UNIVERSITI SAINS MALAYSIA

CDS501 PRINCIPLES & PRACTICES OF DATA SCIENCE & ANALYTICS

Lab 5

Statistical Distribution with R

INTRODUCTION

- This tutorial demonstrate how to use statistical distribution functions in R: normal (norm), binomial (binom) and multinomial (multinom) distributions.
- Each function (except for multinomial) have a single-letter prefix that defines the type of function. These prefixes are d, p, q and r which refer to density, cumulative, quantile and sampling respectively. For example, normal distribution has four functions which are dnorm(), pnorm(), qnorm() and rnorm(). Similarly for binomial distribution (binom).

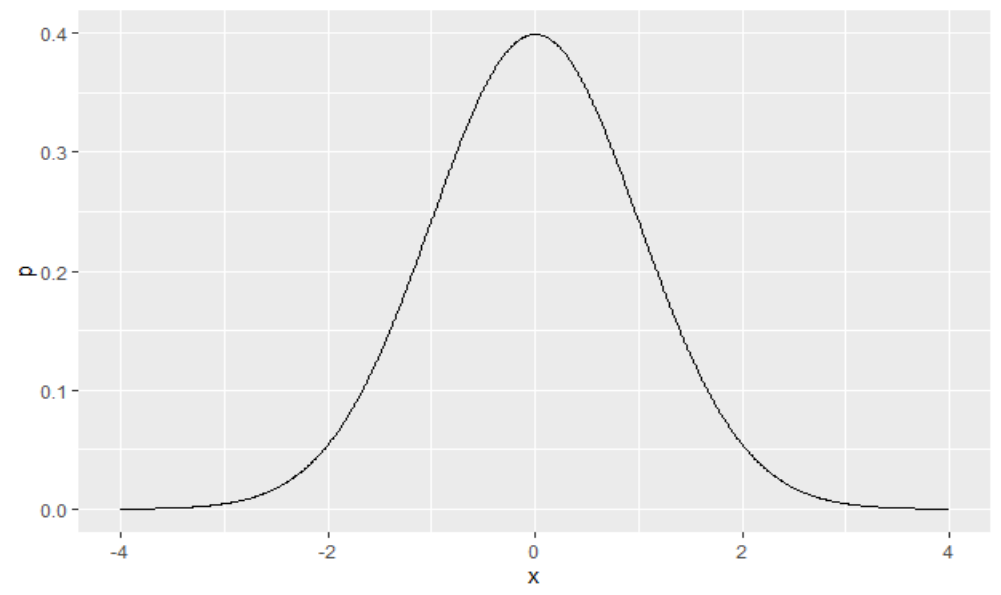
NORMAL DISTRIBUTION

- The normal distribution is defined by the following probability density function, where μ is the population mean and σ is the standard deviation

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

NORMAL DISTRIBUTION

- the normal distribution with μ=0 and σ=1 is called the standard normal distribution, and is denoted as $N(0,1)$. It can be plotted as follows.



EXAMPLE

- Assuming the daily revenue of a grocery shop follows a normal distribution $N(100,20)$ with μ=RM100 and σ=RM20. The maximum revenue is RM200. Let's plot the probability distribution of the daily revenue.
- **> mu <- 100 , > sd <- 20**
- **> n <- seq(0, 200, 0.01) , > f <- dnorm(n, mu, sd)**
- **> ggplot(data.frame(p=f), aes(x=n, y=f)) + geom_line() + ylab("Probability") + xlab("Daily revenue")**

EXAMPLE…

- Let's calculate what is the probability that the revenue of today will be less or equal to RM125, $P(X \leq 125)$?
- **> p <- pnorm(125, mu, sd)** we use pnorm() function
- What is the probability that the revenue of today will be more than RM125, $P(X > 125)$?
- **> p <- 1-pnorm(125, mu, sd)**
- We can set the argument lower.tail = FALSE in order to get the same answer as the above statement.
- **> p <- pnorm(125, mu , sd, lower.tail = FALSE)**

EXAMPLE…

- Let's plot the cumulative distribution of the daily revenue.
- **> f <- pnorm(n, mu, sd)**
- **> ggplot(data.frame(p=f), aes(x=n, y=f)) + geom_line() + ylab("Probability") + xlab("Daily revenue")**
- We can reverse the question to ask what is the daily revenue if the probability is 0.6? To obtain the revenue, we use qnorm() function as follows.
- **> r <- qnorm(0.6, mu, sd)**

BINOMIAL DISTRIBUTION

- The binomial distribution is a discrete probability distribution that describes the outcome of n independent trials in an experiment. Each trial is assumed to have only two outcomes, either success or failure. If the probability of a successful trial is p, then the probability of having x successful outcomes in n independent trials is as follows.

$$f(x = success) = \binom{n}{x} p^x (1-p)^{(n-x)}$$

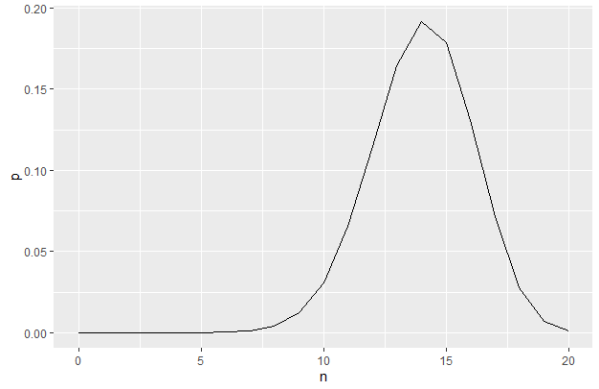
$$\text{where } \binom{n}{x} = \frac{n!}{x!(n-x)!}$$

