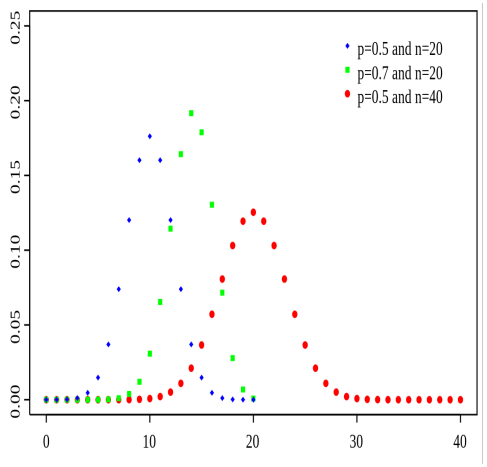
Binomial processes arise from sequences of Bernoulli trials. Specifically, the **binomial random variable**, Y, is the number of "successes" in a sequence of n independent Bernoulli trials. Therefore we have that:

$$Y = X_1 + X_2 + \ldots + X_n. \tag{16}$$

The PMF for the binomial random variable is given by:

$$P(Y = k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{17}$$

To deduce by reason, the probability that there are exactly k successes in n independent trials is $p^k(1-p)^{n-k}$, but there are $\binom{n}{k}$ ways to have k successes in n trials.



Binomial Processes

To compute the expected value, we remind ourselves that $Y = X_1 + X_2 + \ldots + X_n$. Then,

$$E(Y) = E(X_1) + E(X_2) + \ldots + E(X_n) = np, \tag{18}$$

because the expectation is a linear operator.

$$var(Y) = var(X_1) + var(X_2) + \ldots + var(X_n) = np(1-p). \tag{19}$$

The above is true only if X_i for $i = 1, 2, ..., n$ are mutually independent random variables.

Binomial Random Variable

$$E(x) = np, \qquad \sigma_x^2 = np(1-p)$$

Poisson Processes

A **Poisson process** is a stochastic process of binary events on a continuum that satisfies the following properties:

1. The probability distribution for the number of events in an interval is given by a Poisson distribution:

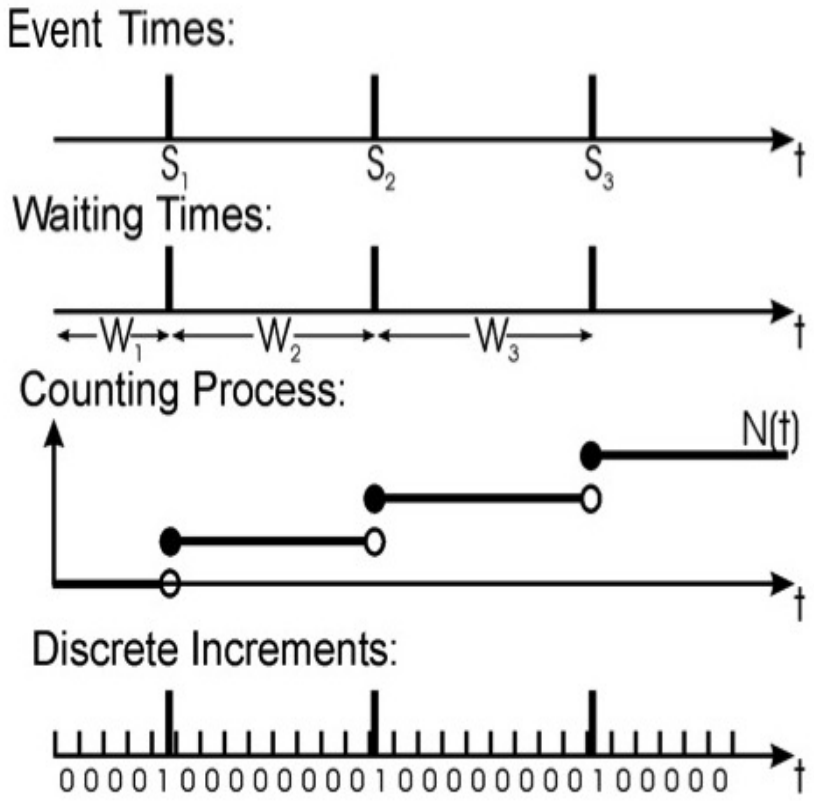
$$P(n \text{ events in } \ (t_i, t_i + \Delta t]) = \frac{(\lambda_i \Delta t)^n \cdot e^{-\lambda_i \Delta t}}{n!}. \qquad (20)$$

2. The number of events in non-overlapping intervals are independent.

3. If λ_i is constant, i.e.$\lambda_i = \lambda$, the waiting time between events is exponentially distributed with parameter λ.

