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 02

Graphical Representation of Data (A&T 1.2-1.3, Supp #1)

1. The records of annual peak stream flow (ft^3/sec) of Addison Creek at Bellwood, IL are listed in the following table for the period 1955-2004.

Year	Date	Flow Rate	Year	Date	Flow Rate	Year	Date	Flow Rate
1955	10-11	598	1972	09-17	706	1989	09-01	541
1956	07-19	228	1973	09-24	509	1990	05-09	895
1957	07-13	560	1974	04-13	428	1991	11-27	675
1958	06-13	405	1975	04-18	598	1992	10-25	351
1959	06-25	273	1976	03-04	454	1993	08-23	517
1960	07-02	295	1977	06-30	359	1994	06-24	425
1961	09-14	588	1978	07-20	607	1995	10-31	273
1962	07-02	260	1979	05-03	591	1996	07-17	400
1963	07-16	325	1980	07-21	662	1997	08-16	753
1964	04-06	284	1981	04-28	548	1998	09-07	466
1965	06-08	434	1982	08-07	839	1999	04-09	385
1966	05-12	499	1983	12-03	639	2000	04-20	477
1967	04-01	264	1984	10-22	464	2001	09-20	379
1968	08-16	584	1985	03-04	556	2002	10-13	652
1969	07-17	553	1986	11-19	391	2003	08-01	567
1970	10-10	449	1987	08-14	1120	2004	08-28	516
1971	08-24	336	1988	08-08	370			

* Source: National Water Information System Web Site (NWISWeb); http://waterdata.usgs.gov/nwis

(*) Sort the peak flow rate data in the ascending order.

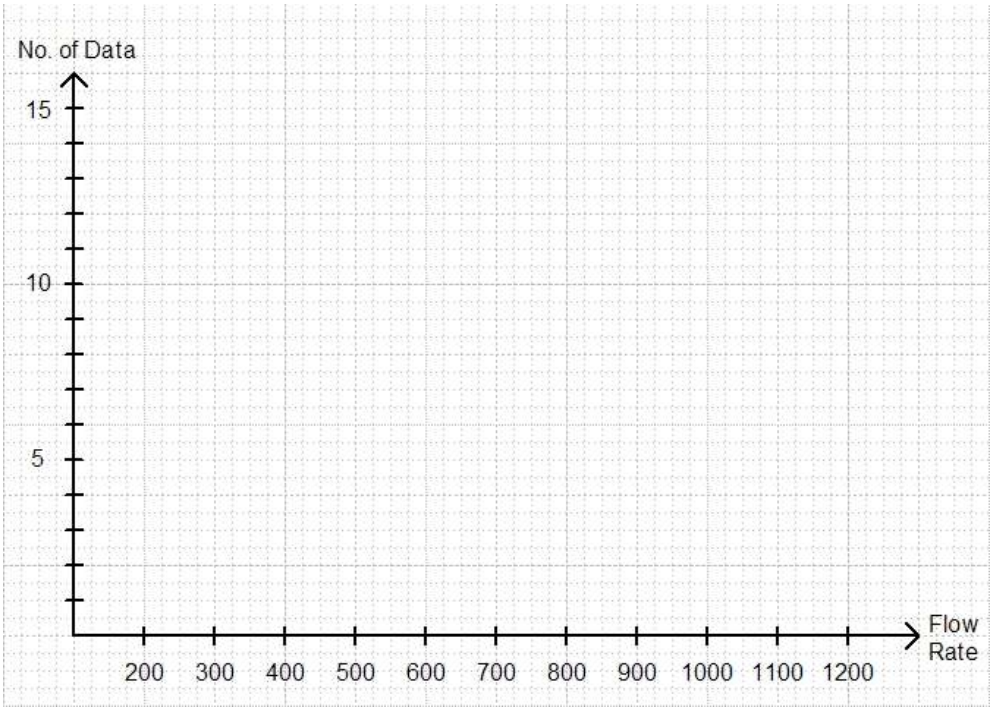
228	260	264	273	273	284	295	325	336	351
359	370	379	385	391	400	405	425	428	434
449	454	464	466	477	499	509	516	517	541
548	553	556	560	567	584	588	591	598	598
607	639	652	662	675	706	753	839	895	1120

(**) Complete the following table.

