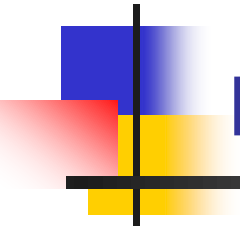
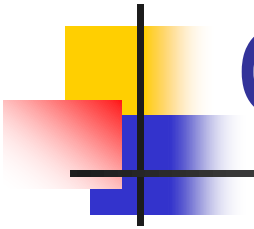
# Introduction to Data Science CS61 June 12 - July 12, 2018



Dr. Ash Pahwa

Lesson 4: Statistics

Lesson 4.2: Normal Distribution



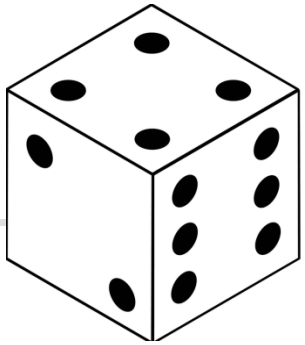
# Outline

- Distributions
- Probability Distribution Function
- Distributions: Uniform, Normal
- Properties of Normal Distribution
- Standard Normal Distribution
- Testing Normality

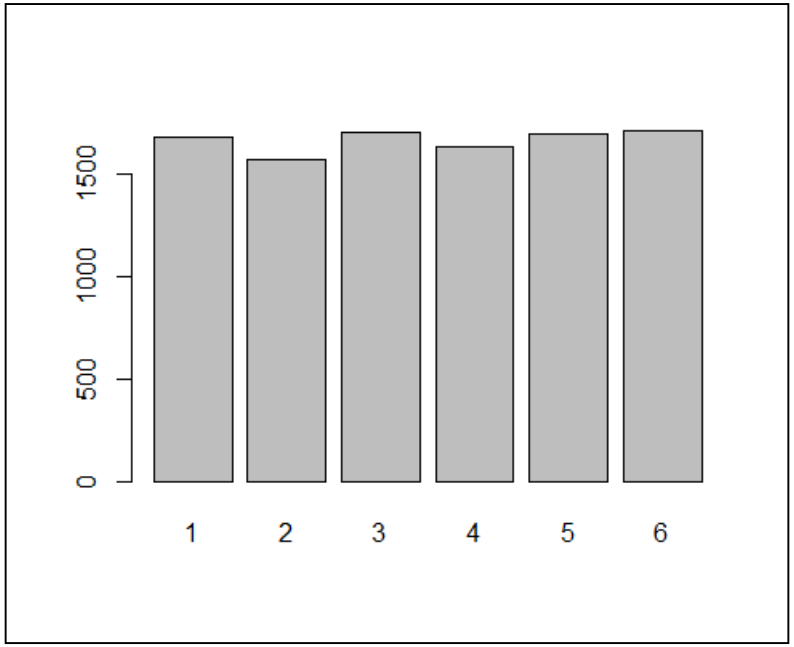
# Distribution

- Histogram of sample space
- X-axis
  - All possible values in a sample space
- Y-axis
  - Frequency of the sample value

# Distribution



- Toss die 10,000 times
- X-axis
  - Sample space
    - 1,2,3,4,5,6
- Y-axis
  - Frequency of each number



# Example: Discrete Data

- Number of arrivals at a restaurant within 15 minutes period
- Total 40 observations

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 7 | 6 | 6 | 6 | 4 | 6 | 2 | 6 |
| 5 | 6 | 6 | 11 | 4 | 5 | 7 | 6 |
| 2 | 7 | 1 | 2 | 4 | 8 | 2 | 6 |
| 6 | 5 | 5 | 3 | 7 | 5 | 4 | 6 |
| 2 | 2 | 9 | 7 | 5 | 9 | 8 | 5 |

| | Number of customers came in this period |
|---|---|
| 7:00 AM – 7:15 AM | 7 |
| 7:15 AM – 7:30 AM | 6 |
| … | |

sort(arrival)
1 2 2 2 2 2 2 3 4 4 4 4 5 5 5 5 5 5 5 6 6 6 6 6 6 6 6 6 6
6 7 7 7 7 7 8 8 9 9 11

# Histogram: Discrete Data

Raw Data

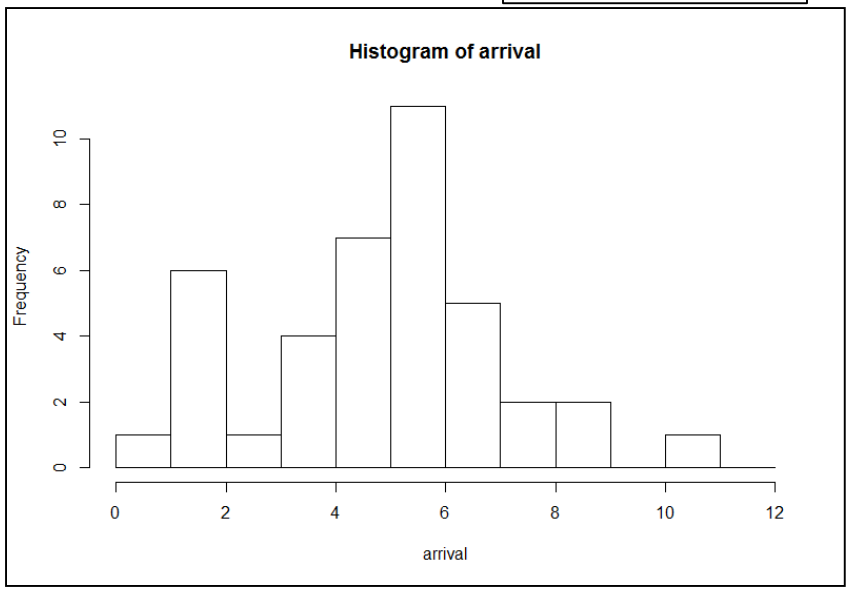
| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 7 | 6 | 6 | 6 | 4 | 6 | 2 | 6 |
| 5 | 6 | 6 | 11 | 4 | 5 | 7 | 6 |
| 2 | 7 | 1 | 2 | 4 | 8 | 2 | 6 |
| 6 | 5 | 5 | 3 | 7 | 5 | 4 | 6 |
| 2 | 2 | 9 | 7 | 5 | 9 | 8 | 5 |

sort(arrival)

1 2 2 2 2 2 2 3 4 4 4 4 5 5 5 5 5 5 5 6 6 6 6 6 6 6 6 6 6 6 6 7 7 7 7 7 8 8 9 9 11

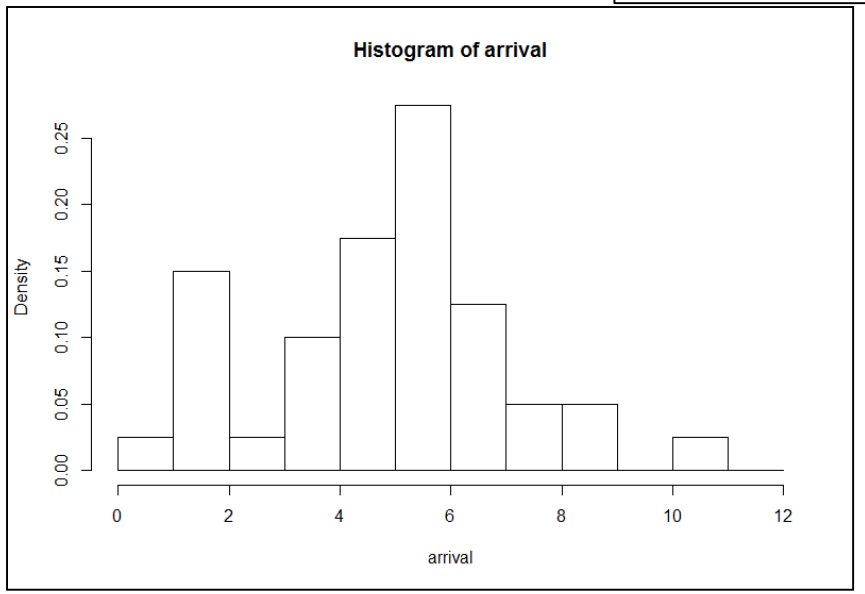
Y-axis = Frequency

Bin size = 1




Y-axis = Relative frequency

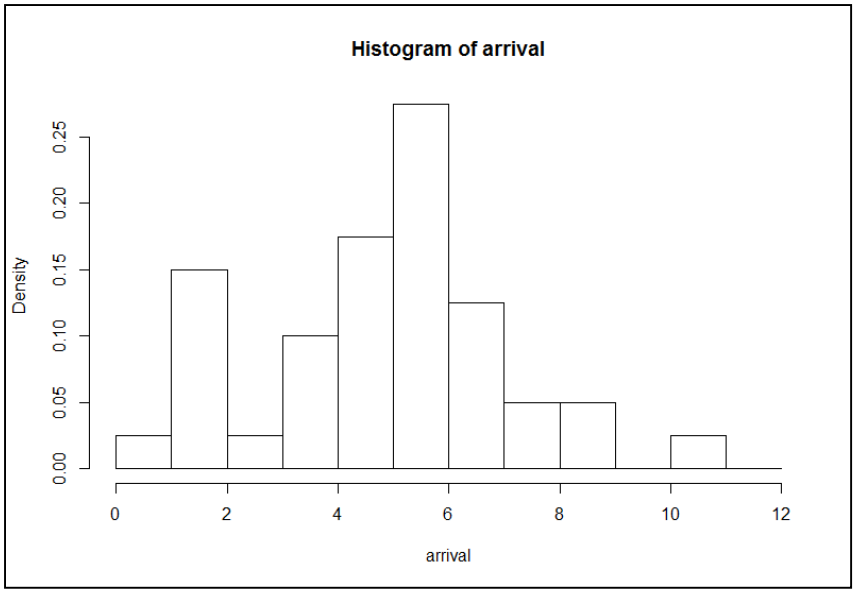
Bin size = 1



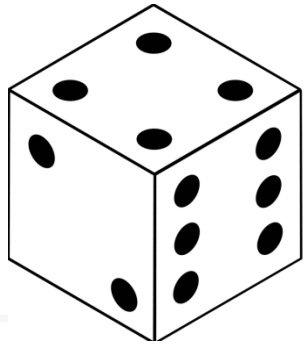

$$Relative\ Frequency = \frac{Frequency}{Sum\ of\ all\ frequencies}$$

# Probability Distribution Function (PDF)

- The relative frequency histogram is also called
  - Probability Distribution Function (PDF)
- The area of a PDF is always 1

