

$$\begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ 0 & & & \sigma_n \end{bmatrix}$$

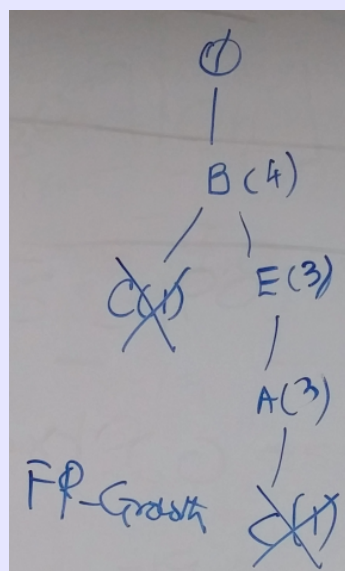
$$X = \sum_{i=1}^{\text{rank}(X)} \sigma_i u_i v_i^T = U \Sigma V^T$$

σ_i : i^{th} singular value of X
 u_i : i^{th} left singular value of X (i^{th} column of U)
 v_i : i^{th} right singular vector of X (i^{th} column of V^T)

Captures the patterns among attributes
 Captures the patterns among the objects

CS 422: Data Mining
 Vijay K. Gurbani, Ph.D.,
 Illinois Institute of Technology

Lecture 2: Random variables, measures of central tendency and distributions

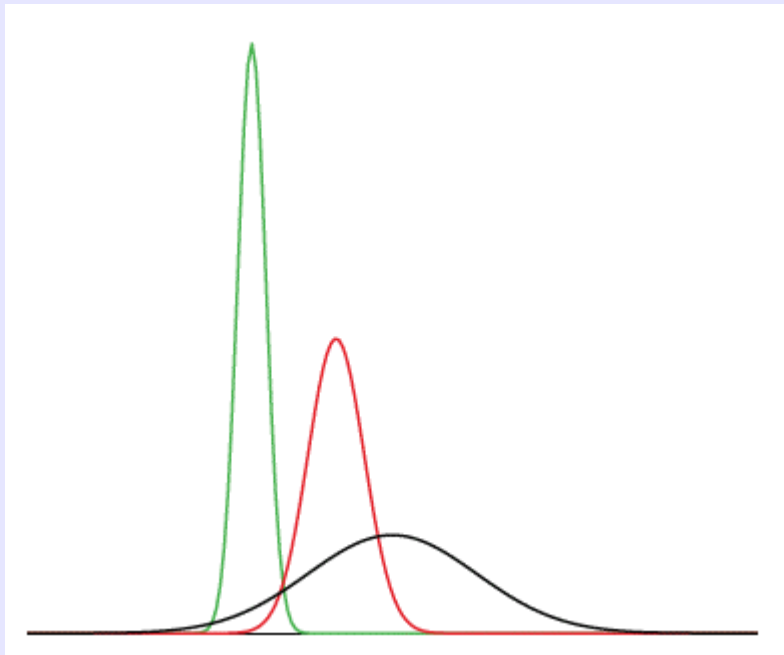


CS 422
 vgurbani@iit.edu



Distributions

- Normal / Gaussian distribution



Green: $\mu=-3.0$, $\sigma=0.5$

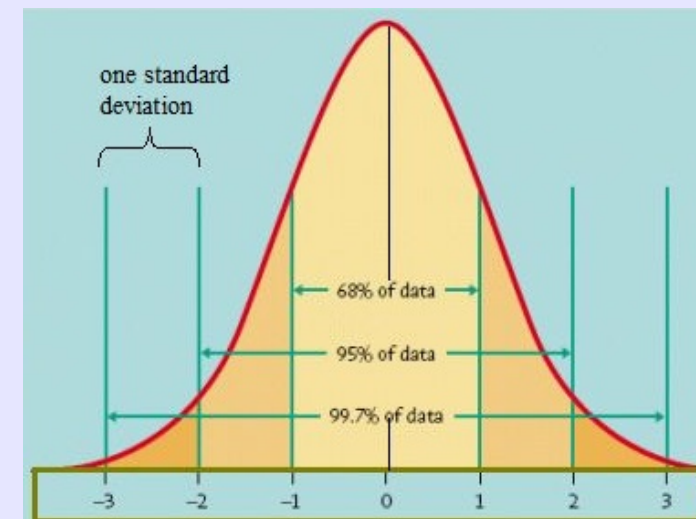
Red: $\mu=0.0$, $\sigma=1.0$

Black: $\mu=2.0$, $\sigma=3.0$

- Parameterized by mean (μ) and standard deviation (σ)

