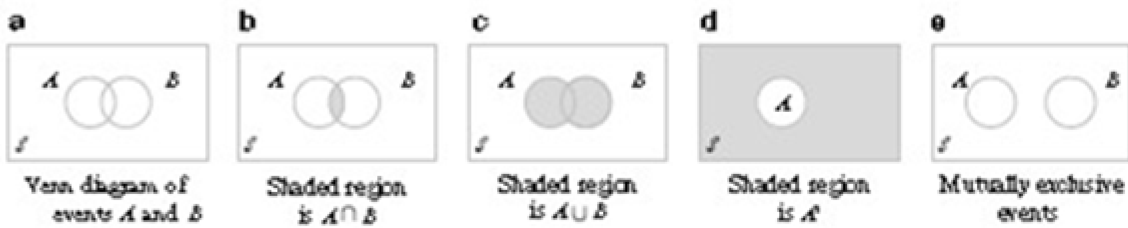
# Session Agenda

- Probability laws
- Discrete probability calculations
- Bayes' Theorem
- The Monty Hall Problem
- ROC curve
- Normal distribution
- Using the normal distribution
- Binomial distribution
- Normal approximation to binomial
- Sampling distributions
- Central Limit Theorem
- Assessing normality
- QQ Plot Examples
- Test #2 topics

# Probability Concepts

The table below describes the smoking habits of a group of asthma sufferers.

| | Nonsmoker | Occasional smoker | Regular smoker | Heavy smoker | Total |
|---|---|---|---|---|---|
| Men | 334 | 50 | 68 | 32 | 484 |
| Women | 357 | 30 | 89 | 37 | 513 |
| Total | 691 | 80 | 157 | 69 | 997 |




**d) Let A denote Nonsmoker. What is P[A]? 691/997 = 0.693**

$$P[A^c] = 1 - P[A] = 1 - 0.693 = 0.307$$

**e) Let B denote Heavy smoker. What is P[A and B]?**

$$P[A \text{ and } B] = P[A \cap B] = 0$$

$$P[A \text{ or } B] = P[A] + P[B] = (691 + 69)/997 = 0.762$$

**c) Let C denote Men. What is P[A or C]?**

$$P[A \cup C] = P[A] + P[C] - P[A \cap B] =$$

$$(484 + 691 - 334)/997 = 0.844$$

**b) What is P[A and C]? 334/997 = 0.335 Does $P[A \cap C] = P[A]P[C]$?**

$$P[A]P[C] = (0.693)(484/997) = 0.336$$

**Conditional Probability: What is the probability of a woman being selected if we restrict attention to nonsmokers? P[ woman | nonsmoker ] = 357/691 = 0.517.**


**Bluman, Allan G., "Probability DeMystified" 2nd ed. (2012), McGraw Hill**

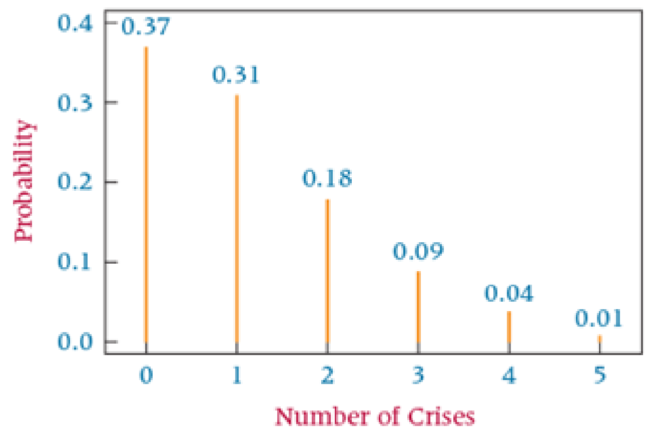
# Discrete Probability Calculations

## Mean, Variance and Standard Deviation

**An executive is considering out-of-town business travel for a given Friday. She recognizes that at least one crisis could occur on the day that she is gone and she is concerned about that possibility.**

Table 5.2 Discrete Distribution of Occurrence of Daily Crises

| NUMBER OF CRISES | PROBABILITY |
|---|---|
| 0 | .37 |
| 1 | .31 |
| 2 | .18 |
| 3 | .09 |
| 4 | .04 |
| 5 | .01 |




```
> average <- sum(crises*probability)
> average
[1] 1.15
> variance <- sum(probability*(crises - average)^2)
> round(variance, digits = 2)       # Variance
[1] 1.41
> round(sqrt(variance), digits = 2)  # Standard deviation
[1] 1.19
```

The mode (most frequent value) is 0 crises.
The median is 1 crisis. (37% less than 1, and 63% 1 or more).

## Using the complement for probability calculations

**Suppose the probability of one or more crises on a day is 0.05. How many days may the executive be gone before the probability of one or more days with a crisis exceeds 0.1 for that interval of time?**

**Let n denote the number of days the executive is gone. We will assume the days are independent so that the chances of a crisis on one day does not affect the chances on another. The probability of a string of uneventful days is (1 – 0.05)^n. The probability that one or more days in a string of n days has one or more crises is 1 – (1 – 0.05)^n. We solve 1- (1 – 0.05)^n >= 0.1 for n.**

```
> n <- ceiling(log(0.9)/log(1 - 0.05))
> n
[1] 3
```

# Bayes' Theorem Calculation

Approximately 1% of women aged 40-50 have breast cancer. A woman with breast cancer has a 90% chance of a positive test from a mammogram, while a woman without has a 10% chance of a false positive result. What is the probability a woman has breast cancer given that she just had a positive test? The answer is 9/(9 + 99)=9/108 = 0.0833.

http://www.math.cornell.edu/~mec/2008-2009/TianyiZheng/Bayes.html

| Confusion Matrix | | |
|---|---|---|
| | Has Breast Cancer | No Breast Cancer |
| | 1% | 99% |
| Negative Test | False Negative Rate = 10% | True Negative Rate = 90% |
| Positive Test | True Positive Rate = 90% | False Positive Rate = 10% |

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(A)P(B \mid A) + P(A')P(B \mid A')}$$

$A$ denotes the event a randomly selected woman has breast cancer.

$B$ denotes the event a randomly selected woman has a positive test.

$P(A) = 0.01$, $P(A') = 0.99$, $P(B|A) = 0.9$, $P(B|A') = 0.1$.

$$P(A|B) = \frac{0.9(0.01)}{0.9(0.01) + 0.1(0.99)} = \frac{0.009}{0.009 + 0.099} = \frac{9}{108}.$$

Increasing the true positive rate to 99%, increases the probability a woman has breast cancer given a positive test result to 50%.

# The Monty Hall Problem

**A contestant picks one of three doors at random. The probability of picking a door with a car behind it is 1/3. The contestant is shown no car is behind one of the doors not picked. What is the probability of winning if the contestant switches the choice of doors?**

**This is a decision problem. The prior probability the car is behind any door is 1/3.**

| | Behind Door | Not Behind Door |
|---|---|---|
| | 1/3 | 2/3 |
| Switch | P[winning\|door] = 0 | P[winning\|not door] = 1 |
| Don't Switch | P[winning\|door] = 1 | P[winning\|not door] = 0 |

**The probability of winning if the contestant switches:**

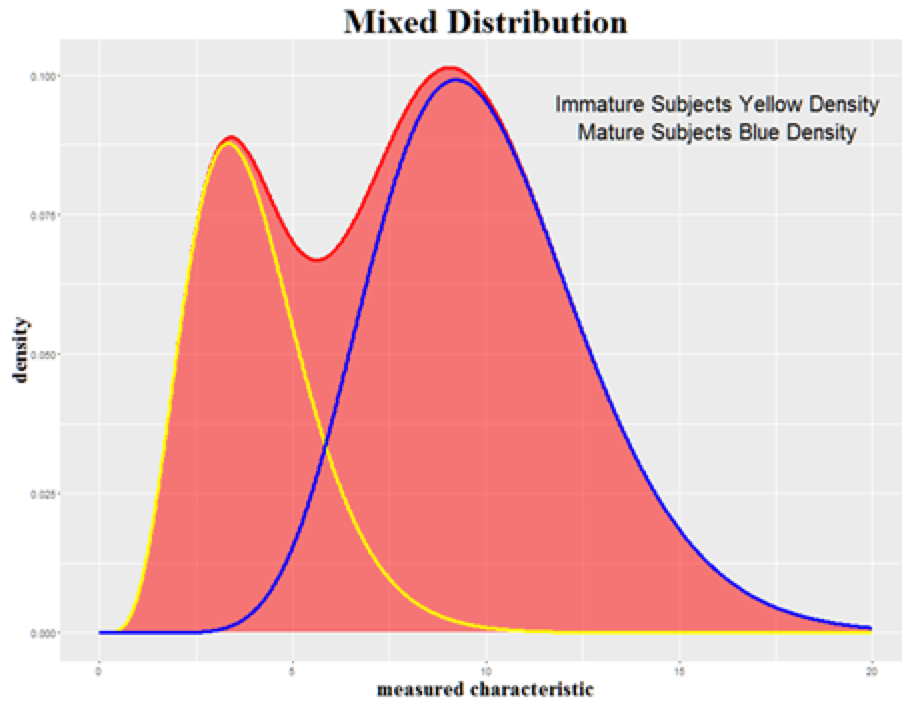
$$(0 \times 1/3) + (1 \times 2/3) = 2/3.$$

