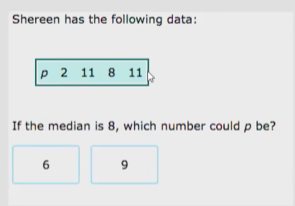
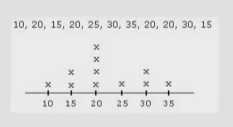
**KAHOOT İLE ÖNCEKİ KONULARIN ÖZETİ**

Mean outlierlara karşı sensitivedir.

Soru: eğer median 8 ise p hangi rakam olmalıdır? 

Cevap: Sort edersek –> 2, 8, 11, 11… Eğer 9 olursa 8 iyice sola kayacak ve ortada 9 kalacak. 6 olmalı ki 8 ortadaki rakam olsun.

Soru : Mode burada kaçtır? 

Cevap : Mode en fazla tekrar eden eleman sayısı idi. Bu durumda mode, 20’dir.,

Soru : Ne zaman mean ile median aynı olur?

Cevap : When the distribution is symmetric. Hatta bu durumda mode da eşit olur.

Soru : ……….is a way to explain how a dataset is distributed.

Cevap : dispersion

**Central tendency** de bir tane metrik değer bir datanın dağılımının gösterirken (ya meandir ya mediandır ya mode dur) ama dağımımızın şekli saçılımı farklı olduğu halde mean ve median aynı olduğu durumlar oluyordu. O zaman bu iki datasetini birbirinden nasıl birbirinden ayırt edebileceğim? O zaman dipersion dediğimiz özellikler imdadımıza yetişiyor.

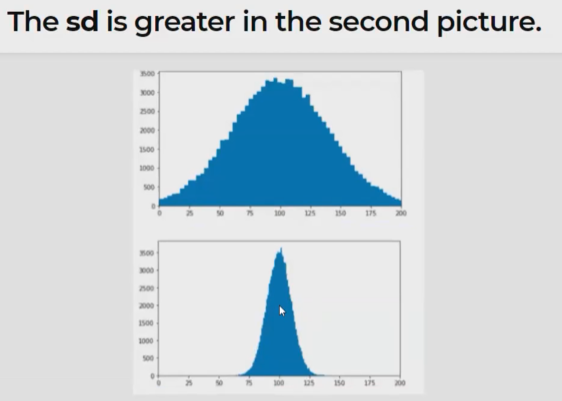
Bu özellikler: Range, Standard Deviation ve IQR.

**Range** dediğimiz şey minimum ve maksimum değerler sözkonusu oluyor. Elimdeki datasetimin minimum ve maksimum aldığı değerlere bakıyorum. Demek ki bu aralıkta değişen değerler var elimde diyebiliyorum.

**Standard Deviation**’da bu dağılımın mean’ini hesap ediyordum. Data setiminin her bir elemanı mean’den ne kadar uzaklaşmış ona bakıyorum. Ama ne kadar uzaklaştığına sadece farktan bakarsak olmaz. Elemanların karelerini alıyor odan devam ediyorduk. Onları topluyor ve eleman sayısı-1’e bölüyorduk sample tarafında. Ama popülasyonda normal n’e (eleman sayısına) bölüyorduk.

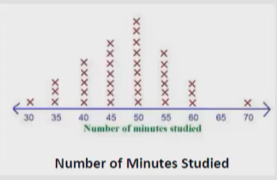
**IQR**: Interquartile Range bize biraz daha farklı birşeyler söylüyor. Datasetimi quartile lara bölüp (4’te birlere bölüp) her bir 4 te bir içerisine hangi değerlerin düştüğünü görüp, burdan dışarda outlierlarımın kalıp kalmadığını görüp bana data setim hakkında önemli şeyler söylüyor.

Histogram ile IQR arasındaki fark neydi? His. 2 boyutluydu frekansı veriyordu ve alt tarafta interval’ları veriyordu ama IQR ise tamamen bir summarizationdur. Elimizdeki data setinin özetlenmiş rakamlarını ifade eder.

