Statistics week 5

Expected Value

- Expectation or expected value of a random variable X is its mean, the average value.
- We know that X can take different values with different probabilities. For this reason, its average value is not just the average of all its values. Rather, it is a weighted average.

Expectation, discrete case

$$\boxed{\mu = \mathbf{E}(X) = \sum_x x P(x)}$$

Example-Expected value

- If you roll a 6 sided die , each side has a probability of 1/6 for landing.
- The Expected value is given as:

$$\left(\frac{1}{6}\times 1\right)+\left(\frac{1}{6}\times 2\right)+\left(\frac{1}{6}\times 3\right)$$
$$+\left(\frac{1}{6}\times 4\right)+\left(\frac{1}{6}\times 5\right)+\left(\frac{1}{6}\times 6\right)=3.5$$

Expectation of function

$$\mathbf{E}\{g(X)\} = \sum_x g(x)P(x)$$

Properties of Expectation

$$\mathbf{E}(aX + bY + c) = a\mathbf{E}(X) + b\mathbf{E}(Y) + c$$

In particular,

$$\mathbf{E}(X + Y) = \mathbf{E}(X) + \mathbf{E}(Y)$$

$$\mathbf{E}(aX) = a\mathbf{E}(X)$$

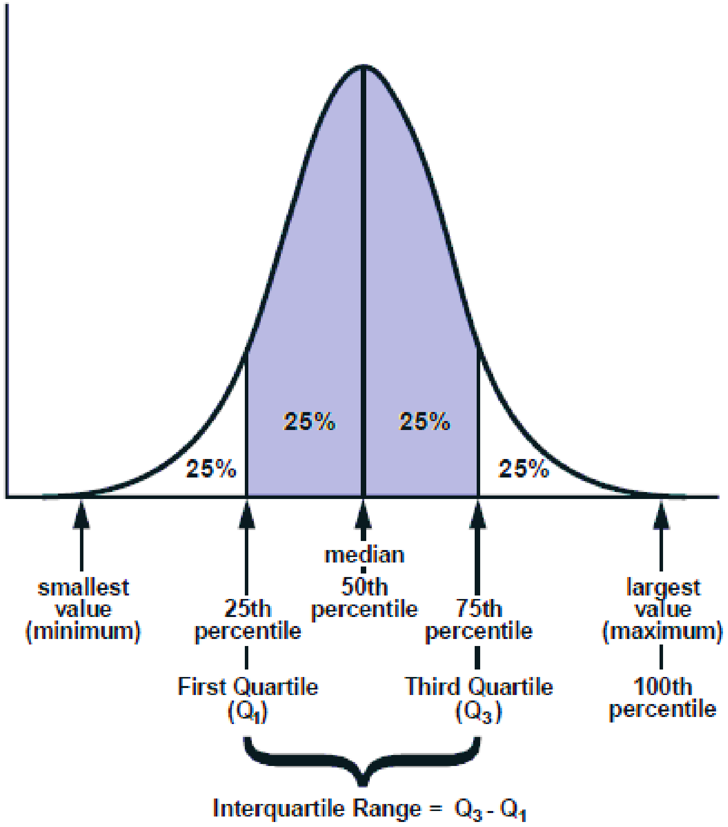
$$\mathbf{E}(c) = c$$

For **independent** X and Y,

$$\mathbf{E}(XY) = \mathbf{E}(X)\mathbf{E}(Y)$$

Inter quartile range

- The interquartile range (IQR) is a measure of the spread of the data.
- The IQR may also be called the midspread, middle 50%.
- It is defined as the difference between the 75th and 25th percentiles of the data.
- $IQR = Q_3 - Q_1$



Inter quartile range: In terms of CDF (for continuous function)

- A random variable X has density f_X, where f_X is a non-negative function, if:

$$\Pr[a \leq X \leq b] = \int_a^b f_X(x)dx$$

where the right-hand side represents the probability that the random variable X takes on a value less than or equal to x.