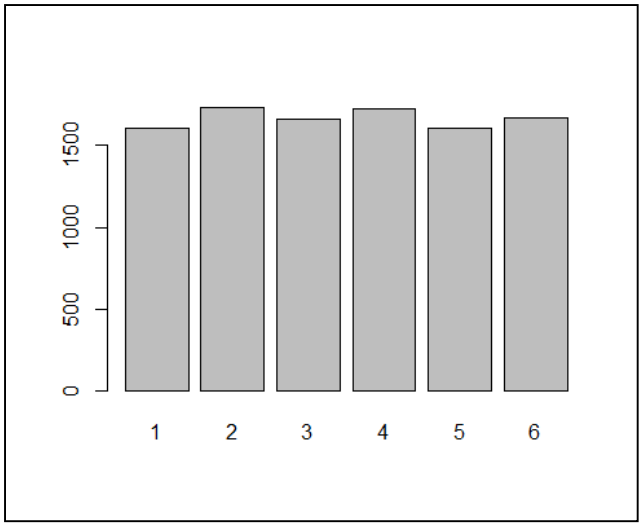


# Uniform Distribution Function

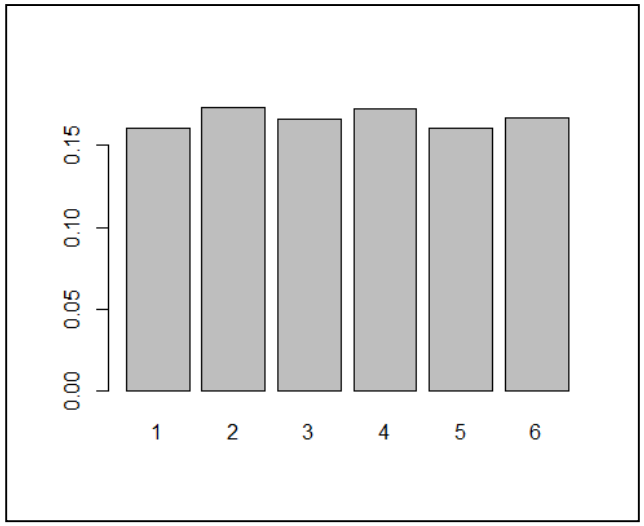


- When the probability of all possible events in the sample space is same
  - Uniform Distribution Function

Histogram




Probability Distribution Function (PDF)



# Distributions

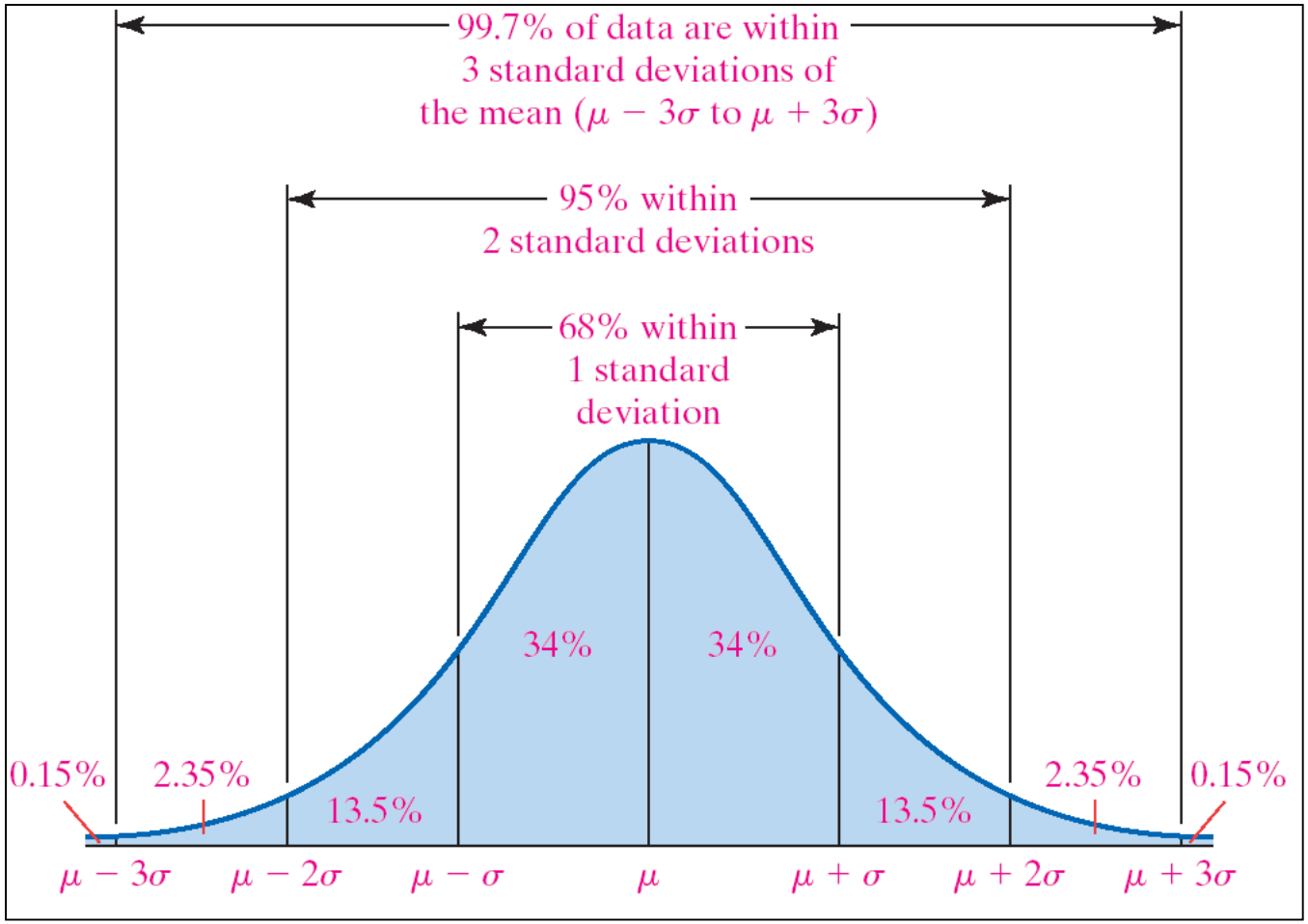
## Continuous distributions

- Uniform
- Normal
- Chi Square
- Fisher's F
- Student's t
- Gamma
- Exponential
- Beta
- Cauchy
- Lognormal
- Logistics
- Weibull

## Discrete distributions

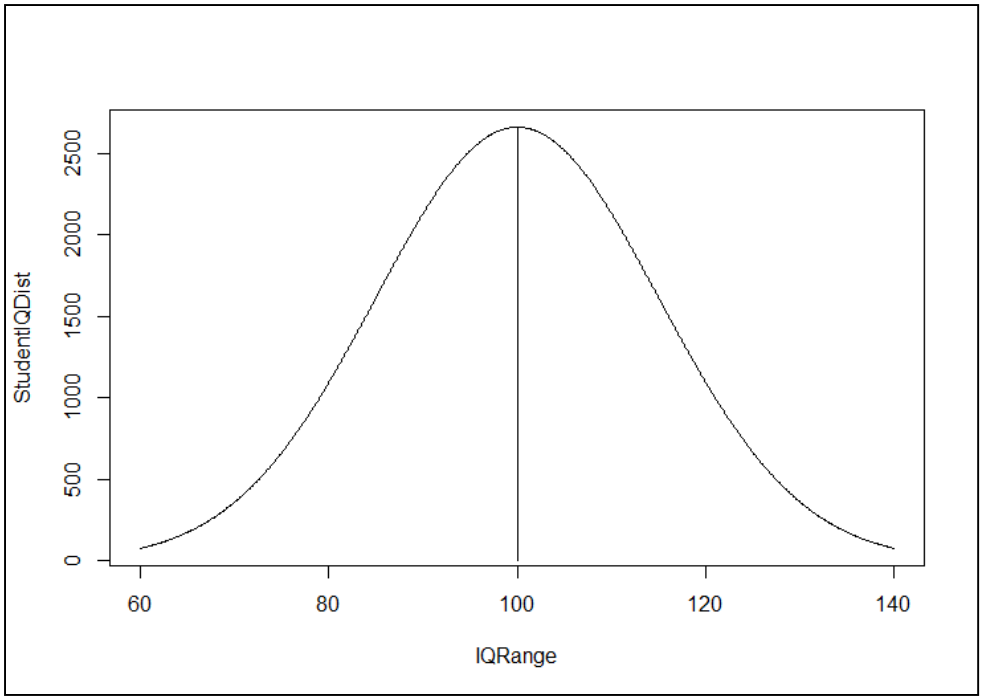
- Binomial
- Poisson
- Hypergeometric
- Negative binomial
- Wilcox

# The Empirical Rule
# Normal Distribution



# Data: Student IQ Test Population

- Population
  - Student IQ Test
  - Normally Distributed
  - Mean = 100
  - Standard Deviation = 15
  - 68% of the data will be within
    - 100 – 15 = 85
    - 100 + 15 = 115

