[Video Notes](https://zoom.us/rec/play/Ff0c-UnpirCvX_YG6nfzg5Cnl-WjsMAuDEabXwuBer796np1HiJl3-FxvWll3-tGad6aOYDLL0N2qZ1D.P14IZZdRsV_Ctg0H)

<https://docs.google.com/document/d/1zI5vlATgrhKntawOxfXV7K25JYvhzzZqjRIp5FjmzdE/edit?usp=sharing>

[00:00:36]

In last class:

Modeling variable dependencies using linear models (two or more variables)

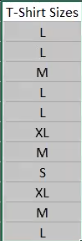
Data -> Find a Model -> Predict using the model.

[00:01:13]

* We will learn about the **One Variable Model**

[00:02:01]

* Below given are the T-Shirt sizes.



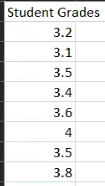
**Question**: What is the most common t-shirt size?

**Answer**: L

* The mode is the value that appears most often in a set of data values.

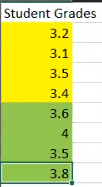
[00:04:29]

* Below given are the student grades.



**Question**: Which number separates the top half from the bottom half of the grades?

**Answer**: 3.5



Median - The value that splits the set in two by 50/50 difference

In R:

grades = c(3.2, 3.1, 3.5, 3.4, 3.6, 4, 3.5, 3.8)

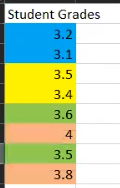
median(grades)

[00:06:36]

**Question**: Which value divides the set into the top 25% vs the bottom 75%?

**Answer**: 3.6 or 3.8 depending on whether we include the boundary or not.

In R → quantile(grades, .75)



Quartile - The value that divides the set into quarters and take one of the quarters

Top 100% - 4th Quartile

Top 25% - 3rd Quartile

Top 50% - 2nd Quartile

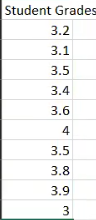
Top 75% (bottom 25%) - 1st Quartile

4th Quartile - Maximum

2nd Quartile - Median

* *Here we calculated the 3rd quartile.*

[00:10:33]



**Question**: What is the number that produces a top 10%, bottom 90% split?

**Answer**: 3.9 or 4 depending on whether the boundary is inclusive

**Percentile** - Divides the set on an arbitrary percentage line (x% vs (100-x)%)

50% percentile - 2nd quartile - Median

100% percentile - 4th quartile - Maximum

In R:

grades1 = c(3.2, 3.1, 3.5, 3.4, 3.6, 4, 3.5, 3.8, 3.9, 3)

quantile(grades1, .90)

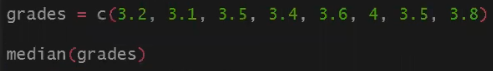
[00:13:11]

* Storing the values in the grades variable in RStudio.



[00:14:14]

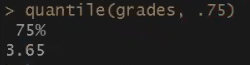
* Use median() in R to calculate the median of the values.





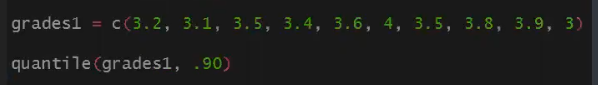
* Use quantile() in R to calculate the quartiles of the values.

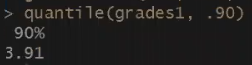


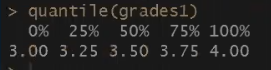


[00:16:10]

* Calculate the top 10% of the grades:

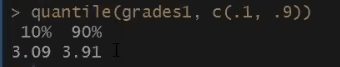




 -> 

* The above function will calculate all the quartiles.
* To calculate the bottom 10% and top 10%:





[00:20:56]

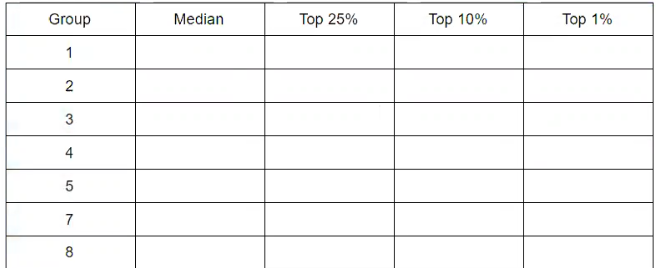
* To read a CSV file in R use, read.csv() function.
* Use the $ symbol to read the columns in that dataset.

data = read.csv("...")

