1 Continuous Random Variable - NORMAL DISTRIBUTION

1.1 Probability Density Functions

Continuous random variables have supports that look like

$$S_X = [a, b] \; or (a, b) \tag{1}$$

or unions of intervals of the above form. Examples of random variables that are often taken to be continuous are:

- the height or weight of an individual,
- other physical measurements such as the length or size of an object, and
- duration of time (usually).

Every continuous random variable X has a probability density function (PDF) denoted f_X associated with it that satisfies three basic properties:

- $f_X(x) > 0$ for $x \in S_X$,
- $\int_{x \in S_X} f_X(x)dx = 1$, and
- $P(X \in A) = \int_{x \in A} f_X(x)dx$, for an event $A \subset S_X$.

1.1.1 Graphing Functions

curve(expression, xmin, xmax)

```
expression an expression
or function involving x
xmin min value of x to p
lot
xmax max value of x to p
lot
```

In [ ]:

```
par(mfrow = c(2,2))
curve(x^2,-10, 10)
curve(x^3 , -10, 10)
curve(sin(x), -2*pi,2*pi )
curve(sin(pi*x )/(pi*x),-5,5)
```

1.2 Cumulative distribution function F_X

It is defined by $F_X(t) = P(X \leq t)$ for $-\infty < t < \infty$ and

$$F_X(t) = P(X \leq t) = \int_{-\infty}^{t} f_X(x)dx, \ -\infty < t < \infty \tag{2}$$

.

1.3 The Continuous Uniform Distribution

A random variable X with the continuous uniform distribution on the interval (a, b) has PDF

$$f_X(x) = \frac{1}{b-a}, \ a < x < b \tag{3}$$

The associated R function is $dunif(min = a, max = b)$. We write $X \sim unif(min = a, max = b)$. Due to the particularly simple form of this PDF we can also write down explicitly a formula for the CDF F_X:

$$F_X(t) = \begin{cases} 0 & t < 0, \\ \frac{t-a}{b-a} & a \leq t < b, \\ 1 & t \geq b. \end{cases} \tag{4}$$

The continuous uniform distribution is the continuous analogue of the discrete uniform distribution; it is used to model experiments whose outcome is an interval of numbers that are "equally likely" in the sense that any two intervals of equal length in the support have the same probability associated with them.

Example 1 Choose a number in [0 1] at random, and let X be the number chosen. Then $X \sim unif(min = 0, max = 1)$. In this case, the mean of X is relatively simple to calculate:

$$\mu = EX = \int_a^b x\frac{1}{b-a}dx = \frac{b+a}{2} \tag{5}$$

1.3.1 R command for uniform distribution

Uniform density: **dunif(x, min=0, max=1)** Useful for graphing, not useful for directly finding probabilities.

In R, all PDF's have a "d" prefix for density

In [ ]:
```
curve ( dunif( x , min = 2 , max = 6) , 0 , 8 , ylim = c( 0 ,0.5 ),
       ylab = "f(x)" , main = "Uniform Density f(x)")
```

Probability is area!

Finding Probabilities

Area represents probability!

- $$P(x < a) = \int_{-\infty}^{a} f(x)dx \text{ (area to the left of a)} \tag{6}$$

- $$P(a < x < b) = \int_a^b f(x)dx \text{ (area between a and b)} \quad (7)$$
- $$P(x > a) = \int_a^{\infty} f(x)dx \text{ (area to the right of a)} \quad (8)$$

1.3.2 Cumulative distribution function (cdf) $F_X(x)$

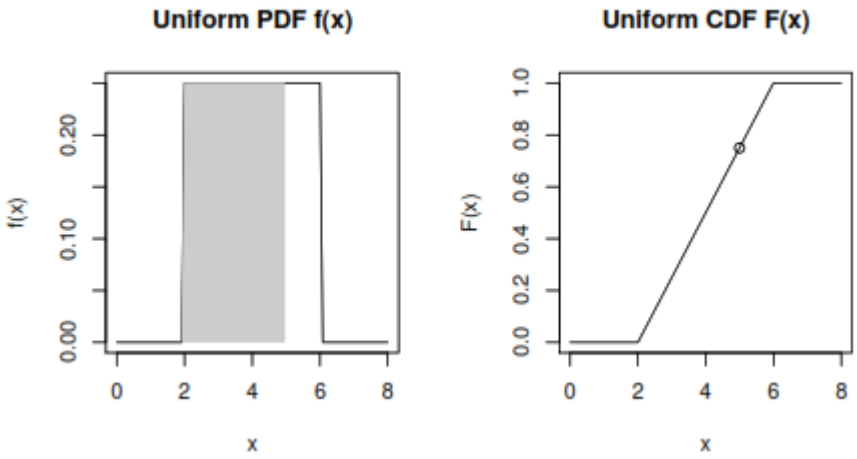
Gives the area to the left of x on the probability density function.

1.3.2.1 R command

Uniform CDF: punif(x, min=0, max=1) Gives the area to the left of the uniform density at x.

In R, all CDF's have a "p" prefix for probability.

Example 2 Find $P(X < 5)$, $P(X > 5$, and $P(3 < x < 5)$



In [ ]:
```
# P(x<5)
punif(5,min=2,max=6)
```

In [ ]:
```
# P(X>5) = 1 - P(X<5)
1-punif(5, min =2, max =6)
```

In [ ]:
```
# P(3<X<5) = P(X <5) - P(X<3)
punif(5,min=2,max=6)-punif(3,min=2,max=6)
```

1.4 INVERSE CUMULATIVE DISTRIBUTION FUNCTIONS

Definition 1 *Inverse cumulative distribution functions* CDF^{-1} *Finds the value of x that has an area left. (Inverse operation of CDF).*

R command UNIFORM INVERSE CDF:

qunif(p, min=0, max=1) # Finds x with area p to the left on the density function.

In R, all inverse CDFs have a "q" prefix for quantile.

Example 3 Find x such that $P(X < x) = 0.75$ (the value of x that has an area to the left of it equal 0.75) and $P(X > x)$ = 0.25 (the value of x that has an area to the right of 0.25).

In [ ]:
```
qunif(0.75, min=2,max=6)
```

In [ ]:
```
#P(X>x)=1-P(X<x)=0.25 implies P(X<x)=0.75
```

1.5 The Normal Distribution

Imagine that you buy a bag with 500 g of coffee. You are curious and empty the contents onto a weight to check whether the bag actually contains 500 g. If you have very precise weight, you will hardly expect that the content weighs exactly 500 g and you are probably not surprised if it is a little more or less. If you are repeating the experiment with many bags, you might expect that the weight of a bag will be very close to 500 g on average. You may also expect that there will not be too much spread. For instance, you do not expect to get less than 450 or more than 550 g, not even once. The weight of a bag can perhaps vary around 490–510 g, but it will rarely be more than 510 g or less than 490 g. This variation can be described by a statistical distribution. The most important statistical distribution is the normal distribution (*). Several statistical techniques require that data “follow” (i.e., can be described by) a normal distribution. If data do not follow a normal distribution, it becomes more difficult to analyze the data.