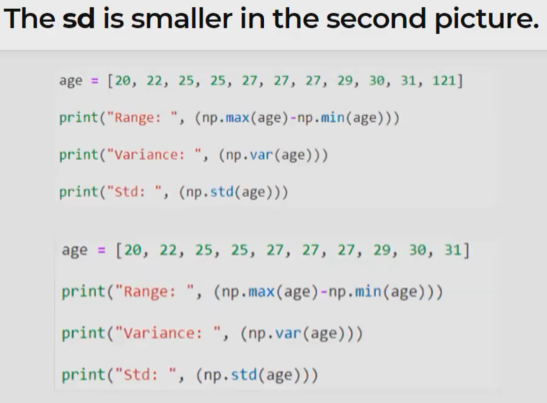
Soru : 

Cevap : False.. Alttaki resimde standart deviation küçüktür. Yani dağılım azdır.

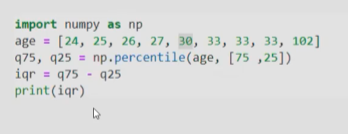
Standart deviation’ın yol arkadaşı meandir. Her elemanın mean’den uzaklaşma durumunu verir. Alttaki resimde elemanların mean’ etrafında toplandığını görüyoruz.

Soru : What is the range of this data? 

Cevap : Max ile min arasdındaki farka bakacağız. 70-30 = 40

Soru: True/False ? 

Cevap : True. İlk resimde sağda büyük bir rakam var ve o rakam dağılımı arttırırıyor. Dolayısıyla sd de artıyor.

Soru: what is the output? 

Cevap : median 30. Sol tarafın medianı yani Q1🡪 26, sağ tarafın medianı yani Q3🡪 33.

IQR : Q3-Q1 = 7

Diyelim ortada 30’dan 2 tane olsaydı. 🡪 24, 25, 26, 27, 30, 30, 33, 33, 33, 102

O zaman soldaki Q1 i hesaplarken soldaki 30’u baz alacaktım. Sağdaki Q3 ü hesaplarken de sağdaki 30’u baz alacaktım.

**Temel noksasında olay şöyle:**

Elimize gelen datanın hangi kaynaktan geldiğini tam olarak bilemeyebiliyoruz. Bazen arcaical ??? arşiv verilerinden olabilir, dijital verilerden olabilir, bazen anket verilerinden olabilir. Türlü türlü kaynaklardan bu verileri elde edebiliriz. Mümkünse bu verinin toplanma şekliyle alakalı, her verinin bir data dictionarysi vardır, bir hikayesi vardır. Bu hikayeyi okumam lazım. Bu datanın hangi tekniklerle toplandığına dair bir bilgi vardır, buna dair hangi sample teknikleri ile toplandığına dair bir bilgi vardır. Bu bilgileri toparlamak gerekiyor.

Ama kendiniz bir data toplamak durumunda kaldığınız zaman da eğer üzerinde ciddi bir hipotez testi yapacaksanız mutlaka **probability sampling** çalışmalısınız. Eğer hızlı, **quick and dirty** çalışma yapacaksanız, kendine göre analitik olarak yani daha önceden **pre-defind** bazı kurallara göre, beni tatmin edecek kadar, amacıma matuf bir sampling yapacaksam orada bir probability sampling yapacaksam tabiri caiz ise ey data üyeleri sen sen sen sen gel yola çıkıyoruz dercesine bir non-probability data setini elinize alır ve devam edersiniz.

**Probability & Non-probability sampling**

**Non-probabilty** data setini daha çok **qualitatif** çalışmalarda tercih edersiniz. Daha az bir sampling üzerinde çalışmanız gerekiyorsa zaten kalitatif çalışmaların çoğu az samplingle olur arkadaşlar. Ama çok **büyük bir veri ile çalışacaksanız**, çok büyük bir popülasyonla çok önemli tamamen random ve her bir üyeye **eşit sans** (equal chance) verecek şekilde çalışacak ve elde ettiğiniz datadan da mutlaka **hipotez testleri** yapacaksanız, üzerinde bir **machine learning** yapacaksanız o zaman **probability sampling** üzerinden gitmekte fayda var arkadaşlar.

Probability sampling dediğimiz zaman mutlaka aklımıza **randomization** gelsin non-probability dediğimiz zaman aklımıza **arbitrary**, herhangi bir randomization yok.

Probability sampling dediğimiz zaman aklımıza **equal chance** gelsin, non-probability dediğimiz zaman **equal chance yok**. Popülasyonun çok önemli kısımlarından çok önemli verileri alamayabilirsin.

