I. THEME #1. BASIC CONCEPTS FROM STATISTICS

A. Lecture #1. Random Variables: Characterization & Description.

1. *Probability of an event*

Discrete vs Continuous events. State/phase/sample space (for events), Σ.

Example of discrete events: two states, $\Sigma = \{0, 1\}$ -also called Bernoulli random variable (derived from a "process", i.e. dynamics - to be discussed later in the course a lot). Probability of a state, σ,

$$\forall \sigma \in \Sigma : \quad \text{Prob}(\sigma) = P(\sigma) \tag{I.1}$$
$$0 \leq P(\sigma) \leq 1 \tag{I.2}$$
$$\sum_{\sigma \in \Sigma} P(\sigma) = 1 \tag{I.3}$$

For Bernoulli process, $P(1) = \beta$, $P(0) = 1 - \beta$.

Question: Can you give an example of the Bernoulli distribution from life/science?

Answer: A biased coin.

Another important discrete event distribution is the Poisson. An event can occur $k = 0, 1, 2, \cdots$ times in an interval. The average number of events in an interval is λ - called event rate. The probability of observing k events within the interval is

$$\forall k \in \mathbb{Z}^* = \{0\} \bigcup \mathbb{Z} : \quad P(k) = \frac{\lambda^k e^{-\lambda}}{k!}. \tag{I.4}$$

(Check that the probability is properly normalized, in the sense of Eq. (I.3).) The distribution is also called exponential distribution (for obvious reason - look at the expression).

Questions: Are Bernoulli and Poisson distributions related? Can you "design" Poisson from Bernoulli? Can you give an example of the Poisson process from life/science?

Answer: Consider repeating Bernoulli, each time independently, thus drawing a Bernoulli process. You get sequence of zeros and ones. Then check only for ones and record times/slots associated with arrivals of ones. Study probability distribution of t arrivals in n step, and then analyze $n \to \infty$, to get the Poisson distribution. We will discuss it in details in Lect.# 5. Example of Poisson — arrival of customers at the shop.

The domain can be continuous, bounded or unbounded. Example of a distribution which is bounded - is the uniform distribution from the $[0, 1]$ interval:

$$\forall x \in [0, 1] : \quad p(x) = 1, \tag{I.5}$$
$$\int_0^1 dx p(x) = 1, \tag{I.6}$$

where $p(x)$ is the probability density. Gaussian distribution is the most important (also most frequently used) continuous distribution:

$$\forall x \in \mathbb{Z} : \quad p(x|\sigma, \mu) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{I.7}$$

$p_{\sigma,\mu}(x)$ another possible notation. It is also called "normal distribution" - where "normality" refers to significance of the distribution for the central limit theorem (law of large numbers), which we will be discussing shortly. The distribution is parameterized by μ and λ - what is the significance of the two parameters? (mean and variance) Standard notation in math for the Gaussian/normal distribution is $N(\mu, \sigma^2)$.

There are many more 'standard' distributions but Bernoulli, Poisson and Gaussian are the 'golden' three. One can generate practically any other distribution from the 'golden set' (possibly extended by the uniform distribution).

Discussion of the notations, e.g. $P_X(x)$, $\mathbb{E}[\cdots] = \langle \cdots \rangle$, $s \sim P(x)$.

2. *Sampling. Histograms.*

Random process generation. Random process is generated/sampled. Any computational package/software contains a random number generator (in fact a number of these). Designing a good random generation is important. In this course, however, we will mainly be using the random number generators (in fact pseudo-random generators) already created by others.

To illustrate let us switch to IJulia/Julia notebook of the first two lectures.

Histogram. To show distributions graphically, you may also "bin" it in the domain - thus generating the histogram, which is a convenient way of showing $p(\sigma)$ (plot with Julia: breaking $[0,1]$ interval in $N > 1$ bins). This is an opportunity for us to introduce statistical computational package to be used in the lectures and recitations IJulia/Julia. The package is available both online (web) and offline (your computer). We will mainly be using the web-version, see `https://juliabox.com/`.

3. *Moments. Generating Function.*

Expectations.

$$\mathbb{E}\left[A(\sigma)\right]_p = \langle A(\sigma)\rangle_p = \sum_{\sigma\in\Sigma} A(\sigma)p(\sigma).$$

Examples: mean,

$$\mathbb{E}\left[\sigma\right],$$

variance,

$$\text{Var}[\sigma] = \mathbb{E}\left[(\sigma - \mathbb{E}[\sigma])^2\right].$$

We have already discussed these for the Gaussian process. What are mean and variance for Bernoulli process?

Moments.

$$k = 0,\cdots, \quad \mathbb{E}\left[\sigma^k\right]_p = \langle\sigma^k\rangle_p = \sum_{\sigma\in\Sigma}\sigma^k p(\sigma) = m_k(\Sigma)$$

Moment Generating Function.

$$M_X(t) = \mathbb{E}\left[\exp(tx)\right], \quad t\in\mathbb{R}$$

One can also view it as a Laplace transform of the probability density function,

$$M_X(t) = \int dx p(x)\exp(tx).$$

Examples of the moment generating functions for aforementioned (and other) distributions — derive it yourself. More on the generating functions at the recitations.