1.5.1 Characteristics of the Normal Distribution

The normal distribution curve is a symmetrical, "bell-shaped" curve similar to a histogram—in this case showing the weight of a very large number of coffee bags. It has been proven in practice that the normal distribution often gives a good description of many types of measurement data, such as weight, height, etc. But the normal distribution is very important also for economic and administrative data.

A normal distribution is completely described by its mean (average) and standard deviation.

The normal distribution in the example above describes the weight of all the coffee bags manufactured by the factory. Since we do not know the mean and standard deviation, they are often written in Greek letters:

- Mean: μ (read “mju”) representing the “center”
- Standard deviation: σ (read “sigma”) representing the “spread”

1.5.2 Normal Distribution Density function $f_X(x)$

$$f_X(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} \tag{9}$$

characterized by mu and σ.

Occurs frequently in nature.

R COMMAND NORMAL DENSITY:

```
dnorm(x,mean=0, sd=1)
   By default it is standard normal density.
```

In [ ]:
```
curve(dnorm(x,mean=0,sd=1),-5,5)
```

1.5.3 Visualizing the effect of μ and σ

In [ ]:
```
par(mfrow = c(2,2))
curve(dnorm(x,mean=1,sd=5),-30,30)
curve(dnorm(x,mean=1,sd=10),-30,30)
curve(dnorm(x,mean=2,sd=5),-30,30)
curve(dnorm(x,mean=2,sd=10),-30,30)
```

Area under curve is always 1.

The following Figure shows the interpretation of the standard deviation in a normal distribution. Here is shown a normal distribution representing the histogram of a population or a very large sample. We observe that:

- 68 % of the data values are in an interval around mean $\pm$ standard deviation
- 95 % of the data values are in an interval around mean $\pm$ 2 standard deviations
- 99.7 % of the data values are in an interval around mean $\pm$ 3 standard deviations

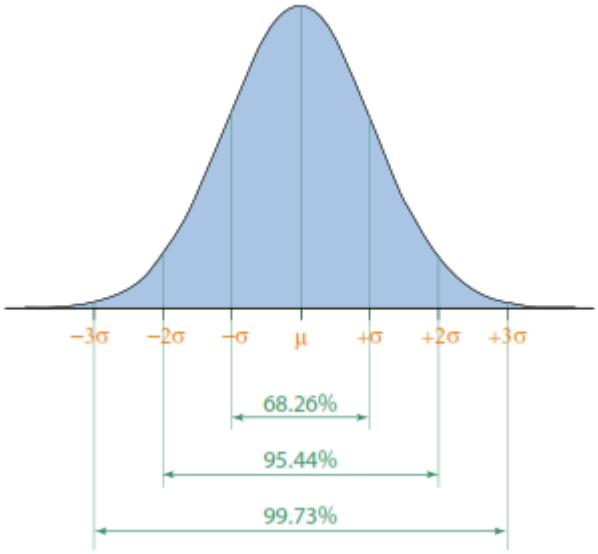
These percentages are unique to the normal distribution! In a way, the normal distribution is "thinking" as if 0 corresponds to the mean and 1 unit corresponds to the standard deviation.

Definition 2 *If X follows a normal distribution with mean μ and standard deviation σ, then*

$$\frac{X-\mu}{\sigma} \text{ z-score} \tag{10}$$

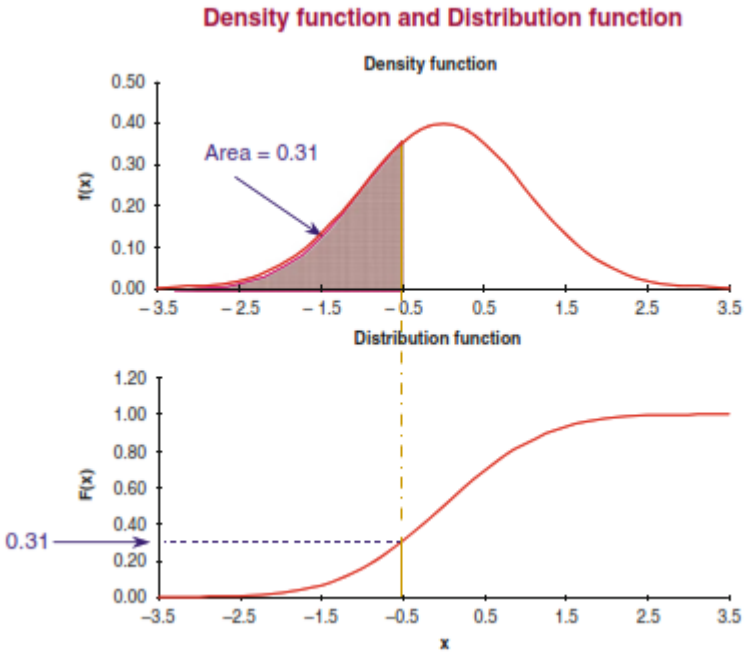
, follows a normal distribution with mean 0 and standard deviation 1.

Remark 1 *There exists in a way, only one normal distribution! The normal distribution with mean 0 a standard deviation 1 is therefore called the standardized normal distribution. All other normal distributions can be transformed to the standard normal distribution.*



1.6 Density Function and Distribution Function

In practice, it is therefore areas under normal distribution curve, which are interesting because they can be interpreted as probabilities. Therefore, we are usually interested in the curve showing areas under the normal distribution curve. The relationship between these two graphs is shown below. The bell-shaped curve in the following Figure is called the density function (*), while the curve showing areas (probabilities) is called the distribution function ()*. The distribution function is often written using the letter F. We can interpret the distribution function by noticing that F(x) is the probability of observing data values up to and including x.



In real-world problems, we almost always need the distribution function.