- PDF:
$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

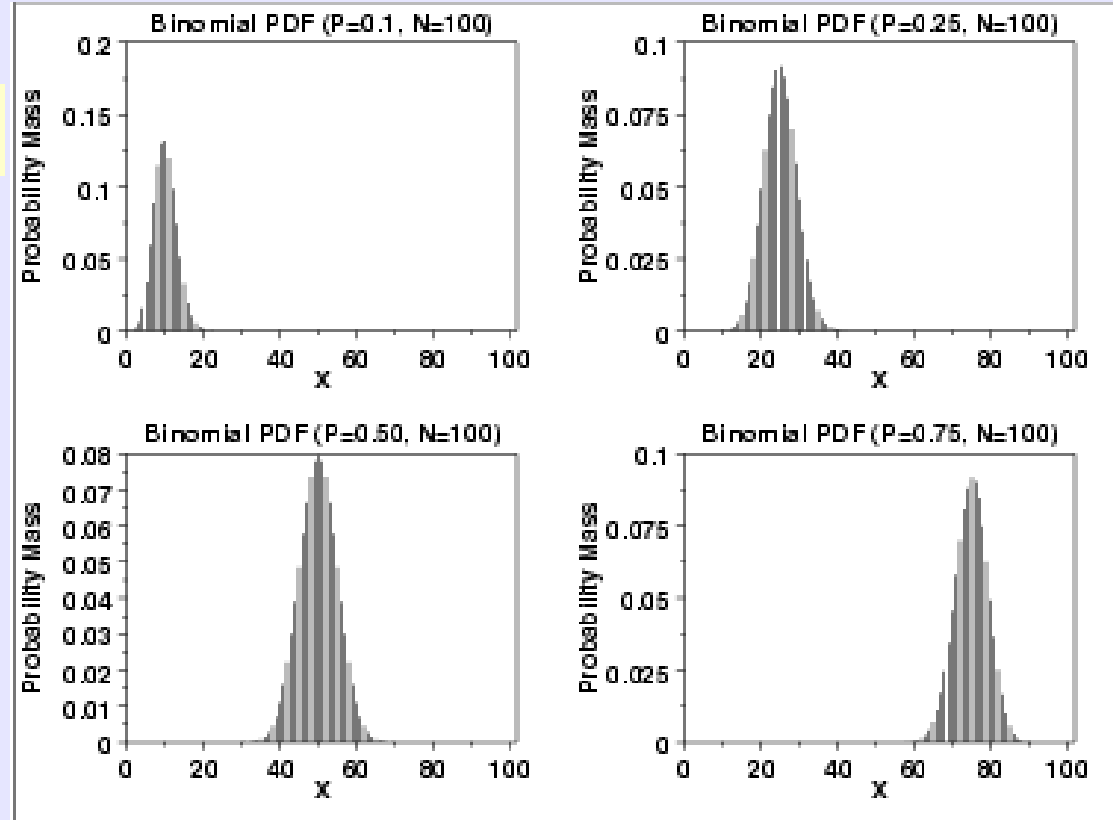
- mean = median = mode



Distributions

- Binomial distribution: parameterized by n (number of trials) and p (probability of success in each trial)
- PMF: $\binom{n}{x} (p)^x (1-p)^{(n-x)}$ for $x = 0, 1, 2, \dots, n$
for observing x successes in n trials.
- Mean: np
- Median: $\lfloor np \rfloor$ or $\lceil np \rceil$
- Variance: $np(1-p)$