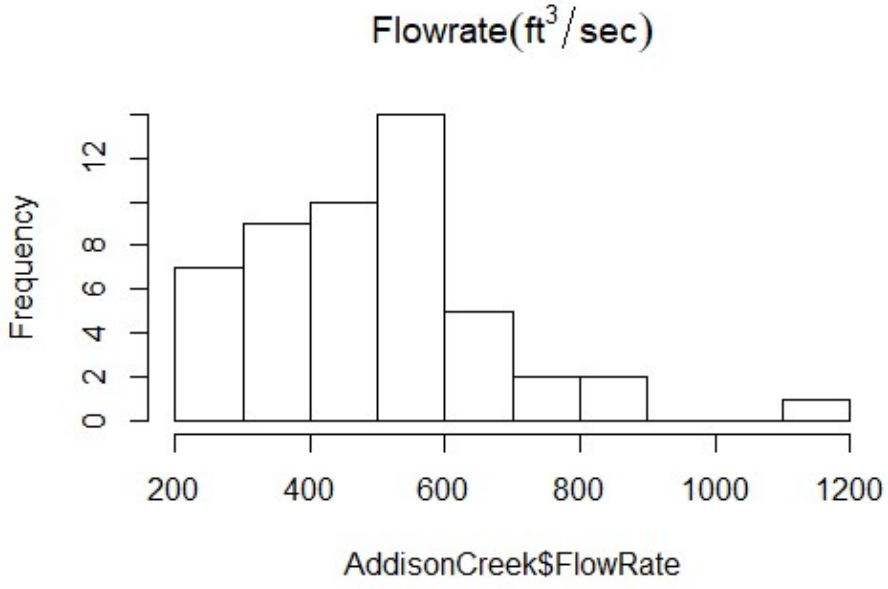
Range, $a < x \le b$	① No. of Occurrences	② Frequency =①/(Total No. of Observations)	③ Normalized Freq. =②/(Length of Bin)
$200 - 300$	7	0.14	0.0014
$300 - 400$			
$400 - 500$			
$500 - 600$			
$600 - 700$			
$700 - 800$			
$800 - 900$			
$900 - 1200$			

Instructor: Junho Song
junhosong@snu.ac.kr

(a) **Histogram**: the number of occurrences that fall within each bin.