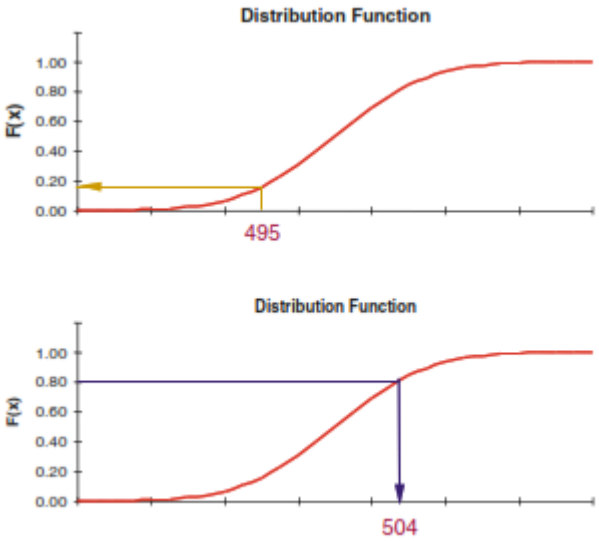
1.7 FINDING PROBABILITIES INVOLVING THE NORMAL DISTRIBUTION IN R

1.7.1 NORMAL CDF

```
pnorm(x,mean=0, sd=1)
  gives the area to the left of the normal density at x.
```

1.7.2 NORMAL INVERSE CDF:

```
qnorm(p, mean=0, sd = 1)
  finds x with area p to the left on the density function
```



We can use the normal distribution function in “both ways”.

When we are using the distribution function in the “reverse” way, for example from 0.80=80 % on the y-axis to a value on the x-axis, we are talking about finding a fractile (*) (also called a quantile or a percentile) in the distribution. The figure above shows the 80 % fractile in a normal distribution. We have seen that the most important fractiles: The quartiles are the 25 % and 75 % fractiles, and the median is the 50 % fractile. The fractiles corresponding to 10 %, 20 %, 30 %, etc., are called the deciles.

1.8 Calculations in the Normal Distribution

Example 4 Let us assume that the weight, X, of (the coffee in) a bag of coffee follows a normal distribution with a mean $\mu = 500g$ and standard deviation $\sigma = 5g$, then the z-score

$$\frac{X - \mu}{\sigma} \tag{11}$$

follows a normal distribution with mean 0 and standard deviation 1. Now let's answer the following questions:

What is the probability that a random coffee bag weighs at most 510 g?

In [ ]:
```
pnorm(510,mean=500,sd=5)
```

In [ ]:
```
x =510
m = 500
s = 5
z = (x - m)/s
pnorm(z,mean=0,sd=1)
```

2) What is the 95 % fractile in the distribution? Assuming $P(X \leq x) = 0.95$ what is the value of x?

In [ ]:
```
qnorm(.95,mean=500,sd=5)
```

This means that the 95 % fractile in the distribution of a randomly chosen coffee bag is 508.25. In other words, the probability that a random coffee bag weighs less than 508.25 g is exactly 95 %.

1.8.1 Tips for solving probabilities involving normal dist.

1. Determine μ and σ.
2. Sketch the PDF & area representing probability.
3. If asked to find probability use CDF: R function pnorm(x,...) to find probability p.

4. If asked to find value of x corresponding to probability use $CDF^{-1} : R function qnorm(p,...) to find the value of x.
5. If working with upper tail be sure to take compliment! Be careful, if you want to find the value of x that has an area p to the right you need to use qnorm(1-p, ...).

Example 5 Given $\mu = 2$ and $\sigma = 1$, find $P(X > 3)$

$P(X > 3) = 1 - P(X < 3)$

\begin{verbatim} p = 1 - pnorm(3, mean = 2, sd = 1) p 0.15866

\end{verbatim}

Thus, $P(X > 3) = 0.159$

Now, let's look at the inverse problem: Given $\mu = 2$ and $\sigma = 1$, what value of x satisfies $P(X > x) = 0.159$? $x = F_X^{-1}(1 - 0.159)$

\begin{verbatim}

p 0.15866 x=qnorm(1-p, mean =2, sd = 1) x 3

\end{verbatim}

1.9 Summary

- For discrete random variables, probability is given by the distribution $p = P(X = x_i)$. ("d" prefix in R)
- For the binomial distribution, the probability of a specific number of successes x is $\mathbf{dbinom(x, n, p)}$.
- For continuous variables, probability is **area** on density $f(x)$.
- Use CDF's $F_X(x)$ to find probabilities. ("p" prefix in R)
  - $P(X < x') = F_X(x')$ (area to the left of x')
  - $P(X > x') = 1 - F_X(x')$ (area to the right of x')
  - $P(a < X < b) = F_X(b) - F_X(a)$ (area between a and b)
- Use inverse CDF's $F_X^{-1}(p)$ to find specific value of x' in $p = P(X < x')$ given probability p. ("q" prefix in R) i.e $x' = F_X^{-1}(p)$

For the normal distribution,

- Use CDF to compute the probabilities:
  - $p = F_X(x')$: $\mathbf{p = pnorm(x', mean = 0, sd = 1)}$
  - $P(X < x') = \mathbf{pnorm(x', \ldots)}$

- $P(X > x') = 1 - \mathbf{pnorm(x', \ldots)}$
- $P(a < X < b) = \mathbf{pnorm(b, \ldots) - pnorm(a, \ldots)}$

  where " $\cdots$ " is " $\mathbf{mean} = \mu, \mathbf{sd} = \sigma$ ".

- use the inverse CDFs $x' = F_X^{-1}(p)$ to find x given probability p. ("q" prefix in R)

\begin{itemize}
- To find x' in $p = P(X < x')$: $\mathbf{x' = qnorm(p, \ldots)}$
- To find p = P(X>x'): \bf x'= qnorm(1-p,\dots)$

\end{itemize}

For continuous variables in general:

- Carefully determine location of area: to left, to right, interval.
- Always make a sketch when doing problems.
- CDF assumes areas to the left. Take the compliment when finding upper tail!

1.10 Testing for the Normal Distributions

We have studied some key characteristics of the normal distribution, its density function and distribution function, and calculations in the normal distribution. In other words, the assumption has been that the data actually are following a normal distribution. There are several ways to check this. This is the topic of this section.

1) The histogram It is always a good idea to study the histogram. This must show a symmetrical, “bell-shaped” appearance.

2) The average = the median. If data can be described by a normal distribution, the average and median must be nearly identical, because the normal distribution is symmetrical. This is very simple to check.

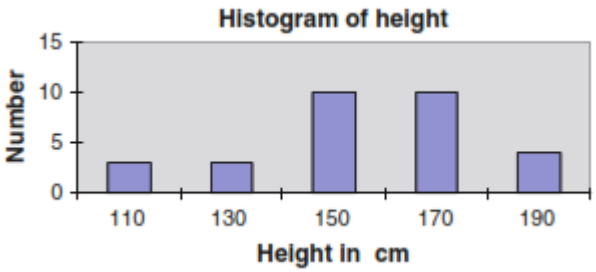
3) Interquartile range larger than the standard deviation. In the normal distribution, the interquartile range (i.e., the difference between the upper and lower quartile) is somewhat larger than the standard deviation; actually, it is around 1.35*the standard deviation, i.e.,

$$IQR = 1 : 35 \times s \tag{12}$$

. By comparison, the standard deviation is precisely 1 in the standardized normal distribution.}

4) Number of data values in symmetric intervals around the mean. We have seen that around 68 % of the data values in a normal distribution are in an interval around mean $\pm$ standard deviation. If the data can be described by a normal distribution, the corresponding proportion for the data values therefore must be relatively close to 68 %. If we have many data values (at least a couple of hundred), one can calculate the proportion of data values between the mean $\pm$ 2 standard deviations. This proportion should be relatively close to 95%.}

Example 6 A histogram of the height of all 30 kids from the Fitness Club survey is shown below. This histogram seems roughly symmetrical. When the sample is small, we must accept some deviation from the ideal appearance.



