Statistical Computing with R: Masters in Data Sciences 503, S18 First Batch, SMS, TU, 2021

Shital Bhandary

Associate Professor

Statistics/Bio-statistics, Demography and Public Health Informatics

Patan Academy of Health Sciences, Lalitpur, Nepal

Faculty, Data Analysis and Decision Modeling, MBA, Pokhara University, Nepal

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

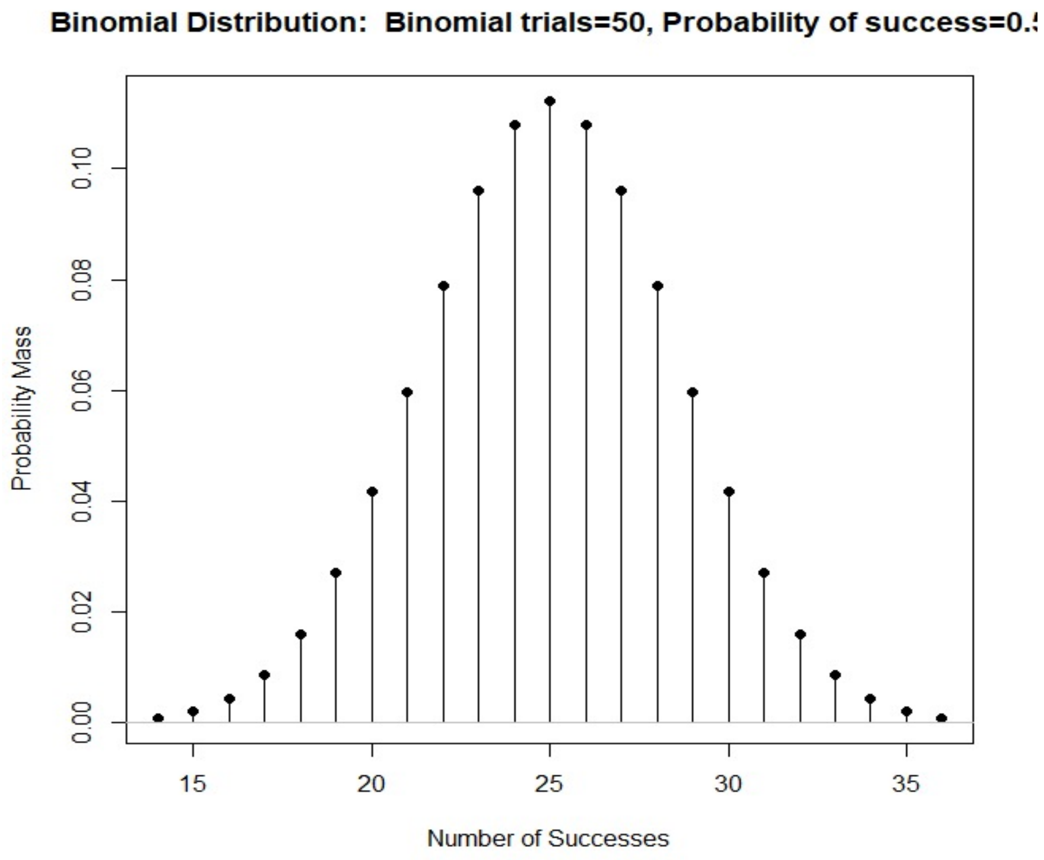
Review Preview

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

Discrete probability distribution:

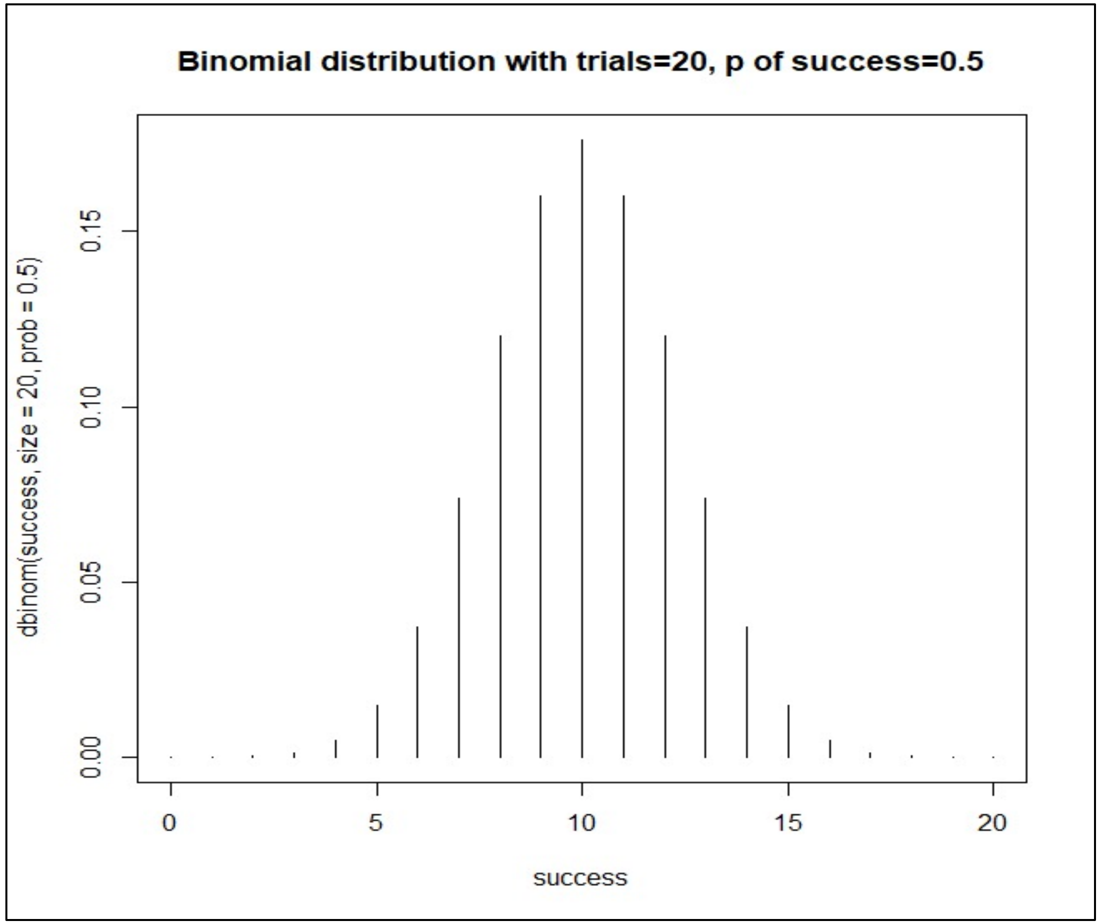
- **Binomial**
- Poisson
- Geometric
- Hypergeometric
- Negative binomial etc.

- Binomial distribution is used heavily in the classification models of supervised learning!



Discrete probability distribution: Binomial
(Check with p=0.3 and p=0.7, any difference?)

```
#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

How was the "binomd" values created?

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

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

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

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

