MATH 3050 – Predictive Analytics

UNC CHARLOTTE

Topic 2: Mathematical Distributions

- ❑ Functions
- ❑ Discrete Distributions
- ❑ Continuous Functions

UNC CHARLOTTE 1

1

Mathematical Distributions – Chapter 7

You can do a lot of math in R. Here we concentrate on the kinds of mathematics that is found most frequently in applications of scientific work and statistical modelling. We will only concentrate on the following topics in the chapter:

- Functions
- Discrete Distributions
- Continuous Distributions

We will study sections 7.1, 7.2, 7.3, 7.4 in the R Book

UNC CHARLOTTE 2

2

Mathematical Functions

These are the most important rules:

- Anything to the power zero is 1: $x^0 = 1.$
- One raised to any power is still 1: $1^x = 1.$
- Infinity plus 1 is infinity: $\infty + 1 = \infty.$
- One over infinity (the reciprocal of infinity, ∞^{-1}) is zero: $\frac{1}{\infty} = 0.$
- A number > 1 raised to the power infinity is infinity: $1.2^{\infty} = \infty.$
- A fraction (e.g. 0.99) raised to the power infinity is zero: $0.99^{\infty} = 0.$
- Negative powers are reciprocals: $x^{-b} = \frac{1}{x^b}.$
- Fractional powers are roots: $x^{1/3} = \sqrt[3]{x}.$
- The base of natural logarithms, e, is 2.718 28, so $e^{\infty} = \infty.$
- Last, but perhaps most usefully: $e^{-\infty} = \frac{1}{e^{\infty}} = \frac{1}{\infty} = 0$

Mathematical Functions

Logarithmic function $y = a\ln(bx)$

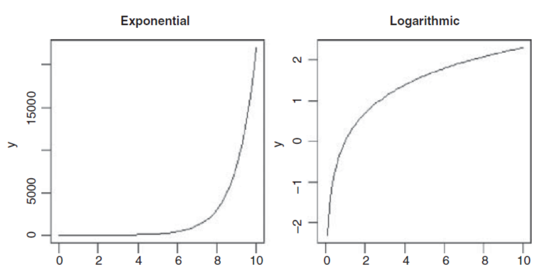
Antilogarithmic function $y = ae^{bx}$

Both are smooth functions

To draw smooth functions in R you need to generate a series of 100 or more regularly spaced *x* values between min(x)and max(x):

```
x <- seq(0,10,0.1)

windows(7,4)
par(mfrow=c(1,2))

y <- exp(x)
plot(y~x,type="l",main="Exponential")

y <- log(x)
plot(y~x,type="l",main="Logarithmic")
```

Type = 'l' is for lines



Mathematical Functions

Gamma function

The gamma function (*t*) is an extension of the factorial function, *t*!, to positive real numbers:

$$\Gamma(t) = \int_0^{\infty} x^{t-1} e^{-x} dx.$$

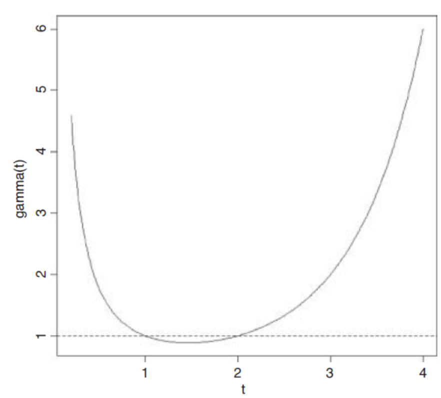
5

Mathematical Functions

Gamma function

It looks like this:

```
par(mfrow=c(1,1))
t <- seq(0.2,4,0.01)

plot(t,gamma(t),type="l")
abline(h=1,lty=2)
```



Note that (*t*) is equal to 1 at both *t* = 1 and *t* = 2.
For integer values of *t*, (*t* + 1) = *t*!

6

Background

Mathematical Functions

Asymptotic Functions

The most commonly used asymptotic function is

$$y = \frac{ax}{1 + bx}$$

which has a different name in almost every scientific discipline. It is called the Michaelis–Menten function in biochemistry. It is called the Holling's Disc Equation in ecology.

The graph passes through the origin and rises with diminishing returns to an asymptotic value at which increasing the value of x does not lead to any further increase in y.

Topic 2: Mathematical Distributions

Background

Mathematical Functions

The other common function is the asymptotic exponential.

$$y = a(1 - e^{-bx})$$

This is a two-parameter model.

For $x = 0$, y = 0. This means the graph goes through the origin.

For x = ∞, y → a. This means the asymptotic value of y is a.

Background

Mathematical Functions

Gompertz Growth Model

$$y = ae^{be^{cx}}$$

The shape of the function depends on the signs of the parameters b and c.

For a negative sigmoid, b is negative and c is positive.
For a positive sigmoid, b is negative and c is negative.

Topic 2: Mathematical Distributions

Background

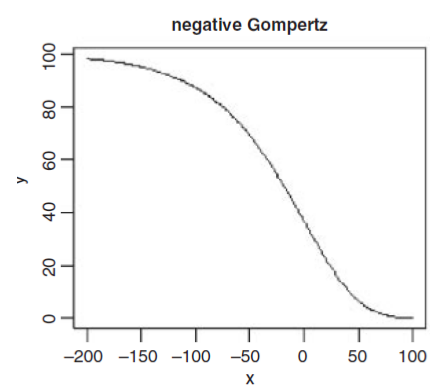
Mathematical Functions

Gompertz Growth Model

$$y = ae^{be^{cx}}$$

Negative Gompertz: $b = -1$ and $c = +0.02$.

```
x <- -200:100
y <- 100*exp(-exp(0.02*x))

plot(x,y,type="l",main="negative Gompertz")
```



Background

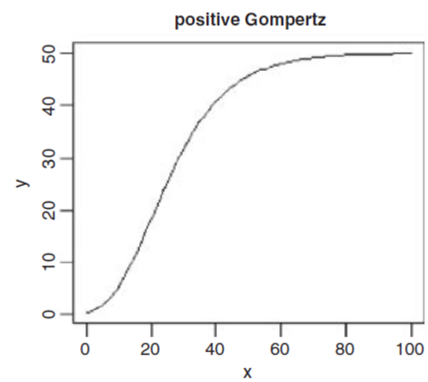
Mathematical Functions

Gompertz Growth Model

$$y = ae^{be^{cx}}$$

Positive Gompertz: $b = -5$ and $c = -0.08$.

```
x <- 0:100
y <- 50*exp(-5*exp(-0.08*x))

plot(x,y,type="l",main="positive Gompertz")
```



Key Concept

Mathematical Functions

Transformations of the Response and Explanatory Variables

We have seen the use of transformation to linearize the relationship between the response and the explanatory variables:

- log(y) against x: Exponential relationships
- log(y) against log(x): Power functions
- exp(y) against x: Logarithmic relationships
- $1/y$ against $1/x$: Asymptotic relationships
- log($p/(1-p)$) against x: Proportion data
- SQRT(y) to stabilize the variance for count data
- arcsin(y) to stabilize the variance of percentage data

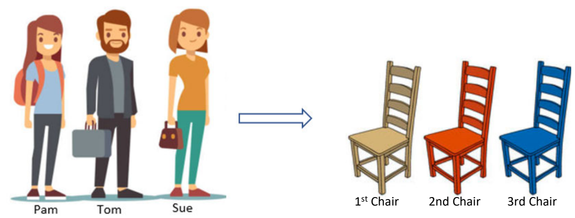
Probability functions

There are two ways to determine the number of ways you can select a sample of size n.

- Permutations – The Numbers Game – Order Matters
- Combinations – The Power Ball – Order Does Not matter

Let's focus on permutations first.

Example: We have 3 chairs and people. How many ways can we fill them when order matters?



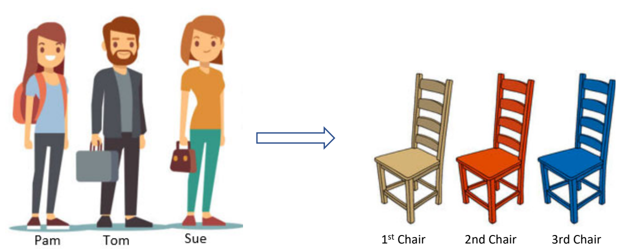
Topic 2: Mathematical Distributions

Probability functions

We could start enumerating all the possibilities:

1. Pam – Tom – Sue
2. Pam – Sue – Tom
3. Tom – Pam – Sue
4. Tom – Sue – Pam
5. Sue – Pam – Tom
6. Sue – Tom – Pam

This could take forever if we have more people and more chairs!



Probability functions

Instead, we could recognize the number of people left to fill remaining seats after seats have been filled.

3 x 2 x 1 = 6

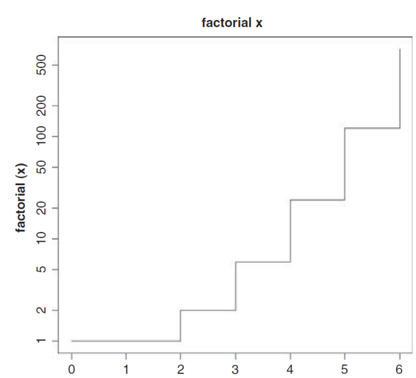
In general, the result is given by the factorial(n) is given by $n! = n(n-1)(n-2)\ldots \times 3 \times 2$ tells how many ways n items can be arranged. In this example n = 3. Therefore 3! = 6.

Probability functions

The R function is `factorial` and we can plot it for values of x from 0 to 10 using the step option `type="s"`, in plot with a logarithmic scale on the y axis `log="y"`

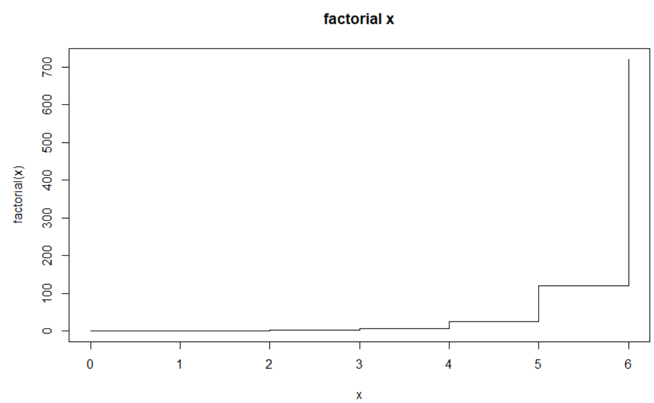
```
par(mfrow=c(1,1))
x <- 0:6

plot(x,factorial(x),type="s"
,main="factorial x",)
```

Note the parameter `log="y".` Because the factorial does not step up in a linear but in a logarithmic way. It would be difficult to measure the effect using a linear scale. Let's look at the graph without this parameter.



Probability functions

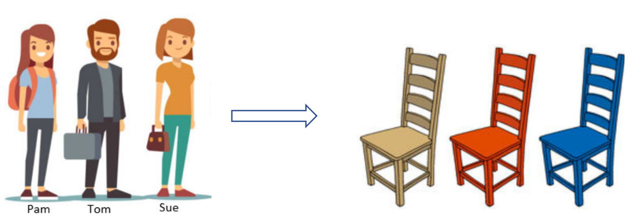
```
par(mfrow=c(1,1))
x <- 0:6

plot(x,factorial(x),type="s"
,main="factorial x",)
```



17

Probability functions

Let's now focus on combinations, where order does not matter.



How many ways can we fill all three chairs with three people when it does not matter where they sit?

18

Probability functions

Let's say we have six people and three chairs.

Answer: $= \dfrac{6\ x\ 5\ x\ 4}{3\ x\ 2\ x\ 1} = \dfrac{\textit{These all the permutations}}{\textit{Get rid of the duplicates}} = \dfrac{6!}{3!\,3!} = \underbrace{\dfrac{6!}{3!\,(6-3)!}}_{\text{Turn the expression into a combination formulas}}$

19

Key Concept

Probability functions

General Combinatorics Formula

Number of combinations (order does not matter) of *n* things taken *r* at a time:

$$C(n,r) = \frac{n!}{(n-r)!\,r!}$$

R Function: Choose (n, r)

20

Probability functions

How many ways to select 3 books from 6?

We can always start to enumerate. Now order does not matter.

1.	#1 #2 #3	11.	#2 #3 #4
2.	#1 #2 #4	12.	#2 #3 #5
3.	#1 #2 #5	13.	#2 #3 #6
4.	#1 #2 #6	14.	#2 #4 #5
5.	#1 #3 #4	15.	#2 #4 #6
6.	#1 #3 #5	16.	#2 #5 #6
7.	#1 #3 #6	17.	#3 #4 #6
8.	#1 #4 #5	18.	#3 #4 #6
9.	#1 #4 #6	19.	#3 #5 #6
10.	#1 #5 #6	20.	#4 #5 #6

This approach is tedious, and it is easy to miss a combination!

Probability functions

The Mathematical Relationship between Permutations and Combinations:

"Combinations get rid of the order in Permutations"

How many ways to select 3 books from 6 if order matters?

1.	#1 #2 #3	11.	#2 #3 #4
2.	#1 #2 #4	12.	#2 #3 #5
3.	#1 #2 #5	13.	#2 #3 #6
4.	#1 #2 #6	14.	#2 #4 #5
5.	#1 #3 #4	15.	#2 #4 #6
6.	#1 #3 #5	16.	#2 #5 #6
7.	#1 #3 #6	17.	#3 #4 #6
8.	#1 #4 #5	18.	#3 #4 #6
9.	#1 #4 #6	19.	#3 #5 #6
10.	#1 #5 #6	20.	#4 #5 #6

For each combination, we would need to enumerate 3! more combinations to list all the permutations of 3.

Probability functions

The permutation is given by:

$$\frac{6!}{3!} = 120$$

This means for every set of three unique numbers there are 3! Or 6 permutations.
For example, for the set #1 #2 #3, we have:

1. #1 #2 #3
2. #1 #3 #2
3. #2 #1 #3
4. #2 #3 #1
5. #3 #1 #3
6. #2 #3 #1

Each of the 20 combinations will have 3! or 6 permutations for a total of 120 permutations for selecting 3 items from a list of 6 items.

Key Concept

Probability functions

This means the number of combinations = The total number of permutations divided by factorial of the subset size (r).