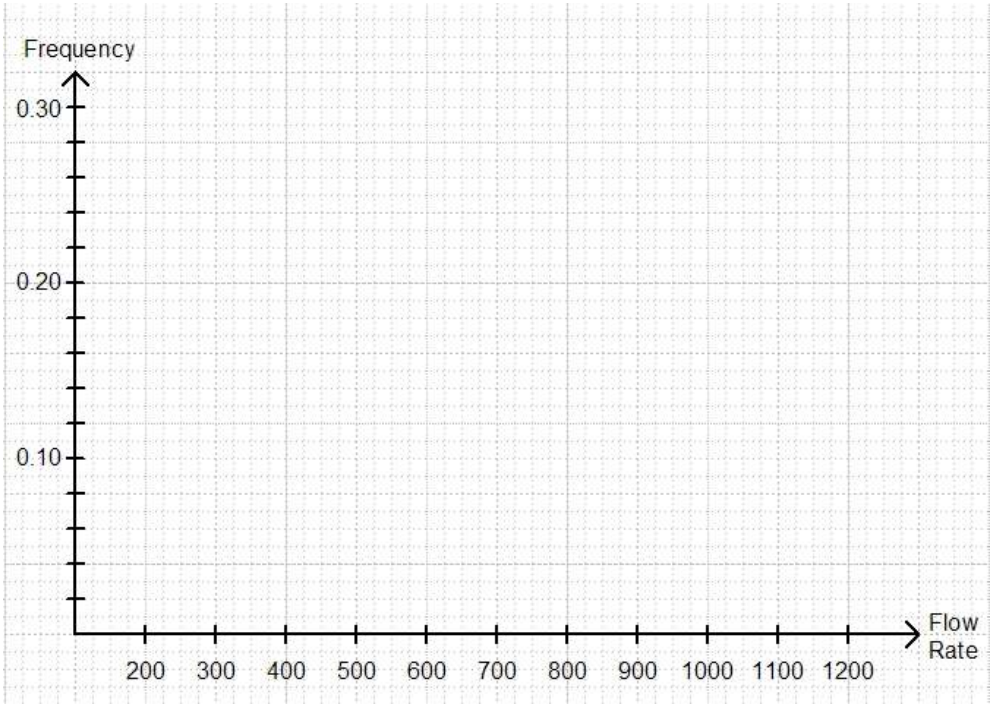


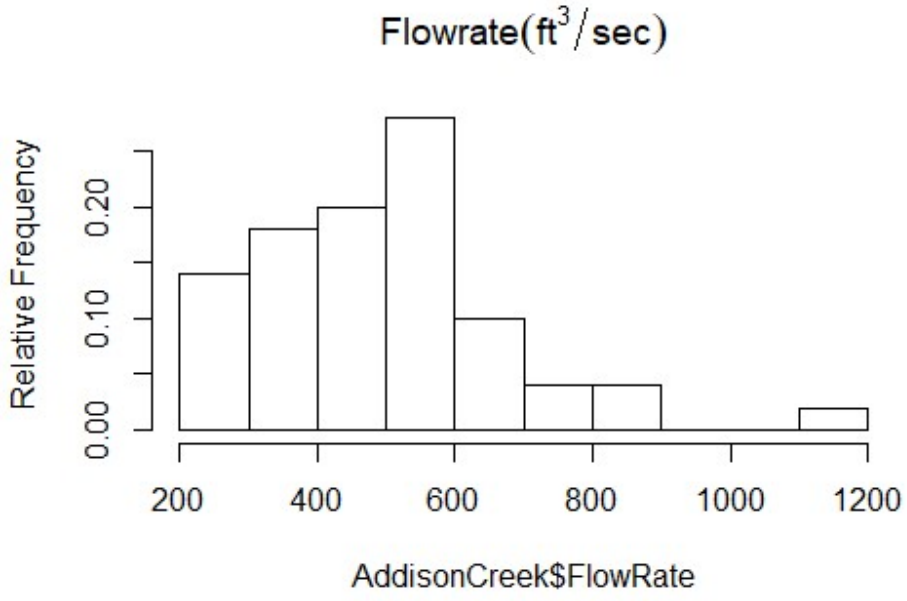

```
AddisonCreek = read.table("AddisonCreek.txt", header=TRUE)
# header: column names are read form the file

hist(AddisonCreek$FlowRate, main=expression(Flowrate(ft^3/sec)))
# main: name of graph
# expression: want to write mathematical symbols and expressions
# compare it with main = "Flowrate(ft3/sec)"
```



