IC105: Probability and Statistics **2021-22-M**

Lecture 11: Some Descriptive Measures (Additional Topics*)

Instructor: Dr. Kuldeep Kumar Kataria *Scribe:*

11.1. Summary of Probabilty Distributions

Let X be a r.v. defined on a probability space $(\Omega, \mathcal{F}, P)$ associated with a random experiment $\mathscr{E}$. Let $F_X(\cdot)$ be its distribution function and $f_X(\cdot)$ be its p.m.f. / p.d.f.

The probabilty distribution of X (*i.e.*, p.m.f. / p.d.f.) describes the manner in which the r.v. X takes values in various sets. It may be desirable to have a set of numerical measures that provide a summary of the prominent features of the probability distribution of X. We call these measures as descriptive measures. Four prominently used descriptive measures are:

(1) Measures of Central Tendency or Location (also called Averages):

This gives us the idea about central value of the probability distribution around which the values of r.v. X are clustered. Commonly used measures of central tendency are:

(a) Mean:

$\mu = \mu_1' = E(X) = \int_{-\infty}^{\infty} x f_X(x) \mathrm{d}x$ or $\sum_{x \in S_X} x f_X(x) \rightarrow$ may or may not exist.

Whenever it exists it gives us the idea about average observed value of X when $\mathscr{E}$ is repeated a large number of times. Note that if distribution of X is symmetric about μ (*i.e.*, $X - \mu \overset{d}{=} \mu - X$), then $E(X) = \mu$, provided it exists.

Mean seems to be the best suited measure of central tendency for symmetric distribution. Because of its simplicity mean is the most commonly used average. However mean may be affected by a few extreme values and also it may not be defined.

(b) Median:

Before defining the median we first inroduce the concept of quantile function or quantile.

The quantile function of r.v. X is a function $Q_X : (0,1) \rightarrow \mathbb{R}$ defined by

$$Q_X(p) = \inf\{x \in \mathbb{R} : F_X(x) \geq p\}, \;\; p \in (0,1).$$

For a fixed $p \in (0,1)$ the quantity $\xi_p = Q_X(p)$ is called the quantile of order p. Note that

$$F_X(\xi_p -) \leq p \leq F_X(\xi_p), \;\; \text{(Exercise)}$$

and $F_X(\xi_p) = p$ provided F_X is continuous at ξ_p. Also note that:

· $Q_X(F_X(x)) \leq x$, provided $0 < F_X(x) < 1$;

· $F_X(Q_X(p)) \geq p, \forall\, 0 < p < 1$;

· F_X is continuous $\implies F_X(Q_X(p)) = p$;

· $Q_X(p) \leq x \iff F_X(x) \geq p$;

· $Q_X(p) = F_X^{-1}(p)$, provided $F_X^{-1}(p)$ exists;

· $Q_X(p_1) \leq Q_X(p_2), \forall\, 0 < p_1 < p_2 < 1$.

The quantile of order 0.5 is called the median of (distribution) of X. If m_e is the median of X, then

$$F_X(m_e-) \leq \frac{1}{2} \leq F_X(m_e).$$

If the random experiment $\mathscr{E}$ is repeated a large number of times about half of the times observed value of X is expected to be less than m_e and about half of the times it is expected to be grearter than m_e.

Suppose that the distribution of X is symmetric about μ. Then

$$\begin{aligned}
&X - \mu \overset{d}{=} \mu - X\\
\implies &P(X - \mu \leq 0) = P(\mu - X \leq 0)\\
\implies &F_X(\mu) = 1 - F_X(\mu-)\\
\implies &F_X(\mu-) \leq \frac{1}{2} \leq F_X(\mu) \implies \mu = E(X) = m_e, \text{ provided } F_X \text{ is continuous at } \mu.
\end{aligned}$$

Merits of Median as a Measure of Central Tendency:

· Unlike mean it is always defined;

· Median is not affected by a few extreme values of X as it takes into account only the probabilities with which different values occur and not their numerical values.

As a measure of central tendency the median is preferred over the mean if the distribution is asymmetric and a few extreme observations occur with positive probabilities.

Demerits of Median as a Measure of Central Tendency:

· Does not at all take into account the numerical values assumed by X;

· For many probability distributions it is not easy to evaluate.

