[DA2https://docs.google.com/document/d/1zI5vlATgrhKntawOxfXV7K25JYvhzzZqjRIp5FjmzdE/edit?usp=sharing](https://docs.google.com/document/d/1zI5vlATgrhKntawOxfXV7K25JYvhzzZqjRIp5FjmzdE/edit?usp=sharing)

**Questions?**

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

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

**One Variable Models**

What is the most common t-shirt size?

l

L

L

L

L

L

L

L

L

L

Mode - The most common value in the set

Which number separates the top half from the bottom half of the grades?

3.5

3.5

3.5

3.5

3.5

3.5

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

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

median(grades)

Which value divides the set in top 25% vs bottom 75%?

quantile(grades, .75)

3.7

Inbetween 3.6 and 3.8

3.7?

3.8

3.

3.6

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

Quartile - The value that divides the set in 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

**3rd Quartile**

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

4.0

3.9

3.9

3.9

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

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

quantile(grades1, .90)

> quantile(grades1)

0% 25% 50% 75% 100%

3.00 3.25 3.50 3.75 4.00

> quantile(grades1, c(.1, .9))

10% 90%

3.09 3.91

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

data$columnName

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Group | Median | Top 25% | Top 10% | Top 1% |
| 1 | 3.585535 | 3.703720 | 3.822291 | 4.079862 |
| 2 | 3.585535 | 3.70372 | 3.822291 | 4.079862 |
| 3 | 3.585535 | 3.70372 | 3.822291 | 4.079862 |
| 4 | 3.585535 | 3.703720 | 3.822291 | 4.079862 |
| 5 | 3.585535 | 3.703720 | 3.822291 | 4.079862 |
| 6 | 3.585535 | 3.703720 | 3.822291 | 4.079862 |
| 7 | 3.585535 | 3.70372 | 3.822291 | 4.079862 |
| 8 | 3.585535 | 3.703720 | 3.822291 | 4.079862 |

Q1 - The school is organizing an honor society for the top 2% of students, what is the cutt-off grade to get in to 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)?

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Group | Q1 | Q2 | Q3 | Q4 |
| 1 | 4.019329 | 3.910711 | 3.440764 | 4.223198 |
| 2 | 4.019329 | 3.910711 | 3.440764 | 4.223198 |
| 3 | 4.019329 | 3.910711 | 3.440764 | 4.223198 |
| 4 | 4.019329 | 3.910711 | 3.440764 | 4.223198 |
| 5 | 4.019329 | 3.910711 | 3.440764 | 4.223198 |
| 6 | 4.019329 | 3.910711 | 3.440762 | 4.223198 |
| 7 | 4.019329 | 3.910711 | 3.440764 | 4.223198 |
| 8 | 4.019329 | 3.910711 | 3.440764 | 4.223198 |

Implicit Assumption - The data set = the whole population (this 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).

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

Approximate the median (middle of 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

It 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**.

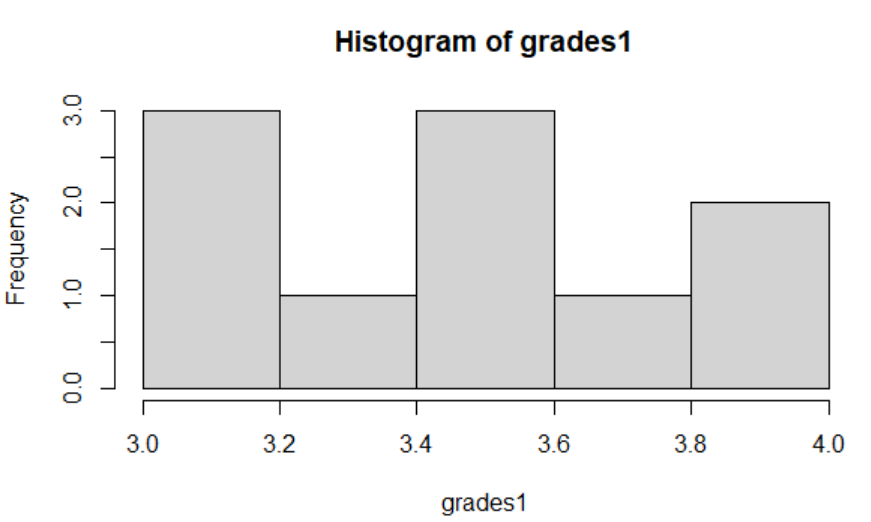
We want to model the behavior of the counts of the values.