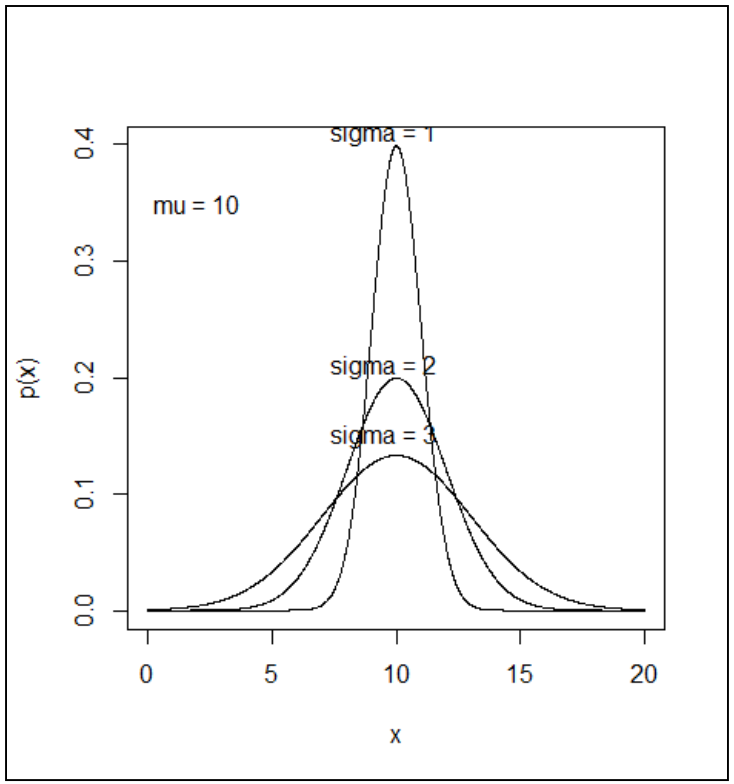


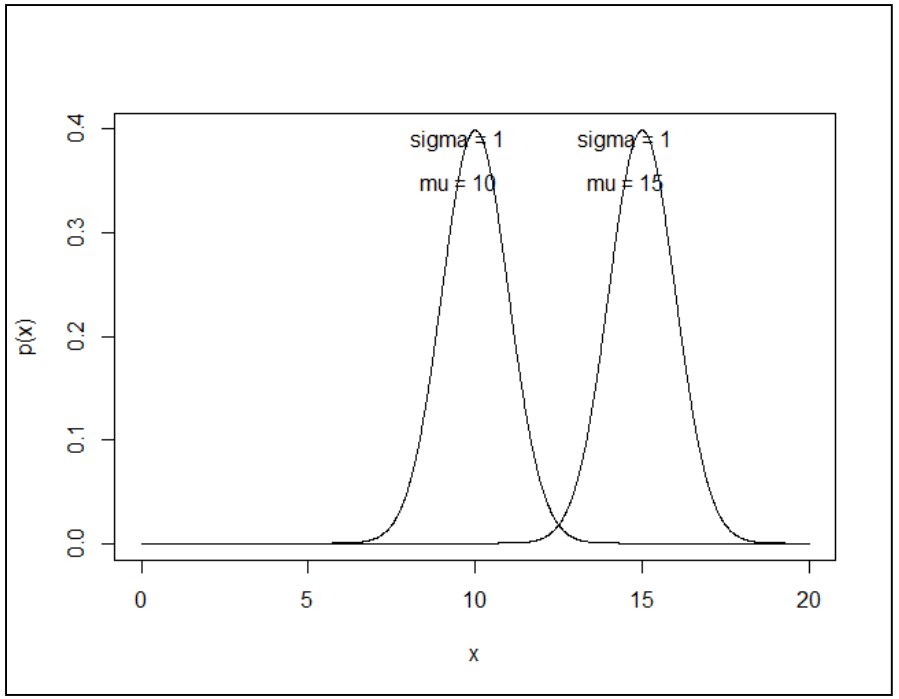
# Normal Distribution

$\mu = \text{Mean}$

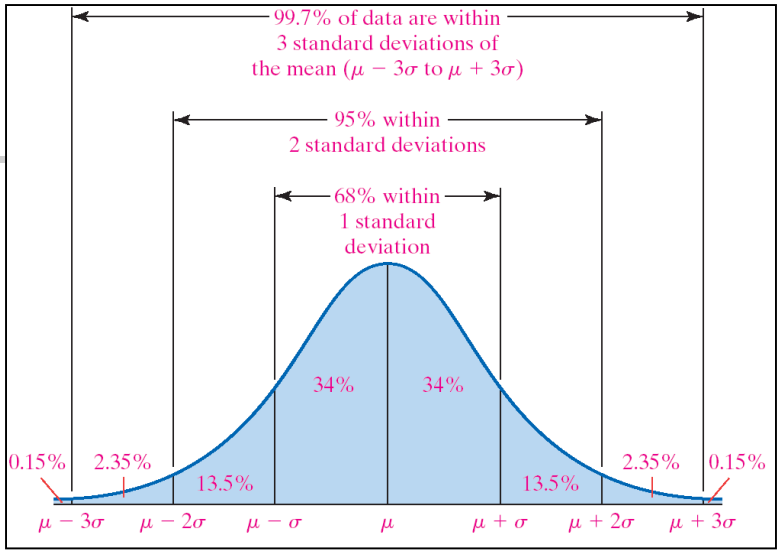
$\sigma = \text{Standard Deviation}$

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

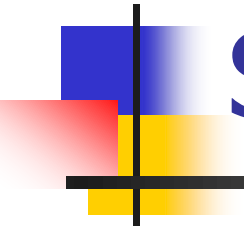


# Properties of Normal Distribution



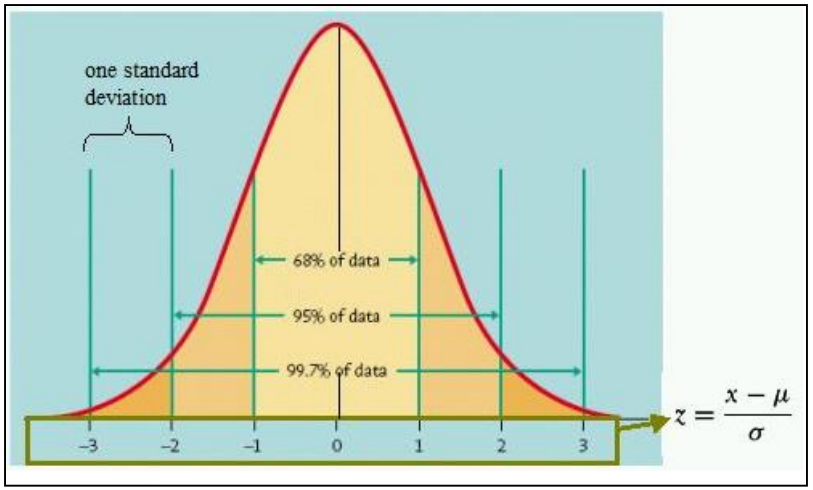

- Symmetric about its mean $\mu$
- Mean = Median = Mode
  - Single peak at $x = \mu$
- Inflection point at $\mu - \sigma \; and \; \mu + \sigma$
- Area under the curve = 1
- Area of left ( $mean \; \mu$ ) =
  - Area of right = ½
- Follows the Empirical Rule



# Standard Normal Distribution

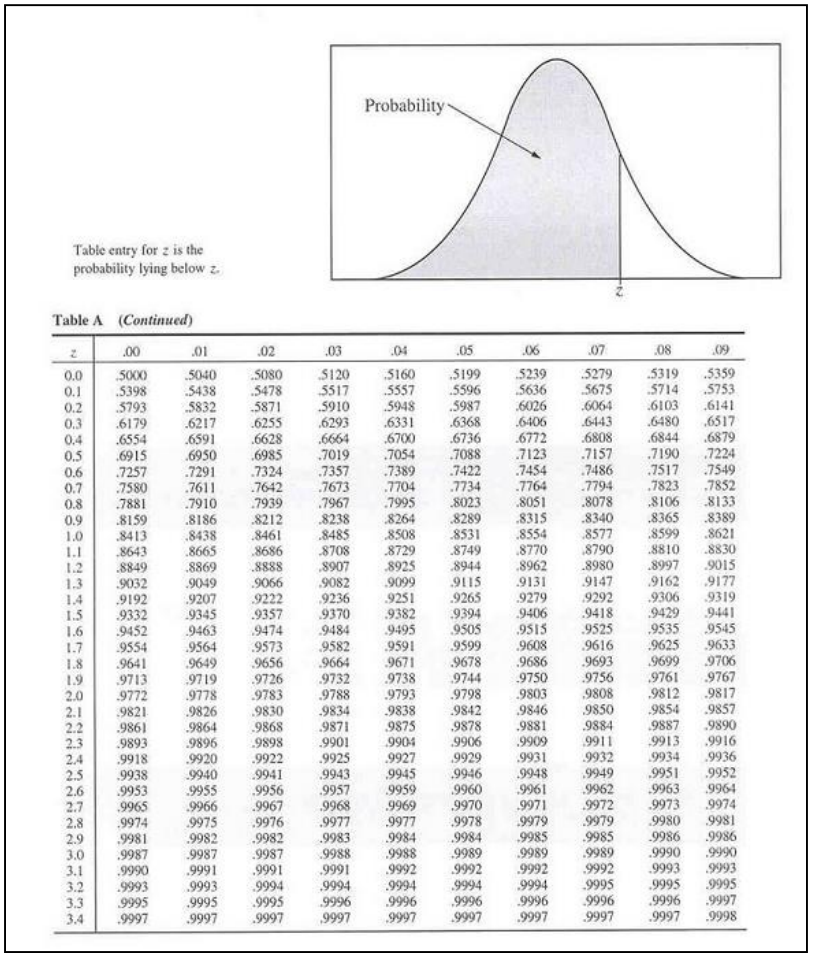
# Standard Normal Curve $\mu = 0, \sigma = 1$

- Symmetric about its mean $\mu = 0, \sigma = 1$
- Mean = Median = Mode
  - Single peak at $z = 0$
- Inflection point at $-1\ and\ +1$
- Area under the curve = 1
- Area of left ( $mean\ \mu = 0$) =
  - Area of right = ½
- Follows the Empirical Rule