- Hence, if F_X is the cdf of X, then:

$$F_X(x) = \int_{-\infty}^{x} f_X(u)du$$

and (if f_X is continuous at x)

$$f_X(x) = \frac{d}{dx}F_X(x)$$

- Then Q_1 = $F_X(x) = 0.25$, and Q_3 = $F_X(x) = 0.75$

Inter quartile range: In terms of CDF (for discrete function)

- Consider a signal having L discrete levels $X_0, X_1, \cdots X_{L-1}$. The probability distribution function (pdf) for k^th level is defined as:

$$p(k) = n(k) \Big/ \sum_{k=0}^{L-1} n(k)$$

where, $k \in [0, L-1]$, $n(k)$ is signal of k^th level.

- The cdf can be defined as:

$$c(k) = \sum_{q=0}^{k} p(q)$$

where, $c(k)$ is the cdf at the k^th level. Note that $c(L-1)$ will always be unity.

- Then $Q_1 = c(x) = 0.25$, and $Q_3 = c(x) = 0.75$

Covariance

- Covariance is a measure of how much two random variables vary together.
- It's similar to variance, but where variance tells you how a single variable varies, covariance tells you how two variables vary together.
- Covariance between two random variables x and y can be calculated using the following formula (for population):

$$cov(x,y) = \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{n}$$

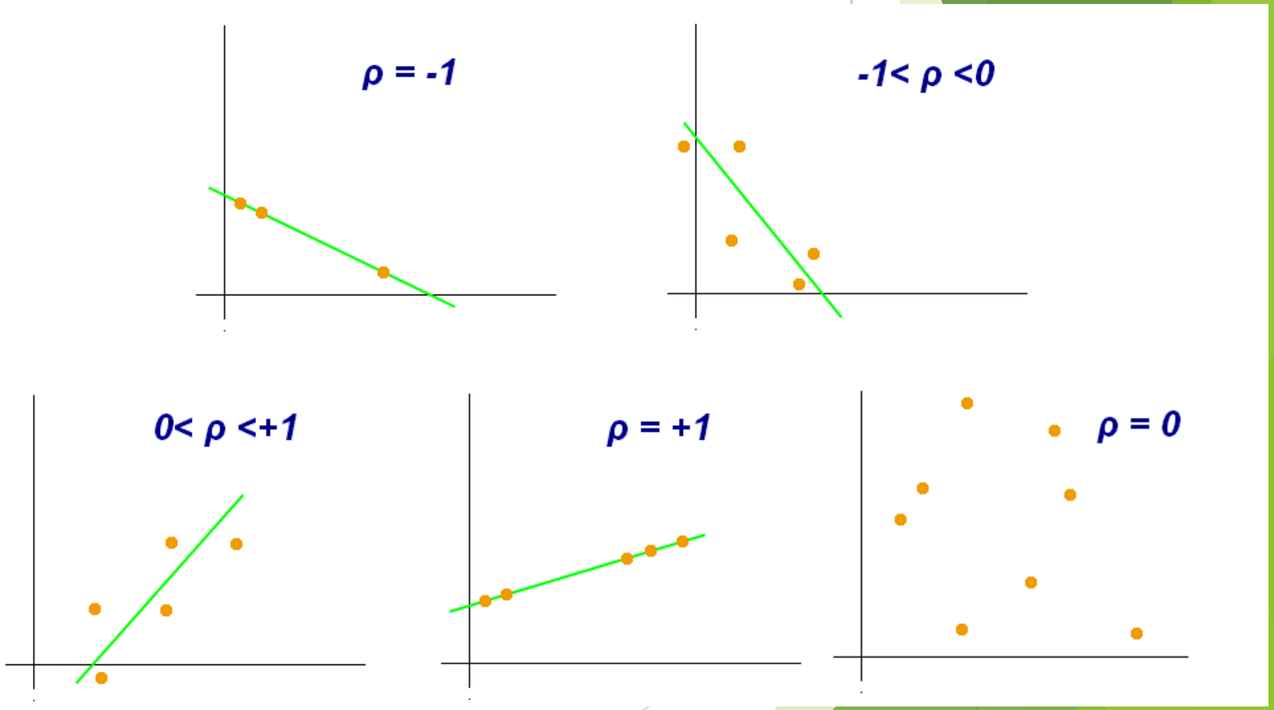
- For a sample covariance, the formula is slightly adjusted:

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

Pearson correlation coefficient

- In statistics, the Pearson correlation coefficient (ρ) is a measure of linear correlation between two sets of data.
- It is the ratio between the covariance of two variables and the product of their standard deviations.
- It is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1.
- The measure can only reflect a linear correlation of variables, and ignores many other types of relationship or correlation.
- As a simple example, one would expect the age and height of a sample of teenagers from a high school to have a Pearson correlation coefficient significantly greater than 0, but less than 1 (as 1 would represent an unrealistically perfect correlation).

