Chapter 1 Statistical Reasoning

- Why statistics?
  - Uncertainty of nature (weather, earth movement, etc.)
  - Uncertainty in observation/sampling/measurement
  - Variability of human operation/error
  - imperfection of machines, etc.
- Section 1.1 Basics of Probability Theory

  (http://www.math.uiuc.edu/~r-ash/BPT/BPT.pdf)
  - The classical definition of probability states that the probability of an event is the number of outcomes favorable to the event, divided by the total number of outcomes, where all outcomes are equally likely. For instance, in coin toss, the probability of seeing head = 1/2. This definition is very restrictive: it considers only experiments with a finite number of outcomes, and, more seriously, circular (no matter how you look at it "equally likely" essentially means "equally probable.") The concept of "equally probable" involves the concept of probability. Thus we are using the concept of probability to define probability itself.

- Mathematical definition of probability:
  - $*\ \Omega$ – a set (e.g., R^n,part of integers, etc.) called sample space representing all possible outcome.

- In coin toss example, $\Omega = \{head,\ tail\}$
- In dice toss, $\Omega = \{1, 2, 3, 4, 5, 6\}$. In an experiment of tossing a dice, one may define other outcomes. For instance, N representing even and O representing odd number. So outcome could be a subset of Ω.This leads to "events"
- An event is a subset of the sample space.

* $\mathcal{F}$ – collection of events
  - $\mathcal{F}$ must be closed in set operations (Boolean algebra), and must contain Ω and $\varnothing$ (empty set). For instance, if $A, B$are two events, so are $A \cap B$, $A \cup B$, $A^c = \Omega \backslash A$, $B^c = \Omega \backslash B$,etc.
  - We call such $\mathcal{F}$ a $\sigma - field$, or Boolean field
* $P(A)$ is a countable additive non-negative function defined for every event A, $P(\Omega) = 1$.
  - countable additive means that

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i) \quad if \quad A_i \cap A_j = \varnothing \quad for \quad any\ i < j$$

· This implies: $P(A^c) = 1 - P(A)$. If $A_1 \subset A_2 \subset A_3 \subset \cdots$, then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = P\left(A_1 \cup \bigcup_{i=1}^{\infty} (A_{i+1} - A_i)\right)$$
$$= P(A_1) + \sum_{i=1}^{\infty} P(A_{i+1} - A_i)$$
$$= P(A_1) + \sum_{i=1}^{\infty} (P(A_{i+1}) - P(A_i)) = \lim_{i \to \infty} P(A_i)$$

· This last property is called lower continuity, or continuous from below.

· If $A_1 \supset A_2 \supset A_3 \supset ...$, and let $A = \bigcap_{n=1}^{\infty} A_n$, then

$$A_1 = A \cup (A_1 - A_2) \cup (A_2 - A_3) \cup (A_3 - A_4) \cup ... = A \cup \bigcup_{n=1}^{\infty} (A_n - A_{n+1})$$

So

$$P(A_1) = P(A) + \sum_{n=1}^{\infty} P(A_n - A_{n+1}) = P(A) + P(A_1) - \lim_{n \to \infty} P(A_{n+1})$$

- This limit $\lim_{n\to\infty} P(A_n) = P(A)$ is called upper continuity.
- Any non-negative countable additive function is called a measure. When the measure of the whole space = 1, it is called a probability measure.
- One (classical) "definition" of P is

$$P(A) = \frac{number\ of\ points\ in\ A}{total\ number\ of\ points\ in\ \Omega} = \frac{favorable\ outcomes}{total\ outcomes}$$

* $(\Omega,\ \mathcal{F},\ P)$ is called a probability space. If you are familiar with real analysis, a probability space is a measurable space with total measure one. An event is called a measurable set. Note that there are non-measurable sets.
* Recall that in real analysis, we may define integration based on this measure P

$$\int_\Omega g(\omega)\, dP = \lim \sum_i x_i P(\{\omega :\ x_i \le g(\omega) \le x_{i+1}\})$$

In particular, if $g(x) = \mathcal{X}_A(x)$, characteristic function of A,then

$$\int_\Omega \mathcal{X}_A(x)\, dP = P(A)$$

* A function $X(\omega)$ defined on $\omega \in \Omega$ is called measurable function if for any real number x, $\{\omega :\ X(\omega) \le x\} \in \mathcal{F}$. We view such a function is a measurement for outcomes associated

with an experiment. In probability theory, it is called a random variable.