# Standard Normal Table

- Tables will help you find the area under the Standard Normal Curve to the LEFT of a z-score

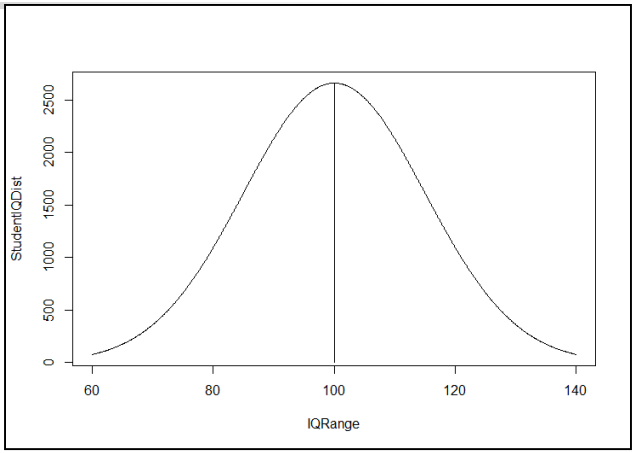



| z | .00 | .01 | .02 | .03 | .04 | .05 | .06 | .07 | .08 | .09 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.0 | .5000 | .5040 | .5080 | .5120 | .5160 | .5199 | .5239 | .5279 | .5319 | .5359 |
| 0.1 | .5398 | .5438 | .5478 | .5517 | .5557 | .5596 | .5636 | .5675 | .5714 | .5753 |
| 0.2 | .5793 | .5832 | .5871 | .5910 | .5948 | .5987 | .6026 | .6064 | .6103 | .6141 |
| 0.3 | .6179 | .6217 | .6255 | .6293 | .6331 | .6368 | .6406 | .6443 | .6480 | .6517 |
| 0.4 | .6554 | .6591 | .6628 | .6664 | .6700 | .6736 | .6772 | .6808 | .6844 | .6879 |
| 0.5 | .6915 | .6950 | .6985 | .7019 | .7054 | .7088 | .7123 | .7157 | .7190 | .7224 |
| 0.6 | .7257 | .7291 | .7324 | .7357 | .7389 | .7422 | .7454 | .7486 | .7517 | .7549 |
| 0.7 | .7580 | .7611 | .7642 | .7673 | .7704 | .7734 | .7764 | .7794 | .7823 | .7852 |
| 0.8 | .7881 | .7910 | .7939 | .7967 | .7995 | .8023 | .8051 | .8078 | .8106 | .8133 |
| 0.9 | .8159 | .8186 | .8212 | .8238 | .8264 | .8289 | .8315 | .8340 | .8365 | .8389 |
| 1.0 | .8413 | .8438 | .8461 | .8485 | .8508 | .8531 | .8554 | .8577 | .8599 | .8621 |
| 1.1 | .8643 | .8665 | .8686 | .8708 | .8729 | .8749 | .8770 | .8790 | .8810 | .8830 |
| 1.2 | .8849 | .8869 | .8888 | .8907 | .8925 | .8944 | .8962 | .8980 | .8997 | .9015 |
| 1.3 | .9032 | .9049 | .9066 | .9082 | .9099 | .9115 | .9131 | .9147 | .9162 | .9177 |
| 1.4 | .9192 | .9207 | .9222 | .9236 | .9251 | .9265 | .9279 | .9292 | .9306 | .9319 |
| 1.5 | .9332 | .9345 | .9357 | .9370 | .9382 | .9394 | .9406 | .9418 | .9429 | .9441 |
| 1.6 | .9452 | .9463 | .9474 | .9484 | .9495 | .9505 | .9515 | .9525 | .9535 | .9545 |
| 1.7 | .9554 | .9564 | .9573 | .9582 | .9591 | .9599 | .9608 | .9616 | .9625 | .9633 |
| 1.8 | .9641 | .9649 | .9656 | .9664 | .9671 | .9678 | .9686 | .9693 | .9699 | .9706 |
| 1.9 | .9713 | .9719 | .9726 | .9732 | .9738 | .9744 | .9750 | .9756 | .9761 | .9767 |
| 2.0 | .9772 | .9778 | .9783 | .9788 | .9793 | .9798 | .9803 | .9808 | .9812 | .9817 |
| 2.1 | .9821 | .9826 | .9830 | .9834 | .9838 | .9842 | .9846 | .9850 | .9854 | .9857 |
| 2.2 | .9861 | .9864 | .9868 | .9871 | .9875 | .9878 | .9881 | .9884 | .9887 | .9890 |
| 2.3 | .9893 | .9896 | .9898 | .9901 | .9904 | .9906 | .9909 | .9911 | .9913 | .9916 |
| 2.4 | .9918 | .9920 | .9922 | .9925 | .9927 | .9929 | .9931 | .9932 | .9934 | .9936 |
| 2.5 | .9938 | .9940 | .9941 | .9943 | .9945 | .9946 | .9948 | .9949 | .9951 | .9952 |
| 2.6 | .9953 | .9955 | .9956 | .9957 | .9959 | .9960 | .9961 | .9962 | .9963 | .9964 |
| 2.7 | .9965 | .9966 | .9967 | .9968 | .9969 | .9970 | .9971 | .9972 | .9973 | .9974 |
| 2.8 | .9974 | .9975 | .9976 | .9977 | .9977 | .9978 | .9979 | .9979 | .9980 | .9981 |
| 2.9 | .9981 | .9982 | .9982 | .9983 | .9984 | .9984 | .9985 | .9985 | .9986 | .9986 |
| 3.0 | .9987 | .9987 | .9987 | .9988 | .9988 | .9989 | .9989 | .9989 | .9990 | .9990 |
| 3.1 | .9990 | .9991 | .9991 | .9991 | .9992 | .9992 | .9992 | .9992 | .9993 | .9993 |
| 3.2 | .9993 | .9993 | .9994 | .9994 | .9994 | .9994 | .9994 | .9995 | .9995 | .9995 |
| 3.3 | .9995 | .9995 | .9995 | .9996 | .9996 | .9996 | .9996 | .9996 | .9996 | .9997 |
| 3.4 | .9997 | .9997 | .9997 | .9997 | .9997 | .9997 | .9997 | .9997 | .9997 | .9998 |

# Normalizing Data: Computing z values in Excel

$$z = \frac{\text{Data Value - Mean}}{\text{Standard Deviation}} = \frac{y - \mu}{\sigma}$$

- Given : Data = 110
- Compute : z value




Clipboard | Font

B6 | =STANDARDIZE(B5,B2,B3)

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | | | | | |
| 2 | **Mean** | 100 | | | |
| 3 | **Standard Deviation** | 15 | | | |
| 4 | | | | | |
| 5 | Score | 110 | | | |
| 6 | z-value | 0.666667 | | | |
| 7 | | | | | |

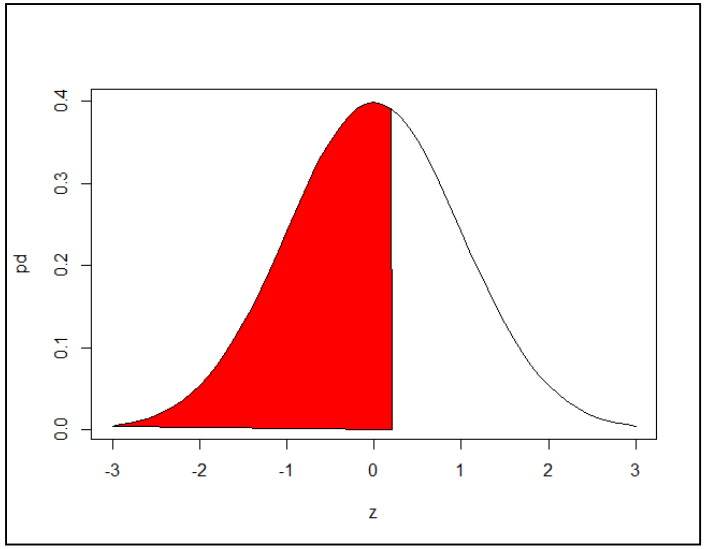
- Population
  - Student IQ Test
  - Normally Distributed
  - Mean = 100
  - Standard Deviation = 15

# Standard Normal Table Excel: $P(z)\ x < z$

- Given : z value = 0.21
- Compute 'Left' Area (Probability) under the Standard Normal Curve

B5 | =NORMSDIST(B4)

| | A | B |
|---|---|---|
| 1 | | |
| 2 | | |
| 3 | Shaded to the left | |
| 4 | z | 0.21 |
| 5 | P(z) Probability z < 0.21 | 0.5832 |
| 6 | | |

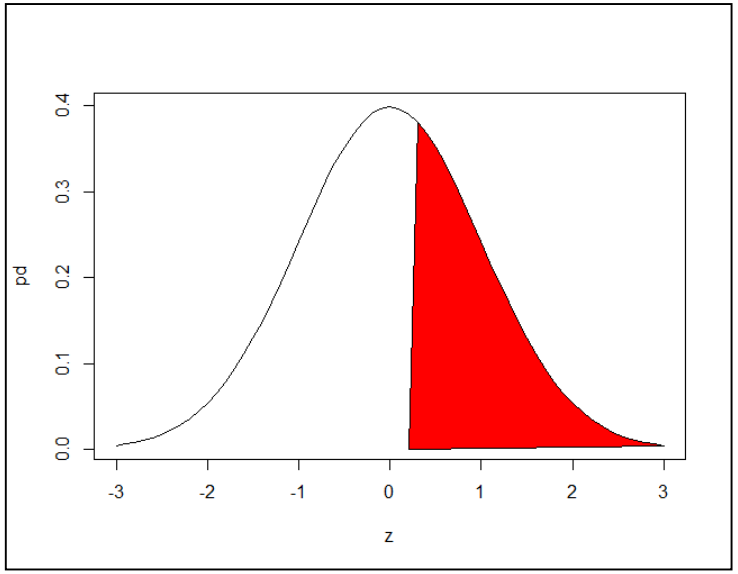


# Standard Normal Table Excel: $P(z)\ x > z$

- Given : z value = 0.21
- Compute 'Right' Area (Probability) under the Standard Normal Curve