Instructor: Junho Song
junhosong@snu.ac.kr

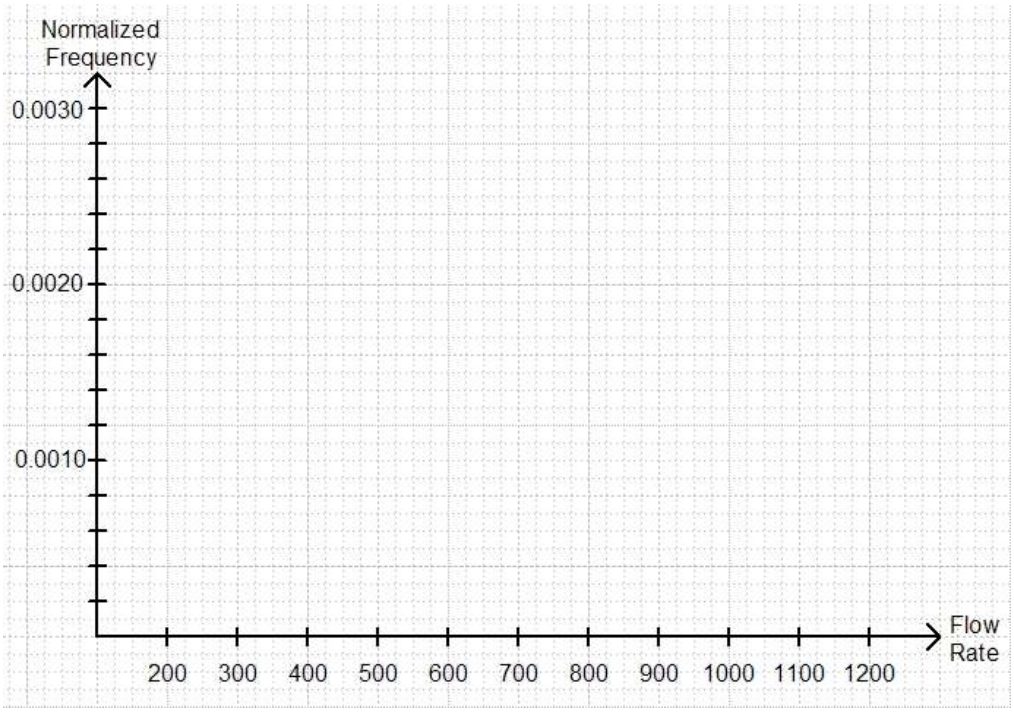
(b) **Frequency diagram**: the <u>percentage</u> (frequency) of occurrences that fall within each bin. → (# of occurrences) / (# of total observations)



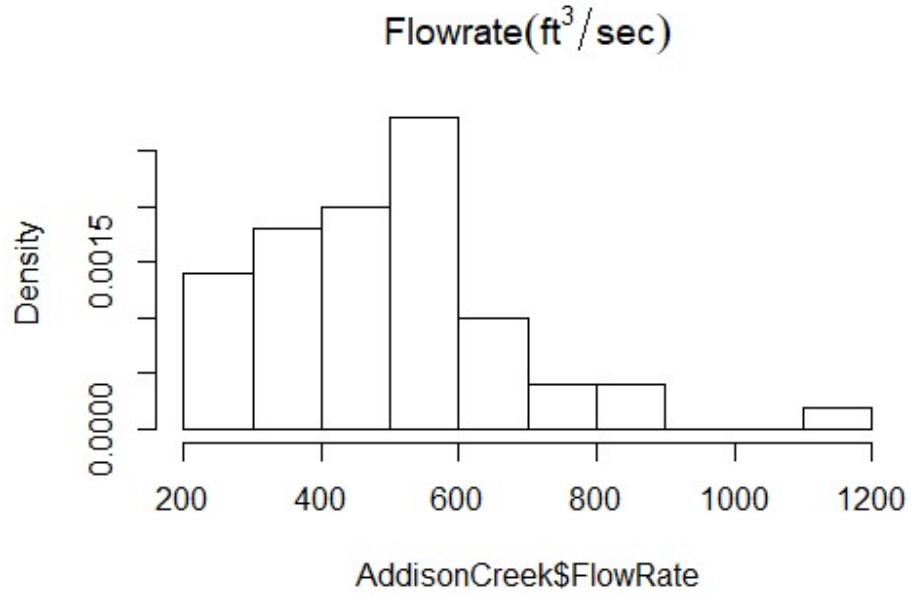


Instructor: Junho Song
junhosong@snu.ac.kr

(c) **Normalized frequency diagram** & **normalized polygon**: the percentage of observations that fall within each bin per unit value of the variable.
→ (# of occurrences) / (# of total observations) / (length of each bin)

