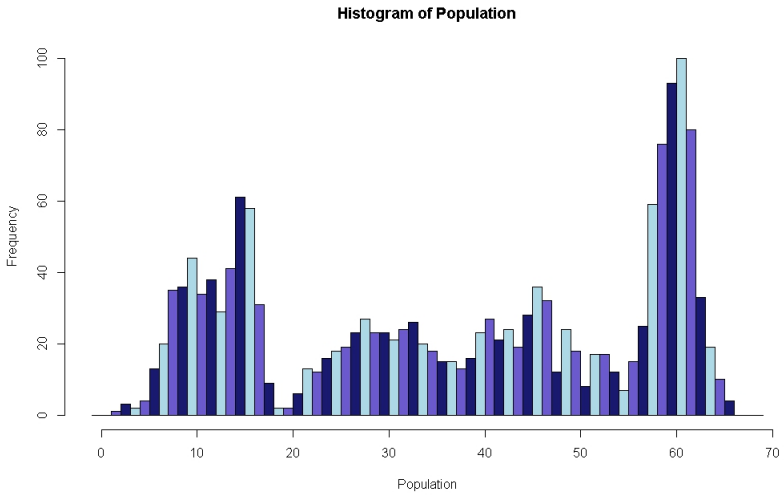


Sampling Distribution

- A probability distribution of a statistic obtained through a large number of samples drawn from a specific population.
- The sampling distribution of a given population is the distribution of frequencies of a range of different outcomes that could possibly occur for a statistic of a population.

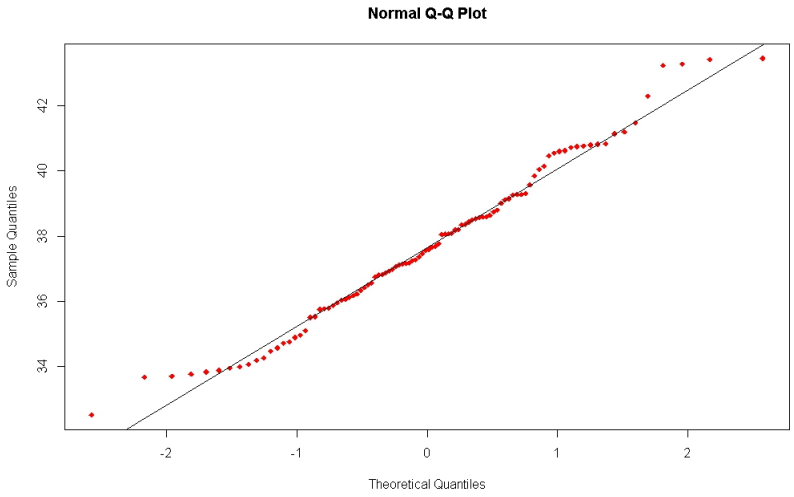
0.1 The Central Limit Theorem

The central limit theorem allows statisticians to use sample statistics to make inferences about the population parameters without knowing about the distribution of the parent population .

- The main aspect of the CLT that we shall consider is that many statistics (e.g sample mean and other related statistics, such as the sample variance) are normally distributed.
- Consider the population, characterized by the histogram on the next slide.
- While the population is not normally distributed, the population of sample statistic will be normally distributed.

Central Limit Theorem
Central Limit Theorem

- Consider an experiment whereby a sample of 60 members of this population was taken, and the following sample statistics were computed
  - Sample mean $\bar{X}$



  - Sample variance s^2
  - Sample median $\tilde{X}$

- This experiment was performed 100 times (i.e. 100 independent samples were taken).
- The sample statistics were collated by type and examined to determine normality.
- A data set can be tested for normality using a very simple graphical procedure known as the 'Normal Probability Plot', or Q-Q plot.
- A data set can be assumed to be normally distributed if the points on the Q-Q plot follow the trendline.

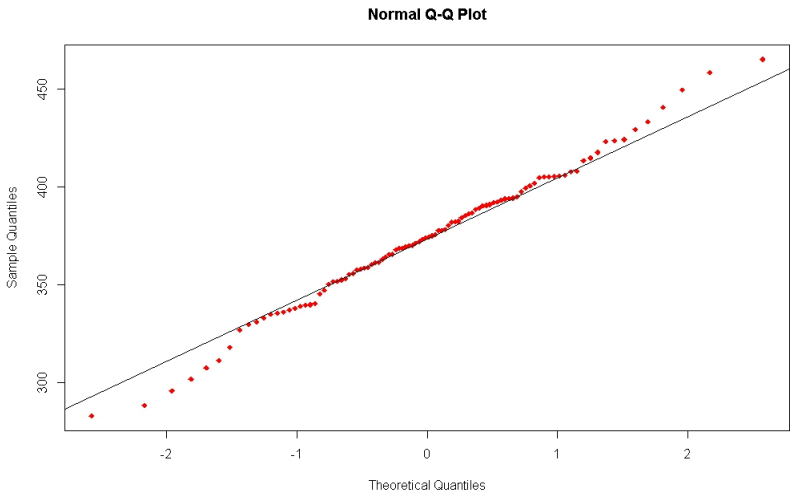
CLT: Sample Mean Q-Q plot
CLT: Sample Variance Q-Q plot
Central Limit Theorem: Sampling Distributions