Probability sampling dediğimiz zaman aklımıza **ambient (ortam)** gelsin, non-probality dediğimiz zaman mutlaka **diest** olsun ???? çünkü zaten subjectif olarak daha önce belirlediğiniz kriterlere göre o samplingi yapmak durumunda kalıyorsunuz. Dolayısıyla analist olarak oraya mutlaka bir diestiniz karışmış oluyor.

Probability sampling dediğimiz zaman bunun bir statistical process olduğunu bilmeniz gerekir ama non-probability sampling dediğimiz zaman tamamen sizin analitik zekanıza yönelik daha önceden belirlediğiniz analitik kurallarınıza yönelik prosedürlerin olduğunu bilmelisiniz.

Probality samplingte objektif dedik ama no-probabilty samplingte sübjektif dedik.

Probability samplingin en önemli kavramlarından bir tanesi representive olması yani popülasyonu tam temsil ya da tama yakın kıvamda temsil edebileceğini biliyoruz. Ama non-probability de bu söz konusu değil.

***(quick and dirty:***

*describing a research design or data analysis that is admitted to be informal and imperfect. The researcher who adopts such an approach recognizes that there is a more refined way to set up the experiment or to analyze the findings. In research, a quick-and-dirty approach is more informal than a*[*pilot study*](https://dictionary.apa.org/pilot-study)*; in data analysis, it is regarded as a first peek at data from a specific analysis.)*

***(What is non-probability sampling?***

***Definition:****Non-probability sampling is defined as a sampling technique in which the researcher selects samples based on the subjective judgment of the researcher rather than random selection. It is a less stringent method. This sampling method depends heavily on the expertise of the researchers. It is carried out by observation, and researchers use it widely for* ***qualitative research.***

*Non-probability sampling is a sampling method in which not all members of the population have an equal chance of participating in the study, unlike probability sampling. Each member of the population has a known chance of being selected. Non-probability sampling is most useful for exploratory studies like a pilot survey (deploying a survey to a smaller sample compared to pre-determined sample size). Researchers use this method in studies where it is impossible to draw random probability sampling due to time or cost considerations****)***

**Stratified sampling ve cluster sampling**

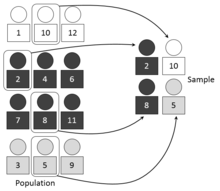
Cluster sampling dediğimiz zaman aklınıza lokasyon gelsin. Bu lokasyondan kastettiğim şey şehirler olabilir, bu şehirin içirisinde dağılmış olan hastaneler olabilir, okullar olabilir

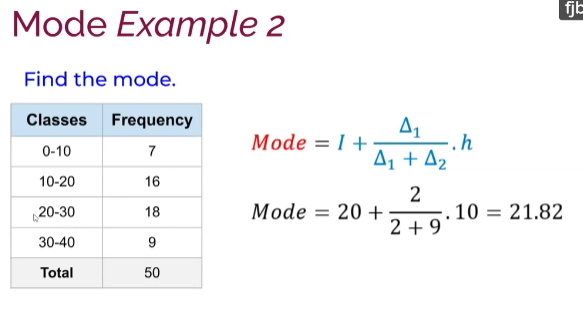
Farketmez. Ve gruplar arası homojen benzerlik varsa ben buna cluster sampling uygulayabilirim. Yani okulların hepsinin aynı özellikte olduğunu varsayıyorum. Şehirlerin hepsinin aynı özellikte olduğunu varsayıyorum. İşte burda cluster sampling kullanabilirim.

Fakat stratify sampling dediğimiz zaman ise grubun içinde homojenlik varsa yani sadece kadınlardan bahsediyorsam, sadece erkeklerden bahsediyorsam, (sadece üniversitelilerden bahsediyorsam olmaz çünkü içinde kadın ve erkek var ama onun alt seviyesinde sadece erkekler üzerinde çalışacak olursam orda yine stratify dan bahsedebilirim.) bakıldığında eğer datanızda herhangi bir stra… olduğunu bilitorsanız, bu cinsiye olabilir ya da çalışan çalışmayan şeklinde de olabilir hiç farketmez, yaş gruplarını ayırmışsınızdır orda da stratify yapmış olabilirsiniz.