The most important statistics for the height of the 30 kids are given in the following table:

Height	
Mean	157.1
Median	159.5
Standard deviation	22.1
$Q1$	146.5
$Q3$	170.0
Interquartile range	23.5

We see that the average and the median are roughly equal.

The interquartile range is slightly larger than the standard deviation, though not by a factor of 1.35.

The interval mean pm standard deviation corresponds to the interval from 135.0 to 179.2.

In this interval, you can count 21 of the 30 data values, equivalent to 70 %, i.e., very close to 68 %.

Overall, we conclude that it appears reasonable that data can be described by a normal distribution.

1.11 Skewness and Kurtosis

The two statistics skewness and kurtosis can be used to check whether data follow a normal distribution. They are, however, complicated to calculate and therefore require the use of a spreadsheet or other statistical software. They provide an easy opportunity to check whether the data can be described by a normal distribution.

Skewness (*) is a measure of how skewed the distribution is compared to a symmetrical distribution:

- If data can be described by a symmetrical distribution, the skewness must be close to 0.
- Positive skewness indicates a right-skewed distribution.
- Negative skewness indicates a left-skewed distribution.

As a very rough guide as to how large deviations from 0 can be accepted for the skewness for different sample sizes n, you can use the expression:

\begin{equation}2\sqrt{\frac{6}{n}\end{equation}

where n is the sample size.

The smaller sample, the greater deviations from 0 you have to accept. When the sample size is multiplied by 4, the maximum acceptable deviation from 0 is divided by 2. So, you check whether the skewness is within the maximum acceptable deviation from 0. If not, we do not have a symmetrical distribution.

If the distribution is symmetrical, you can supplement with evaluating another statistic:

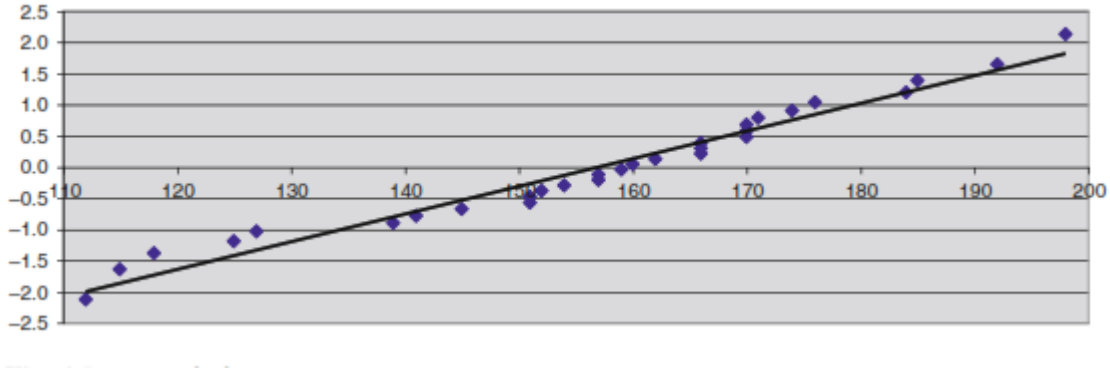
Kurtosis (*) indicates how big “tails” the distribution has:

- A normal distribution has kurtosis 0.
- A positive kurtosis indicates larger “tails” than in the normal distribution.
- A negative kurtosis indicates smaller “tails” than in the normal distribution.

A distribution with positive kurtosis is often more “steep” in the top than the normal distribution. Conversely, a distribution with negative kurtosis is often more “flat” in the top than the normal distribution.

1.12 Normal Plot

Finally, there exists a simple graphical tool to check if your data follow a normal distribution. This is called a normal plot (probability plot or quantile plot). It is a built-in feature of many statistical software packages, e.g., SAS, JMP, SPSS, Minitab, R. If you have a statistical software package, you can produce a plot like the one shown the figure below.



It is also quite easy to do in a spreadsheet. If data follow a normal distribution, the points should be randomly scattered around the straight line. This seems to be the case here. This confirms that data can be described by a normal distribution as in the above figure.

In [ ]:
```
dat = read.csv("fitness_club.csv", header = TRUE)
names(dat)
```

In [ ]:
```
hist(dat$Height)
```

In [ ]:
```
library(e1071)
skewness(dat$Height)                # apply the skewness function
"
Intuitively, the skewness is a measure of symmetry.
As a rule, negative skewness indicates that the mean of the data values
is less than the median, and the data distribution is left-skewed.
Positive skewness would indicate that the mean of the data values is large
and the data distribution is right-skewed.
"

kurtosis(dat$Height)                # apply the kurtosis function

"
Intuitively, the kurtosis describes the tail shape of the data distributio
The normal distribution has zero kurtosis and thus the standard tail shape
It is said to be mesokurtic. Negative kurtosis would indicate a thin-taile
data distribution, and is said to be platykurtic. Positive kurtosis
would indicate a fat-tailed distribution, and is said to be leptokurtic.
"
```

In [ ]:
```
# summary statistics for Height
summary(dat$Height)
```

In [ ]:
```
#install.packages("pastecs")
library(pastecs)
head(dat)
```

In [ ]:
```

options(digitσ=2)
stat.desc(dat)
```

In [ ]:
```
stat.desc(dat,basic=F)
```

1.12.1 Understanding Q-Q Plots

The Q-Q plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential. For example, if we run a statistical analysis that assumes our dependent variable is Normally distributed, we can use a Normal Q-Q plot to check that assumption. It's just a visual check, not an air-tight proof, so it is somewhat subjective. But it allows us to see at-a-glance if our assumption is plausible, and if not, how the assumption is violated and what data points contribute to the violation.

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. Here's an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.

In [ ]:
```
# creating a normal probability plot in R
qqnorm(dat$Height, pch=1, frame = TRUE)
qqline(dat$Height,col=2)
```

1.13 The Normal Distribution and Spreadsheets

There are in Microsoft Excel and OpenOffice Calc two important functions for the normal distribution.

- NORMDIST: Provides the distribution function or density function for a normal distribution.
- NORMINV: Gives fractiles in a normal distribution.

