Lecture 8: Population Summary Characteristics

Ye Tian

Department of Statistics, Columbia University
Calculus-based Introduction to Statistics (S1201)

July 18, 2022

Recap: random variables

Given an experiment and the sample space Ω, a **random variable** is a **function** mapping an outcome ($\omega \in \Omega$) into a real number, i.e.

$$X : \omega \in \Omega \mapsto X(\omega) \in (-\infty, +\infty).$$

- When the possible values of a random variable are countable[1], the random variable is **discrete**.
  Examples: the number of heads/tails of coin flipping, the number of dice etc.
- When *both* of the following apply, the random variable is **continuous**.
  - The range is uncountable (e.g.: an interval on the number line)
  - No possible value of the variable has positive probability, i.e. $\mathbb{P}(X = c) = 0$ for any number c.

  Examples: the temperature at a random location

[1] either constitute a finite set or else can be listed in an infinite sequence in which there is a first element, a second element, and so on ("countably" infinite)

Review: Probability distribution

The **(probability) distribution** of a random variable X describes how the total probability of 1 is **distributed** among all possible values of X. It tells us $\mathbb{P}(X \in \mathcal{B}) = \mathbb{P}(\{\omega : X(\omega) \in \mathcal{B}\})$ for any subset $\mathcal{B}$ of number line [2].

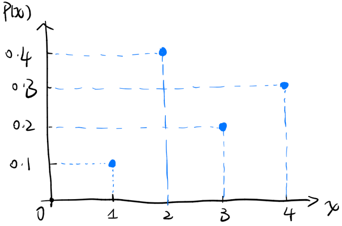
Three characterizations of the distribution of a random variable:

- cdf: $F(x) = \mathbb{P}(X \leq x)$
- pmf (discrete variables only): $p(x) = \mathbb{P}(X = x)$
- pdf (continuous variables only): f such that (any one of the following holds, actually they're equivalent definitions):
  - $\mathbb{P}(a < X \leq b) = F(b) - F(a) = \int_a^b f(x)dx$, for any numbers a, b
  - $f(x) = F'(x)$, for any number x
  - $F(x) = \int_{-\infty}^x f(t)dt$, for any number x

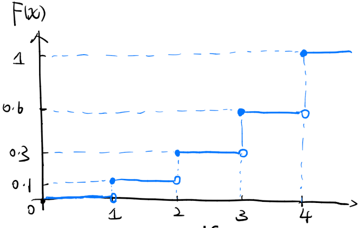
[2] Currently, $\mathcal{B}$ can be any union/intersection of intervals/points on the real line. E.g., $\mathcal{B}$ can be $(0,1)$, $[-2,+\infty)$, $(5,5.5]$, $\{1\}$, $\{-1,2.5\}$, $(-3,-1)\cup(9,10]$ etc.

Comparison: discrete and continuous variables

Discrete distribution:

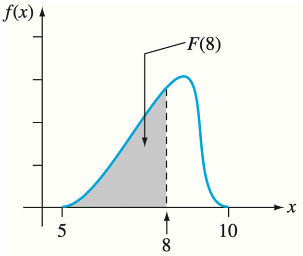


pmf

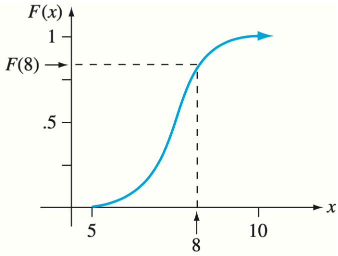


cdf

Continuous distribution:



pdf



cdf

Today's goal

- Understand the definition and interpretation of expectation and variance
- Given a distribution, know how to calculate the expectation and variance
- Understand the definition of other population characteristics (e.g.: median, quantile etc.)
- Have a better understanding on population and samples (today, tomorrow and Wednesday)

Joint Distribution of Multiple Variables

Joint distribution and independence between random variables

The **joint distribution** of two random variables X and Y describes how the total probability of 1 is **distributed** among all possible values of pair (X, Y). It tells us

$$\mathbb{P}(X \in \mathcal{B}_1, Y \in \mathcal{B}_2) = \mathbb{P}(\{\omega : X(\omega) \in \mathcal{B}_1\} \cap \{\omega : Y(\omega) \in \mathcal{B}_2\})$$

for any subsets $\mathcal{B}_1$ and $\mathcal{B}_2$ of number line.