**The probability of winning if the contestant doesn't switch:**

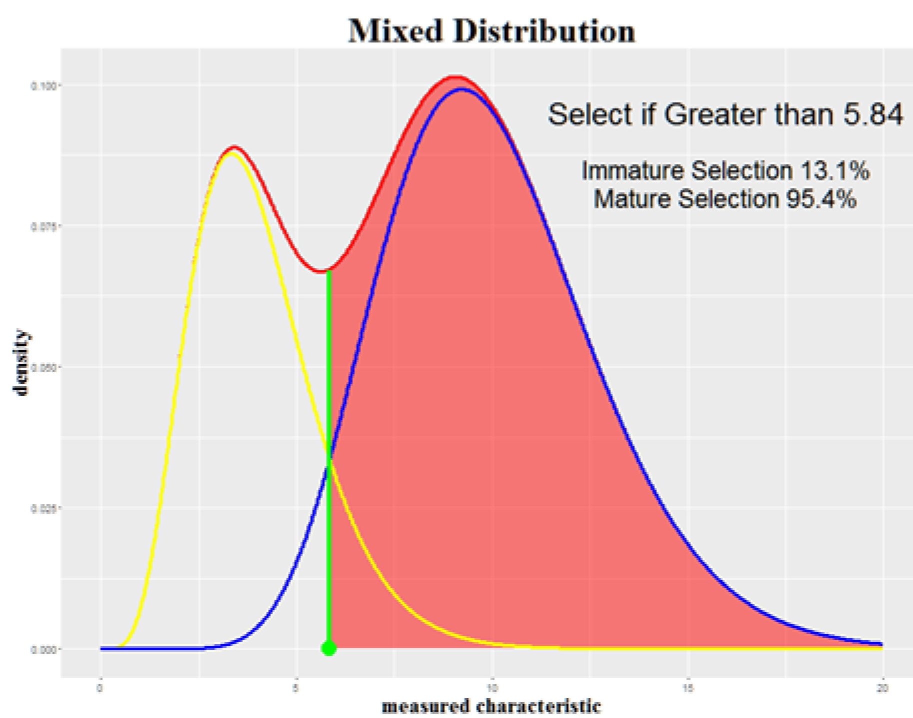
$$(1 \times 1/3) + (0 \times 2/3) = 1/3.$$

**http://angrystatistician.blogspot.com/2012/06/bayes-solution-to-monty-hall.html**

# A Selection Problem




To avoid sacrificing the subject, selection is based on a measured physical characteristic. A subject is classified as mature if this characteristic is greater than some agreed upon value such as 5.84.

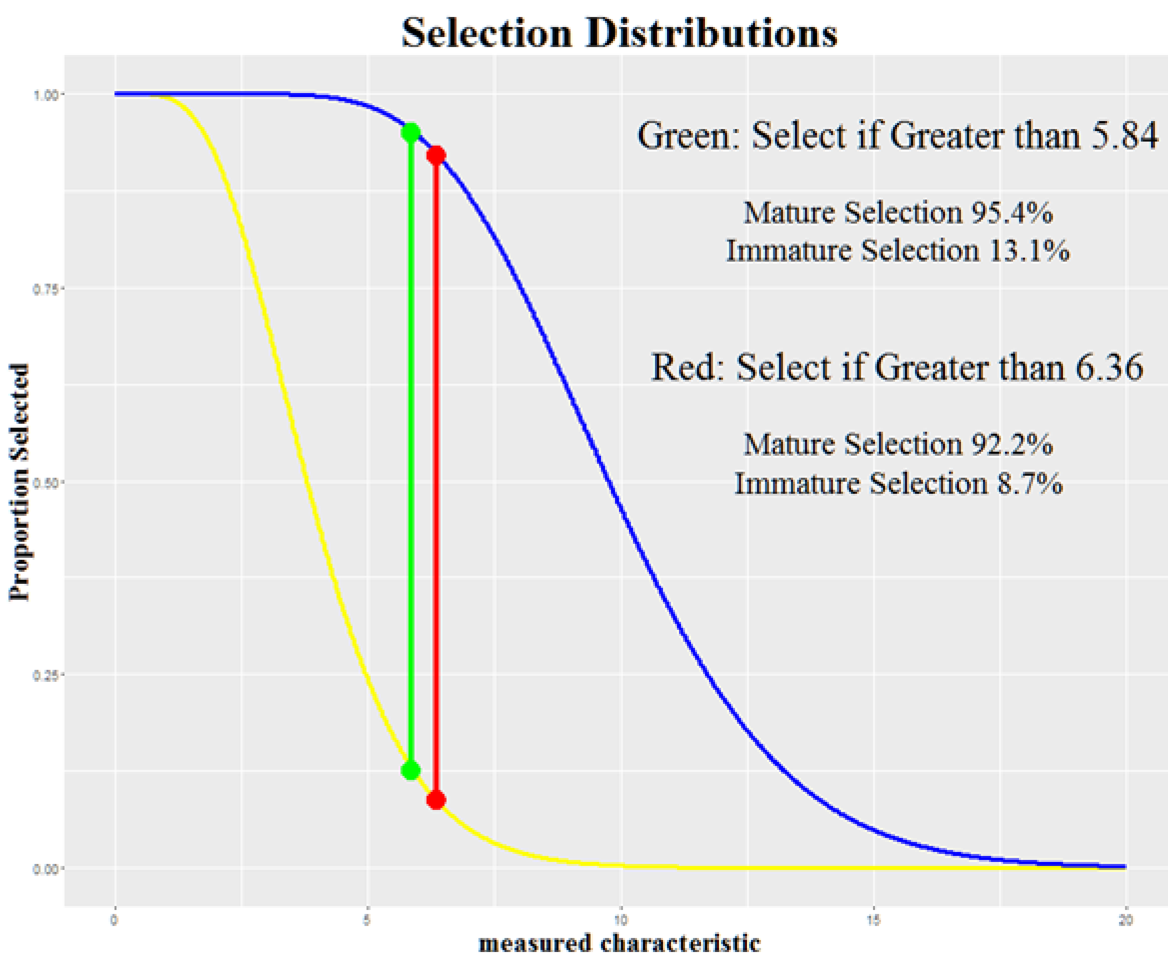


# Confusion Matrix and ROC Curve

| Confusion Matrix (select > 5.84) | | |
|---|---|---|
| | Immature Subject | Mature Subject |
| | 1/3 | 2/3 |
| Don't Select | True Negative Rate = 86.9% | False Negative Rate = 4.6% |
| Select | False Positive Rate = 13.1% | True Positive Rate = 95.4% |

Probability a randomly selected subject that exceeds 5.84 is mature:

$$\left[\frac{0.954*(2/3)}{0.954*(2/3)+0.131*(1/3)}\right]=0.935$$



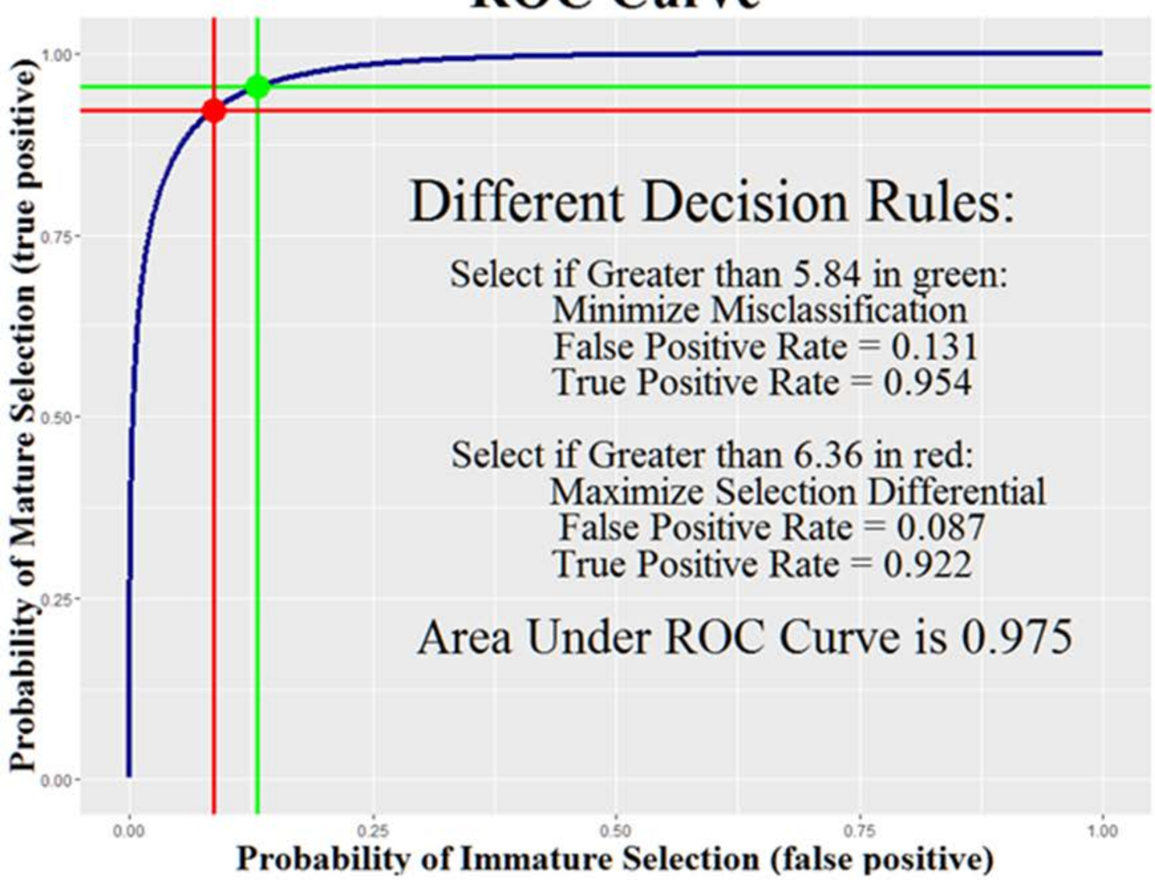
# Confusion Matrix and ROC Curve Continued

| Confusion Matrix (select > 6.36) | | |
|---|---|---|
| | **Immature Subject** | **Mature Subject** |
| | **1/3** | **2/3** |
| **Don't Select** | **True Negative Rate = 92.3%** | **False Negative Rate = 7.8%** |
| **Select** | **False Positive Rate = 8.7%** | **True Positive Rate = 92.2%** |