B9 | =1-NORMSDIST(B8)

| | A | B | C |
|---|---|---|---|
| 1 | | | |
| 2 | | | |
| 3 | Shaded to the left | | |
| 4 | z | 0.21 | |
| 5 | P(z) Probability z < 0.21 | 0.5832 | |
| 6 | | | |
| 7 | Shaded to the right | | |
| 8 | z | 0.21 | |
| 9 | P(z) Probability z > 0.21 | 0.4168 | |
| 10 | | | |

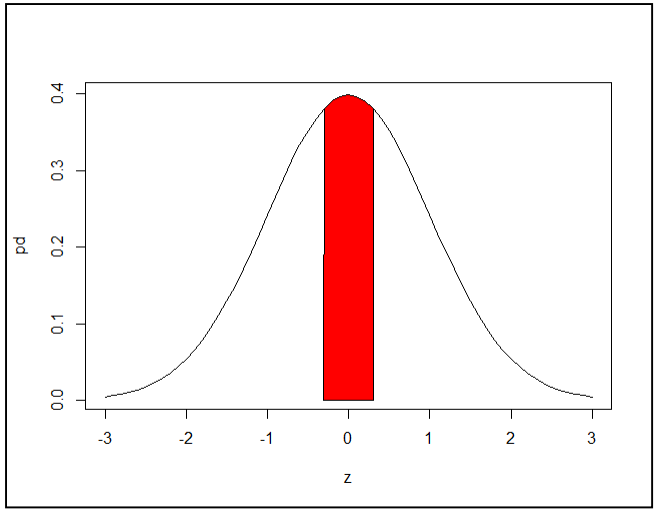


# Standard Normal Table Excel: $P(z)\ x_1 < z < x_2$

- Given : z values (z1 and z2)
- Compute 'In Between' Area (Probability) under the Standard Normal Curve

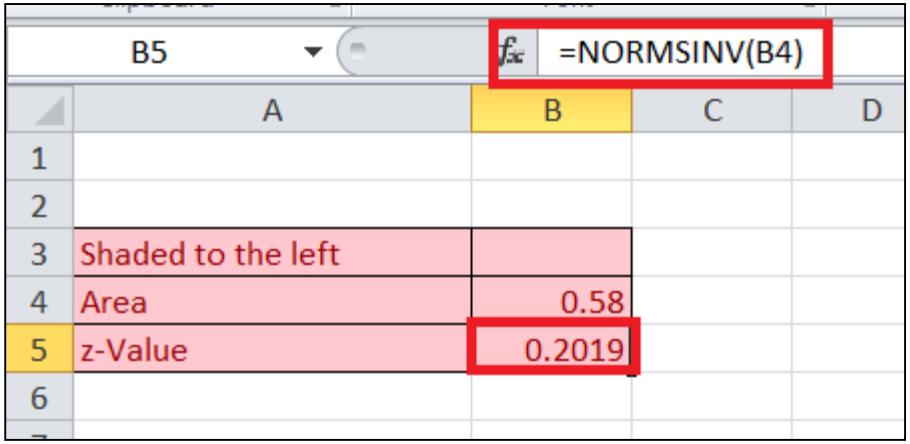
B14 =NORMSDIST(B12)-NORMSDIST(B13)

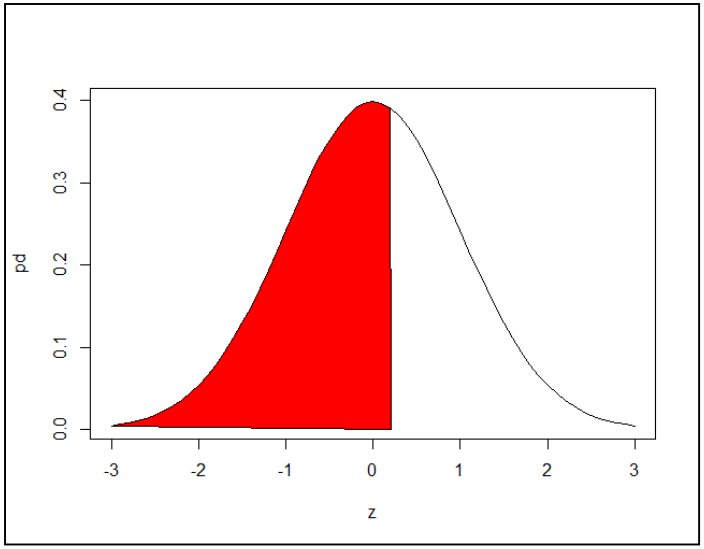
| | A | B | C | D |
|---|---|---|---|---|
| 1 | | | | |
| 2 | | | | |
| 3 | Shaded to the left | | | |
| 4 | z | 0.21 | | |
| 5 | P(z) Probability z < 0.21 | 0.5832 | | |
| 6 | | | | |
| 7 | Shaded to the right | | | |
| 8 | z | 0.21 | | |
| 9 | P(z) Probability z > 0.21 | 0.4168 | | |
| 10 | | | | |
| 11 | Shaded in between | | | |
| 12 | z1 | 0.31 | | |
| 13 | z2 | -0.31 | | |
| 14 | P(z) Probability -0.31 < z <0.31 | 0.2434 | | |
| 15 | | | | |



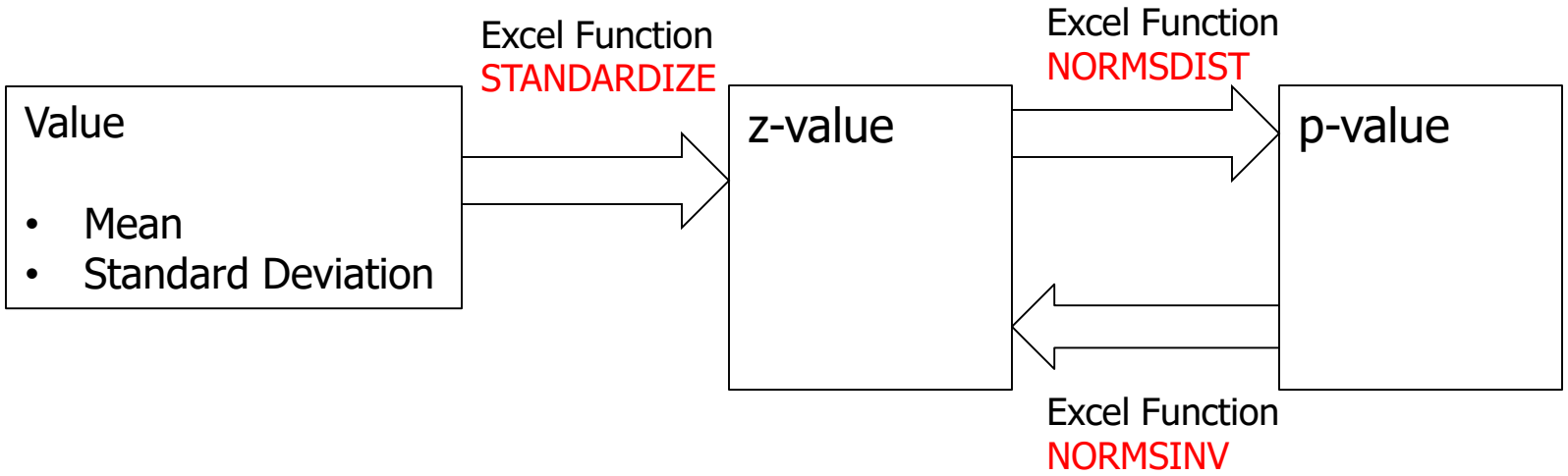
# Standard Normal Table Excel: *Inverse Function*

- Given 'Left' Area (Probability) under the Standard Normal Curve = 0.58
- Compute : z value



# Excel Functions

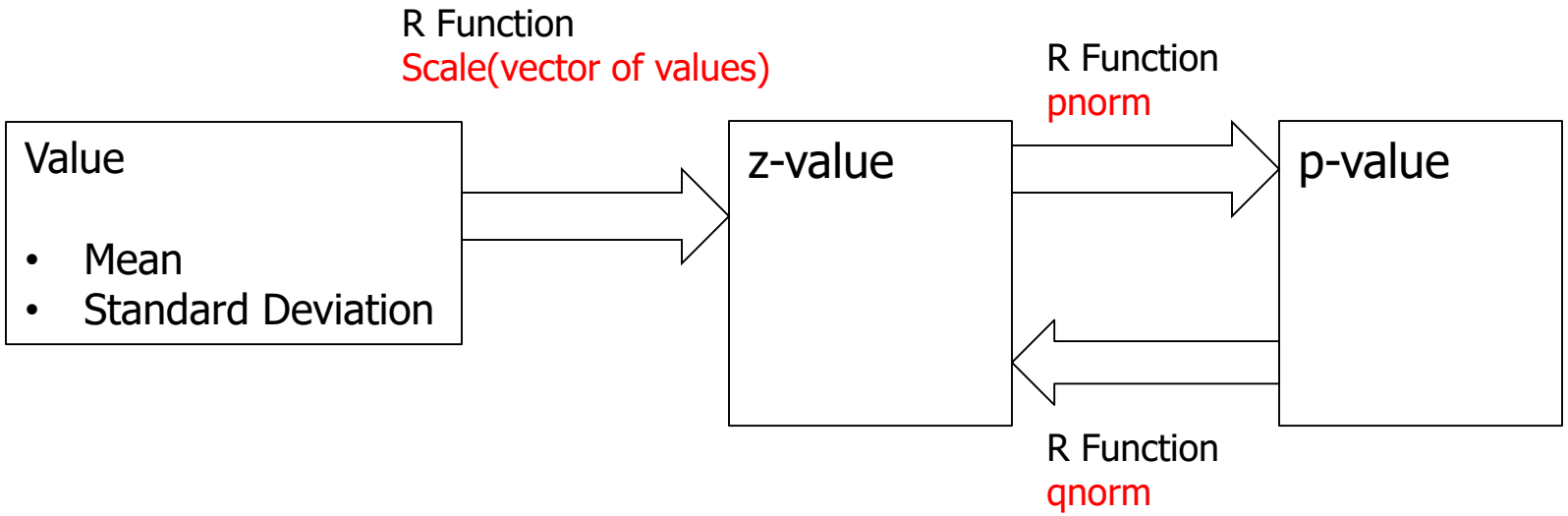


# R Functions for Standard Normal Distributions

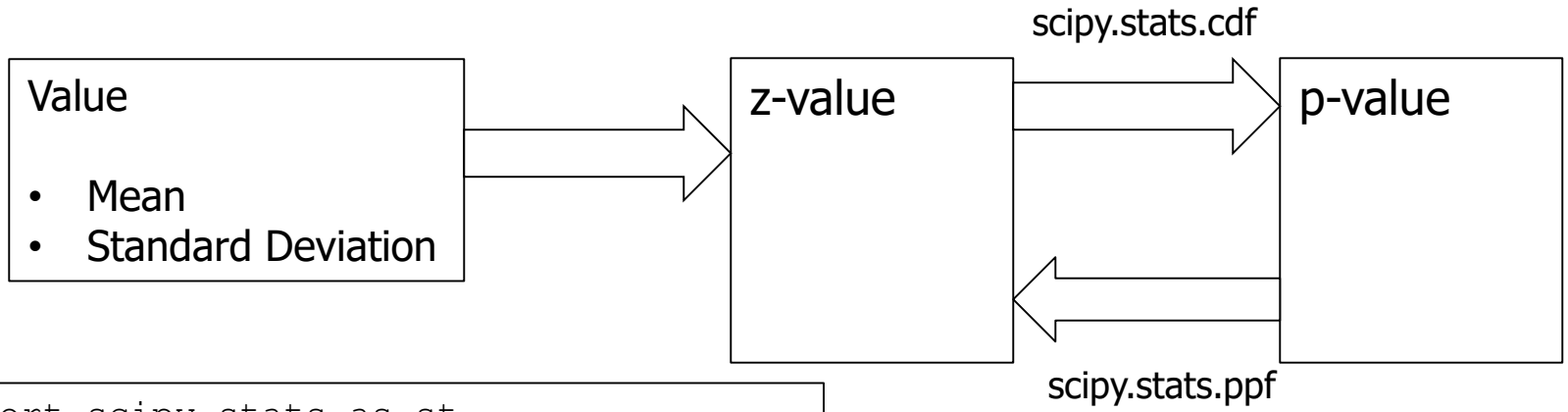
| Distribution | Normal |
|---|---|
| Density | dnorm() |
| Cumulative density Function (cdf) | pnorm() |
| Quantiles | qnorm() |
| Random Numbers | rnorm() |

```
> m = 100
> s = 15
> d = 110
> (zValue <- (d - m)/s)
[1] 0.6666667
> ###################################
> z = 0.21
> (pnorm(z))
[1] 0.5831662
> ###################################
> (1-pnorm(z))
[1] 0.4168338
> ###################################
> z1 = -0.31
> z2 = 0.31
> (pnorm(z2) - pnorm(z1))
[1] 0.243439
> ###################################
> a = 0.58
> (qnorm(a))
[1] 0.2018935
```

# R Functions

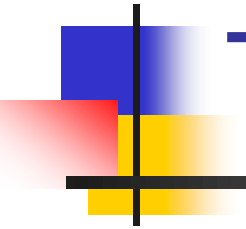

R Function
Scale(vector of values)
R Function
pnorm
Value
• Mean
• Standard Deviation
z-value
p-value
R Function
qnorm

# Python Functions
# Package: scipy.stats




```
import scipy.stats as st
#####################################
# Convert z-value of 1.64 into p-value
st.norm.cdf(1.64)
Out[7]: 0.94949741652589625

#####################################
# Convert  p-value of 0.95 into z-value
st.norm.ppf(0.95)
Out[11]: 1.6448536269514722
```

```
cdf: Cumulative Distribution Function

ppf: Percentage Point Function
```



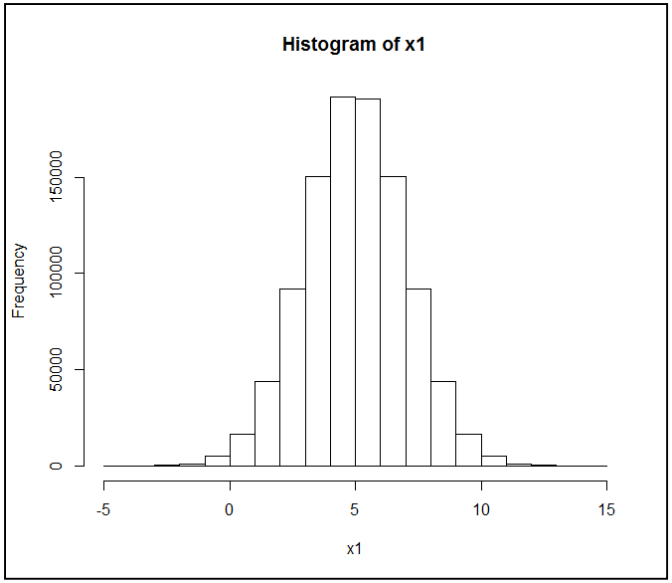
# Testing Normality

# Why Test Normality?

- Many statistical inference procedures assume that we are sampling from a normally distributed population
- We should develop a test to check if data is normally distributed
- Example:
  - The residuals of a regression should be normally distributed
  - The residuals should be tested for normality

# Testing Normality: Histogram

- How to check if data is normally distributed
- Graphical technique
  - Plot the histogram of the data
  - You should see a normal distribution
- Problem with this technique
  - Histograms shape change with different bin sizes



# Testing Normality: QQ Plot: Theory

- How to check if data is normally distributed
- Analytical/Graphical technique
  - Data is plotted against a theoretical normal distribution
  - If you see a straight line
    - Data is normally distributed
  - This technique is called QQ plot – Quantile-Quantile plot

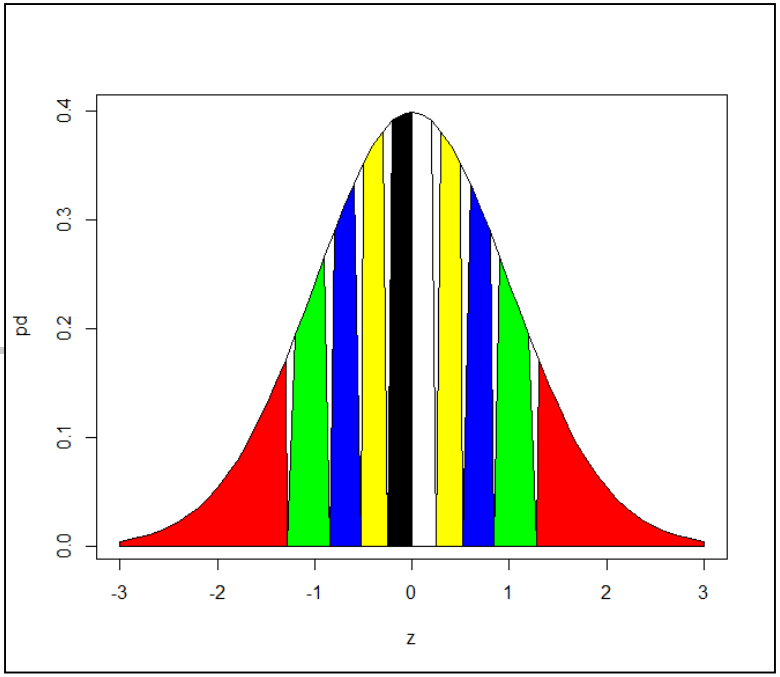
# QQ Plot – Sort the data

| | Data |
|---|---|
| 1 | 3.89 |
| 2 | 4.75 |
| 3 | 6.33 |
| 4 | 4.75 |
| 5 | 7.21 |
| 6 | 5.78 |
| 7 | 5.80 |
| 8 | 5.20 |
| 9 | 6.64 |

- Question
  - Is this data normally distributed?
- Testing Procedure
- First Sort the data
- Plot against appropriate quantiles from the standard normal distribution

| | Sorted Data |
|---|---|
| 1 | 3.89 |
| 2 | 4.75 |
| 3 | 4.75 |
| 4 | 5.20 |
| 5 | 5.78 |
| 6 | 5.80 |
| 7 | 6.33 |
| 8 | 7.21 |
| 9 | 7.90 |