(In [statistics](https://en.wikipedia.org/wiki/Statistics" \o "Collectively exhaustive events), **stratified sampling** is a method of [sampling](https://en.wikipedia.org/wiki/Sampling_(statistics)" \o "Statistics) from a [population](https://en.wikipedia.org/wiki/Population_(statistics)" \o "Population (statistics)) which can be [partitioned](https://en.wikipedia.org/wiki/Partition_of_a_set) into [subpopulations](https://en.wikipedia.org/wiki/Subpopulation" \o "Subpopulation).

In [statistical surveys](https://en.wikipedia.org/wiki/Statistical_survey), when subpopulations within an overall population vary, it could be advantageous to sample each subpopulation (stratum) independently. **Stratification** is the process of dividing members of the population into homogeneous subgroups before sampling. The strata should define a partition of the population. That is, it should be **[collectively exhaustive](https://en.wikipedia.org/wiki/Collectively_exhaustive_events" \o "Statistical survey)** and **[mutually exclusive](https://en.wikipedia.org/wiki/Mutual_exclusivity" \o "Mutual exclusivity)**: every element in the population must be assigned to one and only one stratum. Then [simple random sampling](https://en.wikipedia.org/wiki/Simple_random_sampling) is applied within each stratum. The objective is to improve the precision of the sample by reducing [sampling error](https://en.wikipedia.org/wiki/Sampling_error" \o "Sampling error). It can produce a [weighted mean](https://en.wikipedia.org/wiki/Weighted_mean" \o "Weighted mean) that has less variability than the [arithmetic mean](https://en.wikipedia.org/wiki/Arithmetic_mean" \o "Arithmetic mean) of a [simple random sample](https://en.wikipedia.org/wiki/Simple_random_sample" \o "Partition of a set) of the population.)

[](https://en.wikipedia.org/wiki/File:Stratified_sampling.PNG)



I = class’lar arasındaki orta satırdaki (20-30 satırı) küçük olan değer 🡪 20

Delta 1 = en yüksek frekastan bir önceki frekans değerini çıkarınca bulunan sonuç 18-16 = 2

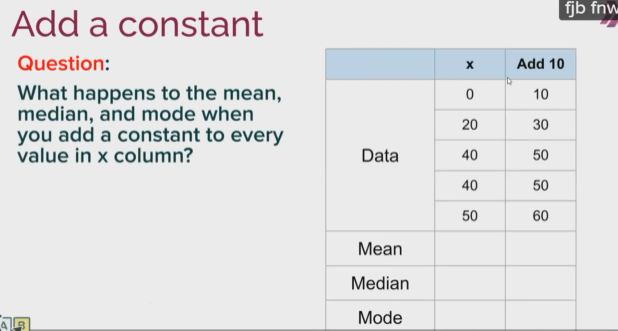
Delta 2 = en yüksek frekans ile bir sonraki frekans arasında fark 18 – 9 = 9

\*10 🡪 her bir interval’ın genişliği.

Hiç hesaplama yapmadan sırf tabloya bakarak aşağı yukarı 20-22 sonucunu söyleyebilirdim. Çünkü mode elimdeki datasetimdeki en çok tekrar eden veridir. Bir önceki class ta 16 var, bir sonrakinde ise yalnızca 9. Dolayısıyla 20 ile 30 arasında olması gereken mode değerinin (16 dan dolayı yani ağırlık aşağı yönde olduğundan ) alt sınıra yakın olması gerektiğini söyleyebilirim.

Küçük bir datasetim varsa mode çok stabil bir değer değildir. Ona ekstra bir değer eklediğinde hemen mode da bir kayma olabilir. Bir de elimizdeki datasetimizde tekrar eden value lar yoksa orda da mode dan bahsedemeyiz. Bir de datasetinde iki ayrı değer aynı sayıda tekrar ediyorsa her iki değer de mode dur. Mesela aynı sayıda 8 ve 9 varsa mode, hem 8 hem de 9 dur.

Örnek: bu sayılara bir değer eklediğimde mean, median ve mode nasıl eklenir?



