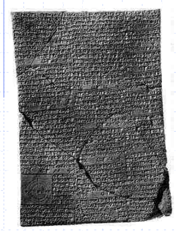
Statistics

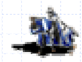
Statistics deals with the collection and analysis of data to solve real-world problems.

History

Ancient Babylonians & Egyptians

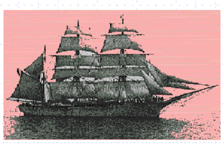


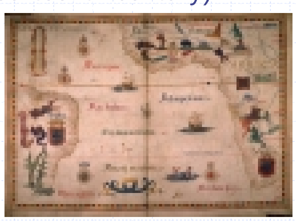
$



History

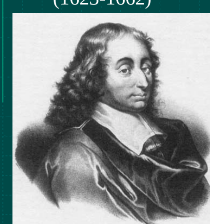
Marine insurance rates determined by past statistical data (fourteenth century)





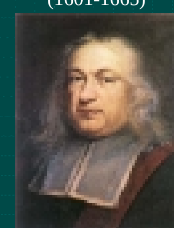
People

Blasé Pascal
(1623-1662)



Pr(E) = ?

Pierre de Fermat
(1601-1665)



People

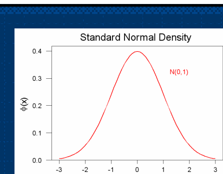
Jacob Bernoulli
(1654-1705)

$\text{relative frequency} \longrightarrow \text{probability}$

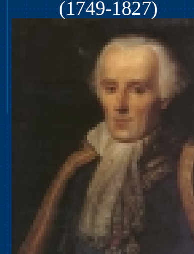
Law of large numbers

$$\frac{1}{n}(X_1 + X_2 + \cdots + X_n) \to E(X)$$

People



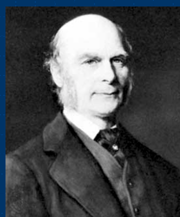
Pierre Simon Laplace
(1749-1827)

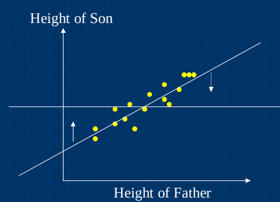


Karl Friedrich
Gauss
(1777-1855)

People

Sir Francis Galton
(1822 – 1911)





Height of the sons of fathers regressed towards the mean height of the population

Regression Analysis

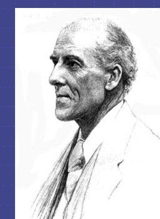
Founders of modern statistics

Ronald Fisher
(1892-1962)

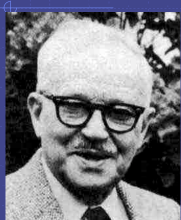
Study
- Biology
- Evolution
- Genetics

Develop
- MLE
- ANOVA
- Correlation
- Chi-square test

Karl Pearson
(1857-1936)



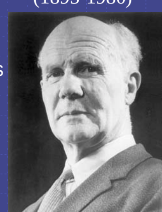
Founders of modern statistics



Jerzy Neyman
(1894-1981)

Develop
- Testing Hypothesis
- Confidence Intervals

Egon Pearson
(1895-1980)



Era of information

Huge datasets

Demographics

Transaction data

Customer behavior

Human genome

Satellite photographs

....

....

Data mining   Data reduction

Multidisciplinary
statistics methodology

Elements of Statistics

- Statistics
  - Data Collection
    - Experimental Design
    - Survey Sampling
    - Observational Study
  - Data Analysis
    - Descriptive Statistics And Statistical Graphics
    - Statistical Inference
      - Estimation
      - Testing Hypothesis

