

Statistical Computing with R: Masters in Data Sciences 503 (S18) Fourth Batch, SMS, TU, 2025

Shital Bhandary

Associate Professor

Statistics/Bio-statistics, Demography and Public Health Informatics

Patan Academy of Health Sciences, Lalitpur, Nepal

Faculty, Masters in Medical Research, NHRC/Kathmandu University

Faculty, FAIMER Fellowship in Health Professions Education, India/USA

Unit 4: Review Preview 1

- Average
 - Geometric mean
 - Harmonic mean
- Moments
 - First order
 - Second order
 - Third order
 - Fourth order
- Normal distribution
 - Skewness
 - Kurtosis
- Relative measures
 - Quintiles
 - Percentiles
 - Standard or z-score
- Breakdown analysis: “apply”

Measures of central tendencies

- Computing measure of central tendency, dispersion, moments and relations position in R using packages and functions/scripts
- Measures of Central Tendency – mean, median, mode, **geometric mean, harmonic mean**
- Measure of Dispersion – standard deviation, inter-quartile range, range
- Moments – mean, standard deviation, **skewness, kurtosis**
- Relative position – percentile, quartiles and z-score

Geometric mean in R:

<https://www.r-bloggers.com/2021/08/calculate-geometric-mean-in-r/>

- **$GM = (x_1, x_2, x_3, \dots, x_n)^{1/n}$**
- **`exp(mean(log(x)))`**
- `data <- c(1, 15, 12, 5, 18, 11, 12, 15, 18, 25)`
- `exp(mean(log(data)))`
- 10.37383 (Interpretation?)
- `data <- c(1, 15, 12, 5, 0, 18, 11, 12, 15, 18, 25, 0, -11)`
- `exp(mean(log(data[data>0])))`
- 10.37383
- **Interpretation?**
- **Better for summarizing simple rates, ratios and proportions!**

Harmonic mean in R:

<https://www.geeksforgeeks.org/harmonic-mean-in-r/>

- Harmonic mean (H) is defined as:

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \dots + \frac{1}{x_n}}$$

- It is available in “psych” package
- Syntax: *harmonic.mean(x)*

Harmonic mean in R:

<https://www.geeksforgeeks.org/harmonic-mean-in-r/>

load the library

- `library("psych")`

create dataframe

- `data=data.frame(col1=c(12,2,3,4),`
- `col2=c(34,32,1,0),`
- `col3=c(2,45,3,2))`

display

- `print(data)`

harmonic mean of column1

- `print(data$col1)`

harmonic mean of column2

- `print(data$col2)`

harmonic mean of column3

- `print(data$col3)`

- **Interpretation?**

- **Better for summarizing instantaneous/marginal rates!**

Moments in R:

ChatGPT plug-in for Google Chrome

- To calculate moments in R, you can use the moments package, which provides functions for various statistical moments.
- Here's an example of calculating the first four moments (mean, variance, skewness, and kurtosis) for a numeric vector x:

```
# Load the 'moments' package
library(moments)

# Calculate the moments of x <- c(1,
2, 3, 4, 5)

• mean_x <- mean(x) #First moment
• var_x <- var(x) #Second moment
• skewness_x <- skewness(x) #Third
• kurtosis_x <- kurtosis(x) #Fourth
• Interpretations?
```

Third and fourth moments

<https://www.itl.nist.gov/div898/handbook/eda/section3/eda35b.htm>

- Skewness – symmetry
- Pearson's coefficient of skewness
- Bowley's coefficient of skewness
- Coefficient = 0 = Symmetrical
- Kurtosis – peakedness
- Pearson's coefficient of kurtosis
- Coefficient of excess kurtosis
- 0 = Mesokurtic for excess kurtosis
- <0 = Platykurtic for excess kurtosis
- >0 = Leptokurtic for excess kurtosis

Percentile & quintiles in R:

ChatGPT plug-in for Google Chrome

- In R, you can calculate percentiles or quartiles or quintiles using the `quantile()` function.
- The function takes two arguments: the data vector and the desired percentile.
- Here's an example:

```
# Create a sample vector of data
```

```
• data <- c(12, 5, 9, 17, 3, 8, 10)
```

```
# Calculate the 75th percentile
```

```
• percentile_75 <- quantile(data,  
0.75)
```

```
# Print the result
```

```
• print(percentile_75)
```

• The output will be the 75th percentile of the data vector, in this case, 11.5.

Standard score or z-score in R: ChatGPT plug-in for Google Chrome

- In R, you can calculate the z-score using the `scale()` function. The `scale()` function standardizes a numeric vector by subtracting the mean and dividing by the standard deviation.
- `# Create a vector of numeric values`
- `values <- c(10, 15, 12, 8, 20)`
- `# Calculate the z-score`
- `z_scores <- scale(values)`
- `# Print the z-scores`
- `print(z_scores)`
- This will give you the standardized z-scores for each value in the vector.

Breakdown analysis with “apply” function

<https://www.geeksforgeeks.org/apply-lapply-sapply-and-tapply-in-r/>

- **tapply() function**
- The [tapply\(\)](#) helps us to compute **statistical measures** (mean, median, min, max, etc..) or a self-written function operation **for each factor variable in a vector**.
- It helps us to create a subset of a vector and then apply some functions to each of the subsets.
- # load library tidyverse
- library(tidyverse)
- # print head of diamonds dataset
- head(diamonds)
- # apply tapply function to get average price by cut
- tapply(diamonds\$price, diamonds\$cut, mean)

Question/queries so far?

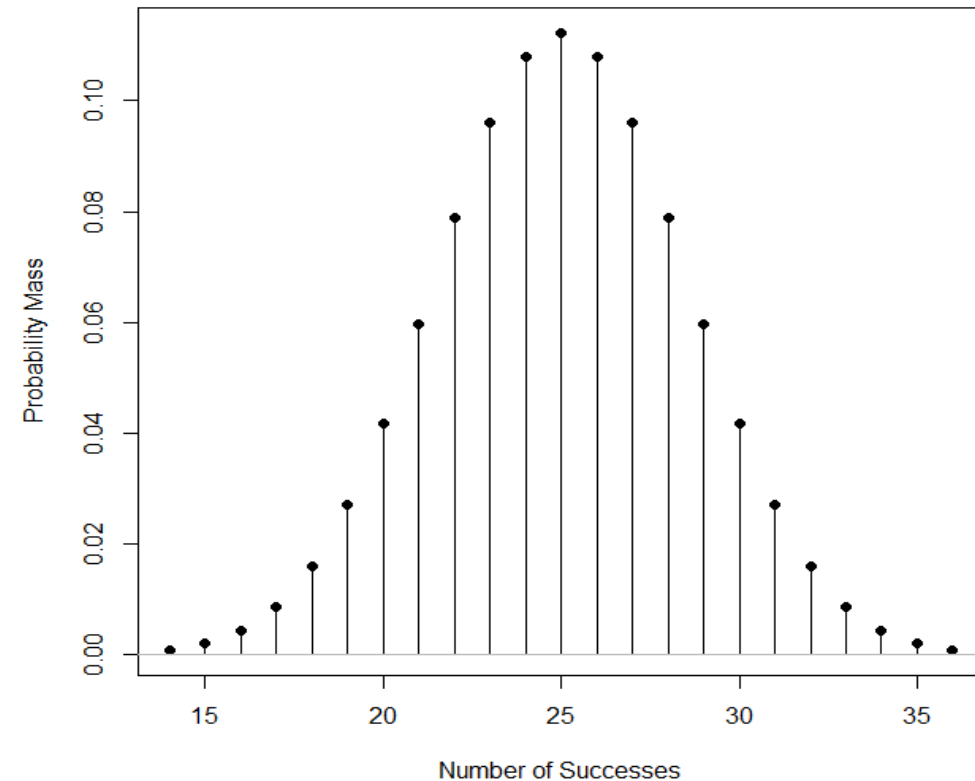
Unit 4: Review Preview 2

- Probability distribution functions
 - Discrete
 - Continuous
- Demo with selected distributions
- Normal approximations of binomial distribution
- Test of normality
 - Graphical
 - Test

Discrete probability distribution:

- **Binomial**
 - Poisson
 - Geometric
 - Hypergeometric
 - Negative binomial etc.
-
- Binomial distribution is used heavily in the classification models of supervised learning!

Binomial Distribution: Binomial trials=50, Probability of success=0.5



Discrete probability distribution: Binomial

(Check with $p=0.3$ and $p=0.7$ vs $p=0.5$ below!)

```
#Number of trials
```

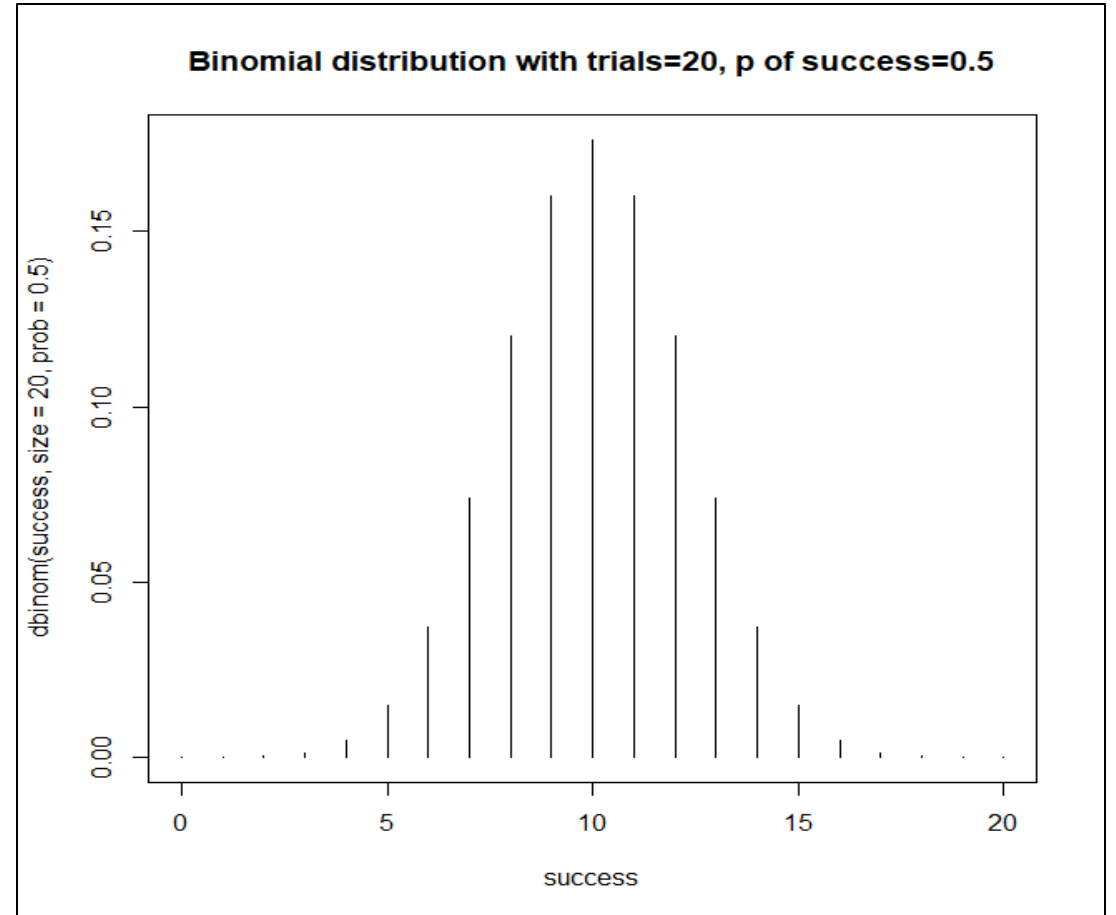
```
success <- 0:20
```

```
# Binomial Probability distribution  
with success probability of 0.5
```

```
dbinom(success, size=20, prob=0.5)
```

```
#Plot
```

```
plot(success, dbinom(success,  
size=20, prob=0.5), type="h", main =  
"Binomial distribution with n=20 and  
p of success=0.5")
```



Let's get/check the data of success and binomial probabilities (**Do this in excel**):

```
binomc <- cbind(success, binomd)
binomc (Results are on the right side >>>)
How was the “binomd” values created?
```

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x} \quad ; x = 0, 1, 2, \dots, n.$$

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

$$n! = n * (n - 1) * \dots * 3 * 2 * 1$$

Prove that: sum of “binomd” = 1 in R and Excel! Why is this important?

	success	binomd
[1,]	0	0.00000009536743
[2,]	1	0.0000190734863
[3,]	2	0.0001811981201
[4,]	3	0.0010871887207
[5,]	4	0.0046205520630
[6,]	5	0.0147857666016
[7,]	6	0.0369644165039
[8,]	7	0.0739288330078
[9,]	8	0.1201343536377
[10,]	9	0.1601791381836
[11,]	10	0.1761970520020
[12,]	11	0.1601791381836
[13,]	12	0.1201343536377
[14,]	13	0.0739288330078
[15,]	14	0.0369644165039
[16,]	15	0.0147857666016
[17,]	16	0.0046205520630

What is normal approximation of binomial distribution? When to use it??

- Is it related to the sample size of the successes and failures?
- Earlier when $n \cdot p > 5$ and $n \cdot q > 5$, it was considered that it will approximate the normal distribution
- Now, it is set at $n \cdot p > 10$ and $n \cdot q > 10$! Important for class imbalance problems!!
- Which regression model is used when we need to use normal distribution for dichotomous or binary dummy dependent variable (Yes = 1 and No = 0)
- Logistic regression model because log transformation was used to convert the exponential equation form to make it linear model!
- What to do if we have class imbalance problem?

Q2: When and how to use?

- Poisson distribution?
- Zero-inflated Poisson distribution. When to use?
- Negative binomial distribution. When to use?
- Hypergeometric distribution?
 - Fisher's exact test??
 - When to use???

Continuous probability distributions:

- Normal
- T
- Chi-square
- F
- Exponential
- Logistic etc.

Normal Distribution Formula

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

μ = mean of x

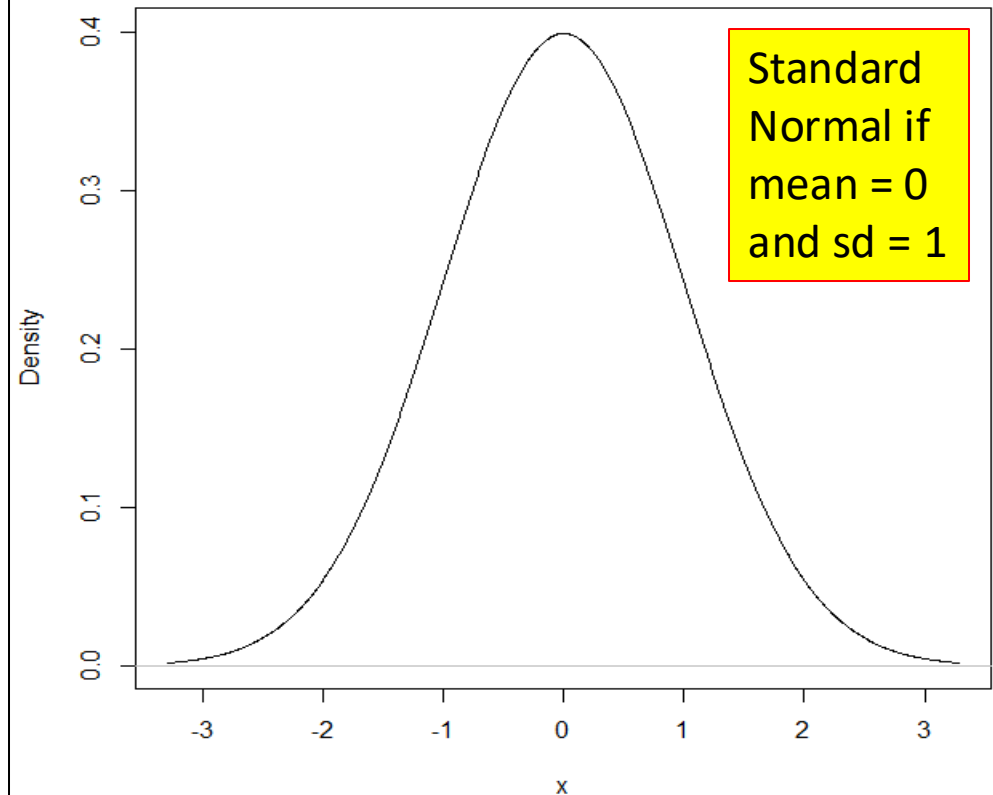
σ = standard deviation of x

$\pi \approx 3.14159 \dots$

$e \approx 2.71828 \dots$

- Normal/Standard Normal Distribution is used in the linear and general linear regression models of supervised learning!

Normal Distribution: Mean=0, Standard deviation=1



Normal Distribution of values between -4 and +4 with pre-defined population mean and sd:

#Define mean and SD

```
pop_mean <- 50
```

```
pop_sd <- 5
```

#Define lower and upper limits

```
LL <- pop_mean - pop_sd
```

```
UL <- pop_mean + pop_sd
```

#Create a sequence of 100 x values based on pop mean and sd

```
x <- seq(-4,4,  
length=100)*pop_sd+pop_mean
```

```
y <- dnorm(x, pop_mean, pop_sd)
```

Normal Distribution Formula

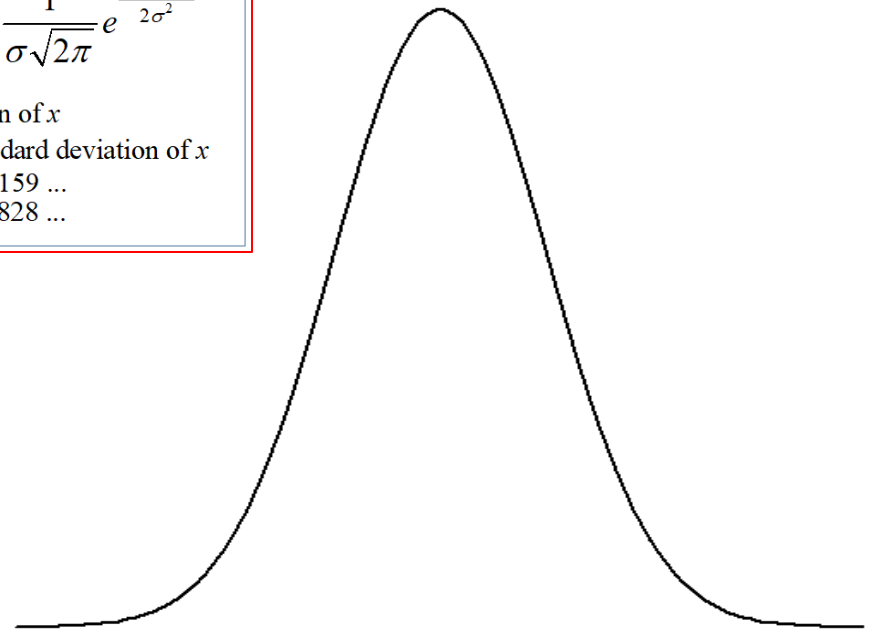
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

μ = mean of x

σ = standard deviation of x

$\pi \approx 3.14159 \dots$

$e \approx 2.71828 \dots$



```
plot(x,y, type="l", lwd=2, axes=F, xlab="", ylab="")
```

Adding x-axis values and mean in the curve:

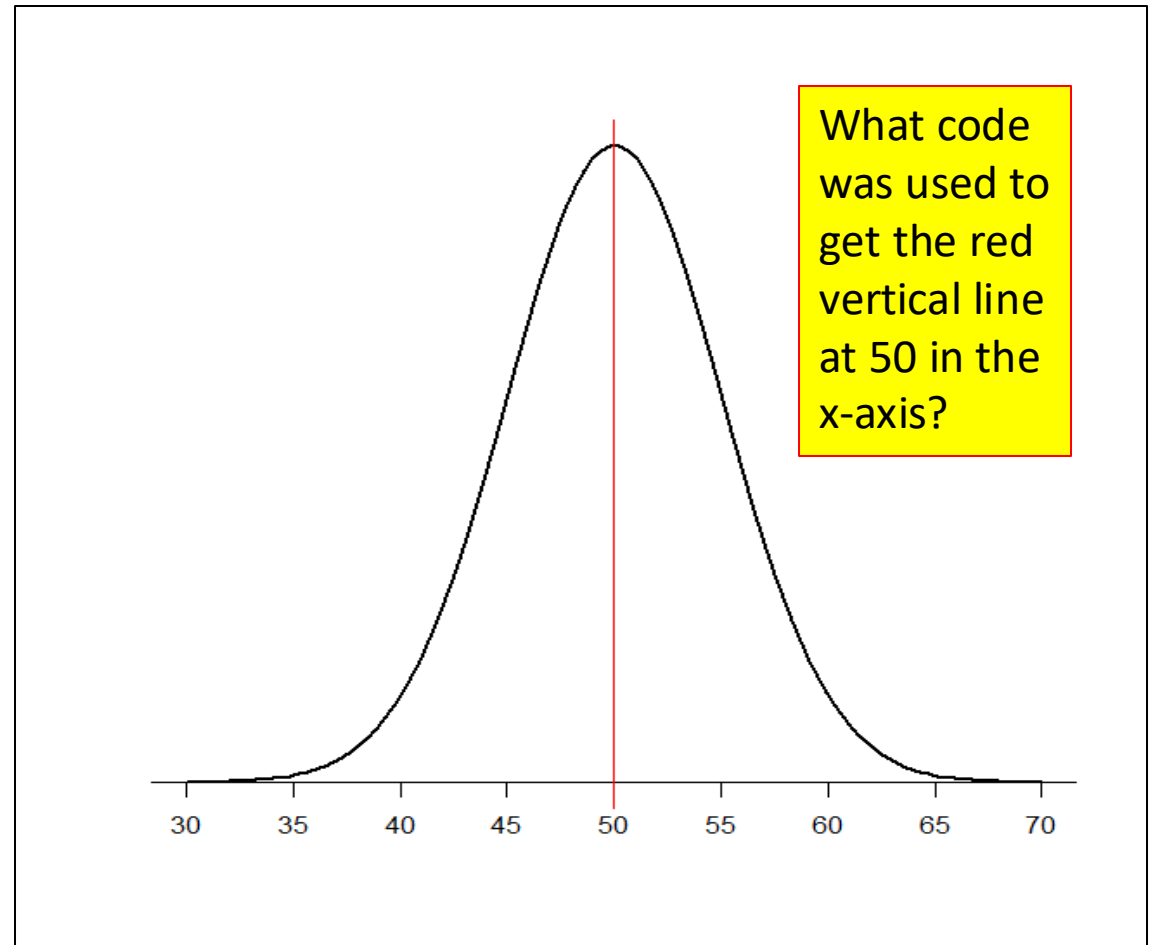
```
plot(x,y, type="l", lwd=2, axes=F,  
xlab="", ylab="")
```

```
sd_axis_bounds = 5
```

```
axis_bounds <- seq(-  
sd_axis_bounds*pop_sd +  
pop_mean,  
sd_axis_bounds*pop_sd +  
pop_mean, by=pop_sd)
```

```
axis(side=1, at=axis_bounds,  
pos=0)
```

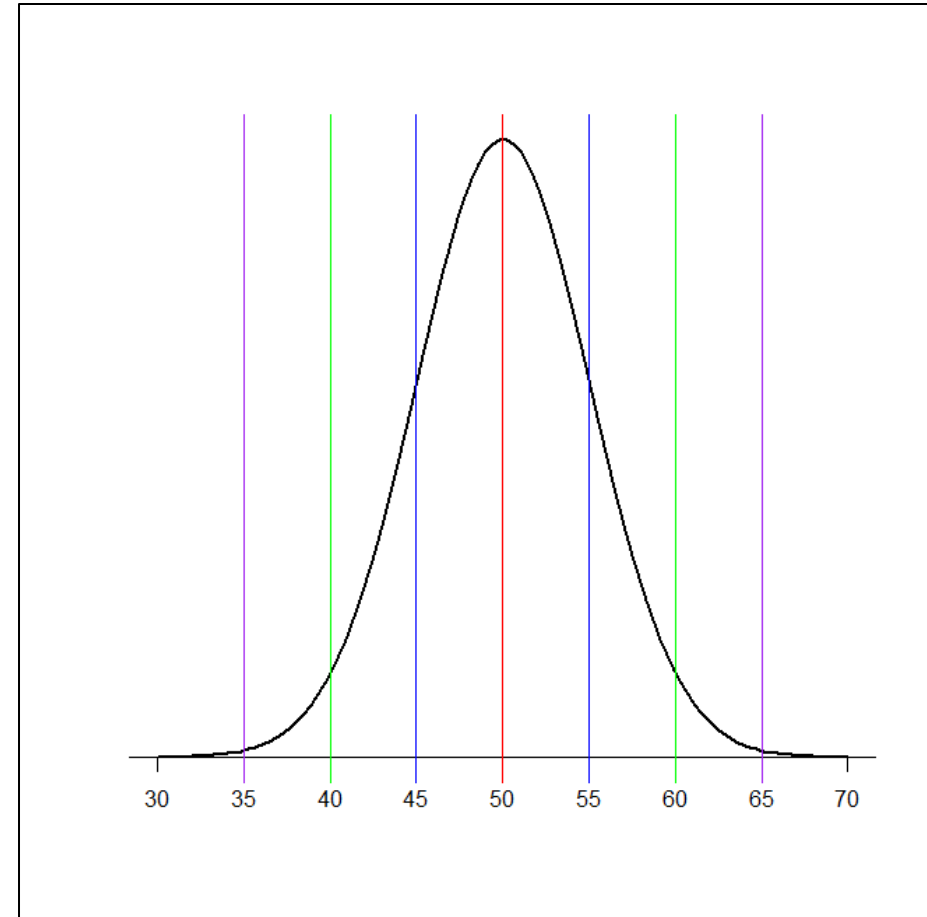
```
abline(??)
```



Class work 1:

- Get this graph and **provide annotation in** it as follows:
- 45-55: mean \pm 1SD = 67% data
- 40-60: mean \pm 2SD = 95% data
- 35-65: mean \pm 3SD = 99% data

Note: You can use ggplot2 package, if required!



Why normal distribution is important?

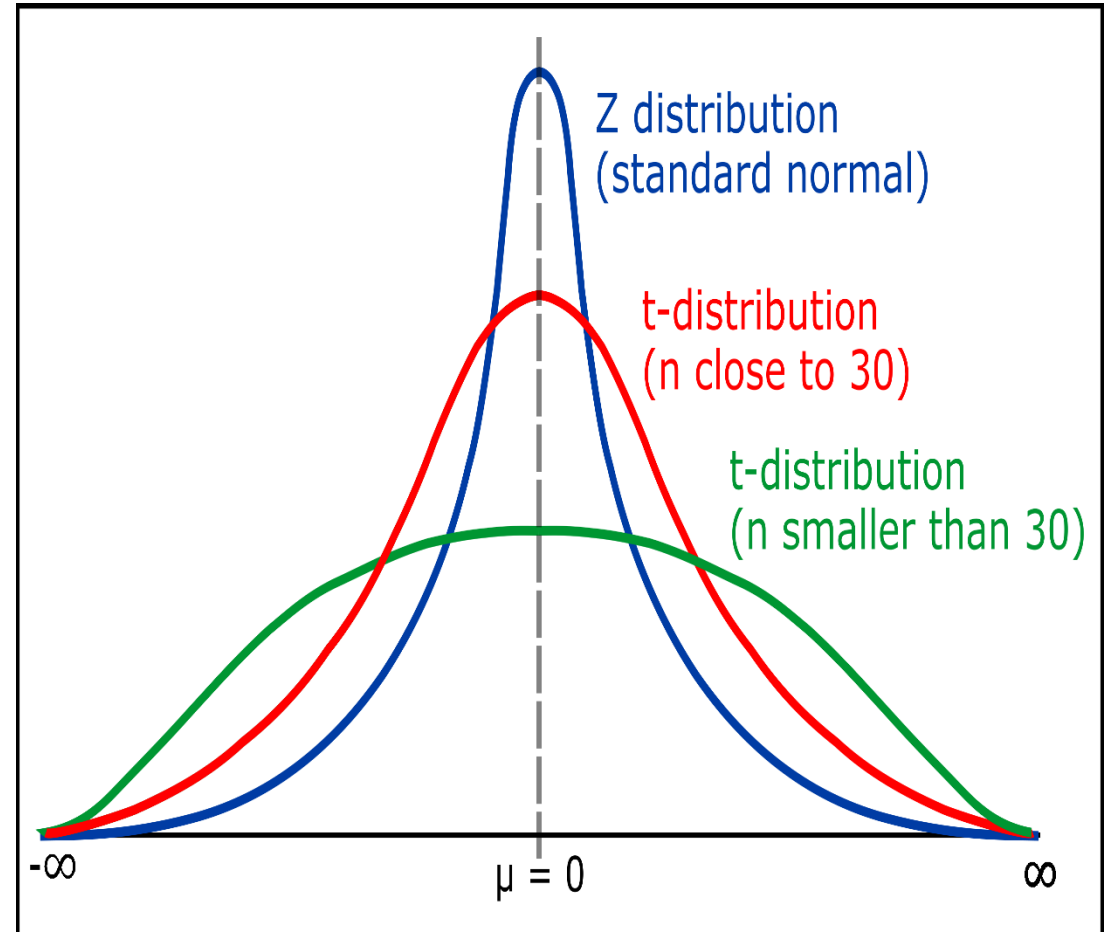
- When continuous variable follows the theoretical normal distribution then we must summarize that variable using mean and standard deviation
- We can also use t-test and 1-way ANOVA to compare means across two or more categories of categorical variables respectively
- When continuous variable do not follow the theoretical normal distribution then we must summarize that variable using median and inter-quartile range
- We can only use median test to compare median across two or more categories of the categorical variables

Q3: Why these test must NOT be used?

- Mann-Whitney U test **must not be used** to compare medians across two categories of a categorical variable?
- e.g. comparing age by sex as sex variable normally has two categories “male” and “female” if age is not normally distributed
- Kruskal-Wallis W test **must not be used** to compare medians across two categories of a categorical variable?
- e.g. comparing age by socio-economic status (SES) variable as SES has 3 categories (low, middle, high) or 5 categories (lowest, low, middle, high, highest) if age is normal!

T and Z distributions: Old school approach!

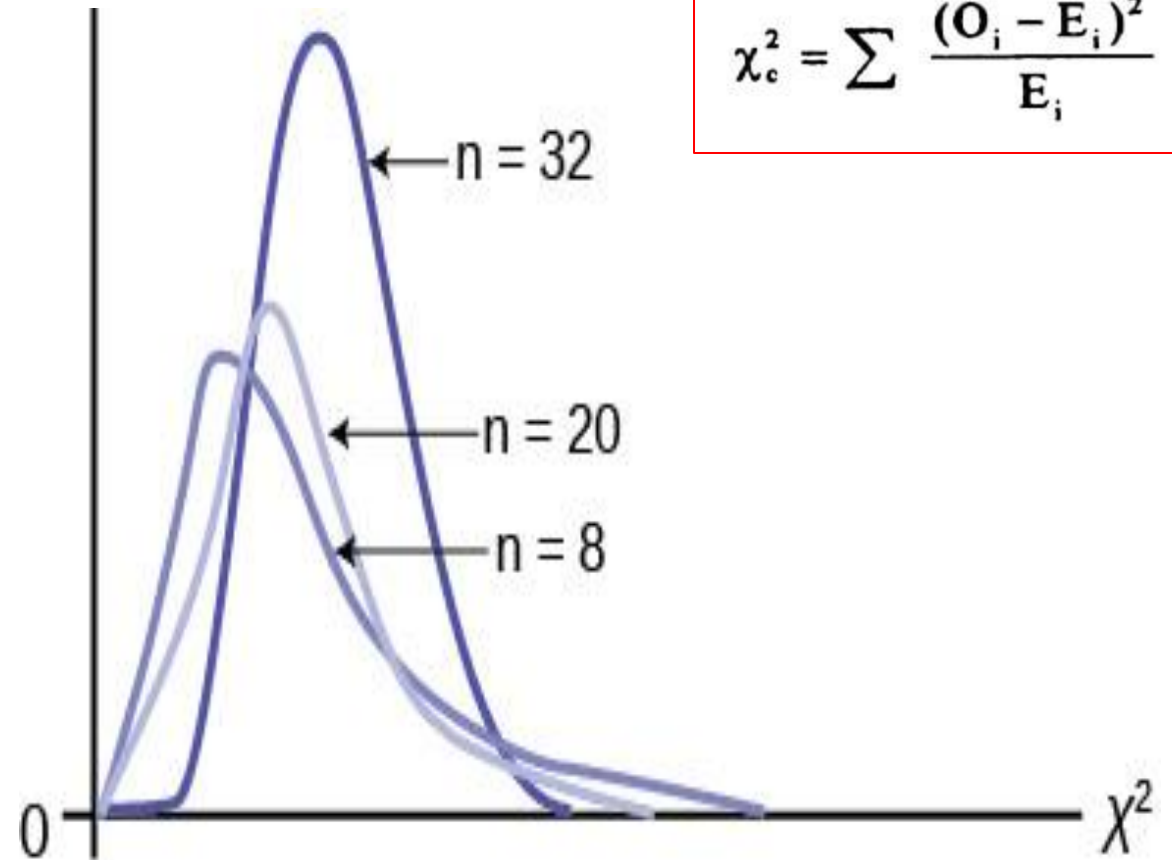
- T distribution is **normally** used when there is small sample size, say, random samples < 30
- As the sample size increases t-distribution behaves like normal distribution so we can use it for large samples too!
- Linear regression is extension of t-test and linear correlation! **Thus, two assumptions are important!**



Chi-square and Z distributions:

- Chi-square distribution is normally used in contingency tables or cross-tabulations to find “association” between dependent and independent variable categories. It is also used for goodness-of-fit and comparing proportions across categories!
- As the sample size increases chi-square distribution also behaves like normal distribution
- **Logistic regression is extension of chi-square test!**

$$\chi^2_c = \sum \frac{(O_i - E_i)^2}{E_i}$$



Q4: Why?

- Logistic regression is described as the extension of the Pearson's chi-square test?
- Both are used to get/test the association between two (or more variables)
- **p-value < 0.05 means association is statistically significant!**
- **Prove it with an example!**
- **Hint:** Create a two-by-two table e.g. smoking vs lung cancer
- Get p-value from chi-square test
- Get p-value from bivariate logistic regression
- Are they same? If yes then good!

Class work 2: Statistical tests are “robust”!

- Get stem-leaf plot, histogram and normal q-q plot of mpg variable of the “mtcars” data
- Test the normality of mpg variable of mtcars data using shapiro wilk test (**Why this test?**)
- `shapiro.test(data)`

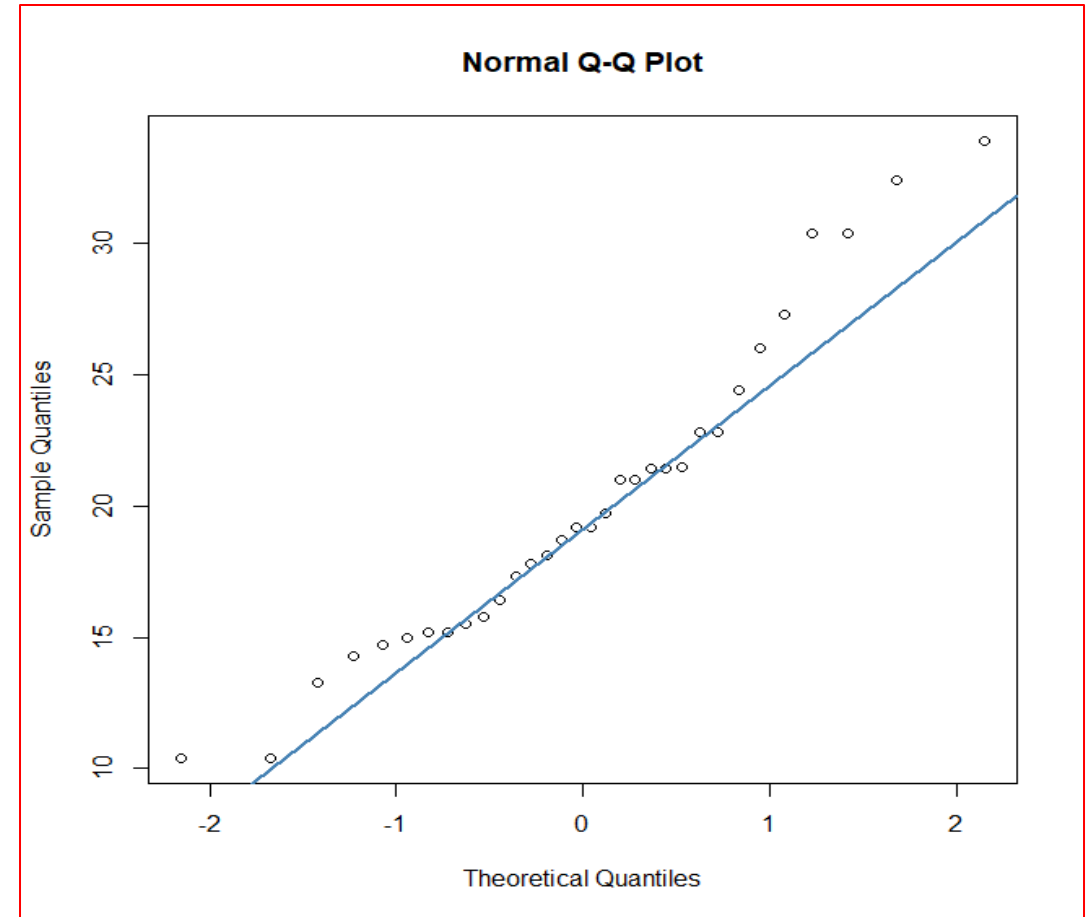
Shapiro-Wilk normality test

data: mtcars\$mpg

$W = 0.94756$, $p\text{-value} = 0.1229$

H_0 : Data follows normal distribution ($p > 0.05$)

H_1 : Data do not follow normal distribution ($p \leq 0.05$)



H_0 : No difference between data and normal distribution

H_1 : Difference between data and normal distribution

Question/queries so far?

- Next class:
- Hypothesis testing with:
 - Z-test
 - T-test
 - Proportion test ...

Thank you!

@shitalbhandary