* Two events A and B are called independent if $P(A \cap B) = P(A) P(B)$
* Conditional probability of B given A: $P(B \mid A) = P(A \cap B) / P(A)$
* Theory of Total Probability: Let $B_1,\ B_2,\ ...$be a finite or countably infinite family of mutually exclusive and exhaustive events (i.e., disjoint and their union is Ω). Then

$$P(A) = \sum_i P(A\ \cap\ B_i) = \sum_i P(B_i) P(A \mid B_i).$$

* We define

$$F_X(x) = P(\{\omega :\ X(\omega) \leq x\}).$$

This is called cumulative $probability$ distribution function CDF of the random variable.

* F_X is continuous from the right. This is because for integers $n > m$, $1/n < 1/m$

$$\{\omega :\ X(\omega) \leq x\} \subset \left\{\omega :\ X(\omega) \leq x + \frac{1}{n}\right\} \subset \left\{\omega :\ X(\omega) \leq x + \frac{1}{m}\right\}$$

So

$$\{\omega :\ X(\omega) \leq x\} \subset \bigcap_{n=1}^{\infty} \left\{\omega :\ X(\omega) \leq x + \frac{1}{n}\right\}$$

On the other hand, for any ω_0 such that

$$\omega_0 \in \bigcap_{n=1}^{\infty} \left\{ \omega : \ X(\omega) \leq x + \frac{1}{n} \right\},$$

it must hold

$$X(\omega_0) \leq x$$

since otherwise

$$X(\omega_0) > x.$$

Then one can find a large integer n such that (e.g., $n = [(X(\omega_0) - x)^{-1} + 2] > (X(\omega_0) - x)^{-1}$)

$$X(\omega_0) > x + \frac{1}{n} \quad \Longrightarrow \quad \omega_0 \notin \left\{ \omega : \ X(\omega) \leq x + \frac{1}{n} \right\}$$

a contradiction. Thus

$$\{\omega : \ X(\omega) \leq x\} = \bigcap_{n=1}^{\infty} \left\{ \omega : \ X(\omega) \leq x + \frac{1}{n} \right\}$$

According to the upper continuity of P, it follows that

$$P\left(\{\omega:\ X(\omega)\leq x\}\right)=\lim_{n\to\infty}P\left(\left\{\omega:\ X(\omega)\leq x+\frac{1}{n}\right\}\right)$$

or equivalently

$$F_X(x)=\lim_{n\to\infty}F_X\left(x+\frac{1}{n}\right)$$

∗ Exercise 1: Define

$$G(x)=P\left(\{\omega:\ X(\omega)<x\}\right). \qquad \text{(Exercise 1)}$$

Show that $G(x)$ is continuous from left.

∗ For right-continuous function $F_X(x)$, one may define the Lebesgue–Stieltjes integration for any function $g(x)$

$$\begin{aligned}\int_\Omega g(x)\,dF_X&=\lim\sum g(x_i)\left(F_X(x_{i+1})-F_X(x_i)\right)\\&=\lim\sum g(x_i)\,P\left(\{\omega:\ x_i\leq X(\omega)\leq x_{i+1}\}\right)=\int_\Omega g(X)\,dP\end{aligned}$$

In particular, when $g = \mathcal{X}_A$,the indicator function of an event A,

$$\int_A dF_X = P\left(\{\omega : \ X(\omega) \in A\}\right)$$

Since $\{\omega : \ a < X(\omega) \leq b\} = \{\omega : \ X(\omega) \leq b\} \setminus \{\omega : \ X(\omega) \leq a\}$,we have

$$P\left(\{\omega : \ a < X(\omega) \leq b\}\right) = F_X(b) - F_X(a)$$

Note that since $F_X(x)$ may not be continuous from the left,

$$P\left(\{\omega : \ a \leq X(\omega) \leq b\}\right) = F_X(b) - F_X\left(a^-\right)$$