Or, we can say

The number of permutations equals the number of combinations times r!

Permutations and Combinations

Number of permutations (order matters) of *n* things taken *r* at a time:	Number of combinations (order does not matter) of *n* things taken *r* at a time:
$P(n,r) = \frac{n!}{(n-r)!}$	$C(n,r) = \frac{n!}{(n-r)!r!}$

Probability functions

The last example is not the most interesting case! We are more interesting in determining how many ways we can choose x items out of a total of n items. Let's think of a card game like poker:

N = 52 – A deck of cards has 52 cards
X = 5 – A poker hand is made up of 5 cards

How many ways can we choose 5 cards from a deck of 52?

25

Topic 2: Mathematical Distributions

Probability functions

This is not the most interesting case! We are more interesting in determining how many ways we can chose x items out of a total of n items. Let's think of a card game like poker:

N = 52 – A deck of cards has 52 cards
X = 5 – A poker hand is made up of 5 cards

How many ways can we choose 5 cards from a deck of 52?
What is the notation?

$$\binom{52}{5} \longrightarrow \binom{52}{5} = \frac{52!}{5!\,47!} = 2,598,960$$

26

Probability functions

What would the relationship be if we selected 4 items from a list of 8 items?

How many permutations of unique sets of 4 would there be?

27

Topic 2: Mathematical Distributions

Probability functions

Calculating combinations in R:

```
>choose(8,4)
[1] 70
```

$$\binom{8}{4} = \frac{8!}{4!\,(8-4)!} = \frac{8\,x\,7\,x\,6\,x\,5}{4\,x\,3\,x\,2} = 70$$

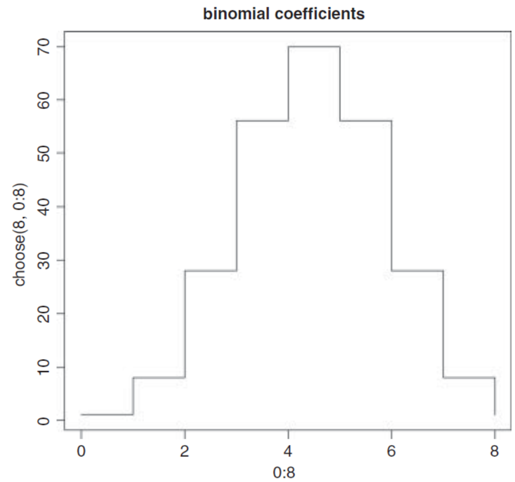
Note: 0! = 1

28

Probability functions

Plotting combinations in R:

```
plot(0:8,choose(8,0:8),type="s"
,main="binomial coefficients")
```



HW: Use R Functions for the following

1. Calculate the probabilities of the following poker hands:
    a. A hand with all diamonds
    b. A hand that has 3 diamonds and 2 hearts
    c. All blacks cards

2. The number of signals that can be sent by 6 flags of different colors taking one or more at a time is
    (A) 63
    (B) 1956
    (C) 720
    (D) 21

HW: Use R Functions for the following

3. The Florida Lotto Saturday night drawing used to work like this: There are 49 ping-pong balls in a machine, each bearing a number from 1 to 49. The machine randomly spits out 6 ping-pong balls. If the numbers on the ping-pong balls match the six numbers that you chose, YOU WIN! How many different outcomes are possible?

4. Now, the Lotto works like this: there are 53 balls instead of 49. How many outcomes are possible under this new scheme?

5. In how many ways can a group of 5 men and 2 women be made out of a total of 7 men and 3 women?
   A. 64
   B. 1
   C. 126
   D. 63

Topic 2: Mathematical Distributions

HW: Use R Functions for the following

6. How may ways can you pick a team of three people from a group of 10?

7. In a group of 6 boys and 4 girls, four children are to be selected. In how many different ways can they be selected such that at least one boy should be there?
   A. 212
   B. 209
   C. 159
   D. 201

8. From a group of 7 men and 6 women, five persons are to be selected to form a committee so that at least 3 men are there in the committee. In how many ways can it be done?
   A. 702
   B. 624
   C. 756
   D. 812

Continuous Probability Distributions

Built-in probability distributions:

- `d` – the probability density function
- `p` - the cumulative probability
- `q` - the quantiles of the distribution
- `r` – the random numbers generated from the distribution

Simply prefix the name of the distribution with one of these functions to generate the values.

For example:

```
curve(dnorm(x),-3,3) #Produces the density function
curve(pnorm(x),-3,3) #Produces the cumulative distribution
curve(qnorm(x),-3,3) #Produces the quantile plot
curve(rnorm(x),-3,3) #Produces plot of normally dist. random numbers
```


Topic 2: Mathematical Distributions

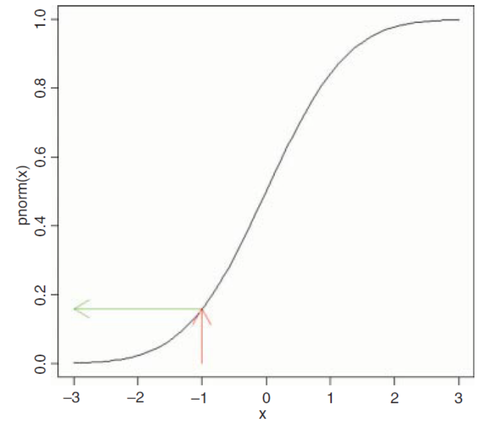
Probability distributions supported by R

R function	Distribution	Parameters
beta	beta	shape1, shape2
binom	binomial	sample size, probability
cauchy	Cauchy	location, scale
* exp	exponential	rate (optional)
* chisq	chi-squared	degrees of freedom
* F	Fisher's *F*	df1, df2
* gamma	gamma	shape
geom	geometric	probability
hyper	hypergeometric	*m*, *n*, *k*
* lnorm	lognormal	mean, standard deviation
* logis	logistic	location, scale
nbinom	negative binomial	size, probability
* norm	normal	mean, standard deviation
* pois	Poisson	mean
signrank	Wilcoxon signed rank statistic	sample size *n*
t	Student's *t*	degrees of freedom
unif	uniform	minimum, maximum (opt.)
* weibull	Weibull	shape
wilcox	Wilcoxon rank sum	*m*, *n*

* Common ones used in insurance modeling. There are others we will introduce later.

The Cumulative Probability

```
curve(pnorm(x),-3,3)
```



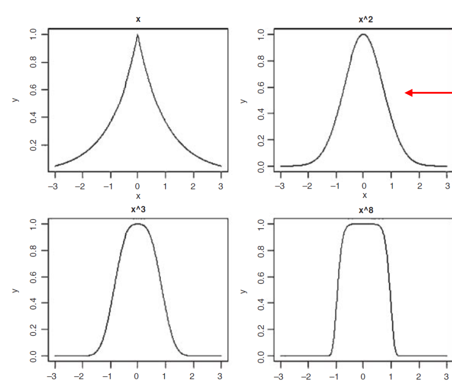
35

Normal Distribution

This distribution is central to the theory of parametric statistics. Consider the following simple exponential function:

$$y = \exp(-|x|^m)$$

```
par(mfrow=c(2,2))
x <- seq(-3,3,0.01)

y <- exp(-abs(x))
plot(x,y,type="l",main= "x")

y <- exp(-abs(x)^2)
plot(x,y,type="l",main= "x^2")

y <- exp(-abs(x)^3)
plot(x,y,type="l",main= "x^3")

y <- exp(-abs(x)^8)
plot(x,y,type="l",main= "x^8")
```



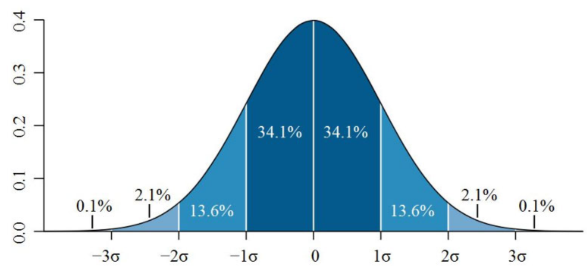
The basis of the Normal Distribution

36

The Standard Normal Distribution

The normal distribution with a mean 0 and standard deviation 1 is called the **standard normal** distribution and the equation is:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$



Problem: Not all distributions have a 0 mean & SD =1.

37

Key Concept

Converting to a Standard Normal Distribution

Step 1: Calculate the mean of the data

Step 2: Calculate the standard deviation of the data

Step 3: For each x, calculate its z value as follows

$$z = \frac{x - \bar{x}}{\sigma_x} \quad \text{A Z-Score}$$

Now we can use the normal distribution tables to calculate any needed probability. The standard normal tables are assumed in the R functions.

What is the big drawback from this approach?

Examples:

```
>pnorm(-1.25)
[1] 0.1056498

>pnorm(1.875)
[1] 0.9696036

>1-pnorm(1.875)
[1] 0.03039636
```


38

Example

Suppose we have measured the heights of 100 people. The mean height was 170 cm and the standard deviation was 8 cm. We can ask three sorts of questions about data like these: what is the probability that a randomly selected individual will be:

a. shorter than 160 cm?
b. taller than a 180 cm?
c. between 160 cm and 180 cm?

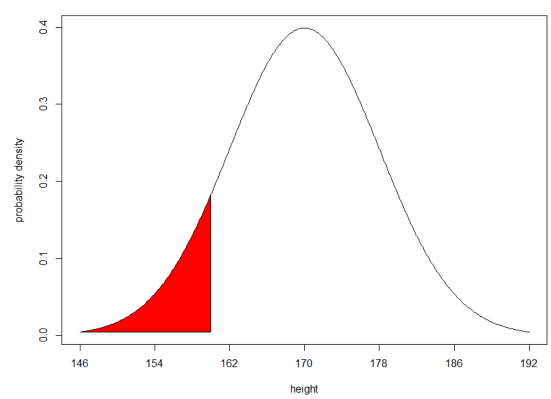
Step 1: Convert 160 cm and 180 cm to Z-Scores

```
Z1 <- (160 – 170)/8    #Z1 = – 1.25
Z2 <- (180 – 170)/8  #Z2 = + 1.75
```


Topic 2: Mathematical Distributions

Solution (a): Prob (X < 160)

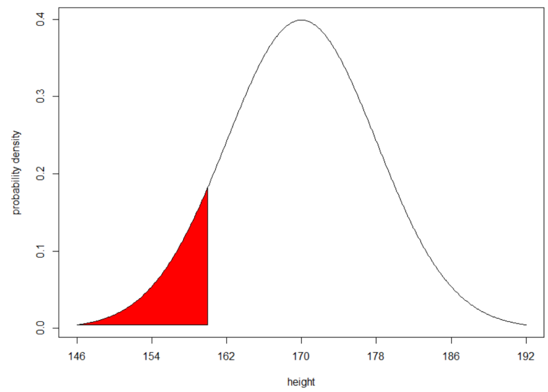
```
pnorm(Z1) = 0.1056498
```



Background

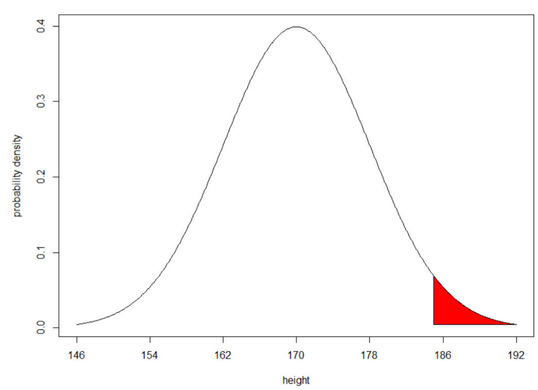
The R Code for this problem

```
x <- seq(-3,3,0.01)
z <- seq(-3,-1.25,0.01)
p <- dnorm(z)
z <- c(z,-1.25,-3)
p <- c(p,min(p),min(p))

plot(x,dnorm(x),type="l",xaxt="n",ylab="pro
bability density",xlab="height")

axis(1,at=-3:3, labels=c("146","154","162",
"170","178","186","192"))

polygon(z,p,col="red")
```



Solution (b): Prob(X > 180)

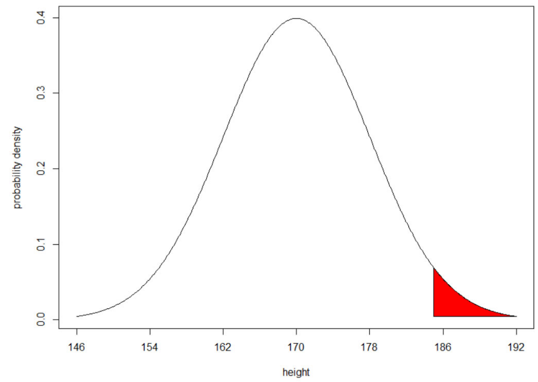
```
1 - pnorm(Z2) = 0.03039636
```



Background

The R Code for this problem

```
z <- seq(1.875,3,0.01)
p <- dnorm(z)
z <- c(z,3,1.875)
p <- c(p,min(p),min(p))

plot(x,dnorm(x),type="l",xaxt="n",ylab="pro
bability density",xlab="height")

axis(1,at=-
3:3,labels=c("146","154","162","170","178",
"186","192"))

polygon(z,p,col="red")
```



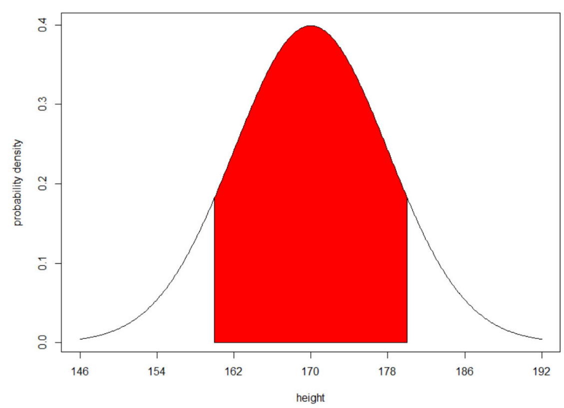
Solution (c): Prob(160 < X < 180)

Given our first two answers how do we calculates this probability?