We can define cdf, pmf and pdf for the joint distribution of (X, Y) similarly:

- cdf: $F(x, y) = \mathbb{P}(X \leq x, Y \leq y)$
- pmf (discrete variables only): $p(x, y) = \mathbb{P}(X = x, Y = y)$
- pdf (continuous variables only): f such that (any one of the following holds, actually they're equivalent definitions):
    - $\mathbb{P}(a < X \leq b, c < Y \leq d) = \int_c^d \int_a^b f(x, y) dx dy$, for any numbers a, b, c and d
    - $f(x, y) = \frac{\partial^2 F}{\partial x \partial y}(x, y)$, for any number x
    - $F(x, y) = \int_{-\infty}^y \int_{-\infty}^x f(r, t) dr dt$, for any numbers x and y

Joint distribution and independence between random variables

Similar to our study on the independence relationship between multiple **events**, we can also study the independence between **random variables**.

But I do want everyone to get a sense what this independence between variables means.

- Recall that $\mathbb{P}(X \in \mathcal{B}) = \mathbb{P}(\{\omega : X(\omega) \in \mathcal{B}\})$. Actually r.v.'s X and Y are independent iff $\mathbb{P}(X \in \mathcal{B}_1, Y \in \mathcal{B}_2) = \mathbb{P}(X \in \mathcal{B}_1)\mathbb{P}(Y \in \mathcal{B}_2)$ for any subsets $\mathcal{B}_1$ and $\mathcal{B}_2$ of the number line, i.e.
$$\{\omega : X(\omega) \in \mathcal{B}_1\} \perp\!\!\!\perp \{\omega : Y(\omega) \in \mathcal{B}_2\}.$$
This means independence between r.v.'s is essentially the independence between multiple pairs of events (n-"tuple" for n r.v.'s)!
- An equivalent definition for two **continuous** r.v.'s X and Y: $X \perp\!\!\!\perp Y$ iff
$$f_{(X,Y)}(x, y) = f_X(x) f_Y(y),$$
where $f_{(X,Y)}(x, y)$ are the joint pdf of (X, Y), f_X and f_Y are the marginal pdf of X and Y.

Independence between random variables

Suppose the joint pmf or pdf of $X_1, \ldots, X_n$ is $f(x_1, \ldots, x_n)$. And the marginal pmf or pdf of X_i is $f_i(x_i)$. Then $X_1, \ldots, X_n$ are independent if and only if

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_i(x_i),$$

for any numbers $x_1, \ldots, x_n$.

Things you need to know about the joint distribution and independence between r.v.'s for this course

- r.v.'s $X \perp\!\!\!\perp Y$ iff $\mathbb{P}(X \in \mathcal{B}_1, Y \in \mathcal{B}_2) = \mathbb{P}(X \in \mathcal{B}_1)\mathbb{P}(Y \in \mathcal{B}_2)$ for any subsets $\mathcal{B}_1$ and $\mathcal{B}_2$ of the number line, i.e.
$$\{\omega : X(\omega) \in \mathcal{B}_1\} \perp\!\!\!\perp \{\omega : Y(\omega) \in \mathcal{B}_2\}.$$
This means independence between r.v.'s is essentially the independence between multiple pairs of events (n-"tuple" for n r.v.'s)!
- Know that we can similarly define the cdf, pmf and pdf for joint distribution as we did for a single variable
- Know how to calculate the expectation and variance of sum of independent variables
- Understand the definitions of covariance and correlation of two variables
- Know the distribution of sum of independent normally-distributed variables

Expectation and Variance

Population universe and sample universe

Sample universe (what we see)

Data ⇒

- Sample mean
- Sample variance
- Sample standard deviation
- Sample quantiles (sample median, quartiles...)
- Sample maximum/minimum
- Sample covariance/correlation
- Bar chart (discrete r.v.)
- Histogram (continuous r.v.)

Population universe (inaccessible)

Probability distribution ⇒

- Expectation/Mean
- Variance
- Standard deviation
- Quantiles (median, quartiles ...)
- Maximum/minimum
- Covariance/correlation
- pmf (discrete r.v.)
- pdf (continuous r.v.)