Median = 40, mean = 150/5 = 30 , mode = 40

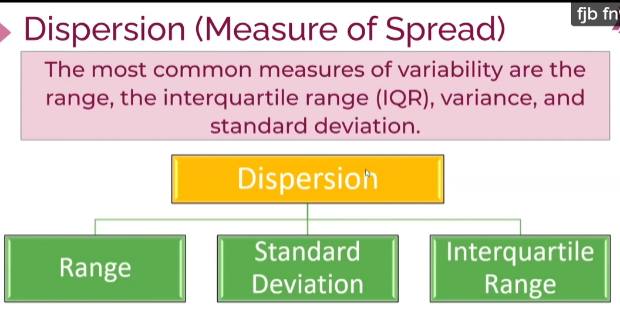
Median ortadaki değerdir. Tüm verilere aynı sayıyı eklersem ortadaki değer değişemeyeceğinden kaç eklersem median = median + n olacaktır. 10 eklersem 50, 10 eklersem 60 gibi. Mode da aynı şekilde olur. Mean de o eklediğim sayı nispetinde artar.

Peki datasetimdeki bu skorlarımı sabit bir sayıyla mesela 10 la çarparsam;

Mean’im, medianı’ım ve mode’um 10 katı artar.

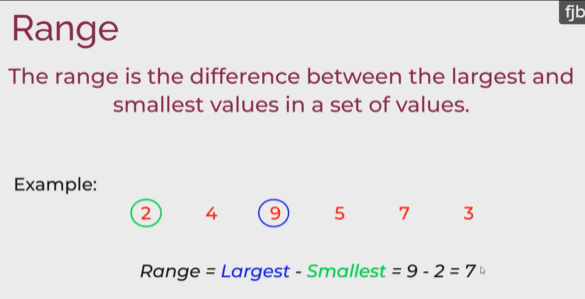
**DISPERSION**

**Range, Standard Deviation, Interquartile Range**



Mean, median ve mode aynı ise yani değelerim simetrik ise grafiğimin basıklık seviyelerinde datasetimin değerleri aynıdır diyebilir miyiz? Hayır. Bu durumda farklı değerler de bakmamız gerekiyor. Bu durumda **Range, Standard Deviation, Interquartile Range** önem arz ediyor.

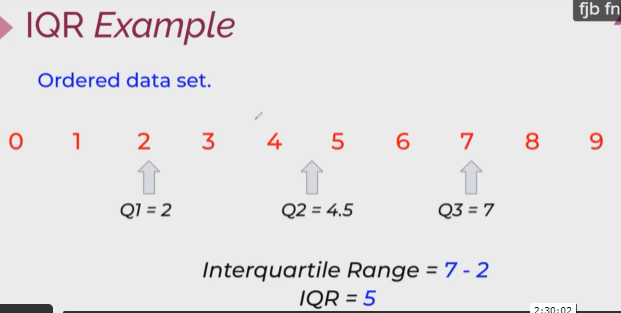
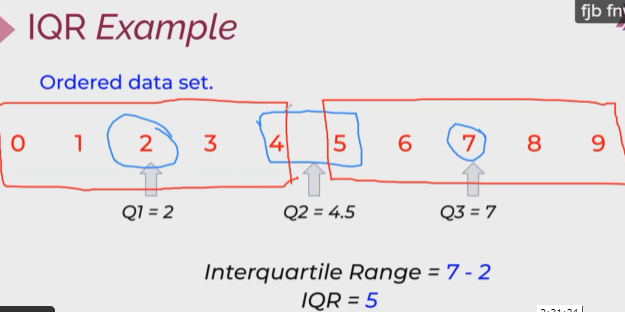
Aşağıdaki örnekte min ve max arasındaki farkı biliyorum (9-2 = 7), dolayısıyla değerlerimin hangi aralıkta dağıldığını bu range’e bakara görebiliyorum ama saflar nerelerde sıklaşmış nerelerde gevşek, hangi değerim meanden daha kadarta ya da yakında bundan bir fikrim yok. Dolayısıyla bu datanın künyesini çıkartmam için bu değerler yeterli değil.



