Seoul National University
Dept. of Civil and Environmental Engineering

Instructor: Junho Song
junhosong@snu.ac.kr

457.212 Statistics for Civil & Environmental Engineers

In-Class Material: Class 03

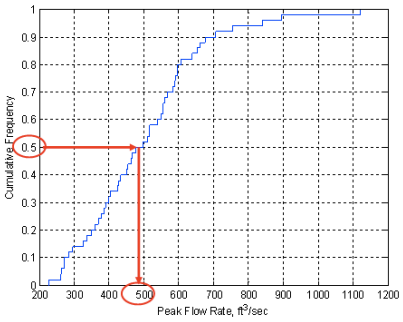
Numerical Descriptors of Data (A&T 1.2-1.3, Supp #1)

Measures for

i) Central tendency: median, s. mean
ii) Dispersion: range, IQR, mean absolute deviation, s. variance, s. standard dev., s.c.o.v.
iii) Asymmetry: skewness
iv) Linear dependence: s. covariance, s. correlation coeff.

1. Measure of Central Tendency

(a) **Median** ($x_{0.5}$): the middle value of the data set, ( )-percentile, ( )-quantile, ( )-quartile



N	odd	even
median	$x_{\left[\frac{N+1}{2}\right]}$	$\frac{x_{[N/2]} + x_{[N/2+1]}}{2}$
	{10, 29, 35}	{10, 29, 35, 49}
	$x_{0.5} =$	$x_{0.5} =$

(b) **Sample mean** ($\bar{x}$): the average of the sample values

$$\bar{x} = \text{———————————} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

* **Example 1:** ( ) is less sensitive to “outliers” (extreme values) than ( )

{1, 2, 3, …, 100, 10^6}

$x_{0.5} =$
$\bar{x} =$

*** Example 2:** In the case of a multi-peak distribution, median and sample mean can be significantly different.

Data Set (N = 2,001)	$x_{0.5}$	$\bar{x}$
{1, ……, 1, 25, 100, ……, 100}		
{24, ……, 24, 25, 26, ……, 26}		

Seoul National University
Dept. of Civil and Environmental Engineering

Instructor: Junho Song
junhosong@snu.ac.kr

2. Measure of Dispersion

(a) **Range**: $r =$

~ depends on ( ), therefore not stable.
~ e.g. range of golf driving distances for 100 and 1,000 hits

(b) **IQR** (Inter Quartile Range) =

~ more stable
~ spread of ( )% population at the center
~ generally, $(x_{1-q} - x_q)$ for small q can be used as a measure of dispersion ($q = 0.25$ for IQR)

How about using "the average of the deviations from the mean" as a measure of dispersion?

- Data set 1: {10, 20, 30, 40}
- Data set 2: {10, 10, 40, 40}

Question 1: Which data set has larger dispersion?

Question 2: What are the sample means?

Question 3: What is the average of the deviations for each data set?

Since "the average of the deviations" does not work …

(c) **Mean Absolute Deviation** (d): average of absolute deviations

$$d = \text{———————————} = \frac{1}{N}\sum_{i=1}^{N} | x_i - \bar{x} |$$

(d) **Sample Variance** (s^2): average of squared deviations

$$s^2 = \text{———————————} = \frac{1}{N}\sum_{i=1}^{N} (x_i - \bar{x})^2$$

(e) **Sample Standard Deviation** (s): square root of sample variance

	d	s^2	s
Data Set 1 {10, 20, 30, 40}			
Data Set 2 {10, 10, 40, 40}			

Instructor: Junho Song
junhosong@snu.ac.kr

(f) "**Unbiased**" sample variance and standard deviations: divide by (*N*–1) instead of (*N*)

> Comparison of dispersion of data sets with different units or quantities? Consider unbiased sample variances of {1, 2, 3} and {2, 4, 6}.

We need a measure of dispersion that is not affected by "scaling" or "unit changes"

(g) **Sample Coefficient of Variation** (C.O.V.; $\hat{\delta}$)

$$\hat{\delta} = \text{——}$$

- dimensionless
- independent of ( ) or ( )
- useful for comparing ( ) of data sets with different magnitude or quantity
- does not work when $\bar{x}$ is close to ( )

> Sample c.o.v. of {1, 2, 3} and {2, 4, 6}?

3. Measure of **Asymmetry**

(a) **Sample Coefficient of Skewness** ($\hat{\theta}$)

$$\hat{\theta} = \text{———————}$$

- Symmetric distribution:
- Asymmetric distribution:
  If positive: "positive skewness" or "skewed to the ( )"
  If negative: "negative skewness" or "skewed to the ( )"

4. Measure of **Linear Dependence** between Two Data Samples

Data given in pairs, i.e. $(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)$ and interested in the dependence.

- "the larger x_i, the larger y_i": ( ) linear dependence
- "the larger x_i, the smaller y_i": ( ) linear dependence

Can be seen from "scatter plots." Numerically?

(a) **Sample Covariance**

$$s_{XY} = \frac{1}{N-1}\left(\qquad\qquad\qquad\qquad\right)$$

~ the sign tells us the trend, but not the ( ) of the dependence

Instructor: Junho Song
junhosong@snu.ac.kr

(b) **Sample Correlation Coefficient**: divide the sample covariance by the product of sample standard deviations

$$r_{XY} = \text{————}$$

- dimensionless
- Bounded by ( ) and ( ): [ ] $\leq r_{xy} \leq$ [ ]
- $r_{XY} \cong -1$: strong ( ) linear dependence
- $r_{XY} \cong 1$: strong ( ) linear dependence
- $r_{XY} \cong 0$: no significant linear dependence

Sketches of scatter plots of these three cases?

5. Matlab® Demonstrations (not required; for your information)

- Quantiles: ~~`quantile`~~ → `prctile_step(x,p)*`
- Median: ~~`median`~~ → use `prctile_step(x,50)*`
- Mean: `mean(x)`
- Maximum: `max(x)`
- Minimum: `min(x)`
- Mean Absolute Deviation: `mad(x)`
- Variance and Covariance: `cov(x)`
- Standard Deviation: `std(x)`
- Skewness: `skewness(x)`
- Sample Correlation Coefficient: `corrcoef(x)`

(*) available at the “Matlab® Routines” section of the course website.