**Probability a randomly selected subject that exceeds 6.36 is mature:**

$$\left[\frac{0.922*(2/3)}{0.922*(2/3)+0.087*(1/3)}\right]=0.955$$

## ROC Curve



# Distributions in R

**Table 14.1** Statistical Distributions and their Functions

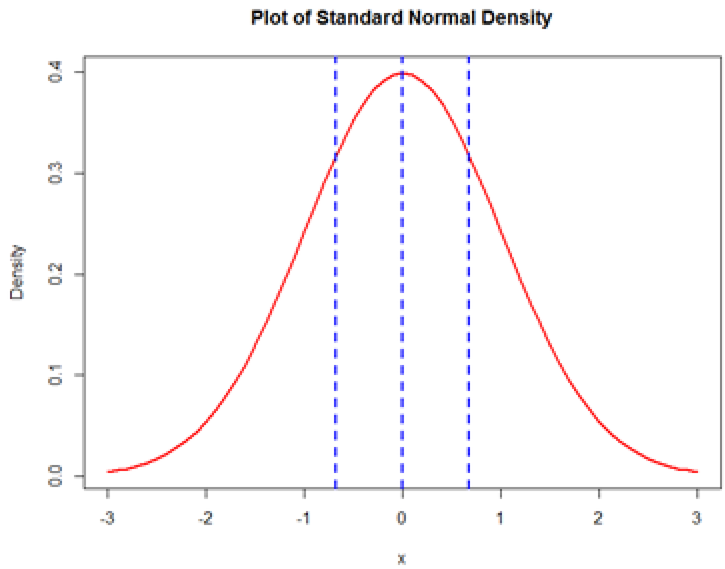
| Distribution | Random Number | Density | Distribution | Quantile |
|---|---|---|---|---|
| Normal | rnorm | dnorm | pnorm | qnorm |
| Binomial | rbinom | dbinom | pbinom | qbinom |
| Poisson | rpois | dpois | ppois | qpois |
| t | rt | dt | pt | qt |
| F | rf | df | pf | qf |
| Chi-Squared | rchisq | dchisq | pchisq | qchisq |
| Gamma | rgamma | dgamma | pgamma | qgamma |
| Geometric | rgeom | dgeom | pgeom | qgeom |
| Negative Binomial | rnbinom | dnbinom | pnbinom | qnbinom |
| Exponential | rexp | dexp | pexp | qexp |
| Weibull | rweibull | dweibull | pweibull | qweibull |
| Uniform (Continuous) | runif | dunif | punif | qunif |
| Beta | rbeta | dbeta | pbeta | qbeta |
| Cauchy | rcauchy | dcauchy | pcauchy | qcauchy |
| Multinomial | rmultinom | dmultinom | pmultinom | qmultinom |
| Hypergeometric | rhyper | dhyper | phyper | qhyper |
| Log-normal | rlnorm | dlnorm | plnorm | qlnorm |
| Logistic | rlogis | dlogis | plogis | qlogis |

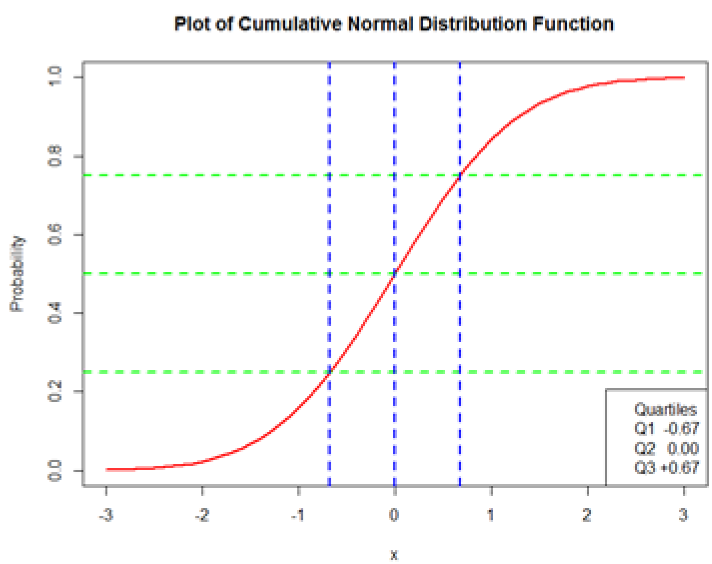
Lander *R for Everyone* page 185

## Useful for many purposes:

**Simulation**
**Exact calculations**
**QQ Charts**
**Displays**

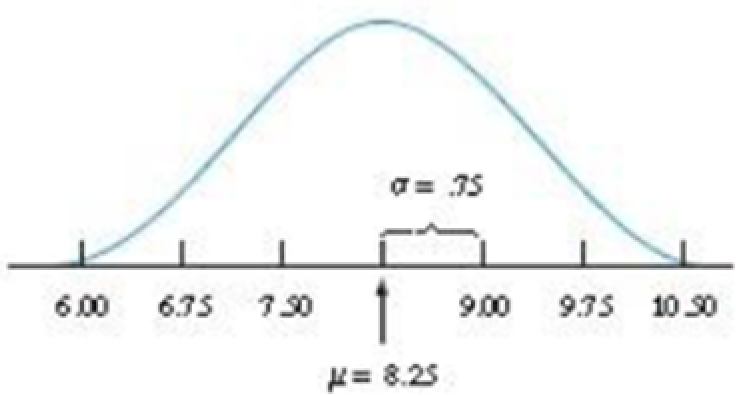
# Normal Distribution

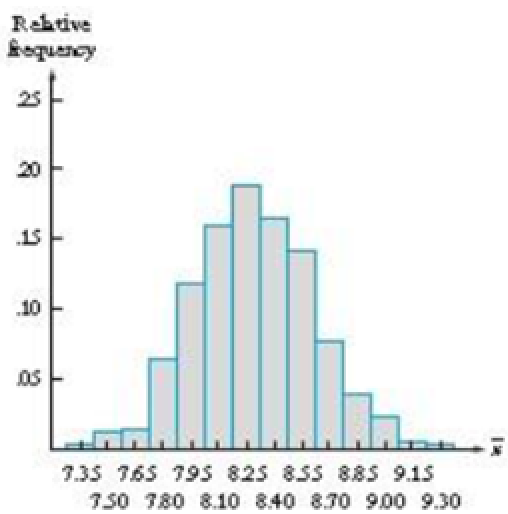


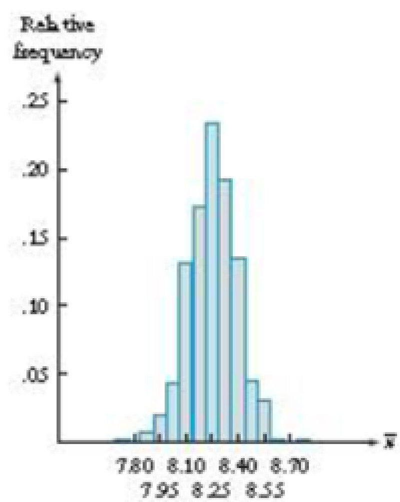

```
> dnorm(0, 0, 1)
[1] 0.3989423
> qnorm(c(0.25, 0.5, 0.75), mean = 0, sd = 1, lower.tail = TRUE)
[1] -0.6744898  0.0000000  0.6744898
> pnorm(c(-0.6744898, 0.0000000,0.6744898), 0, 1, lower.tail = TRUE)
[1] 0.25 0.50 0.75
> pnorm(c(-0.6744898, 0.0000000,0.6744898), 0, 1, lower.tail = FALSE)
[1] 0.75 0.50 0.25
```

# Sampling Distribution of a Normal Random Variable




**n = 5**




**n = 30**

Given a random variable $X$. Suppose that the **population distribution of $X$** is known to be normal, with mean $\mu$ and variance $\sigma^2$, that is, $X \sim N(\mu, \sigma)$. Then, for <u>any</u> sample size $n$, it follows that the **sampling distribution of $\bar{X}$** is normal, with mean $\mu$ and variance $\dfrac{\sigma^2}{n}$, that is, $\bar{X} \sim N\left(\mu, \dfrac{\sigma}{\sqrt{n}}\right)$.

# Using the Normal Distribution

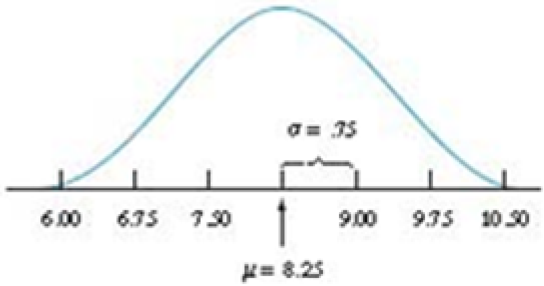



Figure 6.5 Normal distribution, with $\mu$ = 8.25 and $\sigma$ = .75

$$\frac{x-\mu}{\sigma} = \frac{9.0-8.25}{0.75} = 1.0$$

**Probability a standard normal variable >= 1.0?**

```
> pnorm(1, 0, 1, lower.tail = FALSE)
[1] 0.1586553
> 1 - pnorm(1, 0, 1, lower.tail = TRUE)
[1] 0.1586553
> pnorm(9, 8.25, 0.75, lower.tail = FALSE)
[1] 0.1586553
```

---

1) For women aged 18-24, systolic blood pressures (in mm Hg) are normally distributed with a mean of 114.8 and a standard deviation of 13.1. If 36 women are selected at random from the USA population of women aged 18-24, find the probability that their mean systolic blood pressure will be less than 110 mm Hg. Assume that the sampling is done without replacement.