Slide source: <http://www.itl.nist.gov/div898/handbook/eda/section3/eda366i.htm>



Distributions

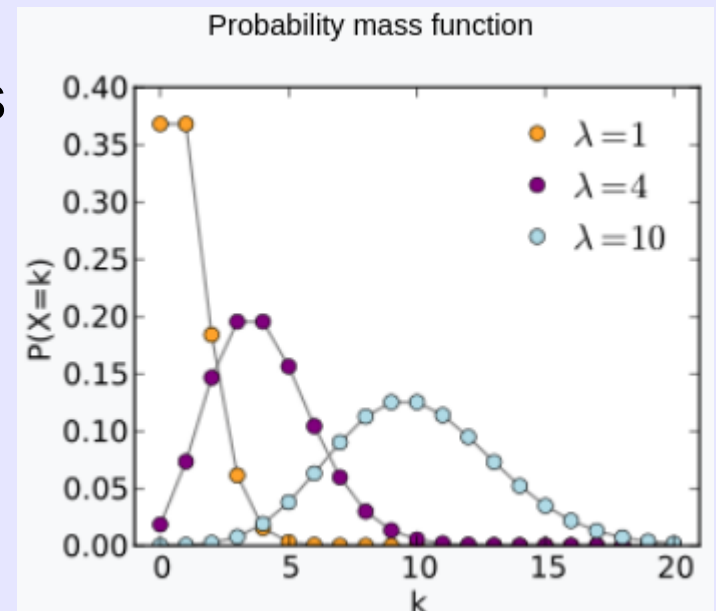
- Poisson distribution: expresses the probability of a given number of events (k) occurring in a fixed interval of time if these events occur with a known constant mean rate (λ) and independently of the time since the last event.

- PMF: $\frac{\lambda^k e^{-\lambda}}{k!}$

- Mean: λ

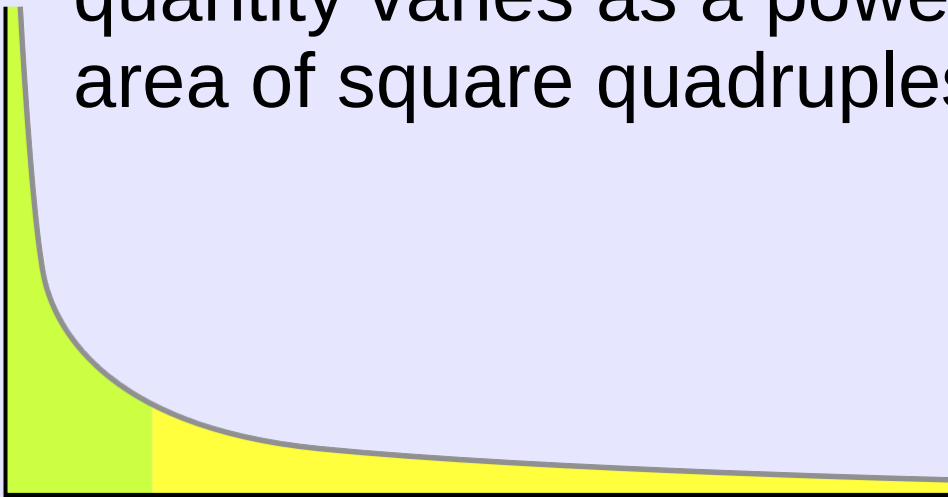
- Median: $\lambda - \ln 2 \leq \nu < \lambda + \frac{1}{3}$.