Prob (X < 160) = 0.1056498

Prob(X > 180) = 0.03039636

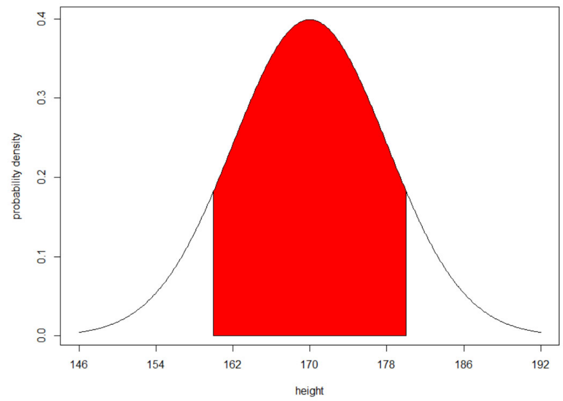
Prob(160 < X < 180) = 1 - 0.1056498 - 0.03039636
= 0.863954



Background

The R Code for this problem

```
z <- seq(-1.25,1.25,0.01)
p <- dnorm(z)
z <- c(z,1.25,-1.25)
p <- c(p,0,0)

plot(x,dnorm(x),type="l",xaxt="n",ylab="probability density",xlab="height")

axis(1,at=-3:3,labels=c("146","154","162","170","178","186","192"))

polygon(z,p,col="red")
```



45

Key Concept

The Normal Distribution Functions in R

dnorm(x, mean = 0, sd = 1, log = FALSE)
pnorm(q, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
qnorm(p, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
rnorm(n, mean = 0, sd = 1)

Note: These functions require the standard deviation. If given the variance, you need to take the square root.

Arguments
x, q – vector of quantiles
p – vector of probabilities.
n – number of observations. If length(n) > 1, the length is taken to be the number required
mean – vector of means
sd – vector of standard deviations
log, log.p – logical; if TRUE, probabilities p are given as log(p)
lower.tail – logical; if TRUE (default), probabilities are P[X ≤ x], otherwise, P[X > x]

46

Example

Suppose widgit weights produced at Acme Widgit Works have weights that are normally distributed with mean 17.46 grams and variance 375.67 grams. What is the probability that a randomly chosen widgit weighs more then 19 grams?

Answer:

```
Prob <-pnorm(19, mean=17.46, sd=sqrt(375.67))
Ans <- 1 - Prob
Ans
[1] 0.4683356
```

Answer using Z-scores:

```
z=(19 - 17.46)/sqrt(375.67)
Prob<-pnorm(z)
Ans<- 1 - Prob
Ans
[1] 0.4683356
```

Note: We get the same answer!

Homework

1. Suppose IQ scores are normally distributed with mean 100 and standard deviation 15. What is the 95th percentile of the distribution of IQ scores?

2. Generates 1000 independent and identically distributed normal random numbers (first line), plots their histogram (second line), and graphs the p. d. f. of the same normal distribution (third and forth lines). Assume the mean = 100 and the sd = 15.

The Central Limit Theorem

If you take repeated samples from a population with finite variance and calculate their averages, then the averages will be normally distributed. It turns out it does not matter what distribution the data comes from!

The mean of the means will be normally distributed!!!

This is an important concept for sampling distributions.

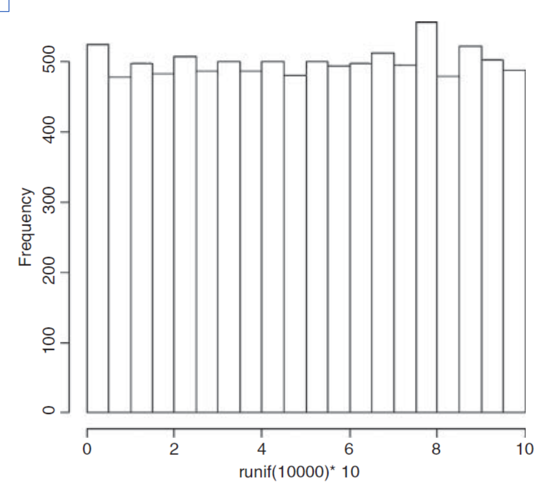
Example 1: Calculating 10,000 Random Numbers

Take 10,000 samples of 5 random numbers form the uniform Distribution between 0 and 10.

```
Multiplier<-10
par(mfrow=c(1,1))
hist(runif(10000)*multiplier,main="")
```

"runif" means random uniform

It will create 10,000 numbers between 0 and 1, then multiply each one by 10 to get them between 0 and 10.

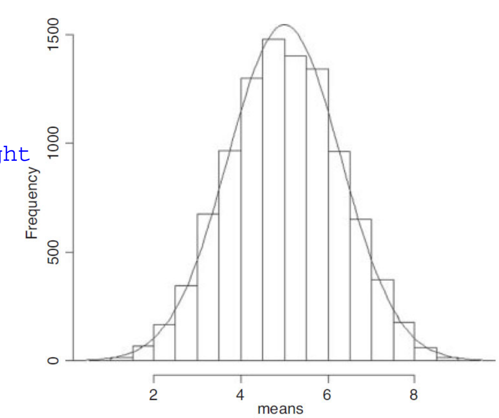


Example 1: Calculating 10,000 Means

```
means <- numeric(10000)

for (i in 1:10000)
{
means[i] <- mean(runif(5)*10)
}

yHeight<- 1600
hist(means,ylim=c(0, yHeight),main="")

xbar<-mean(means)
xbar
[1] 4.998581

Sdbar<-sd(means)
Sdbar
[1] 1.289960
```


Example 1: Plot Histogram

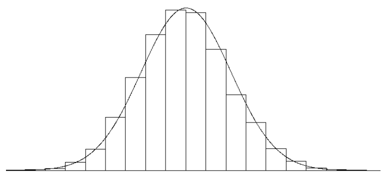
```
seqFrom <-0
seqTo<-10
segInc<-0.1
normheight<-5000

xv <- seq(seqFrom, seqTo, segInc)
yv <- dnorm(xv,mean=4.998581,sd=1.28996)* normheight
lines(xv,yv)
```



Homework

Repeat Example 1 (slides 50 – 52) using a Poisson Distribution with lambda = 25. Use rpois(10000, 10)

Notes:
1. Pay attention to the multiplier not all distribution given random numbers between 0 and 1.
2. You need to use the parameter.
3. You will have to adjust your Height since this is the Poisson Distribution
4. Other parameters you need to adjust to get the curve to fit nicely:
    a. seqFrom
    b. seqTo
    c. segInc
    d. normheight



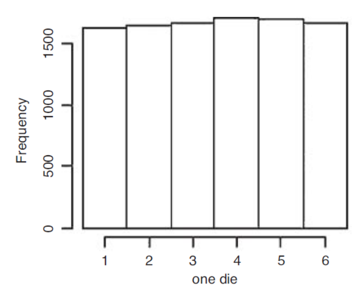
You should get a fit like this or better.

53

Other examples of Uniform Distributions

Throw one die lots of times and each of the six numbers should come up equally often.

```
par(mfrow=c(2,2))
hist(sample(1:6,replace=T,10000),breaks=0.5:6.5,main="",xlab="one die")
```



54

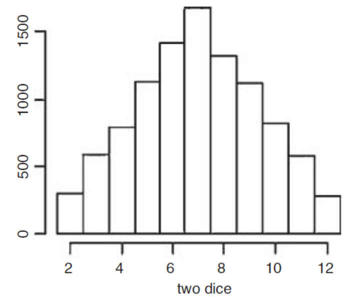
The Game of Craps

Now throw two dice and add the scores together. There are 11 possible scores from a minimum of 2 to a maximum of 12.

The most likely score is 7 because there are six ways that this could come about: (1,6) (6,1) (2,5) (5,2) (3,4) (4, 3)

For many throws of craps we get a **triangular** distribution of scores, centered on 7:

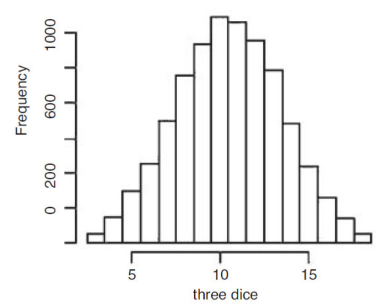
```
a <- sample(1:6,replace=T,10000)
b <- sample(1:6,replace=T,10000)

hist(a+b,breaks=1.5:12.5,main="", xlab="two dice")
```



Three Dice

For three dice we get

```
c <- sample(1:6,replace=T,10000)

hist(a+b+c,breaks=2.5:18.5,main="", xlab="three dice"
```



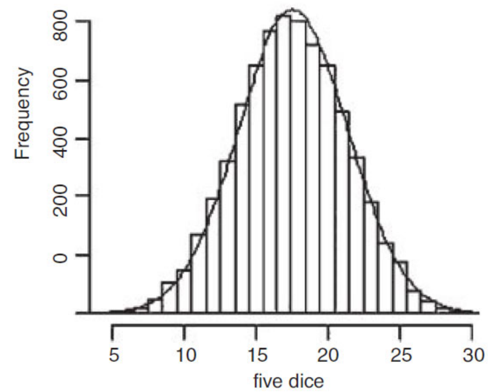
Five Dice

The **binomial** distribution is virtually indistinguishable from the normal:

```
d <- sample(1:6,replace=T,10000)
e <- sample(1:6,replace=T,10000)

hist(a+b+c+d+e,breaks=4.5:30.5,main="",
xlab="five dice")
```

The smooth curve is given by a normal distribution with the same mean and standard deviation:

```
xbar<- mean(a+b+c+d+e)
sdbar<- sd(a+b+c+d+e)

lines(seq(1,30,0.1),dnorm(seq(1,30,0.1),xbar,
sdbar)*10000)
```



As we introduce more & more coins, the we converge to normal distributions

Normal Probability Density Function

The probability density of the normal is

$$f(y|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right]$$

Maximum Likelihood Function

The likelihood function is the product of the probability densities, for each of the values of the response variable, *y*

$$L(\mu, \sigma) = \prod_{i=1}^{n} \left(\frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(y_i - \mu)^2}{2\sigma^2} \right] \right)$$

for $y_1, y_2, y_3, y_4, ,,, y_n$

With a little algebra this expression can be simplified

Maximum Likelihood Function

There are n factors of $\frac{1}{\sigma\sqrt{2\pi}} \longrightarrow \frac{1}{\left(\sigma\sqrt{2\pi}\right)^n}$

And are n factors of $\exp\left[-\frac{(y_i - \mu)^2}{2\sigma^2} \right] \longrightarrow \exp\left[-\frac{\sum_{i=1}^{n} (y_i - \mu)^2}{2\sigma^2} \right]$

Maximum Likelihood Function

The likelihood function simplifies to:

$$L(\mu, \sigma) = \frac{1}{\left(\sigma\sqrt{2\pi}\right)^n} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right]$$

Log Likelihood Function

The Log Likelihood Function is

$$l(\mu, \sigma) = -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \sum(y_i - \mu)^2/2\sigma^2$$

Log Likelihood Function

Apply partial derivatives to solve for the parameters

$$\frac{\mathrm{d}l}{\mathrm{d}\mu} = \sum (y_i - \mu)/\sigma^2$$

$$\frac{\mathrm{d}l}{\mathrm{d}\sigma} = -\frac{n}{\sigma} + \frac{\sum (y_i - \mu)^2}{\sigma^3}$$

Log Likelihood Function

The solutions are

$$\mu = \frac{\sum y_i}{n}$$

$$\sigma^2 = \frac{\sum (y_i - \mu)^2}{n}$$

The maximum likelihood estimate of μ is the arithmetic mean.

This is a biased estimate of the variance, however, because it does not take account of the fact that we estimated the value of μ from the data. To unbias the estimate, we need to lose 1 degree of freedom to reflect this fact, and divide the sum of squares by $n - 1$ rather than by n

Key Concept

Log Likelihood Function

When the distribution in the likelihood function is the Normal Distribution,

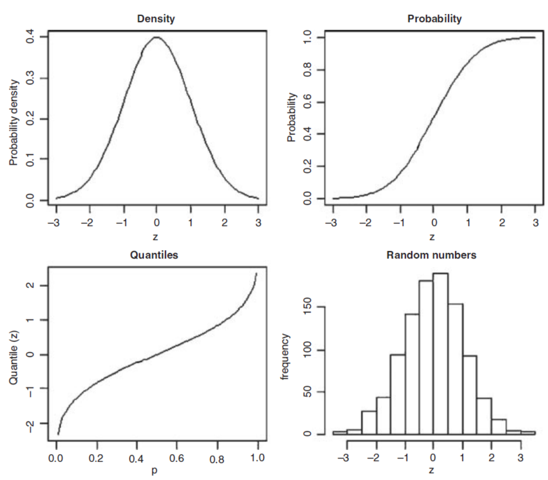
$$L(\mu, \sigma) = \prod_{i=1}^{n} \left(\frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(y_i - \mu)^2}{2\sigma^2} \right] \right)$$

for y_1, y_2, y_3, y_4, ,,, y_n

The parameter estimates are the same as those from Ordinary Least Squares Regression.

The 4 R Functions for the Normal Distribution

```
par(mfrow=c(2,2))

curve(dnorm,-3,3,xlab="z",ylab="Probability
density",main="Density")

curve(pnorm,-3, 3, xlab="z", ylab="Probability",
main="Probability")

curve(qnorm,0,1,xlab="p",ylab="Quantile
(z)",main="Quantiles")

y <- rnorm(1000)
hist(y,xlab="z",ylab="frequency",main="Random
numbers")
```



Background

Generating random numbers with exact mean and standard deviation

```
yvals <- rnorm(100,24,4)

mean(yvals)
[1] 24.2958

sd(yvals)
[1] 3.5725
```

Not quite spot on!