Poisson Processes

There are many random variables that can be defined on a Poisson process as shown below.



Poisson Processes

The number of events in an interval of length Δt has the following expected value:

$$E(n) = \sum_{n=0}^{\infty} n \frac{(\lambda_i \Delta t)^n \cdot e^{-\lambda_i \Delta t}}{n!} = e^{-\lambda_i \Delta t} \sum_{n=1}^{\infty} \frac{(\lambda_i \Delta t)^n}{(n-1)!} \tag{21}$$

$$= (\lambda_i \Delta t) e^{-\lambda_i \Delta t} \sum_{n=1}^{\infty} \frac{(\lambda_i \Delta t)^{n-1}}{(n-1)!} = (\lambda_i \Delta t) e^{-\lambda_i \Delta t} \sum_{m=0}^{\infty} \frac{(\lambda_i \Delta t)^m}{(m)!} \tag{22}$$

$$= (\lambda_i \Delta t). \tag{23}$$

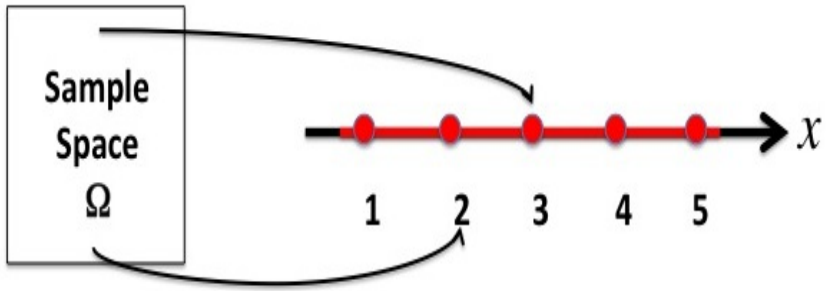
The moment generating function allows us to easily show that $var(n) = E(x) = (\lambda_i \Delta t)$.

Poisson Random Variable

$$E(x) = \sigma_x^2 = (\lambda_i \Delta t)$$

Continuous Random Variables

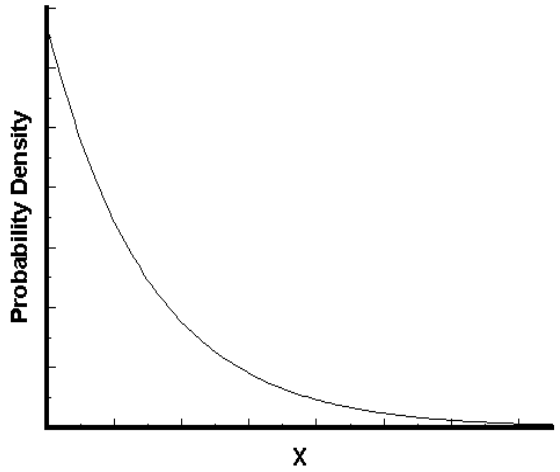
If the image is uncountably infinite then X is called a *continuous random variable*. In the special case that it is absolutely continuous (all of our cases in this course!), its distribution can be described by a *probability density function*, which assigns probabilities to intervals; in particular, each individual point must necessarily have probability zero for an absolutely continuous random variable.



Exponential Random Variable

The exponential distribution (also known as negative exponential distribution) is the probability density function (pdf) that describes the **time,** X_i **between events** in a Poisson point process:

$$f(x;\lambda) = \lambda e^{\lambda x} \quad x \geq 0. \tag{24}$$



We write that $X \sim exp(\lambda)$.

Expected Value and Variance of a PDF

The **expected value** or **mean** of a random variable is intuitively the average value of infinitely many repetitions of the experiment it represents. Formally, the expected value of a continuous random variable is the density-weighted average of all possible values:

$$E(x) = \int_x x f(x) dx. \tag{25}$$

The **variance** of a random variable is the square of the average deviation from the expected value:

$$var(x) \equiv \sigma_x^2 = \int_x (x - E(x))^2 f(x) dx. \tag{26}$$

Mean and Variance of the Exponential RV

$$E(x) = \int_x x\lambda e^{\lambda x} dx = \frac{1}{\lambda}. \tag{27}$$

One can show the above equality by integration by parts.

$$var(x) \equiv \sigma_x^2 = \int_x (x-\lambda)^2 \lambda e^{\lambda x} dx = \frac{1}{\lambda^2}. \tag{28}$$

One can show the above equality by using the fact that $\sigma_x^2 = E(x^2) - E(x)$ where $E(x^2) = \int_x x^2 f(x) dx$, which can also be solved by integration by parts.