Later, we will know why we can use the sample information to infer the population universe.

Expectation or expected values

Sample mean: Data $x_1, \dots, x_n \Rightarrow \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

Definition: The **expectation** of a r.v. X is defined as

$$\mathbb{E}X = \mu_X = \begin{cases} \sum_j x_j \times \mathbb{P}(X = x_j), & X \text{ is discrete,} \\ \int_{-\infty}^{+\infty} x f(x) dx, & X \text{ is continuous and } f \text{ is the pdf} \end{cases}$$

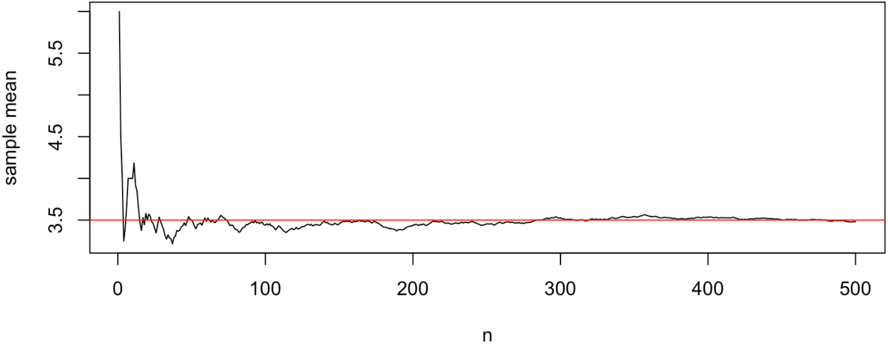
- Expectation (Or expected value) of a random variable is its **long-run average**. When $x_1, \dots, x_n \overset{i.i.d.}{\sim} X$ and $n \to +\infty$, $\bar{x} \approx \mathbb{E}X$. [3]
- $\mathbb{E}X$ indicates the "center" of X. It is not a value that is likely/expected to get.

[3] $i.i.d.$ means "identically and independent distributed (as)"

Expectation or expected values

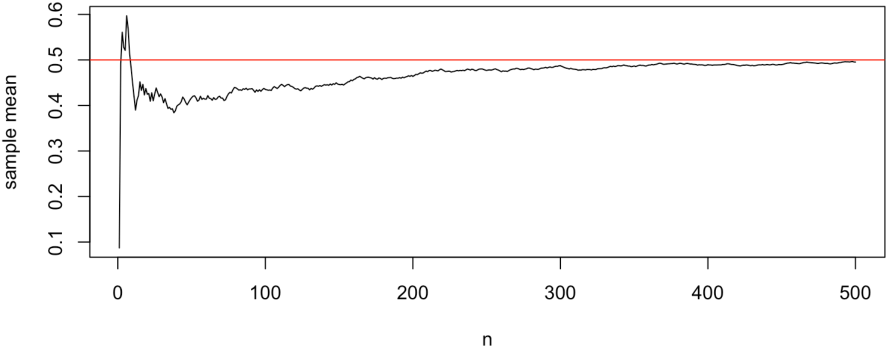
Example 1: Rolling a die. What is the expected number of spots X on the top?

$\mathbb{P}(X = 1) = \cdots = \mathbb{P}(X = 6) = \frac{1}{6} \Rightarrow \mathbb{E}X = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + \cdots + 6 \times \frac{1}{6} = 3.5.$



Expectation or expected values

Example 2: X follows a uniform distribution on $[0,1]$. (i.e. pdf $f(x) = 1$ on $[0,1]$ and 0 elsewhere)
$\mathbb{E}X = \int_0^1 xf(x)dx = \int_0^1 xdx = \frac{1}{2}x^2 \mid_0^1 = \frac{1}{2}$.



Expectation or expected values

Example 3: For a rare disease with 1% prevalence rate, the following group testing is used.

- Pull the blood sample of 10 people together.
- If the result is negative, all of them are negative.
- If the result is positive, test them individually.

If each test costs \$1, what is the expected cost?

Let X = the cost. Then the pmf of X is

- $\mathbb{P}(X = 1) = \mathbb{P}(\{\text{all 10 people are negative}\}) = 0.99^{10} = 0.9044$
- $\mathbb{P}(X = 11) = 1 - \mathbb{P}(\{\text{all 10 people are negative}\}) = 0.0956$

$\Longrightarrow \mathbb{E}X = 1 \times \mathbb{P}(X = 1) + 11 \times \mathbb{P}(X = 11) = 1.956$ (dollars).