- Variance: λ



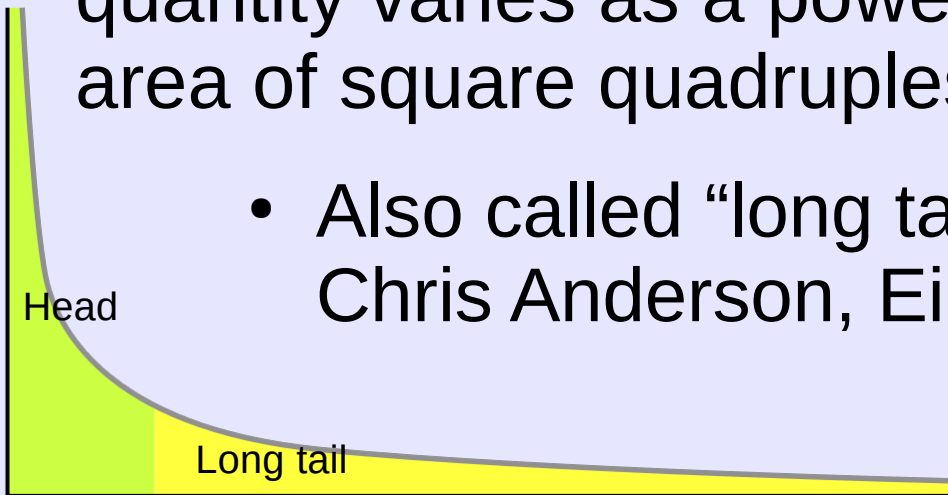
Distributions

- Power-law distributions: relationships where one quantity varies as a power of another. (Example: area of square quadruples when length is doubled.)



Distributions

- Power-law distributions: relationships where one quantity varies as a power of another. (Example: area of square quadruples when length is doubled.)
 - Also called “long tailed” distributions (coined by Chris Anderson, EiC of *Wired*), or the 80-20 rule.
- Unlike other distributions, the moments of Power Law are hard to define: under certain conditions the first moment is defined, but rest are infinite!

