INSE6320		Jia Yuan Yu
	1: Probabilities	
Concordia		January 14, 2020

Risk comes from uncertainty. Uncertainty is quantified using probability theory.

1 Review of probabilities

Probability theory and statistics is the most widely used set of mathematical tools for modeling uncertainty.

- Probability space,
  - Set of outcomes Ω,
  - Set of events $\mathcal{F}$ containing subsets of Ω,
  - A function (probability measure) $P : \mathcal{F} \to \mathbb{R}$.
- Random variables, distribution functions,
  - Real-value random variable is a (measurable) function $X : \Omega \to \mathbb{R}$.
  - Probability distribution function or cumulative distribution function: $F(x) = P(X \leq x)$ for all x.
  - Probability density function, if it exists: $F(x) = \int_{-\infty}^{x} f(z)dz$.
- Examples (Bernoulli, Normal, Uniform, Exponential etc.),
  - Bernoulli with parameter p: $P(X = 1) = p$, $P(X = 0) = 1 - p$.
  - Exponential with rate λ: $F(x) = (1 - e^{-x})1_{[x\geq 0]}$ for all $x \in (-\infty, \infty)$.
  - Normal $N(0, 1)$: $f(x) = (2\pi)^{-1/2}e^{-x^2/2}$ for all $x \in (-\infty, \infty)$.
- Expectation,
  - Bernoulli: $\mathbb{E}X = p * 1 + (1 - p) * 0$

Statistics is the study of data using probability theory: we have access to random variables $X_1, X_2, \ldots$, but we don't know their distributions. We want to use $X_1, X_2, \ldots$ to infer their distributions.

- Classical: Let P denote the joint distribution of $X_1, X_2, \ldots, X_n$. We assume that P belongs to a known set $\{P_\theta : \theta \in \Theta\}$. The goal is to find the true value θ^* or a subset containing it.
- Bayesian: Assume that θ^* is a random variable from a known distribution.

Example 1.1. Consider $X_1, X_2, \ldots, X_n$ corresponding to the measured lifespans of n light bulbs (no decisions involved). Assume that they are independent and identically distributed according to a normal distribution P. Estimate the mean of P.

1.1 Probabilistic guarantees

Suppose that for a given decision a, we observe the following sequence of performance measurements: $X_1^a, X_2^a, \ldots, X_n^a$. Suppose that these measurements are i.i.d. with distribution F. Given δ, we would like to find an λ give guarantees of the form:

$$\mathbb{P}(X_{n+1}^a > \lambda) \geq 1 - \delta. \tag{1}$$

Observe that the above guarantee is equivalent to $1 - F(\lambda) \geq 1 - \delta$ or $\lambda \leq F^{-1}(\delta)$.

One approach is to find an estimate $\hat{F}_n$ of the distribution F using the data $X_1^a, X_2^a, \ldots, X_n^a$. If $\hat{F}$ is a very good estimate of F, then we can say that

$$\mathbb{P}(X_{n+1}^a > \hat{F}_n^{-1}(\delta)) \geq 1 - \delta.$$

Remark 1 (Relation to Newsboy problem). Recall that F^{-1} appears also in the solution to the Newsboy problem. There, the distribution F is assumed to be known, whereas in this course, we need to estimate F from data.

1.2 Estimating distributions

Figure 1: From `http://candywow.weebly.com/`

Estimation with candies. Suppose that your supply chain produces candies and that the color of each candy corresponds to a quality value:

- Green = 1
- Yellow = 2
- Orange = 3
- Red = 3
- Purple = 5

Students observe $X_1, X_2, \ldots$ The (unknown) true distribution F is

- $\mathbb{P}(X_i \le 1) = 21/95$
- $\mathbb{P}(X_i \le 2) = (21 + 15)/95$
- $\mathbb{P}(X_i \le 3) = (21 + 15 + 49)/95$
- $\mathbb{P}(X_i \le 5) = (21 + 15 + 49 + 10)/95$

We can estimate the distribution F for a sequence of i.i.d. random variables as follows. Let $X_1, X_2, \ldots, X_n$ denote the samples. Construct the following empirical distribution function, for every x:

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{[X_i \le x]}.$$

How well does $\hat{F}_n$ estimate F? Good news: exceptionally well!

Theorem 1.1 (Dvoretzky–Kiefer–Wolfowitz Inequality[1])**.** *For every $\varepsilon > 0$, we have*

$$\mathbb{P}\left(\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| > \varepsilon \right) \le 2e^{-2n\varepsilon^2}.$$

Remark 2. The ε above is analogous to the notion of "margin of error," whereas $1 - \delta$ is analogous to "confidence."

Homework: Combine DKW Inequality with (1).

1.3 Estimating normal distributions

Normal random variables are entirely characterized by the mean μ and variance σ^2. The following sample-mean is an unbiased mean estimator:

$$\hat{\mu}_n = \frac{1}{n} \sum_{k=1}^{n} X_k.$$

The following is an unbiased variance estimator:

$$\hat{\sigma}_n^2 = \frac{1}{n-1} \sum_{k=1}^{n} (X_k - \hat{\mu}_n)^2.$$

How well do $\hat{\mu}_n$ and $\hat{\sigma}_n^2$ estimate μ and σ^2? Very well, thank you!

[1]Dvoretzky, A.; Kiefer, J.; Wolfowitz, J. (1956), "Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator", Annals of Mathematical Statistics 27 (3): 642–669.

Theorem 1.2 (Hoeffding)**.** *Let $X_1, X_2, \dots$ be i.i.d. random variables that take values in the interval $[a, b]$, and have mean μ. Let*

$$\hat{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Then, for every n and $\varepsilon > 0$:

$$\mathbb{P}\left(\left|\hat{X}_n - \mu\right| \geq \varepsilon\right) \leq 2 \exp\left(-\frac{2n\varepsilon^2}{(b-a)^2}\right).$$