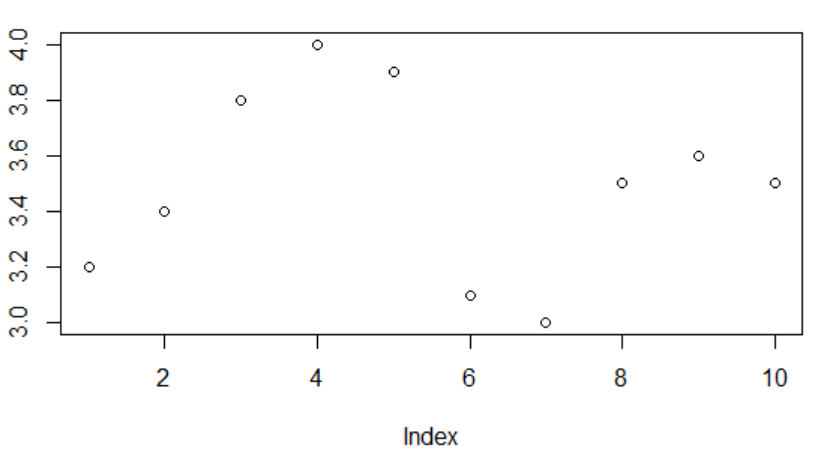
mean(grades1)

sd(grades1)

hist(grades1)

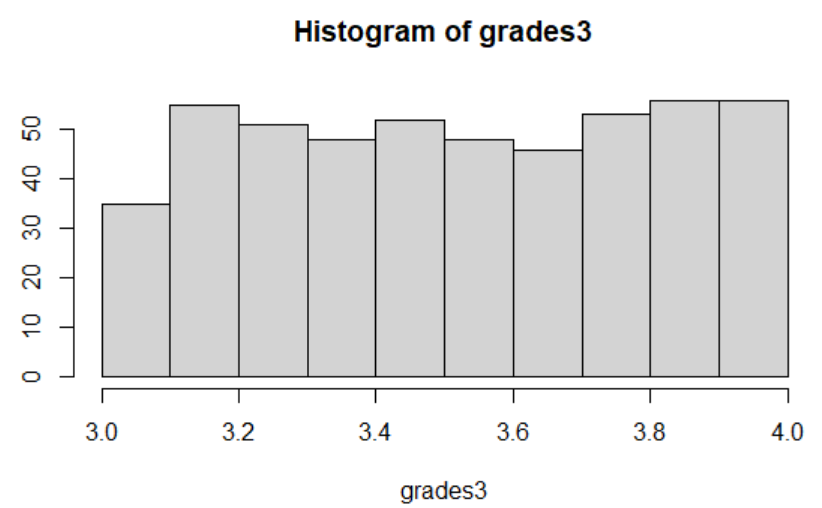
plot(sample(grades1))



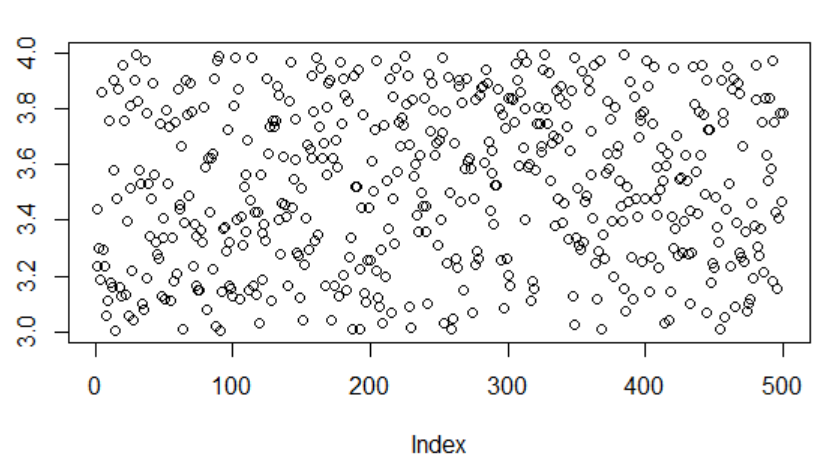


Random distribution → Uniform → Any point can appear with same chance

hist(grades3)