Characteristic function is a related object, defined as a Fourier transform of the probability density:

$$\mathbb{E}\left[\exp(itx)\right] = \int dx p(x)\exp(itx)$$

where $i^2 = -1$.

4. *Probabilistic Inequalities.*

Here are some useful probabilistic inequalities.

- (Markov Inequality)

$$P(x \geq c) \leq \frac{\mathbb{E}[x]}{c} \tag{I.8}$$

- (Chebyshev's inequality)

$$P(|x-\mu| \geq k) \leq \frac{\sigma^2}{k^2} \tag{I.9}$$

- (Chernoff bound)

$$P(x \geq a) = P(e^{tx} \geq e^{ta}) \leq \frac{\mathbb{E}[e^{tx}]}{e^{ta}} \tag{I.10}$$

where μ and σ are mean and variance of x.

We will get back to discussion of these and some additional inequalities in the third lecture.

Exercise: Play in IJulia checking the three inequalities for the distributions mentioned through out the lecture.

Exercise: Prove the Markov inequality. Chebyshev inequality will follow from the Markov. Prove it too. Chernoff is trickier, can you prove it too? [See `http://jeremykun.com/2013/04/15/probabilistic-bounds-a-primer/` to get additional info and also check your answers.]

Exercise: Provide examples of the distributions for which the tree inequalities are saturated (becomes equalities)?

5. *Recitation. Random Variables. Moments. Characteristic Function.*

B. Lecture #2. Random Variables: from one to many.

1. *Law of Large Numbers*

Take n samples $x_1, \cdots, x_n$ generated i.i.d. from a distribution with mean μ and variance, $\sigma > 0$, and compute $y_n = \sum_{i=1}^n x_i/n$. What is Prob(y_n)? $\sqrt{n}(y_n - \mu)$, converges in distribution to Gaussian with mean, μ, and variance, σ:

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n x_i - \mu\right) \xrightarrow{d} N(0,\ \sigma^2). \tag{I.11}$$

This is so-called Weak Version of the Central Limit Theorem.

Let us prove the weak-CLT (I.11) in a simple case $\mu = 0$, $\sigma = 1$. Obviously, $m_1(Y_n\sqrt{n}) = 0$. Compute

$$m_2(Y_n\sqrt{n}) = \mathbb{E}\left[\left(\frac{x_1 + \cdots + x_n}{\sqrt{n}}\right)^2\right] = \frac{\sum_i \mathbb{E}\left[x_i^2\right]}{n} + \frac{\sum_{i\neq j}\mathbb{E}\left[x_i x_j\right]}{n} = 1.$$

Now the third moment:

$$m_3(Y_n\sqrt{n}) = \mathbb{E}\left[\left(\frac{x_1 + \cdots + x_n}{\sqrt{n}}\right)^3\right] = \frac{\sum_i \mathbb{E}\left[x_i^3\right]}{n^{3/2}} \to 0,$$

at $n \to \infty$, assuming $\mathbb{E}\left[x_i^3\right] = O(1)$. Can you guess what will happen with the fourth moment? $m_4(Y_n\sqrt{n}) = 3 = 3m_2(Y_n)$. This is related to the so-called Wick's theorem (used in theoretical physics a lot). And how about higher odd/even moments?

Exercise: Check IJulia notebok for the lecture and experiment with the law of large numbers for different distributions.

The theorem holds for independent but not identically distributed variables, usually denoted as i.i.d. too.

If one is interested in not only the asymptotic itself, $n \to \infty$, but also in how the asymptotic is approached, the so-called strong version of CLT (can also be found under the name of Cramér theorem) states

$$\forall z > \mu: \quad \lim_{n\to\infty} \frac{1}{n}\log \text{Prob}\left(y_n \geq z\right) = -\Phi^*(z) \tag{I.12}$$

$$\Phi^*(z) \doteq \sup_{\lambda\in\mathbb{R}}\left(\lambda z - \Phi(z)\right) \tag{I.13}$$

$$\Phi(\lambda) \doteq \log\left(\mathbb{E}\exp\left(\lambda z\right)\right) \tag{I.14}$$

$\Phi^*(z)$ is a convex function (also called Cramér function). This was a formal (mathematical) statement. A less formal (physical) version of Eq. (I.12) is

$$n \to \infty: \quad \text{Prob}\left(y_n\right) \propto \exp\left(-n\Phi^*(x)\right) \tag{I.15}$$

One of our journal club projects is on this subject.

Note, that the weak version of the CLT (I.11) is equivalent to approximating the Cramer function (asymptotically exact) by a Gaussian around its minimum.

Exercise (bonus): Prove the strong-CLT (I.12,I.13). [Hint: use saddle point/stationary point method to evaluate the integrals.]