Try ➔

```
ydevs <- rnorm(100,0,1)
ydevs <- (ydevs-mean(ydevs))/sd(ydevs)

mean(ydevs)
[1] -2.449430e-17
sd(ydevs)
[1] 1
```

$$ydevs = \frac{yval - mean(ydevs)}{sd(ydevs)}$$

```
yvals <- 24 + ydevs*4
mean(yvals)
[1] 24
sd(yvals)
[1] 4
```

We get the exact μ and σ

HW: Run the code to the right and verify the results ->

Comparing data with a normal distribution

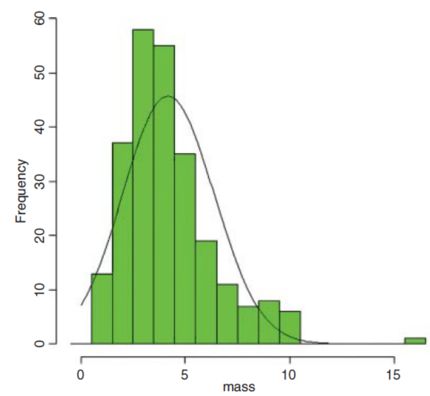
Here we are concerned with the task of comparing a histogram of real data with a smooth normal distribution with the same mean and standard deviation, in order to look for evidence of non-normality (e.g. skew or kurtosis)

```
par(mfrow=c(1,1))
fishes <- read.table("c:\\temp\\fishes.txt",header=T)
attach(fishes)
names(fishes)
[1] "mass"

mean(mass)
[1] 4.194275

max(mass)
[1] 15.53216
```

Comparing data with a normal distribution

Now the histogram of the mass of the fish is produced, specifying integer bins that are 1 gram in width, up to a maximum of 16.5 g:

```
hist(mass,breaks=-0.5:16.5,col="green", main="")
lines(seq(0,16,0.1),length(mass)*dnorm(seq(0,16,0.1)
,mean(mass),sqrt(var(mass))))
```



The distribution of fish sizes is clearly *not* normal. There are far too many fish of 3 and 4 grams, too few of 6 or 7 grams, and too many really big fish (more than 8 grams). This kind of skewed distribution is probably better described by a gamma distribution than a normal distribution.

Other distributions used in hypothesis testing

The main distributions used in hypothesis testing are:

1. **Chi-squared**, for testing hypotheses involving count data;
2. **Fisher's *F***, in analysis of variance (ANOVA) for comparing two variances
3. **Student's *t***, in small sample work for comparing two parameter estimates.

These distributions tell us the size of the test statistic that could be expected by chance alone when nothing was happening (i.e. when the null hypothesis was true).

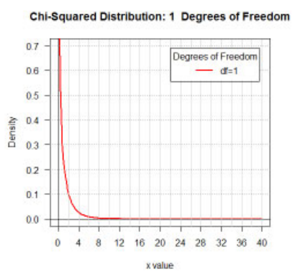
Given the rule that a big value of the test statistic tells us that something *is* happening, and hence that the null hypothesis is false, these distributions define what constitutes a big value of the test statistic (its **critical value**).

Background

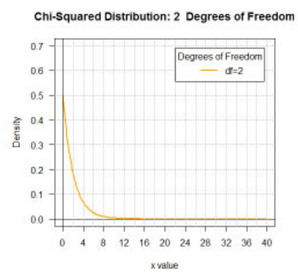
Chi Square Distribution

The χ^2 distribution also changes for different **degrees of freedom**. It is not symmetric and it is defined only for positive values.

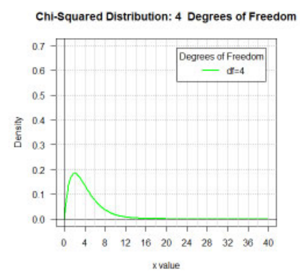
Graph 1



Graph 2



Graph 3

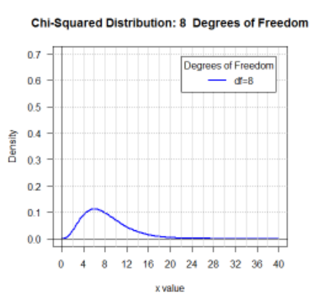


Background

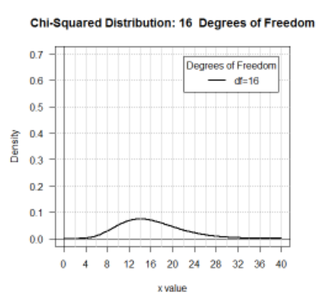
Chi Square Distribution

The χ^2 distribution also changes for different **degrees of freedom**. It is not symmetric and it is defined only for positive values.

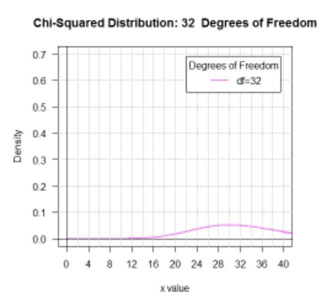
Graph 4



Graph 5



Graph 6

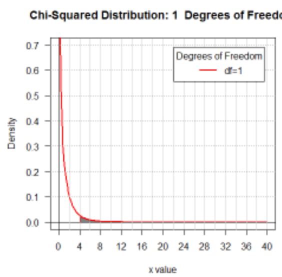


Background

Chi Square Distribution: Area & Probability

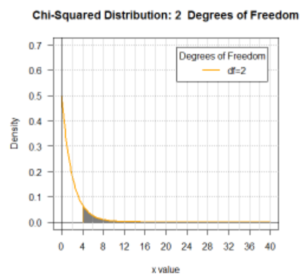
The χ^2 distribution also changes for different **degrees of freedom**. It is not symmetric and it is defined only for positive values.

Graph 7



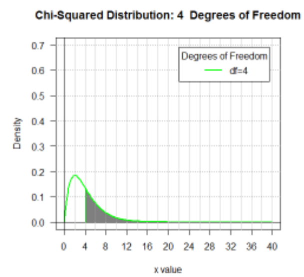
P(x > 4) ≈ 0.04550026

Graph 8



P(x > 4) ≈ 0.1353353

Graph 9



P(x > 4) ≈ 0.4060058

73

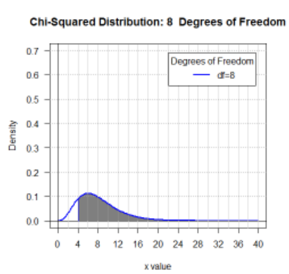
Topic 2: Mathematical Distributions

Background

Chi Square Distribution: Area & Probability

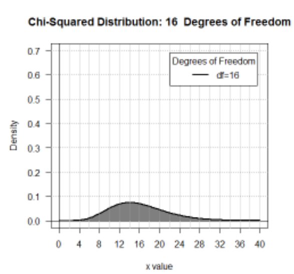
The χ^2 distribution also changes for different **degrees of freedom**. It is not symmetric and it is defined only for positive values.

Graph 10



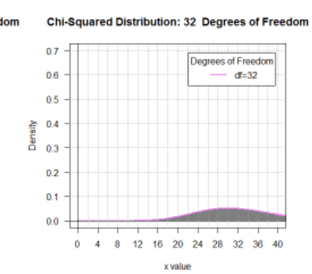
P(x > 4) ≈ 0.8571235

Graph 11



P(x > 4) ≈ 0.9989033

Graph 12



P(x > 4) ≈ 1

74

Key Concept

The Chi-Squared Distribution

The second-best known of all the statistical distributions. It is a special case of the gamma distribution characterized by a single parameter, the number of degrees of freedom.

The mean is equal to the degrees of freedom v and the variance is equal to $2v$. The density function is

$$f(x) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\nu/2-1} e^{-x/2}$$

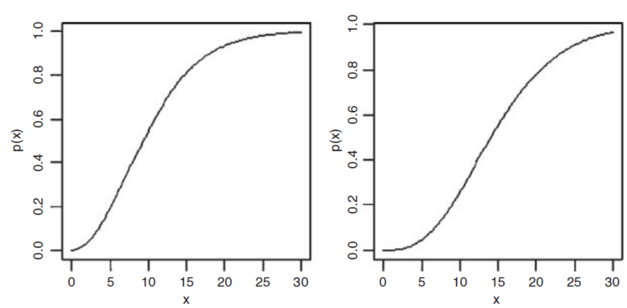
If the non-central chi-squared is the sum of v independent normal random variables, then the non-centrality parameter is equal to the sum of the squared means of the normal variables.

A common distribution used to model claim severity or size of claims.

The Chi-Squared Distribution

Cumulative probability plots for a non-centrality parameter (`ncp`) based on three normal means (of 1, 1.5 and 2) and another with 4 means and `ncp` = 10:

```
par(mfrow=c(1,2))
x <- seq(0,30,.25)

plot(x,pchisq(x,3,7.25),type="l",ylab
="p(x)",xlab="x")

plot(x,pchisq(x,5,10),type="l",ylab="
p(x)",xlab="x")
```



The non-centrality parameter (ncp) means the chi-squared distribution is not centered at zero.

Key Concept

The Chi-Squared Distribution

95% Confidence Interval for σ^2

$$\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{\alpha/2}}$$

Suppose the sample variance $s^2 = 10.2$ on 8 d.f. Then the interval on σ^2 is given by

```
8*10.2/qchisq(.975,8)
[1] 4.65367

8*10.2/qchisq(.025,8)
[1] 37.43582
```

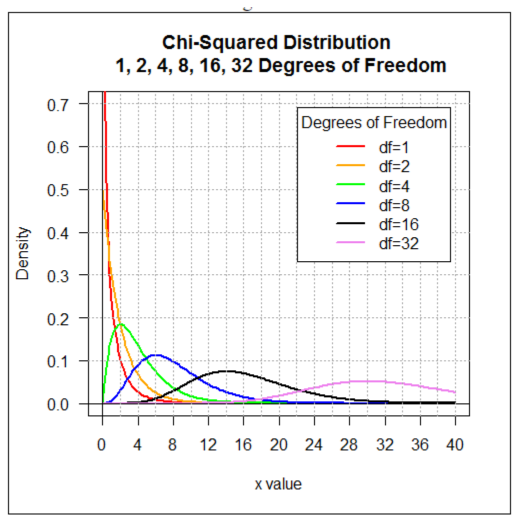
which means that we can be 95% confident that the population variance lies in the range $4.65 \leq \sigma^2 \leq 37.44$.

77

Key Concept

Chi Square Distribution: Area & Probability

Observe the progression of the χ^2 distribution as degrees of freedom increase



pchisq(q, df, ncp = 0, lower.tail = FALSE, log.p = FALSE

q = test value
df = degrees of freedom
ncp = non-centrality parameter
lower.tail = determines area to right or left
log.p = determines probabilities as log probabilities

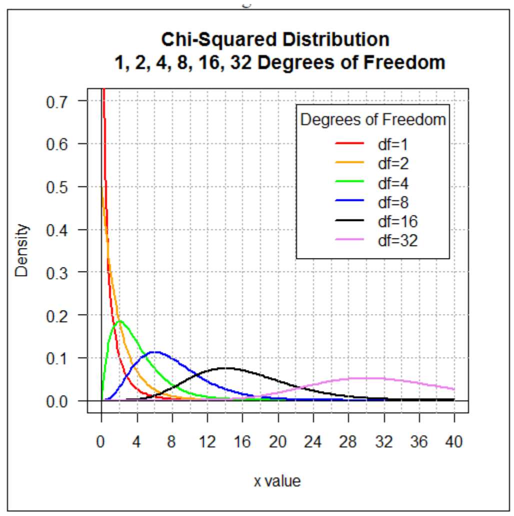
78

Chi Square Distribution: Area & Probability

This is the quantile function

qchisq(q, df, ncp = 0, lower.tail = FALSE, log.p = FALSE

q = area value
df = degrees of freedom
ncp = non-centrality parameter
lower.tail = determines area to right or left
log.p = determines probabilities as log probabilities



Topic 2: Mathematical Distributions

Chi Square Test Statistics

```
pchisq(Test statistic, Degrees of Freedom) #Calculates the
cumulative distribution function

qchisq(Area Value, Degrees of freedom) #Calculates the desired
critical value given an area to the right of it.

Area Value means area under the curve - 0 ≤ Area Value ≤ 1
```

Homework

1. For a χ^2 distribution with 6 degrees of freedom, what is the probability of having a random event **X** be less than 2.34?
2. For a χ^2 distribution with 9 degrees of freedom, what is the probability of having a random event **X** be greater than 15.34?
3. For a χ^2 distribution with 17 degrees of freedom, what is the probability of having a random event **X** be less than 6.66 or greater than 27.34?
4. For a χ^2> distribution with 14 degrees of freedom, what is the probability of having a random event **X** be between 5.25 and 25.41?

Homework

5. For a χ^2 distribution with 5 degrees of freedom, what is the **quantile** that has 0.0333 square units under the curve and to the **left** of that **quantile**?
6. For a χ^2 distribution with 25 degrees of freedom, what is the **quantile** that has 0.125 square units under the curve and to the **right** of that **quantile**?
7. For a χ^2 distribution with 11 degrees of freedom, what are the **quantiles** that have 0.75 square units under the curve and between those **quantiles** with the tails having equal areas?
8. For a χ^2 distribution with 23 degrees of freedom, what are the **quantiles** that have 0.0333 square units under the curve and to the outside the interval between those **quantile** where the tails have equal areas?

Key Concept

Fisher's *F* distribution

This is the famous variance ratio test that occupies the penultimate column of every ANOVA table. This the ratio of treatment variance to error variance and it follows the *F* distribution.

Use the quantile **qf** to look up critical values of *F*.

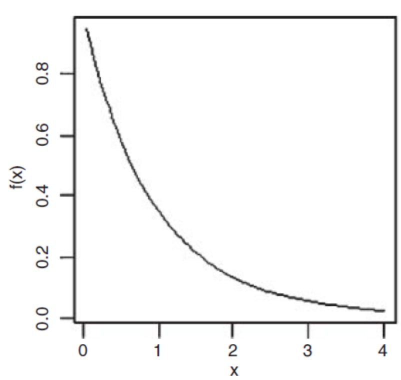
```
qf(.95,2,18)
[1] 3.554557
```

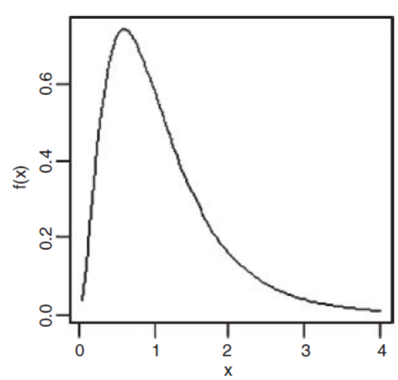
The F Statistic is the ratio of two variances. The numerator has degrees of freedom (d.f.) and the denominator has degrees of freedom.