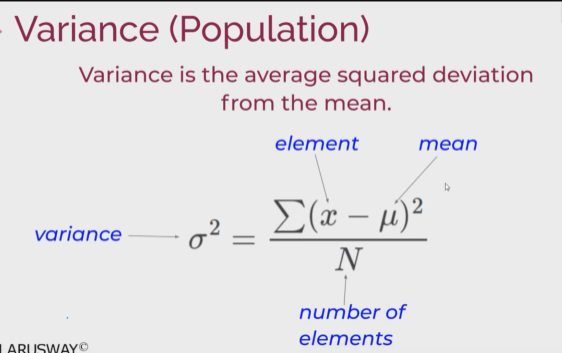
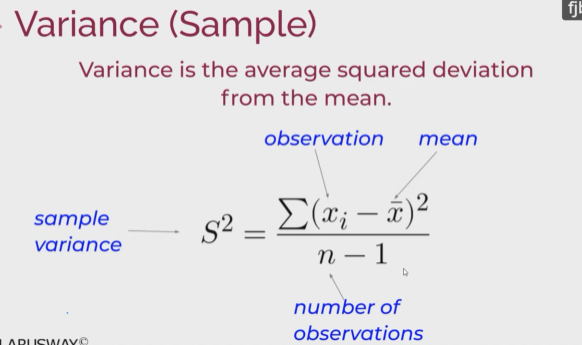
Q2 = Tüm datasetimin median’ı

Q1 = Q2 nin altında kalan kısmının median’ı ( Q2 - min. arasındaki orta değer değer)

Q3 = Q2 nin üstünde kalan kısmının median’ı ( Q2 - max. arasındaki orta değer değer)

**Variance (Population)**

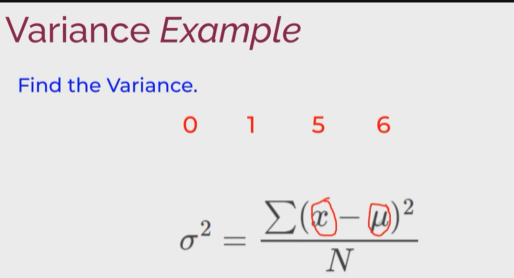
**** ****

Yukarda anlatılanlar ile ben datamın hangi aralıklarda dağıldığını gördüm, mediana göre sağa kayıklığını veya sola kayıklığını görüp skewnessı hakkında da fikir elde ediyorum ama ben halen daha data setimin içersinde neler değiştinin nasıl değiştiğini, mean’e göre datamın nasıl değer aldığı hakkında bir fikrim yok. Buna variability deniyor. Ben variabilitimi bunlarla anlayamıyorum. Datamın içindeki variability i anlayabilmem için variance diye bir metrik değere ihtiyacım var.

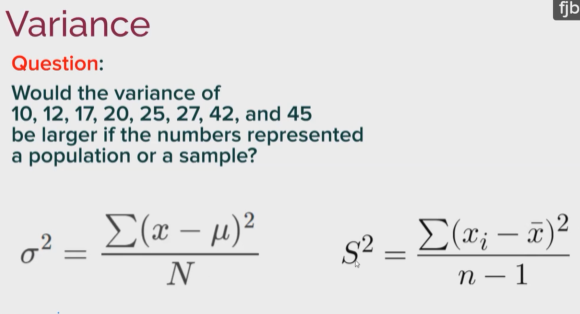
Variance= ortalama değerle datasetimdeki her elemanın arasındaki farkın karelerinin toplamının datasetimin içindeki skorların sayısına bölümüne denir.

Sample’ın variance’ı –> küçük x, popülasyonun variance’ı 🡪 büyük X

Sample variace hesaplarken n’yi 1 eksiltiyoruz. Yani variance değerini büyütmüş oluyorum. Yani mantık olarak, değişkenin içerisindeki varyansın (variability) hidden olmuş saklanmış kıyıda köşede kalmış varyansı da çıkartmış oluyorum.



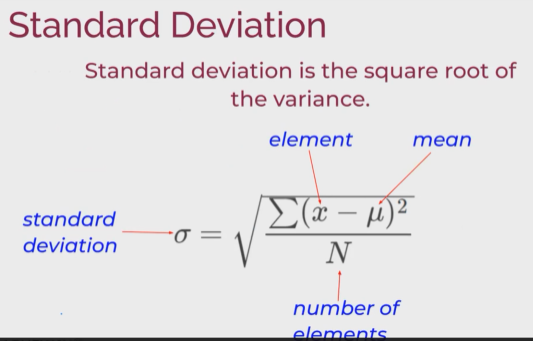
Her bir elemanın mean’den farkına bakıyorum. Elemanlar mean’in sağında ve solunda kümelenmiş olacağından bu farkların toplamı da 0 çıkacak ve anlam ifade etmeyecektir. Anlamlı bir şey bulabilmek için bu farkların karesini alıp öyle N’e bölüyorum.