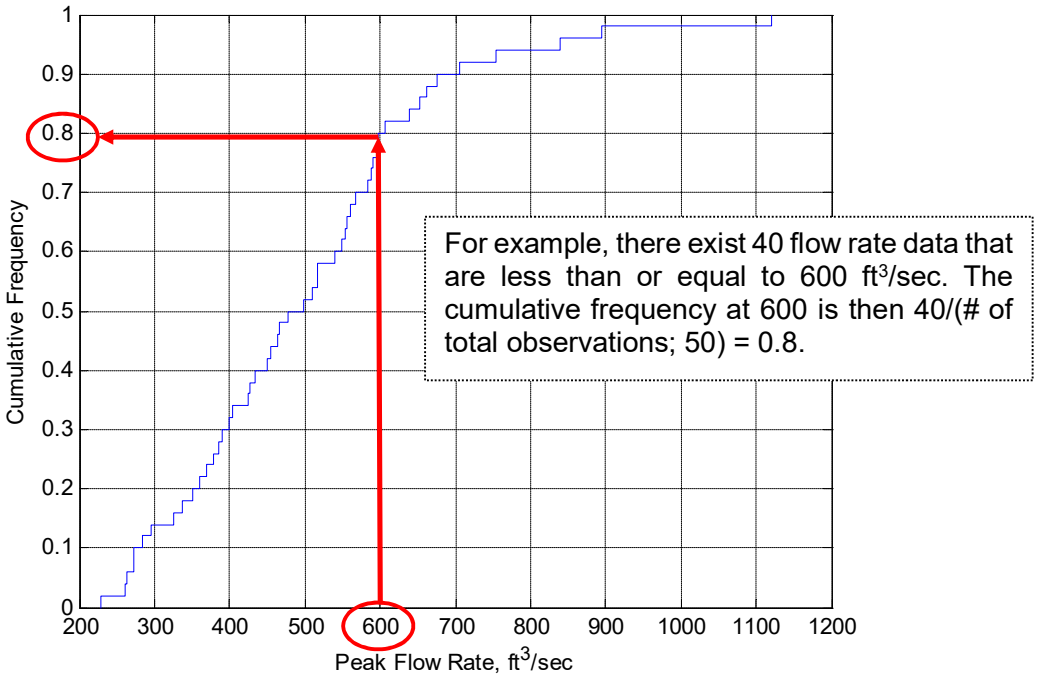



```
hist(AddisonCreek$FlowRate, freq=FALSE,
main=expression(Flowrate(ft^3/sec)))
```

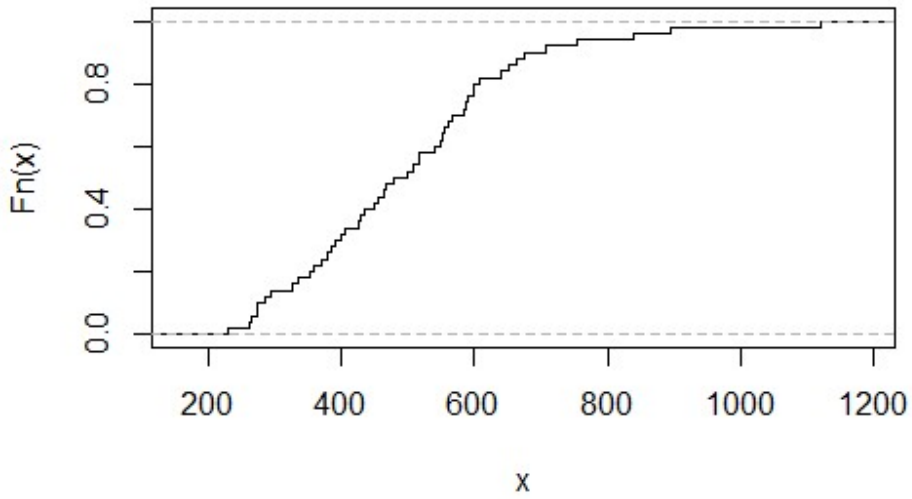


Instructor: Junho Song
junhosong@snu.ac.kr

(d) **Cumulative frequency**: the percentage of the data which are smaller than or equal to a given value.