Solution:

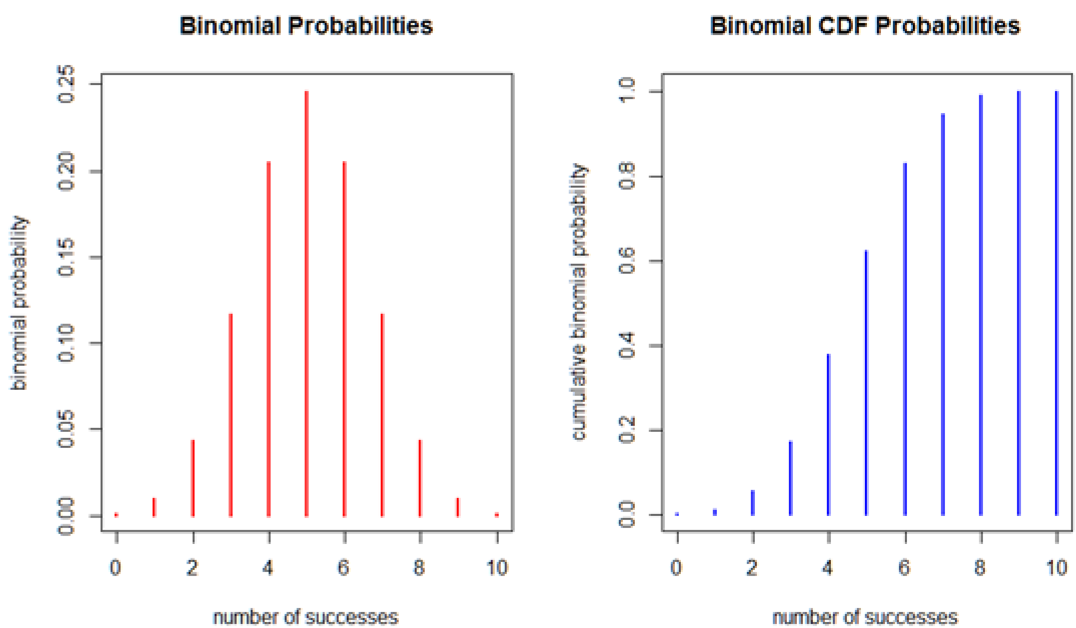
The sampling distribution for the mean of a random sample of size 36 has an expected mean of 114.8 mm Hg and a standard deviation of 13.1/sqrt(36) = 2.183 mm Hg. Calculate the z-score using 110 mm Hg. z = (110 – 114.8)/2.183 = -2.199. Using the normal distribution tables the probability is approximately 0.0139.

```
> z <- ((110-114.8)/(13.1/sqrt(36)))
> pnorm(z,0,1,lower.tail = TRUE)
[1] 0.0139577
> pnorm(110, 114.8, 13.1/sqrt(36), lower.tail = TRUE)
[1] 0.0139577
```

# Binomial Distribution

- The experiment involves n identical trials.
- Each trial has only two possible outcomes.
- Each trial is independent of the previous trials.
- The probability p of success is constant and q = (1-p) is the probability of failure.

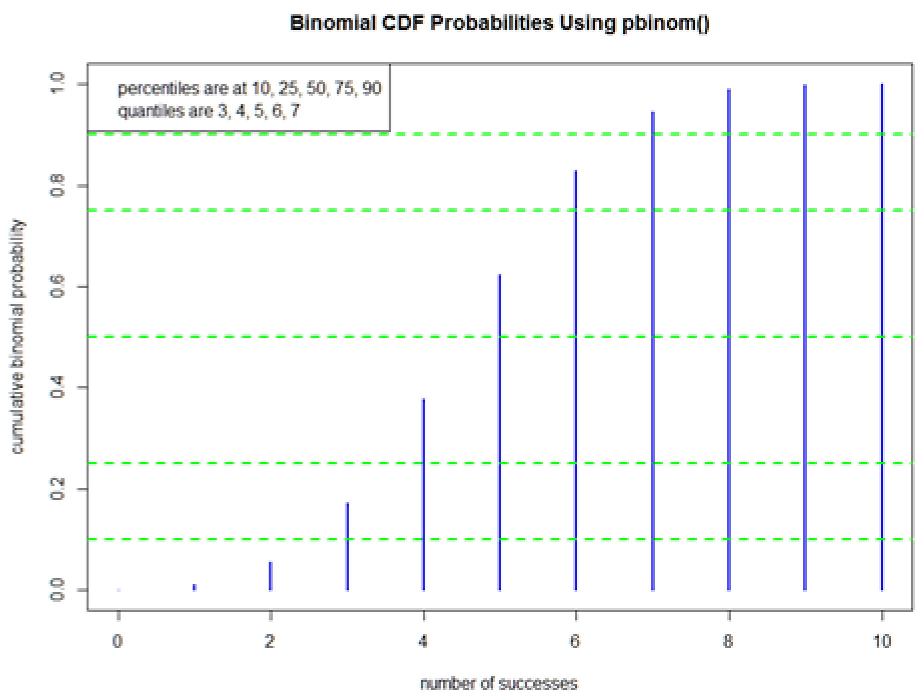
**Example with x = number of successes, n = 10 and p = 0.5. Use dbinom(x, n, p) for the binomial probabilities, and pbinom(x, n, p) for the cumulative distribution function.**




```
> pbinom(5,10,0.5,lower.tail=TRUE)
[1] 0.6230469
> sum(dbinom(c(0,1,2,3,4,5),10,0.5))
[1] 0.6230469
> pbinom(5,10,0.5,lower.tail=FALSE)
[1] 0.3769531
> sum(dbinom(c(6,7,8,9,10),10,0.5))
[1] 0.3769531
```

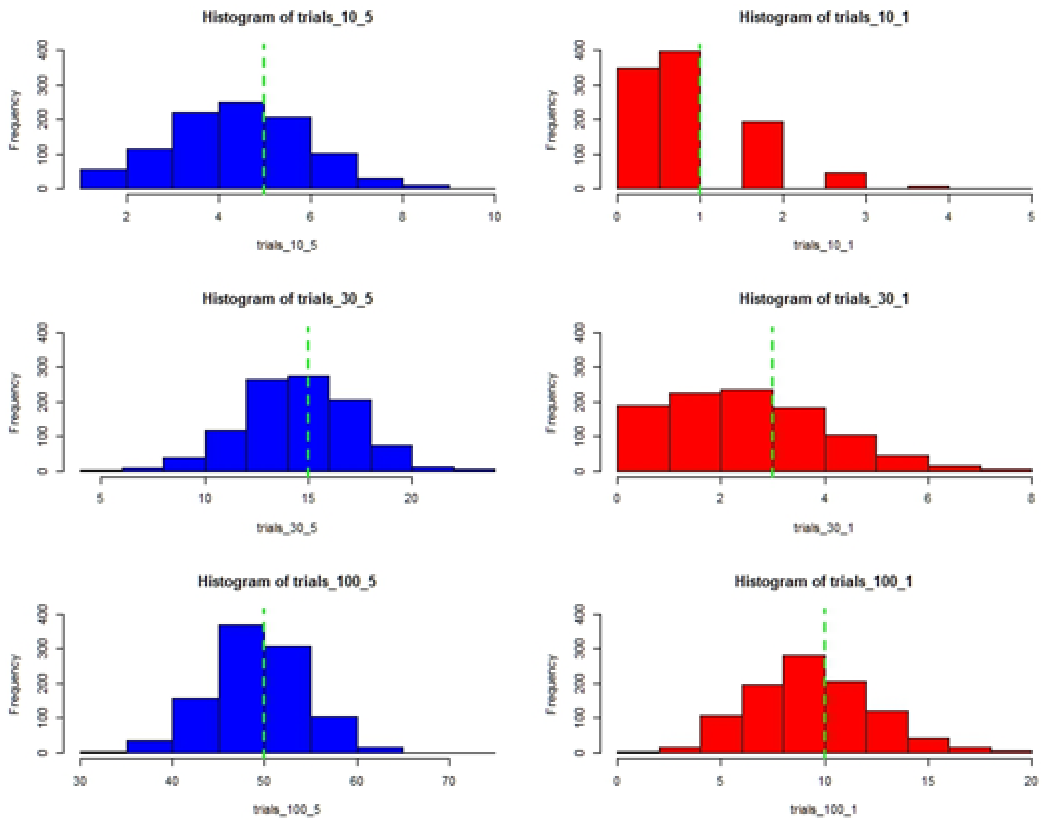
# Binomial Distribution Continued

**Quantiles can be determined using qbinom(). Random binomial outcomes can be generated using rbinom(). Use help() in R to obtain basic information on these functions.**




```
> qbinom(c(0.1, 0.25, 0.5, 0.75, 0.9), 10, 0.5, lower.tail=TRUE)
[1] 3 4 5 6 7
> pbinom(c(3, 4, 5, 6, 7), 10, 0.5, lower.tail=TRUE)
[1] 0.1718750  0.3769531  0.6230469  0.8281250  0.9453125
>
> qbinom(c(0.1, 0.25, 0.5, 0.75, 0.9), 10, 0.5, lower.tail=FALSE)
[1] 7 6 5 4 3
> sum(dbinom(c(8, 9, 10), 10, 0.5))
[1] 0.0546875
> sum(dbinom(c(7, 8, 9, 10), 10, 0.5))
[1] 0.171875
> pbinom(c(7, 6, 5, 4, 3), 10, 0.5, lower.tail=FALSE)
[1] 0.0546875  0.1718750  0.3769531  0.6230469  0.8281250
```