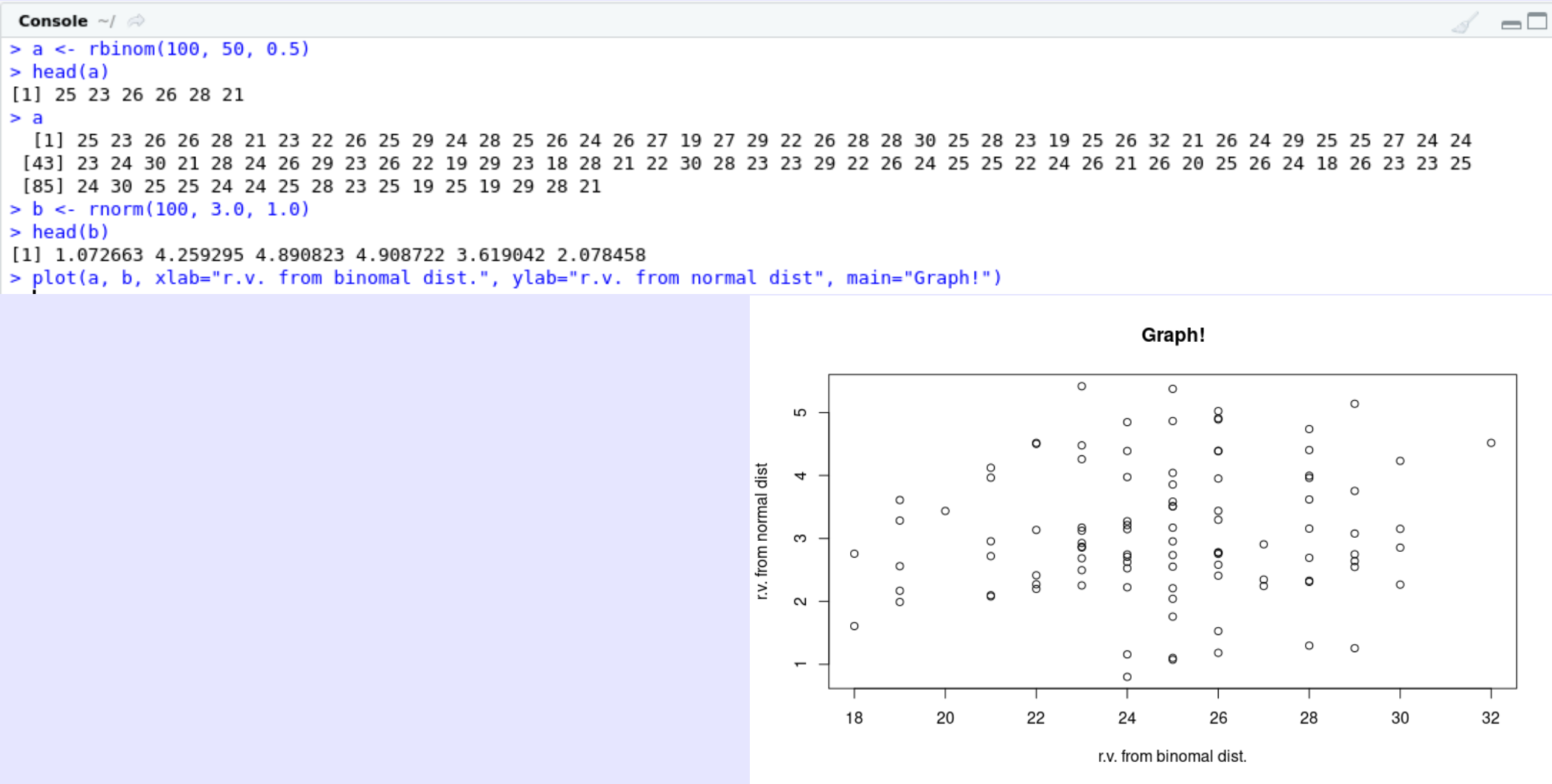


Graphs

- Visualization is important!
- Excellent resource: Claus O. Wilke, *Fundamentals of Data Visualization*, O'Reilly Publishing. (Uses R/RStudio/R Markdown.)
- No time to look at details into graphs, but:
 - Boxplots
 - CDFs
 - Normal x-y plots

