Some Probability and Statistics

Probability

- The goal of probability is to quantify our belief about the potential occurrence of an event.
- Ω = sample space = set of all possible outcomes
- event = subset of Ω
- $A \subseteq \Omega$ occurs if we do an experiment and get $x \in A$.

$\Omega = \{\text{rain, no rain}\}$

$A = \{\text{rain}\}$ = event that it rains

$A = \Omega$ = event that it either rains or doesn't.

- We quantify our belief in the occurrence of A via a probability measure.

Definition ③ A probability measure on a countable sample space Ω is a map

(power set of Ω)

$$P : \mathcal{P}(\Omega) \longrightarrow [0,1] \quad ; \quad P : A \longmapsto P(A)$$

such that $P(\Omega) = 1$ and

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i) \quad \text{if} \quad A_i \cap A_j = \emptyset \quad \forall i \neq j.$$

- If Ω uncountable, replace $\mathcal{P}(\Omega)$ with $\mathcal{U} \subseteq \mathcal{P}(\Omega)$ called a σ-algebra. Each $A \in \mathcal{U}$ is called a measurable set.
- $\mathcal{U}$ usually contain sets that play well with Lebesgue integration. This is because we define many of our favorite probability measures via integration of a density function:

Definition ④ A density function is an integrable function $f : \mathbb{R}^m \to [0,\infty)$ with

$$\int_{\mathbb{R}^m} f = 1.$$

Given density f, we define a probability measure P by

$$P(A) := \int_A f \quad , \quad A \in \mathcal{U}.$$

The triple $(\Omega, \mathcal{U}, P)$ is a probability space.

$\mathbb{R}^m \sim \mathcal{B}_m$.

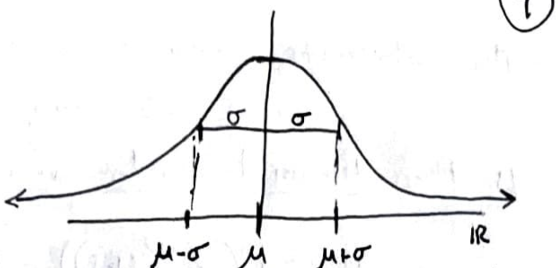
Example (Normal Distribution) $\Omega = \mathbb{R}$

$\mathcal{U} = \mathcal{B} := \{$subsets of $\mathbb{R}$ given by countable union, intersection and relative complement of open sets (intervals)$\}$

- $\mu \in \mathbb{R}$ is the mean
- $\sigma > 0$ is the variance

Define the density

$$\phi_{\mu,\sigma}(x) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}$$



- Denote distribution as $N(\mu,\sigma)$.

- Given a probability space we can also ask if one event is more or less likely given knowledge about another event. This is the notion of dependence / independence...

Defintion ⑤ $A, B \in \mathcal{U}$ such that $P(B) > 0$. The conditional probability of

- A given B is $P(A \mid B)$ where
$$P(A \cap B) = P(A \mid B)\, P(B).$$

- We say A and B are independent if either $P(B) > 0$ and $P(A \mid B) = P(A)$ or $P(B) = 0$.

Exercise: A, B independent $\Leftrightarrow P(A \cap B) = P(A)P(B)$.

Random Variables

- Most experiments we do yield numerical quantitis that we may want to modify. Random variables give us the language for doing that.

Definition ⑥ $(\Omega, \mathcal{U}, P)$ a probability space. A random variable is a

- function $X: \Omega \to \mathbb{R}$ such that for every measurable set $B \in \mathbb{B}$ $X^{-1}(B) \in \mathcal{U}$. The distribution P^X associated to X is
$$P^X(B) := P(X^{-1}(B)), \quad B \in \mathbb{B}.$$

-

Example 1. ① Continnous random variable $\Rightarrow \Omega$ uncountable. For example, $X: \mathbb{R} \to \mathbb{R}$ the identity and $P = \int_B \phi_{\mu,\sigma}$ for $B \in \mathbb{B}$ is called a Gaussian random variable; i.e., $X \sim N(\mu,\sigma)$

② Discrete random variable $\Rightarrow \Omega$ contable. For example:

. A coin has chance $p \in (0,1)$ of landing Heads up (H).