# Normal Distribution Approximation to the Binomial Distribution




**Rules: np > 5 and n(1-p) > 5**

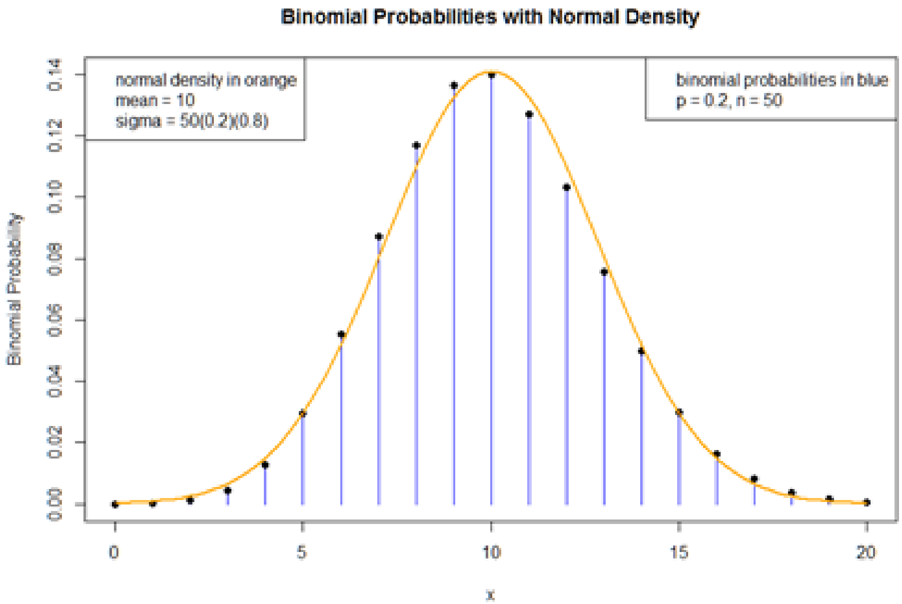
**binomial: mean = np   variance = np(1-p)**

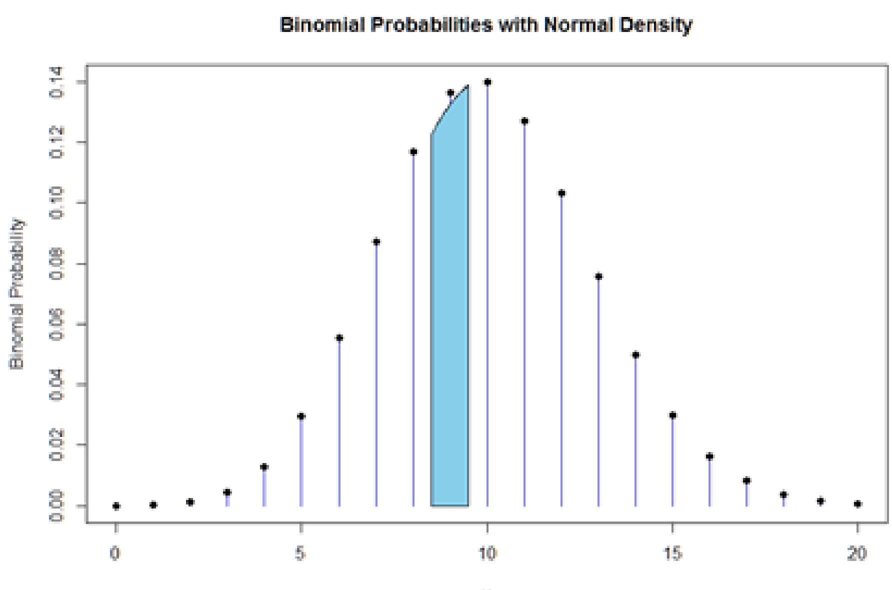
Find the probability of less than 9 successes, and the probability of 9 or fewer successes if n = 100 and p = 0.1.

```
> pnorm(8.5, 10, 3, lower.tail = TRUE)      [1] 0.3085375
> pbinom(8, 100, 0.1, lower.tail = TRUE)    [1] 0.3208739
> sum(dbinom(seq(0,8), 100, 0.1))           [1] 0.3208739

> pnorm(9.5, 10, 3, lower.tail = TRUE)      [1] 0.4338162
> pbinom(9, 100, 0.1, lower.tail = TRUE)    [1] 0.4512902
> sum(dbinom(seq(0,9), 100, 0.1))           [1] 0.4512902
```

# Continuity Correction

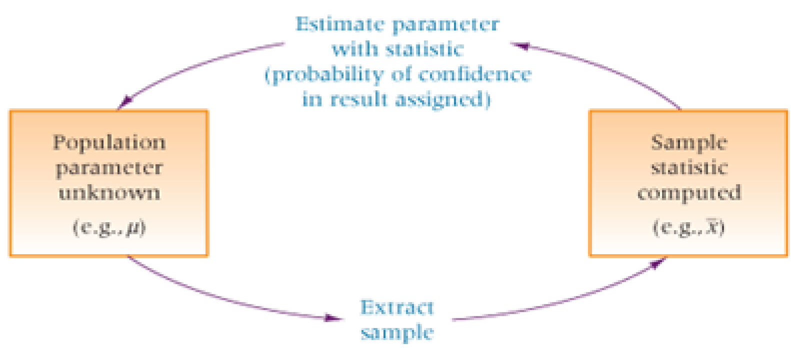



```
> dbinom(9, 50, 0.2)
[1] 0.1364088
> pnorm(9.5, mean = 10, sd = sqrt(10*(1-0.2)), lower.tail = TRUE )-
  pnorm(8.5, mean = 10, sd = sqrt(10*(1-0.2)), lower.tail = TRUE )
[1] 0.1319004
```

# Sampling Distributions

**FIGURE 4.1** Probability in the Process of Inferential Statistics

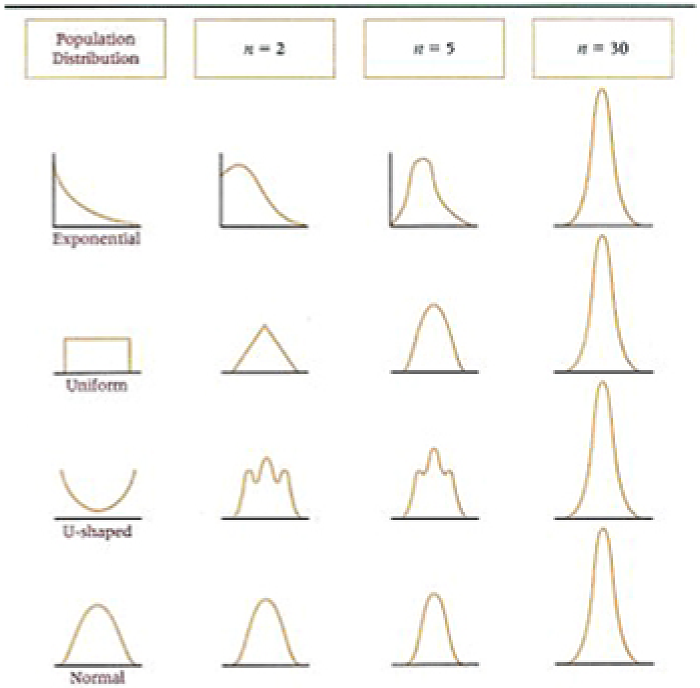



**Definition: A sampling distribution is the probability distribution of a given statistic based on a random sample. It represents the distribution of the statistic computed for all possible random samples from a given population.**

A sampling distribution depends on the population distribution, the statistic being considered, the sampling procedure used and the sample size.

Depending on the population and the statistic, the formulas for the sampling distribution may be complicated and not exist in closed-form. Approximations become necessary through Monte-Carlo simulations, bootstrap methods or asymptotic distribution theory.