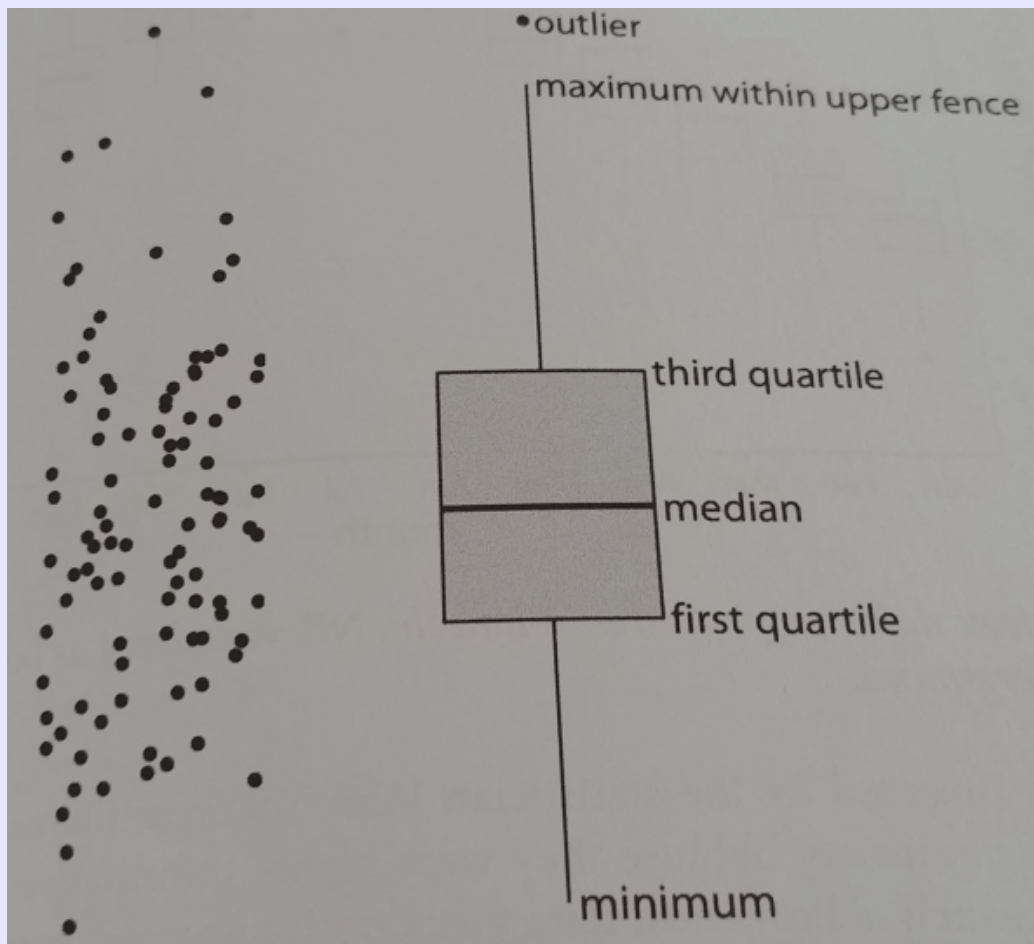
Graphs

- x-y plots

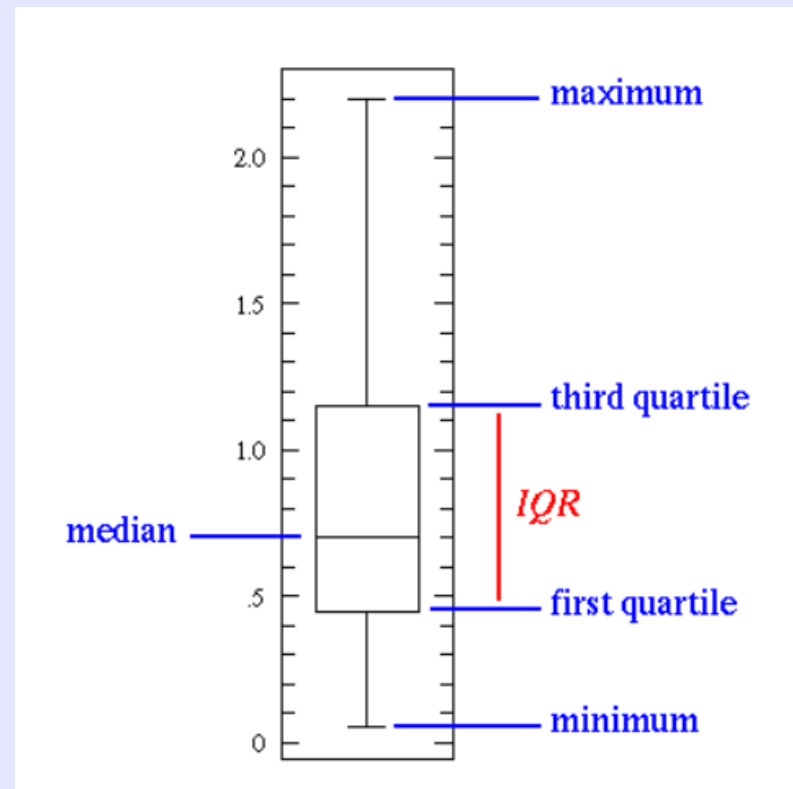


Graphs

- Boxplots




Graphic source: Fundamentals of Data Visualization, Claus O. Wilke, O'Reilly Publishing, 2019.