Key Concept

The Density Function of *F* distribution

This is what the density function of *F* looks like for 2 and 18 d.f. (left) and 6 and 18 d.f. (right):





Key Concept

The Density Function of *F* distribution

The F distribution is a two-parameter distribution defined by the density function

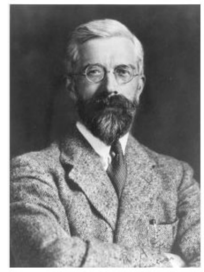
$$f(x) = \frac{r\Gamma(1/2(r+s))}{s\Gamma(1/2r)\Gamma(1/2s)} \frac{(rx/s)^{(r-1)/2}}{[1+(rx/s)]^{(r+s)/2}}$$

where,

r is the degrees of freedom in the numerator
s is the degrees of freedom in the denominator

Used to *assess the significance of the differences between two variances.*

The distribution is equal to the square of Student's t: $F = t^2$.



R. A. Fisher (1890–1962)

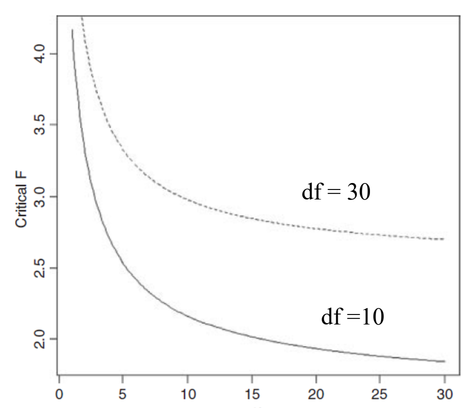
The distribution is named after R.A. Fisher, the father of analysis of variance & modern-day statistics, and principal developer of quantitative genetics.

The Density Function of *F* distribution

While the rule of thumb for the critical value of Student's t is 2, the rule of thumb for $F = t_2 = 4$.

To see how well the rule of thumb works, we can plot critical F against d.f. in the numerator:

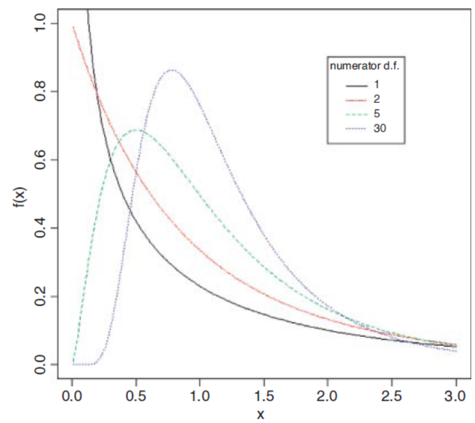
```
windows(7,7)
par(mfrow=c(1,1))
df <- seq(1,30,.1)
plot(df,qf(.95,df,30),type="l",ylab="Critical F")
lines(df,qf(.95,df,10),lty=2)
```



The Density Function of *F* distribution

The shape of the density function of the *F* distribution depends on the degrees of freedom in the numerator.

```
x <- seq(0.01,3,0.01)
plot(x,df(x,1,10),type="l",ylim=c(0,1),ylab
="f(x)")

lines(x,df(x,2,10),lty=6,col="red")
lines(x,df(x,5,10),lty=2,col="green")
lines(x,df(x,30,10),lty=3,col="blue")

legend(2,0.9,c("1","2","5","30"),col=(1:4),
lty=c(1,6,2,3), title="numerator d.f.")
```



The ANOVA Table

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	9543.721	4	2385.930	46.695	.000[a]
	Residual	9963.779	195	51.096		
	Total	19507.500	199			

a. Predictors: (Constant), reading score, female, social studies score, math score

b. Dependent Variable: science score

Topic 2: Mathematical Distributions

Key Concept

Student's *t* Distribution

This famous distribution was first published by W.S. Gossett in 1908 under the pseudonym of 'Student' because his then employer, the Guinness brewing company in Dublin, would not permit employees to publish under their own names.

It is a model with one parameter, r, with density function

$$f(x) = \frac{\Gamma(1/2(r+1))}{(\pi r)^{1/2}\Gamma(1/2r)}\left(1+\frac{x^2}{r}\right)^{-(r+1)/2}$$

If you remove all the constants, you get

$$f(x) = \left(1+x^2\right)^{-1/2}$$

UNC CHARLOTTE 89

89

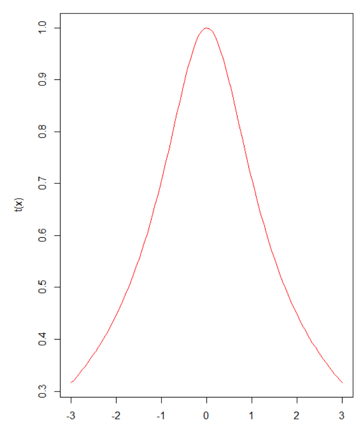
Topic 2: Mathematical Distributions

Key Concept

Student's *t* Distribution

We can plot this for values of x from –3 to +3 as follows:

```
curve( (1+x**2)**(-0.5), -3, 3,ylab="t(x)",col="red")
```



The main thing to notice is how fat the tails of the distribution are, compared with the normal distribution. →

Note error in text. "^" used for exponentiation but should be **

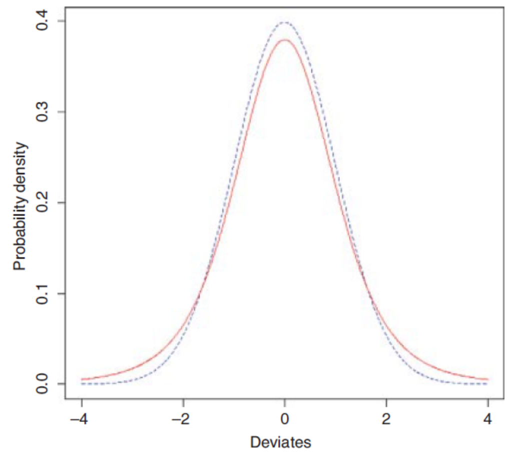
UNC CHARLOTTE 90

90

Key Concept

Student's t vs. Normal Distribution

The difference between the normal (blue dashed line) and Student's *t* distributions (solid red line) is that the *t* distribution has 'fatter tails.' This means that extreme values are more likely with a *t* distribution than with a normal, and the confidence intervals are correspondingly broader.



The Gamma Distribution

The gamma distribution is useful for describing a wide range of processes where the data are positively skew. Insurance claim amount follow this distribution. Its density is given by

$$f(x) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}$$

It has two parameters: α – The shape parameter
$1/\beta$ – The scale parameter

The mean of the distribution is = $\alpha\beta$
The variance of the distribution is = $\alpha\beta^2$
The skewness of the distribution is $2/\sqrt{\alpha}$
The kurtosis of the distribution is $6/\alpha$

Key Concept

The Gamma Distribution

The gamma distribution gives rise to two special distributions for specified value of the parameters.

When $\alpha = 1$, we get the *Exponential Distribution*
When $\alpha = v/2$ and $\beta = 2$, we get the *Chi-Squared Distribution*

Exponential Distribution

The mean of the distribution is $= \beta$
The variance of the distribution is $= \beta^2$
The skewness of the distribution is 2
The kurtosis of the distribution is 6

Chi-Squared Distribution

The mean of the distribution is $= v$
The variance of the distribution is $= 2v$
The skewness of the distribution is $2\sqrt{2/v}$
The kurtosis of the distribution is $12/\alpha$.

Key Concept

Observation of Parameters & Mean & Variance

$$\frac{1}{\beta} = \frac{\text{mean}}{\text{variance}}$$

$$\text{shape} = \frac{1}{\beta} \times \text{mean}$$

Key Concept

The Gamma Density Function in R

```
dgamma(x,
       shape,
       rate = 1,
       scale = 1/rate,
       alpha = shape,
       beta = scale,
       log = FALSE)
```

You can shape and scale in alternate ways.

where,
x – vector of quantiles.
rate – an alternative way to specify the scale.
alpha, beta – an alternative way to specify the shape and scale.
shape, scale – shape and scale parameters.

The following are equivalent:

>dgamma(x, 2, 2)
>dgamma(x, alpha =2, beta = ½)

Example

What density value for x = 1.5 is expected from a gamma distribution with mean = 2 and variance = 3?

alpha =
beta =

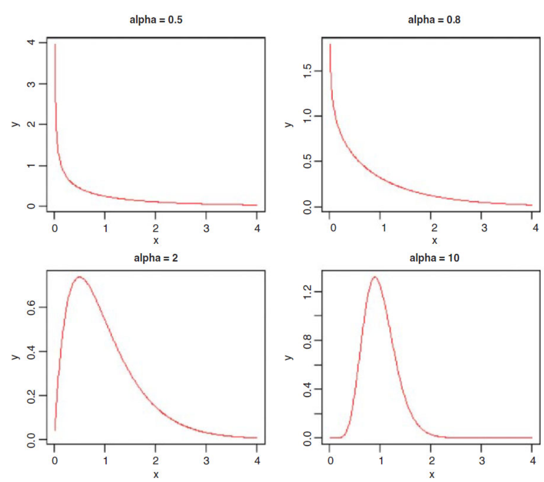
```
#Need to find alpha & beta first.
>pgamma(1.5,alpha = ,beta = )
```

Investigating the Shape of the Gamma Density Function

```
x <- seq(0.01,4,.01)
par(mfrow=c(2,2))

y <- dgamma(x,.5,.5)
plot(x,y,type="l",col="red",main="alpha = 0.5")

y <- dgamma(x,.8,.8)
plot(x,y,type="l",col="red", main="alpha = 0.8")

y <- dgamma(x,2,2)
plot(x,y,type="l",col="red", main="alpha = 2")

y <- dgamma(x,10,10)
plot(x,y,type="l",col="red", main="alpha = 10")
```

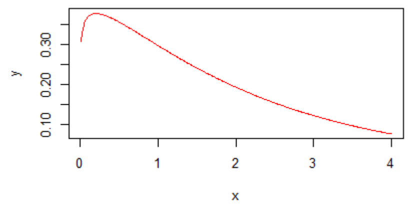

The Results

What are your observations about the graph as alpha increases?



Homework

For what value of α does the graph look like the graph below. Keep the rate parameter = 0.5.



Example

What is the value of the 95% quantile expected from a gamma distribution with mean = 2 and variance = 3?

```
qgamma(0.95,4/3,2/3)
```

Key Concept

The Gamma Cumulative Probability Density Function in R

```
pgamma(q,
       shape,
       rate = 1,
       scale = 1/rate,
       alpha = shape,
       beta = scale,
       lower.tail=TRUE
       log.p = FALSE)
```

where,
- x – vector of quantiles
- rate – an alternative way to specify the scale
- alpha, beta – an alternative way to specify the shape and scale
- shape, scale – shape and scale parameters
- log.p – logical; if TRUE, probabilities p are given as log(p).
- lower.tail –logical; if TRUE (default), probabilities are P[X<=x],=""otherwise,=""p[x="">x]

Example:

```
>alpha = 10
>beta = 15 / 60
>x = 3

># exact
>pgamma(q = x, shape = alpha, scale = beta)
[1] 0.7576078
```


Key Concept

The Gamma Quantile or Inverse Function in R

```
qgamma(p,
       shape,
       rate = 1,
       scale = 1/rate,
       alpha = shape,
       beta = scale,
       lower.tail=TRUE
       log.p = FALSE)
```

where,
- p – probability
- rate – an alternative way to specify the scale
- alpha, beta – an alternative way to specify the shape and scale
- shape, scale – shape and scale parameters
- log.p – logical; if TRUE, probabilities p are given as log(p).
- lower.tail –logical; if TRUE (default), probabilities are P[X<=x],=""otherwise,=""p[x="">x]

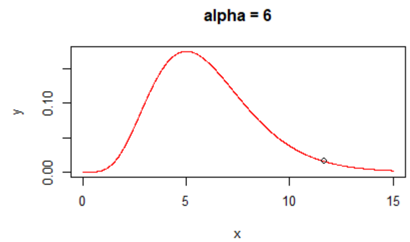
Example:

```
> mu2=qgamma(.975, 6, 1)
> mu2
[1] 11.66833
```

The Gamma Quantile or Inverse Function in R

Example:

```
x <- seq(0.01,15,.01)
par(mfrow=c(2,2))

y <- dgamma(x,6,1)
plot(x,y,type="l",col="red", main="alpha = 6")

mu2=qgamma(.975, 6, 1)
mu2
[1] 11.66833

y <- dgamma(mu2,6,1)
y
[1] 0.01543016
points(mu2, y, color ="blue")
```



The Gamma Random Number Function in R

```
rgamma(n,
       shape,
       rate = 1,
       scale = 1/rate,
       alpha = shape,
       beta = scale,
       lower.tail=TRUE
       log.p = FALSE)
```

where,
- x – vector of quantiles
- n – number of random numbers
- rate – an alternative way to specify the scale
- alpha, beta – an alternative way to specify the shape and scale
- shape, scale – shape and scale parameters
- log.p – logical; if TRUE, probabilities p are given as log(p).
- lower.tail –logical; if TRUE (default), probabilities are P[X<=x],=""otherwise,=""p[x="">x]

Example:

```
#Generate 10 Random Numbers
>alpha = 10
>beta = 15 / 60

>RanNums <-rgamma(n = 10, shape = alpha, scale = beta)
>RanNums
[1] 2.499340 1.218013 1.781416 2.176373 1.324579
[2] 1.944151 3.113932 1.371185 2.107525 1.210983
```

Non-Normally Distributed Data

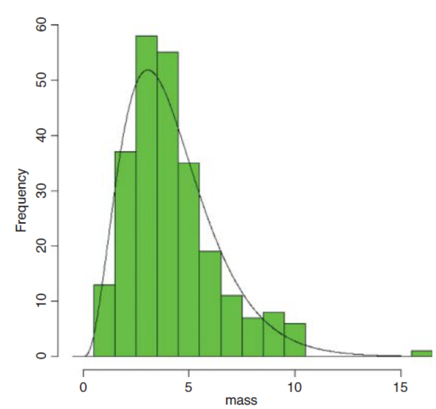
An important use of the gamma distribution is in describing continuous measurement data that are *not* normally distributed.

Here is an example where body mass data for 200 fish are plotted as a histogram and a gamma distribution with the same mean and variance is overlaid as a smooth curve:

```
fishes <- read.table("c:\\temp\\fishes.txt",header=T)
attach(fishes)
names(fishes)
[1] "mass"
```


Non-Normally Distributed Data

```
rate <- mean(mass)/var(mass)
shape <- rate*mean(mass)
rate
[1] 0.8775119

shape
[1] 3.680526

max(mass)
[1] 15.53216

par(mfrow=c(1,1))
hist(mass,breaks=-0.5:16.5,col="green",main="")
lines(seq(0.01,15,0.01),length(mass)*dgamma(seq(0.01,15,0.01),shape,rate))
```



Key Concept

The Exponential Distribution

This is a one-parameter distribution that is a special case of the gamma distribution.

Used in survival analysis. The random number generator of the exponential is useful for Monte Carlo simulations of time to death when the hazard rate (the instantaneous risk of death) is constant with age.

$$f(x;\, \beta) = \frac{e^{-\frac{x}{\beta}}}{\beta}$$

where,

$\beta = Mean\ number\ of\ events\ per\ unit\ time$

$\frac{1}{\beta} = Rate = Waiting\ time\ to\ next\ event$

Example:
Suppose the mean number of customers to arrive at a bank in a 1-hour interval is 10.

Then, the average (waiting) time until the next customer is 1/10 of an hour, or 6 minutes.

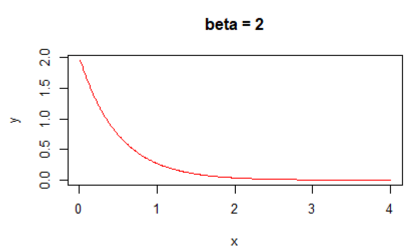
The Exponential and Poisson are related by this parameter.

The Exponential Distribution

This is a one-parameter distribution that is a special case of the gamma distribution.

Used in survival analysis. The random number generator of the exponential is useful for Monte Carlo simulations of time to death when the hazard rate (the instantaneous risk of death) is constant with age.


```
x <- seq(0.01,4,.01)
y <- dexp(x, 2)

plot(x,y,type="l",col="red",main="beta = 2")
```



Key Concept

The Exponential Distribution Functions in R

```
dexp(x, rate = 1, log = FALSE)
pexp(q, rate = 1, lower.tail = TRUE, log.p = FALSE)
qexp(p, rate = 1, lower.tail = TRUE, log.p = FALSE)
rexp(n, rate = 1)
```

Arguments
x, q – vector of quantiles
p – vector of probabilities.
n – number of observations. If length(n) > 1, the length is taken to be the number required.
rate – vector of rates.
log, log.p – logical; if TRUE, probabilities p are given as log(p).
lower.tail – logical; if TRUE (default), probabilities are P[X ≤ x], otherwise, P[X > x].

Key Concept

The Beta Distribution

This has two positive constants, a and b, and x is bounded in the range $0 \le x \le 1$:

$$f(x)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1}$$

One of its most common uses is to model one's uncertainty about the probability of success of an experiment.

- The time it takes to complete a task
- The proportion of defective items in a shipment

The Beta Distribution

Example:
Suppose that DVDs in a certain shipment are defective with a Beta distribution with α = 2 and β = 5.

Compute the probability that the shipment has 20% to 30% defective DVDs.

$$P(0.2 \le X \le 0.3) = \sum_{x=0.2}^{0.3} \frac{x^{2-1}(1-x)^{5-1}}{\mathrm{B}(2,5)} = 0.235185$$

```
> pbeta(.3,2,5) - pbeta(.2,2,5)
[1] 0.235185
```


The Beta Distribution

Generating a family of density functions

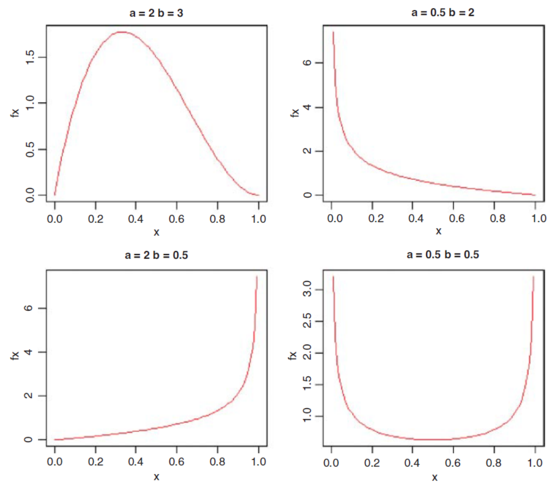
```
par(mfrow=c(2,2))
x <- seq(0,1,0.01)
fx <- dbeta(x,2,3)
plot(x,fx,type="l",main="a=2 b=3",col="red")
fx <- dbeta(x,0.5,2)
plot(x,fx,type="l",main="a=0.5 b=2",col="red")
fx <- dbeta(x,2,0.5)
plot(x,fx,type="l",main="a=2 b=0.5",col="red")
fx <- dbeta(x,0.5,0.5)
plot(x,fx,type="l",main="a=0.5 b=0.5",col="red")
```

Topic 2: Mathematical Distributions

Key Concept

The Beta Distribution – Observations

- When both are greater than 1, we get an *n*-shaped curve which becomes more skew as $b > a$ (top left).
- If $0 < a < 1$ and $b > 1$ then the slope of the density is negative (top right)
- If $a > 1$ and $0 < b < 1$ the slope of the density is positive (bottom left).
- The function is U-shaped when both a and b are positive fractions.
- If $a = b = 1$, then we obtain the uniform distribution on [0,1].



UNC CHARLOTTE 113

113

Topic 2: Mathematical Distributions

Key Concept

The Beta Distribution Functions in R

```
dbeta(x, shape1, shape2, ncp = 0, log = FALSE)
pbeta(q, shape1, shape2, ncp = 0, lower.tail = TRUE, log.p = FALSE)
qbeta(p, shape1, shape2, ncp = 0, lower.tail = TRUE, log.p = FALSE)
rbeta(n, shape1, shape2, ncp = 0)
```

Arguments
x, q – vector of quantiles
p – vector of probabilities.
n – number of observations. If length(n) > 1, the length is taken to be the number required
shape1, shape2 – non-negative parameters of the Beta distribution
ncp – non-centrality parameter
log, log.p – logical; if TRUE, probabilities p are given as log(p)
lower.tail – logical; if TRUE (default), probabilities are P[X ≤ x], otherwise, P[X > x]

UNC CHARLOTTE 114

114

The Beta Distribution – Random Numbers

Here are 10 random numbers from the beta distribution with shape parameters 2 and 3:

```
rbeta(10,2,3)

[1] 0.2908066 0.1115131 0.5217944 0.1691430 0.4456099
[6] 0.3917639 0.6534021 0.3633334 0.2342860 0.6927753
```


Key Concept

The Lognormal Distribution

The lognormal distribution takes values on the positive real line. If the logarithm of a lognormal deviate is taken, the result is a normal deviate, hence the name.

Applications for the lognormal include the distribution of particle sizes in aggregates, flood flows, concentrations of air contaminants, failure times, and insurance claim sizes.

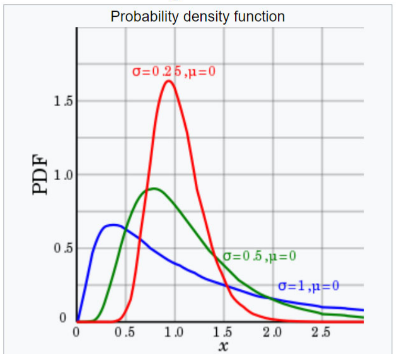
The hazard function of the lognormal is increasing for small values and then decreasing. A mixture of heterogeneous items that individually have monotone hazards can create such a hazard function.

$$\begin{aligned}
f_X(x) &= \frac{\mathrm{d}}{\mathrm{d}x}\Pr(X \le x) = \frac{\mathrm{d}}{\mathrm{d}x}\Pr(\ln X \le \ln x) \\
&= \frac{\mathrm{d}}{\mathrm{d}x}\Phi\left(\frac{\ln x - \mu}{\sigma}\right) \\
&= \varphi\left(\frac{\ln x - \mu}{\sigma}\right)\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\ln x - \mu}{\sigma}\right) \\
&= \varphi\left(\frac{\ln x - \mu}{\sigma}\right)\frac{1}{\sigma x} \\
&= \frac{1}{x}\cdot\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right)
\end{aligned}$$

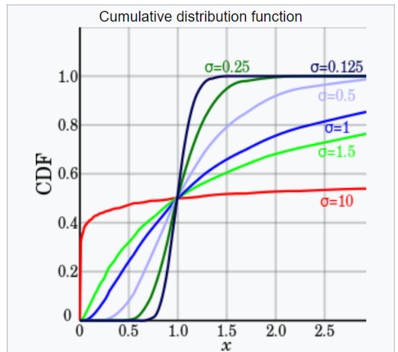
Key Concept

The Lognormal Distribution

Log-normal



Some log-normal density functions with identical parameter μ but differing parameters σ



Cumulative distribution function of the log-normal distribution (with $\mu = 0$)

Topic 2: Mathematical Distributions

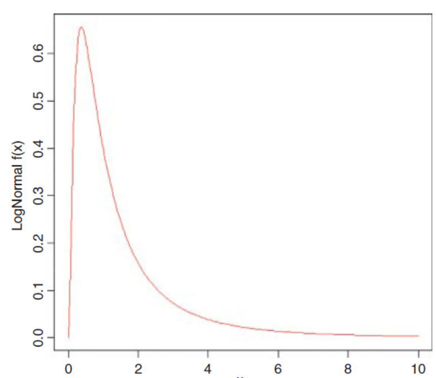
Key Concept

The Lognormal Distribution

```
windows(7,7)

plot(seq(0,10,0.05),dlnorm(seq(0,10,0.05)),
type="l",xlab="x",ylab="LogNormal
f(x)",col="x")
```

The extremely long tail and exaggerated positive skew are characteristic of the lognormal distribution.

Logarithmic transformation followed by analysis with normal errors is often appropriate for data such as these.



Key Concept

Special Properties of the Lognormal Distribution

The most important relationship between the Normal and Lognormal distributions:

"If X follows a lognormal distribution, then Log(X) follows a normal distribution."

This important link allows us to apply linear modeling methods to non-linear problems!

Mean Variance of Lognormal

$$m = \exp\left(\mu + \frac{\sigma^2}{2}\right)$$

$$v = \left(\exp(\sigma^2) - 1\right)\exp(2\mu + \sigma^2)$$

Mean Variance of Normal

$$\mu = \ln\left(\frac{m}{\sqrt{1 + \frac{v}{m^2}}}\right)$$

$$\sigma^2 = \ln\left(1 + \frac{v}{m^2}\right)$$

Key Concept

The Lognormal Distribution Functions in R

```
dlnorm(x, meanlog = 0, sdlog = 1, log = FALSE)
plnorm(q, meanlog = 0, sdlog = 1, lower.tail = TRUE, log.p = FALSE)
qlnorm(p, meanlog = 0, sdlog = 1, lower.tail = TRUE, log.p = FALSE)
rlnorm(n, meanlog = 0, sdlog = 1)
```

Arguments
x, q – vector of quantiles
p – vector of probabilities.
n – number of observations. If length(n) > 1, the length is taken to be the number required
meanlog, sdlog – mean and standard deviation of the distribution on the log scale with default values of 0 and 1 respectively
log, log.p – logical; if TRUE, probabilities p are given as log(p)
lower.tail – logical; if TRUE (default), probabilities are P[X ≤ x], otherwise, P[X > x]

Key Concept

The Logistic Distribution

The logistic is the standard link function in generalized linear models with binomial errors. We will delve into this distribution more latter.

PDF

$$f(x;\mu,s) = \frac{e^{-\frac{x-\mu}{s}}}{s\left(1+e^{-\frac{x-\mu}{s}}\right)^2}$$

CDF

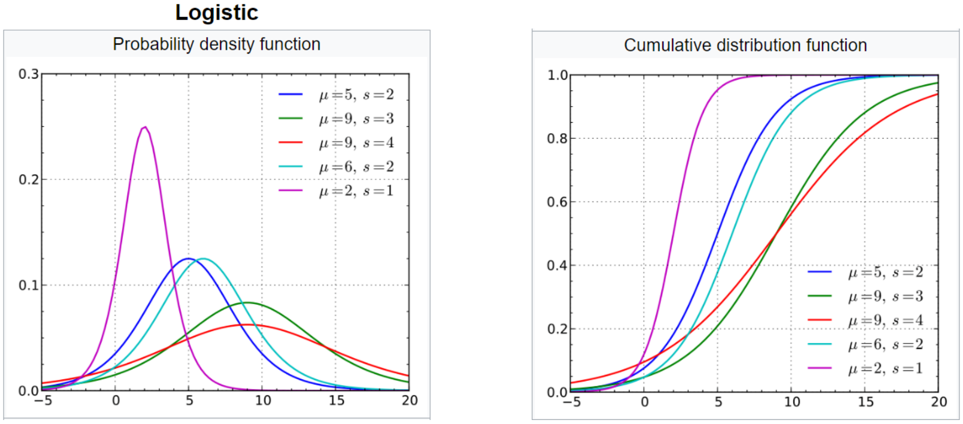
$$F(x;\mu,s) = \frac{1}{1+e^{-\frac{x-\mu}{s}}}$$

Parameters	μ, location (real) $s > 0$, scale (real)

Key Concept

The Logistic Distribution

Logistic



Key Concept

The Logistic Distribution Functions in R

logistic(x,d=0, a=1,c=0, z=1) #The density function
logit(p) #The inverse of the logistic function

Arguments
x – Any integer or real value
d – Item difficulty or delta parameter

a – The slope of the curve at x=0 is equivalent to the discrimination parameter in 2PL models or alpha parameter. Is either 1 in 1PL or 1.702 in 1PN approximations.

c – Lower asymptote = guessing parameter in 3PL models or gamma
z – The upper asymptote --- in 4PL models
p – Probability to be converted to logit value

Key Concept

The Logistic Distribution Functions in R

dlogis(x, location = 0, scale = 1, log = FALSE)
plogis(q, location = 0, scale = 1, lower.tail = TRUE, log.p = FALSE)
qlogis(p, location = 0, scale = 1, lower.tail = TRUE, log.p = FALSE)
rlogis(n, location = 0, scale = 1)

Arguments
x, q – vector of quantiles
p – vector of probabilities.
n – number of observations. If length(n) > 1, the length is taken to be the number required
location, scale – location and scale parameters
log, log.p – logical; if TRUE, probabilities p are given as log(p)
lower.tail – logical; if TRUE (default), probabilities are P[X ≤ x], otherwise, P[X > x]

The Logistic Distribution

The logistic is a unimodal, symmetric distribution on the real line with tails that are longer than the normal distribution.

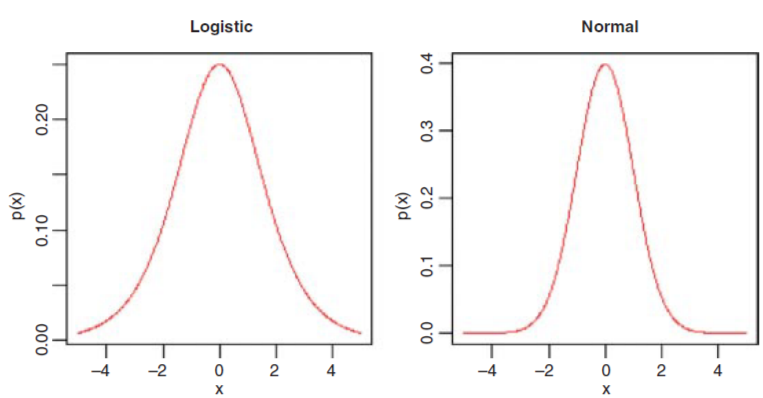
```
windows(7,4)
par(mfrow=c(1,2))

plot(seq(-5,5,0.02),dlogis(seq(-5,5,.02)),
type="l",main="Logistic",col="red",xlab="x",ylab="p(x)")

plot(seq(-5,5,0.02),dnorm(seq(-5,5,.02)),
type="l",main="Normal",col="red",xlab="x",ylab="p(x)")
```


The Logistic Distribution

Here, the logistic density function dlogis (left) is compared with an equivalent normal density function dnorm (right) using the default mean 0 and standard deviation 1 in both cases.

Note the much fatter tails of the logistic (there is still substantial probability at +/– 4 standard deviations).

Note also the difference in the scales of the two y axes (0.25 for the logistic, 0.4 for the normal).



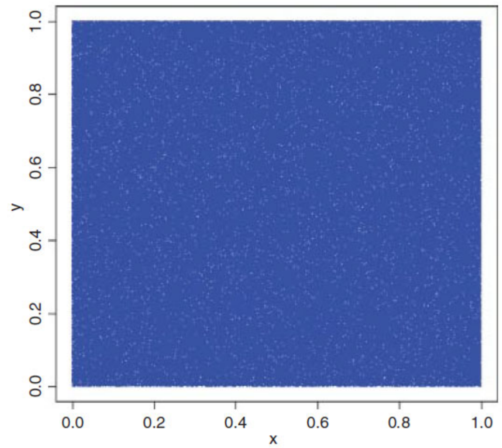
Key Concept

The Uniform Distribution

R's random-number generator.

```
x <- runif(1000000)
y <- runif(1000000)
plot(x,y,pch=".",col="blue")
```

The scatter of unfilled space (white dots amongst the sea produced by 1,000,000 blue dots `pch="."`) shows no evidence of clustering.



Key Concept

The Uniform Distribution

For a more thorough check we can count the frequency of combinations of numbers: with 36 cells, the expected frequency is 1 000 000/36 = 27 777.78 numbers per cell. We use the `cut` function to produce 36 bins:

```
table(cut(x,6),cut(y,6))
```

	(-0.001,0.166]	(0.166,0.333]	(0.333,0.5]	(0.5,0.667]	(0.667,0.834]	(0.834,1]
(-0.000997,0.166]	27667	28224	27814	27601	27592	27659
(0.166,0.333]	27604	27790	27922	27687	27990	27701
(0.333,0.5]	27951	27668	27683	27773	27999	27959
(0.5,0.667]	27550	27767	27951	27912	27619	27577
(0.667,0.834]	27527	28106	27868	28262	27804	27460
(0.834,1]	27617	27662	27863	27867	27727	27577

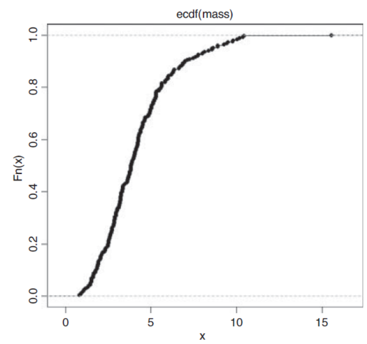
```
range(table(cut(x,6),cut(y,6)))
[1] 27460 28262
```

Not Bad!

Plotting Empirical Cumulative Distribution Functions

The function **ecdf** is used to compute or plot an empirical cumulative distribution function. Here it is in action for the fishes data:

```
fishes <- read.table("c:\\temp\\fishes.txt",header=T)
attach(fishes)
names(fishes)
[1] "mass"
plot(ecdf(mass))
```



The pronounced positive skew in the data is evident from the fact that the left-hand side of the cumulative distribution is much steeper than the right-hand side

Discrete Probability Distributions

1. The Bernoulli Distribution
2. The Binomial Distribution
3. The Geometric Distribution
4. The Hypergeometric Distribution
5. The Multinomial Distribution
6. The Poisson Distribution
7. The Negative Binomial Distribution

Key Concept

Bernoulli Distribution

This is the distribution underlying tests with a binary response variable. The response takes one of only two values:

1 with probability p (a 'success')
0 with probability $1 - p$ (a 'failure')

The density function is given by:

$$p(X) = p^x(1-p)^{1-x}$$

Flipping a coin follows a Bernoulli Distribution!

Key Concept

Statistical Mean & Variance Definitions

The theoretical definitions of the mean and variance of a distribution are given by :

$$\mu = E[X]$$

$$\sigma^2 = E[X^2] - (E[X])^2$$

Note: $E[X^n] = \sum_{i=1}^{n} x^n \cdot \Pr(X = x)$

Key Concept

Bernoulli Distribution: Mean & Variance

$$\mathrm{E}(X) = \sum x f(x) = 0 \times (1-p) + 1 \times p = 0 + p = p$$

$$\mathrm{E}(X^2) = \sum x^2 f(x) = 0^2 \times (1-p) + 1^2 \times p = 0 + p = p$$

$$\mathrm{var}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2 = p - p^2 = p(1-p) = pq$$

Therefore, the mean and variance for the Bernoulli distribution are:

$\mu = p$

$\sigma^2 = p(1-p) = pq$

Key Concept

Binomial Distribution

The Binomial distributions is modelled after the Bernoulli for multiple events where each event has one of two outcomes.

Bernoulli Distribution

$$p(X) = p^x(1-p)^{1-x}$$

Binomial Distribution

$$p(x) = \binom{n}{x} p^x(1-p)^{n-x}$$

The mean of the binomial distribution is np and the variance is $np(1-p)$.

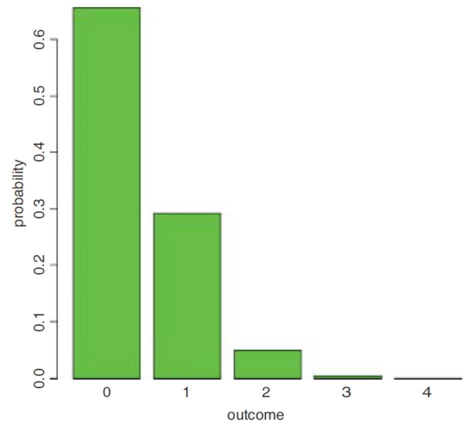
Notice the only difference is the combinatoric factor for n events versus 1 event.

The Bernoulli could also be written as: $p(x) = \binom{1}{x} p^x(1-p)^{1-x}$

Binomial Distribution

Example:

```
p <- 0.1
n <- 4
x <- 0:n

px <- choose(n,x)*p**x*(1-p)**(n-x)
barplot(px,names=x,xlab="outcome",y
lab="probability",col="green")
```



Key Concept

Binomial Distribution Functions in R

dbinom(x, size, prob, log = FALSE)
pbinom(q, size, prob, lower.tail = TRUE, log.p = FALSE)
qbinom(p, size, prob, lower.tail = TRUE, log.p = FALSE)
rbinom(n, size, prob)

Arguments
x, q – vector of quantiles
p – vector of probabilities.
n – number of observations. If length(n) > 1, the length is taken to be the number required
size – number of trials (zero or more)
prob – probability of success on each trial
log, log.p – logical; if TRUE, probabilities p are given as log(p)
lower.tail – logical; if TRUE (default), probabilities are P[X ≤ x], otherwise, P[X > x]

Topic 2: Mathematical Distributions

Key Concept

Binomial Distribution

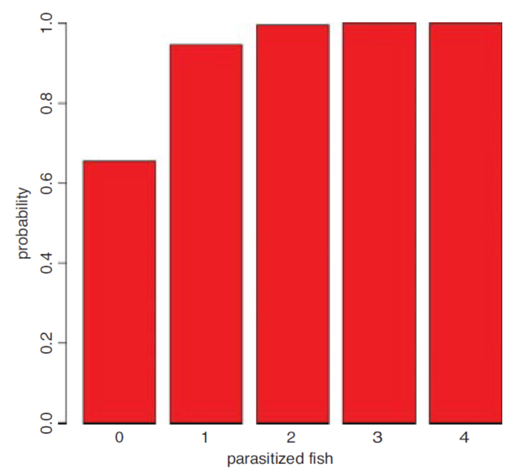
Cumulative Distribution Example:

```
barplot(pbinom(0:4,4,0.1),names=0:4,
xlab="parasitized fish",
ylab="probability",col="red")
```

Notice how the cumulative probability distribution was achieved by the first argument.

```
pbinom(0:4,4,0.1)
```

"**p**" is a cumulative function and argument tells over what numbers to perform the accumulation



UNC CHARLOTTE 137

137

Topic 2: Mathematical Distributions

Key Concept

Binomial Distribution

The 95% Confidence Interval is given by:

```
qbinom(.025,4,0.1)
[1] 0
qbinom(.975,4,0.1)
[1] 2
```

This means that with 95% certainty we shall catch between 0 and 2 parasitized fish out of 4 if we repeat the sampling exercise.

We are very unlikely to get 3 or more parasitized fish out of a sample of 4 if the proportion parasitized really is 0.1.

UNC CHARLOTTE 138

138

Key Concept

HW: Binomial Distribution

Rerun the code below for each p in the sequence seq(0.2, 1.0, 0.1) and describe how the shape of the two barplots change. What do you think this means?

```
p <- 0.1
n <- 4
x <- 0:n

px <- choose(n,x)*p**x*(1-p)**(n-x)
barplot(px,names=x,xlab="outcome",ylab="probability",col="green")

barplot(pbinom(0:4,4,p),names=0:4,
xlab="parasitized fish",
ylab="probability",col="red")
```


Key Concept

Binomial Distribution – Sample Size Determination

It is important to know the likelihood that no sample has a success, when the probability of success is p.

In our example, the probability that a fish has a parasite is 0.1. This means the probability a fish does not have a parasite is 0.9.

With our sample size of $n = 4$,we have a probability of missing the parasite of $0.9^4 = 0.6561$. This mean there is a 65.61% chance of not finding the parasite.

That is too high and means we should rethink our sample size.

Binomial Distribution – Sample Size Determination

Let's say we want this probability to be 0.05 or less. What is the minimum sample size we need?

We need to solve:

$$0.05 = (0.9)^n$$

Taking logs,

$$\log(0.05) = n \log(0.9),$$

so

$$n = \frac{\log(0.05)}{\log(0.9)} = 28.433\,16$$

Key Concept

Binomial Distribution – Sample Size Determination

Random numbers are generated from the binomial distribution like this

```
rbinom(10,4,0.1)
[1]  0  0  0  0  0  1  0  1  0  1
```

Here we repeated the sampling of 4 fish ten times. We got 1 parasitized fish out of 4 on three occasions, and 0 parasitized fish on the remaining seven occasions. We never caught 2 or more parasitized fish in any of these samples of 4.

Key Concept

Geometric Distribution

Suppose that a series of independent Bernoulli trials with probability p are carried out at times 1, 2, 3, ...

Now let W be the waiting time until the first success occurs. So

$$P(W > x) = (1 - p)^x$$

which means that

$$P(W = x) = P(W > x - 1) - P(W > x)$$

Topic 2: Mathematical Distributions

Key Concept

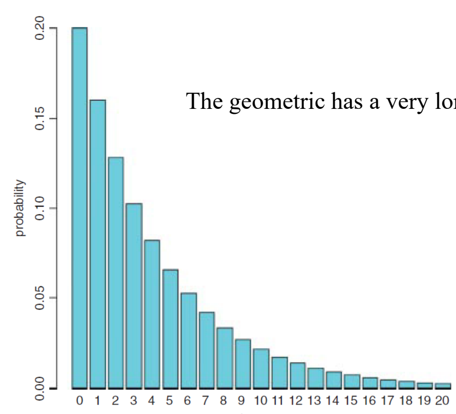
Geometric Distribution

The density function is $f(x) = p(1-p)^{x-1}$

```
fx <- dgeom(0:20,0.2)

barplot(fx,names=0:20,xlab="outcome",
ylab="probability",col="cyan")
```

the mean is $\frac{1-p}{p}$

the variance is $\frac{1-p}{p^2}$



The geometric has a very long tail.

Geometric Distribution

Here are 100 random numbers from a geometric distribution with $p = 0.1$. The modes are 0 and 1, but outlying values as large as 33 and 44 have been generated:

```
table(rgeom(100,0.1))

 0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 17 18 21 22 24 28 29 31 33 44
14 14  8  5  1 13  3  5  3  5  2  5  3  2  3  1  1  2  1  1  2  2  1  1  1  1
```


Key Concept

Hypergeometric Distribution

'Balls in urns' are the classic sort of problem solved by this distribution. The density function of the hypergeometric is

$$f(x) = \frac{\binom{b}{x}\binom{N-b}{n-x}}{\binom{N}{n}}$$

Suppose that there are N coloured balls in the statistician's famous urn: b of them are blue and $r = N - b$ of them are red.

