# Sampling Distribution & Confidence Interval Estimation

# Topics

- Sampling Distribution + Central Limit Theorem

- Normal Approximation of the Binomial

# Sampling Distribution

Suppose we have a population of size $N = 9$, consisting 9 proteins with amino acid length : 80, 100, 90, 120, 140, 110, 150, 160, 130

*Sampling distribution*: distribution of sample means

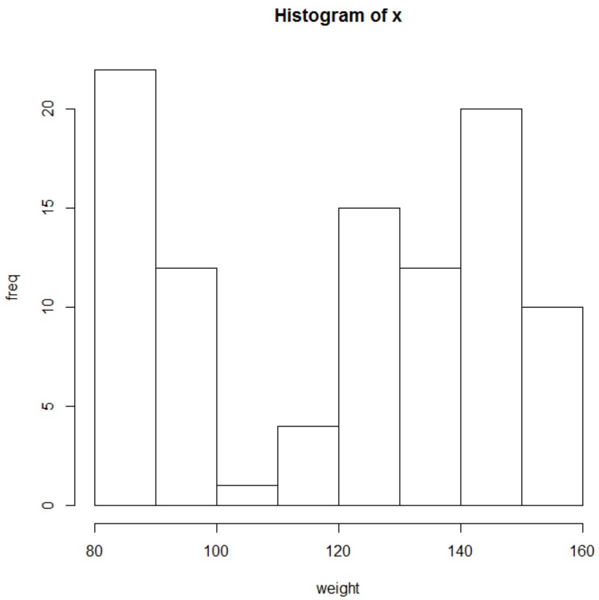
Construct sampling distribution of the sample mean, $\bar{x}$

Based on sample size $n = 2$ drawing from the population.

# Sampling Distribution

| samples | $\bar{x}$ | samples | $\bar{x}$ | samples | $\bar{x}$ |
|---|---|---|---|---|---|
| 80, 100 | 90 | 100, 150 | 125 | 120, 160 | 140 |
| 80, 90 | 85 | 100, 160 | 130 | 120, 130 | 125 |
| 80, 120 | 100 | 100, 130 | 115 | 140, 110 | 125 |
| 80, 140 | 110 | 90, 120 | 105 | 140, 150 | 145 |
| 80, 110 | 95 | 90, 140 | 115 | 140, 160 | 150 |
| 80, 150 | 115 | 90, 110 | 100 | 140, 130 | 135 |
| 80, 160 | 120 | 90, 150 | 120 | 110, 150 | 130 |
| 80, 130 | 105 | 90, 160 | 125 | 110, 160 | 135 |
| 100, 90 | 95 | 90, 130 | 110 | 110, 130 | 120 |
| 100, 120 | 110 | 120, 140 | 130 | 150, 160 | 155 |
| 100, 140 | 120 | 120, 110 | 115 | 150, 130 | 140 |
| 100, 110 | 105 | 120, 150 | 135 | 160, 130 | 145 |

# Central Limit Theorem (CLT)



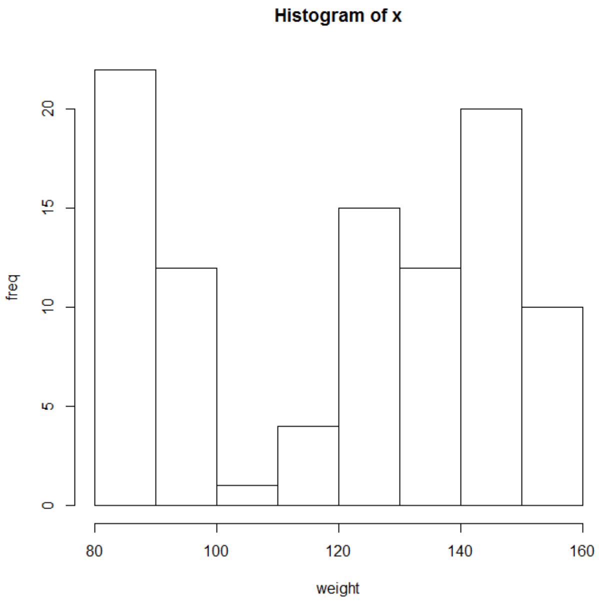
| | |
|---|---|
| 80 | 20 |
| 100 | 12 |
| 90 | 2 |
| 120 | 4 |
| 140 | 12 |
| 110 | 1 |
| 150 | 20 |
| 160 | 10 |
| 130 | 15 |

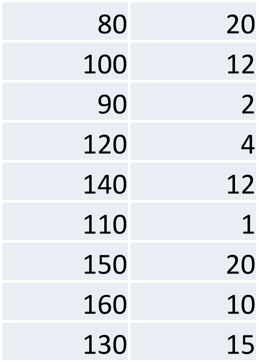
Suppose this is the true distribution of the proteins

80, 100, 90, 120, 140, 110, 150, 160, 130

How would you enter these numbers into a vector in R?

# Central Limit Theorem (CLT)





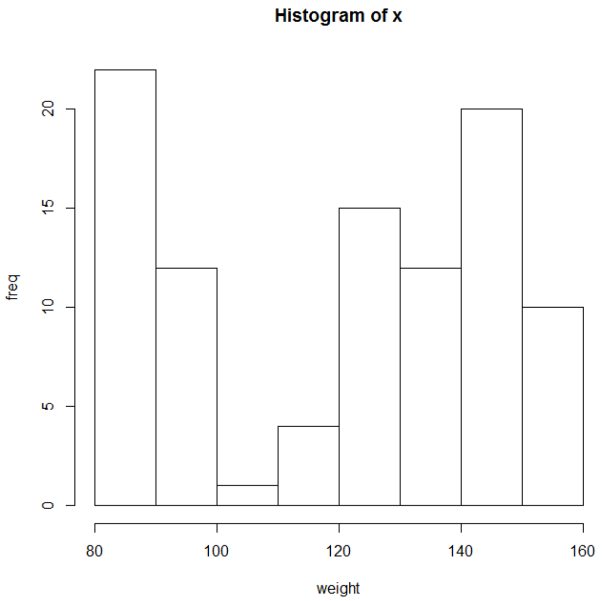
| | |
|---|---|
| 80 | 20 |
| 100 | 12 |
| 90 | 2 |
| 120 | 4 |
| 140 | 12 |
| 110 | 1 |
| 150 | 20 |
| 160 | 10 |
| 130 | 15 |

Suppose this is the true distribution of the proteins

80, 100, 90, 120, 140, 110, 150, 160, 130

```
x<-
c(rep(80,20),rep(100,12),rep(90,2),rep(120,4),rep(140,12),rep(110,1),rep(150,20),
rep(160,10),rep(130,15))
```

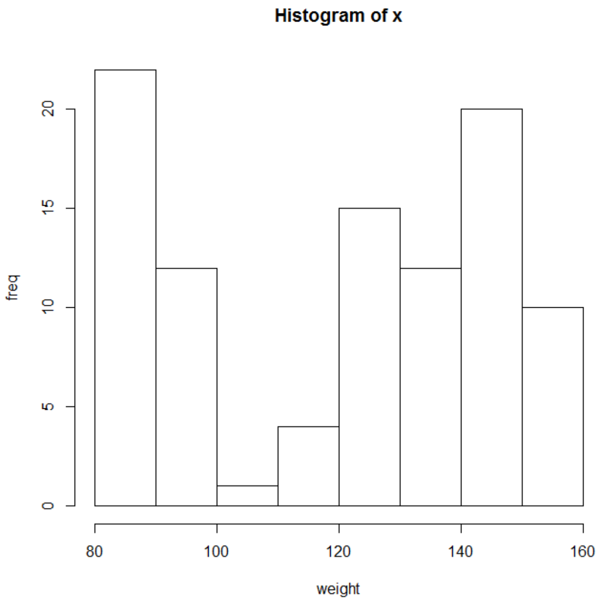
# Central Limit Theorem (CLT)



| | |
|---|---|
| 80 | 20 |
| 100 | 12 |
| 90 | 2 |
| 120 | 4 |
| 140 | 12 |
| 110 | 1 |
| 150 | 20 |
| 160 | 10 |
| 130 | 15 |

```
>x
[1]  80  80  80  80  80  80  80  80  80  80  80  80  80  80  80  80  80  80  80
[20]  80 100 100 100 100 100 100 100 100 100 100 100 100  90  90 120 120 120 120
[39] 140 140 140 140 140 140 140 140 140 140 140 140 110 150 150 150 150 150 150
[58] 150 150 150 150 150 150 150 150 150 150 150 150 150 150 160 160 160 160 160
[77] 160 160 160 160 160 130 130 130 130 130 130 130 130 130 130 130 130 130 130
[96] 130
```

# Central Limit Theorem (CLT)



| | |
|---|---|
| 80 | 20 |
| 100 | 12 |
| 90 | 2 |
| 120 | 4 |
| 140 | 12 |
| 110 | 1 |
| 150 | 20 |
| 160 | 10 |
| 130 | 15 |

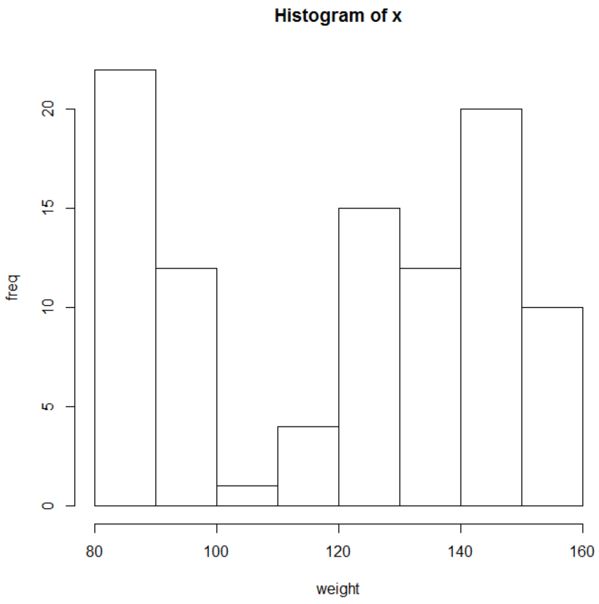
Now, you can sample 10 individuals from this population and find the mean

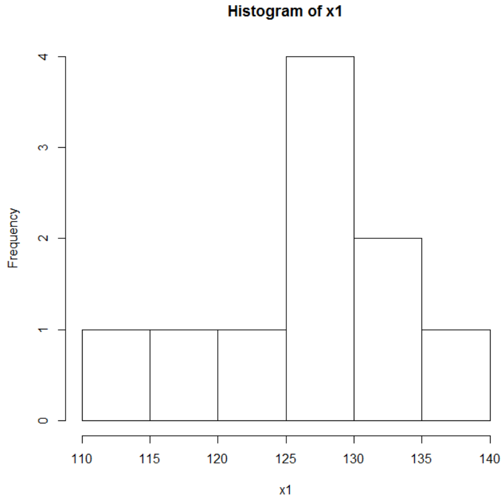
```
>sample(x,10,replace=TRUE)
[1]  80 100 100 130 150  80 150  80 140 140
>mean(sample(x,10,replace=TRUE)
[1] 124
```