Expectation or expected values

Expected value of a **function** of a random variable:

Theorem: Suppose new r.v. $Y = g(X)$, where g is some function on the number line. Then the expectation of Y equals

$$\mathbb{E}Y = \mathbb{E}(g(X)) = \begin{cases} \sum_j g(x_j)\mathbb{P}(X = x_j), & X \text{ is discrete,} \\ \int_{-\infty}^{+\infty} g(x) f(x) dx, & X \text{ is continuous and } f \text{ is the pdf} \end{cases}$$

You will prove the discrete case in HW3!

Expectation or expected values

Example 1 (revisited): Rolling a die. Suppose X is the number of spots on the top. What's the expected value of X^2?

Define $Y = X^2$. We know $\mathbb{P}(X=1) = \cdots = \mathbb{P}(X=6) = \frac{1}{6}$

Way 1: (previous theorem) $\mathbb{E}X^2 = 1^2 \times \frac{1}{6} + 2^2 \times \frac{1}{6} + \cdots + 6^2 \times \frac{1}{6} = 15.17$

Way 2: (definition of $\mathbb{E}Y$) We know $\mathbb{P}(Y = 1^2) = \mathbb{P}(Y = 2^2) = \cdots = \mathbb{P}(Y = 6^2) = \frac{1}{6}\mathbb{E}Y = 1^2 \times \frac{1}{6} + 2^2 \times \frac{1}{6} + \cdots + 6^2 \times \frac{1}{6} = 15.17$

Example 2 (revisited): X follows a uniform distribution on $[0,1]$. (i.e. pdf $f(x) = 1$ on $[0,1]$ and 0 elsewhere)
$\mathbb{E}X^2 = \int_0^1 x^2 f(x)dx = \int_0^1 x^2 dx = \frac{1}{3}x^3 \mid_0^1 = \frac{1}{3}$.

Expectation or expected values

Linearity of expectations: Suppose a and b are two constants (numbers), then $\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$.

Can be easily proved by considering a function of r.v. X as $Y = aX + b$, then calculating the expected value of function of r.v. You will prove it in HW3.

Variance and standard deviation

Sample variance: Data $x_1, \ldots, x_n \Rightarrow s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

Definition: The **variance** of a r.v. X is defined as

$$\mathsf{Var}(X) = \mathbb{E}(X - \mathbb{E}X)^2$$
$$= \begin{cases} \sum_j (x_j - \mathbb{E}X)^2 \times \mathbb{P}(X = x_j), & X \text{ is discrete,} \\ \int_{-\infty}^{+\infty} (x - \mathbb{E}X)^2 f(x) dx, & X \text{ is continuous and } f \text{ is pdf} \end{cases}$$

The **standard deviation** of X is $\mathsf{SD}(X) = \sqrt{\mathsf{Var}(X)}$

- The standard deviation shows the likely size of the deviation of the r.v. X from $\mathbb{E}X$.
- When $x_1, \ldots, x_n \overset{i.i.d.}{\sim} X$ and $n \to +\infty$, $s^2 \approx \mathsf{Var}(X)$.
- The variance and standard deviation is always non-negative.
- When $\mathsf{Var}(X) = 0$, we have $\mathbb{P}(X = \mathbb{E}X) = 1$.

Variance and standard deviation

A shortcut formula to calculate Var(X) : $\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2$

Remark: Note that $\mathbb{E}(X^2)$ and $(\mathbb{E}X)^2$ are totally different!

$$\mathbb{E}(X^2) = \begin{cases} \sum_j x_j^2 \times \mathbb{P}(X = x_j), & X \text{ is discrete,} \\ \int_{-\infty}^{+\infty} x^2 f(x) dx, & X \text{ is continuous and } f \text{ is the pdf} \end{cases}$$

$$(\mathbb{E}X)^2 = \begin{cases} \left[\sum_j x_j \times \mathbb{P}(X = x_j) \right]^2, & X \text{ is discrete,} \\ \left[\int_{-\infty}^{+\infty} x^2 f(x) dx \right]^2, & X \text{ is continuous and } f \text{ is the pdf} \end{cases}$$

Proof:

$$\begin{aligned} \mathrm{Var}(X) = \mathbb{E}(X - \mathbb{E}X)^2 &= \mathbb{E}[X^2 - 2X(\mathbb{E}X) + (\mathbb{E}X)^2] \\ &= \mathbb{E}(X^2) - 2(\mathbb{E}X) \times (\mathbb{E}X) + (\mathbb{E}X)^2 \\ &= \mathbb{E}(X^2) - (\mathbb{E}X)^2 \end{aligned}$$

Question: Why does "$=$" hold?