```
      success          binomd
 [1,]       0 0.000009536743
 [2,]       1 0.000190734863
 [3,]       2 0.001811981201
 [4,]       3 0.010871887207
 [5,]       4 0.046205520630
 [6,]       5 0.014785766016
 [7,]       6 0.036964416503
 [8,]       7 0.073928833007
 [9,]       8 0.120134353637
[10,]       9 0.160179138183
[11,]      10 0.176197052002
[12,]      11 0.160179138183
[13,]      12 0.120134353637
[14,]      13 0.073928833007
[15,]      14 0.036964416503
[16,]      15 0.014785766016
[17,]      16 0.004620552063
```

Q1: What is normal approximation of binomial distribution? When to use it??

- Is it related to the sample size of the successes and failures?
- Which regression model is used when we need to use normal distribution for dichotomous or binary dummy dependent variable (Yes = 1 and No = 0)

Q2: When and how to use?

- Poisson distribution?
- Hypergeometric distribution?

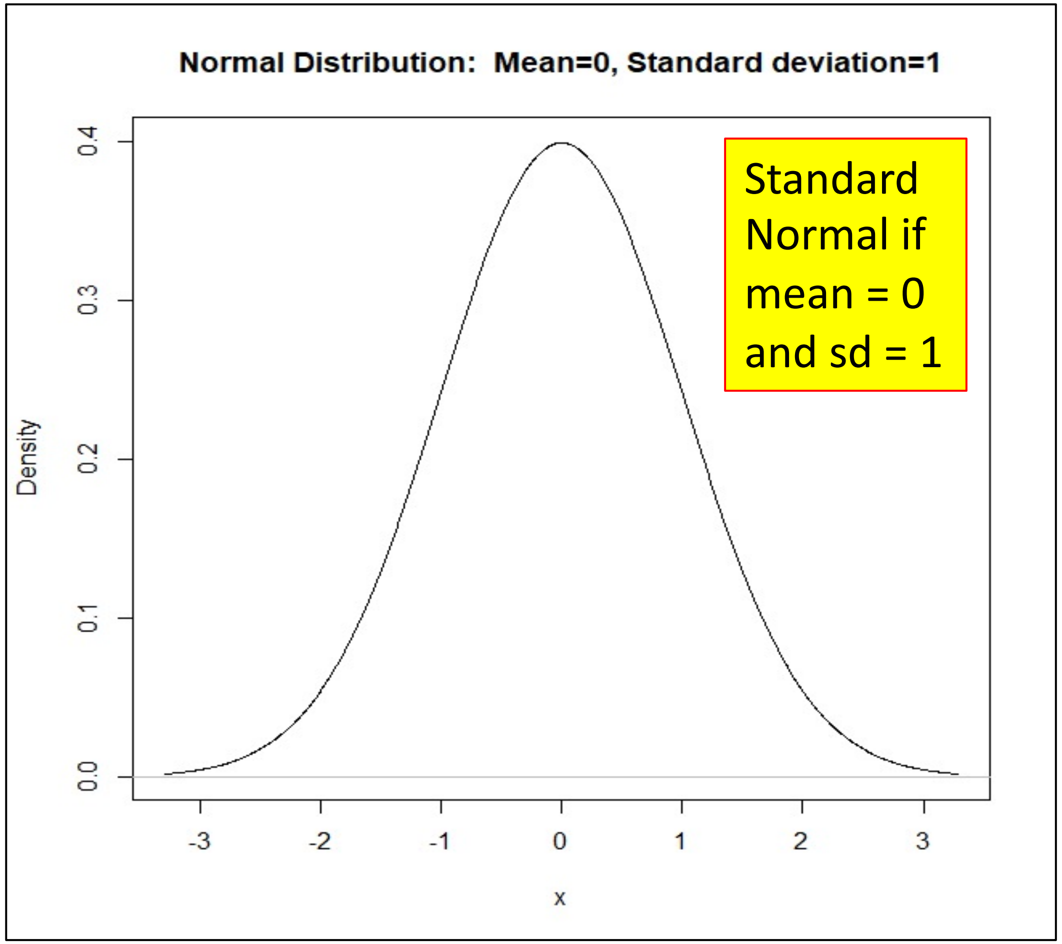
Continuous probability distributions:

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

Normal Distribution Formula

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

$\mu =$ mean of x
$\sigma =$ standard deviation of x
$\pi \approx 3.14159 \ldots$
$e \approx 2.71828 \ldots$



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

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

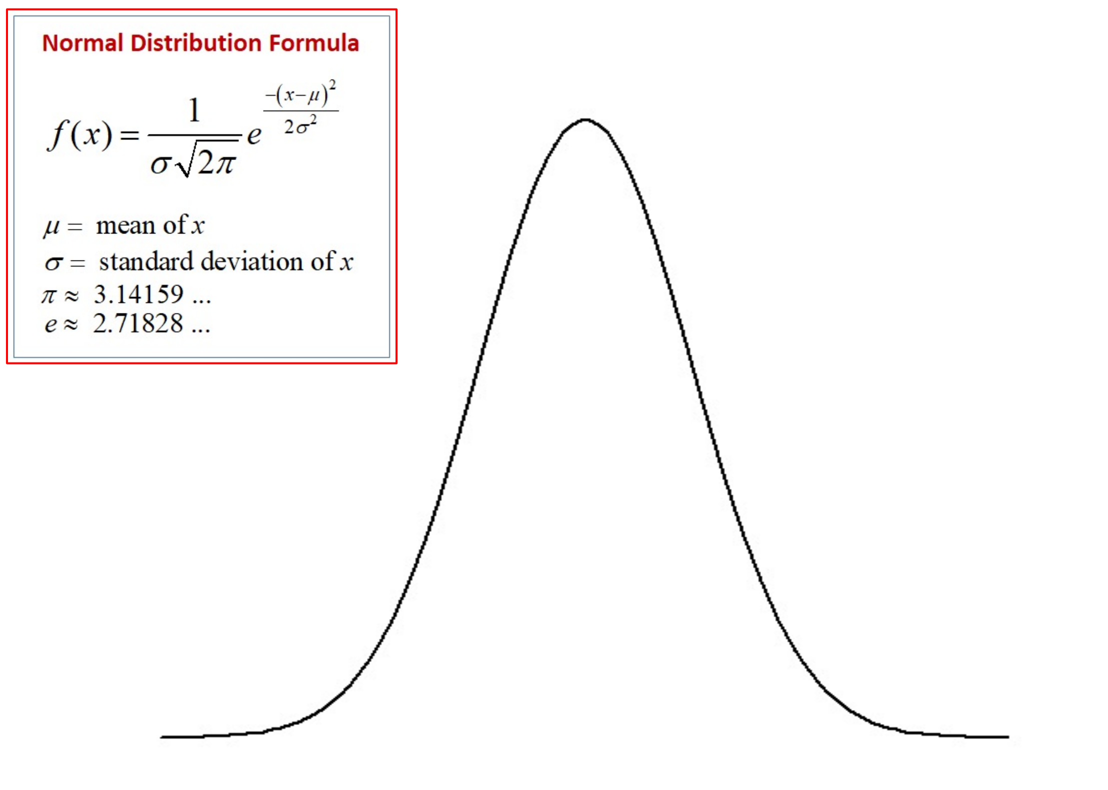
LL <- pop_mean – pop_sd