1.13.1 NORMDIST (X; Mean; Stdev; Cumulative)

- X: The number of which the value of the distribution function (or density function) is desired
- Mean: The mean of a normal distribution
- Stdev: The standard deviation of a normal distribution
- Cumulative: Cumulative = 0 calculates the density function

```
Cumulative = 1 calculates the distribution function
```

It is only if you have to make figures with the "bell-shaped" curve, that you need the density function. Therefore, you should almost always use Cumulative = 1. The distribution function for the standardized normal distribution (with mean 0 and standard deviation 1) can also be obtained by using the function NORMSDIST. This function only has one parameter: X.

1.13.2 NORMINV (Probability; Mean; Stdev)

- Probability: The probability for which a fractile in the normal distribution is wanted
- Mean: The mean of a normal distribution
- Stdev: The standard deviation of a normal distribution

The NORMINV function is used to find a fractile in the normal distribution with a given mean and standard deviation. NORMINV thus has 3 parameters. For the standardized normal distribution (with mean 0 and standard deviation 1), you can use the function NORMSINV. This function only

(with mean 0 and standard deviation 1), you can use the function NORMSINV. This function only has one parameter: probability.

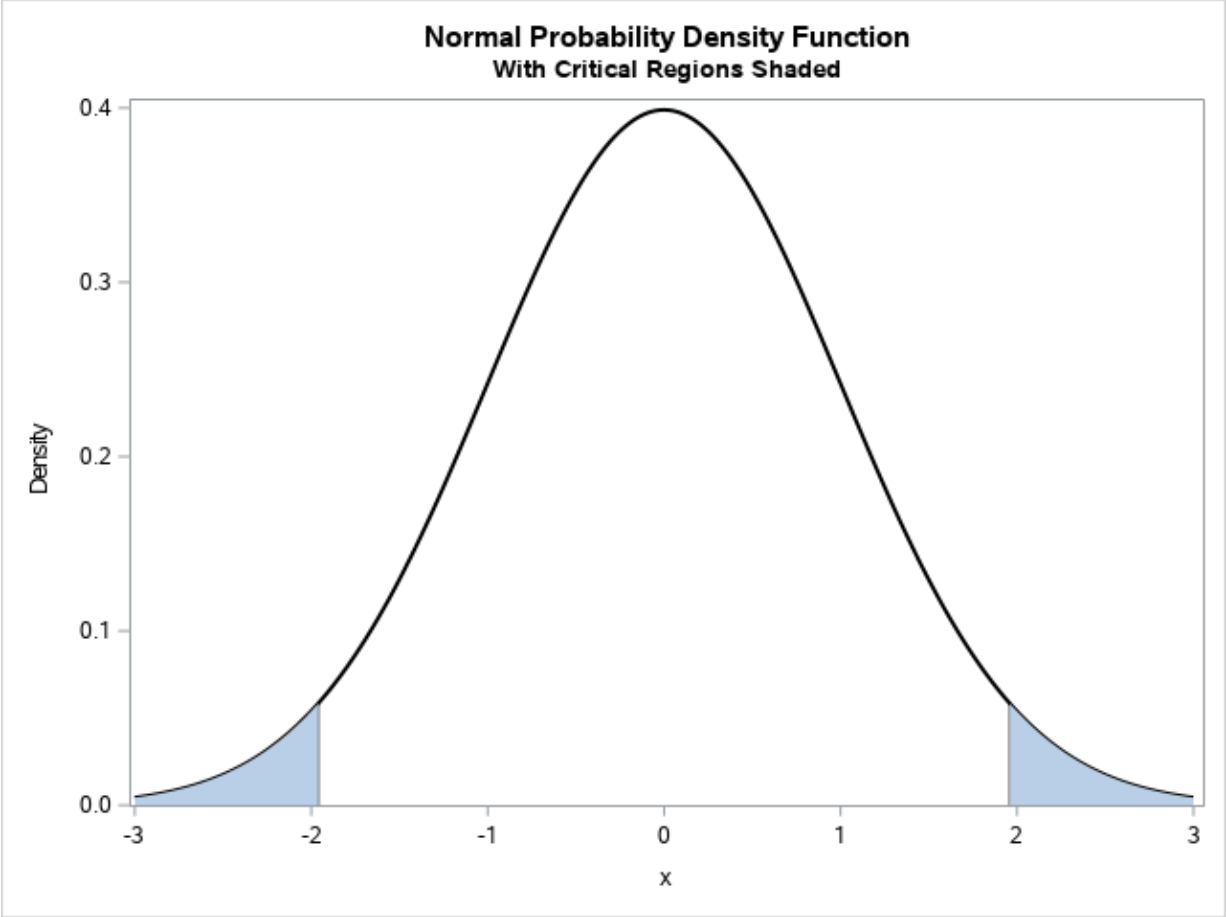
Example 7 Let us assume that the weight, X, of (the coffee in) a bag of coffee follows a normal distribution with a mean $\mu = 500g$ and standard deviation $\sigma = 5g$.

1. What is the probability that a random coffee bag weighs at most 490 g? We use NORMDIST(490; 500; 5; 1) and get the result 0.023=2.3 %.
2. What is the probability that a random coffee bag weighs at most 510 g? Similarly, we find the probability that a random coffee bag weighs at most 510 g: We use NORMDIST(510; 500; 5; 1) and get the result 0.977 = 97.7 %.
3. What is the 5 % fractile in the distribution? Remember that 5 % equals 0.05. In the spreadsheet, we do not use percentages when writing probabilities. We therefore use NORMINV(0.05; 500; 5) and get the result 491.8. This means that the probability that the weight of a random bag of coffee $\leq 491.8g$ is precisely 0.05, which is equivalent to 5 %. In other words, there is a 95 % chance that a random coffee bag weighs more than 491.8 g.
4. What is the 80 % fractile in the distribution? Similarly, we find the 80 % fractile as NORMINV(0.80; 500; 5) and get the result 504.2. This means that the probability that the weight of a random bag of coffee is 504.2 g is precisely 0.80, which is equivalent to 80 %. In other words, there is a 20 % chance that a random coffee bag weighs more than 504.2 g.