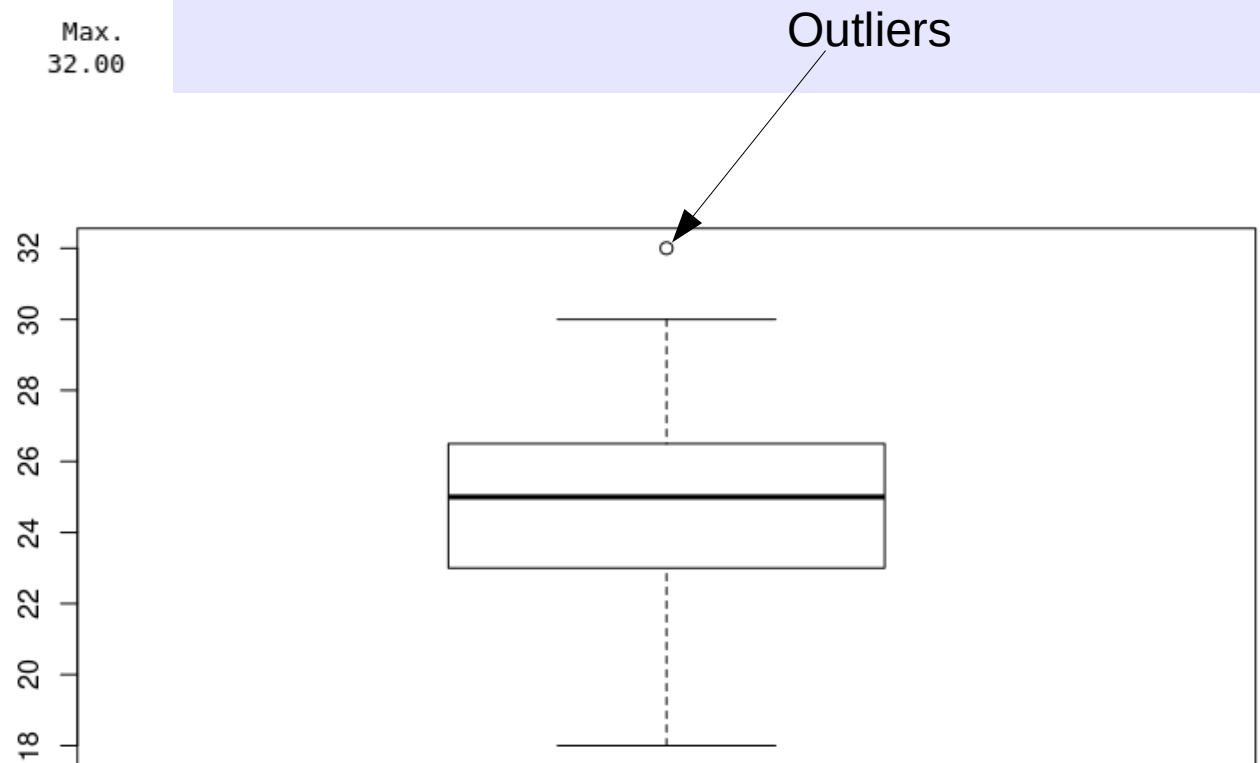


Graphs

- Boxplots

```
Console ~/   
> summary(a)  
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   
  18.00   23.00   25.00   24.80   26.25   32.00   
> boxplot(a)  
> |
```

Variability outside
the upper and lower
quartiles are the
"whiskers"