- Flip coin n times get outcome $\underline{\omega} = (\omega_1, \omega_2, \dots, \omega_n) \in \Omega := \{H,T\}^n$.

. Since outcomes ω_i independent
$$P(\underline{\omega}) = P(\omega_1)P(\omega_2)\cdots P(\omega_n) = p^k(1-p)^{n-k}; \quad k = \#\text{ of H's in } \underline{\omega}$$

- The discrete random variable

$$X : \Omega \longrightarrow \mathbb{R} \quad ; \quad \underline{\omega} \longmapsto \Sigma_{i=1}^{n} \chi(\omega_i = H)$$

is the binomial random variable.

$$P_k := P(X^{-1}(\{k\})) = \binom{n}{k} p^k (1-p)^{n-k} \quad , \quad k \in \{0, \dots, n\}$$

is P^X, called the binomial distribution $B(n,p)$.

$\boxed{\text{Definition ⑦}}$ A random vector in $\mathbb{R}^m$ is a vector of random variables $(X_1, \dots, X_m)$ all defined on the same $(\Omega, \mathcal{U}, P)$. The joint distribution of $(X_1, \dots, X_m)$, denoted $P^{(X_1, \dots, X_m)}$, defined like P^X but for $\mathbb{B}_m$.

- Given jointly distributed random variables, we will often be interested in the dependencies amongst the variables encoded by the distribution, generalizing how we can ask about the dependencies of one event upon another.

$\boxed{\text{Definition ⑧}}$ Given random variables (or vectors) $X_1, \dots, X_m$ on the probability space $(\Omega, \mathcal{U}, P)$, we say $X_1, \dots, X_m$ are mutually independent if

$$P(X_1 \in B_1, \dots, X_m \in B_m) = \prod_{i=1}^{m} P(X_i \in B_i) \quad \forall \text{measurable } B_1, \dots, B_m.$$

$\boxed{\text{Lemma ①}}$ Let $X_1, \dots, X_m$ random vectors on $(\Omega, \mathcal{U}, P)$ with densities $f_1, \dots, f_m$.

If $X_1, \dots, X_m$ mutually independent then $(X_1, \dots, X_m)$ has density

$$f(\underline{x}) = \prod_{i=1}^{m} f_i(x_i) . \quad ⊛$$

If $(X_1, \dots, X_m)$ has density f then $X_1, \dots, X_n$ independent if ⊛ holds almost surely.

$\boxed{\text{Example}}$ (Multivariate Normal (Gaussian) Distributions) The random vector $(X_1, \dots, X_m)$ is distributed according to the standard multivariate normal distribution if

① $X_i \sim N(0,1) \quad \forall i \in [m]$, and

② $X_1, \dots, X_m$ mutually independent.

$$\text{Lemma ①} \Rightarrow f_{(X_1, \dots, X_m)}(\underline{x}) = \prod_{i \in [m]} \phi_{0,1}(x_i) = \prod_{i \in [m]} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} x_i^2\right) = \frac{1}{(\sqrt{2\pi})^m} \exp\left(\frac{-1}{2} \underline{x}^T \underline{x}\right)$$

- More generally, take $\underline{\mu} \in \mathbb{R}^m$ and $\Sigma \in PD_m := \{ m \times m$ positive definite matrices$\}$.
- $(X_1, \dots, X_m)$ distributed according to the multivariate normal distribution $N(\underline{\mu}, \Sigma)$ if it has density

$$f_{(X_1,\dots,X_m)}(\underline{x}) = \phi_{\mu,\Sigma}(\underline{x}) := \frac{1}{(2\pi)^{m/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\underline{x}-\underline{\mu})^T \Sigma^{-1} (\underline{x}-\underline{\mu})\right).$$

Expectation, Variance, and Covariance

- We we start working with data (i.e., statistics) we will want numerical features of the data-generating distribution we can try and estimate.
- The first such feature is the "center" or expectation of the distribution...