# QQ Plot: z-values



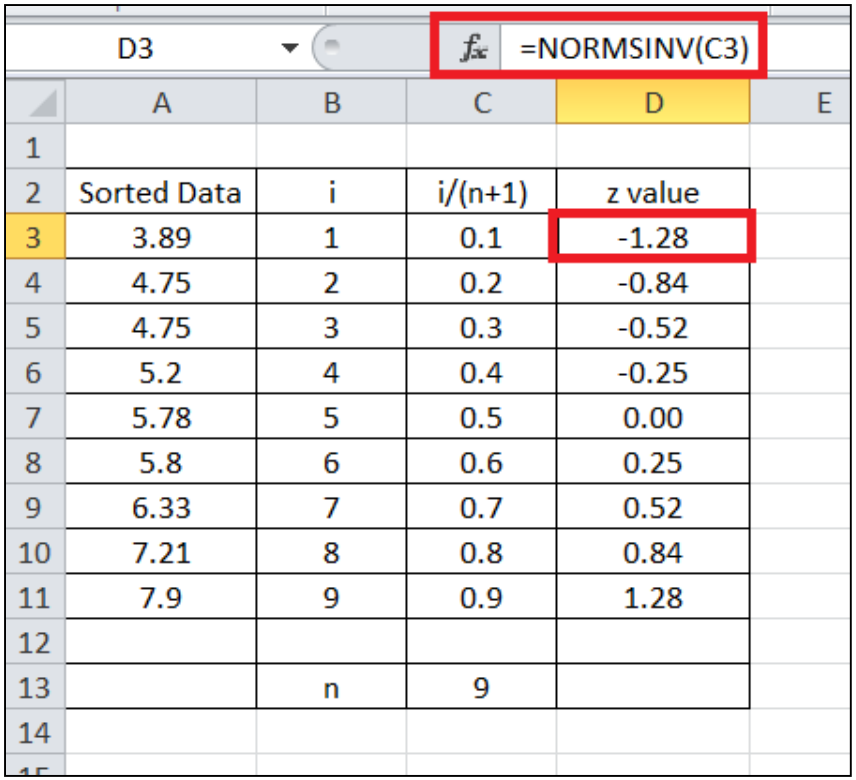

- Divide the normal distribution curve into (n+1=10) parts
- Each part represents 10% of the area
- Compute the corresponding z-values
- QQ plot is
  - X axis: z-values taken from the standard normal distribution curve
  - Y-axis: Sorted Data values

# QQ Plot: z Values

| | Data |
|---|---|
| 1 | 3.89 |
| 2 | 4.75 |
| 3 | 6.33 |
| 4 | 4.75 |
| 5 | 7.21 |
| 6 | 5.78 |
| 7 | 5.80 |
| 8 | 5.20 |
| 9 | 6.64 |

| | Sorted Data |
|---|---|
| 1 | 3.89 |
| 2 | 4.75 |
| 3 | 4.75 |
| 4 | 5.20 |
| 5 | 5.78 |
| 6 | 5.80 |
| 7 | 6.33 |
| 8 | 7.21 |
| 9 | 7.90 |

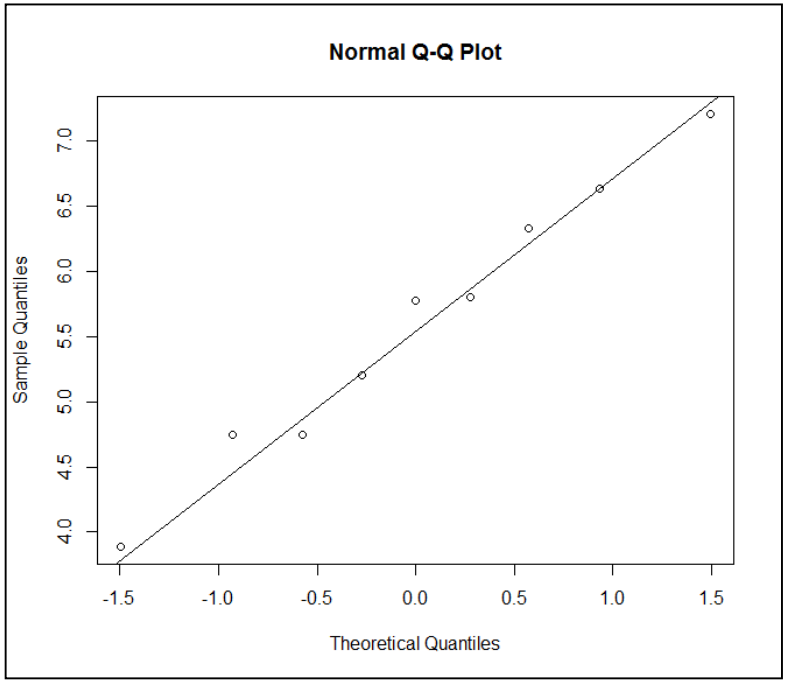
- Plot $i^{th}$ ordered value against $\frac{i}{n+1}$ quantile of Standard Normal Distribution



# QQ Plot: qqnorm, qqline

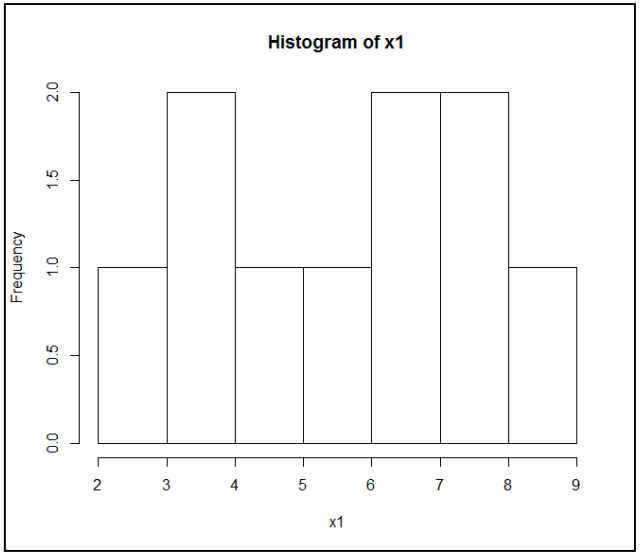
```
(x <- c(3.89,4.75,6.33,4.75,7.21,5.78,5.80,5.20,6.64))
qqnorm(x)
qqline(x)
```

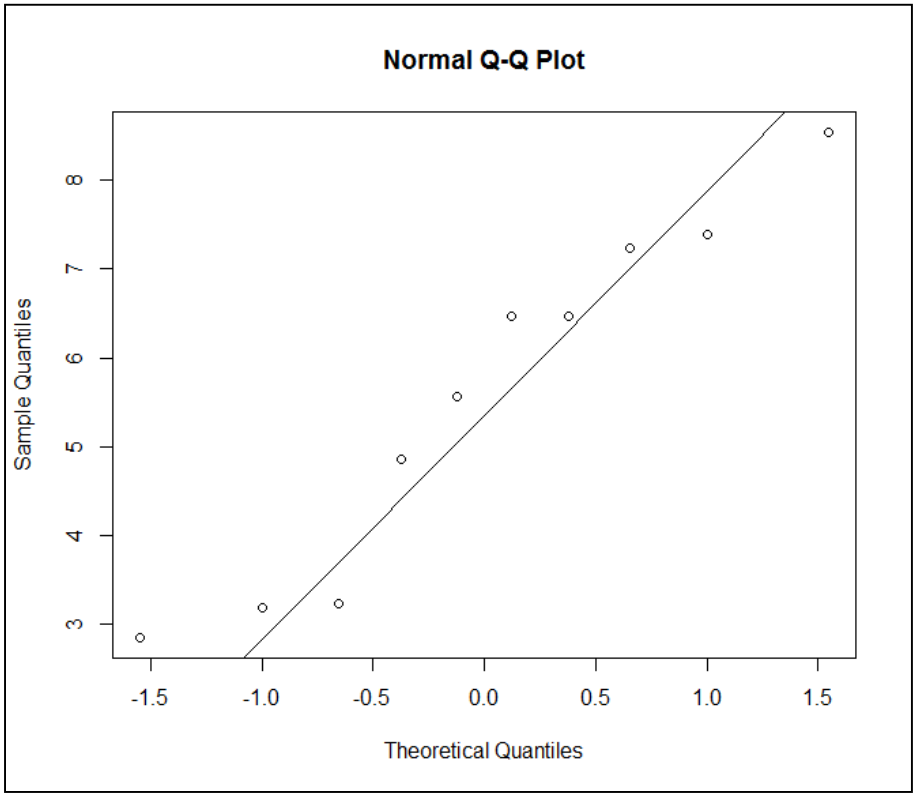
- QQ plot is
  - X axis: z-values taken from the standard normal distribution curve
  - Y-axis: Sorted Data values
- If the line is straight
  - Data is normal



# QQPlot: Known Normal Distribution: N=10

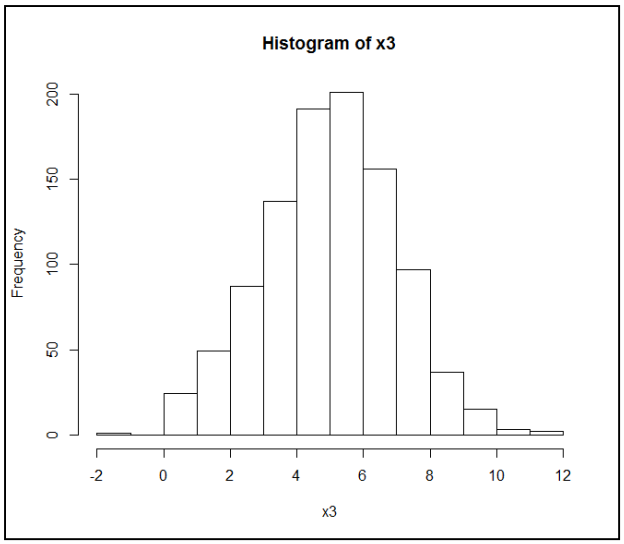
```
x1 <- rnorm(n=10,mean=5,sd=2)
hist(x1)
qqnorm(x1)
qqline(x1)
```