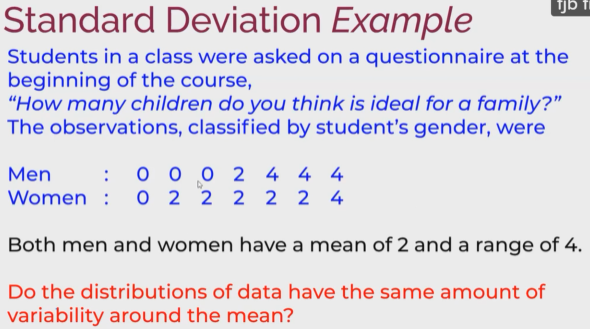
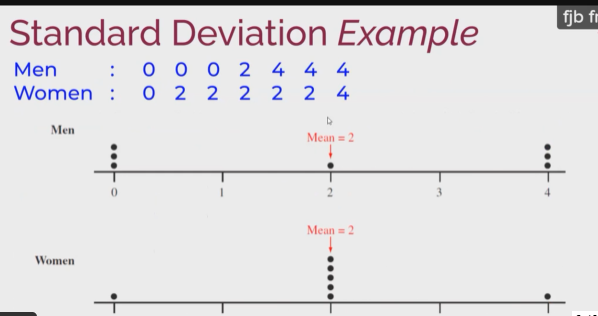


Sample’ın varyansı popülasyonun varyansından her zaman büyüktür. Neden? Çünkü formülde (n-1 yaparak) paydayı küçültüyorum.

**Standart Deviation:**

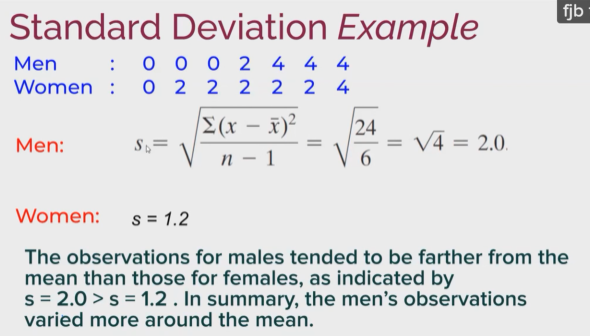
Varyansın kareköküdür.

****

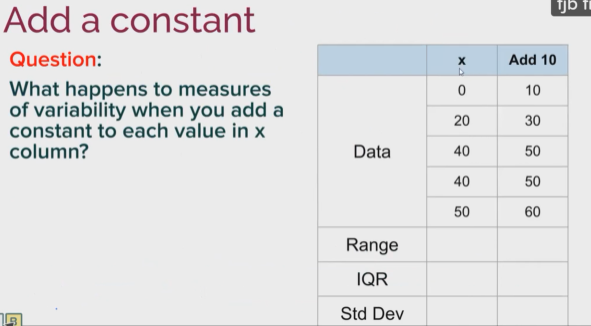
 

Bu sorudaki datasetime baktığımda şunu anlıyorum. Erkekler kendi içlerinde ayrışmış. Bir tarafta çok çocuk isteyenler var, bir tarafta çocuk istemeyenler de var. Ama kadınlar ekseritle 2 çocuğun ideal olduğunu düşünüyorlar.

Ben datasetime hiç bakmadan sağdaki görsele bakara da kaç erkeğin çocuk istemediğini, kaçının 3 çocuk istediğini ve bu şekilde ayrıştıklarını, kadınların da 2 çocukta birleştiklerini anlayabiliyorum.



Standard deviation’a bakarsak erkeklerin gözlemlerinin ortalamanın etrafında daha fazla değiştiğini görebiliyorum



Datasetimin elemanlarına sabit 10 eklersem bu değerler nasıl değişir?

Range : maksimum değer – minimum değer olduğu için değişmeyecektir.

Q2 = 40, Q1=20, Q3 = 40

IQR = Q3-Q1 = 20. Her değere 10 eklersem IQR’da değişmeyecektir.