Descriptive Statistics

Data: a set of numbers representing characteristics of observations

Year	Tutorial	Quiz	Midterm	Final	Overall	Grade
3	B	90.50	96.0	100	96.90	A+
3	B	96.50	99.0	95	96.50	A+
3	A	90.00	95.0	100	96.50	A+
1	A	94.00	99.0	96	96.50	A+
*	C	97.00	98.0	94	95.80	A+
2	B	88.50	97.0	96	94.80	A
2	C	86.50	100.0	94	94.30	A+
3	B	90.75	95.0	95	94.15	A+
1	A	93.00	96.0	93	93.90	A
3	B	94.00	92.0	94	93.40	A
		....	....	....		
		....	....	....		
2	A	53.00	49.5	42	46.45	D
3	A	69.50	35.0	44	46.40	D
1	B	59.00	35.0	47	45.80	D
1	B	48.00	45.0	44	45.10	D
3	B	51.50	65.0	27	43.30	D
3	B	41.00	31.0	43	39.00	F
3	A	56.00	19.0	38	35.90	F
1	B	47.50	32.0	27	32.60	F
1	B	23.00	29.0	27	26.80	F
1	A	39.50	20.0	18	22.90	F

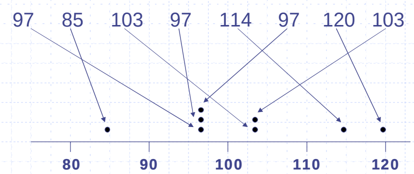
Descriptive Statistics

Scales of Measurement

Scale	Examples	Operations
Nominal Scale	Male/Female Chinese/British/American/...	+ - × ÷ (crossed out)
Ordinal Scale	disagree/agree/strongly agree low/medium/high	+ - × ÷ (crossed out)
Interval Scale	°C, °F, altitude of a place	+ - × ÷ (× ÷ crossed out)
Ratio Scale	length, weight	+ - × ÷

Statistical Graphics

Dot diagram



Frequency Distribution

Table: Flow of vehicles

Vehicles	Frequency	Percentage
Cars	45	59
Lorries	22	29
Motorcycles	6	8
Buses	3	4
Total	76	100

Table: Grade of Students

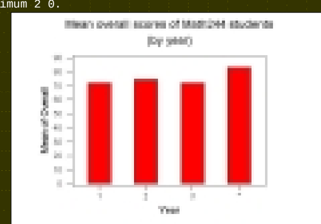
```
MTB > Tally  'Code-Grade';
SUBC>   Counts;
SUBC>   CumCounts;
SUBC>   Percents;
SUBC>   CumPercents.

Summary Statistics for Discrete Variables

Grade Grade_n Count CumCnt Percent CumPct
  A+        1     7      7    1.39   1.39
  A         2    26     33    5.17   6.56
  A-        3    57     90   11.33  17.89
  B+        4    92    182   18.29  36.18
  B         5    89    271   17.69  53.88
  B-        6    76    347   15.11  68.99
  C+        7    58    405   11.53  80.52
  C         8    36    441    7.16  87.67
  C-        9    23    464    4.57  92.25
  D        10    34    498    6.76  99.01
  F        11     5    503    0.99 100.00
          N=    503
```

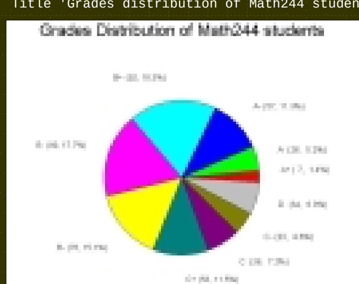
Bar Chart

```
MTB > Chart Mean( Overall ) * 'Year';
SUBC>   Bar;
SUBC>     Type 1;
SUBC>     EColor 1;
SUBC>     Color 2;
SUBC>   Title "Mean overall scores of Math244 students";
SUBC>   Title "(by year)";
SUBC>   Minimum 2 0.
```



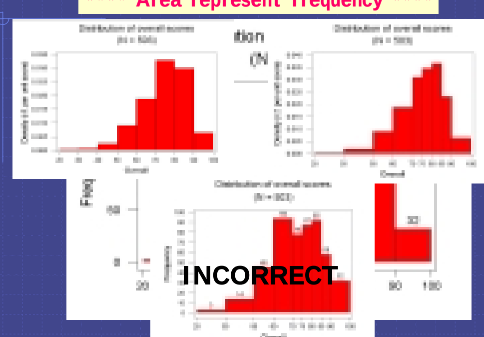
Pie Chart

```
MTB > %Pie 'Grade';
SUBC>   Title 'Grades distribution of Math244 students'.
```



Histogram

****** Area represent frequency ******



Percentile

"Joining Mensa Society requires scoring in the 98th percentile in one of the standard intelligence test."