# Central Limit Theorem Convergence



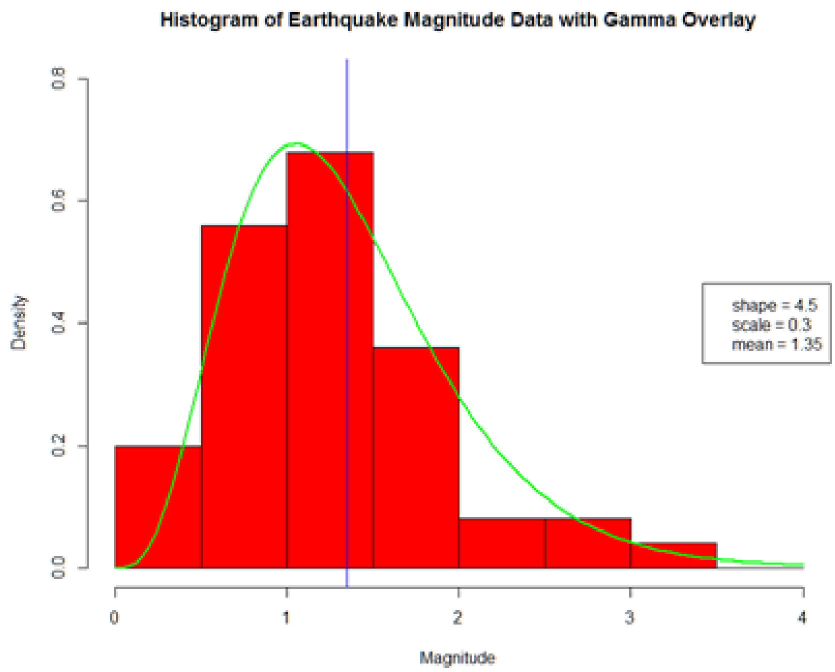

**The mean $\bar{X}$ of a random sample drawn from a population with mean $\mu$ and standard deviation $\sigma$ can be assumed to have approximately a normal distribution with mean $\mu$ and standard deviation $\sigma / \sqrt{n}$ if n is large enough.**

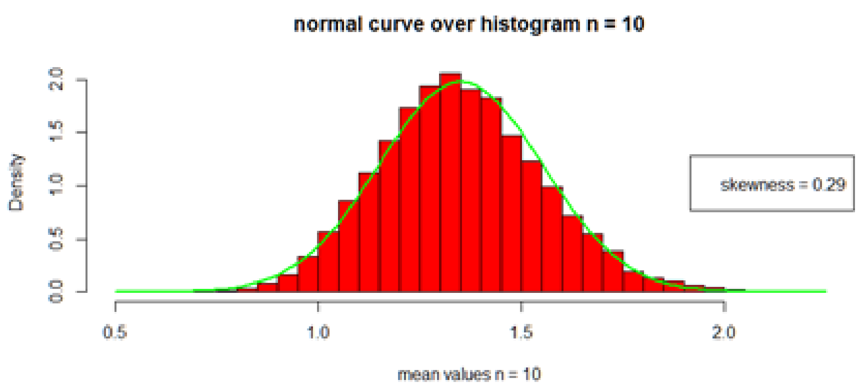
## How large must be the sample size n?

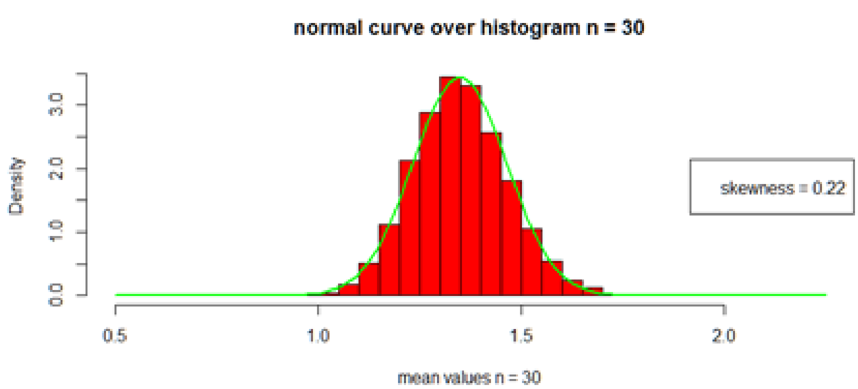
**Black (page 241) "…in this text (as in many others), a sample of size 30 or larger will suffice…." Wilcox states (page 90) in general n >= 40 will suffice.**

**These rules work as long as the population distribution is well behaved as above. This is not always the case. If there is substantial asymmetry, multiple peaks or extreme outliers present, these rules break down.**

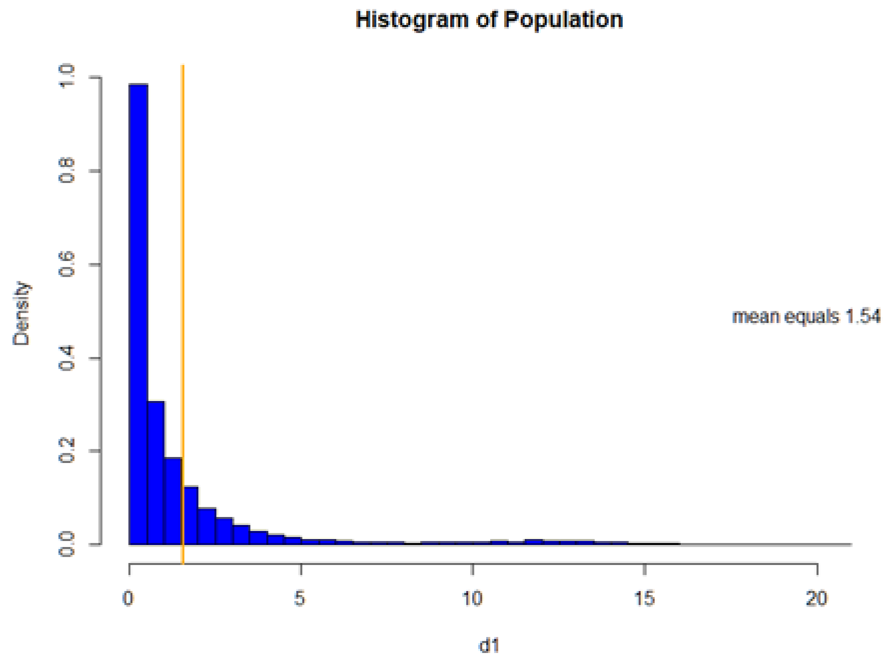
# Sampling Earthquake Magnitude Data

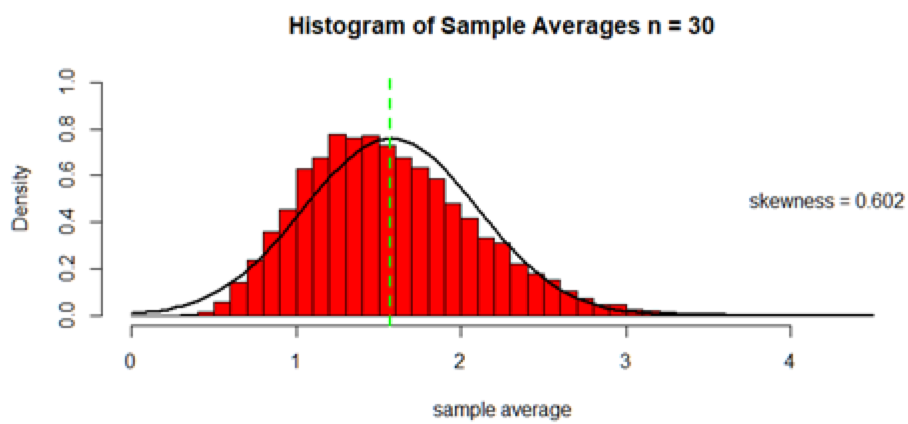


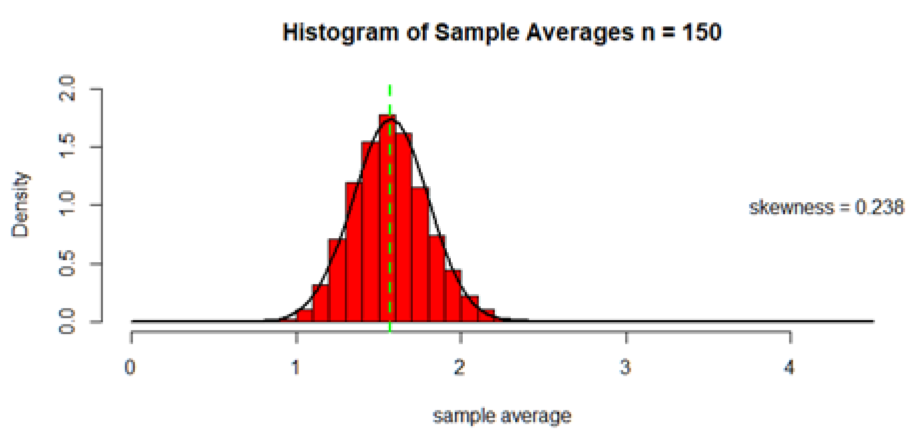



# Sampling Asymmetric Distribution



# What Do These Examples Have in Common?

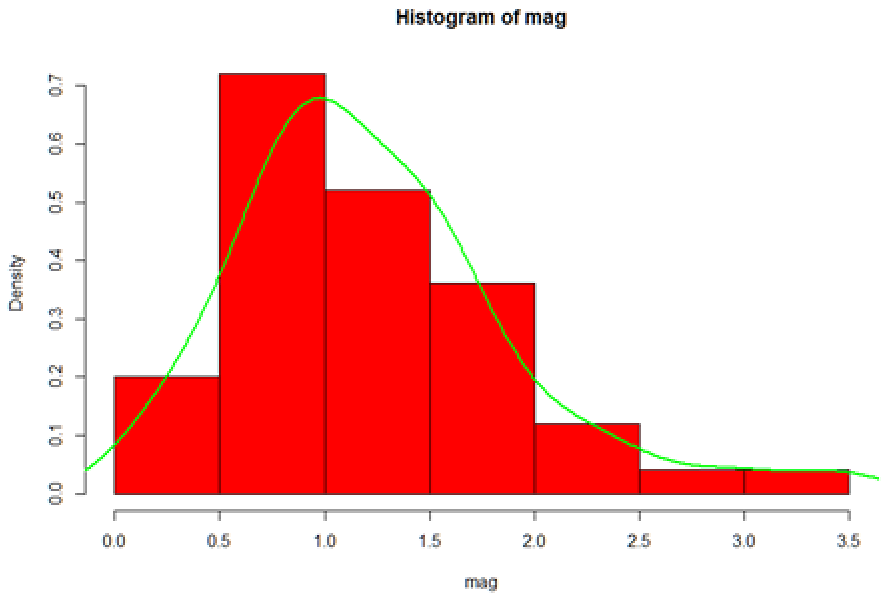
- **Size of loan defaults**
- **Load on web server**
- **Monthly maximum rainfall**
- **Income distribution**
- **Stock price distribution**
- **Maintainable system repair time**