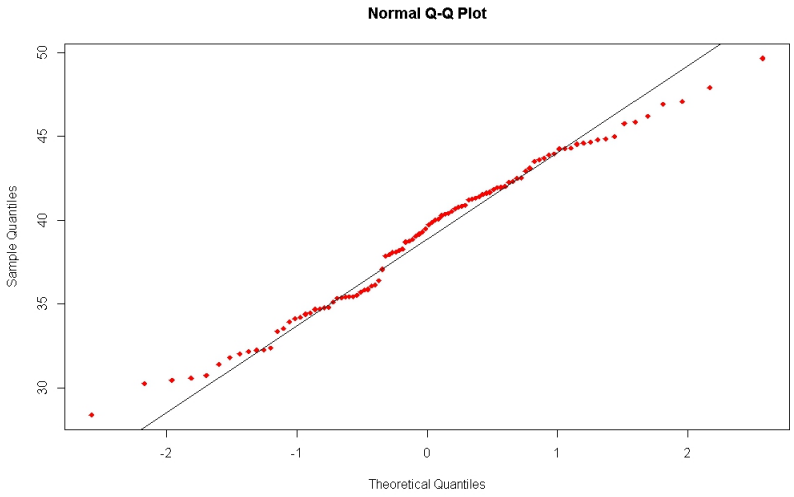
- In each of the three plots, the points follow the trend-line quite closely in each case.
- As we can see, the population of these statistics are normally distributed.
- We refer to these distributions as 'Sampling Distributions'.
- While the statistic that we will be dealing with in this module do have normal sampling distributions, it must be noted that many statistics have sampling distributions other than the normal distribution.

Quantile Functions

- The Cumulative Distribution Function is used to identify the probability of a random variable being below a threshold value k.

$$P(X \leq k)$$


Normal Q-Q Plot
Sample Quantiles
300
350
400
450
-2
-1
0
1
2
Theoretical Quantiles


Normal Q-Q Plot
Sample Quantiles
30
35
40
45
50
-2
-1
0
1
2
Theoretical Quantiles

- In short, we compute a probability values associated with a specified value.
- (In R, this is carried out using the p- family of functions.)
- With Quantile Functions, we are performing the opposite operation, i.e. for a specified probability, we determine the threshold value k.
- For some value p, we computed k such that

$$P(X \leq k) = p$$

- (In R, this is performed using the q- family of functions.)

Quantile Functions

- Recall that, from the Murdoch Barnes Tables, $P(Z \leq 1.96) = 0.975$ and $P(Z \leq 1.28) = 0.8997$

```
> qnorm(0.975)
[1] 1.959964
>
> qnorm(0.8997)
[1] 1.279844

Z = c(rnorm(300,10,3) , rnorm(150,15,1) , rnorm(100,24,3.5),rnorm(200,30,4) , rnorm(400,45,6),rnorm(500
Population =Z
hist(Population, breaks = -1:69, col=c("midnightblue","lightblue","slateblue"))



Var.Sample = numeric()
Median.Sample = numeric()
Mean.Sample = numeric()

for( i in 1:100)
{
Sample = sample(Z,60)


Mean.Sample = c(Mean.Sample,mean(Sample))
Median.Sample = c(Median.Sample,median(Sample))
Var.Sample = c(Var.Sample,var(Sample))

}
qqnorm(Median.Sample, pch =18, col="red")
qqline(Median.Sample)
#
qqnorm(Mean.Sample, pch =18, col="red")
qqline(Mean.Sample)
```

```
#



qqnorm(Var.Sample, pch =18, col="red")
qqline(Var.Sample)

qqnorm(Median.Sample, pch =18, col="red")
qqline(Median.Sample)
```

0.2 The Central Limit Theorem

- This theorem states that as sample size n is increased, the sampling distribution of the mean (and for other sample statistics as well) approaches the normal distribution in form, regardless of the form of the population distribution from which the sample was taken.
- For practical purposes, the sampling distribution of the mean can be assumed to be approximately normally distributed, even for the most non-normal populations or processes, whenever the sample size is $n > 30$.
- (For populations that are only somewhat non-normal, even a smaller sample size will suffice. A variation of the normal distribution can be used for such circumstances.)

0.3 The Central Limit Theorem

The central limit theorem allows statisticians to use sample statistics to make inferences about the population parameters without knowing about the distribution of the parent population .