(c) Mode:

Roughly speaking mode m_0 of a probability distribution is the value that occurs with highest probability and is defined by

$$f_X(m_0) = \sup\{f_X(x) : x \in S_X\}.$$

If the random experiment $\mathscr{E}$ is repeated a large number of times then either mode m_0 or a value in the neighborhood of m_0 is observed with maximum frequency.

Note that mode of a distribution may not be unique. A distribution having single / double / triple / multiple mode(s) is called a unimodal / bimodal / trimodal / multimodal distribution.

Merits of a Mode as a Measure of Central Tendency:

It is easy to understand and easy to calculate. Normally, it can be found by just inspections.

Demerits of Mode as a Measure of Central Tendency:

· A probability distribution may have more than one mode which may be far apart.

As a measure of central tendency, mode is less preferred than mean and median. Clearly for symmetric unimodal distributions mean=median=mode.

(2) Measures of Dispersion:

Apart from measures of central tendency other measures are often required to describe a probability distribution. Measures of dispersion give the idea about the scatter (cluster / dispersion) of probability mass of the distribution about a measure of a central tendency. Some of the measures of dispersion are listed below.

(a) Range:

Let $S_X = [a, b]$. Then range of distribution of X is defined by $R = b - a$. It does not take into account how the probability mass is distributed over $[a, b]$. For this reason it is not a preferred measure of dispersion.

(b) Mean Deviation:

Let A be a suitable measure of central tendency. Define

· $MD(A) = E(|X - A|) \rightarrow$ called the mean deviation of X about A (provided it exists);

· $MD(\mu) = E(|X - \mu|) \rightarrow$ mean deviation about mean $\mu = E(X)$;

· $MD(m_e) = E(|X - m_e|) \rightarrow$ mean deviation about median.

It can be show that $MD(m_e) \leq MD(A), \forall\ A \in \mathbb{R}$. For this reason $MD(m_e)$ seems to be more applicable than $MD(A)$ for any $A \in \mathbb{R}$.

· $MD(A)$ is generally difficult to compute for many distributions;

· $MD(A)$ is sensitive to extreme observations;

· $MD(A)$ may not exist for many distributions.

(c) Standard Deviation (SD):

The standard deviation of distribution of X is defined by $\sigma = \sqrt{\mathrm{Var}(X)} = \sqrt{E(X - \mu)^2}$, where $\mu \in \mathbb{R}$. Clearly $\sigma \leq \sqrt{E(X - A)^2}, \forall A \in \mathbb{R}$. It has same unit as that of X.

Standard deviation σ gives us the idea of average spread of values of X around the mean μ.

· σ is simple to compute for most distributions (unlike $MD(A)$, $A \in \mathbb{R}$);

· SD is most widely used measure of dispersion (especially for nearly symmetric distributions);

· For some distributions SD does not exist;

· SD is sensitive to extreme observations.

(d) Quartile Deviation:

Let $q_1 = \xi_{0.25} =$ quantile of order 0.25 (lower quantile of X),

$q_2 = m_e = \xi_{0.5} =$ quantile of order 0.5=median,

$q_3 = \xi_{0.75} =$ quantile of order 0.75 (upper quantile of X).

So, q_1, q_2, q_3 divide the probability distribution of X into 4 parts so that

$$F_X(q_1-) \leq \frac{1}{4} \leq F_X(q_1),\ \ F_X(q_2-) \leq \frac{1}{2} \leq F_X(q_2)\ \text{ and }\ F_X(q_3-) \leq \frac{3}{4} \leq F_X(q_3).$$

Note that q_1, q_2 and q_3 divide the p.d.f. / p.m.f. of X into 4 parts so that each of them has 25% probability mass.

Define $IQR = q_3 - q_1 \rightarrow$ inter-quantile range, $QD = \frac{q_3 - q_1}{2} \rightarrow$ quantile deviation or the semi-interquantile range.

· Unlike SD, QD is not sensitive to extreme values assumed by X.

· Does not at all take into account numerical values of X.

· Ignores the tail of the probability distribution (constituting 50% of probability diistributin on left side of q_1 and right side of q_3).

· QD depends on the unit of measurements of X and thus it may not be appropriate for comparing dispersions of two probability distributions having different units of measurements. For this purpose one may use $CQD = \frac{q_3 - q_1}{q_3 + q_1} \rightarrow$ coefficient of quartile deviation. It does not depend on units of measurements.