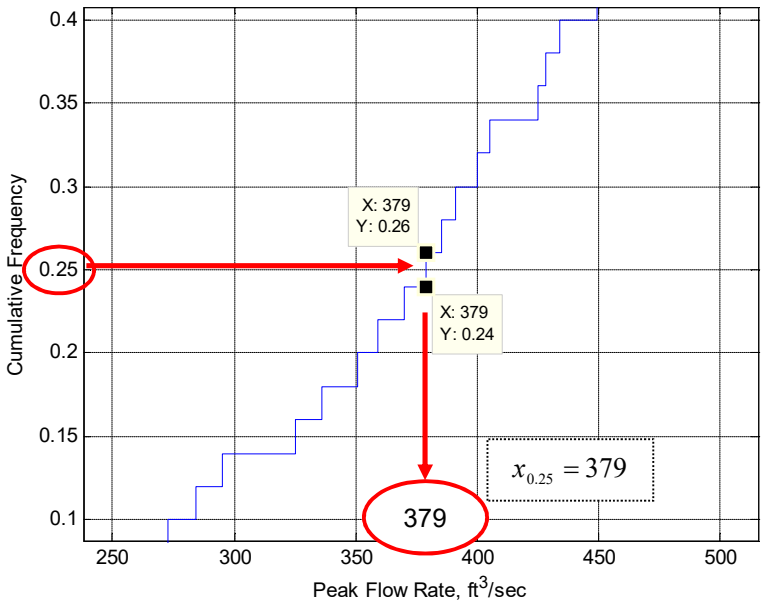


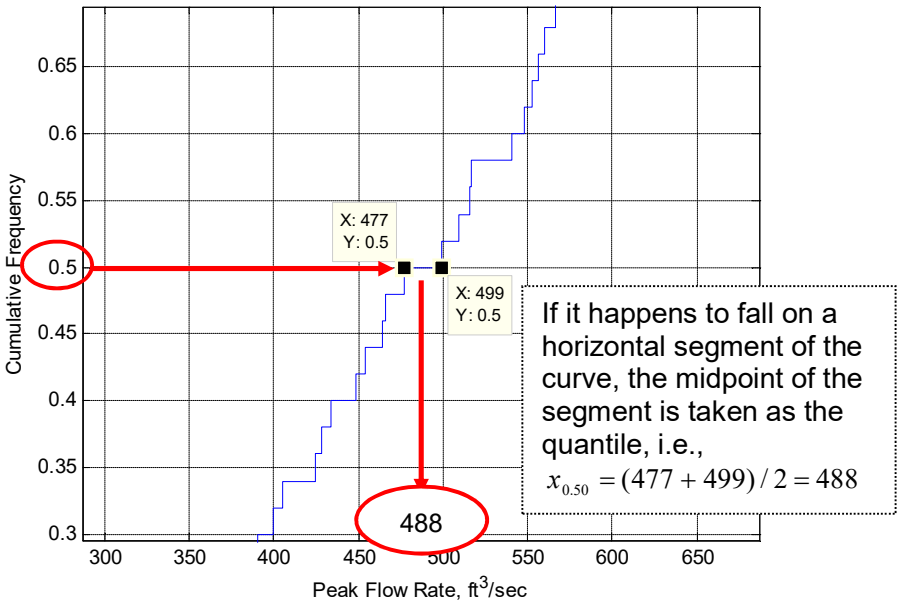

```
plot(ecdf(AddisonCreek$FlowRate), verticals=TRUE, pch="", main="")
# verticals: draw vertical lines at steps
# pch: plotting character
```



Instructor: Junho Song
junhosong@snu.ac.kr

(e) **Quantile**: threshold value for specified frequency. x_q (q-quantile or qx100-percentile) is the value of the variable for which the cumulative frequency is q , i.e., the inverse solution of the cumulative frequency.