And you can replicate sampling using the function replicate()

```
>x1<-replicate(10, sample(x,10,replace=TRUE))
```

# Central Limit Theorem (CLT)





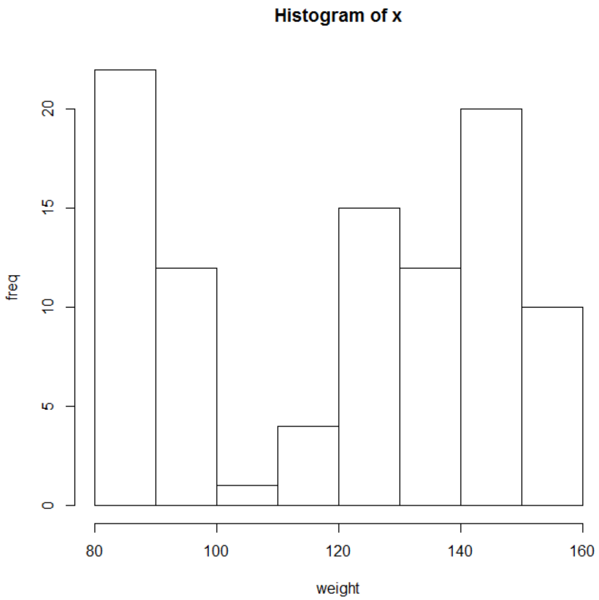
And you can replicate finding the mean of each sample using the function replicate()

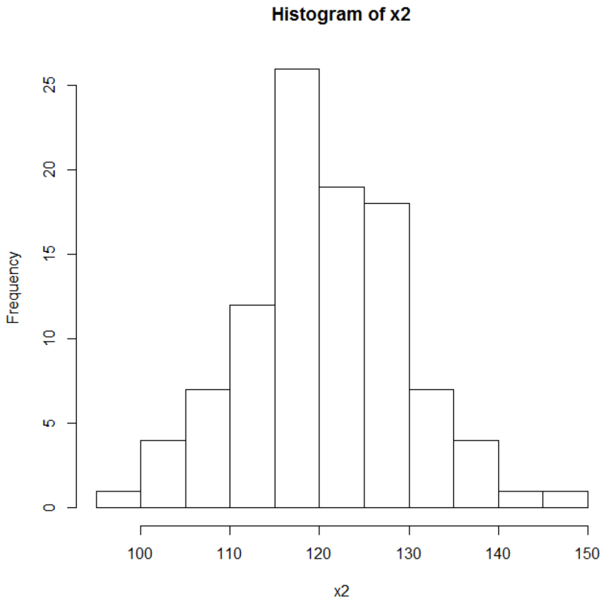
```
>x1<-replicate(10, mean(sample(x,10,replace=TRUE)))
[1] 114 129 135 123 128 119 126 127 132 137
```

And draw a histogram of the sample means – replicated 10 times

```
>hist(x1)
```

# Central Limit Theorem (CLT)





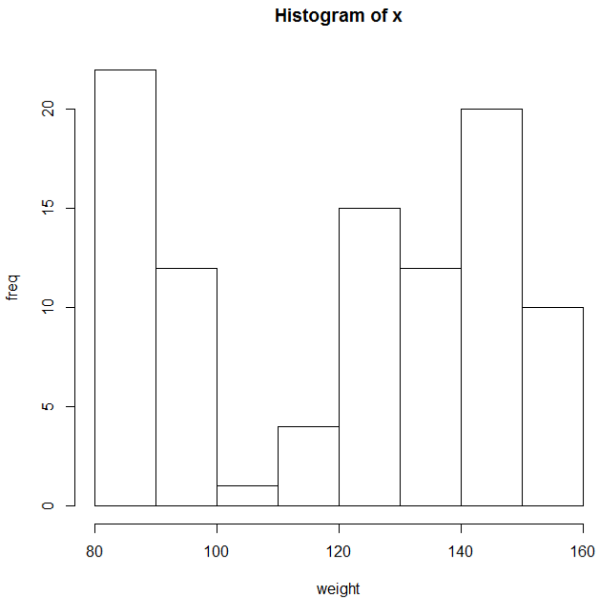
Now lets increase the number of sampling replicates to 100

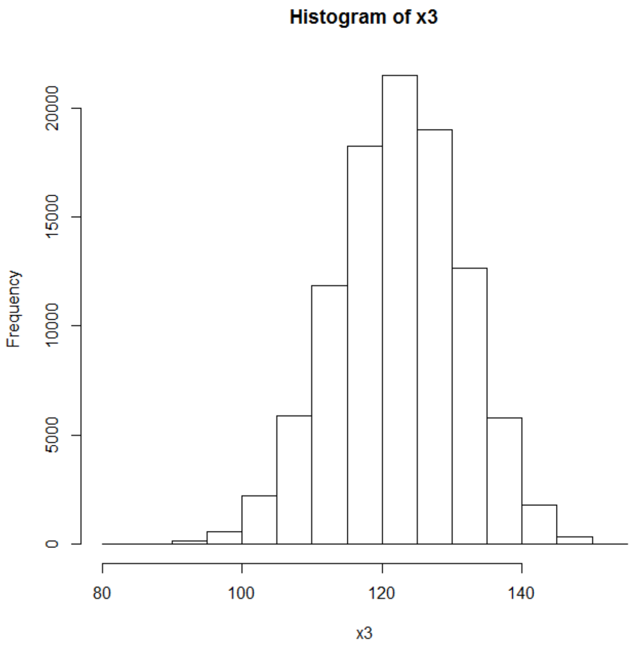
```
>x1<-replicate(100, mean(sample(x,10,replace=TRUE)))
```

And draw a histogram of the sample means replicated 100 times

```
>hist(x1)
```

# Central Limit Theorem (CLT)





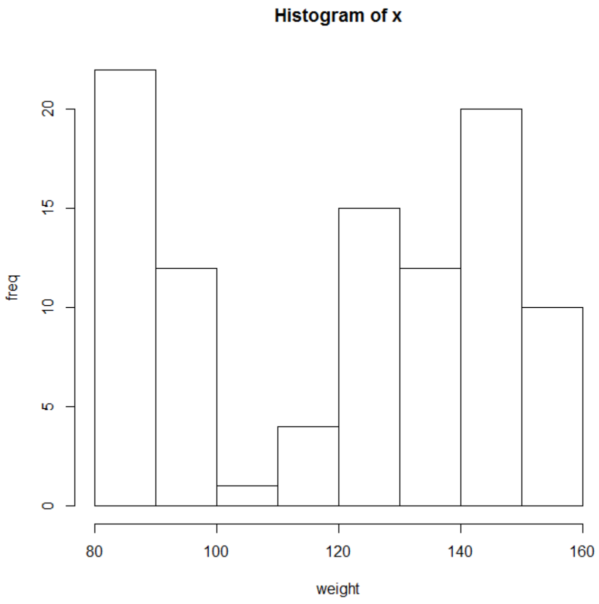
Now lets increase the number of sampling replicates to 100000

```
>x3<-replicate(100000, mean(sample(x,10,replace=TRUE)))
```

And draw a histogram of the sample means replicated 100 times

```
>hist(x3)
```

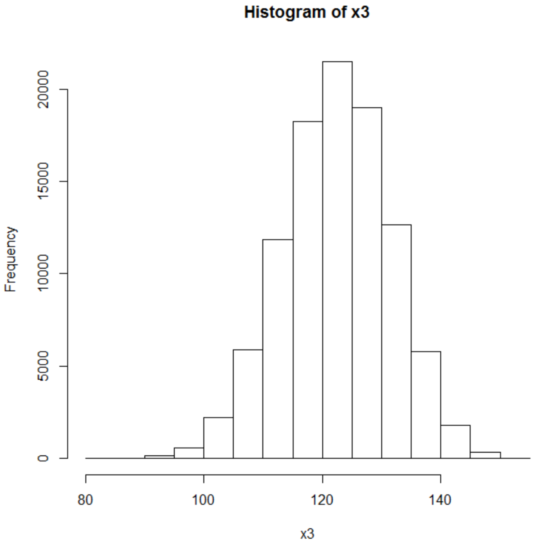
# Central Limit Theorem (CLT)



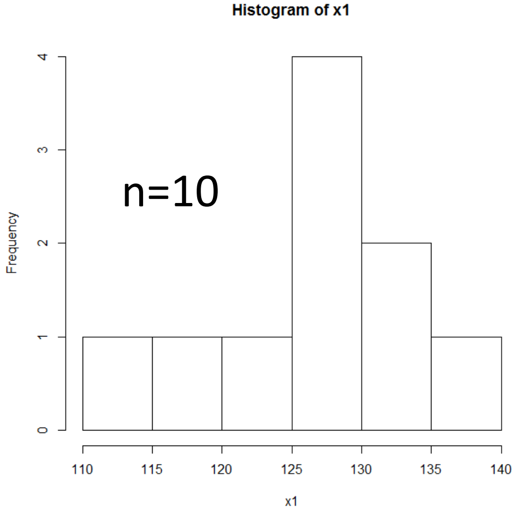
Original distribution

```
> mean(x)
[1] 122.9167
> sd(x)
[1] 28.68767
```

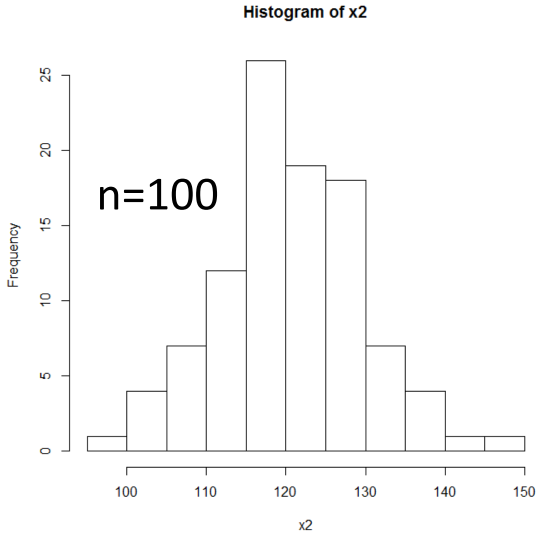
Sampling distribution of Sample means



n=100000



n=10



n=100

# Central Limit Theorem (CLT)

When ***n*** is large, the sampling distribution of $\overline{x}$ will be approximately *normal* with the approximation becoming more precise as ***n*** increases

$$u_{\bar{x}} = u$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

$$\sigma_{\bar{x}} = standard\ error\ of\ \bar{x}$$

$$Z = \frac{\bar{X} - u}{\frac{\sigma}{\sqrt{n}}}$$

$u, \sigma$ = population distribution

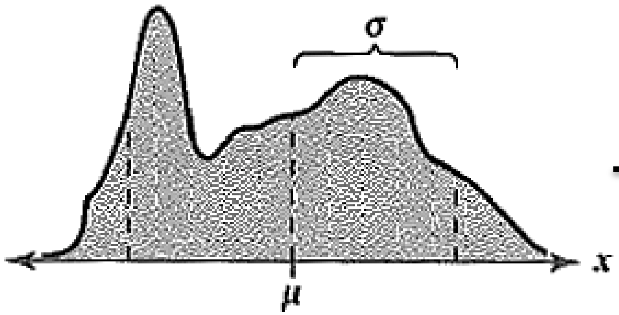
$u_{\bar{x}}, \sigma_{\bar{x}}$ = sampling distribution $\bar{x}$

What number is large enough? ($n$ >30)

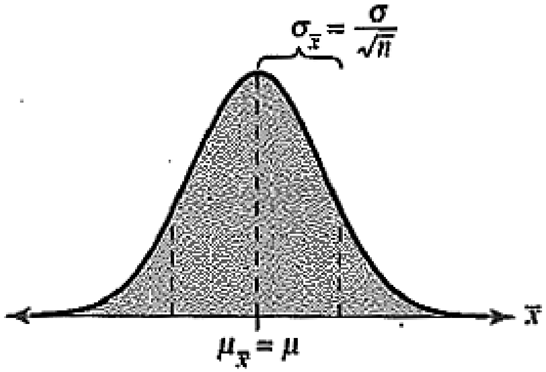
If the population itself is normally distributed, the sampling distribution of sample means is normally distributed for ***any*** sample size *n*.

# Central Limit Theorem (CLT)

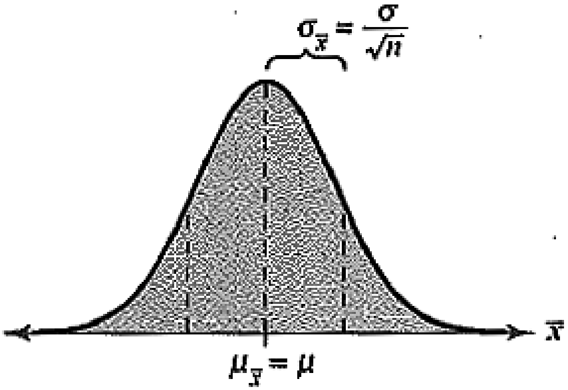
*For any population distribution* by CLT

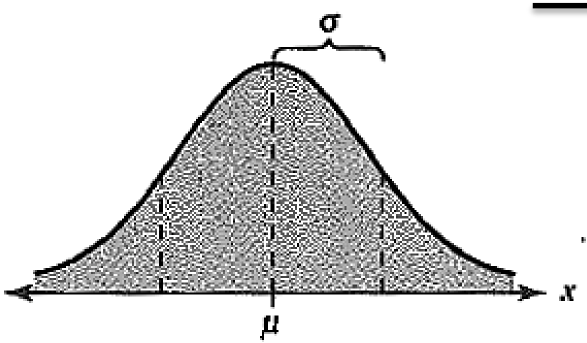


**distribution of sample means**
***n*≥30**



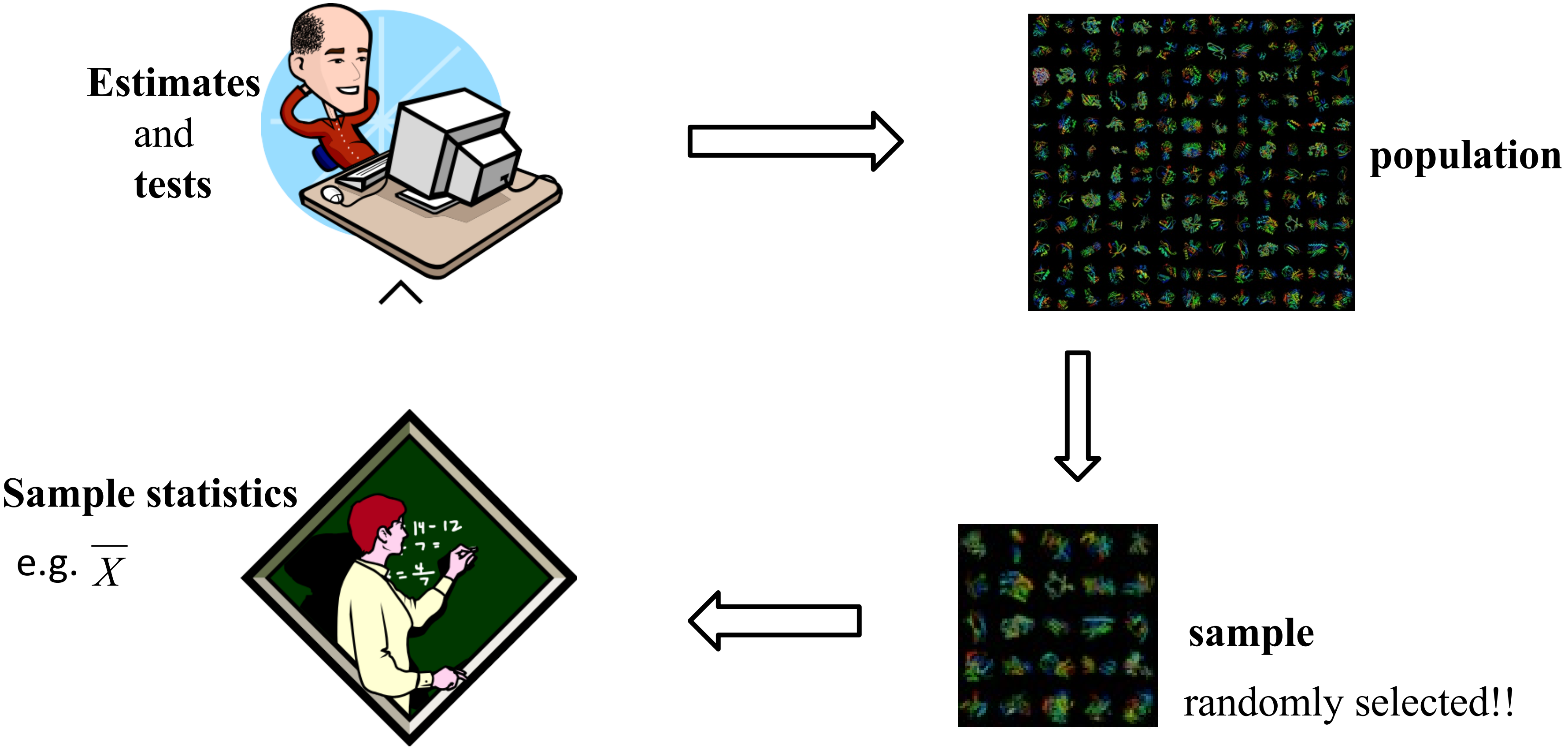
If the population itself is normally distributed, the sampling distribution of sample means is normally distributed for ***any*** sample size *n*.





**distribution of sample means**
**any size *n***

# Statistical Inference



**Estimates**
and
**tests**

**population**

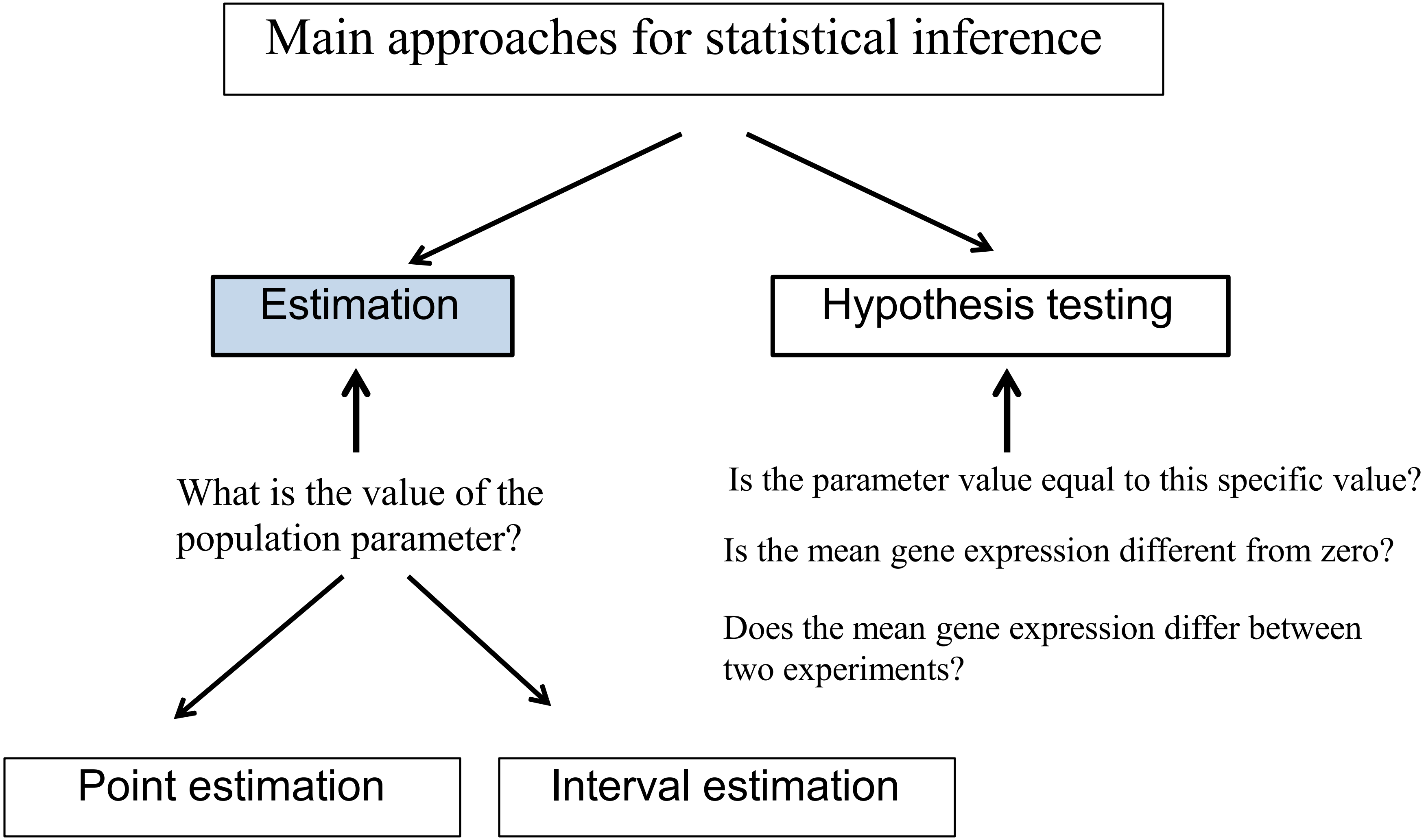
**sample**

randomly selected!!

**Sample statistics**

e.g. $\overline{X}$

# Statistical Inference



Main approaches for statistical inference

Estimation

What is the value of the population parameter?

Point estimation

Interval estimation

Hypothesis testing

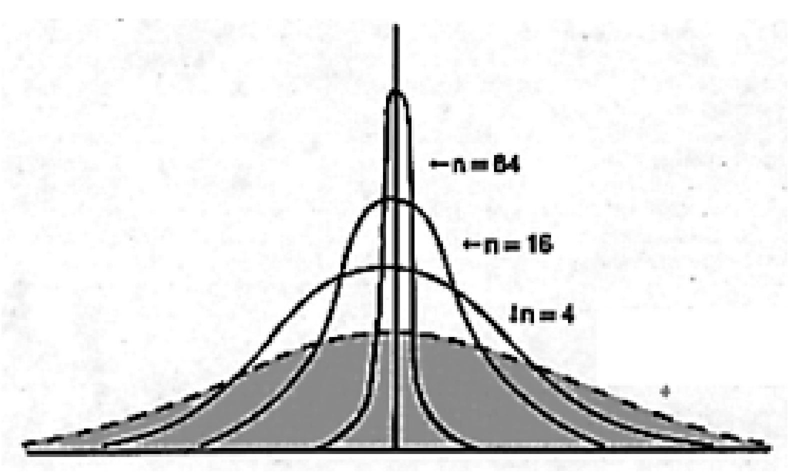
Is the parameter value equal to this specific value?

Is the mean gene expression different from zero?

Does the mean gene expression differ between two experiments?

# Point Estimation

$\bar{x}\ can\ estimate\ u$



Studied through sampling distribution of $\bar{X}$

**It provides a *single* value

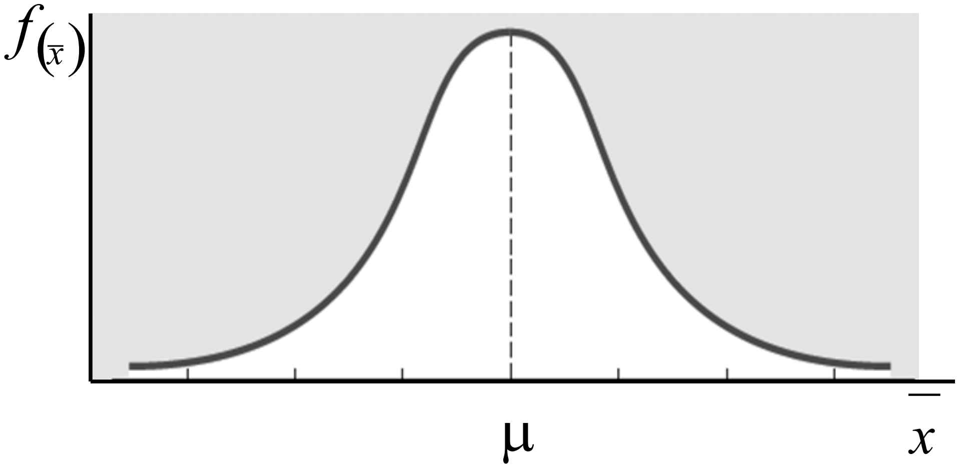
**It does not provide information about how close the value is to the population parameter

# Interval Estimation

*It provides a range of values based on one sample.*

It provides information about the *closeness to the unknown population parameter* in terms of *probability or confidence*

Example: $n = 40$ proteins, and we want to know the mean weight



Let's say $\bar{x} = 290$

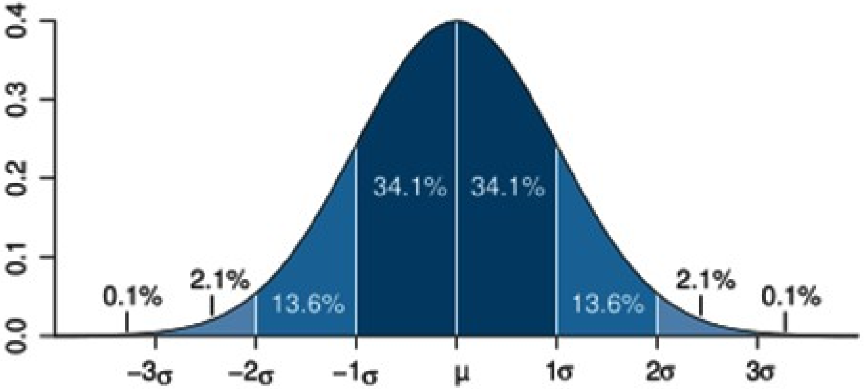
Is μ exactly 290?
Check the sampling distribution→→

How about $1 < \mu < 5000$ ?
--Large range
--We need a small interval with high confidence

*e.g.* unknown population mean of protein size is between 260 and 340 with 95% confidence

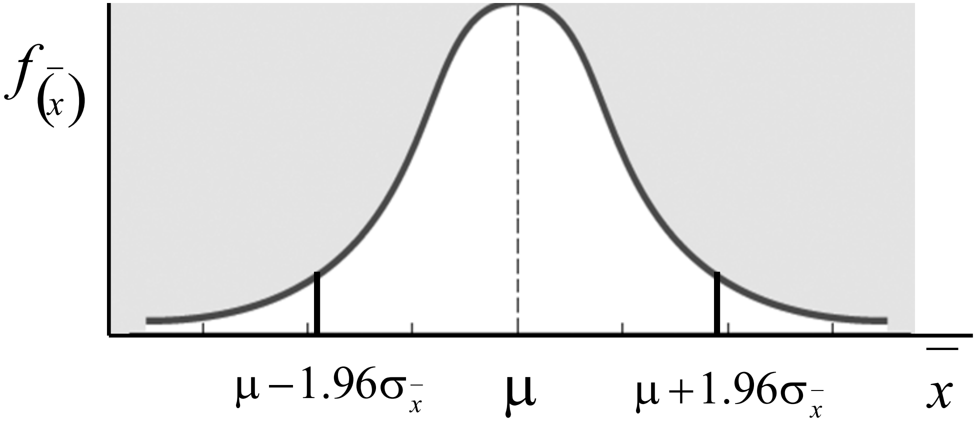
# Interval Estimation

By CLT, we know that for a large *n*, $\bar{x}$ is approximately normally distributed with a mean μ and a standard error $\sigma_{\bar{x}}$



<u>Remember the empirical rules</u>:

~68 % of the values are within 1 standard deviation of the mean
~95 % of the values are within 2 standard deviations of the mean
~99.7 % lie within 3 standard deviations of the mean



The interval $\mu \pm 2\sigma_{\bar{x}}$

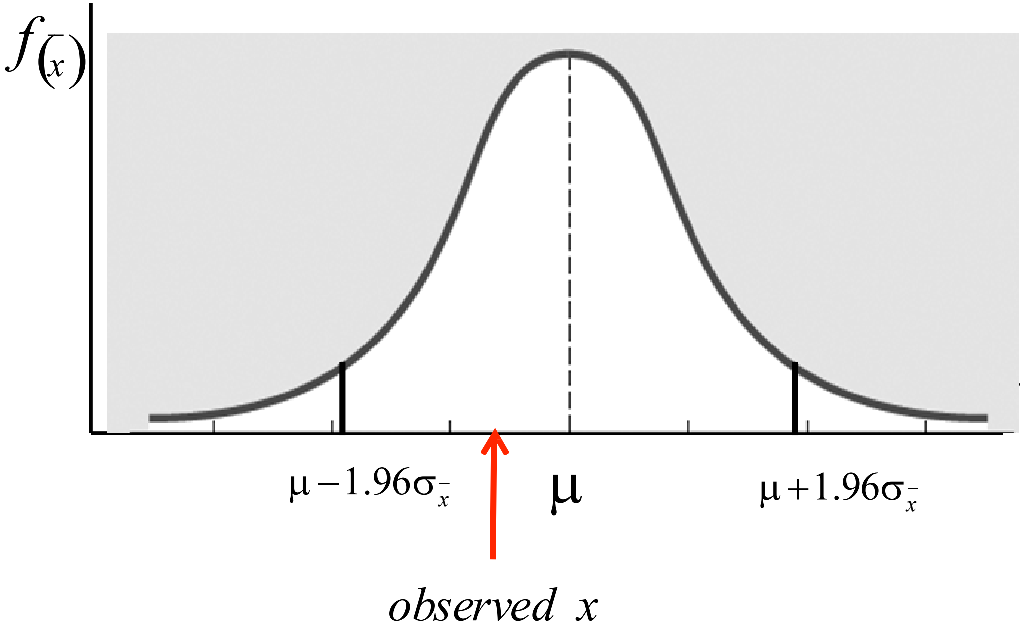
Or more precisely:

$$(\mu - 1.96\sigma_{\bar{x}},\ \mu + 1.96\sigma_{\bar{x}})$$

includes 95% of $\bar{x}$ from the sampling

# Interval Estimation

Another way to look at this:



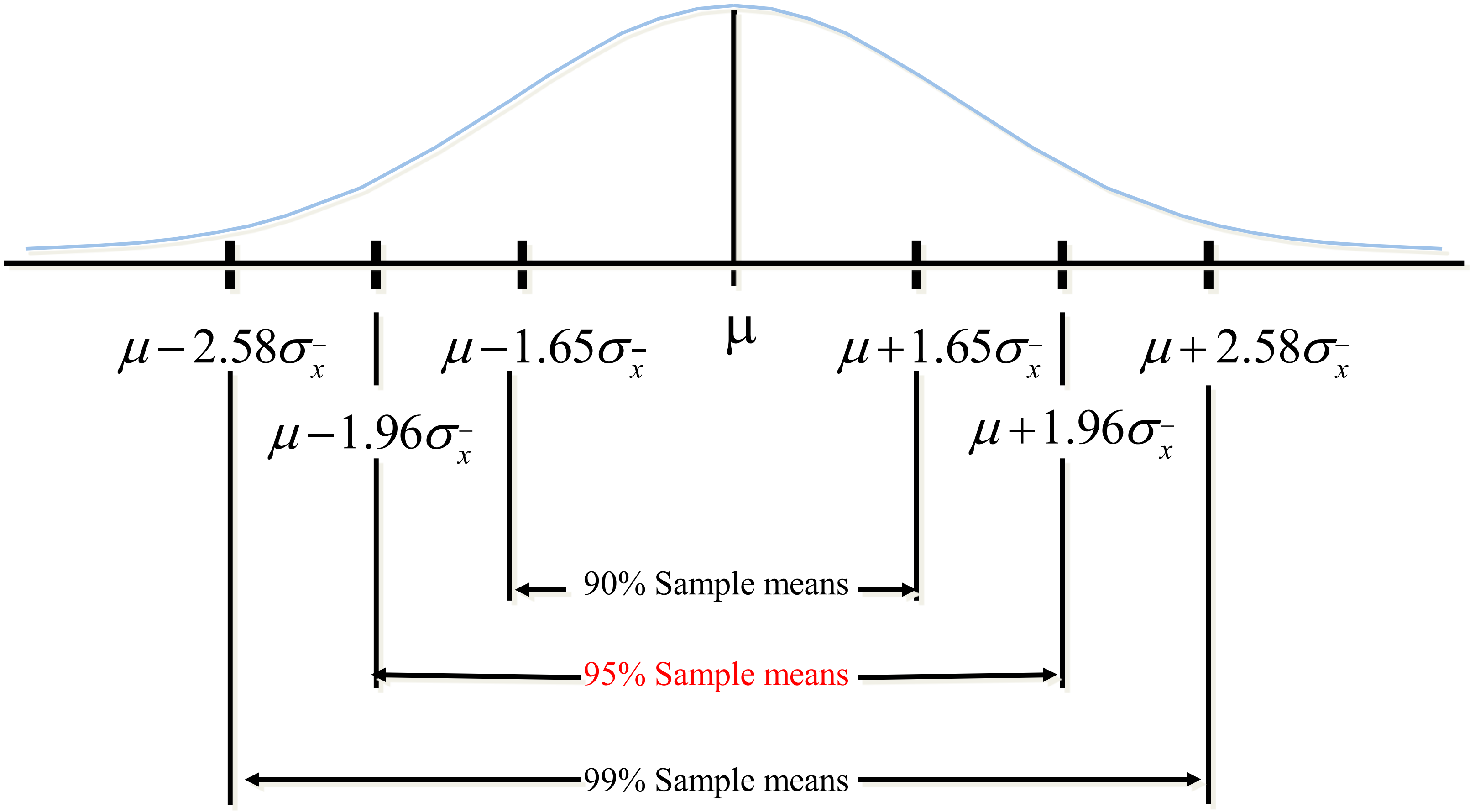
**Any time $\bar{x}$ falls in the interval $\mu \pm 1.96\sigma_{\bar{x}}$, the interval $\bar{x} \pm 1.96\sigma_{\bar{x}}$ will contain the parameter $\mu$ .

Since the probability of $\bar{x}$ falls in the interval $\mu \pm 1.96\sigma_{\bar{x}}$ is 0.95

$$\bar{x} \pm 1.96\sigma_{\bar{x}}$$

is an interval estimate of μ with level of confidence 95%

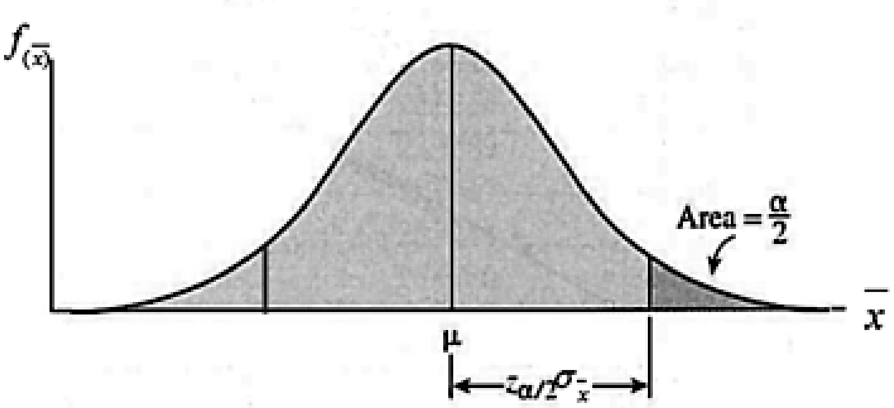
# Confidence Depends on Interval Z


$\mu - 2.58\sigma_{\bar{x}}$
$\mu - 1.65\sigma_{\bar{x}}$
μ
$\mu + 1.65\sigma_{\bar{x}}$
$\mu + 2.58\sigma_{\bar{x}}$
$\mu - 1.96\sigma_{\bar{x}}$
$\mu + 1.96\sigma_{\bar{x}}$
90% Sample means
95% Sample means
99% Sample means

# Confidence Depends on Interval Z

When ***n*** is large and σ is known, a 100*(1-α)% Confidence Interval for μ has bounds

$$\bar{x} \pm z_{\alpha/2}\sigma_{\bar{x}} \qquad \sigma_{\bar{x}} = \sigma/\sqrt{n}$$

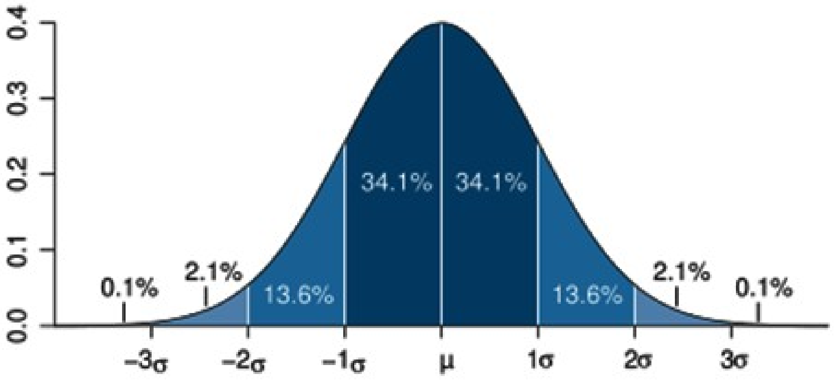


| α | 100 (1-α)% | $Z_{\alpha/2}$ |
|---|---|---|
| 0.10 | 90% | 1.645 |
| 0.05 | 95% | 1.96 |
| 0.01 | 99% | 2.58 |

Confidence Interval

1. The probability of the unknown population parameter falls within interval
2. Defined 100*(1-α)%----**<u>confidence coefficient</u>**
3. Typical values: 99%, 95%, 90%

# Normal Distribution



ZSCORE IS HOW MANY STANDARD DEVIATIONS YOU ARE FROM THE MEAN

There are 414 female biology students. The mean height is 166.8. and the stdev is 6.4cm.

What range of heights include 95% of this population? This also means if you randomly select a person, we are 95% confident that this person will be between 154.23 and 179.34 cm.

$$Z = \frac{X - \mu}{\sigma} \qquad X = z\sigma + u$$

$$X = z\sigma + u = (-1.96 \times 6.4) + 166.8 = 154.23$$

$$X = z\sigma + u = (1.96 \times 6.4) + 166.8 = 179.34$$

$$154.23 < X < 179.34$$

# Topics

- Sampling Distribution + Central Limit Theorem

- Normal Approximation of the Binomial

# Normal Approximation to Binomial Distribution

Recall that we calculated the probability that x has a specific value (x=k) given that we know two parameters:

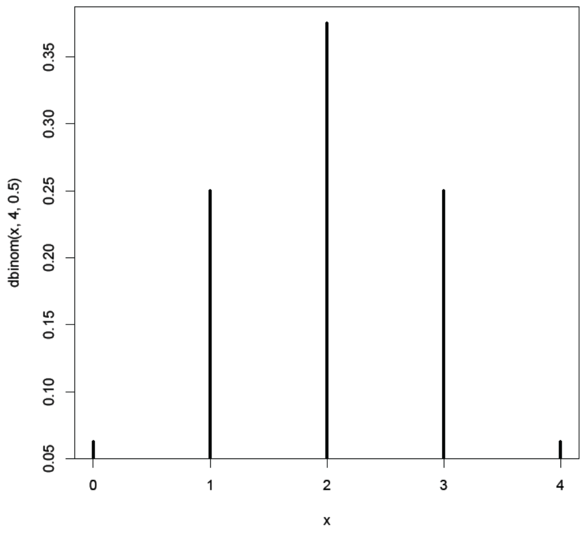
n: the number of cases,

p: the probability of success in any case

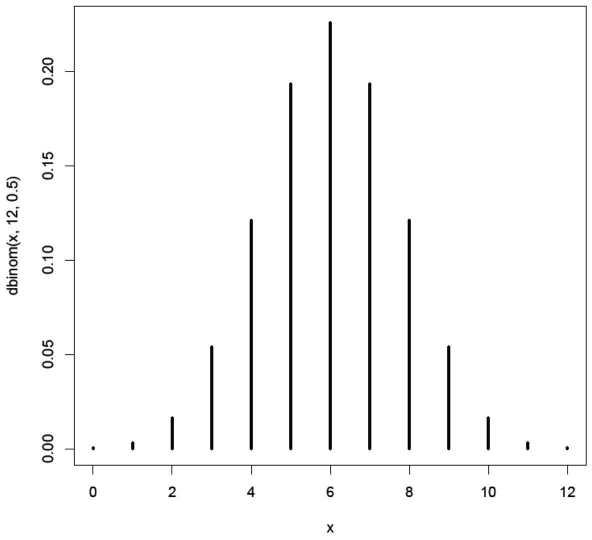
Probability density function of X is:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad for\ k = 0, 1, \ldots\ldots, n$$

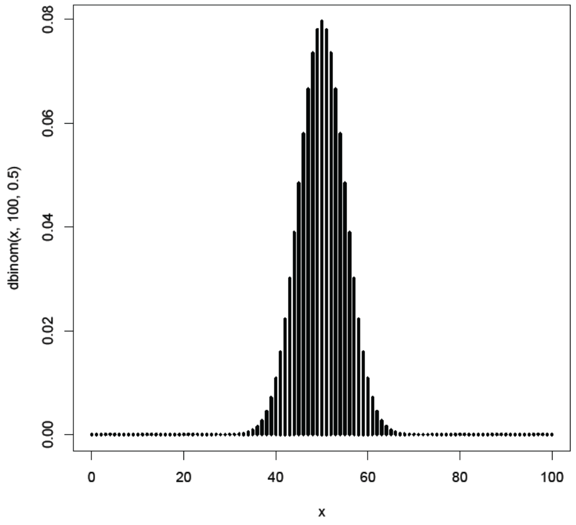
$$u_x = np\ and\ \sigma_x = \sqrt{np(1-p)}$$



*n=4,   p=0.5*



*n=12, p=0.5*



*n=100, p=0.5*

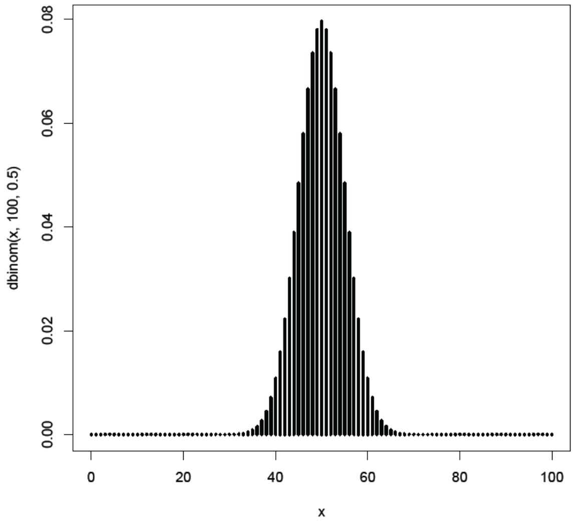
# Normal Approximation to Binomial Distribution

Recall that we calculated the probability that x has a specific value (x=k) given that we know two parameters:

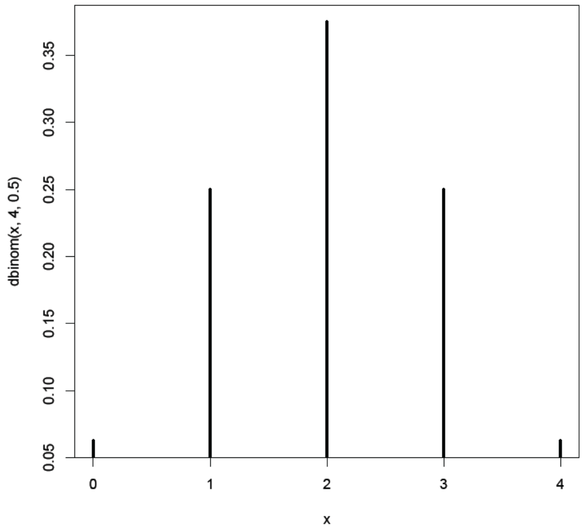
n: the number of cases,

p: the probability of success in any case

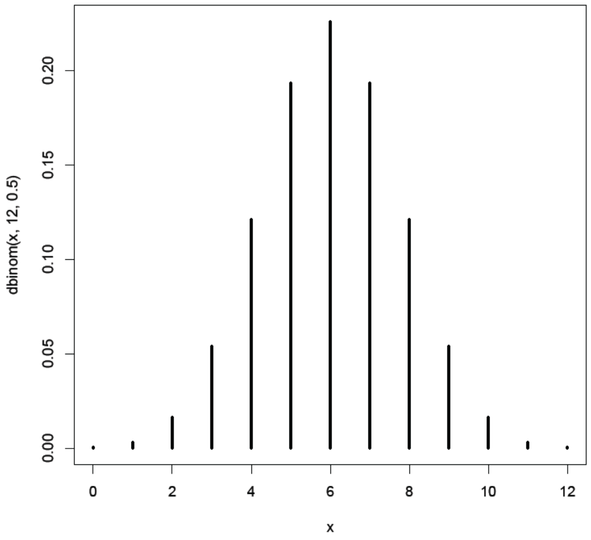
When n is large, and p is not too close to 0 or 1, the binomial distribution begins to approximate a normal distribution --- Z SCORE TIME



*n=100, p=0.5*



*n=4, p=0.5*



*n=12, p=0.5*

# Normal Approximation to Binomial Distribution

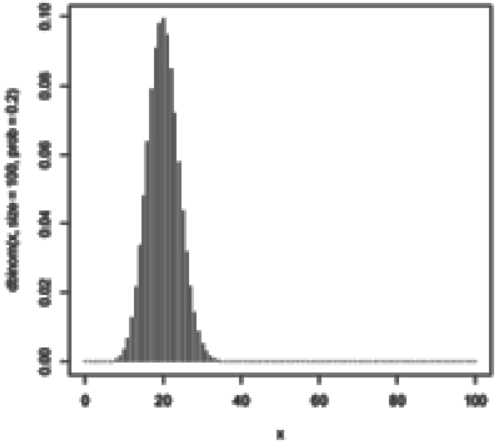
Appropriate conditions for normal approximation to binomial distribution:

***1. n*** is large
***2. p*** is not too near 0 or 1

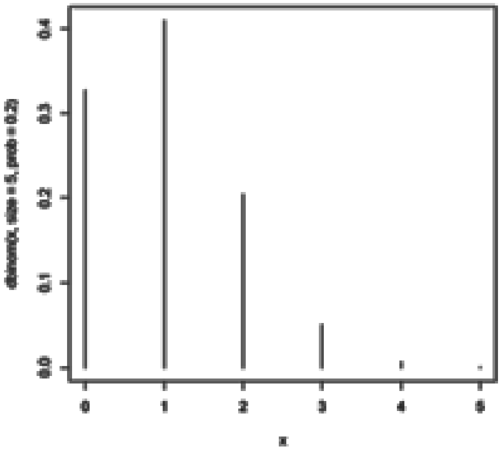
Rule of thumb: $np \geq 5$ AND $n(1-p) \geq 5$

Otherwise the actual binomial distribution is skewed to

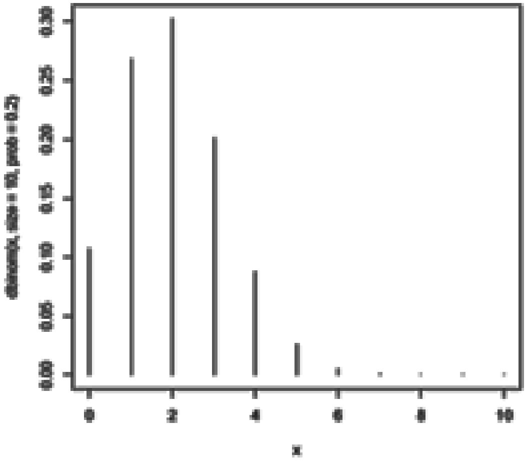
the left or the right



*n=100, p=0.2*



*n=5, p=0.2*



*n=10, p=0.2*

# Normal Approximation to Binomial Distribution

Example: We are going to select 100 protein structures from the protein data bank. About 20% of proteins are membrane proteins. What is the probability that at least 15 of the 100 protein structures are membrane proteins?

# Normal Approximation to Binomial Distribution

Example: We are going to select 100 protein structures from the protein data bank. About 20% of proteins are membrane proteins. What is the probability that at least 15 of the 100 protein structures are membrane proteins?

We can calculate u and σ

$$\mu = 100 \times 0.2 = 20$$

$$\sigma = \sqrt{100 \times 0.2(1 - 0.2)} \quad = 4$$

# Normal Approximation to Binomial Distribution

Example: We are going to select 100 protein structures from the protein data bank. About 20% of proteins are membrane proteins. What is the probability that at least 15 of the 100 protein structures are membrane proteins?

$$\mu = 100 \times 0.2 = 20$$

$$\sigma = \sqrt{100 \times 0.2(1-0.2)} = 4$$

$$P(X \geq 14.5) = P(Z \geq \frac{14.5-20}{4})$$

We can calculate u and σ

Then convert it into a Z-score

# Normal Approximation to Binomial Distribution

Example: We are going to select 100 protein structures from the protein data bank. About 20% of proteins are membrane proteins. What is the probability that at least 15 of the 100 protein structures are membrane proteins?

$$\mu = 100 \times 0.2 = 20$$

$$\sigma = \sqrt{100 \times 0.2(1-0.2)} \quad = 4$$

$$P(X \geq 14.5) = P(Z \geq \frac{14.5-20}{4})$$

$$= P(Z \geq -1.38)$$

$$= 1 - P(Z < -1.38)$$

$$= 1 - 0.0838$$

$$= 0.9162$$

We can calculate u and σ

Then convert it into a Z-score

Then calc. probability

# Normal Approximation to Binomial Distribution

Example: We are going to select 100 protein structures from the protein data bank. About 20% of proteins are membrane proteins. What is the probability that at least 15 of the 100 protein structures are membrane proteins?

$$\mu = 100 \times 0.2 = 20$$

$$\sigma = \sqrt{100 \times 0.2(1-0.2)} \quad = 4$$

$$P(X \geq 14.5) = P(Z \geq \frac{14.5 - 20}{4})$$

$$= P(Z \geq -1.38)$$

$$= 1 - P(Z < -1.38)$$

$$= 1 - 0.0838$$

$$= 0.9162$$

We can calculate u and σ
Then convert it into a Z-score
Then calc. probability

Note: We converted a binomial (discrete) to a normal (continuous) so we can use non-discrete values for prob.

# BINF6200/8200:Statistics for Bioinformatics
# Lab 4

R - Poisson Distribution

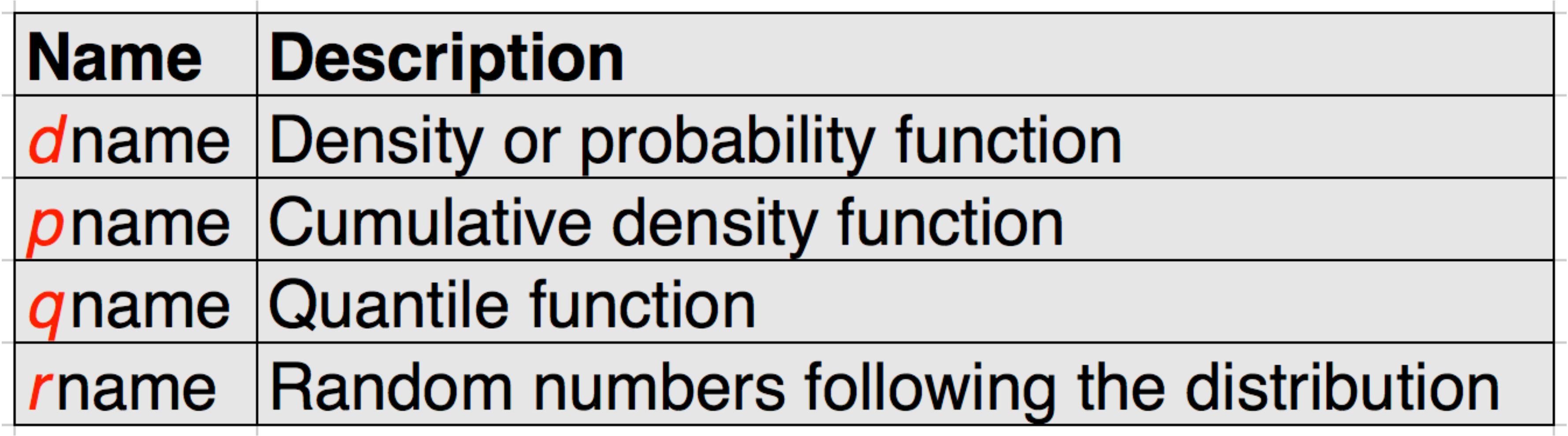
R - Creating figures, labeling axis

R - Testing for Normality

R - Drawing/Shading Normal Dist.

# Poisson Distribution

In general, functions for each distribution…

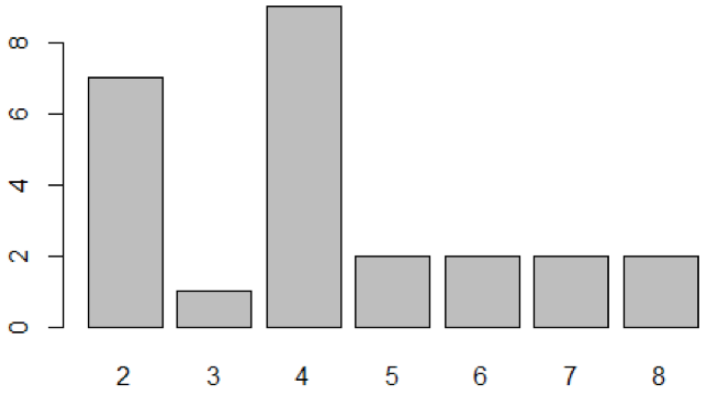


| **Name** | **Description** |
|---|---|
| *d*name | Density or probability function |
| *p*name | Cumulative density function |
| *q*name | Quantile function |
| *r*name | Random numbers following the distribution |

# Poisson Distribution

❖ Generate 25 random numbers following Poisson distribution with λ = 4

```
> set.seed(423)                 # make the generated numbers repeatable
> v <- rpois(25, 4)
> barplot(table(v))             # table() build a contingency table of the counts at
                                  each combination of factor levels
```



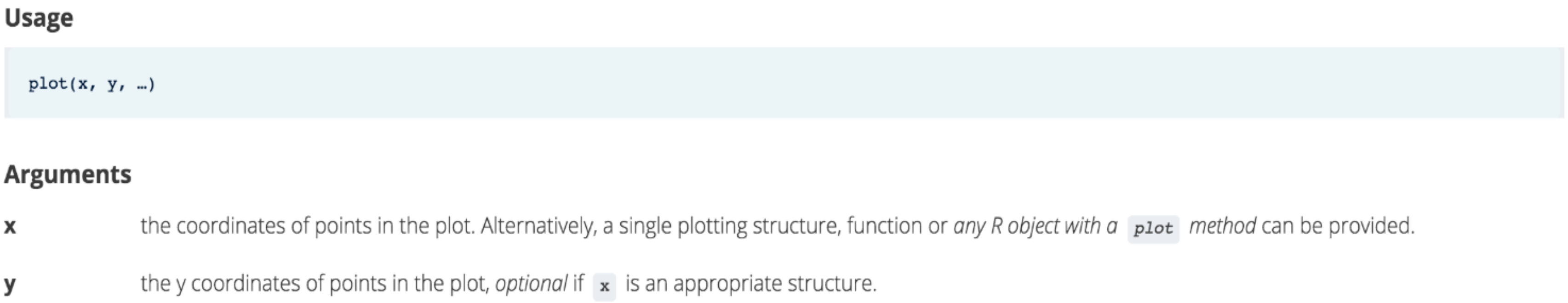
➢ The bar plot is an estimate of the probability distribution. It is more appropriate than a histogram, because the data are discrete, not continuous.

# Plotting

- ❖ Graphing techniques we’ve learned so far ...

> boxplot()
> barplot()
> hist()

➢ The most used plotting function in R is the **plot()** function. It is a generic function, including many methods which are called according to the type of object passed to plot().



**Usage**

```
plot(x, y, …)
```

**Arguments**

**x** the coordinates of points in the plot. Alternatively, a single plotting structure, function or *any R object with a* `plot` *method* can be provided.

**y** the y coordinates of points in the plot, *optional* if `x` is an appropriate structure.

# Plotting

## Changing plot type:

| | |
|---|---|
| type= “p”. | # points |
| “l” | # lines |
| “b” | # both lines and points |
| “h” | # histogram-like vertical lines |
| “s” | # stair steps |

## Adding title & axis label:

| | |
|---|---|
| main= | # overall plot tile |
| sub= | # sub title |
| xlab= | # a title for x axis |
| ylab= | # a title for y axis |

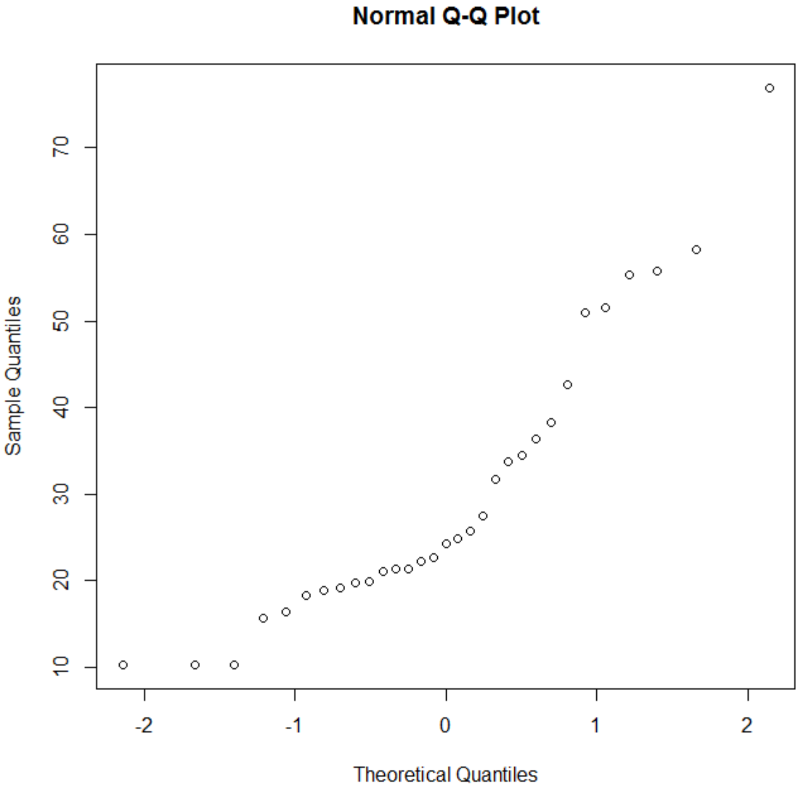
## Changing color & symbol:

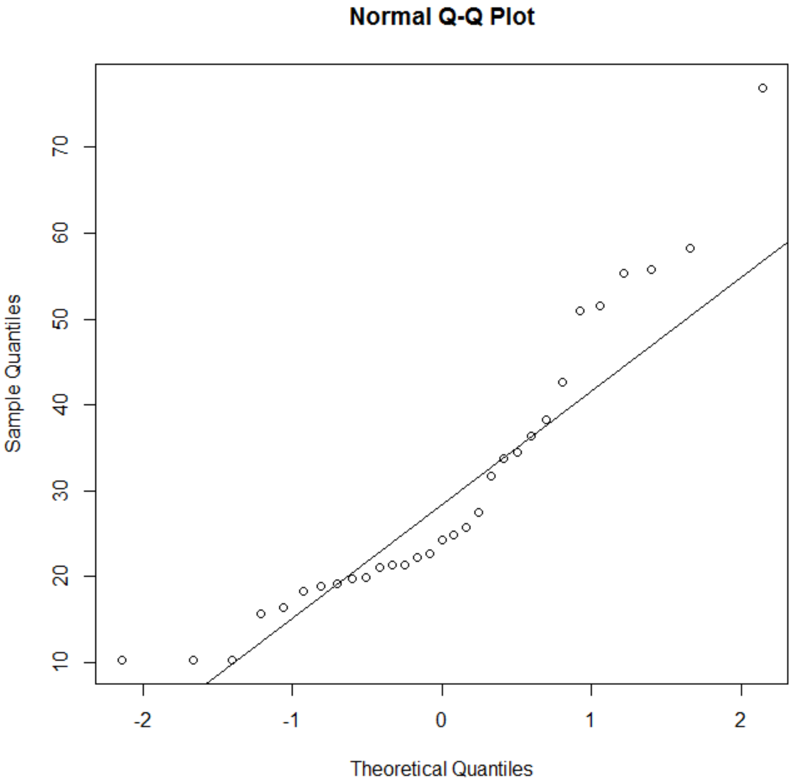
| | |
|---|---|
| col= | # color of the plot |
| pch= | # specify symbols to use when plotting points |

# Testing for Normality

You can test for normality by using a probability plot – qqplot in R

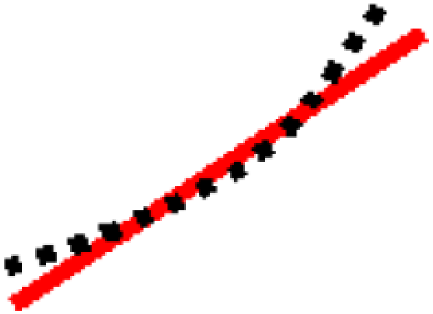
```
> qqnorm(trees.data$Volume)
> qqline(trees.data$Volume)
```



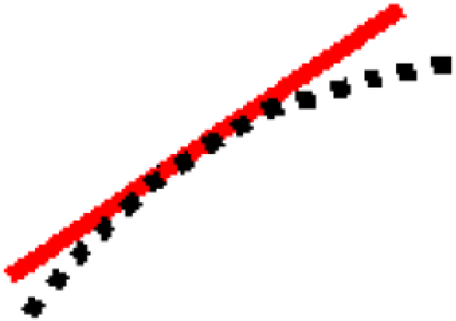


# Testing for Normality

**It indicates that your distribution has:**

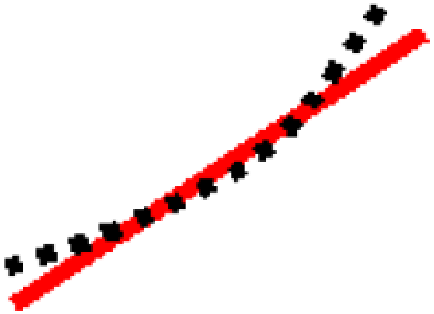


Right Skew – if the plotted points appear to bend up and to the left of the normal line, that indicates a long tail to the right

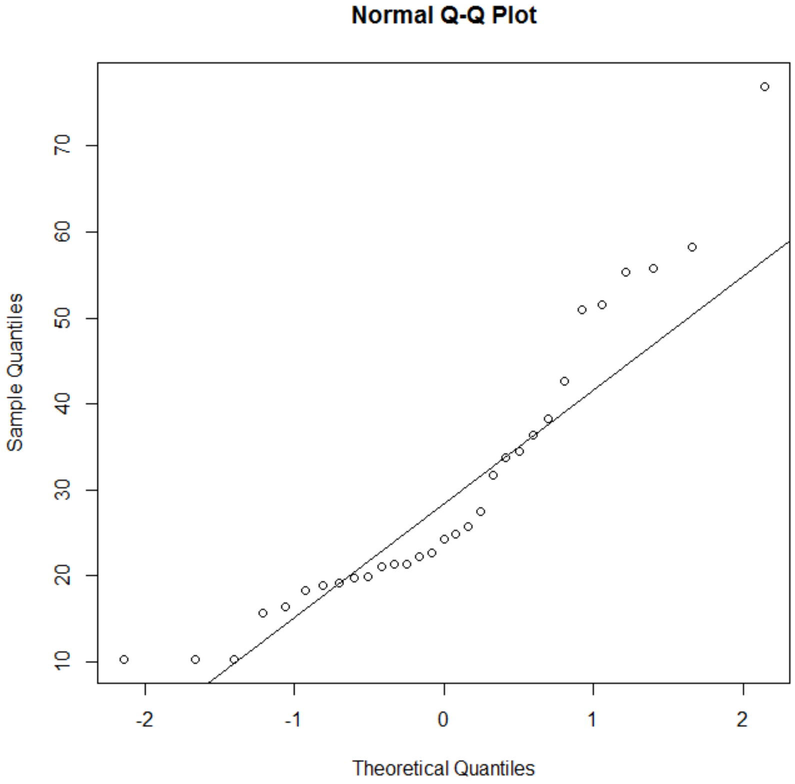


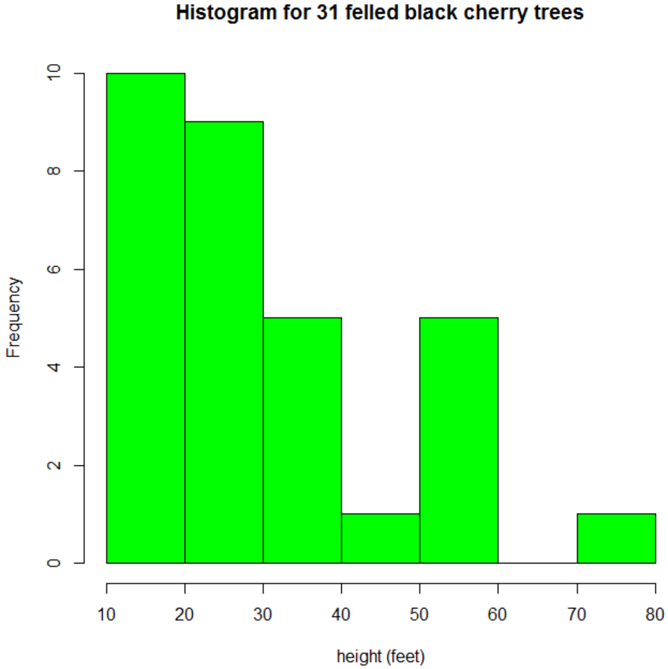
Left Skew – if the plotted points appear to bend down and to the right of the normal line, that indicates a long tail to the left