Variance and standard deviation

Example 1 (revisited): Rolling a die. Suppose X is the number of spots on the top.

We know $\mathbb{P}(X=1) = \cdots = \mathbb{P}(X=6) = \frac{1}{6}$

We have already obtained $\mathbb{E}X^2 = 1^2 \times \frac{1}{6} + 2^2 \times \frac{1}{6} + \cdots + 6^2 \times \frac{1}{6} = 15.17$, $\mathbb{E}X = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + \cdots + 6 \times \frac{1}{6} = 3.5$.Then by the shortcut formula,

$$\text{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2 = 15.7 - 3.5^2 = 3.45,$$
$$\text{SD}(X) = \sqrt{\text{Var}(X)} = 1.86.$$

Example 2 (revisited): X follows a uniform distribution on $[0,1]$. (i.e. pdf $f(x) = 1$ on $[0,1]$ and 0 elsewhere)

We have already obtained $\mathbb{E}X^2 = \int_0^1 x^2 f(x)dx = \int_0^1 x^2 dx = \frac{1}{3}x^3 \mid_0^1 = \frac{1}{3}$, $\mathbb{E}X = \int_0^1 x f(x)dx = \int_0^1 x dx = \frac{1}{2}x^2 \mid_0^1 = \frac{1}{2}$. Then by the shortcut formula,

$$\text{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2 = 1/3 - (1/2)^2 = 1/12,$$
$$\text{SD}(X) = \sqrt{\text{Var}(X)} = 0.289.$$

Variance and standard deviation

"Linearity" of variances: Suppose a and b are two constants (numbers), then:

- $\mathsf{Var}(aX + b) = a^2\mathsf{Var}(X)$
- $\mathsf{SD}(aX + b) = |a|\mathsf{SD}(X)$

Can be easily proved by the definition of variance $\mathsf{Var}(X) = \mathbb{E}(X - \mathbb{E}X)^2$. You will prove it in HW3.

For the variance of a **function** of a random variable: first use the shortcut formula, then use the conclusion for expectations:
$\mathsf{Var}[g(X)] = \mathbb{E}[g^2(X)] - [\mathbb{E}g(X)]^2$.

Other Summary Characteristics of Distributions

Quantiles

Sample quantiles: Data $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)} \Rightarrow k$-th percentile or $k\%$ quantile is the value where $k\%$ of the values are below.

Definition: The k-th **percentile** or $k\%$ **quantile** of a **continuous** r.v. X is defined as the number $\eta(k\%)$ that satisfies

$$k\% = F(\eta(k\%)) = \mathbb{P}(X \leq \eta(k)) = \int_{-\infty}^{\eta(k)} f(x)dx,$$

where f and F are the pdf and cdf of X.

- We don't study the discrete case here. But you can imagine that the trick of interpolation can be applied to help define the quantile.
- **Median** $= \eta(50\%)$.
- When $x_1, \ldots, x_n \overset{i.i.d.}{\sim} X$ and $n \to +\infty$, sample $k\%$ quantile $\approx \eta(k\%)$.

Quantiles

Example 2 (revisited): X follows a uniform distribution on $[0, 1]$. (i.e. pdf $f(x) = 1$ on $[0, 1]$ and 0 elsewhere)

By definition, $F(x) = \int_{-\infty}^{x} 1dt = \int_{0}^{x} 1dt = x$ when $x \in [0, 1]$. Thus by letting

$$F(\eta(k\%)) = 0.01k,$$

we obtain

$$\eta(k\%) = 0.01k,$$

for $k \in [0, 100]$.
$\Rightarrow k\%$ quantile $= 0.01k$.

Covariance

Covariance and correlation help us understand the strength of **linear** association/correlation between two random variables.