$*$ One may view $F_X(x)$ is new measurement for $X : \ [F_X(a), F_X(b)] \quad replaces \quad [a, b]$
In particular,

$$F_X(x) = \int_{-\infty}^{x} dF_X$$

$*$ $F_X(x)$ is always non-negative, increasing, and

$$\lim_{x \to -\infty} F_X(x) = 0, \quad \lim_{x \to \infty} F_X(x) = 1$$

* We call F_X absolutely continuous, if there exists a function such that $\rho(x)$

$$F_X(x) = \int_{-\infty}^{x} dF_X = \int_{-\infty}^{x} \rho(t)\, dt$$

* $\rho_X(x)$ is called probability density function (PDF). One can show that ρ is in fact the derivative of F_X :

$$\begin{aligned}\rho_X(x) &= \lim_{\delta \to 0} \frac{F_X(x+\delta) - F_X(x)}{\delta} \\ &= \lim_{\delta \to 0} \frac{P(\{\omega :\ x < X(\omega) \leq x+\delta\})}{\delta} = F_X'(x)\end{aligned}$$

* $\rho_X(x)$ may be understood as the per-unit probability of $X = x$
* For any interval A

$$P\left(X^{-1}(A)\right) = P(\{\omega : X(\omega) \in A\}) = \int_A dF_X(x) = \int_A \rho_X(x)\, dx.$$

For any function $g(x)$

$$\int_\Omega g(X(\omega))\, dP = \int_{-\infty}^{\infty} g(x)\, \rho_X(x)\, dx$$

* This formula established connection between Probability measure P and PDF p. From this

point, we don't need to care about P, nor about X.All we need is ρ (or F). In application, CDF $F_X(x)$(or PDF) has all information about random variable $X(\omega)$. For instance, $X(\omega) =$ height of a person, $F_X(x)$ = percent of persons height is less than x.

$*$ Expected value, or expectation of X,is defined as

$$\mu = E[X] = \int_{-\infty}^{\infty} x\rho_X(x)\,dx = \int_{\Omega} X(\omega)\,dP$$

$E(X)$ means average value of X,or center of mass of ρ_X

· If X takes only finite many values x_i, then

$$E[X] = \int_{\Omega} X(\omega)\,dP = \sum x_i P(X = x_i)$$

· $E[X]$ is linear

$$E[aX + bY] = aE[X] + bE[Y]$$

· In particular, $E[X - \mu] = 0$

· $R = X - \mu$ is called residue of X. Reynolds decomposition:

$$X = \mu + R$$

· for any function $g(x)$, the expected value of random variable $Y = g(X)$ is

$$E[Y] = \int_{-\infty}^{\infty} g(x)\,\rho_X(x)\,dx = \int_{\Omega} g(X(\omega))\,dP.$$

· n th moment of X

$$E[X^n] = \int_{-\infty}^{\infty} x^n \rho_X(x)\,dx = \int_{\Omega} X(\omega)^n\,dP$$

· Variance or the moment of inertia

$$\begin{aligned} \upsilon = \sigma^2 &= E\left[(X-\mu)^2\right] \\ &= \int_{-\infty}^{\infty} (x-\mu)^2\,\rho_X(x)\,dx = \int_{\Omega} (X(\omega)-\mu)^2\,dP \end{aligned}$$

· σ is called standard deviation of X

· median $m(X)$ is a number such that $F_X(m) = 1/2$

$*$ Joint distribution function: Let X and Y are two random variables. The joint cumulative $probability$ distribution function is

$$\begin{aligned} F(x,y) &= P(X \le x,\ Y \le y) \\ &= P(\{\omega : X(\omega) \le x \ \text{ and } \ Y(\omega) \le y\}) \end{aligned}$$

- We say $F(x,y)$ is absolutely continuous if there exists a probability density function PDF $\rho(x,y)$ such that for any intervals $A,\ B$

$$P(X \in A,\ Y \in B) = \int\int_{A\times B} \rho(x,y)\,dxdy$$

so

$$F(x,y) = \int_{-\infty}^{y}\int_{-\infty}^{x} \rho(x,y)\,dxdy$$

$$\int_{\Omega} g(X,Y)\,dP = \int\int_{R^2} g(x,y)\,\rho(x,y)\,dxdy$$

- Two random variables are called independent if $\{\omega : X \in A\}$ and $\{\omega : Y \in B\}$ are independent events for any intervals $A,\ B$. In this case

$$F(x,y) = F_X(x)\,F_Y(y),\ \rho(x,y) = \rho_X(x)\,\rho_Y(y)$$

* Covariance between random variables: for random variables $X_{1,}X_2, ..., X_n$

$$\begin{aligned} Cov\left(X_i, X_j\right) &= E\left[\left(X_i - \mu_i\right)\left(X_j - \mu_j\right)\right] \\ &= \int_{\Omega} \left(X_i - \mu_i\right)\left(X_j - \mu_j\right) dP \\ &= \int \int_{R^2} \left(x - \mu_i\right)\left(y - \mu_j\right) \rho_{ij}\left(x, y\right) dxdy \end{aligned}$$

  is called covariance between X_i and X_j.

  · $Cov\left(X_i, X_j\right) = \upsilon\left(X_i, X_j\right)$ measures relative dependence between X_i and X_j

  · If X_i and X_j are independent, then $Cov\left(X_i, X_j\right) = 0$

  · $Cov\left(X_i, X_i\right) = \upsilon\left(X_i\right)$ = variance

  · the matrix $V = \left[Cov\left(X_i, X_j\right)\right]$ is called covariance matrix

$$V = \begin{bmatrix} \upsilon\left(X_1\right) & \upsilon\left(X_1, X_2\right) & \cdots & \upsilon\left(X_1, X_n\right) \\ \upsilon\left(X_2, X_1\right) & \upsilon\left(X_2\right) & \cdots & \upsilon\left(X_2, X_n\right) \\ \vdots & \vdots & \ddots & \vdots \\ \upsilon\left(X_n, X_1\right) & \upsilon\left(X_n, X_2\right) & \cdots & \upsilon\left(X_n\right) \end{bmatrix}$$

* random process = stochastic process: $\left\{X\left(t\right)\right\}_{t \geq 0}$. For each $t,\ X\left(t\right)$ is a random variable, and it is "continuous" with respective to t

* Random field $X\left(x\right)$: for each $x,\ X\left(x\right)$ is a random variable. So $\upsilon\left(x, y\right) = Cov\left(X\left(x\right), X\left(y\right)\right)$ is

a function of two variable.

* Example: Let Ω_i for $i = 1, 2, ..., n$, be disjoint regions, and K_i be random variable representing conductivity of Ω_i with mean μ_i and variance σ_i^2. Then the over conductivity of the entire region $\Omega = \cup\Omega_i$ is

$$K(x) = \sum_i \mathcal{X}_{\Omega_i}(x) K_i$$

The over all mean is

$$\mu = E[K] = \sum_i E[\mathcal{X}_{\Omega_i} K_i] = \sum_i E[\mathcal{X}_{\Omega_i}] E[K_i] = \sum_i I_i \mu_i$$

where $I_i = proportion$ of Ω_i, assuming independence of $\mathcal{X}_{\Omega_i}$ and K_i, and

$$E[\mathcal{X}_{\Omega_i}] = \int_\Omega \mathcal{X}_{\Omega_i} dP = \int_{\Omega_i} dP = I_i.$$

The covariance between different locations x and y :

$$\sigma^2(K) = E\left[\sum_{i=1}^{n}\left(\mathcal{X}_{\Omega_i}(x)K_i - I_i\mu_i\right)\right]^2 \qquad \text{(Optional Exercise)}$$

$$= \sum_{i=1}^{n} I_i\sigma_i^2 + \frac{1}{2}\sum_{i=1}^{n} I_i I_j\left(\mu_i - \mu_j\right)^2$$

- # Section 1.2 Uniform Distributions
  - Given interval $[a,b]$,consider cumulative distribution function

$$F_{[a,b]}(x) = \begin{cases} 0 & \text{if } x < a \\ \dfrac{x-a}{b-a} & \text{if } a \le x \le b \\ 1 & \text{if } b < x \end{cases}$$

The random variable X associated with F is said to be uniformly distributed on $[a,b]$,and is denoted as $X \to U[a,b]$.

– PDF is

$$\rho(x)\begin{cases} 0 & \text{if } x < a \\ \dfrac{1}{b-a} & \text{if } a \le x \le b \\ 0 & \text{if } b < x \end{cases}$$

– mean

$$E[X] = \int_{-\infty}^{\infty} x\rho(x)\,dx = \int_a^b \frac{x}{b-a}dx = \frac{b+a}{2}$$

– variance

$$\sigma^2 = \int_{-\infty}^{\infty}\left(x - \frac{b-a}{2}\right)^2 \rho(x)\,dx = \frac{1}{b-a}\int_a^b d\left(x - \frac{b-a}{2}\right)^2 x = \frac{(b-a)^2}{12}$$

– Standard deviation $\upsilon = (b-a)/\sqrt{12}$

– In particular, if $X \to U[0,1]$ is called the standard uniform distribution.

– Let $X \to F$, and $Y = (b-a)X + a \to G$, then

$$G(x) = P\{Y = (b-a)X + a \le x\} = P\left\{X \le \frac{x-a}{b-a}\right\} = F\left(\frac{x-a}{b-a}\right) \qquad (b > a)$$

So if $X \to U[0,1]$, then

$$(b-a)X + a \to U[a,b]$$

• Section 1.3 Gaussian Distributions

– PDF of X

$$\rho = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)}$$

– Exercise 2 : show that

$$mean = \mu,\ variance = \sigma^2 \qquad \text{(Exercise 2)}$$

– Denote it by $N(\mu, \sigma)$

– Let $Y = (X-\mu)/\sigma$.Then

$$E[Y] = \frac{E[X] - \mu}{\sigma}$$

Is called a standard normal distribution with zero mean and standard deviation 1,i.e., $N(0,1)$

– n-σ process:

$$P(|X-\mu| \le n\sigma) = P(|Y| \le n) = \frac{1}{\sqrt{2\pi}} \int_{-n}^{n} e^{-x^2/2} dx \to 1 \text{ as n} \to \infty$$

the probability of $X(x)$ lies in n-σ neighborhood of its mean:

$$\frac{1}{\sqrt{2\pi}}\int_{-2}^{2} e^{-x^2/2}dx = 0.954\,50$$
$$\frac{1}{\sqrt{2\pi}}\int_{-4}^{4} e^{-x^2/2}dx = 0.999\,94$$
$$\frac{1}{\sqrt{2\pi}}\int_{-6}^{6} e^{-x^2/2}dx = 0.9999966 = 1 - 3.4\times 10^{-6}$$

– In industry, a Six Sigma process describes quantitatively how a process is performing. To achieve Six Sigma, a process must not produce more than 3.4 defects per million opportunities. A Six Sigma defect is defined as anything outside of customer specifications.

- Exponential distribution

$$\rho(x) = \frac{1}{\lambda}e^{-x/\lambda} \text{ for } x \geq 0, \quad \rho(x) = 0 \text{ for } x < 0$$

– Exercise 3: Show that

$$E[X] = \lambda, \upsilon = \lambda^2 \qquad \text{(Exercise 3)}$$

– $E[X^n] = \lambda^n n!$

– $m(X) = \lambda \ln 2 < E[X]$

– momorylessness: it satisfies (Exercise 4)

$$P(X > s + t \mid X > s) = P(X > t) \qquad \text{(Exercise 4)}$$

(recall that the conditional probability $P(A \mid B) = P(A \cap B)/P(B)$). When X is interpreted as the waiting time for an event to occur relative to some initial time, this relation implies that, if X is conditioned on a failure to observe the event over some initial period of time s, the distribution of the remaining waiting time is the same as the original unconditional distribution. For example, if an event has not occurred after 30 seconds, the conditional probability that occurrence will take at least 10 more seconds is equal to the unconditional probability of observing the event more than 10 seconds relative to the initial time.

