Managing_Describing_Analyzing_data_Week3-Intro to probability

Me

2023-09-18

Probability is the chance that an event will or will not occur

Sample Space Is the set of all possible outcomes from an experiment

Marginal or unconditional probability

Probability of event A occuring, only 1 event can occur.

Say rolling dice. Or you have a production lot of 100 parts, and can draw 1. What is the probability of draawing a defective?

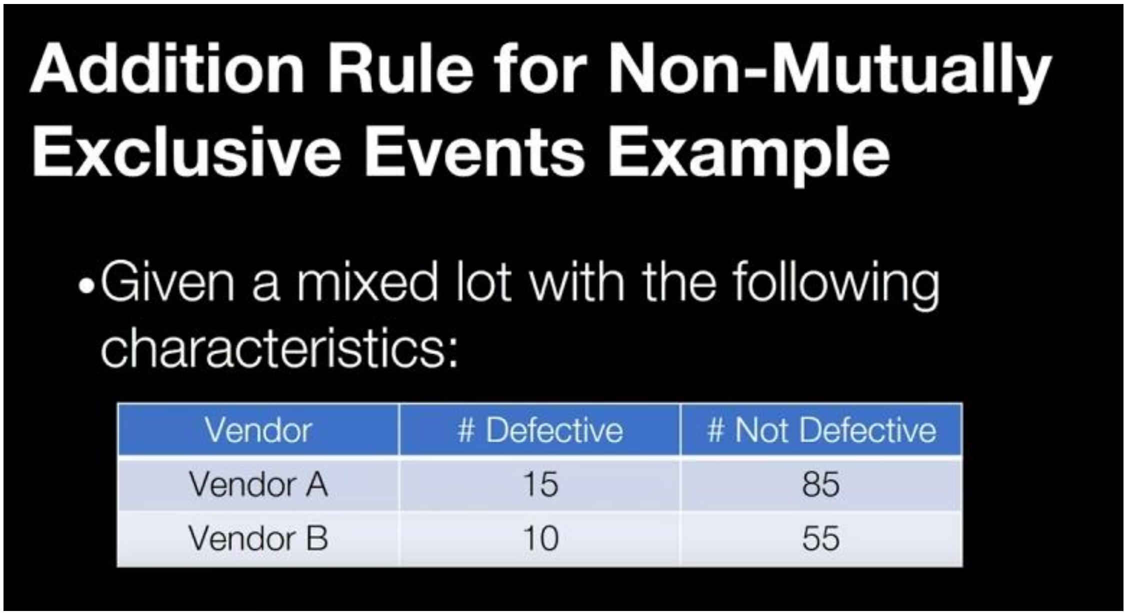
1/P

Addition rule for mutually exclusive events

P(A or B)= P(A)+ P(B)

Addition rule for non mutually exclusive events

P(A or B) = P(A)+ P(B)- P(A+B)



Addition Rule for Non-Mutually Exclusive Events Example

- Given a mixed lot with the following characteristics:

Vendor	# Defective	# Not Defective
Vendor A	15	85
Vendor B	10	55

What is the probability of selecting specific vendors, say vendor 1.

$$\frac{100}{165}+\frac{25}{165}=\frac{125}{165}=0.7575$$

But there are now too many parts

so you have to subtract out the union.

P(A or B) = P(A)+ P(B)- P(A+B)

$$\frac{100}{165}+\frac{25}{165}-\frac{15}{165}=\frac{110}{165}=0.666$$

Joint Probability example 1-independent

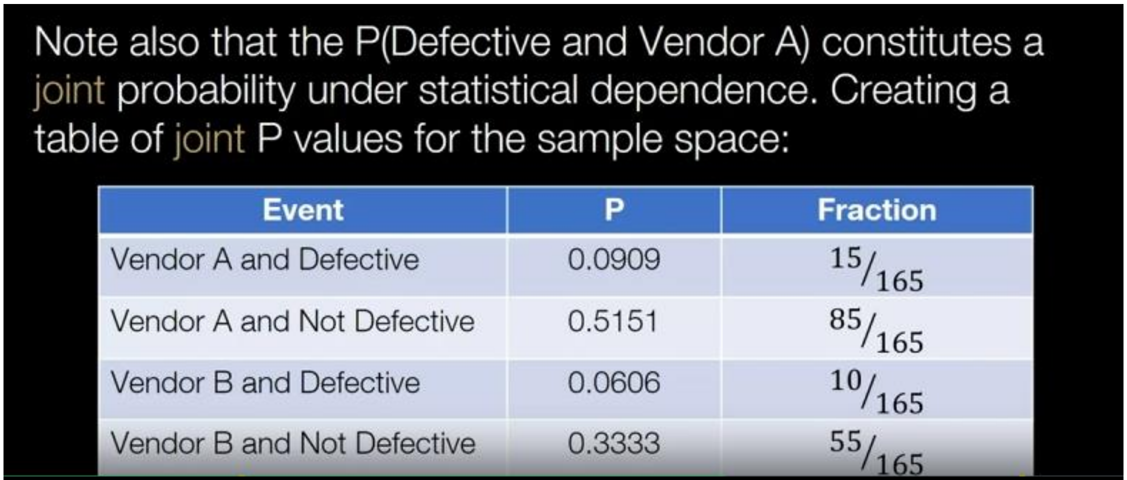
The probability of a machine operator producing a defective part at any point in time is 0.05. What is the probability that three bad parts will be produced in succession or three defective parts?

P(AXBXC) = P(A) X P(B) X P(C)

If a and B are **independent**: The P(B|A) = P(B) because A and B are independent

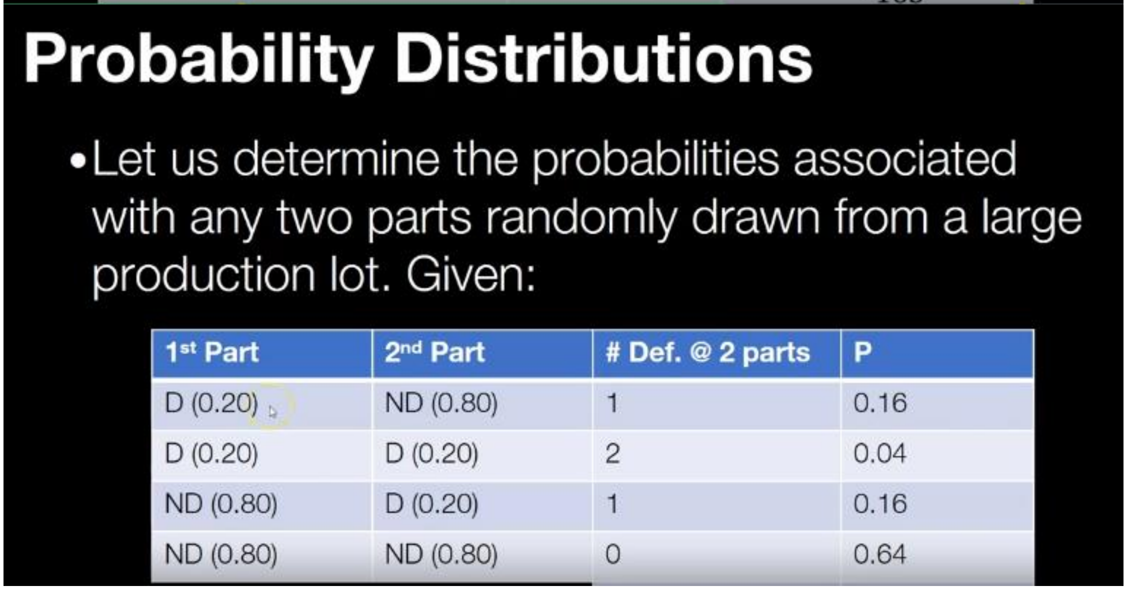
If the conditions are dependent

$$P(B|A)=\frac{P(BA)}{P(A)}$$



Note also that the P(Defective and Vendor A) constitutes a joint probability under statistical dependence. Creating a table of joint P values for the sample space:

Event	P	Fraction
Vendor A and Defective	0.0909	$^{15}/_{165}$
Vendor A and Not Defective	0.5151	$^{85}/_{165}$
Vendor B and Defective	0.0606	$^{10}/_{165}$
Vendor B and Not Defective	0.3333	$^{55}/_{165}$



Probability Distributions

- Let us determine the probabilities associated with any two parts randomly drawn from a large production lot. Given:

1st Part	2nd Part	# Def. @ 2 parts	P
D (0.20)	ND (0.80)	1	0.16
D (0.20)	D (0.20)	2	0.04
ND (0.80)	D (0.20)	1	0.16
ND (0.80)	ND (0.80)	0	0.64

The above total is 1.00 and the table is given above. Since we multiply them together.

For a binomial distribution in R using lolcat

require(lolcat)

table.dist.binomial(n=2,p=0.2)

The above table also gives exact probabilities or above or below.

for example one or less defective is 0.96

More on probability distributions

A discrete probability distribution is one where there are a limited number of possible values and a distribution can only fall into either discrete or continuous probability categories

Let’s assume we have an automated process that produces between 50 and 60 parts per day. During a two month production period, daily production levels will refer to that as DP were noted and the following data were generated.

frequency distribution for discrete random variable

```
freqdist <- frequency.dist.grouped(Daily.Production$X50)
```

Send that to an object that will be smaller

```
probdist <- freqdist[,c(“min”,“freq”,“rel.freq”)]
```

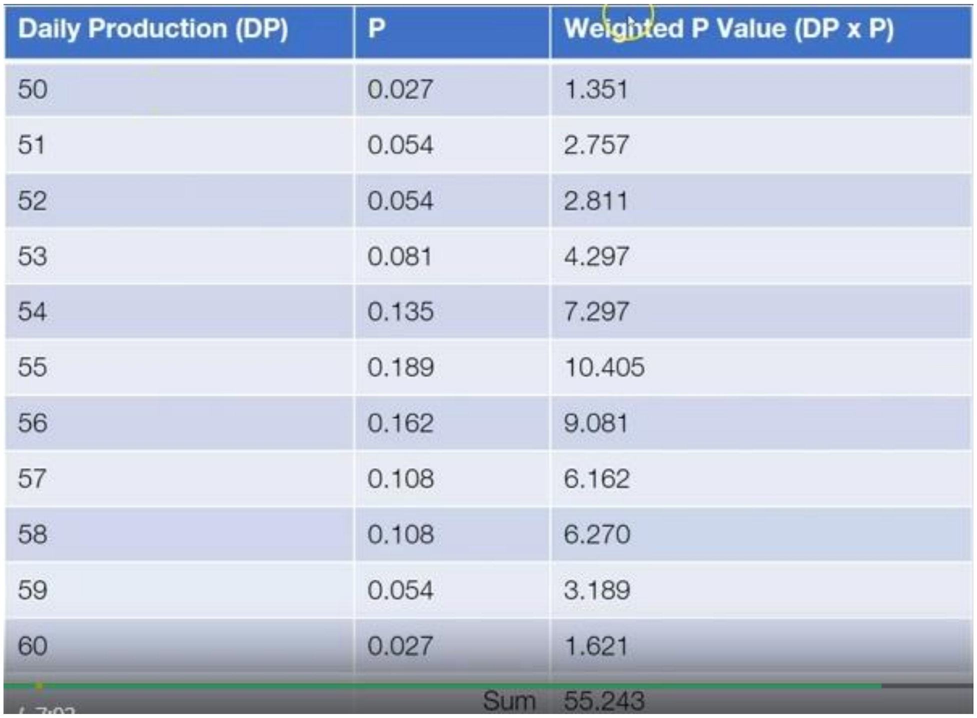
and give the columns names

```
colnames(probdist)<-c(“Daily production”,“# of days”, “P(DP)”)
```

Now a histogram, if you want probability, use freq=F, and use anchor value if the data is discrete

```
hist.grouped(Daily.Production$X50, freq=F,anchor.value=50)
```

The expected value of a discrete random variable is the weighted average of the expected outcomes.



Daily Production (DP)	P	Weighted P Value (DP x P)
50	0.027	1.351
51	0.054	2.757
52	0.054	2.811
53	0.081	4.297
54	0.135	7.297
55	0.189	10.405
56	0.162	9.081
57	0.108	6.162
58	0.108	6.270
59	0.054	3.189
60	0.027	1.621
	Sum	55.243

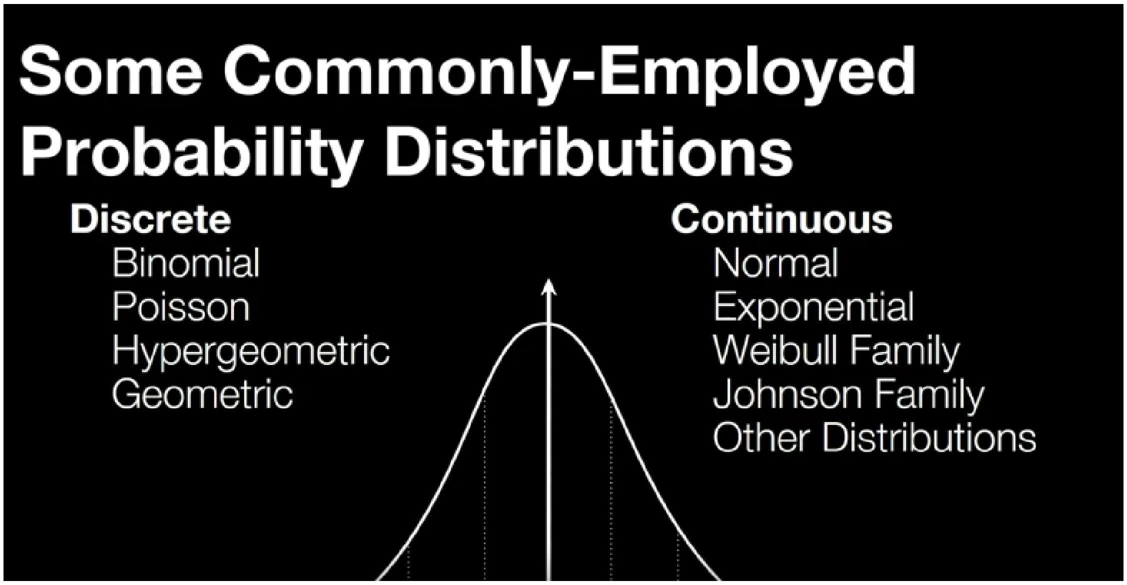
The R function for this is

```
weighted.mean(x,y)
```

Expected value of a discrete random variable

x<-probdist$`Daily production` y <- probdist$`P(DP)` weighted.mean(x,y)

and weighted mean just using weighted.mean(Daily.Production$X50) seems to also work



Some Commonly-Employed Probability Distributions

Discrete
Binomial
Poisson
Hypergeometric
Geometric

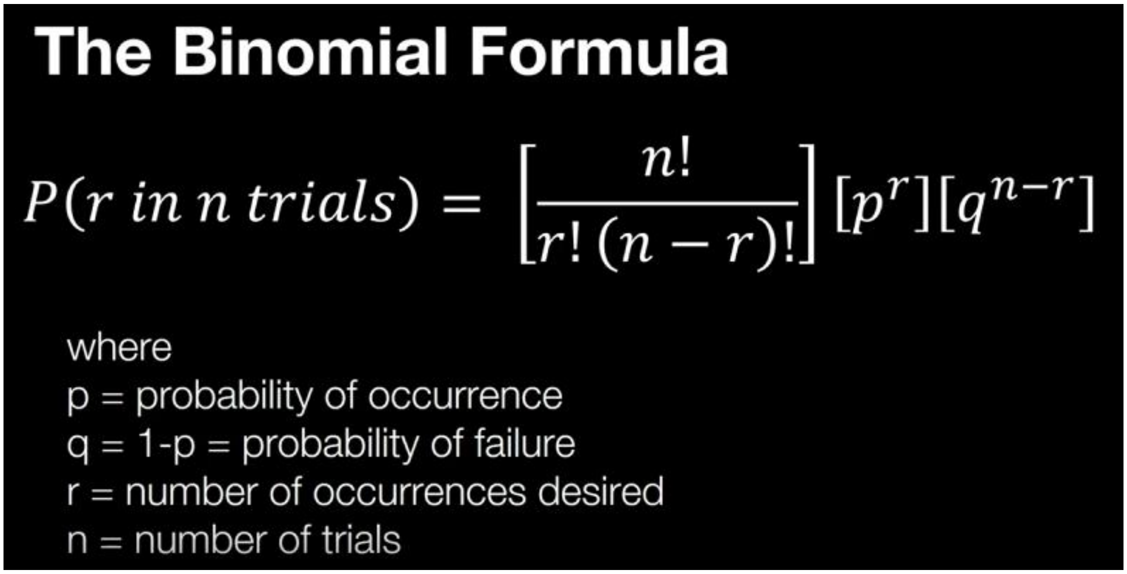
Continuous
Normal
Exponential
Weibull Family
Johnson Family
Other Distributions

Calculating

Binomial Distribution

Relates to a discrete random variable for nominal data

Only two outcomes which remains fixed over time.



The Binomial Formula

$$P(r\ in\ n\ trials) = \left[\frac{n!}{r!\,(n-r)!}\right][p^r][q^{n-r}]$$

where
p = probability of occurrence
q = 1-p = probability of failure
r = number of occurrences desired
n = number of trials

Let's look at an example, let's say we have a vendor that frequently ships 2 bad parts out of 10. Let's suppose the vendor ships our company 50 parts. If we tell them that at least 9 parts out of 10 must be good, and nothing in their manufacturing process has changed, what is the probability that we will receive what we asked for?

For this example, we have to define what we know. We know that they'll ship us 2 bad parts out of 10, which gives us a probability of getting a good part of 80 percent or 0.80. The probability of getting a bad part is q or 1 minus p which is 0.20, but what we're interested in is r or the count of the good parts we would get. Our requirement was 9 out of 10, so our r in this case if we're shipping 50 parts is going to be 0.9 times 50 or 45.

It would become P(45)+P(46)+P(47)+P(48)+P(49)+P(50)

Use DBINOM which gives probability density at exactly a certain value

dbinom(x=45,size=50,prob=0.8)

Or make table (Table.dist.binom)

table.dist.binomial(n=50,p=0.8)

Use PBINOM for X >=45

Values Greater than X, see below for => X

Note that pbinom gives P[X>x] for upper tail probs so to get X to be included, make it 1 smaller. So if looking for 45, do q =44

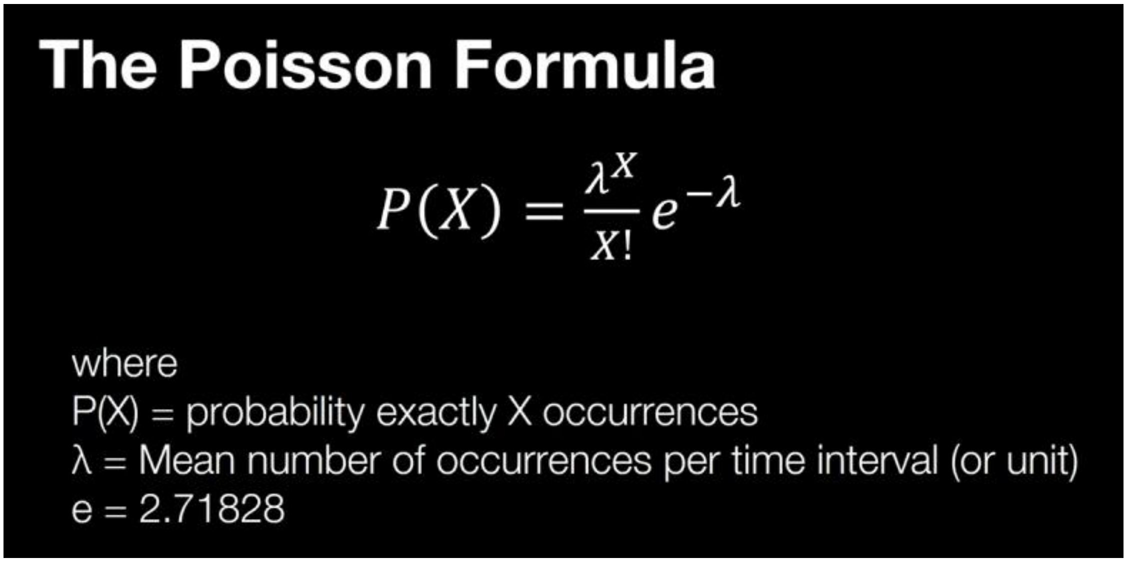
pbinom(q=44,size=50,prob = 0.80,lower.tail = F)

Poisson Data

Key word to look for is PER

And the mean must be known

Number of parts PER shift Number of breakdowns PER shift Number of failures per 100 cycles



The Poisson Formula

$$P(X) = \frac{\lambda^X}{X!} e^{-\lambda}$$

where
P(X) = probability exactly X occurrences
λ = Mean number of occurrences per time interval (or unit)
e = 2.71828

$$\lambda = 25$$

$$X = 10$$

25 parts per hour, X = probability we are looking for (10)

$$P(10) = \frac{25^{10}}{10!} e^{-25} = 0.0000365$$

R function is DPOIS

dpois(x=10,lambda = 25)

Poisson example table

table.dist.poisson(lambda=25)[7:51,]

Poisson prob for 10 or fewer

ppois(q=10,lambda=25, lower.tail=T)

Poisson prob of at least 20 but no more than 30

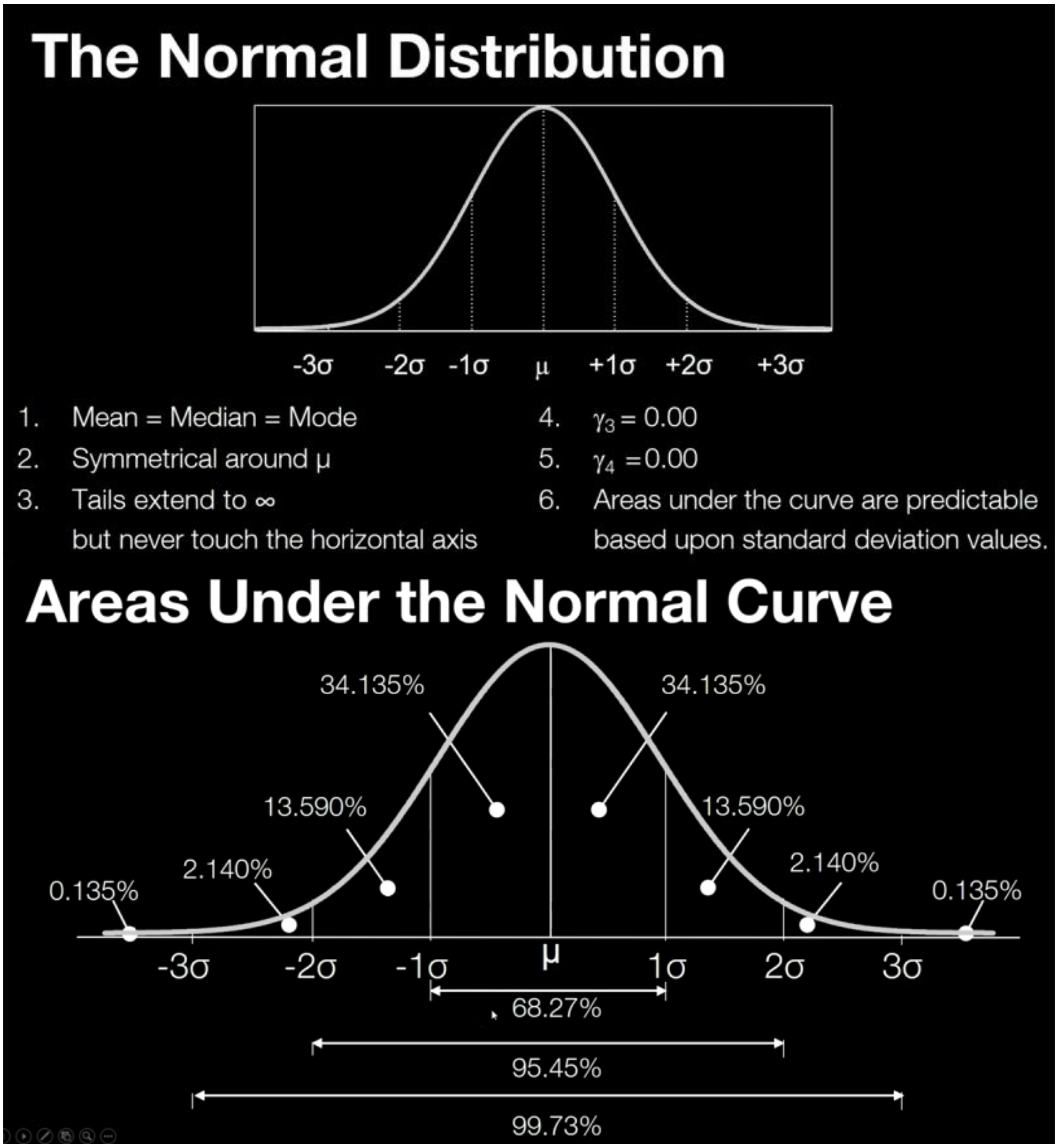
ft20<-ppois(q=19,lambda = 25,lower.tail = T) tw30<-ppois(q=30,lambda = 25,lower.tail = T) tw30-ft20

Poisson dist for testing

poissdist <- rpois(n =100,lambda=25) poisson.dist.test(poissdist)

If data is distributed using poisson dist. The P value will be > 0.05

Normal Dist



The Normal Distribution

1. Mean = Median = Mode
2. Symmetrical around μ
3. Tails extend to ∞ but never touch the horizontal axis
4. $\gamma_3 = 0.00$
5. $\gamma_4 = 0.00$
6. Areas under the curve are predictable based upon standard deviation values.

Areas Under the Normal Curve

The area corresponding to any score can be found using the Z score

Z is the number of stdev from X to u

$$z = \frac{X - \mu}{\sigma}$$

For example, today we have some tooling used on a particular drilling process and it's lasted an average of 180 hours before requiring replacement and the process has a standard deviation of five hours. What is the probability that a tool selected at random from the tool crib will last less than 172 hours before replacement is needed?

$$z = \frac{X - \mu}{\sigma}$$

$$z = \frac{172 - 180}{5} = -1.60$$

Calculate area under normal curve using pnorm

```
pnorm(q=172,mean=180,sd=5, lower.tail=TRUE)
```

Calculate area under normal curve using zscore

```
pnorm(q=-1.6,mean=0,sd=1, lower.tail=T)
```

Each gives same value

Let’s take a look at another example. In this example, we have a stamping operation and it’s been running consistently, punching two holes in a sheet of metal. The distance from center to center between the two holes has been at an average of 5.20 millimeters, with a standard deviation of 0.05 millimeters.

The process produces center to center distances that can be modeled with a normal distribution. The specifications for these parts require a maximum or upper specification limit of 5.35 millimeters and a minimum or lower specification limit of 5.15 millimeters. What percentage of the manufactured parts are likely to fall outside of the specifications?

$$z = \frac{X - \mu}{\sigma}$$

$$z = \frac{5.15 - 5.20}{0.05} = -1.00$$

$$z = \frac{5.35 - 5.20}{0.05} = 3.00$$

Use the Z score (which sets mean to 0 and sd to 1)

```
pnorm(q=-1.0,mean=0,sd=1, lower.tail=T)
```

```
pnorm(q=3.0,mean=0,sd=1, lower.tail=F)
```

use pnorm and regular data, and make them into variables apperently

```
pnorm(q=5.15, mean=5.20, sd=0.05, lower.tail=T) pnorm(q=5.35, mean=5.20, sd=0.05,
lower.tail=F)
```

```
lower <-pnorm(q=5.15, mean=5.20, sd=0.05, lower.tail=T) upper <-pnorm(q=5.35,
mean=5.20, sd=0.05, lower.tail=F)
```

```
lower+upper
```

Another example using Dnorm

x = seq(5,5.45,length=200) y=dnorm(x,mean=5.22,sd=0.05)

plot(x,y,type=‘l’)

Test for normality

If data normally distributed it will have a P value greater than 0.05 for anderson/wilk/etc

if n <25, use anderson darling test for normality,shapirowilk too if n>25 use skewness and kurtosis tests (D’Agostino)

n <25 anderson.darling.normality.test() shapiro.wilk.normality.test() summary.continuous()

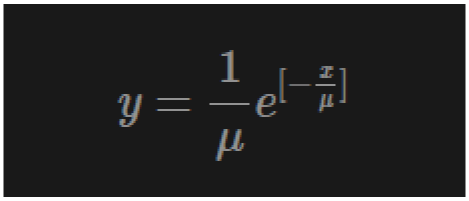
n >25 dagostino.normality.omnibus.test() summary.continuous()

normdata1 <-rnorm(n=24,mean=10,sd=2) anderson.darling.normality.test(normdata1) shapiro.wilk.normality.test(normdata1)

Exponential Distribution

Time to failiure for example follows an exponential dist

When xmin = 0



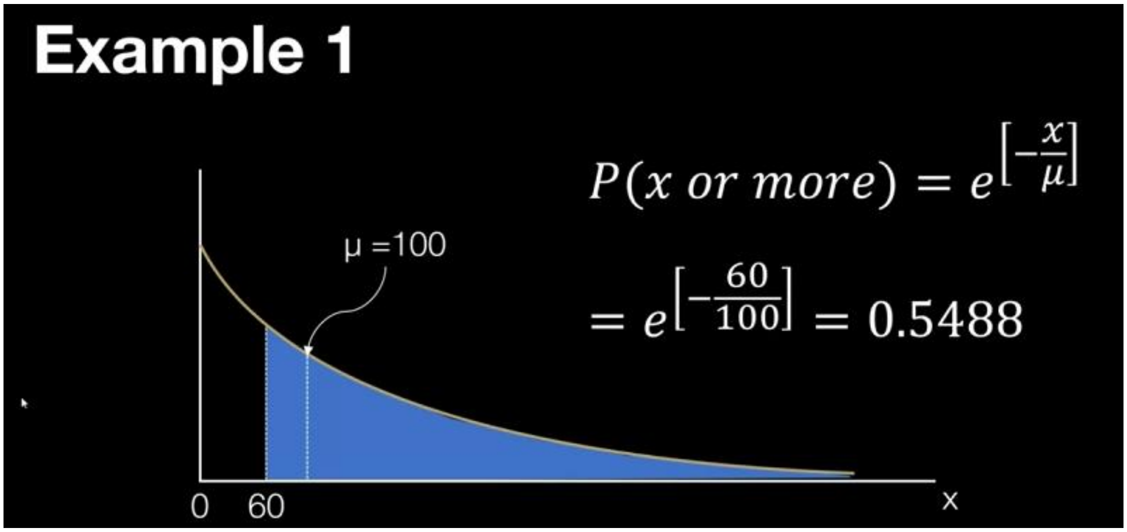
$$y = \frac{1}{\mu} e^{[-\frac{x}{\mu}]}$$

The normal distribution contains an area 50 percent above and 50 percent below the population mean. With the exponential distribution, 36.8 percent of the area under the curve is above the average and 63.2 percent is below.

Only the mean and min is required for calculations, often used for between failure analysis, and related to a continuous random variable.

Example

an in-plant study has shown that an engine control module laboratory tester is capable of operating on an average of 100 hours between breakdowns or a mean time between failure (MTBF) of 100 hours. What is the probability that the tester will run for at least 60 successive hours without a breakdown is distributed exponentially?



Example 1

$$P(x\ or\ more) = e^{\left[-\frac{x}{\mu}\right]}$$

$$= e^{\left[-\frac{60}{100}\right]} = 0.5488$$

Calculating

in R use the PEXP function

```
pexp(q=60,rate=1/100,lower.tail=F)
```

dexp is to make a picture of an exponential distribution

shape first x=seq(60,800,length=200) y =dexp(x,rate=1/100) plot(x,y,type=‘l’)

now fill it in

```
x=seq(60,800,length=100) y =dexp(x,rate=1/100) plot(x,y,type=‘l’)
polygon(c(60,x,800),c(0,y,0),col=‘red’)
```

Another example

The mean time between breakdowns has been established at 50 minutes or mu equal to 50. The origin parameter (X_min) is estimated to be five minutes, what is the probability of this machine running 20 minutes or less before a breakdown? Now, when we have a minimum value or origin parameter, sometimes it’s referred to as Omicron, we have to account for this in our equation, essentially shifting the entire distribution back to the point or the origin parameter. In this case, the origin parameter is five or X_min is five.

For a lower rate,

```
pexp(q = (20-5), rate = 1/(50-5),lower.tail=T)
```

```
plot(pexp.low(q=20,low=5,mean=50,lower.tail=T))
```

Test for exponentionalati when n <=100, shapiro wilk

```
expdata <-rexp(n=100,rate=1/50)
```

```
shapiro.wilk.exponentiality.test(expdata)
```

if n>100 use epps and pulley

```
expdata1 <-rexp(n=101,rate=1/50) shapetest.exp.epps.pulley.1986(expdata1)
```