Recall that the variance is defined as $\text{Var}(X) = \mathbb{E}(X - \mathbb{E}X)^2$. The **covariance** between X and Y is "similarly" defined by

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)] = \mathbb{E}(XY) - (\mathbb{E}X)(\mathbb{E}Y)$$

- $\text{Cov}(X, Y)$ can be positive, negative, or zero
- When $\text{Cov}(X, Y) = 0$, we say X and Y are **(linearly) uncorrelated**.
- Independence $\Rightarrow$ uncorrelation, but uncorrelation $\not\Rightarrow$ independence

Correlation

Correlation is a "standardized" version of covariance.

The **correlation** between X and Y is defined by

$$\text{Corr}(X, Y) = \rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

- $-1 \leq \text{Corr}(X, Y) \leq 1$ (can be shown by the integral version of Cauchy-Schwarz inequality, which we will not cover)
- $\rho_{X,Y}$ measures the strength of **linear** association between X and Y. The larger the $|\rho_{X,Y}|$, the stronger the linear association.
- $\rho_{X,Y} > 0$ means positive association/correlation, i.e. larger X tends to associate with large Y and smaller X tends to associate with smaller Y
- When $\rho_{X,Y} = 0$, X and Y are **(linearly) uncorrelated**.
- Independence $\Rightarrow$ uncorrelation, but uncorrelation $\not\Rightarrow$ independence
- $\rho_{X,Y}$ is **invariant** under different measurement units, i.e. $\text{Corr}(aX + b, cY + d) = \text{Corr}(X, Y)$ for any numbers $a \neq 0$, $c \neq 0$, b and d.

Sample covariance and sample correlation

Sample covariance and correlation: Data $(x_1, y_1), \ldots, (x_n, y_n) \Rightarrow$

$$\widehat{\mathrm{Cov}}(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \left(\frac{1}{n-1}\sum_{i=1}^{n} x_i y_i\right) - \bar{x}\bar{y},$$

$$\hat{\rho}_{X,Y} = \frac{\widehat{\mathrm{Cov}}(X,Y)}{s_x s_y}$$

where sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, sample mean $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, sample variance $s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$, $s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$

Compare with the population covariance and correlation:

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

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}.$$

The "linearity" of expectation and variance (A general version)

Suppose $X_1, \ldots, X_n$ are random variables. $a_1, \ldots, a_n, b_1, \ldots, b_n$ are constants.

- $\mathbb{E}[\sum_{i=1}^n (a_i X_i + b_i)] = \sum_{i=1}^n (a_i \mathbb{E} X_i + b_i)$
- If $X_1, \ldots, X_n$ are **independent**, then $\text{Var}[\sum_{i=1}^n (a_i X_i + b_i)] = \sum_{i=1}^n a_i^2 \text{Var}(X_i)$

Remark:

- The conclusion of expectation holds even when $X_1, \ldots, X_n$ are dependent!
- Some simple examples: For r.v. X and Y:
  - $\mathbb{E}(X + 2Y) = \mathbb{E}X + 2\mathbb{E}Y$
  - $\mathbb{E}(-X + 3Y + 10) = -\mathbb{E}X + 3\mathbb{E}Y + 10$
  - $\text{Var}(-X + 3Y + 10) =$ **We don't know!**
  - If $X \perp\!\!\!\perp Y$, then $\text{Var}(-X + 3Y + 10) = \text{Var}(X) + 9\text{Var}(Y)$

Population universe and sample universe

Sample universe (what we see)

Data ⇒

- Sample mean
- Sample variance
- Sample standard deviation
- Sample quantiles (sample median, quartiles...)
- Sample maximum/minimum
- Sample covariance/correlation
- Bar chart (discrete r.v.)
- Histogram (continuous r.v.)

Population universe (inaccessible)

Probability distribution ⇒

- Expectation/Mean
- Variance
- Standard deviation
- Quantiles (median, quartiles ...)
- Maximum/minimum
- Covariance/correlation
- pmf (discrete r.v.)
- pdf (continuous r.v.)

Later, we will know why we can use the sample information to infer the population universe.

Many thanks to

- Chengliang Tang
- Joyce Robbins
- Yang Feng
- Owen Ward
- Wenda Zhou
- And all my teachers in the past 25 years