(d) Coefficient of Variation:

Like QD, the SD σ also depends on units of measurements of r.v. X and thus it is not an appropriate measure of dispersion for comparing distributions having different units of measurements. For this purpose we consider

$$CV(\text{coefficient of variation}) = \frac{\sigma}{\mu},$$

where $\mu = E(X), \;\; \sigma = \sqrt{\text{Var}(X)}$. Here, we assume $\mu \neq 0$.

· CV measures variation per unit of mean.

· CV does not depend on the unit of measurements of r.v. X.

· CV is very sensitive to small changes in μ when μ is near 0.

(3) Measure of Skewness:

Skewness of a probability distribution is a measure of its asymmetry (lack of symmetry).

Recall that: Distribution of X is symmetric about $\mu \iff X - \mu \overset{d}{=} \mu - X \iff f_X(\mu + x) = f_X(\mu - x), \forall\, x \in \mathbb{R}$ and in that case

· $\mu = E(X) = m_e$ (median);

· The shape of the p.d.f. / p.m.f. on the left of μ is the mirror image of that on the right side of μ.

Positively Skewed Distributions:

· Have more probability mass to the right side of p.d.f. / p.m.f.

· Have longer tails on the right side of p.d.f.

For unimodal positively skewed distribution, normally

$$\text{Mode} < \text{Median} < \text{Mean}$$

since the positive mass to large values of X pulls up the values of mean μ.

Negatively Skewed Distributions:

· Have more probability mass to the left side of the p.d.f. / p.m.f.

· Have longer tails on the left side of p.d.f.

For unimodal negatively skewed distributions, normally

$$\text{Mean} < \text{Median} < \text{Mode}.$$

Let $E(X) = \mu,\ \sqrt{\text{Var}(X)} = \sigma$ and $Z = \dfrac{X-\mu}{\sigma}$: standardized variable (independent of units). Define

$$\text{Coefficient of skewness} = \beta_1 = E(Z^3) = \frac{E((X-\mu)^3)}{\sigma^3} = \frac{\mu_3}{\mu_2^{3/2}},\ \text{ where } \mu_r = E((X-\mu)^r),\ \ r = 1, 2, \ldots$$

· For symmetric distributions $\beta_1 = 0$. Converse may not be true.

· For positively skewed distributions, normally β_1 is large positive quantity.

· For negatively skewed distributions, normally β_1 is a small negative quantity.

A measure of skewness can also be based on quantiles. Let q_1 : first quantile, m_e : Median (or second quantile q_2), q_3 : third quantile, μ : mean.

· For symmetric distributions: $q_3 - m = m - q_1 \left(m = \dfrac{q_1+q_3}{2}\right)$.

· For positively skewed distributions: $q_3 - m > m - q_1$.

· For negatively skewed distributions: $q_3 - m < m - q_1$.

Thus a measure of skewness can be based on $(q_3 - m) - (m - q_1) = q_3 - 2m + q_1$. Define

$$\text{Yule coefficient of skewness} = \beta_2 = \frac{(q_3-m)-(m-q_1)}{q_3-q_1} = \frac{q_3-2m+q_1}{q_3-q_1}\ \ \text{(independent of units)}.$$

Clearly for positively / negatively skewed distribution $\beta_2 > 0/\beta_2 < 0$ and for symmetric distributions $\beta_2 = 0$.

(4) Measures of Kurtosis:

For $\mu \in \mathbb{R}$ and $\sigma > 0$, let $Y_{\mu,\sigma}$ be a r.v. having p.d.f.

$$f_{Y_{\mu,\sigma}}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},\ \ -\infty < x < \infty\ \ \text{(Normal distribution, } Y_{\mu,\sigma} \sim N(\mu, \sigma^2)).$$

It can be shown that

· $E(Y_{\mu,\sigma}) = \mu$, $\text{Var}(Y_{\mu,\sigma}) = \sigma^2$;

· $Y_{\mu,\sigma} - \mu \overset{d}{=} \mu - Y_{\mu,\sigma}$ and hence $\beta_1 = 0$, $E((Y_{\mu,\sigma} - \mu)^4) = 3\sigma^4$;

· $f_{Y_{\mu,\sigma}}(\cdot)$ is unimodal and symmetric.

Kurtosis of the probability distribution of X is a measure of peakedness and thickness of tails of p.m.f. / p.d.f. of X relative to that of normal distribution.

A disribution is said to have higher (lower) kurtosis than the normal distribution if its p.m.f. / p.d.f. in comparison with p.d.f. of a normal distribution, has a sharper (rounded) peak and longer, fatter (shorter, thinner) tails.

Define $Z = \dfrac{X-\mu}{\sigma}$ (independent of units)

$$\nu_1 = E(Z^4) = \frac{E((X-\mu)^4)}{\sigma^4} = \frac{\mu_4}{\mu_2^2} \to \text{Kurtosis of the probability distribution of } X.$$

ν_1 is used as a measure of kurtosis for unimodal distributions. For $N(\mu, \sigma^2)$ distribution, $\nu_1 = 3$. The quantity $\nu_2 = \nu_1 - 3$ is called the excess kurtosis of the distribution of X. Obviously for normal distributions, $\nu_2 = 0$.

Mesokurtic distributions: Distributions with $\nu_2 = 0$, Leptokurtic distributions: Distributions with $\nu_2 > 0$ (has sharper peak and longer, fatter tails).

Platykurtic distributions: Distributions with $\nu_2 < 0$ (has rounded peak and shorter, thinner tails).

Example 11.1. *For $\alpha \in [0,1]$, let X_α has the p.d.f.*

$$f_\alpha(x) = \begin{cases} \alpha e^x, & x < 0, \\ (1-\alpha)e^{-x}, & x \geq 0. \end{cases}.$$

Recall that for $r \in \{1, 2, \dots\}$

$$I_r = \int_0^\infty x^{r-1} e^{-x} \mathrm{d}x = (r-1)! \quad (\textit{using integration by parts}).$$

Thus, for $r \in \{1, 2, \dots\}$

$$\begin{aligned} \mu_r'(\alpha) = E(X_\alpha^r) &= \int_{-\infty}^0 \alpha x^r e^x \mathrm{d}x + \int_0^\infty (1-\alpha) x^r e^{-x} \mathrm{d}x \\ &= ((-1)^r \alpha + 1 - \alpha) \int_0^\infty x^r e^{-x} \mathrm{d}x \\ &= \begin{cases} (1-2\alpha) r!, & r \in \{1, 3, 5, \dots\}, \\ r!, & r \in \{2, 4, 6, \dots\}. \end{cases}. \end{aligned}$$

Let ξ_p be the quantile of order $p \in (0,1)$. Then $F_\alpha(\xi_p) = p$, where F_α is the d.f. of X_α. Clearly $F_\alpha(0) = \alpha \int_{-\infty}^0 e^x \mathrm{d}x = \alpha$. For $0 \leq \alpha < p$, we have

$$p = F_\alpha(\xi_p) = \int_{-\infty}^0 \alpha e^x \mathrm{d}x + \int_0^{\xi_p} (1-\alpha) e^{-x} \mathrm{d}x = 1 - (1-\alpha) e^{-\xi_p}$$

and for $\alpha \geq p$

$$p = \int_{-\infty}^{\xi_p} \alpha e^x \mathrm{d}x = \alpha e^{\xi_p}.$$

Thus,

$$\xi_p = \begin{cases} \ln\left(\frac{1-\alpha}{1-p}\right), & \textit{if } 0 \leq \alpha < p, \\ -\ln\left(\frac{\alpha}{p}\right), & \textit{if } p \leq \alpha \leq 1, \end{cases}$$

$$q_1(\alpha) = \xi_{1/4} = \begin{cases} \ln\left(\frac{4(1-\alpha)}{3}\right), & \textit{if } 0 \leq \alpha < \frac{1}{4}, \\ -\ln(4\alpha), & \textit{if } \frac{1}{4} \leq \alpha \leq 1, \end{cases}$$

$$m_e(\alpha) = \xi_{1/2} = \begin{cases} \ln(2(1-\alpha)), & \textit{if } 0 \leq \alpha < \frac{1}{2}, \\ -\ln(2\alpha), & \textit{if } \frac{1}{2} \leq \alpha \leq 1, \end{cases}$$

$$q_3(\alpha) = \xi_{3/4} = \begin{cases} \ln(4(1-\alpha)), & \textit{if } 0 \leq \alpha < \frac{3}{4}, \\ -\ln\left(\frac{4\alpha}{3}\right), & \textit{if } \frac{3}{4} \leq \alpha \leq 1, \end{cases}$$

$$\begin{aligned} &\mu_1'(\alpha) = E(X_\alpha) = 1 - 2\alpha, \\ &\textit{Mode} = m_0(\alpha) = \sup\{f_\alpha(x) : -\infty < x < \infty\} = \max\{\alpha, 1-\alpha\}, \\ &\mu_2'(\alpha) = E(X_\alpha^2) = 2, \;\; \sigma(\alpha) = \sqrt{\operatorname{Var}(X_\alpha)} = \sqrt{1 + 4\alpha - \alpha^2}. \end{aligned}$$

Note that, for $0 \le \alpha < \frac{1}{2}$, $m_e(\alpha) = \ln(2(1-\alpha)) \ge 0$ *and for* $\alpha > \frac{1}{2}$, $m_e(\alpha) = -\ln(2\alpha) < 0$. *Thus, for* $0 \le \alpha < \frac{1}{2}$ *(so that* $m_e(\alpha) \ge 0$*)*

$$\begin{aligned}
MD(m_e(\alpha)) &= E(|X - m_e(\alpha)|) \\
&= \alpha \int_{-\infty}^{0} (m_e(\alpha) - x)e^{x}\mathrm{d}x + (1-\alpha)\int_{0}^{m_e(\alpha)} (m_e(\alpha) - x)e^{-x}\mathrm{d}x + (1-\alpha)\int_{m_e(\alpha)}^{\infty} (x - m_e(\alpha))e^{-x}\mathrm{d}x \\
&= m_e(\alpha) + 2\alpha = \ln(2(1-\alpha)) + 2\alpha.
\end{aligned}$$

Similarly, for $\frac{1}{2} \le \alpha \le 1$ *(so that* $m_e(\alpha) \le 0$*)*

$$\begin{aligned}
MD(m_e(\alpha)) &= E(|X - m_e(\alpha)|) \\
&= \alpha \int_{-\infty}^{m_e(\alpha)} (m_e(\alpha) - x)e^{x}\mathrm{d}x + \alpha\int_{m_e(\alpha)}^{0} (x - m_e(\alpha))e^{x}\mathrm{d}x + (1-\alpha)\int_{0}^{\infty} (x - m_e(\alpha))e^{-x}\mathrm{d}x \\
&= 2(1-\alpha) - m_e(\alpha) = \ln(2\alpha) + 2(1-\alpha).
\end{aligned}$$

Thus,

$$MD(m_e(\alpha)) = \begin{cases} \ln(2(1-\alpha)) + 2\alpha, & \textit{if } 0 \le \alpha < \frac{1}{2}, \\ \ln(2\alpha) + 2(1-\alpha), & \textit{if } \frac{1}{2} \le \alpha \le 1, \end{cases}$$

$$IQR \equiv IQR(\alpha) = q_3(\alpha) - q_1(\alpha) = \begin{cases} \ln 3, & \textit{if } 0 \le \alpha < \frac{1}{4} \textit{ or } \frac{3}{4} \le \alpha \le 1, \\ \ln(16\alpha(1-\alpha)), & \textit{if } \frac{1}{4} \le \alpha < \frac{3}{4}, \end{cases}$$

$$QD \equiv QD(\alpha) = \frac{q_3(\alpha) - q_1(\alpha)}{2} = \begin{cases} \ln\sqrt{3}, & \textit{if } 0 \le \alpha < \frac{1}{4}, \\ \ln(4\sqrt{\alpha(1-\alpha)}, & \textit{if } \frac{1}{4} \le \alpha < \frac{3}{4}, \\ ln\sqrt{3}, & \textit{if } \frac{3}{4} \le \alpha \le 1, \end{cases}$$

$$CQD \equiv CQD(\alpha) = \frac{q_3(\alpha) - q_1(\alpha)}{q_3(\alpha) + q_1(\alpha)} = \begin{cases} \dfrac{\ln 3}{\ln\left(\frac{16(1-\alpha)^2}{3}\right)}, & \textit{if } 0 \le \alpha < \frac{1}{4}, \\ \dfrac{\ln(16\alpha(1-\alpha))}{\ln\left(\frac{(1-\alpha)}{\alpha}\right)}, & \textit{if } \frac{1}{4} \le \alpha \le \frac{3}{4}, \\ -\dfrac{\ln 3}{\ln\left(\frac{16\alpha^2}{3}\right)}, & \textit{if } \frac{3}{4} \le \alpha \le 1. \end{cases}$$

For $\alpha \ne \frac{1}{2}$,

$$CV \equiv CV(\alpha) = \frac{\sigma(\alpha)}{\mu_1'(\alpha)} = \frac{\sqrt{1 + 4\alpha - 4\alpha^2}}{1 - 2\alpha},$$

$$\mu_3(\alpha) = E((X_\alpha - \mu_1'(\alpha))^3) = \mu_3'(\alpha) - 3\mu_1'(\alpha)\mu_2'(\alpha) + 2(\mu_1'(\alpha))^3 = 2(1-2\alpha)^3,$$

$$\beta_1 \equiv \beta_1(\alpha) = \frac{\mu_3(\alpha)}{\sigma(\alpha)} = \frac{2(1-2\alpha)^3}{\sqrt{1 + 4\alpha - 4\alpha^2}},$$

$$\beta_2 \equiv \beta_2(\alpha) = \frac{q_3(\alpha) - 2m(\alpha) + q_1(\alpha)}{q_3(\alpha) - q_1(\alpha)} = \begin{cases} \dfrac{\ln(\frac{4}{3})}{\ln 3}, & \textit{if } 0 \le \alpha < \frac{1}{4}, \\ -\dfrac{\ln(4\alpha(1-\alpha))}{\ln(16\alpha(1-\alpha))}, & \textit{if } \frac{1}{4} \le \alpha < \frac{1}{2}, \\ \dfrac{\ln(4\alpha(1-\alpha))}{\ln(16\alpha(1-\alpha))}, & \textit{if } \frac{1}{2} \le \alpha \le \frac{3}{4}, \\ \dfrac{\ln(\frac{3}{4})}{\ln 3}, & \textit{if } \frac{3}{4} \le \alpha \le 1. \end{cases}$$

Clearly, for $0 \leq \alpha < \frac{1}{2}$, $\beta_i(\alpha) > 0$, $i = 1, 2$ *and for* $\frac{1}{2} < \alpha \leq 1$, $\beta_i(\alpha) < 0$, $i = 1, 2$. *For* $\alpha = \frac{1}{2}$, $\beta_i(\alpha) = 0$, $i = 1, 2$. *Thus,*

· *for* $0 \leq \alpha < \frac{1}{2}$, *distribution of* X_α *is positively skewed;*

· *for* $\frac{1}{2} < \alpha \leq 1$, *distribution of* X_α *is negatively skewed;*

· *for* $\alpha = \frac{1}{2}$, *distribution of* X_α *is symmetric (infact in this case* $f_\alpha(x) = f_\alpha(-x)$, $\forall\ x \in \mathbb{R}$*).*

$\mu_4 \equiv \mu_4(\alpha) = E((X_\alpha - \mu_1'(\alpha))^4) = \mu_4'(\alpha) - 4\mu_1'(\alpha)\mu_3'(\alpha) + 6(\mu_1'(\alpha))^2\mu_2'(\alpha) - 3(\mu_1'(\alpha))^4 = 24 - 12(1-2\alpha)^2 - 3(1-2\alpha)^4$

$$\nu_1 \equiv \nu_1(\alpha) = \frac{\mu_4(\alpha)}{(\mu_2(\alpha))^2} = \frac{24 - 12(1-2\alpha)^2 - 3(1-2\alpha)^4}{(2 - (1-2\alpha)^2)^2}$$

and

$$\nu_2 \equiv \nu_2(\alpha) - 3 = \frac{12 - 6(1-2\alpha)^4}{(2 - (1-2\alpha)^2)^2}.$$

Clearly, for any $\alpha \in [0, 1]$, $\nu_2(\alpha) > 0$. *It follows that for any value of* $\alpha \in [0, 1]$ *the distribution of* X_α *is leptokurtic.*