Now a sample of n balls is removed from the urn; this is sampling *without replacement*.

Now $f(x)$ gives the probability that x of these n balls are blue.

Key Concept

Hypergeometric Distribution Functions in R

```
dhyper(x, m, n, k, log = FALSE)
phyper(q, m, n, k, lower.tail = TRUE, log.p = FALSE)
qhyper(p, m, n, k, lower.tail = TRUE, log.p = FALSE)
rhyper(nn, m, n, k)
```

The order matters

Arguments

x, q – vector of quantiles representing the number of white balls drawn without replacement from an urn which contains both black and white balls.
m – the number of white balls in the urn.
n – the number of black balls in the urn.
k – the number of balls drawn from the urn.
p – probability, it must be between 0 and 1.
nn – number of observations. If length(nn) > 1, the length is taken to be the number required.
log, log.p – logical; if TRUE, probabilities p are given as log(p).
lower.tail – logical; if TRUE (default), probabilities are P[X≤x], otherwise, P[X>x].

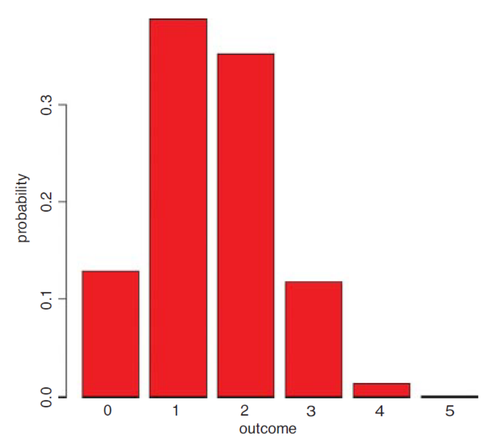
Key Concept

Hypergeometric Distribution Functions in R

Example:

Let the urn contain $N = 20$ balls, of which 6 are blue and 14 are red. We take a sample of $n = 5$ balls so x could be 0, 1, 2, 3, 4 or 5 of them blue.

```
ph <- dhyper(0:5,6,14,5)

barplot(ph,names=(0:5),col="red",
xlab="outcome",ylab="probability")
```



We are very unlikely to get more than 3 red balls out of 5. The most likely outcome is that we get 1 or 2 red balls out of 5.

Homework

An urn contains 4 red balls and 10 blue balls. Five balls are drawn at random without replacement from this urn.

1. What is the probability that exactly two red balls are drawn?
2. What is the probability that exactly three red balls are drawn?
3. What is the probability that at least two red balls are drawn?
4. What is the probability that zero red balls are drawn?

Key Concept

The Multinomial Distribution

Suppose that there are t possible outcomes from an experimental trial, and the outcome i has probability p_i.

Now allow n independent trials where $n = n_1 + n_2 + \ldots + n_t$ and ask what is the probability of obtaining the vector of N_i occurrences of the i^{th} outcome:

$$P(N_i = n_i) = \frac{n!}{n_1!n_2!n_3!\ldots n_t!} p_1^{n_1} p_2^{n_2} p_3^{n_3} \ldots p_t^{n_t}$$

The Multinomial Distribution Function in R

dmultinom(x, size, prob, log = FALSE)
pmultinom(lower = -Inf, upper = Inf, size, probs, method)
qmultinom(p, size = NULL, prob, lower.tail = TRUE, log.p = FALSE)
rmultinom(n, size, prob)

Arguments
x – k-column matrix of quantiles.
n – number of observations. If length(n) > 1, the length is taken to be the number required
P – vector of probabilities
lower – vector
upper – vector
size – numeric vector; number of trials (zero or more).
prob – k-column numeric matrix; probability of success on each trial.
Log – logical; if TRUE, probabilities p are given as log(p).

The Multinomial Distribution Function in R

Example: dmultinom(x, size, prob, log = FALSE)

```
># Compute single pdf values:
>p1<- 0.2; p2<- 0.3; p3<- 0.5;   # 3 possible outcomes
>dmultinom(c(5,5,5), prob=c(p1,p2,p3))  # prob. of 5 each
>dmultinom(c(0,0,9), prob=c(p1,p2,p3))  # prob. of 9 3's
>dmultinom(c(0,2,7), prob=c(p1,p2,p3))  # prob. of 2 2's and 7 3's.
```

```
> # Compute single pdf values:
> p1<- 0.2; p2<- 0.3; p3<- 0.5;    # 3 possible outcomes
> dmultinom(c(5,5,5), prob=c(p1,p2,p3))  # prob. of 5 each
[1] 0.01838917
> dmultinom(c(0,0,9), prob=c(p1,p2,p3))  # prob. of 9 3's
[1] 0.001953125
> dmultinom(c(0,2,7), prob=c(p1,p2,p3))  # prob. of 2 2's and 7 3's.
[1] 0.0253125
```

Key Concept

The Multinomial Distribution Function in R

Example: pmultinom(lower = -Inf, upper = Inf, size, probs, method)

>Install.packages(“pmultinom”) Note: You must install the package “pmultinom” to use pmultinom()
>Library(pmultinom)
>pmultinom(upper=c(5,5,5), size=10, probs=c(1/2, 1/4, 1/4), method="exact“0)

[1] 0.5835915

Key Concept

The Multinomial Distribution Function in R

qmultinom(p, size = NULL, prob, lower.tail = TRUE, log.p = FALSE)

Note: This function has a format, but I have not been able to find the package that contains it.

The Multinomial Distribution Function in R

Example: rmultinom(n, size, prob)

```
>p1<- 0.2; p2<- 0.3; p3<- 0.5;    # 3 possible outcomes
>rmultinom(1, size=9, prob=c(p1,p2,p3))  # 1 run of 9 trials
>rmultinom(5, size=9, prob=c(p1,p2,p3))  # 5 runs of 9 trials
```

```
> rmultinom(1, size=9, prob=c(p1,p2,p3))  # 1 run of 9 trials
     [,1]
[1,]    1
[2,]    3
[3,]    5
> rmultinom(5, size=9, prob=c(p1,p2,p3))  # 5 runs of 9 trials
     [,1] [,2] [,3] [,4] [,5]
[1,]    2    1    4    2    1
[2,]    4    5    2    1    2
[3,]    3    3    3    6    6
```


The Multinomial Distribution Function in R

Examples: An experiment of drawing a random card from an ordinary playing cards deck is done with replacing it back. This was done ten times. Find the probability of getting 2 spades, 3 diamond, 3 club and 2 hearts.

Solution:

- There are n=10 trials
- The probability of drawing a spade, diamond, club or heart is 13/52 = 0.25
- This means $p_1 = p_2 = p_3 = p_4 = 0.25$
- We have $n_1 = 2$, $n_2 = 3$, $n_3 = 3$, $n_4 = 2$

```
>dmultinom(c(2, 3, 3, 2), 10, c(0.25, 0.25, 0.25, 0.25))
[1] 0.02403259
```

Homework

1. Of the 10 widgets produces in a factory, what is the probability that 5 are excellent, 2 are good and 2 are fair and 1 is poor? Assume that the classification of individual bits are independent events and that the probabilities of A, B, C and D are 40%, 20%, 5% and 1% respectively.
2. Suppose we have an urn containing 9 marbles. Two are red, three are green, and four are blue. We randomly select 5 marbles from the urn, with replacement. What is the probability of selecting 3 green marbles, 1 red marble, and 1 blue marbles?

Key Concept

Poisson Distribution

This is one of the most useful and important of the discrete probability distributions for describing count data.

The Poisson is a one-parameter distribution with the interesting property that its variance is equal to its mean.

The density function of the Poisson shows the probability of obtaining a count of x when the mean count per unit is λ:

$$p(x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

Poisson Distribution – Recursive Probabilities

For $x = 0$: $p(0) = e^{-\lambda}$

For $x = 1$: $p(1) = p(0)\lambda = \lambda e^{-\lambda}$

$$\boxed{p(x) = p(x-1)\frac{\lambda}{x}}$$

Key Concept

Poisson Distribution Functions in R

```
dpois(x, lambda, log = FALSE)
ppois(q, lambda, lower.tail = TRUE, log.p = FALSE)
qpois(p, lambda, lower.tail = TRUE, log.p = FALSE)
rpois(n, lambda)
```

Arguments
x – vector of (non-negative) quantiles
q – vector of quantiles
p – vector of probabilities.
n – number of random values to return
lambda – vector of non-negative means
log, log.p – logical; if TRUE, probabilities p are given as log(p)
lower.tail – logical; if TRUE (default), probabilities are P[X ≤ x], otherwise, P[X > x]

Poisson Distribution Functions in R

If we wanted 600 simulated counts from a Poisson distribution with a mean of, say, 0.90 blood cells per slide, we just type:

```
count <- rpois(600,0.9)
```

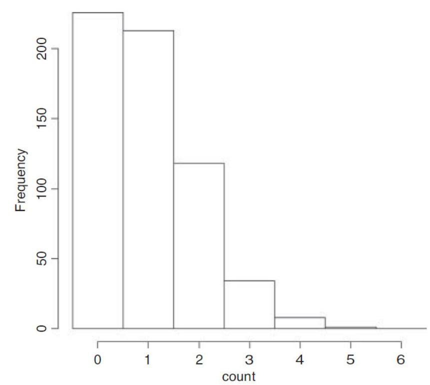
We can use table to see the frequencies of each count generated:

```
table(count)
```

```
count
  0    1    2   3  4  5
244  212  104  33  6  1
```

```
hist(count,breaks = - 0.5:6.5,main="")
```

Note the use of the vector of break points on integer increments from –0.5 to create integer bins for the histogram bars.



Poisson distribution example

The average number of homes sold by the Acme Realty company is 2 homes per day. What is the probability that exactly 3 homes will be sold tomorrow?

Solution: This is a Poisson experiment in which we know the following:

- λ = 2; since 2 homes are sold per day, on average.
- x = 3; since we want to find the likelihood that 3 homes will be sold tomorrow.
- e = 2.71828; since *e* is a constant equal to approximately 2.71828.

We plug these values into the Poisson formula as follows:

$$P(x; \lambda) = (e^{-\lambda}) (\lambda^{x}) / x!$$
$$P(3; 2) = (2.71828^{-2}) (2^{3}) / 3!$$
$$P(3; 2) = (0.13534) (8) / 6$$
$$\mathbf{P(3; 2) = 0.180}$$

Solution using R:

```
dpois(3, 2, log = FALSE)
[1] 0.180447
```

Thus, the probability of selling 3 homes tomorrow is 0.180.

Poisson distribution example

Suppose the average number of lions seen on a 1-day safari is 5. What is the probability that tourists will see fewer than four lions on the next 1-day safari?

Solution: This is a Poisson experiment in which we know the following:

- λ = 5; since 5 lions are seen per safari, on average.
- x = 0, 1, 2, or 3; since we want to find the likelihood that tourists will see fewer than 4 lions; that is, we want the probability that they will see 0, 1, 2, or 3 lions.
- e = 2.71828; since *e* is a constant equal to approximately 2.71828.

Poisson distribution example

We need to calculate the sum of four probabilities: P(0; 5) + P(1; 5) + P(2; 5) + P(3; 5).
To compute this sum, we use the Poisson formula:

$P(x \leq 3, 5) = P(0; 5) + P(1; 5) + P(2; 5) + P(3; 5)$
$P(x \leq 3, 5) = P(0; 5) + P(0; 5)*5/1 + P(0; 5)*5/1*5/2 + P(0; 5)*5/1*5/2 *5/3$ **#Using Recursive Formula!**
$P(x \leq 3, 5) = P(0; 5) + P(0; 5)*5/1 + P(0; 5)*5^2/2! + P(0; 5)*5^2/3!$
$P(x \leq 3, 5) = [(e^{-5})(5^0) / 0!] + [(e^{-5})(5^1) / 1!] + [(e^{-5})(5^2) / 2!] + [(e^{-5})(5^3) / 3!]$
$P(x \leq 3, 5) = [(e^{-5})] + [(e^{-5})(5^1) / 1!] + [(e^{-5})(5^2) / 2!] + [(e^{-5})(5^3) / 3!]$

$P(x \leq 3, 5) = [(0.006738)(1) / 1] + [(0.006738)(5) / 1] + [(0.006738)(25) / 2] + [(0.006738)(125) / 6]$
$P(x \leq 3, 5) = [0.0067] + [0.03369] + [0.084224] + [0.140375]$

$P(x \leq 3, 5) = 0.2650$

Thus, the probability of seeing at no more than 3 lions is 0.2650.

Solution Using R:

```
ppois(3, 5, log = FALSE)
[1] 0.2650259
```

Fitting a Poisson distribution

Consider the two sequences of birth times we saw at the beginning. Both of these examples consisted of a total of 44 births in 24-hour intervals. Therefore the mean birth rate for both sequences is 44/24 = 1.8333

What would be the expected counts if birth times were really random i.e. what is the expected histogram for a Poisson random variable with mean rate $\lambda = 1.8333$?

Using the Poisson formula we can calculate the probabilities of obtaining each possible value.

In practice we group values with low probability into one category.

x	0	1	2	3	4	5	≥ 6
$P(X = x)$	0.159	0.293	0.268	0.164	0.075	0.027	0.011

Then if we observe 24-hour intervals we can calculate the expected frequencies as $24 \times P(X = x)$ for each value of x.

x	0	1	2	3	4	5	≥ 6
Expected freq.	3.837	7.035	6.448	3.941	1.806	0.662	0.271

Homework

1. Suppose there is a disease, whose average incidence is 2 per million people. What is the probability that a city of 1 million people has at least twice the average incidence?

2. Suppose we know that births in a hospital occur randomly at an average rate of 1.8 births per hour.

3. Now suppose we know that in hospital A births occur randomly at an average rate of 2.3 births per hour and in hospital B births occur randomly at an average rate of 3.1 births per hour. What is the probability that we observe 7 births in total from the two hospitals in a given 1-hour period?

4. Suppose disease A occurs with incidence 1.7 per million, disease B occurs with incidence 2.9 per million. Statistics are compiled, in which these diseases are not distinguished, but simply are all called cases of disease "AB". What is the probability that a city of 1 million people has at least 6 cases of AB?