Exponential Random Variable

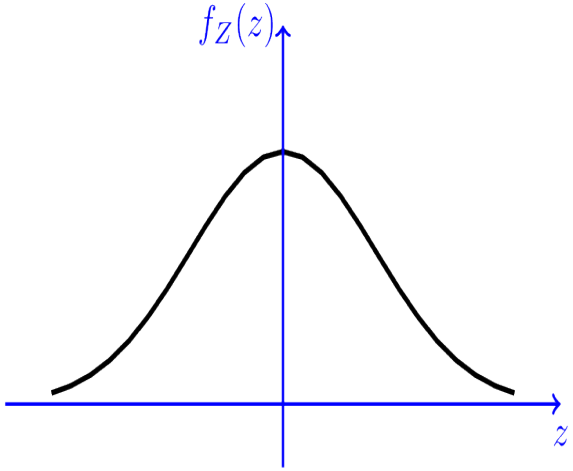
$$E(x) = \tfrac{1}{\lambda}, \quad \sigma_x^2 = \tfrac{1}{\lambda^2}$$

Gaussian Random Variable

The Gaussian random variable is also called the **normal random variable**. We will start by defining the *standard normal* random variable, and then obtain other normal random variables by scaling and shifting a standard normal random variable.

A continuous random variable Z is said to be a standard normal random variable, shown as $Z \sim N(0,1)$, if its PDF is given by

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \tag{29}$$



Mean and Variance of the Standard Normal RV

$$E(x) = \int_{-\infty} z \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = 0. \tag{30}$$

One can show the above equality by a symmetry argument (center of mass of pdf is at 0).

$$\sigma_x^2 = \int_{-\infty} (z-0)^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = 1 \tag{31}$$

One can show the above using integration by parts.

Normal Random Variables

Now that we have seen the standard normal random variable, we can obtain any normal random variable by shifting and scaling a standard normal random variable. In particular, define

$$X = \sigma Z + \mu \quad \text{where} \quad \sigma > 0. \tag{32}$$

Then, we have that

$$E(x) = \sigma E(Z) + \mu = \mu, \tag{33}$$

$$var(x) = \sigma^2 var(Z) = \sigma^2. \tag{34}$$

The latter equality comes from the fact that $var(ax) = a^2 var(x)$ for any real constant a.

Normal Random Variables

$$X = \sigma Z + \mu \quad \text{where} \quad \sigma > 0. \tag{35}$$

Then, we can show that the pdf of X is

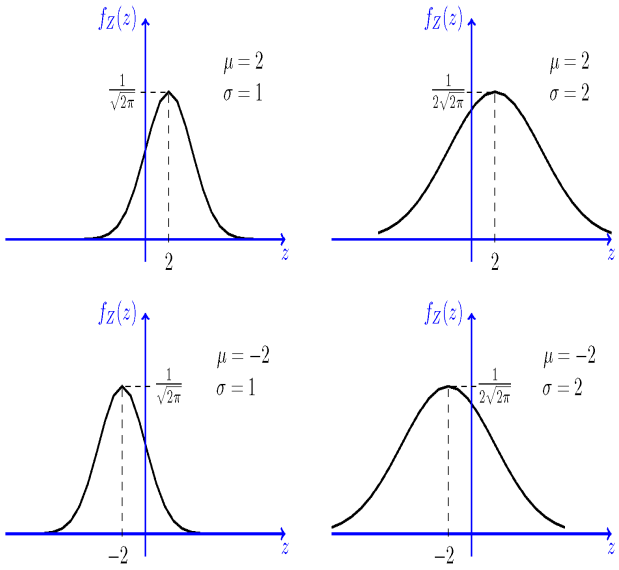
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(z-\mu)^2}{2\sigma^2}} \tag{36}$$

Gaussian Random Variable

$$E(x) = \mu \quad var(x) = \sigma^2$$

We say that $X \sim N(\mu, \sigma^2)$.

Normal Random Variables



An important and useful property of the normal distribution is that a linear transformation of a normal random variable is itself a normal random variable. In particular, we have the following theorem:

Theorem

If $X \sim N(\mu_x, \sigma_x^2)$, **and** $Y = aX + b$ **for** a, b **real constants, then** $Y \sim N(\mu_y, \sigma_y^2)$, **where** $\mu_y = a\mu_x + b$, **and** $\sigma_y^2 = a^2\sigma_x^2$.