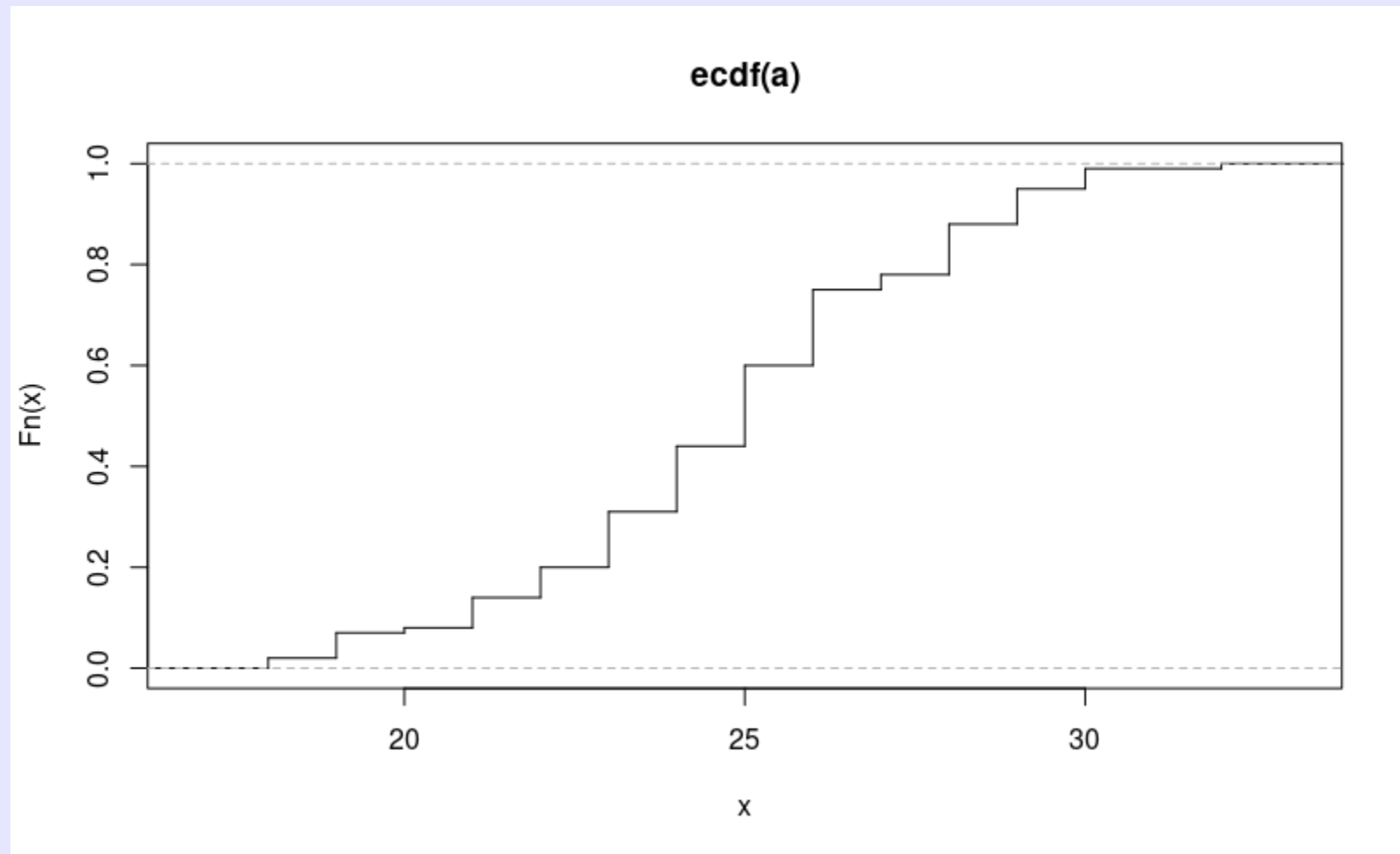
Graphs

- Empirical cumulative distribution function
 - CDF(y) of a dataset X at a value y is the ratio of samples that are **lower** than the value y :
$$\text{CDF}(X, y) = \frac{|\{i : i \leq y\}|}{|X|}$$
 - The derivative of the CDF is the PDF.
 - Example: $X = [2, 7, 8, 9, 10, 15, 16, 20]$
 - $\text{CDF}(X, 15) = 6/8 = 0.75$

Graphs

- ECDF

```
Console ~/   
> plot(ecdf(a), verticals = T, do.points=F)  
> |
```



A note on debugging

- When you run into problems, learn how to debug your code.
 - See
 - <https://support.rstudio.com/hc/en-us/articles/205612627-Debugging-with-RStudio>
 - <https://adv-r.hadley.nz/debugging.html>
- Sometimes, even after debugging, things are not clear. In such cases:
 - Isolate the problem as best as you can.
 - Reproduce the problem on a small dataset.
 - And get in touch with the TA or me.
- Remember: The better you can describe your problem to me or the TA, the quicker it is for us to help you.
- Debugging is an art! Become proficient at it.

```
> df <- read.csv(...)
> index <- sample(1:nrow(df), 0.20*nrow(df))
> small.df <- df[index, ]
> rm(index)
> dim(df)
[1] 11034    31
> dim(small.df)
[1] 2206    31
```