- BMI less than 5th percentile: Underweight*
- BMI greater than 85th percentile: At risk of Overweight*
- BMI greater than 95th percentile: Overweight*
- BMI between 5th and 85th percentile: Healthy Weight*

Dataset of *n* observations

approximately *np* observations less than or equal to (100*p*)th percentile

Percentile from raw data

68, 75, 58, 47, 83, 34, 90, 71, 63, 79

↓ sorted

34 47 58 63 68 71 75 79 83 90

$$X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$$

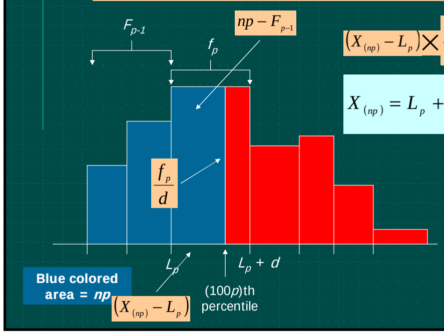
$$(n+1)p = r.f$$

integer fractional

$$(100p)\text{th percentile} = X_{(r)} + f\left(X_{(r+1)} - X_{(r)}\right)$$

Percentile from histogram

area on the left of (100*p*)th percentile = *np*



$$\left(X_{(np)} - L_p\right) \times \frac{f_p}{d} = np - F_{p-1}$$

$$X_{(np)} = L_p + \frac{d\left(np - F_{p-1}\right)}{f_p}$$

Percentile

Table: Frequency table for 20 grain bullet penetration depths into oak wood from a distance of 15 feet

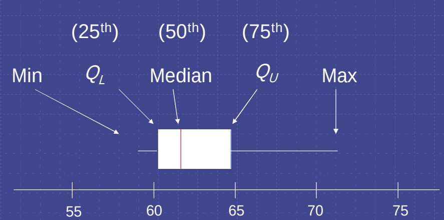
Penetration Depth (mm)	Frequency	Cumulative Frequency
58 – 60	5	5
60 – 62	3	8
62 – 64	6	14
64 – 66	3	17
66 – 68	1	18
68 – 70	0	18
70 – 72	2	20
Total	20	

$p = 0.75$, $np = 15$ $L_p = 64$ $F_{p-1} = 14$ $f_p = 3$, $d = 2$

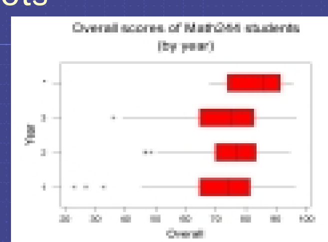
$$75\text{th percentile} = 64 + \frac{2(15-14)}{3} = 64.67$$

Boxplot

Five number summary

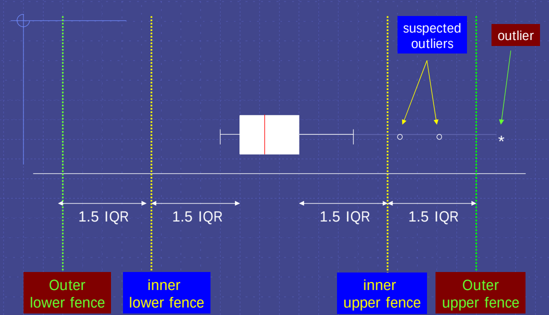


Boxplots




```
MTB > Boxplot 'Overall'*'Year';
SUBC>   Transpose;
SUBC>   Box;
SUBC>     Type 1;
SUBC>     Color 2;
SUBC>   Symbol;
SUBC>     Outlier;
SUBC>   Title "Overall scores of Math244 students";
SUBC>   Title "(by year)".
```

Detecting Outliers



Stem-and-Leaf Display

78 65 90 86 79 51 79 62 84 101

5 *	1
6	2 5
7	8 9 9
8	4 6
9	0
10 *	1

N = 10
Leaf unit = 1

stem
leaf

Stem-and-Leaf Display

3 *	8
4	0 0 1 4 5 5 6 7 7 9
5 *	0 1 3 8

N = 15
Leaf unit = 1

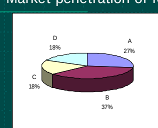
3 +	8
4 *	0 0 1 4
4 +	5 5 6 7 7 9
5 *	0 1 3
5 +	8