Statistics and the Data Likelihood Function

A statistic is any function of data. In this course, you will all be working with data and computing statistics. Then, you will be constructing models from statistics, which entails (i) assuming a distribution that the statistic is drawn from, and then (ii) estimating the PMF or PDF from observed data.

For example, consider an experiment where a biased coin is flipped 100 times. This experiment will generate the following set of *i.i.d* samples of a Bernoulli random variable with some probability of heads p:

$$x_1, x_2, x_3, \ldots, x_{100} \tag{37}$$

The problem we wish to solve is to estimate p, which in turns provides and estimate of the PMF for the random variable. An intuitive choice for p would be to compute the sample mean, which gives:

$$\widehat{p} = \frac{1}{100} \sum_{i=1}^{100} x_i = \frac{n_h}{100} \tag{38}$$

where n_h is the number of heads in the 100 tosses.

The Data Likelihood Function

One formal procedure for estimating p is called **Maximum Likelihood Estimation (MLE).** Formally, MLE is a procedure for estimating a PMF or PDF from data. In our example, we wish to construct a function of p that describes the probability of observing the specific set of the 100 samples (the likelihood of observing), and then maximizing it over p.

We define $l(p; \text{observed data})$ to be the **data likelihood function** for our bernoulli observations:

$$l(p; \text{observed data}) = \prod_{i=1}^{100} p_i^x (1-p)^{1-x_i} \tag{39}$$

Then,

$$\widehat{p}_{ML} = \arg\max_p \; l(p) \quad = \quad \arg\max_p \; \log l(p) \tag{40}$$

Maximum Likelihood Estimation

$$l(p; \text{observed data}) = \prod_{i=1}^{100} p^{x_i}(1-p)^{1-x_i} \tag{41}$$

We can take the derivative of $\log l(p)$ with respect to p to maximize:

$$\log l(p) = \sum_{i=1}^{100} x_i log(p) + (1-x_i) log(1-p) \tag{42}$$

to get $\widehat{p}_{ML}$,

Maximum Likelihood Estimation

$$\frac{d \log l(p)}{dp} = \sum_{i=1}^{100} \frac{x_i}{p} + (1 - x_i)\frac{-1}{(1-p)} \tag{43}$$

$$= \frac{1}{p}\sum_{i=1}^{100} x_i - \frac{1}{(1-p)}\sum_{i=1}^{100}(1 - x_i) \tag{44}$$

$$= \frac{1}{p} \cdot n_h - \frac{1}{(1-p)} \cdot (100 - n_h) \tag{45}$$

$$= n_h(\frac{1}{p} + \frac{1}{(1-p)}) - 100\frac{1}{(1-p)} \tag{46}$$

$$= n_h\frac{1}{p(1-p)} + -100\frac{1}{(1-p)} \tag{47}$$

$$\tag{48}$$

Now we set $\frac{d \log l(p)}{dp} = 0$ and get that

$$n_h\tfrac{1}{p(1-p)} + -100\tfrac{1}{(1-p)} = 0 \tag{49}$$

$$n_h\tfrac{1}{p} + -100 = 0 \tag{50}$$

$$\rightarrow \qquad \widehat{p}_{ML} = \tfrac{n_h}{100} \tag{51}$$

Maximum Likelihood Estimation

Now, let's consider *i.i.d* samples of an exponential random variable $x_1, x_2, x_3, \ldots, x_{100}$, with unknown parameter λ. The data likelihood function is

$$l(\lambda; \text{observed data}) = \prod_{i=1}^{100} \lambda e^{-\lambda x_i} \tag{52}$$

Thus,

$$\log l(\lambda) = \sum_{i=1}^{100} log(\lambda) - \lambda x_i = N log(\lambda) - \lambda \sum_{i=1}^{100} x_i \tag{53}$$

Then, taking the derivative with respect to λ, we get

$$\frac{d \log l(\lambda)}{d\lambda} = \frac{N}{\lambda} - \sum_{i=1}^{100} x_i = 0 \tag{54}$$

which gives $\widehat{\lambda}_{ML} = \frac{N}{\sum_{i=1}^{100} x_i}$. Recall that $E(x) = \frac{1}{\lambda}$, thus the ML estimate of meand of the PDF is the sample mean!