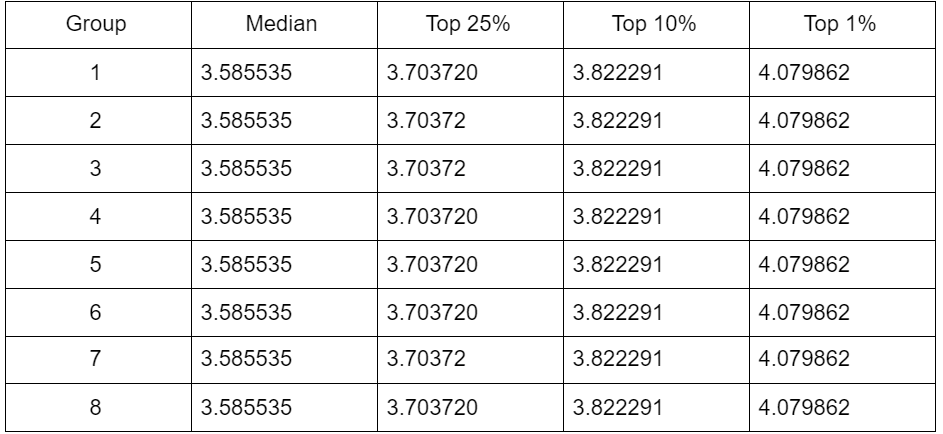
data$columnName

**Activity**:

* Complete the below table:



[00:32:46]



**Activity**:

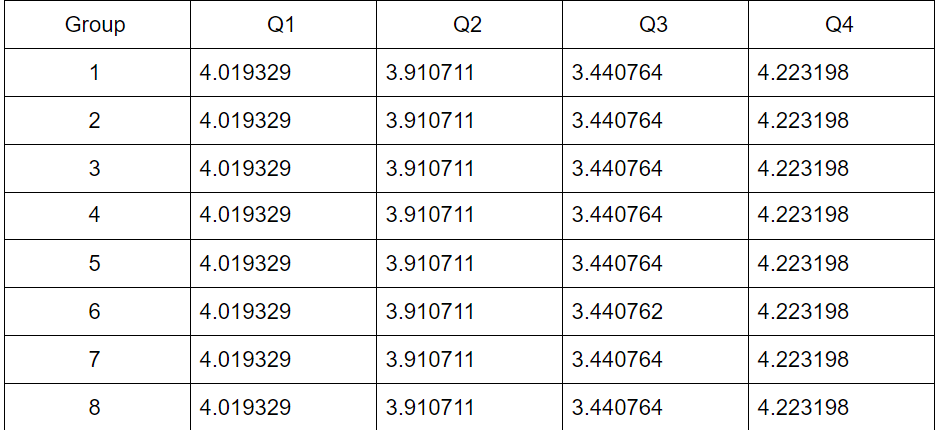
Q1 - The school is organizing an honor society for the top 2% of students, what is the cut-off grade to get into the society

Q2 - A professor wants to curve the grades so that 5% of the students get a 4.0+, what is the cut-off grade for conversion into 4.0

Q3 - The school is planning on providing student support to the bottom 20% of the grades, what is the cut-off grade to get support?

Q4 - The school wants to pick the top student in the cohort, what is the grade of the top student (possibly more than 1)?

[00:42:33]



Implicit Assumption - The data set = the whole population (these grades represent all the students). When that happens then this math can be used.

Many times this doesn’t happen. We need to use samples. Impractical to get every data point (i.e. survey, all height, all weight).

[00:44:47]

Approximations.

Sample:

* Random
* Large Enough

Estimate the average human height?

* Sample needs to be people across the whole world.

Estimate the average human height in the US?

* Sample people across the US

[00:46:14]

Approximate the median (middle of the set) → Average / Mean

Approximate the separations sets (quartiles) → Standard Deviation

Mean → Measure of the center of the data set

Standard Deviation → Measure of the separation from the center → High Standard Deviation → Lots of separation from the center.

Height → Design houses so that people fit.

If height has high standard deviation → Houses being too tall

Its height has a low standard deviation → Ceiling height will be closer to the average people height

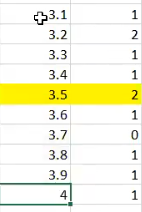
The behavior of the set regarding the average and standard deviation can be **modeled**.

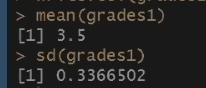
We can use those models to analyze and make predictions.

[00:49:54]

We want to model the behavior of the counts of the values.

* The below shows the number of times that particular has occurred.
* The highlighted row shows the average of the dataset.

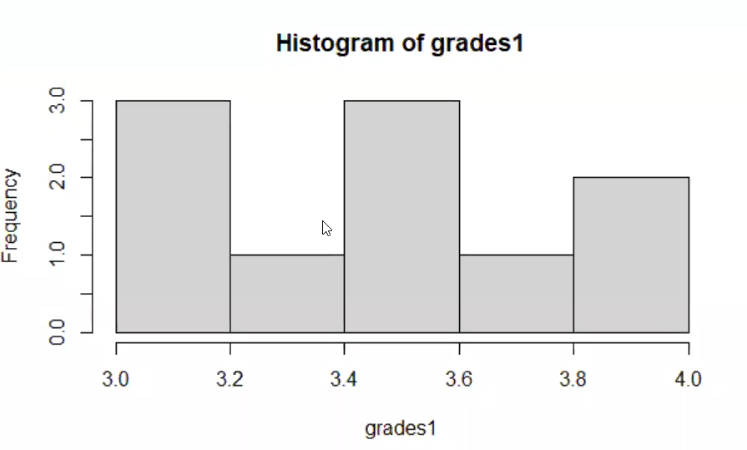


 -> 

[00:51:56]

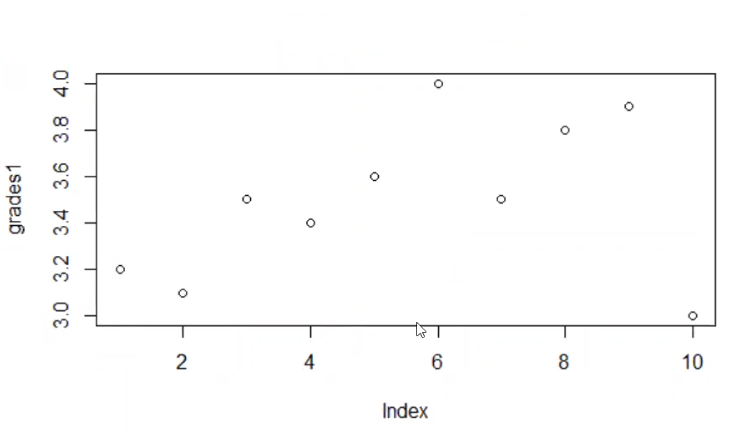
* We can create a histogram that shows where the points are more or less.
* In R, histograms can be created using the hist() function.





* Use the R plot() function to plot the data points.

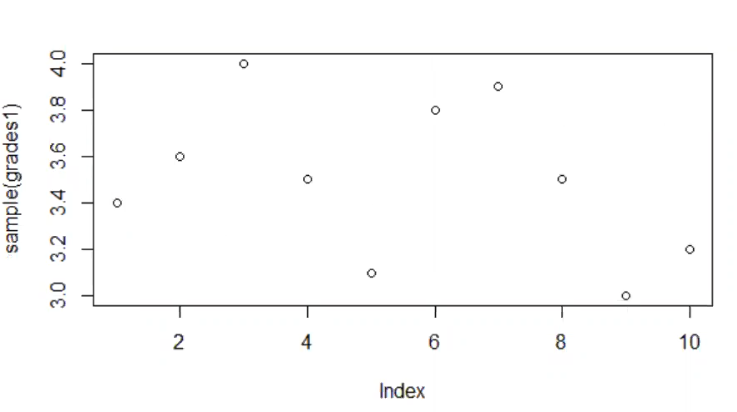




[00:52:45]

* The above data points were plotted in a sequential manner so we can unsort them using sample().





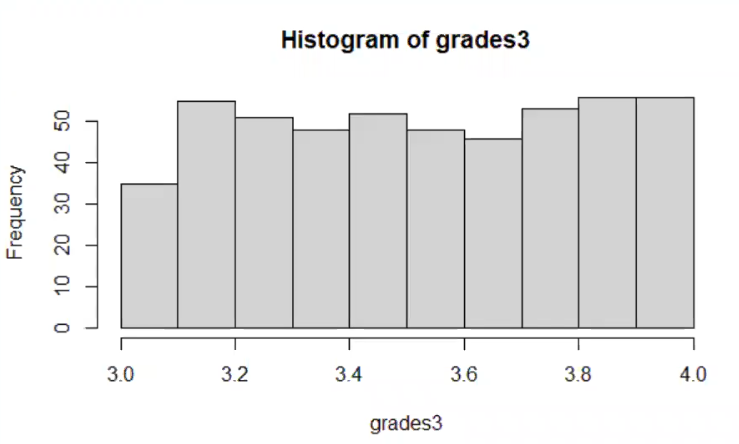
Random distribution → Uniform → Any point can appear with the same chance

*The points are not accumulated at one place i.e. the midpoint is 3.5 but the points are scattered all over the place.*

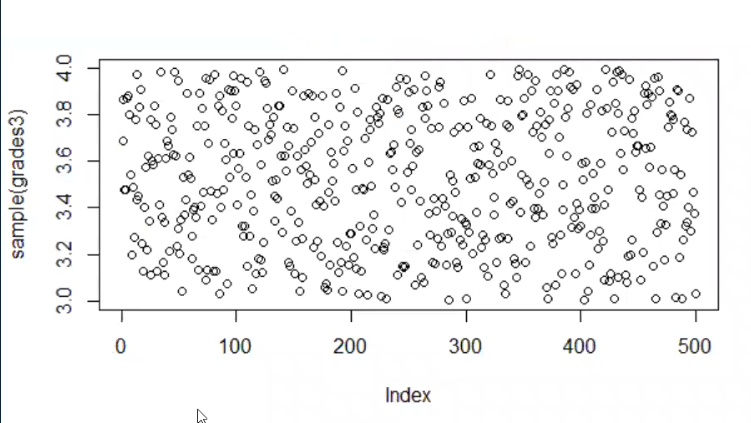
[00:55:33]

* We have a new dataset here named grades3.
* We will create a histogram for that data.









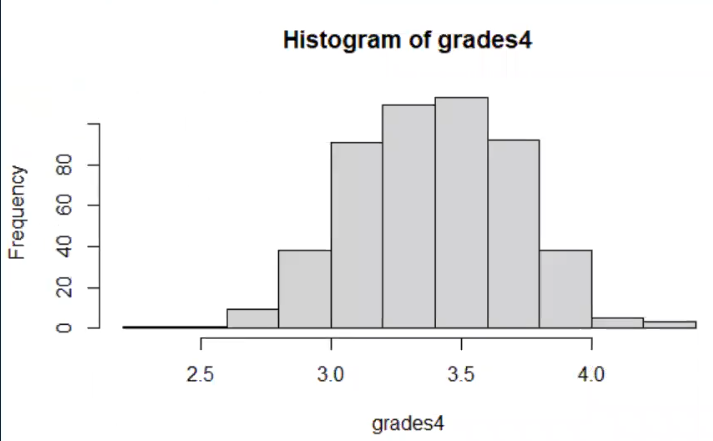
The model is good if you need complete randomness → Casino this should happen. Lottery same thing.

[00:57:23]

Normal Distribution:

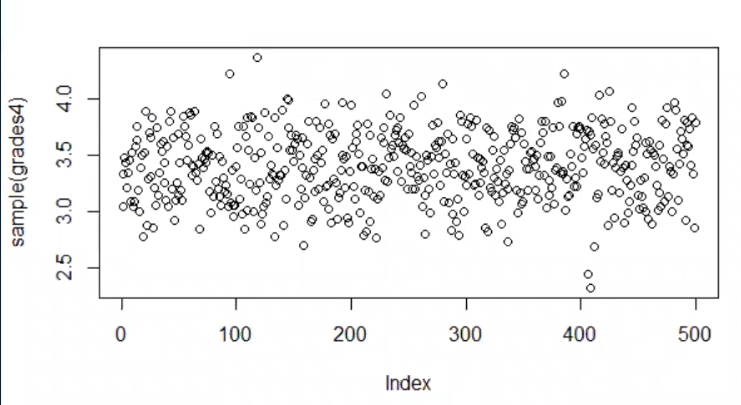
* The normal distribution is a continuous probability distribution that is symmetrical on both sides of the mean, so the right side of the center is a mirror image of the left side.
* The area under the normal distribution curve represents probability and the total area under the curve sums to one.
* Creating a histogram for grades4







(Standard deviation of 0.3)

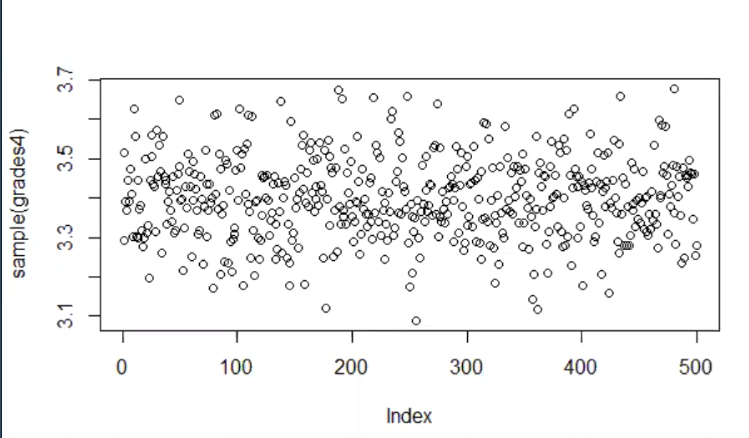


* Here the points seem to be accumulating to the middle.

[00:59:22]

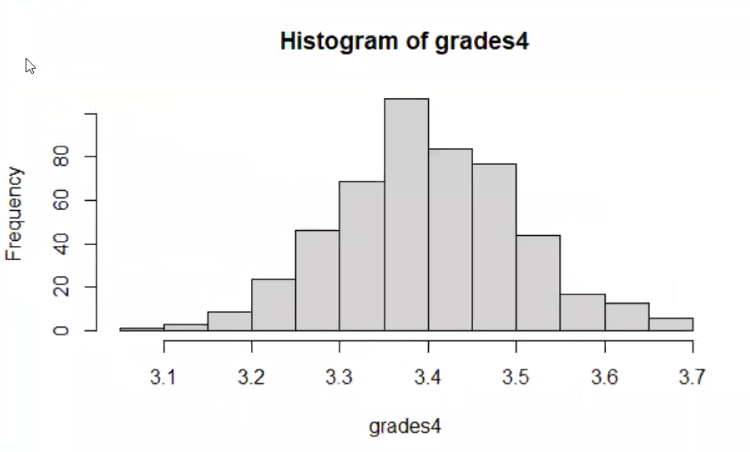
* Taking the same dataset and lowering the standard deviation:

(Standard Deviation of 0.1)



The same way that we looked at a chart and said variables look linear. Look at a histogram and say variables look uniform or normal (also with direct plot, easier with the histogram)

(Standard Deviation of 0.1)



[01:03:16]

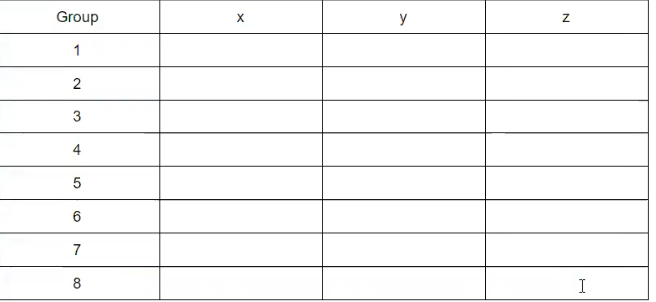
**Activity**:

For each variable:

* Uniform or Normal
* If Uniform find Max and Min
* If Normal find mean and std

Ex- uniform(10, 20) - Looks uniform and min=10, max=20

normal(2, 1) - Looks normal mean = 2, and sd = 1



[01:18:21]



[01:19:18]

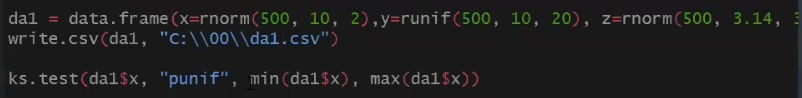
There are statistical tests to identify the distribution

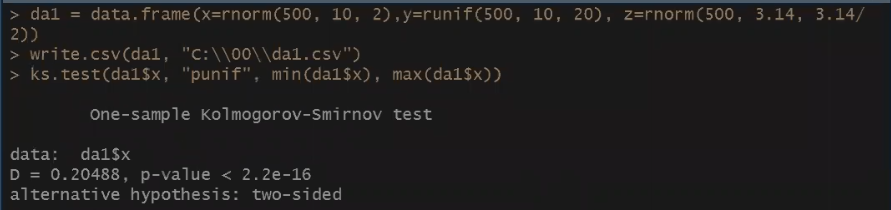
To test whether something is uniform - the Kolmogorov-Smirnov test

To test whether something is normal - Shapiro–Wilk test

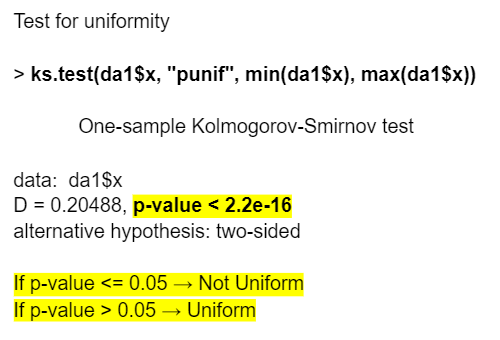
[01:20:54]

* In R, ks.test() performs one or two sample Kolmogorov-Smirnov tests.
* We can use this test to test if something is normal or not.

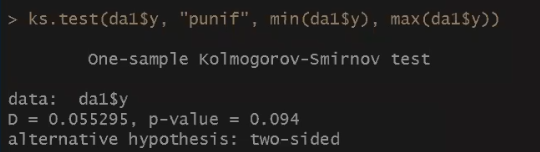


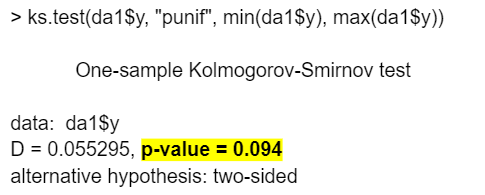


[01:23:01]







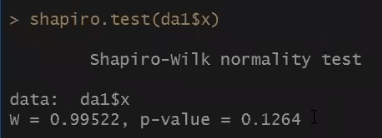


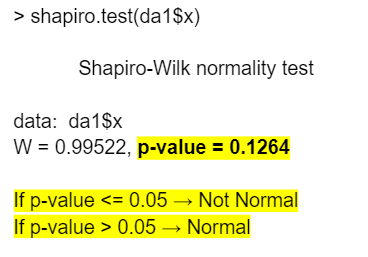
[01:24:53]

Shapiro-Wilks Normality Test:

* The Shapiro-Wilks test for normality is one of three general normality tests designed to detect all departures from normality.
* The test rejects the hypothesis of normality when the p-value is less than or equal to 0.05.

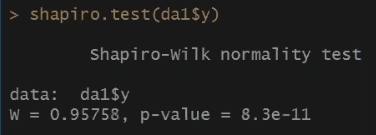


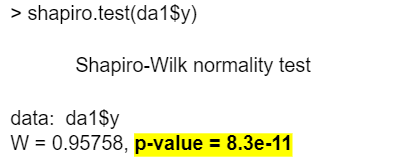




[01:25:50]







[01:27:10]

Data --> Model / Distribution → Use the model

Not everything has a model → Stock Market doesn’t have a model (known so far) - There are partial models

Many things have models → Many things we can discuss

Being random doesn’t mean no model → Roulette has a well-defined model, the model doesn’t help

Statistics - Binomial, Poisson, Lambda

[01:43:19]

**Activity**:

Da2

<https://drive.google.com/file/d/1pTdfQ95RqYqcUxCkxrlV5fV7KHKiWl9Z/view?usp=sharing>

[Da1.csv](https://drive.google.com/file/d/1Z9yg-h82Y2zSiqUhzPXPM6yqdYb-Remx/view?usp=sharing)

The distribution

P-value of the tests

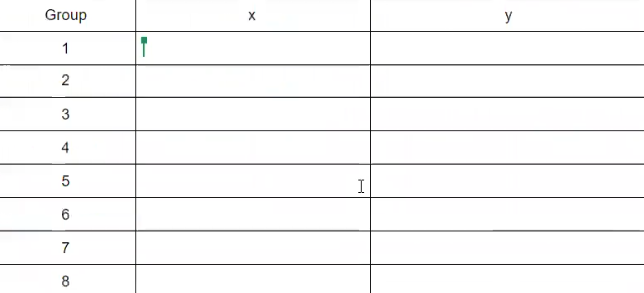
KS: p-value

SW: p-value

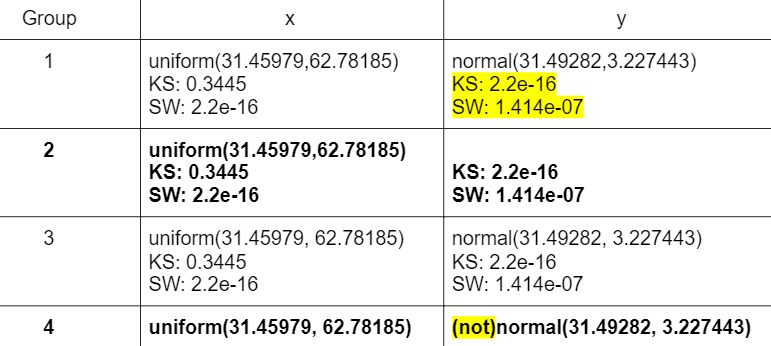
uniform(1,2)

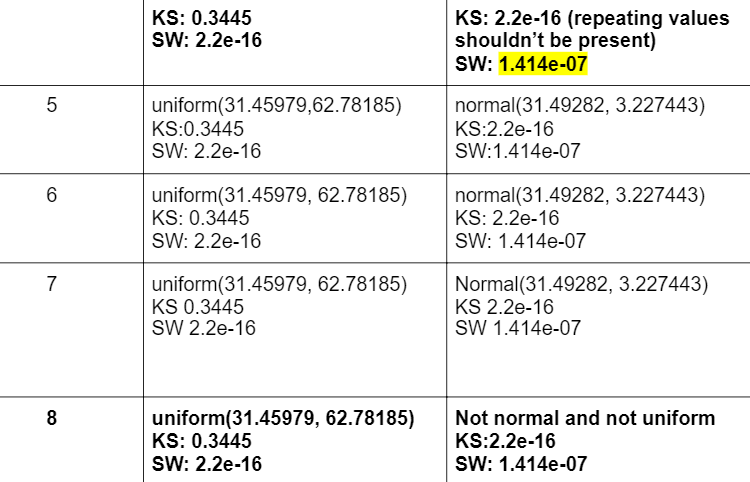
KS: 0.2

SW: 0.0001



[02:02:56]





[02:05:53]

If the data fits a distribution now we can use that to answer the same questions as we did at the beginning of the class.

da1$x

Normal

Mean: 9.92586

Sd:2.076673

Question: What is the cut-off for da1$x so that 10% of the points are above the cut-off?

> qnorm(.9, mean(da1$x), sd(da1$x))

**[1] 12.539**

Question: What is the cut-off for da1$x so that 10% of the points are below the cut-off?

> qnorm(.1, mhist(da1$x), sd(da1$x))

**[1] 7.247636**

qnorm is always calculating below (above = 1 - below)

[02:11:07]

Question: What % of the points are above 7?

> 1-pnorm(7, mean(da1$x), sd(da1$x))

**[1] 0.9194684**

**92%**

Question: What % of the points are below 6?

> pnorm(6, mean(da1$x), sd(da1$x))

**[1] 0.02965435**

**3%**

Question: What % of the points are between 6 and 7?

> pnorm(7, mean(da1$x), sd(da1$x)) - pnorm(6, mean(da1$x), sd(da1$x))

**[1] 0.05087727**

**5%**

Below 6, between 6 and 7 or above 7. The sum of those % needs to be 1 (92 + 3 + 5 = 100)%

[02:15:18]

Uniform distribution - **qunif** and **punif**

Other distributions **qx** **px** where x is the distribution

[02:17:35]

New Data Set

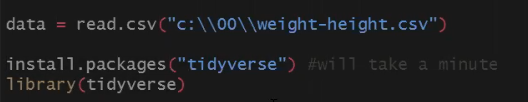
<https://drive.google.com/file/d/1u47Ab_ctM4gG77d8LluzEzdRk-vFuERR/view?usp=sharing>

How to reduce the set for distribution testing

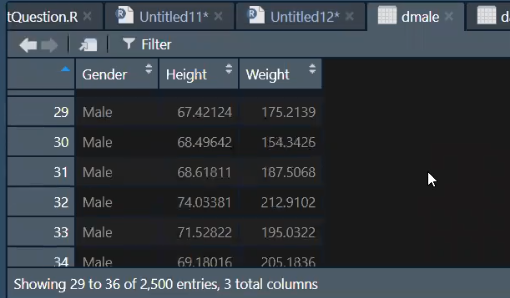
sample(data, 5000)

install.packages("tidyverse") #will take a minute run once

library(tidyverse) #run every you restart rstudio



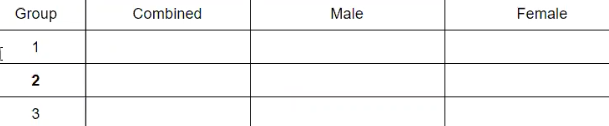
dmale = data %>% filter(Gender == "Male")

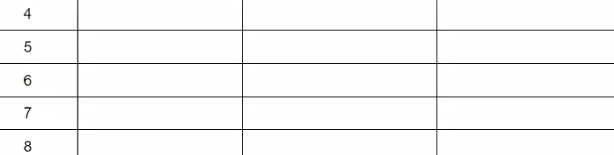


[02:21:11]

**Activity**:

Find distributions:





Q1 - % of total population with height > 70 inches

Q2 - % of male population with height > 70 inches

Q3 - % of female population with height > 70 inches

Q4 - % of the total population with height < 65 inches

Q5 - % of male population with height < 65 inches

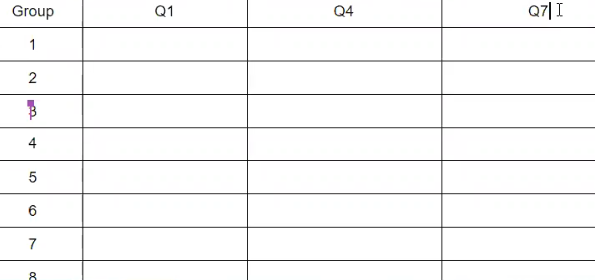
Q6 - % of female population with height < 65 inches

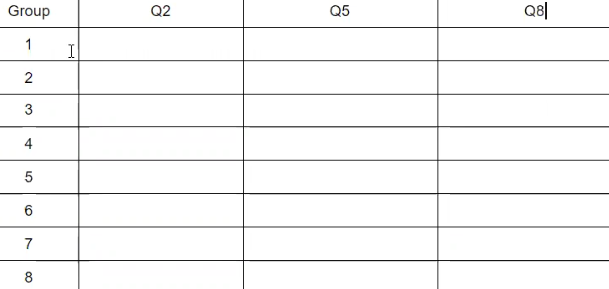
Q7 - What is the height so that only 10% of the population is above

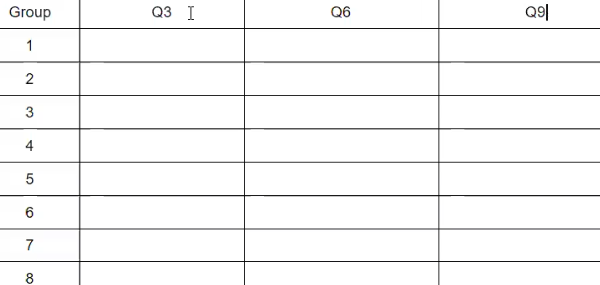
Q8 - What is the height so that only 10% of the male population is above

Q9 - What is the height so that only 10% of the female population is above

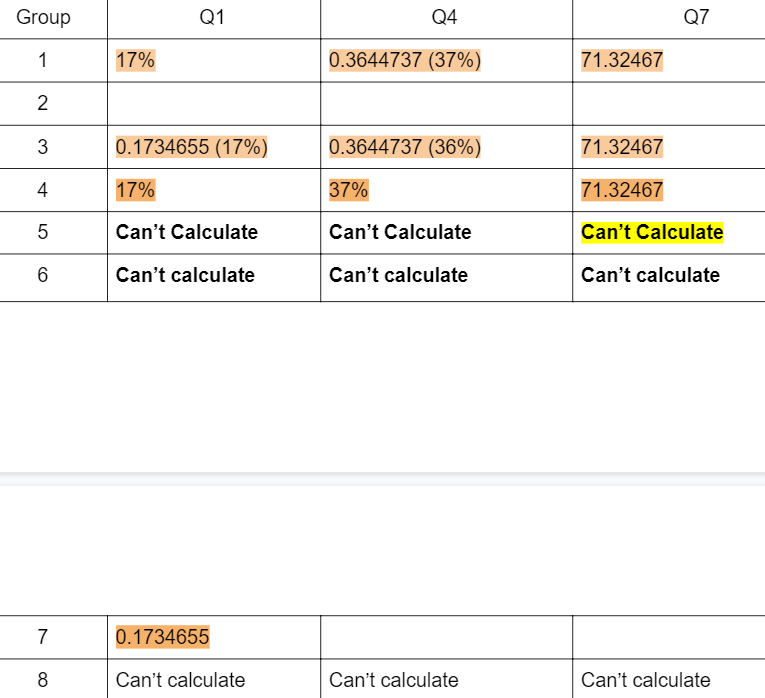
Notice the answer can be “Can’t calculate” (the data set doesn’t follow one of the distributions we use)



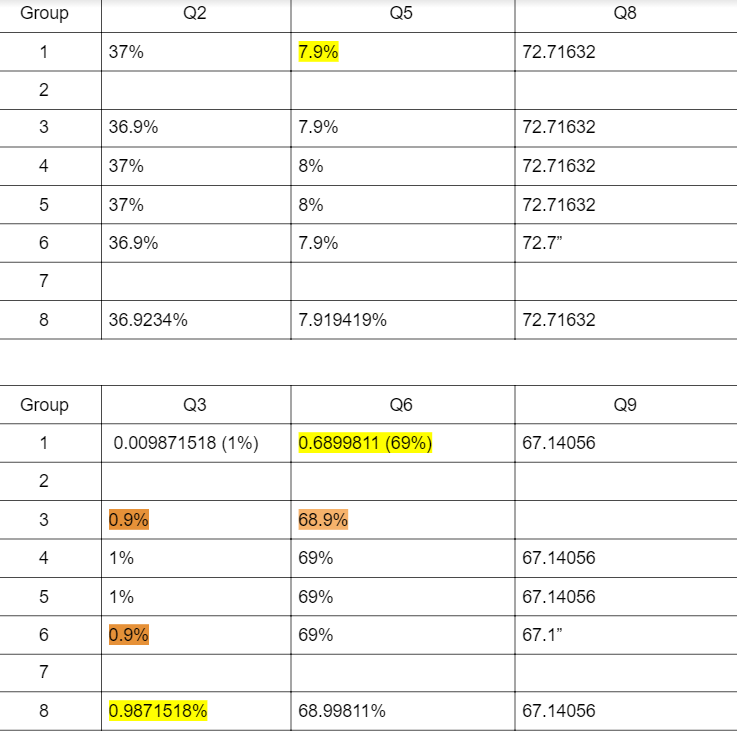




[02:48:33]



[02:50:42]



When simplifying digits -> Rond (don’t truncate)