**Data should <u>not</u> be considered arising from a normal distribution if the following are observed:**

- **Histogram departs dramatically from a bell shape.**
- **Outliers are present (more than a couple).**
- **Normal quantile plot shows one or both of the following:**
    - **The points do not lie reasonably close to a straight line.**
    - **The points show some systematic pattern that is not a straight-line pattern.**

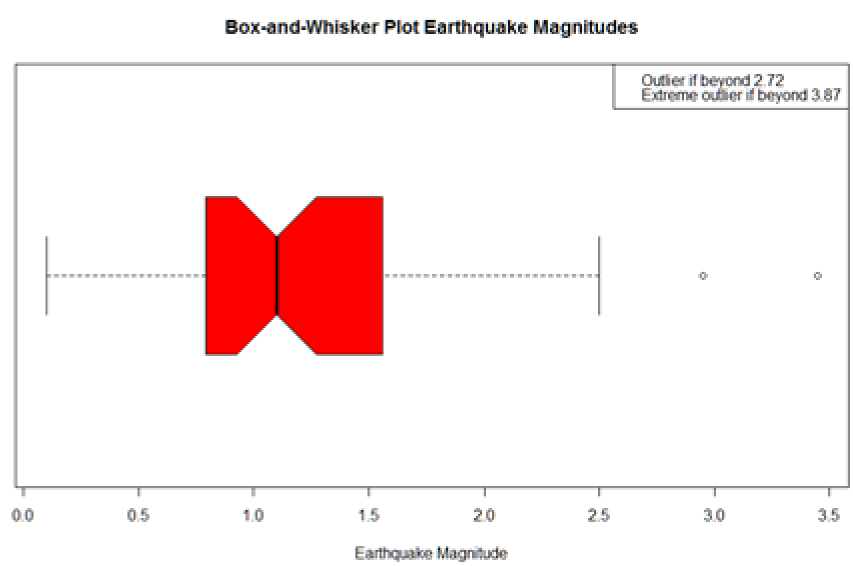
**Judgment and critical thinking are needed to make practical sense of data. Real data usually are not perfect. The presence of outliers is a case in point. The criteria given above may be used to evaluate convergence to normality of a sampling distribution.**

# Kernel Density Estimation and Box-and-Whisker Plot




```
hist(mag,freq=FALSE, col = "red")
lines(density(mag), col = "green", lwd = "2" )

library(moments)
skewness(mag)  [1] 1.068522
kurtosis(mag)    [1] 4.636123
```



# Box-and-Whisker Plot, and the Normal Distribution

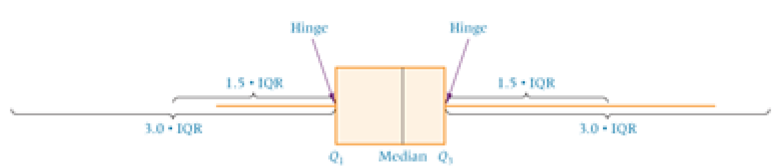
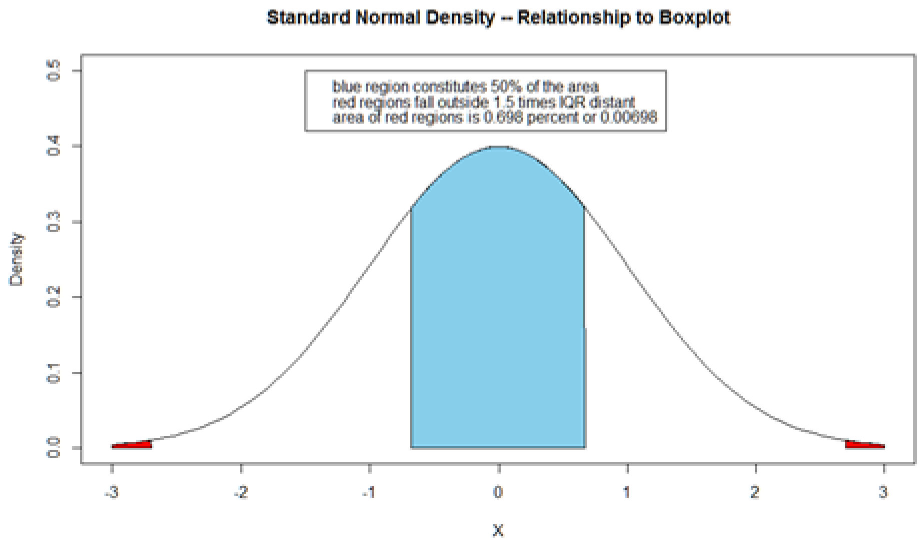
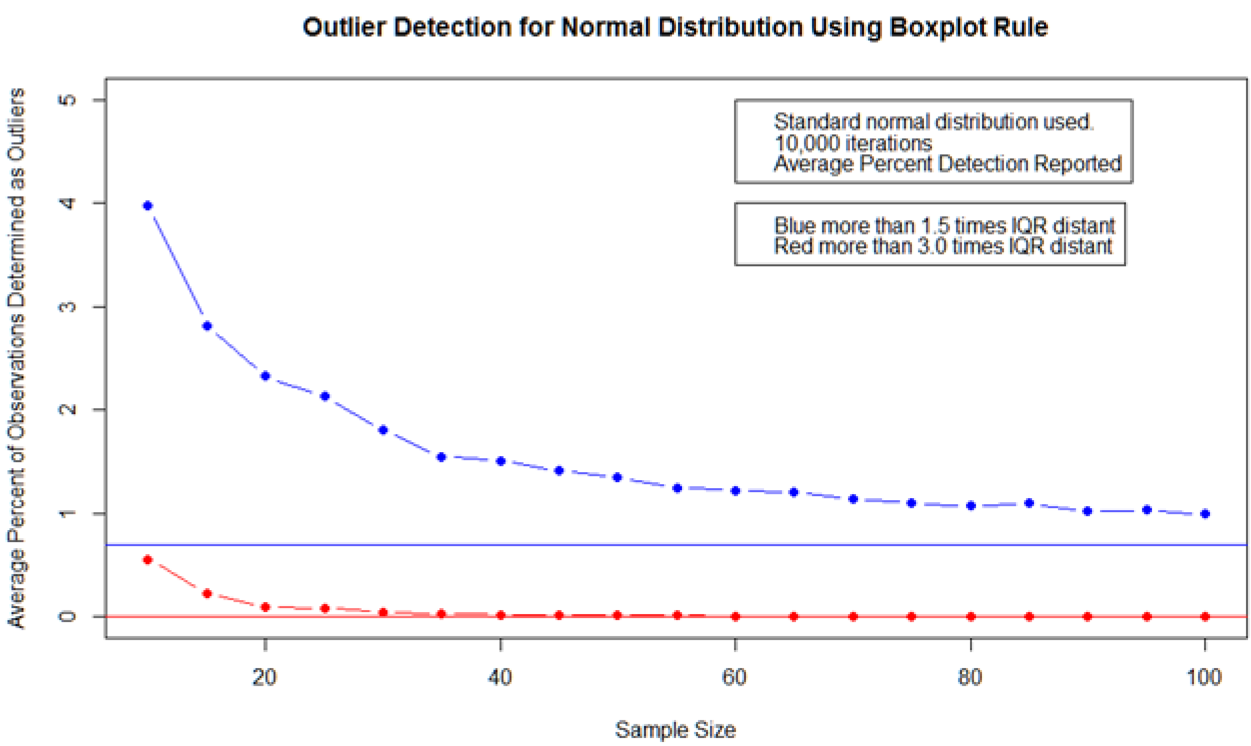


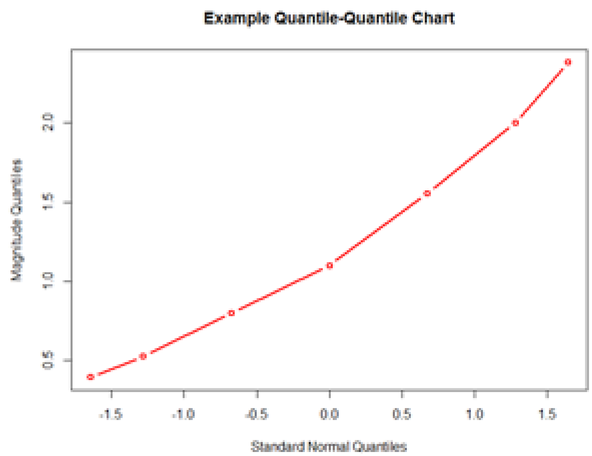

FIGURE 3.13 Box-and-Whisker Plot



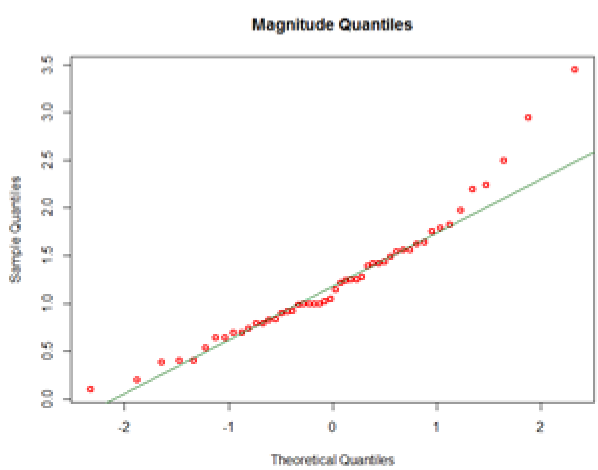



# Quantile-Quantile Charts

**Problem: Find the quantile for the 35th percentile of a standard normal distribution.**
**> qnorm(0.35, mean = 0, sd = 1, lower.tail = TRUE) [1] -0.3853205**




```
                          5%         10%           25%  50%         75%       90%       95%
magnitude_quantiles  0.394500  0.526000  0.8000000 1.1 1.5550000 2.002000 2.383000
normal_quantiles    -1.644854 -1.281552 -0.6744898 0.0 0.6744898 1.281552 1.644854
```

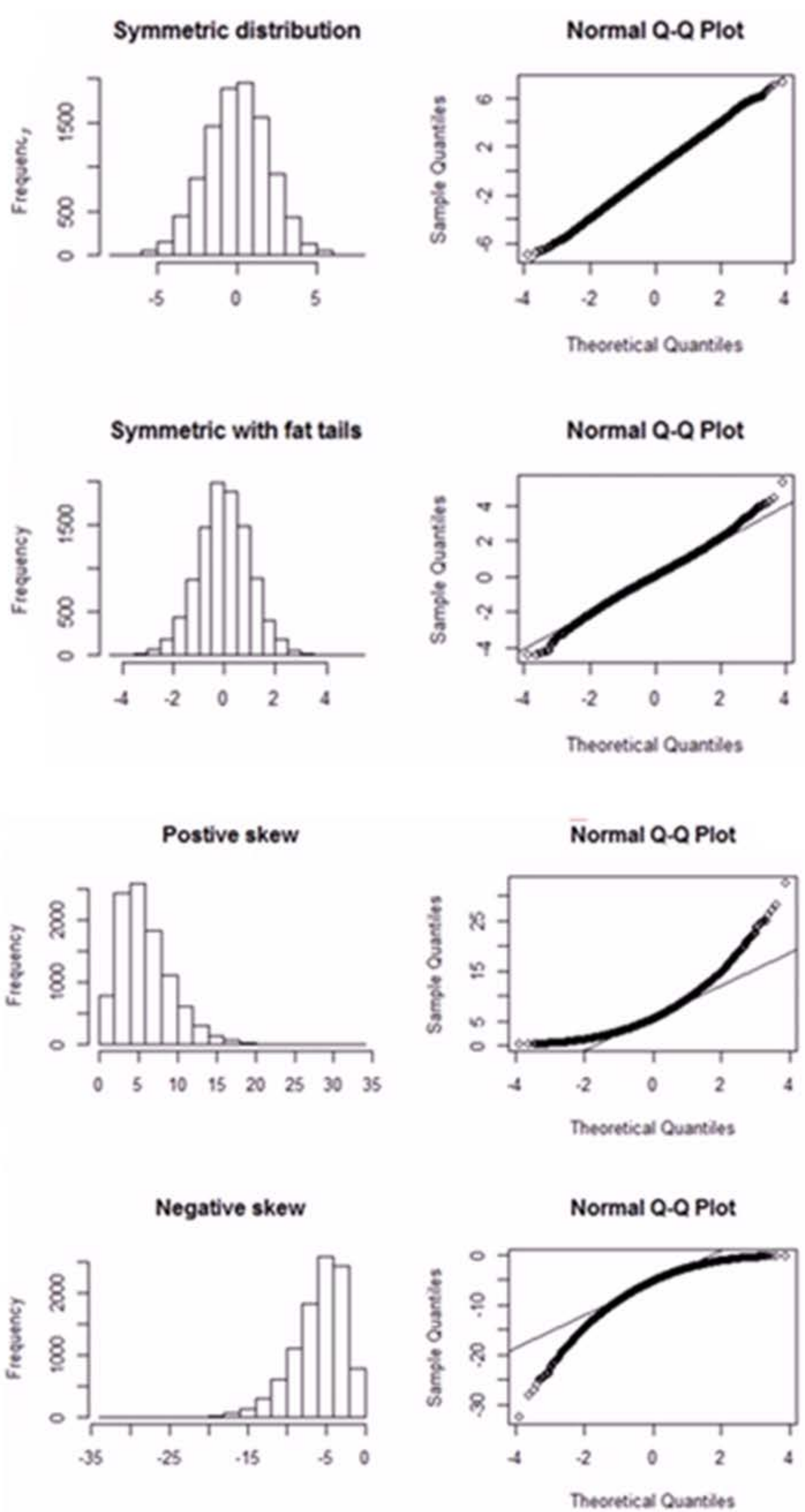



qqnorm()
qqline()

To calculate skewness and kurtosis, use skewness() and kurtosis() in the package ‘moments’.

https://www.youtube.com/watch?v=X9_ISJ0YpGw

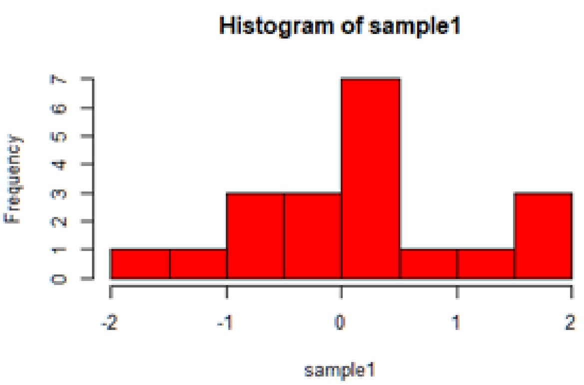
# QQ Plot Examples

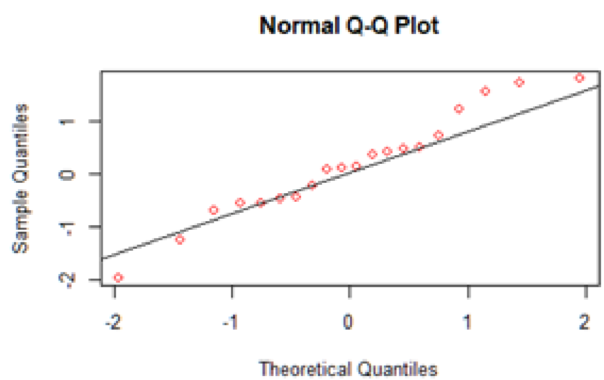


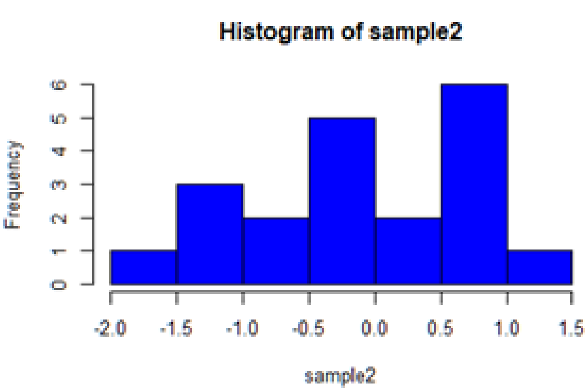
# Extra Credit Problem

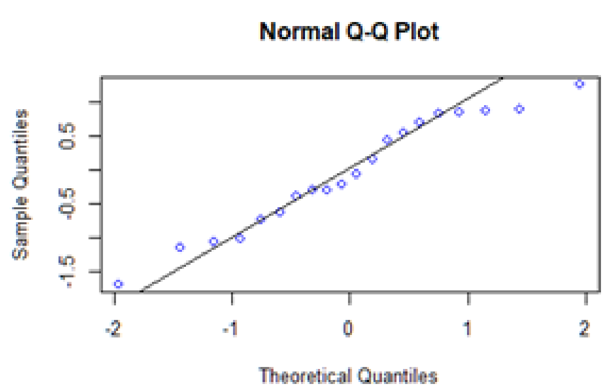
**Investigate the variability in the skewness and kurtosis statistics when sampling from a normal distribution.**

```
> require(moments)
> set.seed(123)
> sample1 <- rnorm(20, mean = 0, sd = 1)
> sample2 <- rnorm(20, mean = 0, sd = 1)
> c(skewness(sample1), kurtosis(sample1))
[1] -0.0674934  2.7170186
> c(skewness(sample2), kurtosis(sample2))
[1] -0.2098007  1.9852928
```



# Some Sync Session Learning Points

- The field of statistics is more a part of science than a branch of mathematics. It is not algorithmic. As more or better data come available a statistical model may be revised and conclusions may change.
- Application of statistical methods to data requires judgment to best represent the data as they are.
- A histogram, constructed from a voluntary response sample, may or may not represent the population distribution for the characteristic being measured.
- Definition: A sampling distribution is the probability distribution of a given statistic based on a random sample. It represents the distribution of the statistic computed for all possible random samples from a given population.
- A sample size of 40 is not always sufficient to justify use of the central limit theorem.
- The sample mean for a random sample drawn from a normal distribution has a sampling distribution which is normal.
- Under certain conditions the binomial, Poisson and Hypergeometric distributions are similar and may be used to approximate each other.
- Useful criteria for judging if a sample does not arise from a normal distribution are:
    - Histogram departs from a bell shape.
    - More than a few outliers are present.
    - Data plotted on a QQ chart does not lie close to a straight line, and/or shows a systematic pattern.
- An ROC curve is a plot of the true positive rate against the false positive rate for the different possible cutpoints of a binary classifier.

# Topics for Test #2

- **Discrete distributions**
  - **Binomial and Poisson**

- **Classical probability calculations**
  - **Combinations and permutations**

- **Bayes' formula**

- **Normal distribution probabilities**
  - **z-formula calculations**

- $\bar{X}$ **sampling distribution calculation**

- **Binomial continuity correction**

- **Normal distribution quantiles**

**Don't forget to start working on the data analysis assignment.**