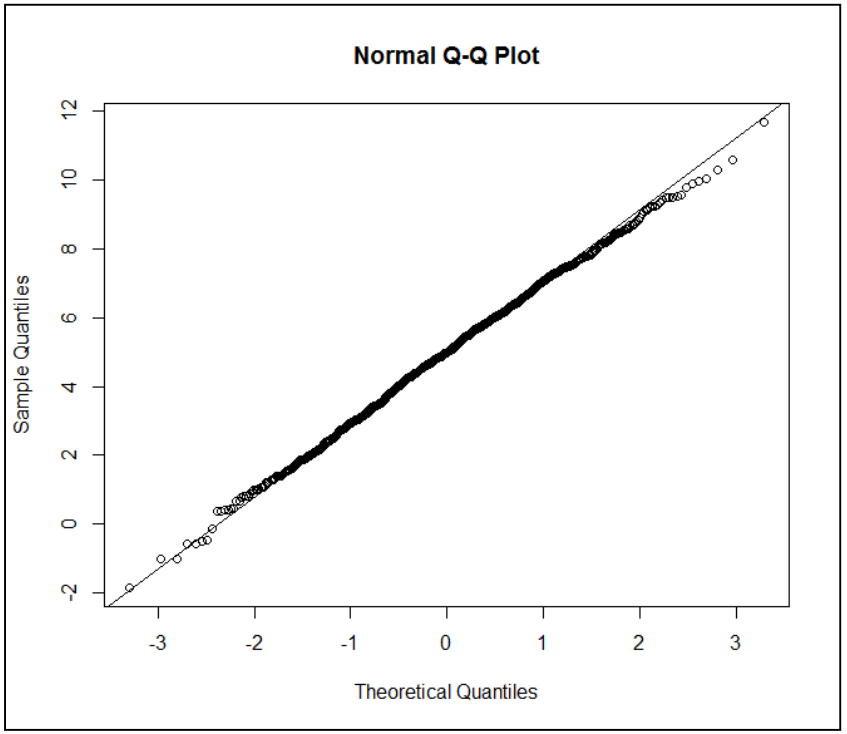


# QQPlot: Known Normal Distribution: N=1000

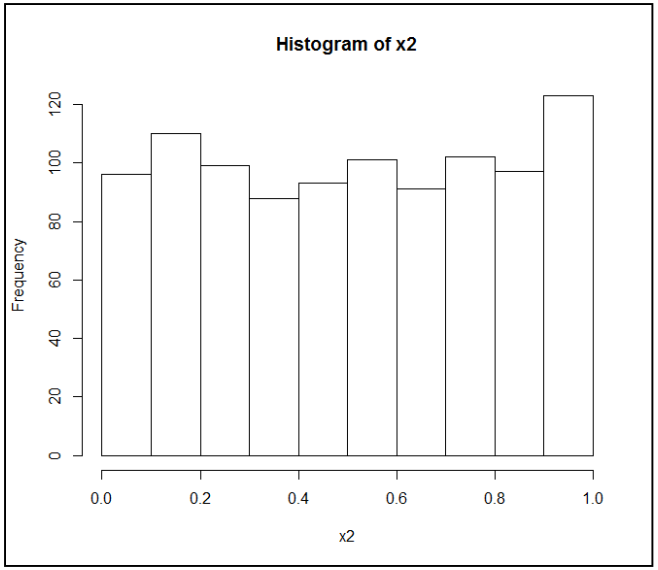
```
x3 <- rnorm(1000,mean=5,sd=2)
hist(x3)
qqnorm(x3)
qqline(x3)
```

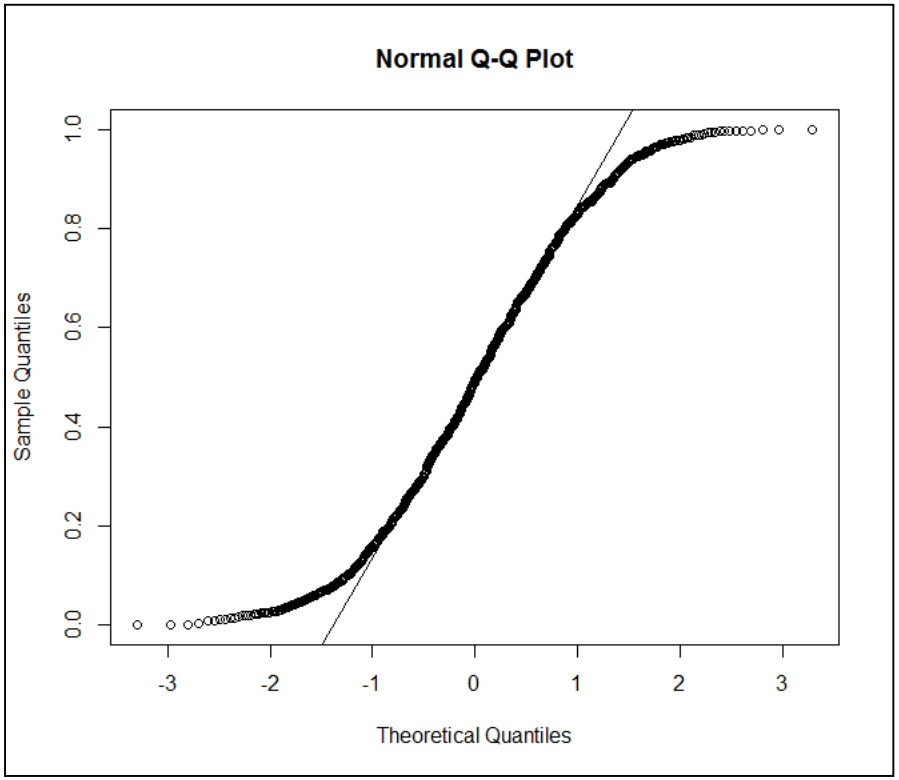


# QQPlot: Known Uniform Distribution: N=1000

```
x2 <- runif(n=1000)
hist(x2)
qqnorm(x2)
qqline(x2)
```

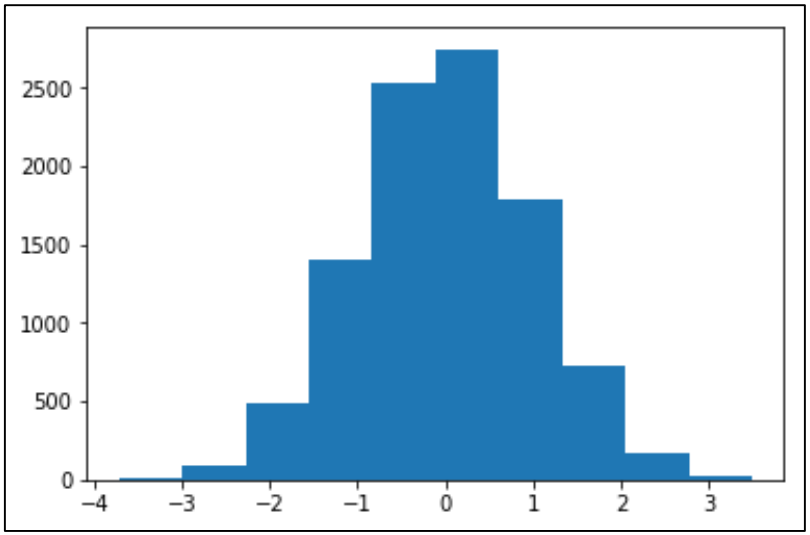


# Python: QQPlot

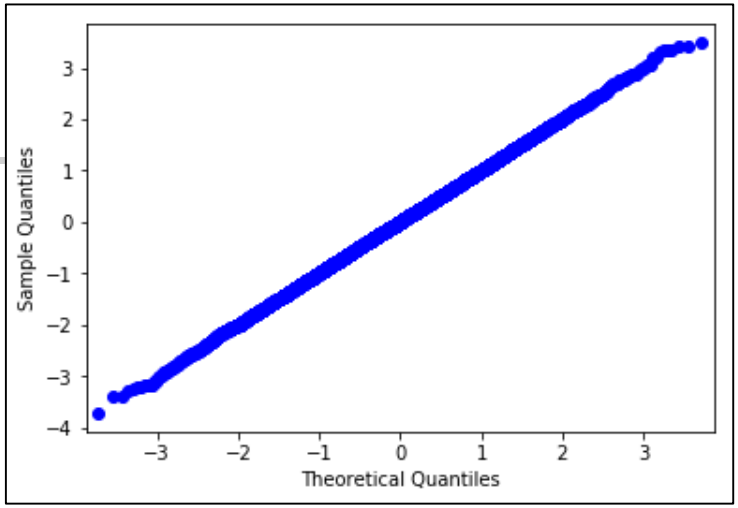
```
import numpy as np

import statsmodels.api as sm
C:\ProgramData\Anaconda3\lib\site-packages\statsmodels\compat\pandas.py:56:
FutureWarning: The pandas.core.datetools module is deprecated and will be
removed in a future version. Please use the pandas.tseries module instead.
  from pandas.core import datetools

import matplotlib.pyplot as plt
```

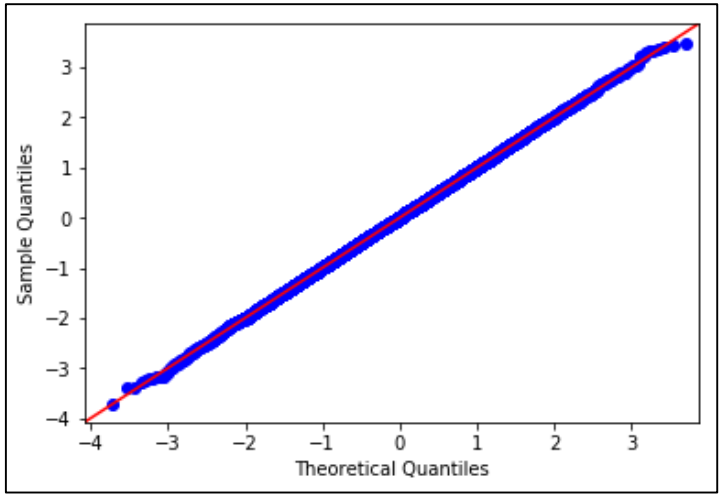
```
test = np.random.normal(0,1,10000)
plt.plot(test)
plt.hist(test)
```

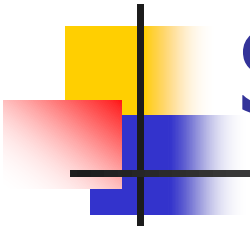


# Python: QQPlot

```
sm.qqplot(test)
sm.qqplot(test,line='45')
```





# Summary

- Distributions
- Probability Distribution Function
- Distributions: Uniform, Normal
- Properties of Normal Distribution
- Standard Normal Distribution
- Testing Normality