# Testing for Normality



Right Skew – if the plotted points appear to bend up and to the left of the normal line, that indicates a long tail to the right





# Testing for Normality

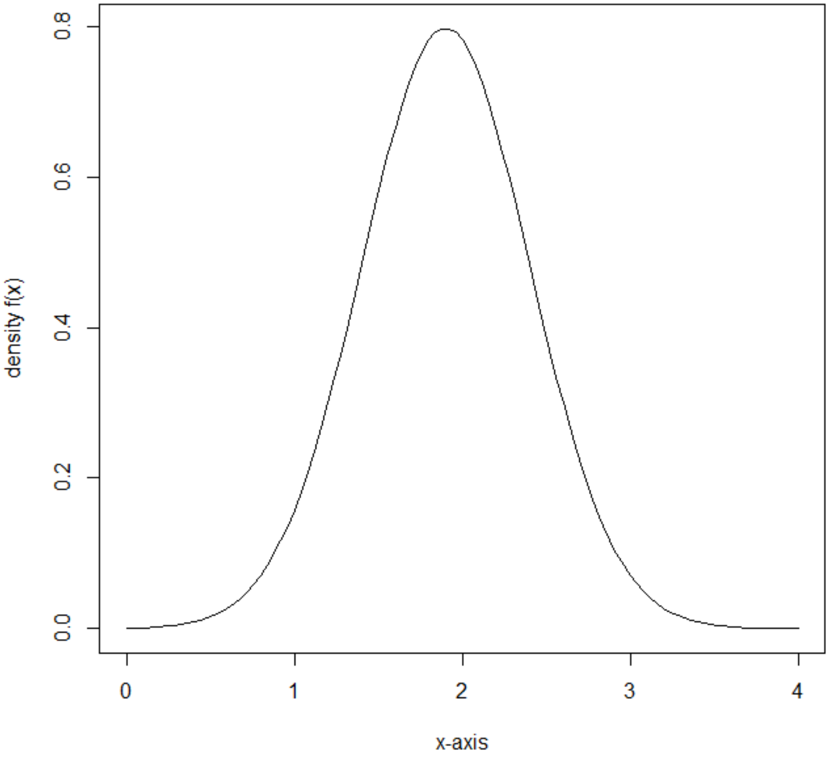
You can test for normality by using a normality test : SW test in R

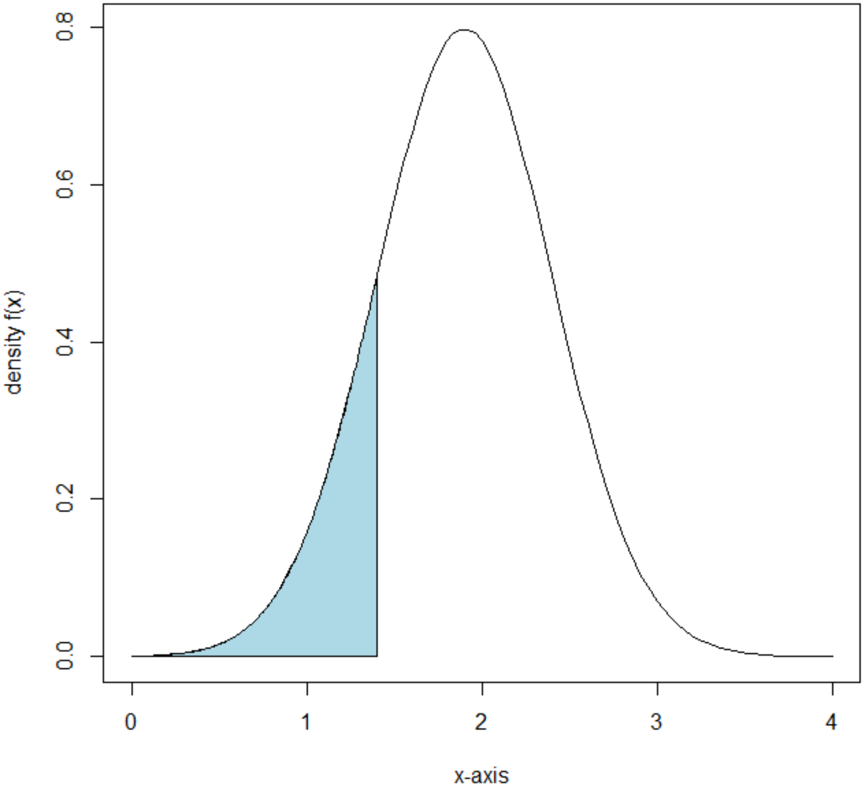
```
> shapiro.test(trees.data$Volume)

        Shapiro-Wilk normality test

data:  trees.data$Volume
W = 0.88757, p-value = 0.003579

>
```

# Plotting a Normal Dist

```
> f<-function(x){dnorm(x,1.9,0.5)}
> plot(f,0,4,xlab="x-axis",ylab="density f(x)")
```

```
> x<-seq(0,1.4,0.01)
> polygon(c(0,x,1.4), c(0,f(x),0), col="lightblue")
```



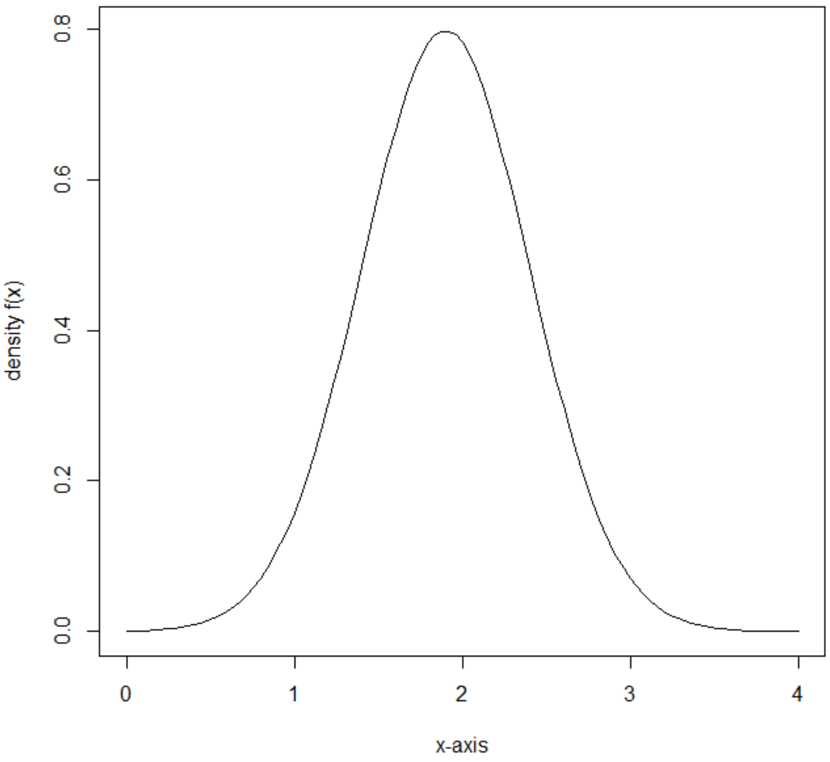


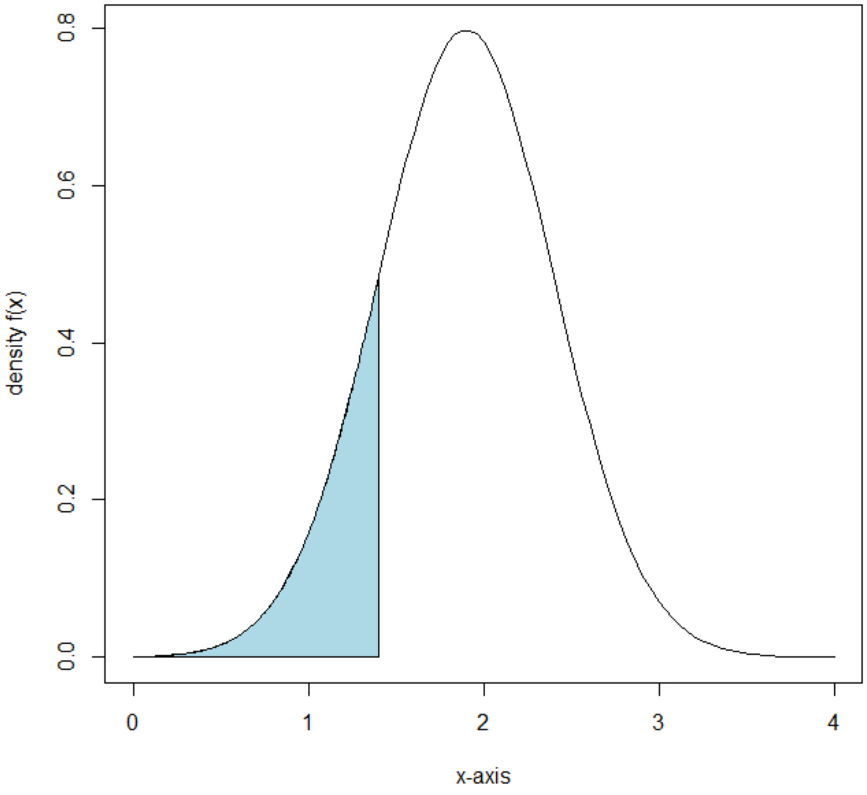
# Plotting a Normal Dist

```
> f<-function(x){dnorm(x,1.9,0.5)}
> plot(f,0,4,xlab="x-axis",ylab="density f(x)")
```

```
> x<-seq(0,1.4,0.01)
> polygon(c(0,x,1.4), c(0,f(x),0), col="lightblue")
```

Polygon allows you to draw (x,y)





# Plotting

- ❖ We can also use package “ggplot2” to plot.

- ❖ Resources (google):
  - ➢ package ggplot 2
  - ➢ 10 reasons to switch to ggplot
    - ➢ https://mandymejia.com/2013/11/13/10-reasons-to-switch-to-ggplot-7/
  - ➢ ggplot vs. base graph
    - ➢ https://flowingdata.com/2016/03/22/comparing-ggplot2-and-r-base-graphics/