Maximum Likelihood Estimation

Now, let's consider *i.i.d* samples of a Gaussian random variable with unknown parameters μ and σ^2. The data likelihood function is

$$L(\lambda; \text{observed data}) = \prod_{i=1}^{100} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$\log l(\lambda) = \sum_{i=1}^{100} log(1) - \frac{1}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x_i - \mu)^2 \tag{55}$$

Then, taking derivatives with respect to μ and σ^2, we get

$$\mu_{ML} = \frac{1}{N} \sum_{i=1}^{100} x_i, \qquad \sigma^2_{ML} = \frac{1}{N} \sum_{i=1}^{100} (x_i - \mu_{ML})^2$$

Maximum Likelihood Estimation Properties

As the sample size, $n \to \infty$, sequences of maximum likelihood estimators have these properties:

- *Consistency:* If the data were generated by $f(\cdot\,;\theta_0)f(\cdot\,;\theta_0)$ and we have a sufficiently large number of observations n, then it is possible to find the value of θ_0 with arbitrary precision. In mathematical terms this means that as n goes to infinity the estimator $\widehat{\theta}_{ML}$ converges in probability to its true value θ_0. *caveat: we never know f and must guess, often models are wrong!*
- *Convergence to normal distribution:* Additionally, if (as assumed above) the data were generated by $f(\cdot\,;\theta_0)$, then under certain conditions, it can also be shown that the maximum likelihood estimator converges in distribution to a normal distribution.

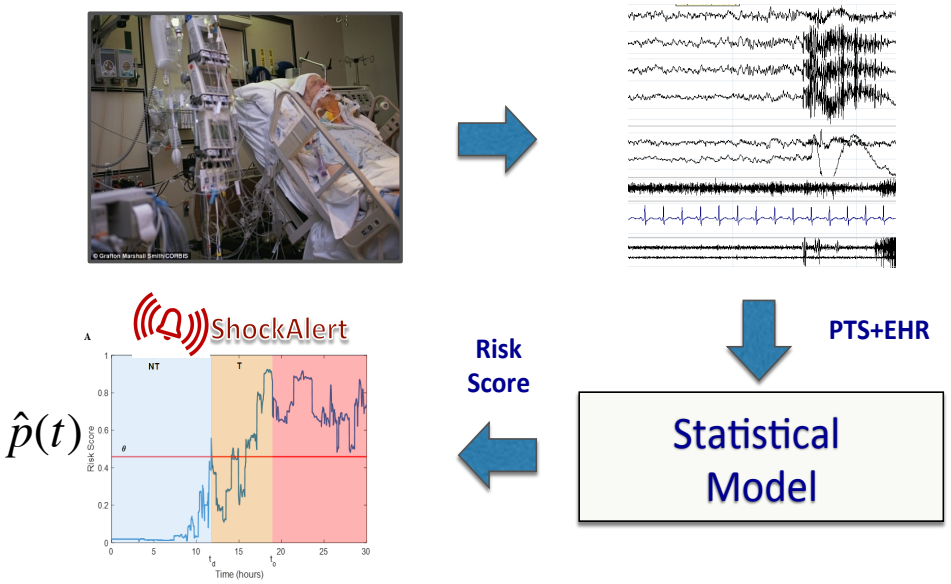
$$\sqrt{n}\big(\widehat{\theta}_{\text{mle}} - \theta_0\big) \xrightarrow{d} \mathcal{N}(0,\, I^{-1}) \tag{56}$$

  where I is the Fisher information matrix.
- *Efficiency:* i.e., it achieves the Cramer-Rao lower bound (lower bound on the variance of unbiased estimators of a deterministic fixed, but unknown parameter) when the sample size tends to infinity: $\text{var}(\widehat{\theta}) \geq \dfrac{1}{I(\theta)}$. This means that no consistent estimator has lower asymptotic mean squared error than the MLE (or other estimators attaining this bound).

MLE Applied in Clinical Application

In this class, several projects will entail estimation a Bernoulli PMF from observed data and then using the MLE of p to perform binary classification.

One example comes from estimating if and when an ICU patient is going to go into septic shock given that he/she has sepsis from the patient's electronic health record (EHR) and vital sign measurements (e.g. arterial blood pressure, heart rate, etc) that may be measured every minute.



MLE Applied in Clinical Application

Let the EHR data be represented by an n-dimensional vector $[x_1, x_2, ..., x_n]$, and let the physiological waveform data be represented by an m-dimensional time-varying vector $[w_1(t), w_2(t), ..., w_n(t)]$. Then, from observations of $\{x_i\}_i$ and $\{w_j(t)\}_j$, we are interested in estimating

$$\widehat{p}(t; \mathbf{x}, \mathbf{w}) = f(\mathbf{x}, \mathbf{w}(t)) \tag{57}$$

for each patient. We call $\widehat{p}(t; x, w)$ the risk score for the patient at time t. This risk score can then be used to detect if and when a patient may transition to shock.

How do we choose f? Stay tuned...