Seoul National University
Dept. of Civil and Environmental Engineering

Instructor: Junho Song
junhosong@snu.ac.kr

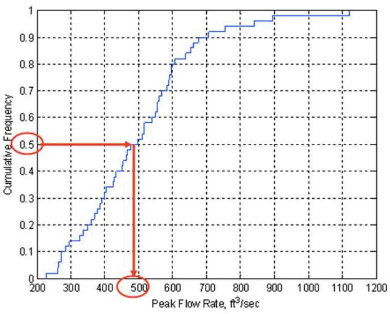
457.212 Statistics for Civil & Environmental Engineers
In-Class Material: Class 03
Numerical Descriptors of Data (A&T 1.2-1.3, Supp #1)

Partial descriptors, measures or descriptors 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	$x_{\left\lceil \frac{N+1}{2} \right\rceil}$ {10, 29, 35} $x_{0.5} =$	$\dfrac{x_{[N/2]} + x_{[N/2+1]}}{2}$ {10, 29, 35, 49} $x_{0.5} =$

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

$$\bar{x} = \rule{6cm}{0.4pt} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

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

{1, 2, 3, …, 100, 10^6}

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

```r
X1 = c(1:100,1000000)

median(X1) # quantile(X1, 0.5) should give the same result
mean(X1)
```

* **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}		

```
X2 = c(array(1,1000),25,array(100,1000))
X3 = c(array(24,1000),25,array(26,1000))

mean(X2)
mean(X3)

median(X2)
median(X3)
```

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)

```
AddisonCreek = read.table("AddisonCreek.txt", header=TRUE)
FR = AddisonCreek$FlowRate

range_FR = diff(range(FR))
IQR_FR = IQR(FR)

# minimum and maximum
min(FR)
max(FR)
```

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?

Instructor: Junho Song
junhosong@snu.ac.kr

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

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

$$d = \rule{4cm}{0.4pt} = \frac{1}{N}\sum_{i=1}^{N}|x_i - \bar{x}|$$

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

$$s^2 = \rule{4cm}{0.4pt} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

(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}			

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

```
x4 = c(10,20,30,40)
x5 = c(10,10,40,40)

mad_x4 = mean(abs(x4-mean(x4)))
mad_x5 = mean(abs(x5-mean(x5)))

var(x4)
var(x5)

sd(x4)
sd(x5)
```

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.; $\hat{\delta}$)

$$\hat{\delta} = \rule{2cm}{0.4pt}$$

- 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}?

```
x6 = c(1,2,3)
x7 = c(2,4,6)

sd(x6)
sd(x7)

sd(x6)/abs(mean(x6))
sd(x7)/abs(mean(x7))
```

3. How to install R packages

- Collections of functions and data sets developed by the community
- Increase the power of R by improving existing base R functions, or by adding new ones

- Example : R package "`moments`"

```
install.packages("moments") # install packages
library(moments) # load and attach add-on packages
```

4. Measure of **Asymmetry**

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

$$\hat{\theta} = \underline{\hspace{4cm}}$$

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

```
skewness(FR) # Compute the skewness coeff. using the function
skewness in "moments" package
```

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

Data given in pairs, i.e. $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$ 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**

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

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

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

$$r_{XY} = \overline{\qquad\qquad}$$

- dimensionless
- Bounded by ( ) and ( ): [ ] $\le r_{xy} \le$ [ ]
- $r_{XY} \cong -1$: 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?

```
HT = AddisonCreek$Height
cov(FR,HT)
cor(FR,HT)
```

6. Example: Computational simulations of steel structures under earthquake ground motions

Download the dataset 'Kim_Collapse.txt' from the eTL website (generated during Mr. Taeyong Kim's PhD research)

Related reference: Deniz, D., J. Song, and J.F. Hajjar (2018). Energy-based sidesway collapse fragilities for ductile structural frames under earthquake loadings. *Engineering Structures*. Vol. 174, 282- 294.

```
# Exercise 01: Make a scatter plot of Velocity Ratio (VR) and Drift Ratio
(DR)

Kim = read.table("Kim_Collapse.txt")
VR = Kim$EquivalentVelocityRatio
DR = Kim$DriftRatio

plot(DR,VR)

# Exercise 02: Compare partial descriptors of two sets - median, mean,
maximum, minimum, variance, standard deviation, and c.o.v.

median(VR); mean(VR); max(VR); min(VR); var(VR); sd(VR);
sd(VR)/abs(mean(VR))
median(DR); mean(DR); max(DR); min(DR); var(DR); sd(DR);
sd(DR)/abs(mean(DR))


# Exercise 03: Compare boxplots of DR and VR (before/after scaling by
means)

boxplot(DR,VR); boxplot(DR/mean(DR),VR/mean(VR))
```