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)



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

Adding x-axis values and mean in the curve:

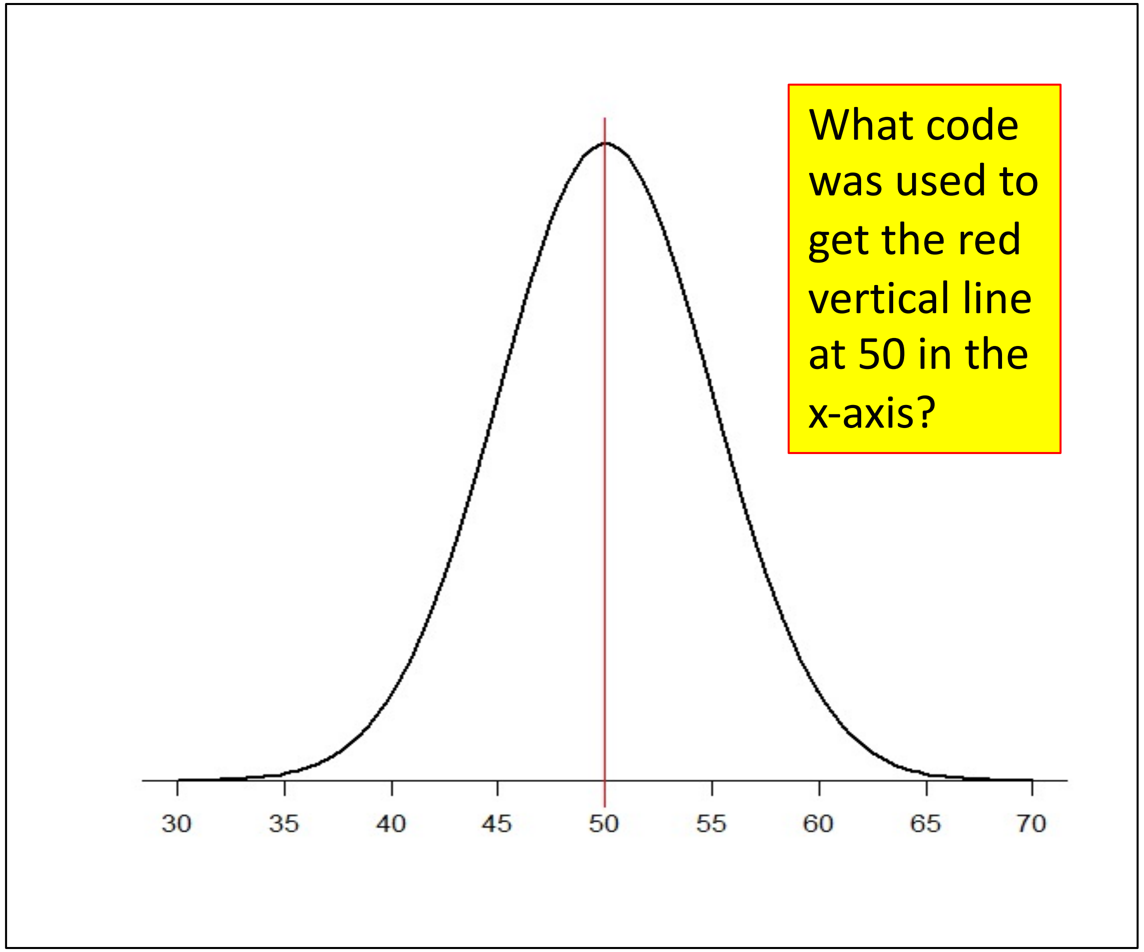
plot(x,y, type=“l", lwd=2, axes=F, xlab="", ylab="")

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(??)



Assignment 1:

- Get this graph and **provide annotation in** it as follows:

- 45-55: mean ± 1SD = 67% data
- 40-60: mean ± 2SD = 95% data
- 35-65: mean ± 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?

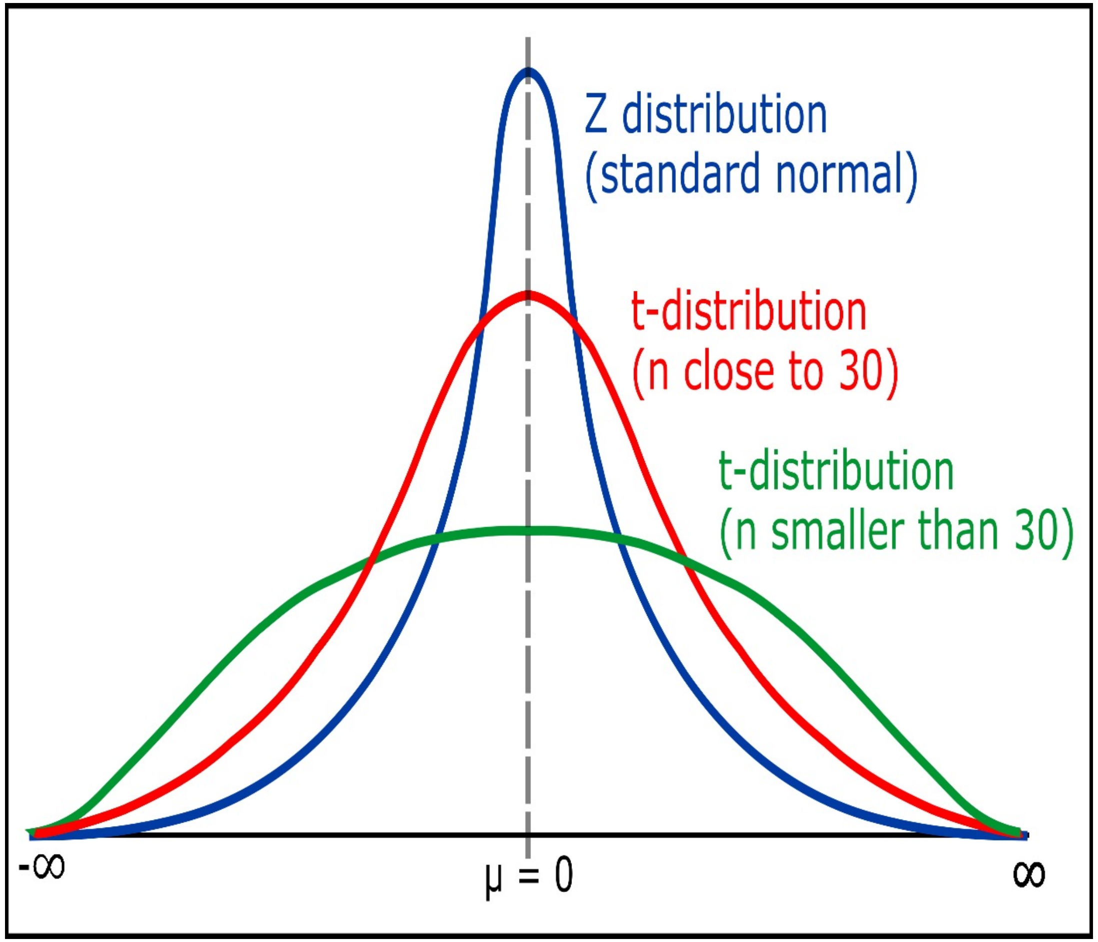
- e.g. comparing age by sex as sex variable normally has two categories “male” and “female” if age is not normally distributed

- Kruskall-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:

- 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 1-way ANOVA!**

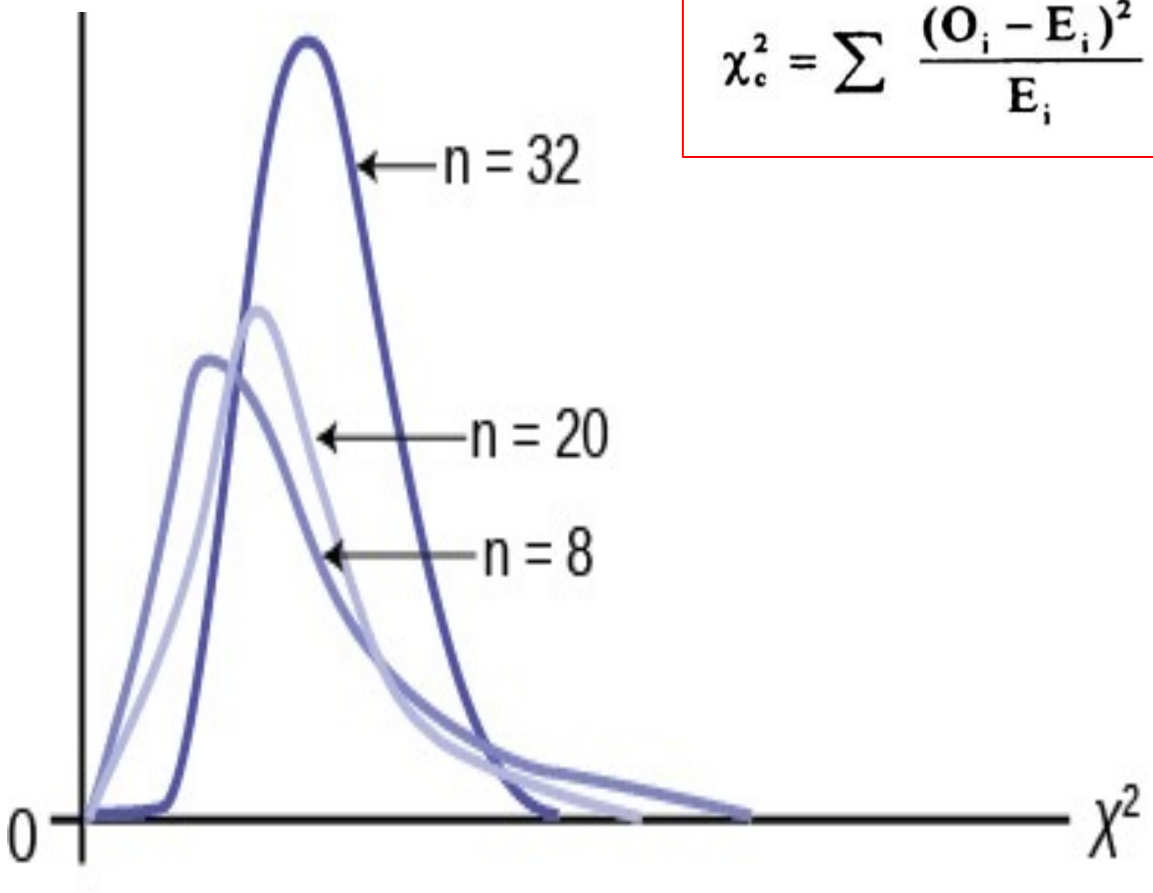


Chi-square and Z distributions:

- 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-test and comparing proportions across categories!

- As the sample size increases chi-square distribution also behaves like normal distribution

- **Logistic regression is extension of <u>chi-square test</u>!**



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

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!

Test of normality: Key point of this lecture! (Goodness-Of-Fit with Chi-square variants):

- This is a goodness-of-fit test for comparing data against the normal distribution

- Most widely used tests are:
  - Jarque-Bera test
  - **Kolmogorov-Smirnov test (large samples i.e. n>100)**
  - **Shapiro-Wilk test (Small samples)**
  - Anderson-Darlington test etc.

- Test of normality is assessed:

- Graphically (suggestive):
  - Stem-leaf plot
  - Histogram
  - Q-Q plot

- Test (confirmative):
  - ?? (depends on sample size!)

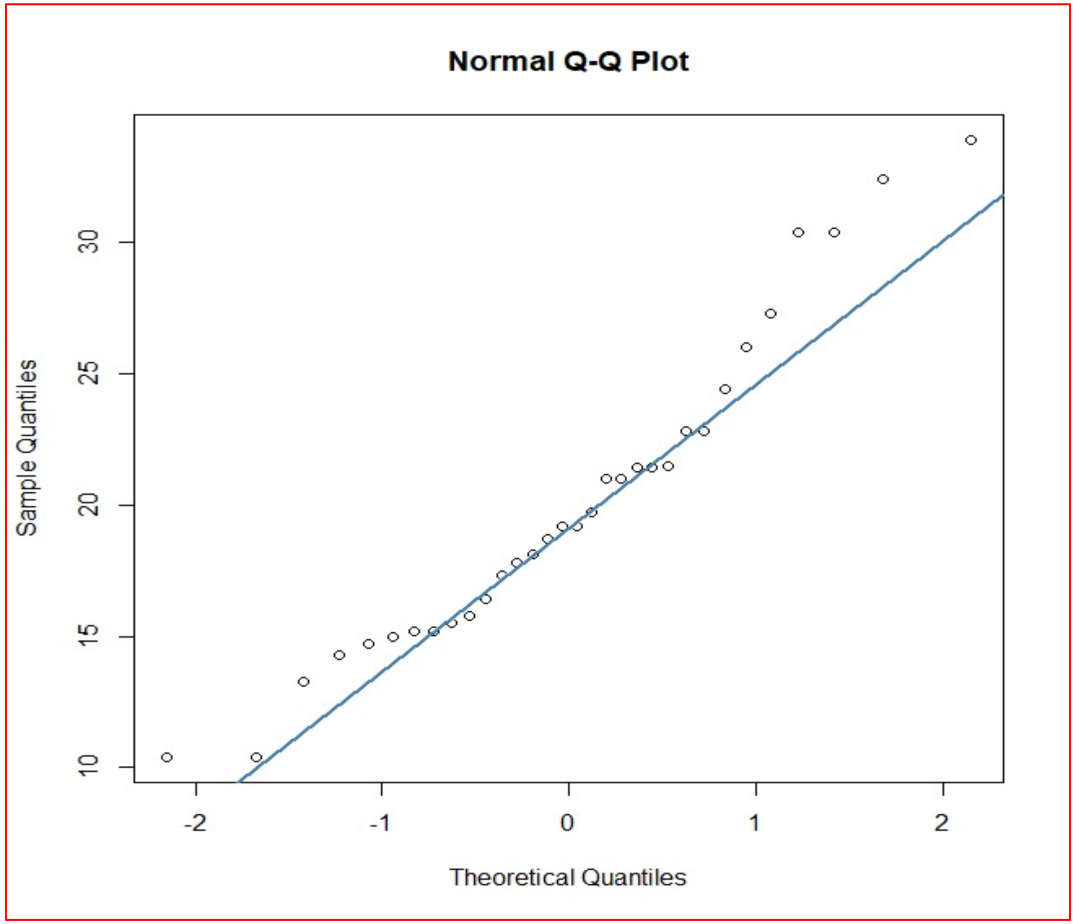
Assignment 2: Statistical tests are “robust”!

- 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-value = 0.1229



H_0: Data follows normal distribution ($p>0.05$)
H_1: Data do not follow normal distribution ($p<=0.05$)

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
- 1-way ANOVA …

Thank you!

@shitalbhandary