– The exponential distribution the only memoryless continuous probability distributions.

- # Section 1.4 The Binomial Distribution (two outcomes)

– Consider n devices, each has a failure probability p.Then the probability that exact k devices fail is

$$p(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

The discrete random variable X with this PDF is called binomial distribution, and

$$F(x) = \sum_{k \leq x} \binom{n}{k} p^k (1-p)^{n-k}$$

– Mean is (Exercise 5)

$$E[X] = \sum_{k=0}^{n} kp(k) = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} = pn \qquad \text{(Exercise 5)}$$

– variance (Exercise 6)

$$\upsilon = \sigma^2 = \sum_{k=0}^{n} (k-np)^2 p(k) = \sum_{k=0}^{n} (k-np)^2 \binom{n}{k} p^k (1-p)^{n-k} = np(1-p) \qquad \text{(Exercise 6)}$$

– Problem A in page 7

– Newsboy problem in page 8 of textbook (Exercise 7: run MATLAB routines)

- Geometric distribution
  – Suppose that each trial the probability of success is p.The probability in a sequence of trials that the first occurrence of success requires k number ($k = 1, 2, ...$) of independent trials is

$$p(k) = P(X = k) = (1-p)^{k-1} p$$

– Exercise 8: show

$$E[X] = 1/p, \sigma^2 = (1-p)/p^2 \qquad \text{(Exercise 8)}$$

– Exercise 9: The geometric distribution is memoryless . This is the only memoryless discrete probability distribution.

$$P(X > s+t \mid X > s) = P(X > t) \qquad \text{(Exercise 9)}$$

- # Section 1.5 The Poisson Distribution (continuum version of binomial)

– Suppose λ random noise spikes occur on channel per unit time. Consider in a time period of length T.We divide $[0,\ T]$ into n subintervals of length T/n. On average, the probability of a noise spike in one time subinterval is $p = \lambda T/n$ Then according to the binomial distribution, the probability that a spike occurs in exactly k subinterval is

$$\binom{n}{k} p^k (1-p)^{n-k} = \binom{n}{k} \left(\frac{\lambda T}{n}\right)^k \left(1 - \frac{\lambda T}{n}\right)^{n-k} \to \frac{(\lambda T)^k}{k!} e^{-\lambda T}$$

So the probability that exact k spikes in $[0, T]$ is the Poisson distribution with PDF

$$p(k) = \frac{(\lambda T)^k}{k!} e^{-\lambda T}$$

– $\mu = \sigma = \lambda T$

– Problem A in page 10 of textbook (Run MATLAB, Exercise 10)

- # Section 1.6 Taguchi quality control
  – X is a random variable measure the quality of a product. The quality of loss function (QLF) is defined as

$$L(X) = k(X-\theta)^2$$

  – k is the loss coefficient, θ is the target value. The goal is to make $L(X)$ as small as possible.

  – $E[L(X)] = kE\left[(X-\theta)^2\right]$

$$\begin{aligned} E[L(X)] &= kE\left[((X-\mu)+(\mu-\theta))^2\right] \\ &= kE\left[(X-\mu)^2 + 2(\mu-\theta)(X-\mu) + (\mu-\theta)^2\right] \\ &= kE\left[(X-\mu)^2\right] + 2k(\mu-\theta)E(X-\mu) + k(\mu-\theta)^2 \\ &= k\sigma^2 + k(\mu-\theta)^2 \end{aligned}$$

  – It reaches minimum when $\sigma = 0$ and $\mu = \theta$

  – Two problems in page 12.

- # Homework

(a) Exercise 1 – 10 (as outlined above in this note, not from textbook)

(b) Let $X \to U[0,1]$ (i.e., the CDF of random variable X is the standard uniform distribution), and $Y(\omega) = aX(\omega) + b$.Find a, b such that $Y \to U[-C, C]$ for any $C > 0$.

(c) From textbook: #1.1, #1.2

- Hints for Exercise #5,6,8: Differentiate the identities (twice if necessary)

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}$$

$$\sum_{k=0}^{\infty} x^k = \frac{1}{1-x}$$