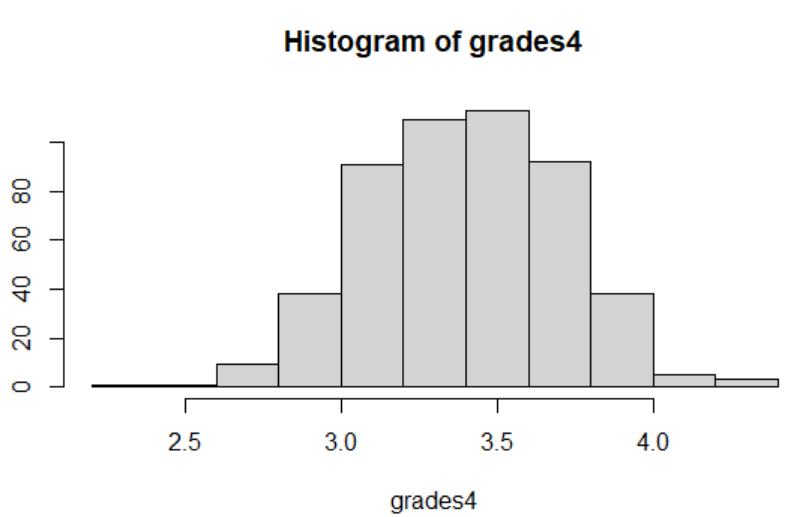


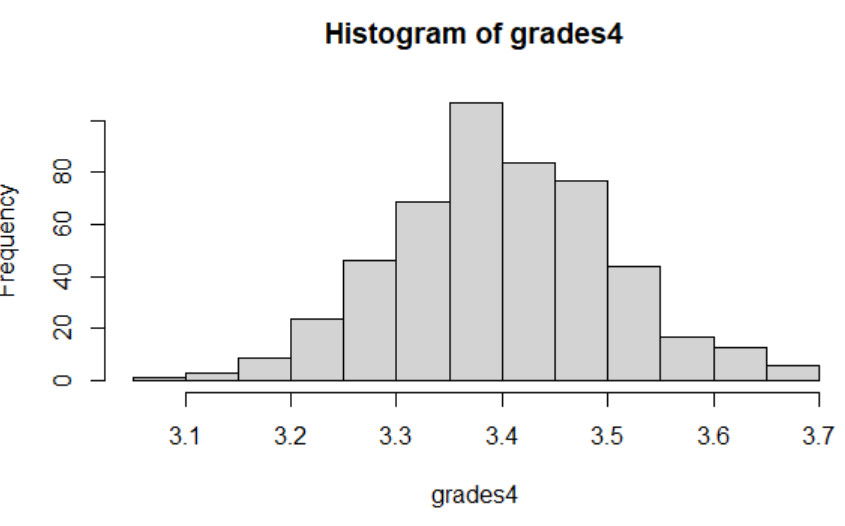
plot(sample(grades3))



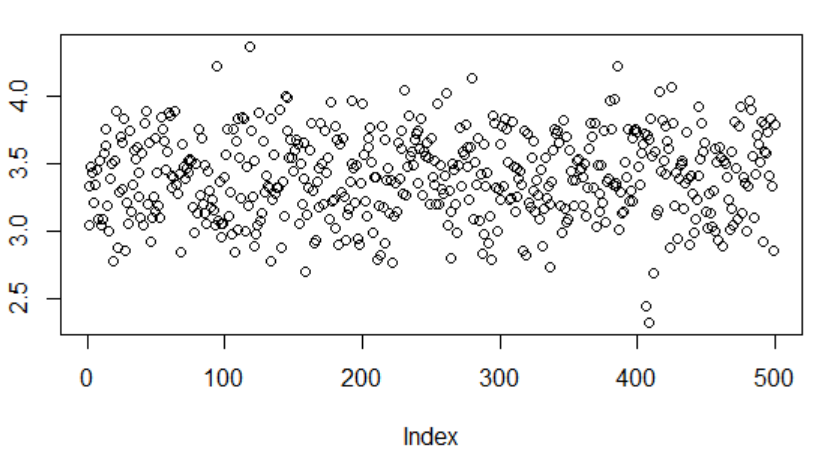
Model is good if you need complete randomness → Casino this should happen. Lottery same thing.

