POBABILITY AND STATISTICS, Lesson 8.

- **Descriptive statistics**: $(\Omega, \mathcal{A}, \mathcal{P})$ is *statistical space*, where $(\Omega, \mathcal{A}, \mathbb{P})$ is probability space. for all $\mathbb{P} \in \mathcal{P}$. *Parametric* case: $\mathcal{P} = \{\mathbb{P}_\theta \,|\, \theta \in \Theta\}$, where $\Theta \subset \mathbb{R}^k$ is the *parameter space*.

  Statistical sample: $X_1, X_2, \ldots, X_n$ i.i.d. *Sample space* ($\mathcal{X}$): set of all possible *realizations* $\mathbf{x} = (x_1, \ldots, x_n)$ of $\mathbf{X} = (X_1, \ldots, X_n)$. *Statistic*: $T = T(\mathbf{X}) = T(X_1, \ldots, X_n)$ measurable function of the sample elements. Basic statistics:

  - *Sample mean*: $\bar{X} = \dfrac{1}{n}\sum_{i=1}^n X_i$. (Sometimes $\bar{X}_n$, $\bar{x}$, $\bar{x}_n$.)
  - *Steiner's Theorem*: $\dfrac{1}{n}\sum_{i=1}^n (x_i - c)^2 = \dfrac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2 + (\bar{x} - c)^2$.
  - *Empirical variance*: $S^2 = \dfrac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \bar{X}^2 = \overline{X^2} - \bar{X}^2$.
  - *Corrected empirical variance*: $S^{*2} = \dfrac{n}{n-1}S^2 = \dfrac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$.
  - *Standard Error of Mean*: $\bar{X}\sqrt{n}/S^*$.
  - *k-th empirical moment*: $M_k = \dfrac{1}{n}\sum_{i=1}^n X_i^k$. *Centered* version: $M_k^c = \dfrac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^k$. ($S^2 = M_2^c = M_2 - M_1^2$.)
  - *Skewness*: $M_3^c/(M_2^c)^{3/2}$. *Curtosis*: $M_4^c/(M_2^c)^2 - 3$.
  - *Empirical covariance*: $C = \dfrac{1}{n}\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) = \dfrac{1}{n}\sum_{i=1}^n X_i Y_i - \bar{X}\bar{Y}$.
  - *Empirical correlation coefficient*: $R = \frac{C}{S_X S_Y} = \frac{\sum_{i=1}^n X_i Y_i - n\bar{X}\bar{Y}}{\sqrt{\left(\sum_{i=1}^n X_i^2 - n\bar{X}^2\right)\left(\sum_{i=1}^n Y_i^2 - n\bar{Y}^2\right)}}$.

- **Order statistics**: $X_1^* \le X_2^* \le \cdots \le X_n^*$ (neither independent, nor identically distributed).

  - *Sample range*: $X_n^* - X_1^*$.
  - *Empirical median*: X_{k+1}^* (if $n = 2k+1$), and $(X_k^* + X_{k+1}^*)/2$ (if $n = 2k$).
  - *Proposition* (Steiner in L_1-norm): $\min_{c\in\mathbb{R}} \dfrac{1}{n}\sum_{i=1}^n |x_i - c| = \dfrac{1}{n}\sum_{i=1}^n |x_i - m|$.
  - *Empirical c.d.f.*: $F_n^*(x) := \frac{\sum_{i=1}^n I(X_i < x)}{n}$ (stochastic process, x is the time).
  - **Glivenko–Cantelli Theorem**: $\sup_{x\in\mathbb{R}} |F_n^*(x) - F(x)| \to 0$, almost surely ($n \to \infty$).
  - **Kolmogorov–Smirnov Theorems** (one-sample case). Let F be a continuous c.d.f., and $X_1, \ldots, X_n$ be i.i.d. sample from the F-distribution with empirical c.d.f. F_n^*. Then

    $$\lim_{n\to\infty} \mathbb{P}\left(\sqrt{n}\sup_{x\in\mathbb{R}}(F_n^*(x) - F(x)) < z\right) = S(z), \quad \lim_{n\to\infty} \mathbb{P}\left(\sqrt{n}\sup_{x\in\mathbb{R}}|F_n^*(x) - F(x)| < z\right) = K(z)$$

    $\forall z \in \mathbb{R}$, where $S(z) = 1 - e^{-2z^2}$, $\quad K(z) = \sum_{i=-\infty}^{\infty} (-1)^i e^{-2i^2 z^2}$, if $z > 0$, and they are 0 otherwise.
  - **Kolmogorov–Smirnov Theorems** (two-sample case). Let F and G be continuous c.d.f.'s. $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ are independent samples from the F- and G-distributions with empirical c.d.f.'s F_n^* and G_m^*, respectively. If $F = G$, then

    $$\lim_{n,m\to\infty} \mathbb{P}\left(\sqrt{\frac{nm}{n+m}}\sup_{x\in\mathbb{R}}(F_n^*(x) - G_m^*(x)) < z\right) = S(z), \qquad \forall z \in \mathbb{R},$$

    $$\lim_{n,m\to\infty} \mathbb{P}\left(\sqrt{\frac{nm}{n+m}}\sup_{x\in\mathbb{R}}|F_n^*(x) - G_m^*(x)| < z\right) = K(z), \qquad \forall z \in \mathbb{R}.$$