Chebyshev's rule applies to any sample of measurements, regardless of the shape of the frequency distribution:

At least $1-\frac{1}{k^2}$ of the measurements will fall within k std deviations of the mean $(\bar{x}-ks, \bar{x}+ks)$ for any whole number $k>1$.

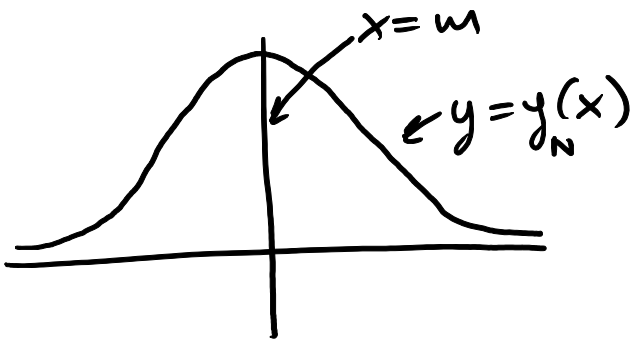
Emprical rule is a rule of thumb that applies to samples with freq distributions which are bell shaped (or "mound shaped" or "approximately normal").

68% of measurements in $(\bar{x}-s, x+\bar{s})$

95% of measurements in $(\bar{x}-2s, \bar{x}+2s)$

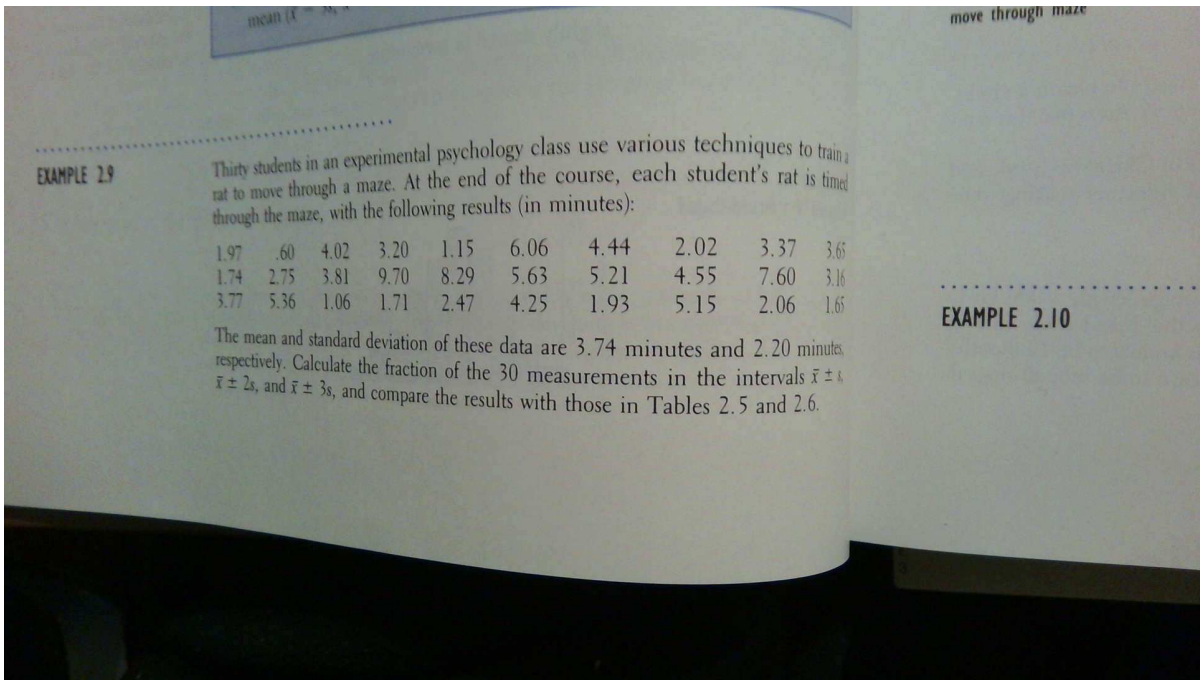
99.7% of measurements in $(\bar{x}-3s, \bar{x}+3s)$

Derived from integral of $y_N(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(x-w)^2}{2\sigma^2}\right]$



$$\int_{-\infty}^{+\infty} y_N(x)\,dx = 1$$

Consider an example with the data below:



EXAMPLE 2.9

Thirty students in an experimental psychology class use various techniques to train a rat to move through a maze. At the end of the course, each student's rat is timed through the maze, with the following results (in minutes):

1.97	.60	4.02	3.20	1.15	6.06	4.44	2.02	3.37	3.65
1.74	2.75	3.81	9.70	8.29	5.63	5.21	4.55	7.60	3.16
3.77	5.36	1.06	1.71	2.47	4.25	1.93	5.15	2.06	1.65

The mean and standard deviation of these data are 3.74 minutes and 2.20 minutes, respectively. Calculate the fraction of the 30 measurements in the intervals $\bar{x} \pm s$, $\bar{x} \pm 2s$, and $\bar{x} \pm 3s$, and compare the results with those in Tables 2.5 and 2.6.

EXAMPLE 2.10

Step 1: construct freq distribution and histogram.

min = 0.60

max = 9.70

Decide on class width say 1.5

class 1: [0.60, 2.09]

class 2: [2.10, 3.59]

class 3: [3.60, 5.09]

class 4 : [5.10, 6.59]

class 5 : [6.60, 8.09]

class 6 : [8.10, 9.59]

class 7 : [9.60, 11.09]

Notice that this necessarily covers all the data as data is given to two decimal figures.

use R to get counts

class	freq
class 1	10
class 2	5
class 3	7
class 4	5
class 6	1
class 7	1
class 8	1

construct class boundaries:

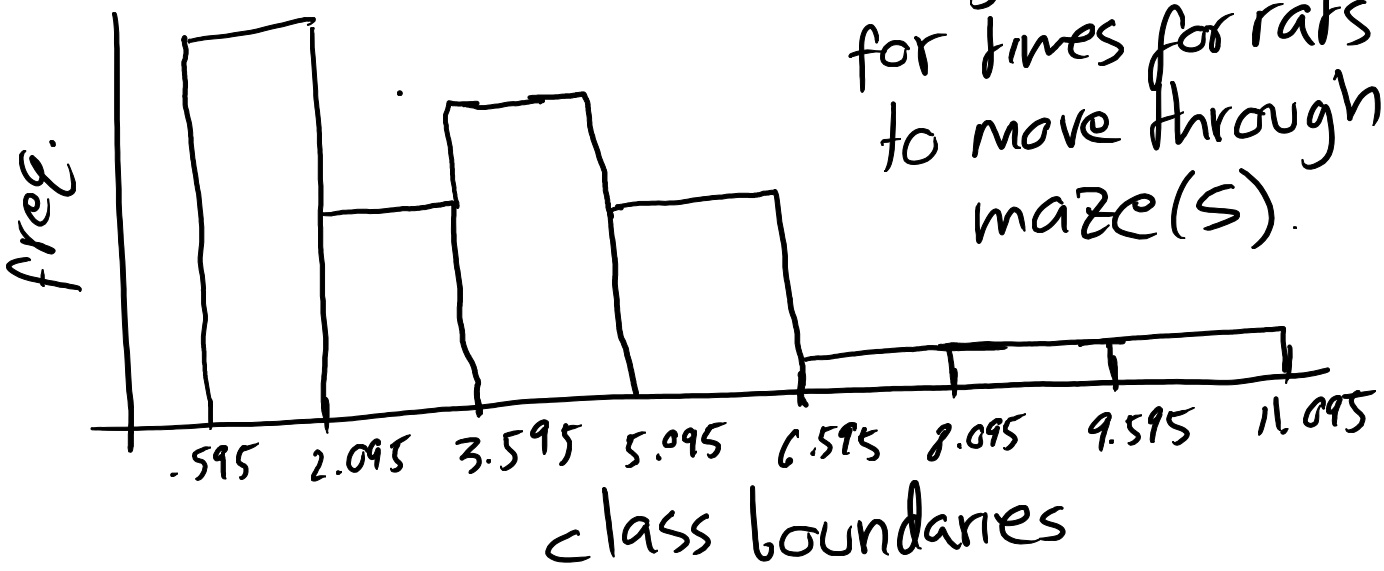
0.595, 2.095, 3.595,

5.095, 6.595, 8.095,

9.595, 11.095

diffs are 1.5 = class width

Histogram for times for rats to move through maze(s).



I: $(\bar{x}-s, \bar{x}+s) = (1.54, 5.94)$ $23/30 \simeq 77\%$

II: $(\bar{x}-2s, \bar{x}+s) = (-.66, 8.14)$ $28/30 \simeq 93\%$

III: $(\bar{x}-3s, \bar{x}+3s) = (-2.86, 10.34)$ $30/30 = 100\%$

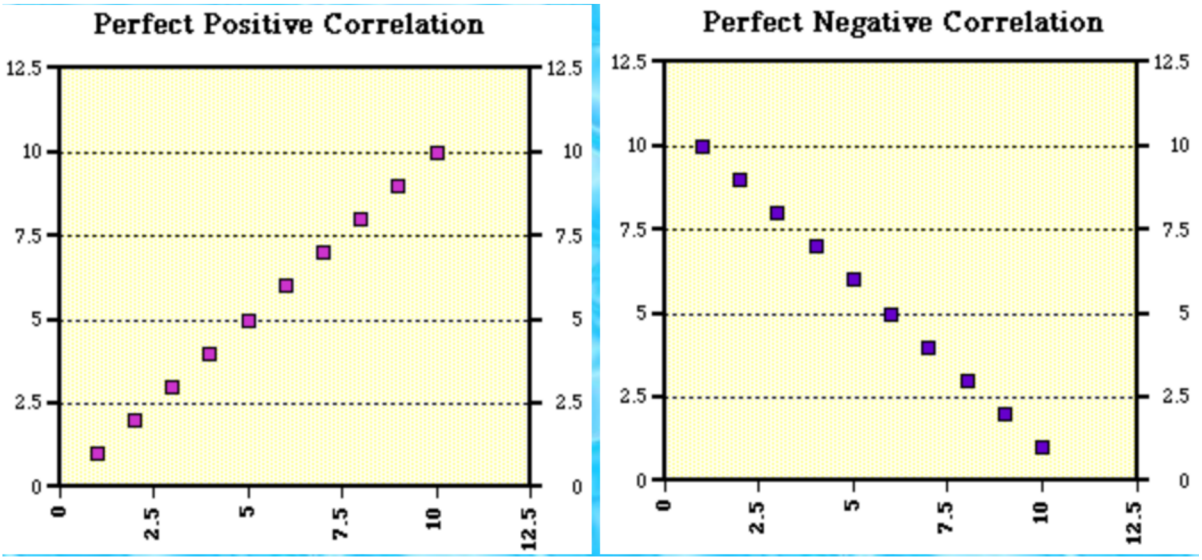
Emprical rule: (68, 95, 99.7)

Good estimate usmg the empirical rule.

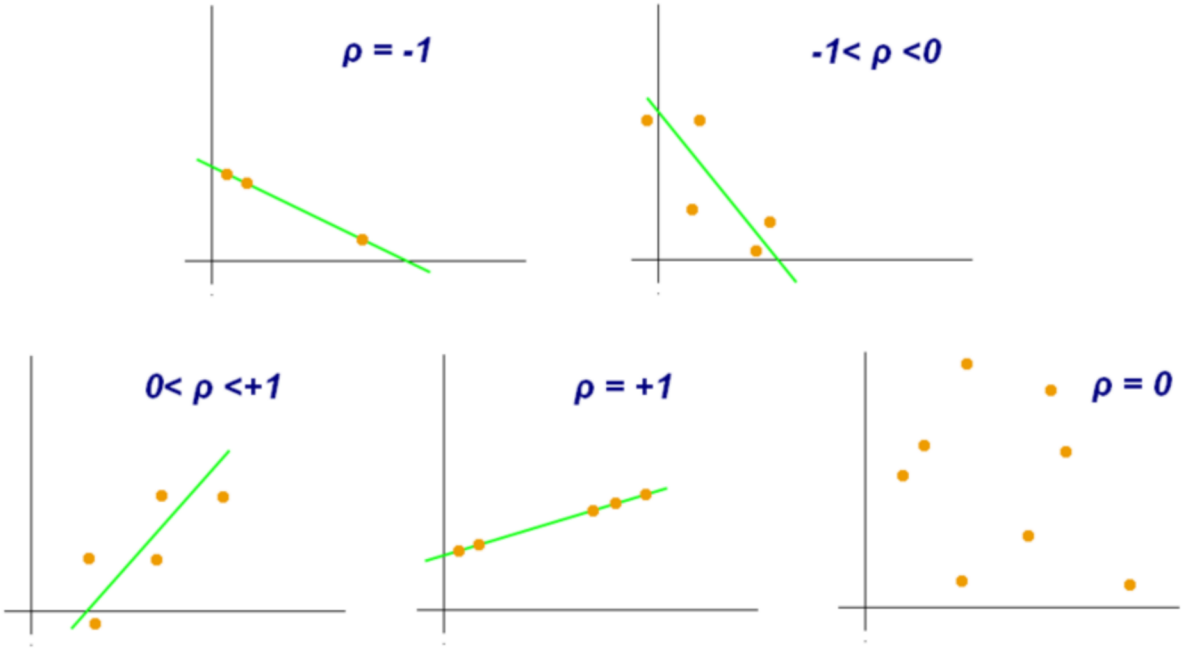
By Chebyshev's rule we expect at least 75%, 89% of measurements to lie within II and III, which holds above. Note that for interval I, Chebyshev's theorem can't be used.

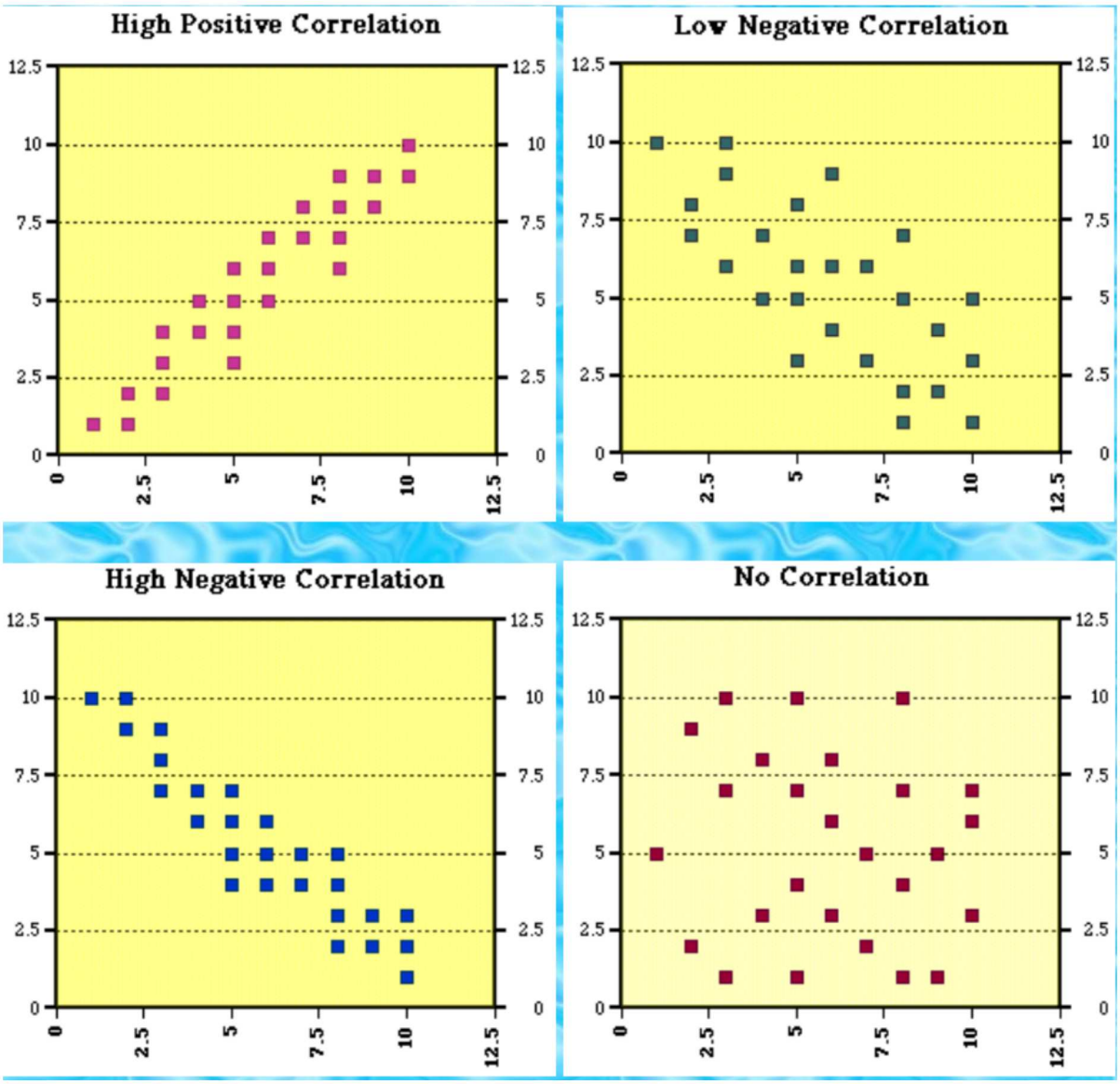
Scatter plots and correlation

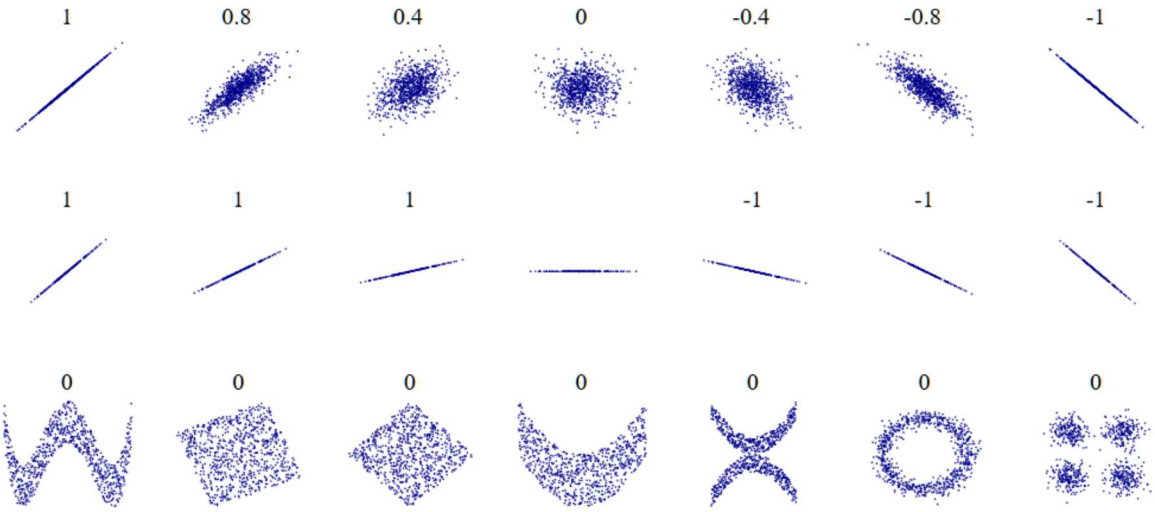
Allow us to compare two sets of data $x = \{x_1, \dots, x_n\}$ and $y = \{y_1, \dots, y_n\}$ and observe if any correlation (not causation) exists between them.



Correlation coefficient $-1 \leq \rho \leq 1$




High Positive Correlation
Low Negative Correlation
High Negative Correlation
No Correlation
0
2.5
5
7.5
10
12.5


1
0.8
0.4
0
-0.4
-0.8
-1
1
1
1
-1
-1
-1
0
0
0
0
0
0
0

Let $s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$

$$s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

Then the Pearson's correlation coefficient may be written as:

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$$

$$= \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y}$$

If the correlation coefficient r is close to 1 or -1, then a straight line can be drawn through the data pts' region; which in "some measure" is the best fit line.

$$y_{LS} = \underbrace{mx}_{\text{slope}} + \underbrace{b}_{y\text{-intercept}}$$

For each data pair (x_i, y_i), we define the residual as: $r_i = mx_i + b_i - y_i$

Then let:

$$R(m,b) = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (mx_i + b_i - y_i)^2$$

we want to determine the constants m and b to make the total residual measure $R(m,b)$ (sometimes called "cost function") as small as possible. This can be accomplished using calculus by setting:

$$\frac{\partial R}{\partial m} = 0 = \frac{\partial R}{\partial b}$$

The result is a system of equations from which m and b can be determined.

$$m = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$b = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i y_i \sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

In R, the command abline can be used to plot a straight line with a specified slope and y-intercept.