**Normal Distribution**

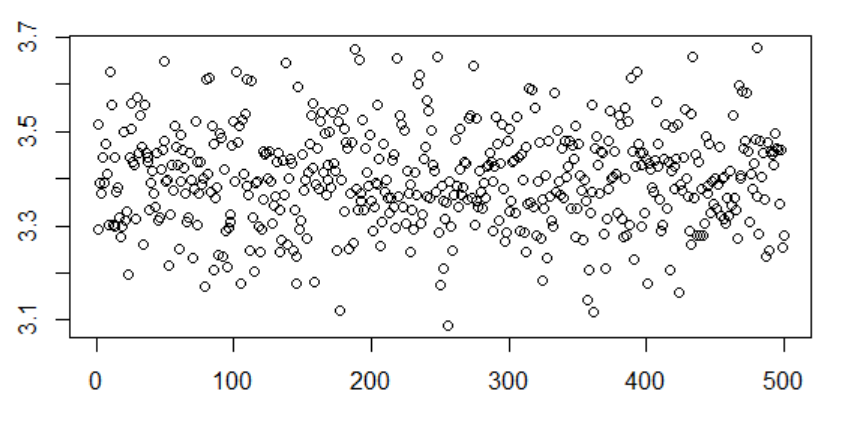




Standard deviation of .3



Standard Deviation of .1



Same way that we looked 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)

**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

|  |  |  |  |
| --- | --- | --- | --- |
| Group | x | y | z |
| 1 | Normal  Mean: 9.92586  sd:2.076673 | Uniform  Min: 10.00268  Max:19.98414 | Normal  Mean: 3.286698  SD:1.552503 |
| 2 | Normal mean:9.92586  std:2.076673 | Uniform  Min 10.00268  Max 19.98414 | Normal  Mean 3.286698  Sd 1.552503 |
| 3 | Normal  Mean: 9.92586  Sd = 2.076673 | Uniform  Min: 10.00268  Max:19.98414 | Normal  Mean: 3.286698  SD: 1.552503 |
| **4** | **Normal(9.92586, 2.076673)** | **Uniform(10.00268, 19.98414)** | **normal(3.286698, 1.552503)** |
| 5 | Normal  Mean:9.92586  SD: 2.076673 | Uniform  Min: 10.00268 Max: 19.98414 | Normal  Mean: 3.286698  SD: 1.552503 |
| 6 | Normal  Mean: 9.92586  Sd: 2.076673 | Uniform  Max: 19.98414  Min: 10.00268 | Normal  Mean: 3.286698  Sd: 1.552503 |
| 7 | Normal  Mean 9.893396  SD 2.076673 | Uniform  Max19.98414  Min: 10.00268 | Normal  Mean : 3.326628  SD:1.552503 |
| **8** | **normal(9.92586, 2.076673)** | **uniform(10.00268, 19.98414)** | **normal(3.286698, 1.552503)** |

Are we supposed to have a dataset? I don't see one in chat

[Marcelo] da1.csv

<https://drive.google.com/file/d/1Z9yg-h82Y2zSiqUhzPXPM6yqdYb-Remx/view?usp=sharing>

There are statistical tests to identify distribution

To test whether something is uniform - Kolmogorov-Smirnov test

To test whether something is normal - Shapiro–Wilk test

Test for uniformity

> **ks.test(da1$x, "punif", min(da1$x), max(da1$x))**

One-sample Kolmogorov-Smirnov test

data: da1$x

D = 0.20488, **p-value < 2.2e-16**

alternative hypothesis: two-sided

If p-value <= 0.05 → Not Uniform

If p-value > 0.05 → Uniform

> ks.test(da1$y, "punif", min(da1$y), max(da1$y))

One-sample Kolmogorov-Smirnov test

data: da1$y

D = 0.055295, **p-value = 0.094**

alternative hypothesis: two-sided

> shapiro.test(da1$x)

Shapiro-Wilk normality test

data: da1$x

W = 0.99522, **p-value = 0.1264**

If p-value <= 0.05 → Not Normal

If p-value > 0.05 → Normal

> shapiro.test(da1$y)