Definition 9 X a random variable taking values in $\mathcal{X} \subseteq \mathbb{R}$ such that $\int_{\mathcal{X}} |x| f(x) < \infty$. The expected value, expectation or mean of X is

$$E[X] := \int_{\mathbb{R}} x f(x) \qquad \left(E[X] := \sum_{x \in \mathcal{X}} x P(X=x) \text{ if } X \text{ discrete} \right)$$

The vector $X = (X_1, \dots, X_m)$ has expectation $E[X] = (E[X_1], \dots, E[X_m])$.

Lemma 2 Expectation is linear: $X = (X_1, \dots, X_m)$, $a_0 \in \mathbb{R}$, $\underline{a} \in \mathbb{R}^m$ then if $Y = a_0 + a^t X$,

$$E[Y] = a_0 + \underline{a}^t E[X] = a_0 + \sum_{i=1}^m a_i E[X_i].$$

Examples $E[X] = \mu$ if $X \sim N(\mu, \sigma)$. (check).

- We also measure the spread of a distribution around its mean:

Definition 10 The variance of a random variable X is

$$\mathrm{Var}[X] := E[(X - E[X])^2] \longleftrightarrow \text{the expectation of the squared deviation from the mean...}$$

- Lemma 2 $\Rightarrow$ $\mathrm{Var}[X] = E[X^2] - E[X]^2$. (check!)
- By generalizing variance, we get a measure of dependency between variables

Definition 11 The covariance of two random variables X and Y is

$$\mathrm{Cov}[X, Y] := E[(X - E[X])(Y - E[Y])].$$

The covariance matrix of $X = (X_1, \dots, X_m)$ is $\mathrm{cov}[X] := [\mathrm{Cov}[X_i, X_j]]_{i,j=1}^m$

Example $\mathrm{Cov}[X] = \Sigma$ for $X \sim N(\mu, \Sigma)$.

Exercise Show that $\mathrm{cov}[X]$ is positive semidefinite for any random vector X.

Statistics

- While probability gives a language for describing our belief in the occurrence of an event when the probability space is known, statistics focuses on how to infer the potential occurrence of an event when it is unknown...
- In statistics, start with only data and try to infer the probability of an event.
- Two Perspectives:

① Frequentist: There is a "true" unknown probability distribution and we want to use our data to learn this distribution precisely as # samples → ∞.

② Bayesian: Use data to update our prior beliefs about an unknown parameter.

- We will mainly consider the frequentist perspectives..

Statistical Models

Definition (12) A statistical model M is a family of probability distributions.

A parametric statistical model M_Θ is a mapping from a finite dimensional parameter space $\Theta \subseteq \mathbb{R}^d$ to a space of distributions or densities:

$$P_\bullet : \Theta \to M_\Theta \; ; \; \theta \longmapsto P_\theta .$$

$M_\Theta := P_\bullet(\Theta) = \{P_\theta : \theta \in \Theta\}$. M_Θ is identifiable if $P_\bullet$ is injective.

Examples ① The Binomial Random Variable Model: $\Theta := [0,1] \subseteq \mathbb{R}^1$.

Fix $r \in \mathbb{Z}_{\geq 0}$. Then

$$M_\Theta := \left\{ \left(\binom{r}{i} \theta^i (1-\theta)^{r-i} \;\middle|\; i \in \{0,\dots,r\} \right) : \theta \in [0,1] \right\}.$$

② Multivariate Gaussian (Normal) Model: X an m-dim'l random vector.

$\Theta := \mathbb{R}^m \times PD_m$. M_Θ is the image of

$$P_\bullet : \theta = (\mu, \Sigma) \longmapsto \phi_{\mu,\Sigma}(x).$$

Definition (13) A statistical model M is implicit if it is defined by a set of constraints.

Example conditional independence models. (Markov chain)

Note A model can be both implicit and parametric..

- When trying to estimate how our data-generating distribution fits a model's parameters or defining constraints we will rely on features of our data that we call statistics...

Definition (14) A statistic is a function from the state space of a random variable (i.e., its image) to another set $T : \mathfrak{X} \to \mathbb{T}$. For a parametric model M_Θ, T is sufficient if it densityies $f_\Theta(x)$ satisfies

$$f_\Theta(x) = h(x) g(T(x), \Theta) \quad \text{for some } g, h.$$

Example X_p is the binomial random variable where $X = 0$ if an event happens and $X = 1$ if it doesn't. Do n i.i.d experiments $X^{(1)}, \dots, X^{(n)}$. Then if $\Theta \in (0,1)$ is $P(X = 0)$,

$$P(X^{(1)} = x_1, \dots, X^{(n)} = x_n) = \prod_{i \in [n]} P(X_i = x_i) = \Theta^{\sum x_i} (1 - \Theta)^{n - \sum x_i}.$$

So $h(x) \equiv 1$, $g(t, \Theta) = \Theta^t (1 - \Theta)^{n-t}$ and so $T(X) = \sum_{i=1}^n X^{(i)}$ is sufficient.

- Here, our data was the outcome of an event and the # of samples was n. If the event was "a flipped coin landed heads up" then each sample would have been drawn from the same distribution and independent.

★ Our data will usually be i.i.d = independent and identically distributed.

- To estimate the probability distribution assuming it comes from our parametric model, we will often use sufficient statistics to estimate the model parameters.

Definition (15) A parameter of a statistical model M_Θ is a function $s : \Theta \to \mathbb{R}$. An estimator of s is a function $\hat{s} : \mathbb{D} \to \mathbb{R}$, where $\mathbb{D}$ is the data. $\hat{s}$ is consistent if $\hat{s} \to s$ as $|\mathbb{D}| \to \infty$.

Example The maximum likelihood estimator (MLE)

$$\hat{\Theta} = \underset{\Theta \in \Theta}{\arg\max}\, L(\Theta \mid \mathbb{D}) = \underset{\Theta \in \Theta}{\arg\max}\, P(\mathbb{D} \mid \Theta).$$

(see last lecture!)

- Parameter estimation only gives an approximation of the model parameters assuming the data-generating distribution is actually in the model. This may be a bad assumption.

Hypothesis Testing

- Hypothesis tests are the statistical framework for rejecting an assumption (such as "the data-generating distribution is in this model") base on data.

- H_o = the null hypothesis

[Q] Given the data can we reject the null hypothesis H_o?

[Example] • X, Y are two random variables.

- $H_o : (X = Y)$; i.e. X and Y are equidistributed (same distribution).
- $H_a : (X \neq Y)$ is the alternative hypothesis $(\neg H_0)$.

- Can reject H_o if $E[X] \neq E[Y]$.
- Collect samples $X^{(1)}, \dots, X^{(n_1)}$ from X and $Y^{(1)}, \dots, Y^{(n_2)}$ from Y.
- Compute the sample means (estimators of the means):

$$\hat{\mu}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} X^{(i)} \quad \text{and} \quad \hat{\mu}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} Y^{(i)}.$$

- Permutation test (exact test):
  - set $\bar{d} = |\hat{\mu}_1 - \hat{\mu}_2|$. and $Z := \{Z_1, \dots, Z_{n_1+n_2}\} = \{X^{(1)}, \dots, X^{(n_1)}, Y^{(1)}, \dots, Y^{(n_2)}\}$.
  - for each partition $A \cup B$ of Z where $|A| = n_1$ compute

$$d_{A\cup B} := \left| \frac{1}{n_1} \sum_{a \in A} a - \frac{1}{n_2} \sum_{b \in B} b \right|.$$

  - If $p := \dfrac{|\{A \cup B \mid d_{A\cup B} \geq \bar{d}\}|}{\binom{n_1+n_2}{n_1}} < \alpha = 0.05$ then reject H_o.

- We can do this because if few $d_{A\cup B}$ are larger than $\bar{d}$, it suggests that $\hat{\mu}_1$ and $\hat{\mu}_2$ are further away that would be possible due to sampling error...