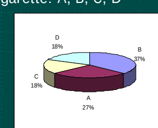
N = 15
Leaf unit = 1

* for 0 – 4
\+ for 5 – 9

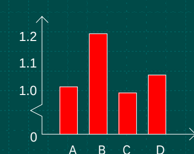
Lying with Statistical Graphics

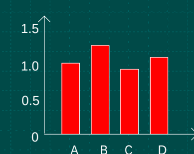
Market penetration of four brands of cigarette: A, B, C, D



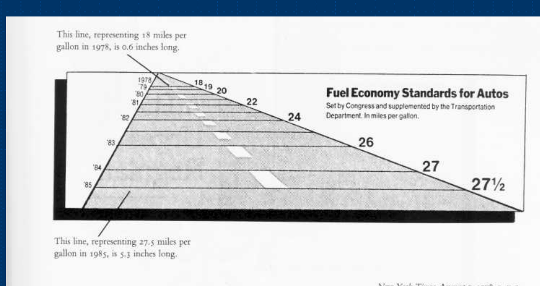


Monthly sales of the four brands (in $m)

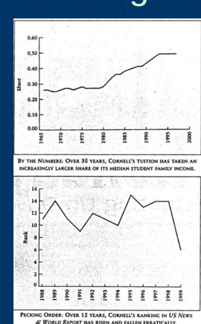




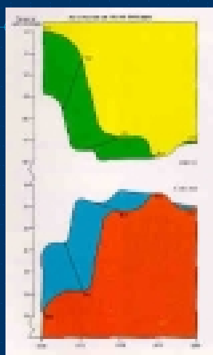
Bad Graphic Designs



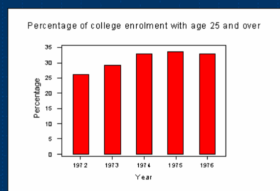
Bad Graphic Designs



Bad Graphic Designs



Percentage of college Enrolment with age 25 and over



Measures of Central Location

mean / average / arithmetic mean

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$ (raw data)

$$\bar{X} = \frac{f_1X_1 + f_2X_2 + \cdots + f_nX_n}{f_1 + f_2 + \cdots + f_n}$$ (ungrouped frequency table)

$$\bar{X} = \frac{f_1m_1 + f_2m_2 + \cdots + f_nm_n}{f_1 + f_2 + \cdots + f_n}$$ (grouped frequency table)

m_i = midpoints of i^{th} class

Measure of Central Location

median / 50th percentile / second quartile

$$M = \begin{cases} X_{\left(\frac{n+1}{2}\right)} & n \text{ odd} \\ \frac{1}{2}\left(X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}\right) & n \text{ even} \end{cases}$$

X_i : 10 12 7 300 15 $\quad \bar{X} = 68.8$

$X_{(i)}$: 7 10 12 15 300 $\quad M = X_{(3)} = 12$

Measures of Variation

78	65	90	86	79	51	79	62	84	101

$\bar{X} = 77.5$ $\quad M = 79$ $\quad \text{mode} = 79$

$range = 50$ $\quad IQR = 22.75$ $\quad MAD = 10.9$ $\quad SD = 13.88$

$$SD = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2} = 13.88$$

rang I .25l = 50

78	45	110	106	79	21	79	32	84	141

$range = 120$ $\quad IQR = 65.25$ $\quad MAD = 26.9$ $\quad SD = 35.02$

Standard Deviation

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Why do we use (n-1) instead of n ?

n=1 $\Rightarrow$ $SD = 0$ $\quad s$ = undefined ← more reasonable

$$X_1 - \bar{X} = n\bar{X} - (X_2 + \ldots X_n) - \bar{X}$$
$$= -(X_2 - \bar{X}) - (X_3 - \bar{X}) - \cdots - (X_n - \bar{X})$$

only (n-1) pieces of information in $\sum_{i=1}^{n}(X_i - \bar{X})^2$

Skewness

skewed to right (+ve skewed)

skewed to left (-ve skewed)

$\gamma_1 > 0$ $\quad \gamma_2 > 0$ $\qquad \gamma_1 < 0$ $\quad \gamma_2 < 0$

$$\gamma_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^3}{(SD)^3}$$

$$\gamma_2 = \frac{mean - median}{SD}$$