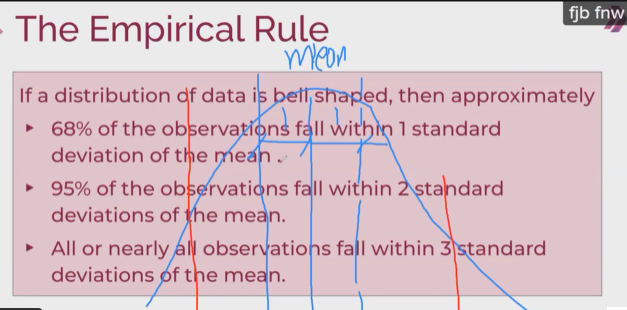
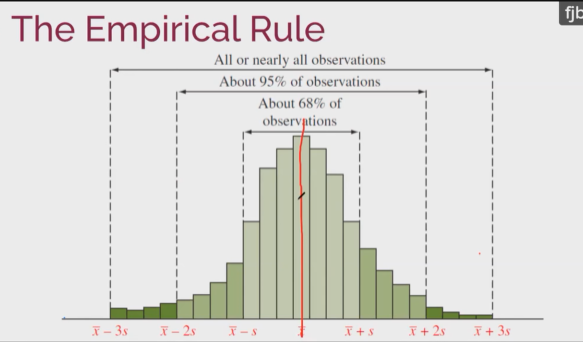
Standard deviation da elemanlara bağlı olduğu için ve elemanların hepsi sabit arttığı için değişmeyecektir.

Peki her elemanı 10 ile çarparsam değeler nasıl değişir?

Range artacak şekilde değişecektir. Çünkü max değerim 500, min değerim 0 olduğu için Range = 500 olacaktır.

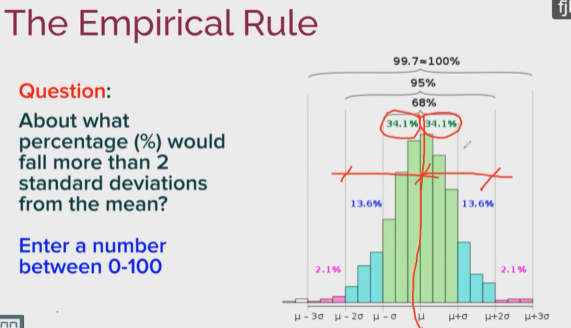
IQR da da aynı şekilde Q3 ve Q1 değeleri 10ar kat artacağı için 350 olarak artmış olacaktır.

Standart dev. de aynı şekilde artacaktır.

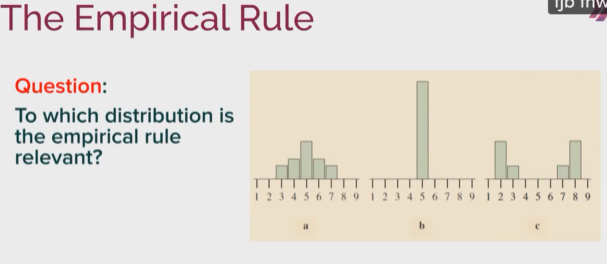
 

Normal bir dağılımda standart sapmadan sağa ve sola 1 uzaklaştığımda arada kalan alan toplam datamın %68’idir.

Standart sapmadan iki yana doğru 2 uzaklaştığımda toplam datanın %95 ini kapsamış oluyorum. Bu değer 3 olduğunda neredeyse tamamına yakınını kapsamış oluyorum.



Hangi distribution da emrical rule çalışır?



Birinci grafikta çalışır.

**Elimdeki veri böyle sağa kayıksa ne yapmam lazım**. Bu veriyle mutlaka çalışmam gerekiyor.

Bu durumda elimdeki kayık olan veride empirical rule’u uygulayıp geçerim diyemezsin. Eğer elinizdeki veri kayıksa benim bunu normal dağılıma bir şekilde benzetmem gerekiyor. Matematiksel işlem yaparım. Ya ½ sini alırım bazen logaritmasını alırım bazen karesini alırım. Bazı durumlarda datesinden veri atarak bu kayıklığı hallederim bazen de dönüşüm yaparak trasformation yaparak yaparım. Buna normalization denir. **Python arka planda bu işlemi yapıyor ve normale benzetiyor ve normale dönmüş halini machine learning’e sokuyor.**

Bu şekilde normale benzetip empirical rule u işletirim. Fakat bu sefer ne oluyor? İşlettiğiniz şey o değerin karelerinden, kareköklerinden, logaritmalarından, ln lerinden oluşacağı için bu sefer bunu sizin anlamlandırmanız biraz daha zorlaşacaktır. Bunu da aklınızın bir köşesinde bulundurun.