EXAMPLE

- Assuming a coin is tossed 20 times and the probability of landing on heads is 0.5. Let's plot the probability distribution of this experiment.
- **> n <- 0:20** # plot for all events (0 success, 1 success etc.)
- **> p <- 0.5**
- **> n_trial = 20**
- **> f <- dbinom(x=n, size=n_trial, prob=p)**
- **> ggplot(data.frame(p=f), aes(x=n, y=p)) + geom_line()**

EXAMPLE…

- We can see that in 20 tosses, we will get 10 heads since the probability is 0.5. The summation of the probability = 1.0
- **> sum(f)**
- Let's change the probability of landing on heads to 0.7. We will see that the chance of getting more heads is higher.



EXAMPLE…

- let's reduce the probability of landing on heads to 0.3. We can see from the plot that the chance of getting heads is also reduced.
- Let's generate binomial data of a fair coin being tossed for 10 times in 20 trials. The head has a probability of 0.5.
- **> n_trial <- 20**
- **> n_toss_trial <- 10 # number of toss per trial**
- **> p <- 0.5**
- **> data <- rbinom(n=n_trial, size=n_toss_trial, prob=p)**

MULTINOMIAL DISTRIBUTION

- Multinomial distribution is defined as follows

$$f = \frac{n!}{n_1!\, n_2! \ldots n_x!} p_1^{n_1} p_2^{n_2} \ldots p_x^{n_x}$$

- where n is the number of trials, n_x is the number of outcome x and p_x is the probability of outcome x.

EXAMPLE

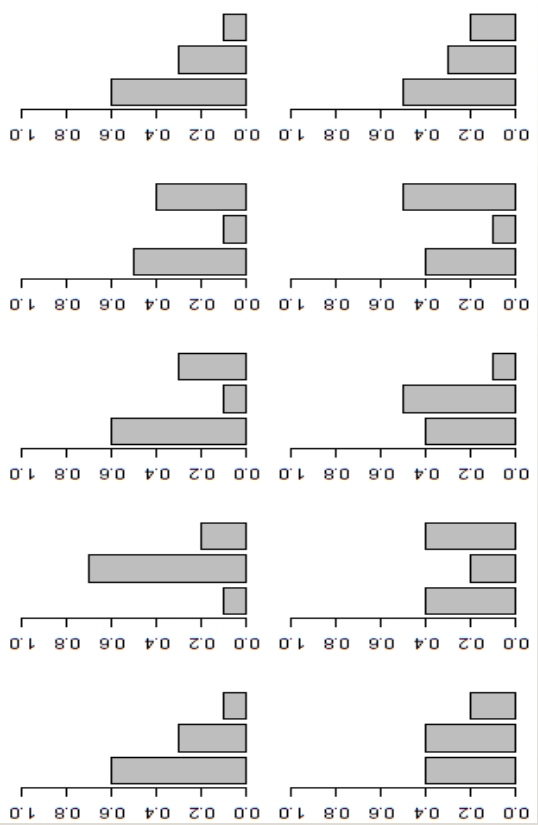
- Assuming a three-sided dice is tossed 10 times. The probability of the dice landing on "1", "2" and "3" are 0.40, 0.35 and 0.25 respectively. To calculate the probability of getting five "1", three "2" and two "3", we use dmultinom() function as follows.
- **> outcome <- c(5,3,2)**
- **> p <- c(0.4,0.35,0.25)**
- **> f <- dmultinom(x=outcome, prob=p)**

EXAMPLE…

- Let's generate multinomial data of the three-sided dice being tossed 10 times for 10 trials.
- **> n_trial <- 10**
- **> n_toss <- 10**
- **> p <- c(0.4,0.35,0.25)**
- **> data <- dmultinom(n=n_trials, size=n_toss, prob=p)**

EXAMPLE…

- The columns correspond to each trial and the rows correspond to each outcome.
- Let's plot bar chart of the data.
- **> df <- data.frame(data)/n_toss**
- **> par(mar=c(1,2,1,2), mfrow=c(2,5))**
- **> for(i in 1:10) {**
- **barplot(df[,i], ylim=c(0,1))**
- **}**



EXERCISES

- Assume a computer science course in university with 1000 students enrolled. The professor has marked the final examination papers and inputting the grades into a spreadsheet. He sees that the average for the final examination is 65% with more than half of the students having grades in the range between 55% and 75%.
- a. Generate a dataset for the grades using rnorm function and display the data distribution using density plot.
- b. Calculate the probability that a randomly chosen exam paper will have a grade of less than 50%?

EXERCISES…

- c. Calculate the probability that a randomly chosen exam paper will have a grade between 65 and 85?
- Suppose a quiz has 10 multiple choice questions. Each question has five possible answers.
- a. If there is only one answer is correct, find the probability of having five or less correct answers if a student attempts to answer every question.
- b. If there are two answers are correct, find the probability of having six or more correct answers if a student attempts to answer every question.

EXERCISES…

- Suppose an election for the president of the student's CS society is being held. Based on a survey, 35% of the eligible voters prefer candidate 1, 40% prefer candidate 2 and 25% have no preference.
- a. Calculate the probability that 20 will prefer candidate 1, 25 will prefer candidate 2 and 5 will have no preference from 50 voters.
- b. If there are 1000 voters, generate the multinomial data for 20 elections