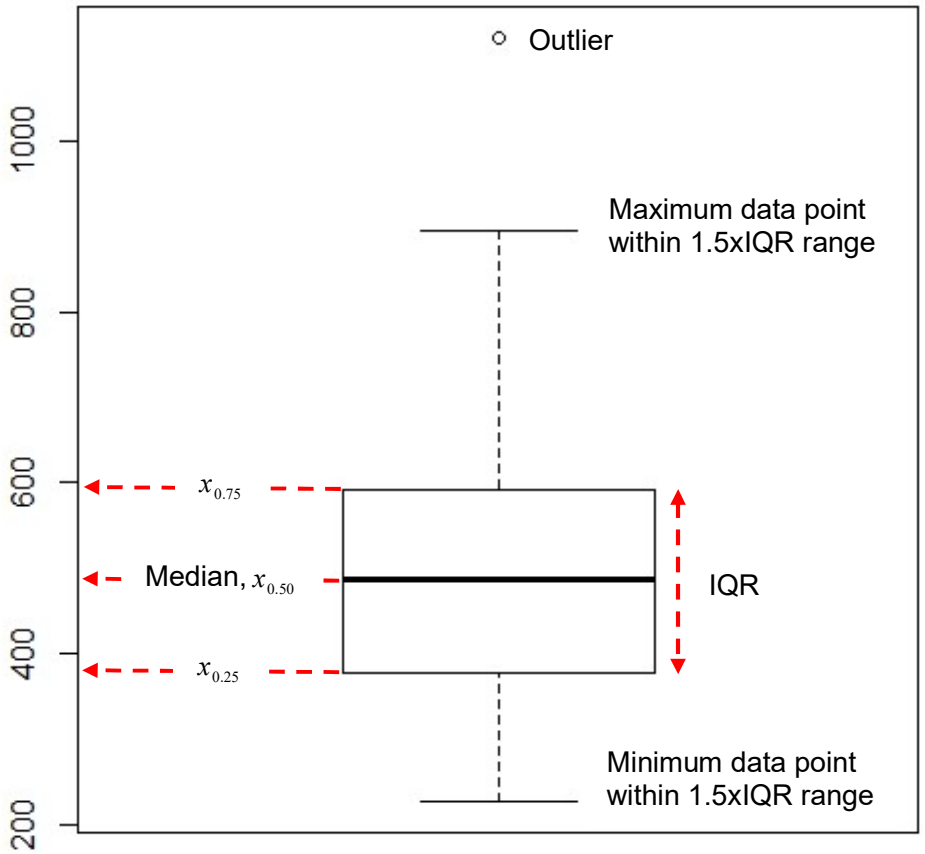


- $x_{0.05}, x_{0.10}, x_{0.90}, x_{0.95}$: quantiles often used in statistical analysis.
- $x_{0.25}$: first quartile.
- $x_{0.50}$: second quartile, median.
- $x_{0.75}$: third quartile.
- *IQR*, $x_{0.75} - x_{0.25}$: inter-quartile range. A measure of dispersion.