1.14 Normal Distribution in SaS

It is important to have a basic understanding of the normal distribution, and how the shape changes with its parameters. Below, there is SAS code example for you to play around with. Insert it into your SAS editor and change the three values defined at the top of the code to see how it affects the shape of the distribution.

```
%let alpha = 0.05; /* Set alpha value */

data normal_PDF(drop = lower_q upper_q);
   lower_q = quantile('normal', &alpha/2 , &mu, &sigma);
 /* Set lower quantile         */
   upper_q = quantile('normal', (1 - &alpha/2), &mu, &sigma);
/* Set upper quantile         */

   do x=&mu - 3*&sigma to &mu + 3*&sigma by 0.01;
      density = pdf('normal',x,&mu,&sigma);
/* Normal Density Function    */
      output;
   end;
   x = .; density = .;

   x_line = upper_q; line = pdf('normal',x_line,&mu,&sigma);output;
/* Line for upper quantile    */
   x_line = lower_q; line = pdf('normal',x_line,&mu,&sigma);output;
/* Line for lower quantile    */
   x_line = .; line = .;

   do lower_x_band = &mu - 3*&sigma to lower_q by 0.01;
      lower_band = pdf('normal',lower_x_band,&mu,&sigma);
/* Lower critical region      */
      output;
   end;
   lower_x_band = .; lower_band = .;

   do upper_x_band = upper_q to &mu + 3*&sigma by 0.01;
      upper_band = pdf('normal',upper_x_band,&mu,&sigma);
/* Lower critical region      */
      output;
   end;
   upper_x_band = .; upper_band = .;
run;

title 'Normal Probability Density Function';
title2 'With Critical Regions Shaded';
proc sgplot data = normal_PDF noautolegend;
   series x = x y = density    / lineattrs = (color = black thickness =
 2);
   dropline x = x_line y = line / lineattrs = (color = black);
   band x = lower_x_band upper = lower_band lower = 0;
   band x = upper_x_band upper = upper_band lower = 0;
   yaxis offsetmin=0 min=0 label="Density";
   xaxis label = 'x';
run;
title;
```



1.15 Four essential distributions functions for statistics with sas

Normal, Poisson, exponential—these and other "named" distributions are used daily by statisticians for modeling and analysis. There are four operations that are used often when you work with statistical distributions. In SAS software, the operations are available by using the following four functions, which are essential for every statistical programmer to know:

PDF function: This function is the probability density function. It returns the probability density at a given point for a variety of distributions. (For discrete distribution, the PDF function evaluates the probability mass function.) CDF function: This function is the cumulative distribution function. The CDF returns the probability that an observation from the specified distribution is less than or equal to a particular value. For continuous distributions, this is the area under the PDF up to a certain point. QUANTILE function: This function is closely related to the CDF function, but solves an inverse problem. Given a probability, P, it returns the smallest value, q, for which CDF(q) is greater than or equal to P. RAND function: This function generates a random sample from a distribution. In SAS/IML software, use the RANDGEN subroutine, which fills up an entire matrix at once.

Study the material in the blog

https://blogs.sas.com/content/iml/2011/10/19/four-essential-functions-for-statistical-programmers.html (https://blogs.sas.com/content/iml/2011/10/19/four-essential-functions-for-statistical-programmers.html)

1.16 Random Numbers

In evaluating how well a data set is consistent with the normal distribution, it can be a good benchmark to do the same calculations and charts for a similar number of random numbers from a normal distribution. In Microsoft Excel, you can construct random numbers from a normal distribution using the add-in menu Data Analysis, which has a sub-item Random number generation. A similar option does not exist in Open Office Calc. The same is true in R and SaS. Following are some examples in R.

In [ ]:
```
# Example : Uniform distribution
runif(1)    # generates 1 random number
runif(3)    # generates 3 random number
runif(3, min=5, max=10)    # define the range between 5 and 10
```

In [ ]:
```
# Example: Normal distribution
rnorm(1)    # generates 1 random number
rnorm(3)    # generates 3 random number
rnorm(3, mean=10, sd=2)    # provide our own mean and standard deviation
```

1.17 Discrete Distributions

Finite populations We can generate samples from a given set of objects in R using the sample function.**

```
sample(1:100)
    # a random permutation of the numbers 1 to 100

sample(x,10,replace=F)
    # a random sample of size N=10
    # drawn from the elements of x without replacement

sample(1:6,10,replace=T,prob=c(1/7,1/7,1/7,1/7,1/7,2/7))
    # a random sample of size 10 from the digits 1 to 6
    # with unequal probabilities of selection
    # eg. a loaded die
```

Bernoulli(p) Trials Bernoulli trials are sequences of independent dichotomous trials, each with probability p of success. The sample space consists of the two possible outcomes. For example, the results of 10 successive (independent) tosses of a fair coin would constitute 10 Bernoulli(1/2) trials. In R: rbinom(N,1,p) generates a sequence of N trials, each with probability p of success. Actually, R returns a sequence of 1's and 0's, but we can identify 1's with successes and 0's with failures if we like. Try generating a sample of 50 bernoulli trials, with p = .1. What is the average number of 0's between each 1? What do you think it would be after a really large number of trials?

Binomial(n,p) This distribution arises from a sequence of n independent Bernoulli trials, each with probability p of success. Since the number of successes in n trials can range from 0 to n, the sample space is just the integers 0,1,2, ... n. For example, the number of heads that occur in 20 independent tosses of a fair coin has a Binomial(20,.5) distribution. In R: rbinom(N,n,p) generates N independent samples from a binomial(n,p) distribution. Generate 100 samples of binomial(20,.5) random values and make a normal plot. What do you see?

In [ ]:
```
data <- rbinom(100,20,.5)
qqnorm(data)
```

1.18 Calculations with t- distribution

Case I: Calculating $P(t < x)$ If $X \sim t(df)$, where df is the degrees of freedom, use the $pt(x, df)$ r function to calculate $P(X < x)$.

Example 8 calculate $P(X < 1.2)$ for 4 dfs

In [ ]:
```
pt(1.2,4)
```

Case II. Calculating $P(X > x)$ If $X \sim t(df)$, where df is the degrees of freedom, use the $pt(x, df, lower.tail = FALSE) function to calculate$P(X > x)$ or subtract pt(x, df) from 1.

In [ ]:
```
pt(1.2, 4, lower.tail = FALSE)
```

In [ ]:
```
1 - pt(1.2, 4)
```

case III. Given percentile, find corresponding t-value, if $X \sim t(df)$, use the qt(percentile, df) function to find the x-value that corresponds with a given percentile.

Example 9 If $X \sim t(12)$, what x-value corresponds with the 75th percentile?

In [ ]:
```
qt(0.75, 12)
```

Case IV: For confidence interval calculations, under certain conditions, one may need to find t*.

Example 10 For confidence level 95% find the critical value t* with 8 dfs

In [ ]:
```
#IV. Determing t*

qt(0.975,12) # finds t that correspondrs for the two tail t distribution t
pt(2.17881282966723,12)
```

In [ ]:
```
# plots the density of the t -distribution
curve(dt(x,12,0,1))
```

1.18.1 R: Normal Distribution

1. Calculating P(X < x) If $X \sim N(\mu, \sigma)$, use the $pnorm(x, mu, sigma)$ function to calculate $P(X < x)$. Example 1.a: If $X \sim N(85, 5)$, use the following R code to calculate P(X < 81).