Exercise: Give an example of an expectation for which not only vicinity of the minimum but also other details of $\Phi^*(x)$ are significant at $n \to \infty$? More specifically give an example of the object which behavior is controlled solely by left/right tail of $\Phi^*(x)$? $\Phi^*(0)$ and its vicinity?

Example of Bernoulli process – a (generally unfair) coin toss

$$x = \begin{cases} 0 & \text{with probability } 1-p \\ 1 & \text{with probability } p \end{cases} \tag{I.16}$$

Then

$$\Phi(\lambda) = \log(pe^\lambda + 1 - p) \tag{I.17}$$

$$0 < x < 1: \quad \Phi^*(z) = z\log\frac{z}{p} + (1-z)\log\frac{1-z}{1-p} \tag{I.18}$$

Eqs. (I.17,I.18) are noticeable for two reasons. First of all, they lead (after some algebraic manipulations) to the famous Stirling formula for the asymptotic of a factorial

$$n! = \sqrt{2\pi n}n^n e^{-n}\left(1 + O(1/n)\right).$$

Do you see how? Second, the $z\log z$ structure is an "entropy" which will appear a number of times in the course - stay tuned.

2. *Multivariate Distribution. Marginalization. Conditional Probability.*

Consider an n-component vector build of components each taking a value from a set, Σ, $\sigma = (\sigma_i \in \Sigma | i = 1, \cdots, n)$. Σ may be discrete, e.g. $\Sigma = \{0, 1\}$, or continuous, e.g. $\Sigma = \mathbb{R}$. Assume that any state, σ, occur with the probability, $P(\sigma)$, where $\sum_\sigma P(\sigma) = 1$.

Consider a simple example of bi-variate distribution.

$$\sigma = (\sigma_i = \pm 1 | i = 1, \cdots, n) : P(\sigma) = Z^{-1} \prod_{i=1}^{n-1} \exp(J\sigma_i\sigma_{i+1}) \tag{I.19}$$

$$Z = \sum_\sigma \prod_{i=1}^{n-1} \exp(J\sigma_i\sigma_{i+1}) \tag{I.20}$$

where Z is the normalization constant, also called the partition function. The partition function s introduced to guarantee that the sum over all the states is unity. For $n = 2$ one gets

$$P(\sigma) = P(\sigma_1, \sigma_2) = \frac{\exp(J\sigma_1\sigma_2)}{4\cosh(J)}. \tag{I.21}$$

$P(\sigma)$ is also called a joint probability distribution function of the σ vector components, $\sigma_1, \cdots, \sigma_n$. It is also useful to consider conditional distribution, say for the example above with $n = 2$,

$$P(\sigma_1|\sigma_2) = \frac{P(\sigma_1, \sigma_2)}{\sum_{\sigma_1} P(\sigma_1, \sigma_2)} = \frac{\exp(J\sigma_1\sigma_2)}{2\cosh(J\sigma_2)} \tag{I.22}$$

is the probability to observe σ_1 under condition that σ_2 is known. Notice that, $\sum_{\sigma_1} P(\sigma_1|\sigma_2) = 1,\ \forall \sigma_2$.

Let us now marginalize the multivariate (joint) distribution over a subset of variables. For example,

$$P(\sigma_1) = \sum_{\sigma \setminus \sigma_1} P(\sigma) = \sum_{\sigma_2, \cdots, \sigma_n} P(\sigma_1, \cdots, \sigma_n). \tag{I.23}$$

We will repeat exercises (joint, conditional, marginal) with multivariate Gaussian distribution at the recitations. The Gaussian distributions are remarkably unique/universal. This is because application of any of the aforementioned operations, joint-to-conditional and joint-to-marginal, will also be Gaussian. Recall that the Gaussian emerges naturally in the result of the CLT.

3. *Bayes Theorem*

We already saw how to get conditional probability distribution and marginal probability distribution from the joint probability distribution

$$P(x|y) = \frac{P(x, y)}{P(y)}, \quad P(y|x) = \frac{P(x, y)}{P(x)}. \tag{I.24}$$

Combining the two formulas to exclude the joint probability distribution we arrive at the famous Bayes formula

$$P(x|y)P(y) = P(y|x)P(x). \tag{I.25}$$

Here, in Eqs. (I.24,I.26) both x and y may be multivariate.

Rewriting Eq. (I.26) as

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)}, \tag{I.26}$$

one often refers (in the field of the so-called Bayesian inference/reconstruction) to $P(x)$ as the "prior" probability distribution which measures the degree of the initial "belief" in x. Then, $P(x|y)$, called the "posterior", measured the degree of the (statistical) dependence of x on y, and the quotient $\frac{P(y|x)}{P(y)}$ represents the "support/knowledge" y provides about x.

A good illustration of the notion of the conditional probability can be found at `http://setosa.io/ev/conditional-probability/`

Let us conclude the lecture playing a bit with a made up case of the bi-variate binary distribution containing the total of $2^2 = 4$ states.

4. *Recitation. Properties of Gaussian Distributions. Laws of Large Numbers.*