Shapiro-Wilk normality test

data: da1$y

W = 0.95758, **p-value = 8.3e-11**

**This data is fake and designed for this class**

**Not everything in the world is either normal or uniform, there are hundreds of other distributions.**

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

**Break till 7:45**

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

|  |  |  |
| --- | --- | --- |
| Group | x | y |
| 1 | uniform(31.45979,62.78185)  KS: 0.3445  SW: 2.2e-16 | normal(31.49282,3.227443)  KS: 2.2e-16  SW: 1.414e-07 |
| **2** | **uniform(31.45979,62.78185)**  **KS: 0.3445**  **SW: 2.2e-16** | **KS: 2.2e-16**  **SW: 1.414e-07** |
| 3 | uniform(31.45979, 62.78185)  KS: 0.3445  SW: 2.2e-16 | normal(31.49282, 3.227443)  KS: 2.2e-16  SW: 1.414e-07 |
| **4hist** | **uniform(31.45979, 62.78185)**  **KS: 0.3445**  **SW: 2.2e-16** | **(not)normal(31.49282, 3.227443)**  **KS: 2.2e-16 (repeating values shouldn’t be present)**  **SW: 1.414e-07** |
| 5 | uniform(31.45979,62.78185)  KS:0.3445  SW: 2.2e-16 | normal(31.49282, 3.227443)  KS:2.2e-16  SW:1.414e-07 |
| 6 | uniform(31.45979, 62.78185)  KS: 0.3445  SW: 2.2e-16 | normal(31.49282, 3.227443)  KS: 2.2e-16  SW: 1.414e-07 |
| 7 | uniform(31.45979, 62.78185)  KS 0.3445  SW 2.2e-16 | Normal(31.49282, 3.227443)  KS 2.2e-16  SW 1.414e-07 |
| **8** | **uniform(31.45979, 62.78185)**  **KS: 0.3445**  **SW: 2.2e-16** | **Not normal and not uniform**  **KS:2.2e-16**  **SW: 1.414e-07** |

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

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**

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)

What % of the points are above 7?

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

**[1] 0.9194684**

**92%**

What % of the points are below 6?

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

**[1] 0.02965435**

**3%**

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)%

Uniform distribution - qunif and punif

Other distributions qx px where x is the distribution

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 rsshatudio

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

Find distributions:

|  |  |  |  |
| --- | --- | --- | --- |
| Group | Combined | Male | Female |
| 1 | Can’t Calculate | normal(69.04298,2.866323) | normal(63.64935,2.724203) |
| 2 | **Neither normal or uniform** | normal(69.04298,2.866323) | normal(63.64935, 2.74203) |
| 3 | Neither normal or uniform | normal(69.04298,2.866323) | normal(63.64935,2.2724203) |
| 4 | **Neither normal nor uniform** | normal(69.04298, 2.866323) | normal(63.64935, 2.724203) |
| 5 | **Not Normal & Not Uniform** | normal(69.04298,2.866323) | normal(63.64935,2.724203) |
| 6 | Can’t calculate | normal(69.04298,2.866323) | normal(63.64935,2.724203) |
| 7 | Not Normal | normal(69.04298,2.866323) | Normal(63.64935, 2.724203) |
| 8 | Can’t calculate | normal(69.04298, 2.866323) | normal(63.64935, 2.724203) |

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 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)

|  |  |  |  |
| --- | --- | --- | --- |
| Group | Q1 | Q4 | Q7 |
| 1 | 17% | 0.3644737 (37%) | 71.32467 |
| 2 |  |  |  |
| 3 | 0.1734655 (17%) | 0.3644737 (36%) | 71.32467 |
| 4 | 17% | 37% | 71.32467 |
| 5 | **Can’t Calculate** | **Can’t Calculate** | **Can’t Calculate** |
| 6 | **Can’t calculate** | **Can’t calculate** | **Can’t calculate** |
| 7 | 0.1734655 |  |  |
| 8 | Can’t calculate | Can’t calculate | Can’t calculate |

|  |  |  |  |
| --- | --- | --- | --- |
| Group | Q2 | Q5 | Q8 |
| 1 | 37% | 7.9% | 72.71632 |
| 2 |  |  |  |
| 3 | 36.9% | 7.9% | 72.71632 |
| 4 | 37% | 8% | 72.71632 |
| 5 | 37% | 8% | 72.71632 |
| 6 | 36.9% | 7.9% | 72.7” |
| 7 |  |  |  |
| 8 | 36.9234% | 7.919419% | 72.71632 |

|  |  |  |  |
| --- | --- | --- | --- |
| Group | Q3 | Q6 | Q9 |
| 1 | 0.009871518 (1%) | 0.6899811 (69%) | 67.14056 |
| 2 |  |  |  |
| 3 | 0.9% | 68.9% |  |
| 4 | 1% | 69% | 67.14056 |
| 5 | 1% | 69% | 67.14056 |
| 6 | 0.9% | 69% | 67.1” |
| 7 |  |  |  |
| 8 | 0.9871518% | 68.99811% | 67.14056 |

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

**Next Sessions is Asynchronous**

**Flow is the same, but there is not synchronous sessions → Group work becomes individual**