In [ ]:
```
pnorm(81, 85, 5)
```

Example 1.b: If X~N(85, 5), use the following R code to calculate P(X < 81). This method involves first calculating z, then using the pnorm function to find the area to the left of z using the standard normal distribution.

In [ ]:
```
# Calculate z
z = (81 - 85)/5
z
# define standard normal
pnorm(z, 0, 1)
# if the mean and sd not provided,  R assumes mean =0, sd = 1
 pnorm(z)
```

II. Calculating P(X > x)

If $X \sim N(\mu, \sigma)$, use the pnorm(x, mu, sigma, lower.tail = FALSE) function to calculate P(X > x). A second method would be to subtract pnorm(x, mu, sigma) from 1. Example 1:

In [ ]:
```
pnorm(81, 85, 5, lower.tail =FALSE)
1 - pnorm(81, 85, 5)
```

III. Given percentile, find corresponding x-value

If $X \sim N(\mu, \sigma)$, use the qnorm(percentile, mu, sigma) function to find the x-value that corresponds with a given percentile. Example 1: If $X \sim N(30, 3)$, what x-value corresponds with the 75th percentile?

In [ ]:
```
qnorm(0.75, 30, 3)
```

Example 2: Jetblaster is a popular game app. Scores on the game are normally distributed with a mean of 1,114 and a stanard deviation of 321. Jack wishes to qualify for the national tournament. Only those who have a score in the top 10% qualify. What is the minimum score Jack needs to qualify for the national tournament? (Hint: The 90th percentile is the cutoff to score in the top 10 percent.)

In [ ]:
```
qnorm(0.9, 1114, 321)
```

IV. Simulating Normal Random Variables

In statistics, one often finds the need to simulate random scenarios that are normally distributed. To do this, we need to use the rnorm(n, u, c) function, where n represents the number of random observations you wish to observe. Example 1: Scores on the Jetblaster app game are normally distributed with a mean of 1,114 and a stanard deviation of 321. Simulate 8 random scores of this game.

In [ ]:
```
rnorm(8, 1114, 321)
```

IV. Assessing Normality

Often one must try to assess if data is actually normally distributed. There are a few methods available to do this. One very popular method is to construct a normal probability plot. Example 1: Sample 1000 observations from a normally distributed random variable that has a mean of 30 and a standard deviation of 3. Then construct a histogram and normal probability plot to assess normality.

In [ ]:
```
# Randomly sample 1000 observations from N(30, 3)
data =
rnorm(1000, 30, 3)
# Construct histogram
hist(data)
# Construct Normal Probability Plot
curve(dnorm(x,0,1))
qqnorm( data, main =
"Normal Probabili ty Plot")
# Add diagonal line to help with normality assessment
qqline( data)
```

Theoretlcal Ouantiles

Conclusion: As one would expect given that the data was drawn from the normal distribution, the histogram clearly looks tobe normally distributed. Upon inspection of the Normal Probability Plot, one can see that the points line up with the line. When the points line up weil with the horizontal line, one can assume the data is normally distributed, as is the case shown above.

1.18.2 Binomial Distribution

1. Calculating Exact Binomial Probabilities

If $X \sim B(n, p)$, one may use the following mathematical formula to calculate $P(X = k)$.

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \text{ where } \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Likewise, if $X \sim B(n, p)$, one may use the **dbinom(x, n, p)** function to calculate P(X = k). In other words, the **dbinom(x, n, p)** function will provide the same answer as using the mathematical formula shown above.

In [ ]:
```
# Example 1:
#If X~B(lO, 0.3), use the following R code to calculate P(X = 3).
# Method 1 - Use dbinom(k, n, p)
dbinom(3, 10,0.3)
```

In [ ]:
```
# Method 2 - Use the mathematica 1 formula . .... a  li ttle more compl ic
choose(10,3)*(0.3)^3*(0.7)^7
```

In [ ]:
```
#Example 2:
#On the way to work, a commuter encounters one stoplight. The probability
#course of 20 days of commuting, find the probability that the commuter en
dbinom(8, 20, 0.3)
```

In [ ]:
```
#II. Calculating Cumulative Binomial Probabilities
#If X~B(n,p), use pbinom(x, n, p) function to calculate P(X s x).
#Example 1:
#If X~B(10, 0.3), use the following to code to find P(X s 3).
pbinom(3,10,0.3) # pbinom(x, n, p)
```

In [ ]:
```
#Example 2: If X~B(lO, 0.3), use the following to code to find P(X > 3).
# Method 1 - Use "lower.tail =FALSE"
pbinom(3, 10, 0.3, lower.tail = FALSE)
```

In [ ]:
```
# Method 2 - Subtract pbinom(x, n, p) from 1
 1 - pbinom(3, 10, 0.3)
```

In [ ]:
```
#III. Simulating Binomial Random Variables
#In statistics, one often finds the need to simulate random scenarios that
#rbinom() function, you need to define three parameters:
#Example 1:
#Let's say you wanted to simulate rolling a dice 5 times, and you wished t
#using the following code:
rbinom(1,2,1/6)
#rbinom(number of experiments, number of observations per experiment, prob
#Conclusion: The above code simulated rolling a die five times. The output
#labeled as a "success".
```

In [ ]:
```
#Example 2:
#Let's say you wanted to simulate a dass of 30 different students flipped
#they observed. You could simulate this experiment using the following cod
rbinom(30,10,0.5) # rbinom(number of experiments, number of observations p
#Conclusion: The above output shows the number of "successes" recorded by
#heads, while the second student recorded seeing 4 heads.
```

1.19 R: Chi-Square Distribution

If $Q \sim X^2(df)$, use the **pchisq(q, df, lower.tail = FALSE)** function to calculate $P(Q > q)$.

In [ ]:
```
#Example 1:
#If Q~X^2(df = 7), use the following R code to calculate P(Q > 13).
 pchisq(13, df = 7, lower.tail = FALSE)
```

In [ ]:
```
#Example 2:
#If Q ~ X^2(df =12), use the following R code to calculate P(Q > 9).
pchisq(9, df =12, lower.tail =FALSE)
```

1.20 II. Given percentile, find corresponding q-value

If $Q \sim X^2(df = 12)$, use the **qchisq(percentile, df)** function to find the q-value that corresponds with a given percentile.

In [ ]:
```
#Example:
#If Q~X^2(df =12), what q-value corresponds with the 75th percentile?
qchisq(.75, df = 12)
```

1.21 III. Simulating Chi-Square Random Variables

In statistics, one may finds the need to simulate random scenarios that have a Chi-Square distribution. To do this, we need to use the **rchisq(n, df)** function, where n represents the number of random observations you wish to observe.