	+.70 or higher	Very strong positive relationship
	+.40 to +.69	Strong positive relationship
	+.30 to +.39	Moderate positive relationship
	+.20 to +.29	Weak positive relationship
	+.01 to +.19	No or negligible relationship
Plot	0	No relationship [zero correlation]
	-.01 to -.19	No or negligible relationship
	-.20 to -.29	Weak negative relationship
	-.30 to -.3	Moderate negative relationship
	-.40 to -.69	Strong negative relationship
	-.70 or higher	Very strong negative relationship



- Given a pair of random variables (x, y), The formula for ρ is:

$$\rho(x, y) = \frac{cov(x, y)}{\sigma_x \sigma_y}$$

where, cov is the covariance, σ_x is the standard deviation of x, σ_y is the standard deviation of y.

- For sample,

$$\rho_s(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Skewness

- Pearson's first skewness coefficient (mode skewness):

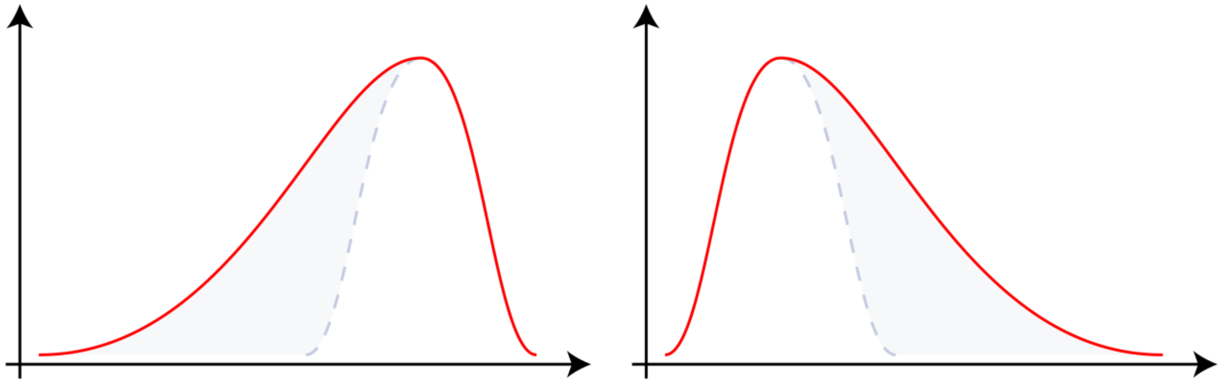
$$\gamma_{mode} = \frac{Mean - Mode}{SD}$$

- Pearson’s second skewness coefficient (median skewness):

$$\gamma_{median} = \frac{3(Mean - Mode)}{SD}$$

- Bowley's measure of skewness (Quartile based skewness):

$$\gamma_{quartile} = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1}$$



Negative Skew   Positive Skew

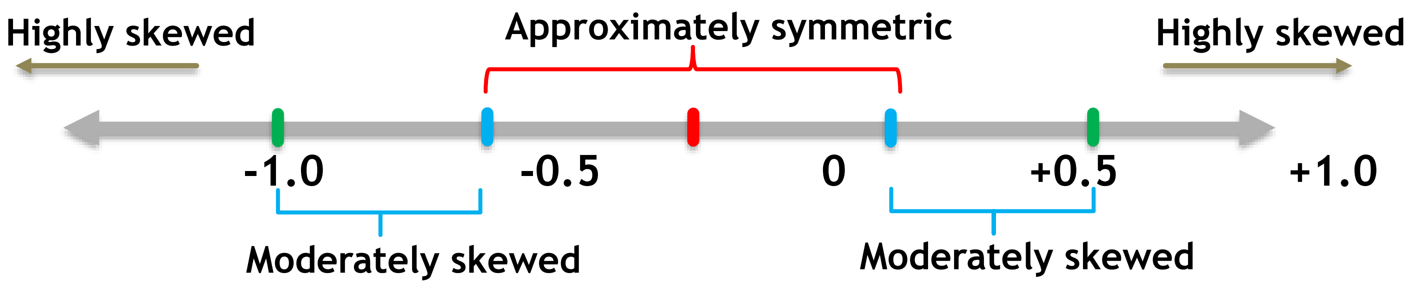
- Moment coefficient of skewness of a population (biased estimator):

$$\gamma = \frac{1}{n}\frac{\sum(x-\mu)^3}{\sigma^3}$$

- Moment coefficient of skewness of n-sample (unbiased estimator):

$$\gamma = \frac{n}{(n-1)(n-2)}\frac{\sum(x-\bar{x})^3}{{s_n}^3}$$

- Visualisation



Example (mode skewness)

Calculate Karl Pearson coefficient of skewness of the following data set ($S = 1.7$).

Value (x)	1	2	3	4	5	6	7
Frequency (f)	2	3	4	4	6	4	2

Solution

$$
\begin{aligned}
\text{mean} = \bar{x} &= \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i}^{n} f_i} \\
&= \frac{1 \times 2 + 2 \times 3 + \ldots + 7 \times 2}{25} \\
&= \frac{104}{25} \\
&= 4.16 \\
\text{mode} &= 5 \\
S_{kp} &= \frac{\text{mean-mode}}{\text{standard deviation}} \\
&= \frac{4.16 - 5}{1.7} = -0.4941
\end{aligned}
$$

Since $S_{kp} < 0$ distribution is skewed left.

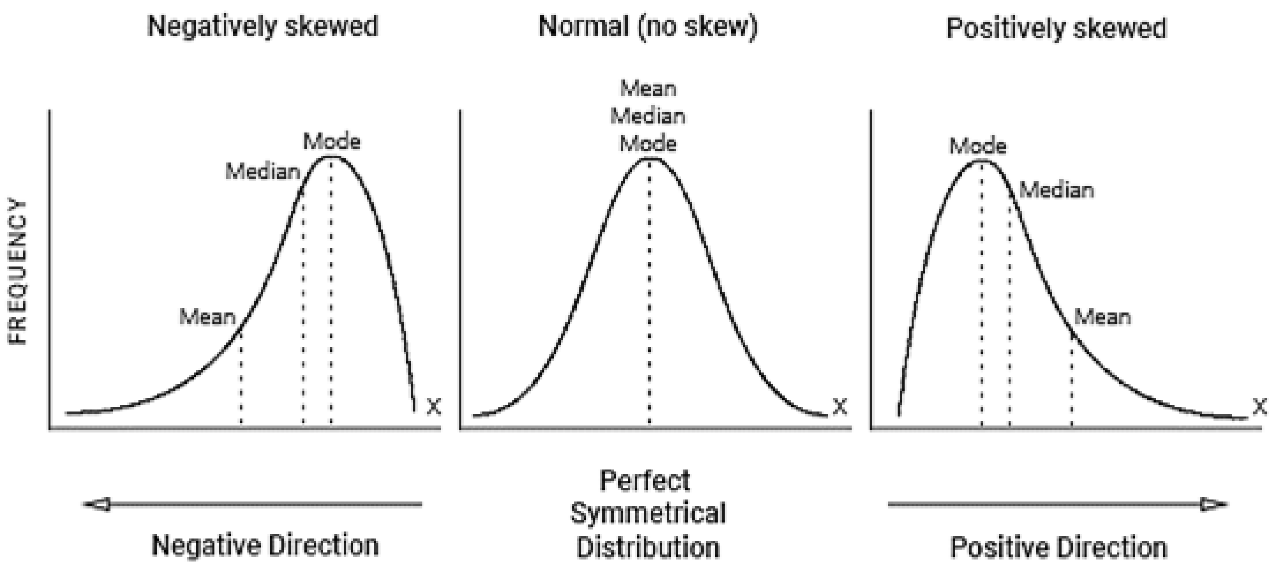
How to visualise skewness

What sign actually means: $-1 \leq S_{kp} \leq 1$.

$S_{kp} = 0 \Rightarrow$ distribution is symmetrical about mean.

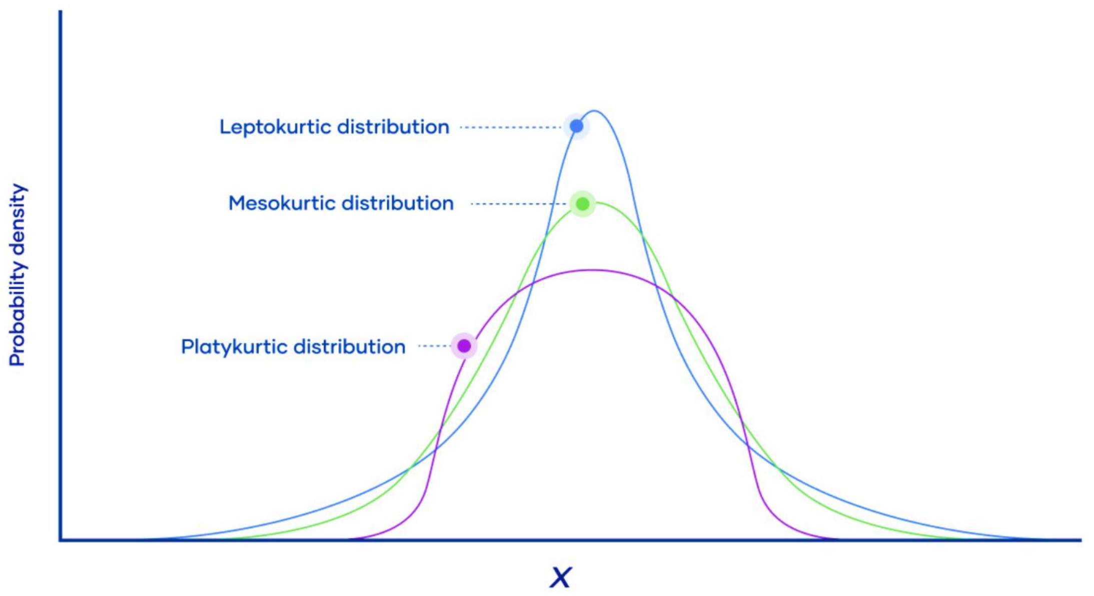
$S_{kp} > 0 \Rightarrow$ distribution is skewed to the right.

$S_{kp} < 0 \Rightarrow$ distribution is skewed to the left.



Kurtosis

- Kurtosis is a measure of the tailedness of a distribution.
- Tailedness is how often outliers occur.
- Excess kurtosis is the tailedness of a distribution relative to a normal distribution.
  - Distributions with medium kurtosis (medium tails) are mesokurtic.
    - Kurtosis=3.
  - Distributions with low kurtosis (thin tails) are platykurtic.
    - Kurtosis<3.
  - Distributions with high kurtosis (fat tails) are leptokurtic.
    - Kurtosis>3.
- Kurtosis can vary from 1 to ∞.



Kurtosis

- Kurtosis of a population (biased estimator):

$$\frac{1}{n}\left(\frac{\sum(x-\mu)^4}{\sigma^4}\right)$$

- Excess kurtosis:

$$\frac{1}{n}\left(\frac{\sum(x-\mu)^4}{\sigma^4}\right)-3$$

Note: -3 is added to make the value of mesokurtic (reference normal distribution) = 0.

- Kurtosis of a n-sample (unbiased estimator):

$$\frac{n(n+1)}{(n-1)(n-2)(n-3)}\left(\frac{\sum(x-\bar{x})^4}{{s_n}^4}\right)-\frac{3(n-1)^2}{(n-2)(n-3)}$$

	Category		
	Mesokurtic	**Platykurtic**	**Leptokurtic**
Tailedness	Medium-tailed	Thin-tailed	Fat-tailed
Outlier frequency	Medium	Low	High
Kurtosis	Moderate (3)	Low (< 3)	High (> 3)
Excess kurtosis	0	Negative	Positive

Moments (of statistical distribution)

- The shape of any distribution can be described by ‘moments’:
- ✓ The 0^{th} moment is a reference point. In statistics it assumed as zero origin.
- ✓ The 1^{st} moment is the center of mass of a probability distribution. Also known as mean, which indicates the central tendency of a distribution.
- ✓ The 2^{nd} moment deals with spread of a distribution. Also known as variance, which indicates the width or deviation.
- ✓ The 3^{rd} moment deals with relation of the two tails of a distribution. Also known as skewness, which indicates any asymmetric ‘leaning’ to either left or right.
- ✓ The 4^{th} moment deals with the combined size of the tails relative to whole distribution. Also known as Kurtosis, which indicates the ‘fatness’ of the outer tails.

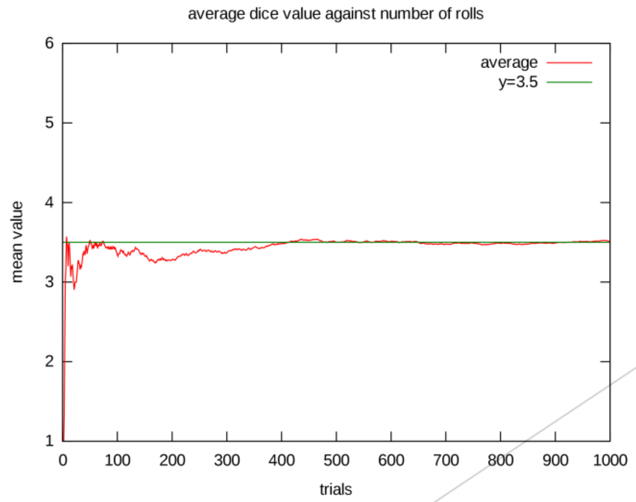
Monte Carlo Method

- Monte Carlo method, also known as the Monte Carlo simulation or a multiple probability simulation, is a mathematical technique, which is used to estimate the possible outcomes of an uncertain event.
- Monte Carlo Simulations have assessed the impact of risk in many real-life scenarios, such as in robotics, drug development, stock prices, sales forecasting, etc.
- Monte Carlo method allows decision-makers to see the impact of individual inputs on a given outcome, and correlation allows them to understand relationships between any input variables.

- Regardless of what tool you use, Monte Carlo techniques involves three basic steps:
  - Set up a predictive model, identifying both the dependent variable to be predicted and the independent variables (also known as the input, risk or predictor variables) that will drive the prediction.
  - Specify probability distributions of the independent variables. Use historical data and/or the analyst’s subjective judgment to define a range of likely values and assign probability weights for each.
  - Run simulations repeatedly, generating random values of the independent variables. Do this until enough results are gathered to make up a representative sample of the near infinite number of possible combinations.

Lab task:
Fair six-sided die: MATLAB mobile

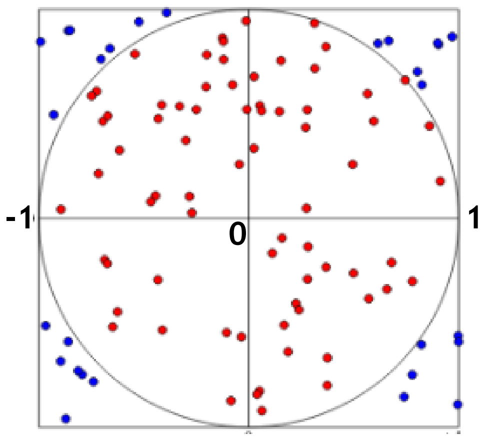
- num_trials = 1000; — % Specify number of trials.
- trials = randi(6, [1 num_trials]); — % Now grab all the dice rolls.
- figure(1); — % Assign specific figure.
- plot(cumsum(trials)./(1:num_trials), 'r-’); — % Cumulative sum of trial results divided by index gives the average.
- hold on;
- plot([1 num_trials], [3.5 3.5], 'color', [0 0.5 0]); — % Let's put a green reference line at 3.5.
- title('average dice value against number of rolls');
- xlabel('trials');
- ylabel('mean value');
- legend('average', 'y=3.5');



Try the same program for different number of trials (e.g. 10, 700, 10000 etc.) and comment on findings.

Lab task:
π: MATLAB mobile

- N = 1000; %Specify number of points
- r = 1; %Circle radius
- n = 0; % Successful event counter
- x = 2*rand(1,N)-1; % N samples between -1 and 1
- y = 2*rand(1,N)-1;
- for i = 1:N
-     if ((x(i)^2+y(i)^2)<=r^2)
-         n = n+1;
-     end
- end
- pi_pred = 4*n/N



From where this formula is coming?

Try the same program for different number of trials (e.g. 10, 700, 10000 etc.) and comment on findings.

Any Questions?