Instructor: Junho Song
junhosong@snu.ac.kr

```
quantile(AddisonCreek$FlowRate, 0.05, type=2)

quantile(AddisonCreek$FlowRate, type=2) # quartiles

IQR(AddisonCreek$FlowRate, type=2)
# type 2: Inverse of empirical distribution function, but with
averaging at discontinuities
```

(f) **Box Plot**: a graphical representation of a data set using quartiles and IQR.




```
boxplot(AddisonCreek$FlowRate)
```

Seoul National University
Dept. of Civil and Environmental Engineering

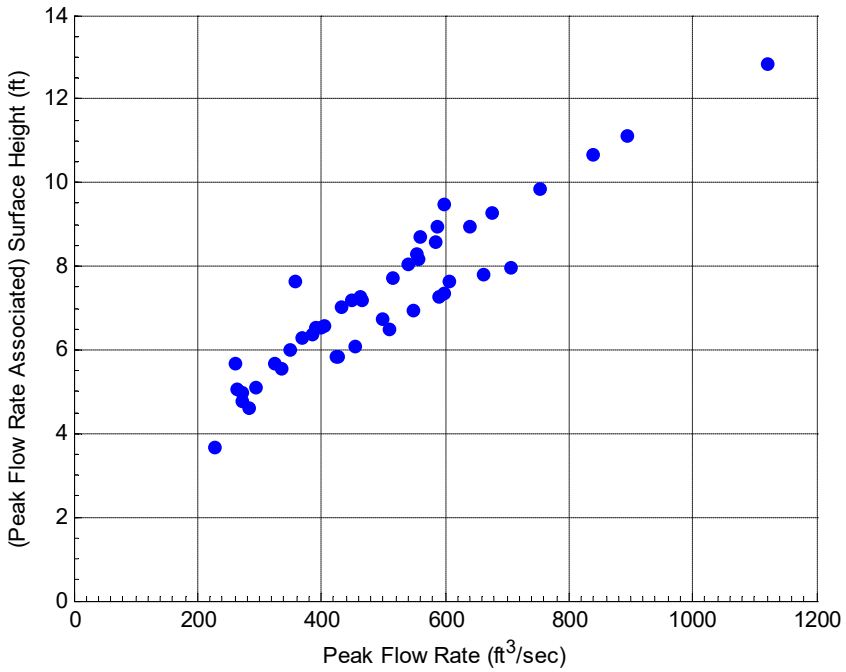
Instructor: Junho Song
junhosong@snu.ac.kr

2. The pairs of annual peak stream flow (ft^3/sec) and surface height associated with the peak stream (ft) of Addison Creek are listed in the following table for the period 1955-2004.

Year	Flow Rate	Height	Year	Flow Rate	Height	Year	Flow Rate	Height
1955	598	9.48	1972	706	7.97	1989	541	8.03
1956	228	3.65	1973	509	6.49	1990	895	11.13
1957	560	8.69	1974	428	5.82	1991	675	9.27
1958	405	6.58	1975	598	7.33	1992	351	5.99
1959	273	4.76	1976	454	6.06	1993	517	7.72
1960	295	5.08	1977	359	7.65	1994	425	5.84
1961	588	8.93	1978	607	7.63	1995	273	4.96
1962	260	5.68	1979	591	7.28	1996	400	6.52
1963	325	5.68	1980	662	7.80	1997	753	9.86
1964	284	4.59	1981	548	6.95	1998	466	7.20
1965	434	7.02	1982	839	10.68	1999	385	6.35
1966	499	6.74	1983	639	8.94	2000	477	7.31
1967	264	5.05	1984	464	7.27	2001	379	6.27
1968	584	8.59	1985	556	8.17	2002	652	8.97
1969	553	8.28	1986	391	6.52	2003	567	8.18
1970	449	7.19	1987	1120	12.84	2004	516	7.69
1971	336	5.56	1988	370	6.30			

* Source: National Water Information System Web Site (NWISWeb); http://waterdata.usgs.gov/nwis

(g) **Scatter Diagram**: plots data pairs to identify relationship or dependence between two sets of data. Mark the last five pairs of the table on the scatter diagram below.




```
flowrate = AddisonCreek[,3]
height = AddisonCreek[,4]

plot(flowrate, height, pch=20)
```