In [ ]:
```
#Example:
#Simulate 16 random variables drawn from the chi-square distribution with
rchisq(16, 7)
```

2 Quiz 3 Exercises

Exercise 1 Calculate the variance of the uniform distribution.

2.1 Central Limit Theorem

Suppose $X_1, X_2, \ldots, X_n, \ldots$ are i.i.d. random variables each having mean μ and standard deviation σ. For each n let S_n denote the sum and let $\bar{X}_n$ be the average of $X_1, \ldots, X_n$

$$\begin{aligned} S_n &= X_1 + X_2 + \ldots + X_n = \sum_{i=1}^{n} X_i \\ \bar{X}_n &= \frac{X_1 + X_2 + \ldots + X_n}{n} = \frac{S_n}{n} \end{aligned} \tag{13}$$

The properties of mean and variance show

$$E(S_n) = n\mu, \mathrm{Var}(S_n) = n\sigma^2, \sigma_{S_n} = \sqrt{n}\sigma \tag{14}$$

$$E\left(\bar{X}_n\right) = \mu, \mathrm{Var}\left(\bar{X}_n\right) = \frac{\sigma^2}{n}, \sigma_{\bar{X}_n} = \frac{\sigma}{\sqrt{n}} \tag{15}$$

since they are multiples of each other, S_n and $\bar{X}_n$ have the same standardization

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \tag{16}$$

Central Limit Theorem: For large n

$$\bar{X}_n \approx \mathrm{N}\left(\mu, \sigma^2/n\right), S_n \approx \mathrm{N}\left(n\mu, n\sigma^2\right), Z_n \approx \mathrm{N}(0,1) \tag{17}$$

Notes:

1. In words: $\bar{X}_n$ is approximately a normal distribution with the same mean as X but a smaller variance.
2. S_n is approximately normal.
3. Standardized $\bar{X}_n$ and S_n are approximately standard normal.

The central limit theorem allows us to approximate a sum or average of i.i.d random variables by a normal random variable. This is extremely useful because it is usually easy to do computations with the normal distribution. A precise statement of the CLT is that the codf's of Z_n converge to $\Phi(z)$:

$$\lim_{n\to\infty} F_{Z_n}(z) = \Phi(z) \tag{18}$$

Exercise 2 Suppose $X_1, \ldots, X_{100}$ are i.i.d. with mean $1/5$ and variance $1/9$. Use the central limit theorem to estimate $P(\sum X < 30)$.

In [ ]:

Exercise 3 The average IQ in a population is 100 with standard deviation 15 (by definition, IQ is normalized so this is the case). What is the probability that a randomly selected group of 100 people has an average IQ above 115?

Type *Markdown* and LaTeX: α^2

In [ ]:

Exercise 4 The execution times of queries on a database is normally distributed with a mean of 5 seconds and a standard deviation of 1 second. Determine the following:

a. What is the probability of the execution time being more than 8 seconds?

b. What is the probability of the execution time being less than 6 seconds?

c. What percentage of responses will take between 4 and 7 seconds?

d. What is the 95-percentile execution time?

In [ ]:

2.2 Exponential Distribution

The exponential distribution describes the arrival time of a randomly recurring independent event sequence. If μ is the mean waiting time for the next event recurrence, its probability density function is:

$$f(x) = \begin{cases} \frac{1}{\mu} e^{-x/\mu} & \text{when } x \geq 0 \\ 0 & \text{when } x < 0 \end{cases}$$

Exercise 5 (10 pts) Suppose the mean checkout time of a supermarket cashier is three minutes. Find the probability of a customer checkout being completed by the cashier in less than two minutes.

In [ ]:

Exercise 6 (10 pts) Assume that the test scores of a college entrance exam fits a normal distribution. Furthermore, the mean test score is 72, and the standard deviation is 15.2. What is the percentage of students scoring 84 or more in the exam?

In [ ]:

Exercise 7 (10 pts) Suppose X is normal with mean 8.0 and standard deviation 5.0 (a) Transform this to standard normal. What is the z-value? (b) Find $\mathrm{P}(\mathrm{X} < 8.6)$ using the standard normal

Type *Markdown* and LaTeX: α^2

In [ ]:

Exercise 8 (10 = 4 3 3 pts)

Suppose X is normal with mean 8.0 and standard deviation 5.0 .

(a) Find the Z value for the known probability.

(b) Find X from the Z value using the formula $X = m + Z\sigma$

(c) Now find the X value so that only 20% of all values are below this X using the normal distribution

In [ ]:

2.3 Normal Distribution Approximation for Binomial Distribution

- The shape of the binomial distribution is approximately normal if n is large
- The normal is a good approximation to the binomial when $\mathrm{np}(1-\mathrm{p}) > 9$
- Standardize to Z from a binomial distribution:

$$Z = \frac{X - E(X)}{\sqrt{\mathrm{Var}(X)}} = \frac{X - np}{\sqrt{np(1-p)}} \tag{19}$$

Let X be the number of successes from n independent trials, each with probability of success p. If $\mathrm{np}(1-\mathrm{p}) > 9$

$$P(a < X < b) = P\left(\frac{a - np}{\sqrt{np(1-p)}} \leq Z \leq \frac{b - np}{\sqrt{np(1-p)}}\right) \tag{20}$$

Exercise 9 40% of all voters support ballot proposition A. What is the probability that between 76 and 80 voters indicate support in a sample of n = 200 ?

In [ ]:

Exercise 10 (10 pts)

Let $\mathrm{X}_1, \mathrm{X}_2, \ldots \mathrm{X}_\mathrm{k}$ be k random variables with means $\mu_1, \mu_2, \ldots \mu_k$ and variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2$.

(a) Compute $\mathrm{E}\left(\mathrm{X}_1 + \mathrm{X}_2 + \cdots + \mathrm{X}_k\right)$

(b) Compute $\mathrm{Var}(X_1 + X_2 + \cdots + X_k) = \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_k^2$ assuming covariance is 0 and different than zero.

Exercise 11

(30 pts) Consider $X_1, X_2, \ldots$ all independent and with distribution $\mathrm{N}(0, 1)$. Let $\bar{X}_m$ be the average of $X_1, \ldots X_n$

(a) Give $E\left(\bar{X}_n\right)$ and $\sigma_{\bar{X}_n}$ exactly.

(b) Use a R simulation to estimate $E\left(\bar{X}_n\right)$ and $\mathrm{Var}\left(\bar{X}_n\right)$ for $n = 1, 9, 100.$ (You should use the rnorm function to simulate 1000 samples of each X_j.)

Type *Markdown* and LaTeX: α^2

In [ ]: