SAKARYA ÜNİVERSİTESİ

Olasılık ve İstatistik
HAFTA 11
Tanımlayıcı İstatistik (devam)

Dr. Öğretim Üyesi Burcu ÇARKLI YAVUZ

bcarkli@sakarya.edu.tr

Merkezi Eğilim Ölçütleri

Geometrik ortalama

- Örnek sayısının az olduğu durumlarda uç değerlerin minimize edilmesi adına aritmetik ortalama ölçütü yerine, Geometrik ortalama tercih edilir.
- Geometrik ortalama kısaca “n” tane değerin birbiri ile çarpımının n. dereceden kökü olarak tanımlanır ve aşağıdaki formül ile hesaplanır.

$$G = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

Örnek

- Bir sınav için final sonuçları aşağıda verilmiştir. Geometrik ve Aritmetik ortalamayı hesaplayınız.

45	37	40	30	35	45	50	95

Çözüm

$$\mu = \frac{45 + 37 + 40 + 30 + 35 + 45 + 50 + 95}{8} = 47{,}13$$

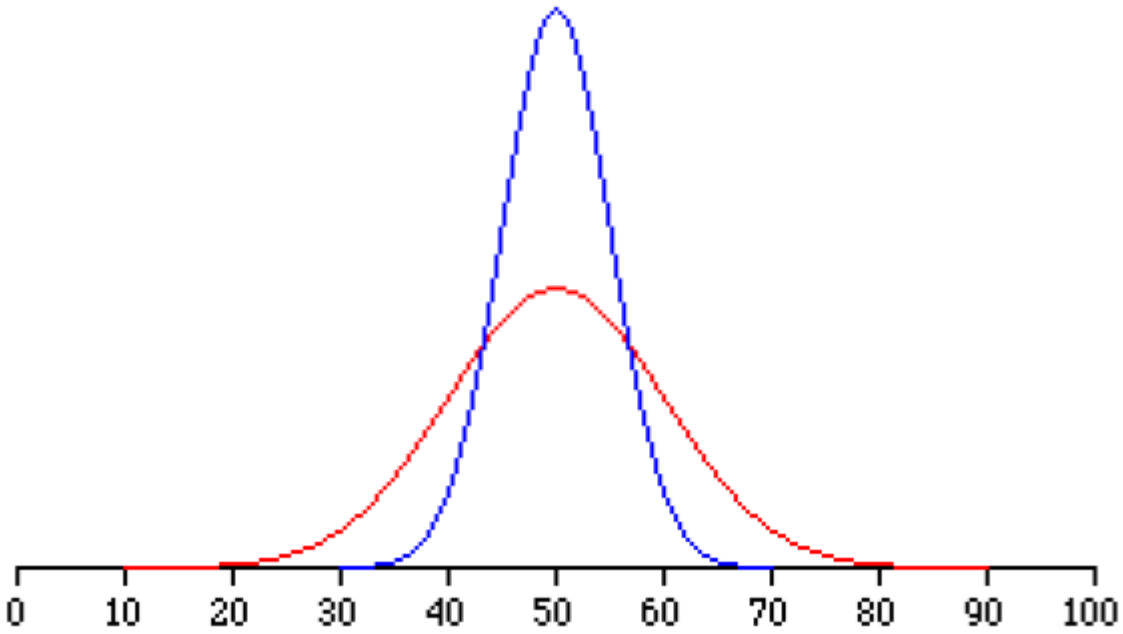
$$G = \sqrt[8]{45 * 37 * 40 * 30 * 35 * 45 * 50 * 95} = 44{,}34$$

- 8. değer olan 95 uç değerdir.
- Bu değer aritmetik ortalamayı daha fazla etkilemekte, veri setindeki 8 değerin 6 sı ortalamanın altında kalmaktadır.
- Geometrik ortalama ise daha az etkilenmekte ve veri setinin bir yarısı Geometrik ortalamadan büyük iken, diğer yarısı daha düşüktür.

Değişkenlik Ölçütleri

- Merkezi eğilim ölçütleri dağılım hakkında bilgi vermez. Bir veri setinin ortalamasının ne olduğu kadar, verilerin bu ortalama etrafında nasıl değişkenlik gösterdiğinin de bilinmesi önemlidir.

- Örnekte mavi ve kırmızı sınıfların bir dersten aldığı ortalamalar aynı olmakla beraber, farkı değişkenlikleri olduğu görsel olarak söylenebilir.



Değişkenlik Ölçütleri

- Örnekte simetrik olan iki dağılımın ortalaması, medyanı ve modu aynı ve birbirine eşittir.
- Sadece merkezi eğilim ölçütleri ile bu örneklemler ile ilgili yorum yaptığımızda, aynı veri setinden bahsedildiği sonucuna dahi ulaşmak olasıdır.
- Bu şekildeki durumlarda veri dağılımlarını tanımlayan değişkenlik ölçütlerine (Değişim aralığı, varyans, standart sapma, değişkenlik katsayısı) ihtiyaç duyulmaktadır.

Değişkenlik Ölçütleri

Değişim aralığı

- En basit değişkenlik ölçütü olan değişim aralığı örneklemin en küçük değeri ile en büyük değerinin farkından hesaplanır.
- En önemli avantajı kolayca hesaplanabilmesi ve yorumlanabilmesi olmakla birlikte, bu basitlik yeterli bilgiyi içermeme şeklinde karşımıza çıktığından diğer bir taraftan önemli bir dezavantajdır.

Değişim aralığı

- Değişim aralığı bazı durumlarda yetersiz kalmaktadır.
- Birbirinden tamamen farklı dağılıma sahip iki veri setinin en küçük ve en büyük değerleri birbirine eşit olabilir.
- Bu iki örneklemi karşılaştırmak sadece değişim aralığı ölçütü ile mümkün değildir.

Veri Seti 1:	4, 4, 4, 4, 4, 50
Değişim Aralığı:	50 – 4 = 46

Veri Seti 2:	4, 8, 15, 24, 39, 50
Değişim Aralığı:	50 – 4 = 46

Değişkenlik Ölçütleri

Varyans

- Bir veri setindeki her bir değerin ortalamadan uzaklıklarının karelerinin ortalaması şeklinde hesaplanır.
- Varyans beklenen değer ile (Bütçe) gözlenen değer (Harcama) arasındaki farktır. Yapılması gereken ile yapılan arasındaki farktır.

$$\text{Ana Kütle Varyansı: } \sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

$$\text{Örneklem Varyansı: } s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Örnek

- 6 mezun tarafından ortalama yapılan iş başvurusu sayısı aşağıda verilmiştir.
- Varyansı hesaplayınız.

17	15	23	7	9	13

Çözüm

- Varyansı hesaplayabilmek için öncelikle örneklem ortalaması hesaplanmalıdır (Tüm mezunların değil, 6 mezunun bilgisi verilmiş). Sonra her bir değerin ortalamadan uzaklıklarının karesi alınarak, toplam örnek sayısının bir eksiğine bölünmelidir.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{17 + 15 + 23 + 7 + 9 + 13}{6} = 14$$

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} = \frac{(17-14)^2 + (15-14)^2 + \cdots + (13-14)^2}{6-1}$$

$$= \frac{166}{5} = 33.2$$

- Bu 6 kişilik örneklem için toplam değişkenlik 33,2 birim olarak bulunur.

Varyans

- Varyansın yorumlanması tek bir örneklem için çoğu zaman kolay değildir.
- Verilerin değişkenliklerinin (negatif veya pozitif yönde) mutlak olarak belirlenebilmesi adına, ortalamadan uzaklaşmaların karelerinin alınması çoğu zaman bu değeri yorumlayanlar için soruna yol açmaktadır.
- Varyans birden fazla değişken için karşılaştırılmalı olarak daha kolay yorumlanabilen bir ölçüttür.

Değişkenlik Ölçütleri

Standart Sapma

- Ortalama veya beklenen değerden ne ölçüde sapma olduğunu gösterir.
- Düşük standart sapma değerleri verilerin ortalamaya daha yakın seyrettiğini gösterir. Yüksek değerlerde ise veriler o kadar ortalamadan uzaklaşır.
- Standart sapma, varyans değerinin karekökü alınarak hesaplanır. En sık kullanılan değişkenlik ölçütüdür.

$$Ana\ Kütle\ Standart\ Sapması$$

$$\sigma = \sqrt{\sigma^2}$$

$$Örneklem\ Standart\ Sapması$$

$$s = \sqrt{s^2}$$

Örnek

- 6 mezun tarafından ortalama yapılan iş başvurusu sayısı aşağıda verilmiştir. Standart sapmayı hesaplayınız.

17	15	23	7	9	13

Çözüm

- Standart sapma, varyans değerinin karekökü olduğundan önceki çözümde hesaplanan varyans değerinin karekökü alınır.

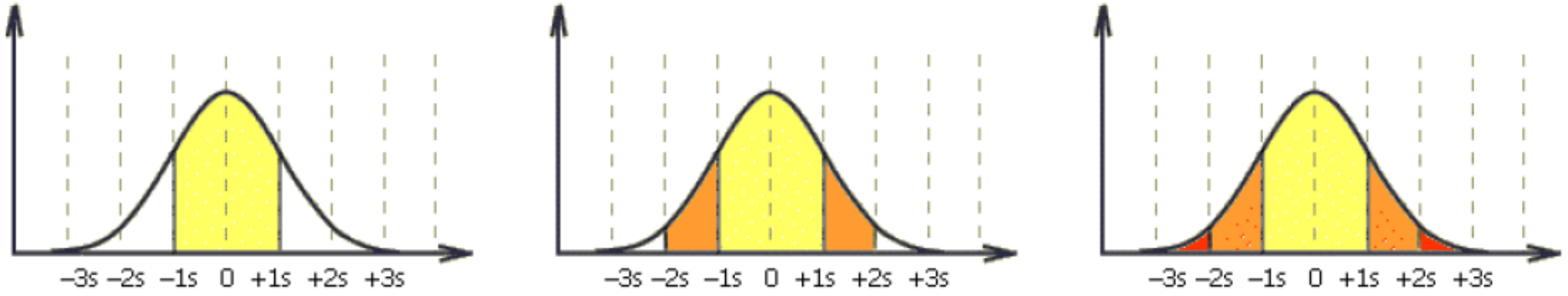
$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = 33.2 \qquad s = \sqrt{s^2} = \sqrt{33.2} = 5.8$$

Standart Sapma

- Standart sapma örneklemdeki değerler ile aynı aralıkta olduğundan (ortalamadan, farklarının karesinin karekökünden hesaplandığı için) yorumlanması oldukça kolaydır.
- Örneğin yukarıdaki örnekte verilerin önemli bir kısmının ortalama olan 14 değerinin 5,8 üstünde ve 5,8 altında olması beklenir.
- Yani verilerin büyük çoğunluğu 8,2 ile 19,8 arasındadır. (Bizim örneğimizde 6 örneğin 4 tanesi yani %66,7 sı)

Standart Sapma

- Yukarıdaki örnek için yaptığımız yorum pratikte çok değerlidir. Çünkü pratikte veri setlerinin çoğunun dağılımı simetrik sürekli bir dağılım olan normal dağılıma uyar ve bu dağılım için en önemli açıklayıcı parametre standart sapmadır.



Standart Sapma

- Normal dağılıma sahip veri setlerinde, bütün verilerin %68 i ortalamadan tek standart sapma uzaklıkla yer alır. Yani verilerin 2/3 ünden fazlası $\mu \pm s$ aralığındadır. Verilerin büyük çoğunluğu (%95) $\mu \pm 2s$ aralığında iken, neredeyse bütün veriler (%99,7) $\mu \pm 3s$ aralığında yer alırlar.
- Örneğin 100 kişilik bir sınıftaki öğrencilerin matematik dersinde aldıkları notların ortalaması 47 ve standart sapması 11 olan normal dağılıma uyuyor ise, bu sınıftaki öğrencileri 68 tanesinin 36 ile 58 arasında, 95 tanesinin ise 25 ile 69 arasında not alması beklenir. Öğrencilerin hemen hemen hepsi (99 tanesi) ise 14 ila 80 arası not alması öngörülebilir.
- Standart sapma sağladığı pratik yorumlama avantajlarından dolayı hemen hemen her sektörde yoğun bir kullanım alanına sahiptir. Varyansa göre daha rahat anlaşılabilir. Tek değişken için yorumlanabileceği gibi, birden fazla değişken için karşılaştırma durumlarında da rahatlıkla yorumlanabilir.

Değişkenlik Ölçütleri

Değişkenlik Katsayısı

- Değişkenlik katsayısı standart sapmanın aritmetik ortalamaya bölünmesi ile elde edilen yüzdesel bir orandır. Aşağıdaki formülle hesaplanır.

$$CV = \frac{\sqrt{\sigma^2}}{\mu} \; veya \; CV = \frac{\sqrt{s^2}}{\bar{x}}$$

Değişkenlik katsayısı

- Bir ana kütleden 3 farklı örneklem çektiğimizi düşünelim. Bu örneklemler için ortalamalar ve standart sapmalar aşağıdaki gibi olsun.

Örneklem 1	Ortalama = 141	Standart Sapma = 12
Örneklem 2	Ortalama = 136	Standart Sapma = 12
Örneklem 3	Ortalama = 136	Standart Sapma = 10

- Sadece standart sapma ile yorum yaparsak Örneklem 1 ve Örneklem 2 için değişkenliklerin aynı olduğu sonucuna varırız. Fakat ikinci örneklemin ortalaması daha düşüktür. Bu durumda ikinci örneklem için bu standart sapma değeri daha fazla değişkenlik göstermektedir.
- Üçüncü örneklemde standart sapma değeri daha düşük olmakla birlikte, ortalama da düşüktür. Bu durumda birinci örneklem mi yoksa üçüncü örneklem mi daha değişkendir. İşte bu şekilde karşılaştırma durumlarındaki zorluklardan kurtulmak adına değişkenlik katsayısı oranları hesaplanır.

Değişkenlik katsayısı

- Değişkenlik katsayıları üzerinden yorum yaparsak; en fazla değişkenlik ortalamanın düşük, standart sapmanın yüksek olduğu Örneklem 2 dedir. Örneklem 3 ise en az değişkenliğe sahip veri setidir.

Örneklem 1	Değişkenlik Katsayısı = 12/141 = 0,0851
Örneklem 2	Değişkenlik Katsayısı = 12/136 = 0,0882
Örneklem 3	Değişkenlik Katsayısı = 10/136 = 0,0735

Göreceli Durum Ölçütleri

Persentil (yüzdebirlik veya yüzdelik)

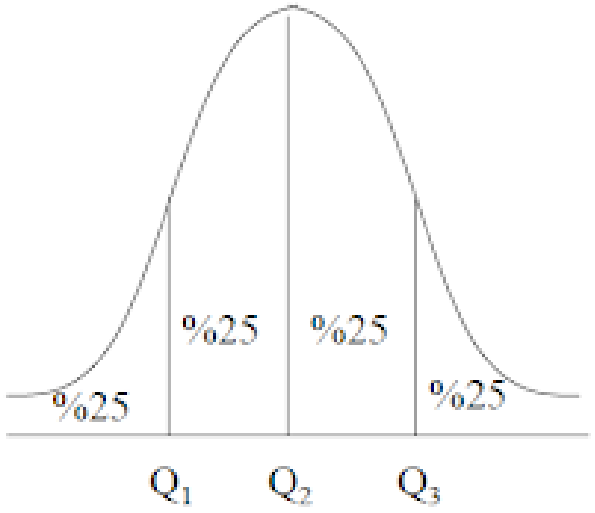
- Büyük veri setlerini tanımlamakta kullanılan göreceli bir durum ölçütüdür.
- Persentili hesaplamak için öncelikle veri seti küçükten büyüğe sıralanmalıdır. Bir yüzdebirlik, sıralanmış bir veri serisini yüz eşit parçaya böler ve böylece her bir bölünen parçanın anakütle veya örneklem verilerinin 1/100'ini kapsar.
- Aşağıdaki formül yardımıyla persentil değerinin konumu belirlenebilir. n veri sayısını, p ise kaçıncı yüzdebirlik olduğunu gösterir.

$$L_P = (n+1)\frac{P}{100}$$

Göreceli Durum Ölçütleri

Çeyreklik (Kartil)

- Küçükten büyüğe doğru sıralanmış verileri dört eşit parçaya bölen değerlere çeyreklikler denir. Örneklem için sırasıyla %25 ine denk gelen değer (%25 lik persentil), Medyan (%50 lik persentil) ve %75 e denk gelen değerlere (%75 lik persentil) karşılık gelir.
- Sırasıyla Q_1, Q_2, Q_3 olarak gösterilir. Q_1ilk veya alt çeyreklik, Q_2 ikinci çeyreklik veya medyan ve Q_3 ise üçüncü çeyreklik veya üst çeyreklik olarak adlandırılır.



Örnek

- Bir grup çalışanın boy değerleri santimetre cinsinden aşağıda verilmiştir. Q_1, Q_2, Q_3 değerlerini belirleyiniz.

173	165	171	175	188
183	177	160	151	169
162	179	145	171	175
168	158	186	182	162
154	180	164	166	157

Çözüm

Çeyreklik değerlerini hesaplayabilmek için önce veriler sıralanmalıdır.

Sıra	Boy	Sıra	Boy
1	145	14	171
2	151	15	171
3	154	16	173
4	157	17	175
5	158	18	175
6	160	19	177
7	162	20	179
8	162	21	180
9	164	22	182
10	165	23	183
11	166	24	186
12	168	25	188
13	169		

Çözüm

- Sonra persentillerin konumlarının hesaplandığı L_p değerleri hesaplanır.

$$L_{25} = (25+1)\frac{25}{100} = 6{,}5$$

$$L_{50} = medyan = 13$$

$$L_{75} = (25+1)\frac{75}{100} = 19{,}5$$

- Bu değerler bize persentilin konumu verir. Yani verilerin % 25 ini içine alan değer 6,5. değerdir. Böyle bir değer olmadığından doğrusal bir interpolasyona ihtiyaç vardır. 6,5. değer 6. değer ile 7. değer eşit uzaklıkta olduğunda 6. değer ile 7. değerin aritmetik ortalamasından hesaplanır. Örneklemin % 75 ini kapsayan değer 19,5. değer de aynı şekilde belirlenir.

Çözüm

- Böylece Q_1, Q_2, Q_3 değerleri bulunur.

$$Q_1 = \frac{6.değer + 7.değer}{2} = \frac{160 + 162}{2} = 161$$

$$Q_2 = medyan = 13.değer = 169$$

$$Q_3 = \frac{19.değer + 20.değer}{2} = \frac{177 + 179}{2} = 178$$

Çeyreklik

- Çeyreklik değerleri belirlenirken persentil konumundan gelen değer ondalıklı ise doğrusal interpolasyonla sonuç çıkarılır.
- Örneğin 24 değerli bir veri setinde %25 lik persentil 6,25. değer olacaktır. Bu durumda persentil değerinin 6. değere 7. değerden daha yakın olacağı görülmektedir.
- 6. değer 8 ve 7. değer 12 ise 6,25. değer aşağıdaki gibi hesaplanır.

$$0{,}75 * 6.değer + 0{,}25 * 7.değer = 0{,}75 * 8 + 0{,}25 * 12 = 6 + 3 = 9$$

5 Nokta Yöntemi

- 5 sayı yöntemi Q_1, Q_2, Q_3 çeyreklik değerleri ile birlikte veri setindeki en küçük (S) ve en büyük değerin (L) birlikte sunulduğu bir tanımlayıcı istatistik yöntemidir. Değişim aralığı değerinin daha gelişmiş hali olarak da ifade edebileceğimiz bu yöntem veri setinin değişkenliği ve konumu hakkında önemli bilgiler sunmaktadır. Aşağıdaki gibi bir tabloda sunulur.

En Küçük Değer (S)	
Birinci Çeyreklik (Q_1)	
Medyan (Q_2)	
Üçüncü Çeyreklik (Q_3)	
En büyük Değer (L)	

Örnek

- Önceki soruda verilen bir grup çalışanın boy değerleri için 5 nokta değerlerini belirleyelim. En küçük değer 145 ve en büyük değer 188 di. Çeyreklikler ise sırasıyla, 161, 169 ve 178 olarak bulunmuştu. Bu şartlar altında 5 nokta tablosu aşağıdaki gibidir.

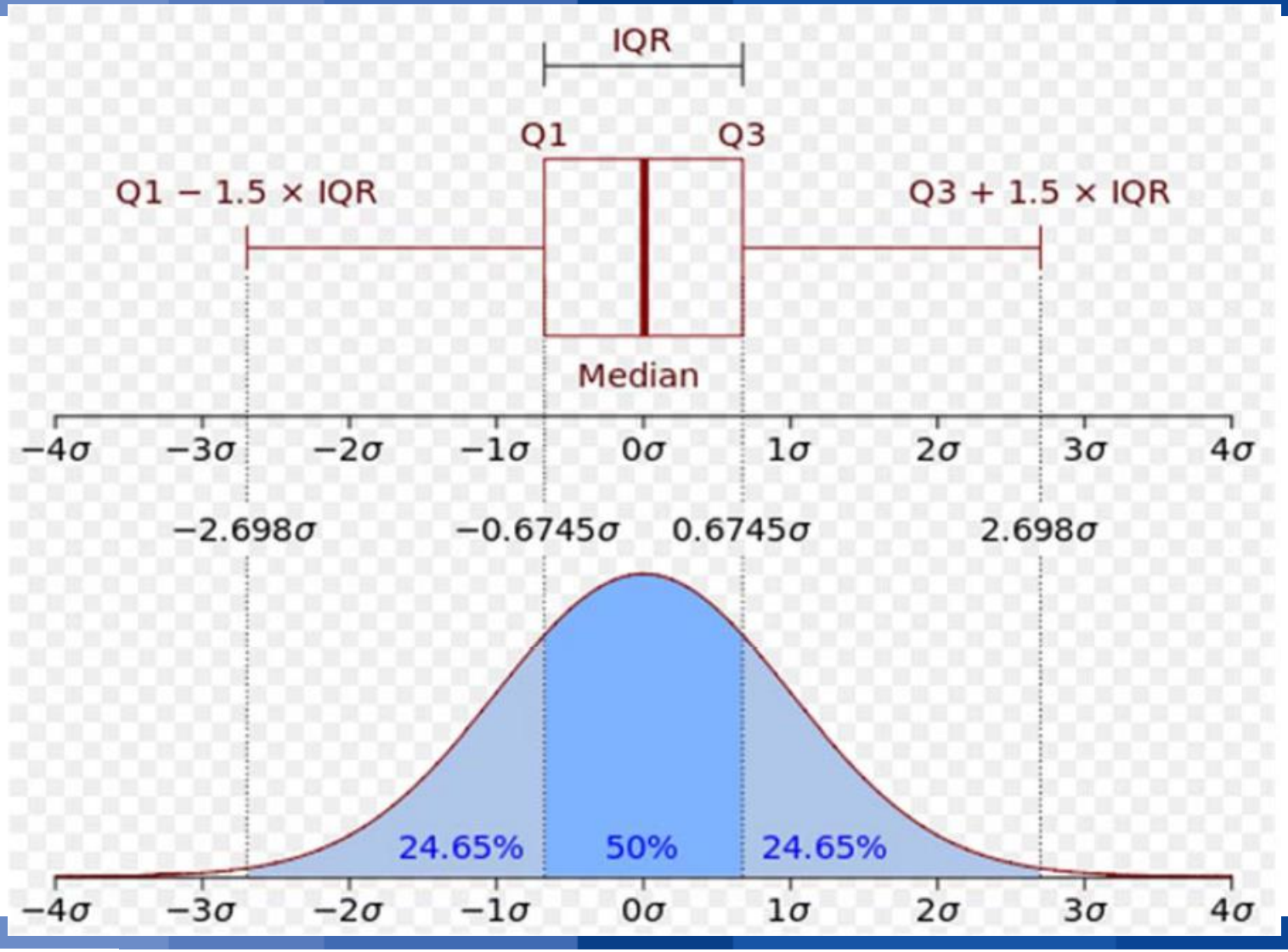
En Küçük Değer (S)	145
Birinci Çeyreklik (Q_1)	161
Medyan (Q_2)	169
Üçüncü Çeyreklik (Q_3)	178
En büyük Değer (L)	188

- Yukarıdaki tablodan verilerin yaklaşık olarak 40 birim içerisinde dağıldığı ve çeyreklikllerin medyana yakın olması durumu da dikkate alınarak çok fazla değişkenlik göstermedikleri sonucuna varılabilir.

Kutu grafiği

- 5 Nokta yöntemi tamamen sayılardan oluşan yapısıyla bazı durumlarda kolayca yorumlanamayabilir.
- Bu yüzden “Kutu Grafiği” geliştirilmiştir.
- Temelde 5 Nokta yöntemine dayanan kutu grafiği veri setinin değişkenliği ve hangi aralıkta değiştiği ile ilgili önemli yorumlar sunar.
- Kutu grafiği uç değerleri, medyanı (dağılım simetrikse ortalamayı) ve dağılımı birlikte sunduğunda kuvvetli bir görsel tanımlama aracıdır.

Kutu grafiği



Kutu grafiği

- Kutu genişliği verilerin %50 sini kapsayacak şekilde belirlenmektedir.
- Ayrıca veri setinin değişkenliğini gösteren kutudan sağa ve sola doğru uzanan çubuklar da (Whiskers) çizilmelidir.
- Kutu 5 nokta değerleri ile çizilebilirken, sağ ve sol çubukları çizmeden önce IQR olarak ifade edilen çeyreklikler arası uzaklığın belirlenmesi gerekmektedir.

Kutu grafiği

- IQR değeri üçüncü çeyreklikten birinci çeyrekliğin çıkarılması ile bulunur. Bu değer 5 nokta tablosundaki $Q_3 - Q_1$ değerine eşittir.
- Çubukların ne kadar uzayacağını belirlemek için öncelikle sol taraftaki çubuk için $Q_1 - 1{,}5 *$ IQR ve sağ taraftaki çubuk için $Q_3 + 1{,}5 *$ IQR değerleri hesaplanmalıdır.
- Hesaplanan değerler sırasıyla S den küçük ve L den büyük ise dikkate alınmazlar. Aksi halde hesaplanan bu değerler dikkate alınarak çizimler yapılır.
- **!!!** Eğer çizimlerde S ve L değerleri yerine IQR dan hesaplanan yeni değerler kullanılırsa birden fazla gözlem değeri aykırı değer olarak görülecektir.

Örnek

- Önceki soruda verilen bir grup çalışanın boy değerleri için kutu grafiğini çiziniz ve aykırı değer olup olmadığını sorgulayınız.

En Küçük Değer (S)	145
Birinci Çeyreklik (Q_1)	161
Medyan (Q_2)	169
Üçüncü Çeyreklik (Q_3)	178
En büyük Değer (L)	188

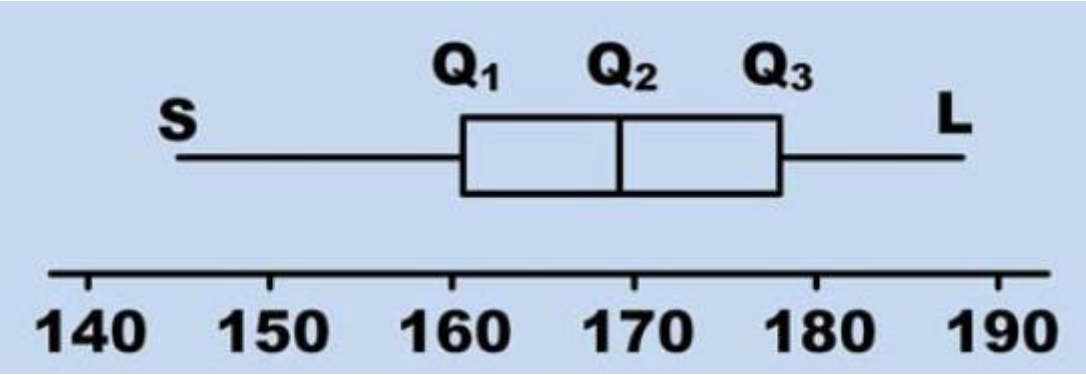
Çözüm

- 5 nokta tablosuna ek olarak IQR, $Q_1 - 1{,}5 * \text{IQR}$ ve $Q_3 + 1{,}5 * \text{IQR}$ değerleri hesaplanmalıdır.

En Küçük Değer (S)	145
Birinci Çeyreklik (Q_1)	161
Medyan (Q_2)	169
Üçüncü Çeyreklik (Q_3)	178
En büyük Değer (L)	188
IQR	178-161=17
$Q_1 - 1,5 * IQR$	161-1,5*17= 135,5
$Q_3 + 1,5 * IQR$	178+1,5*17= 203,5

Çözüm

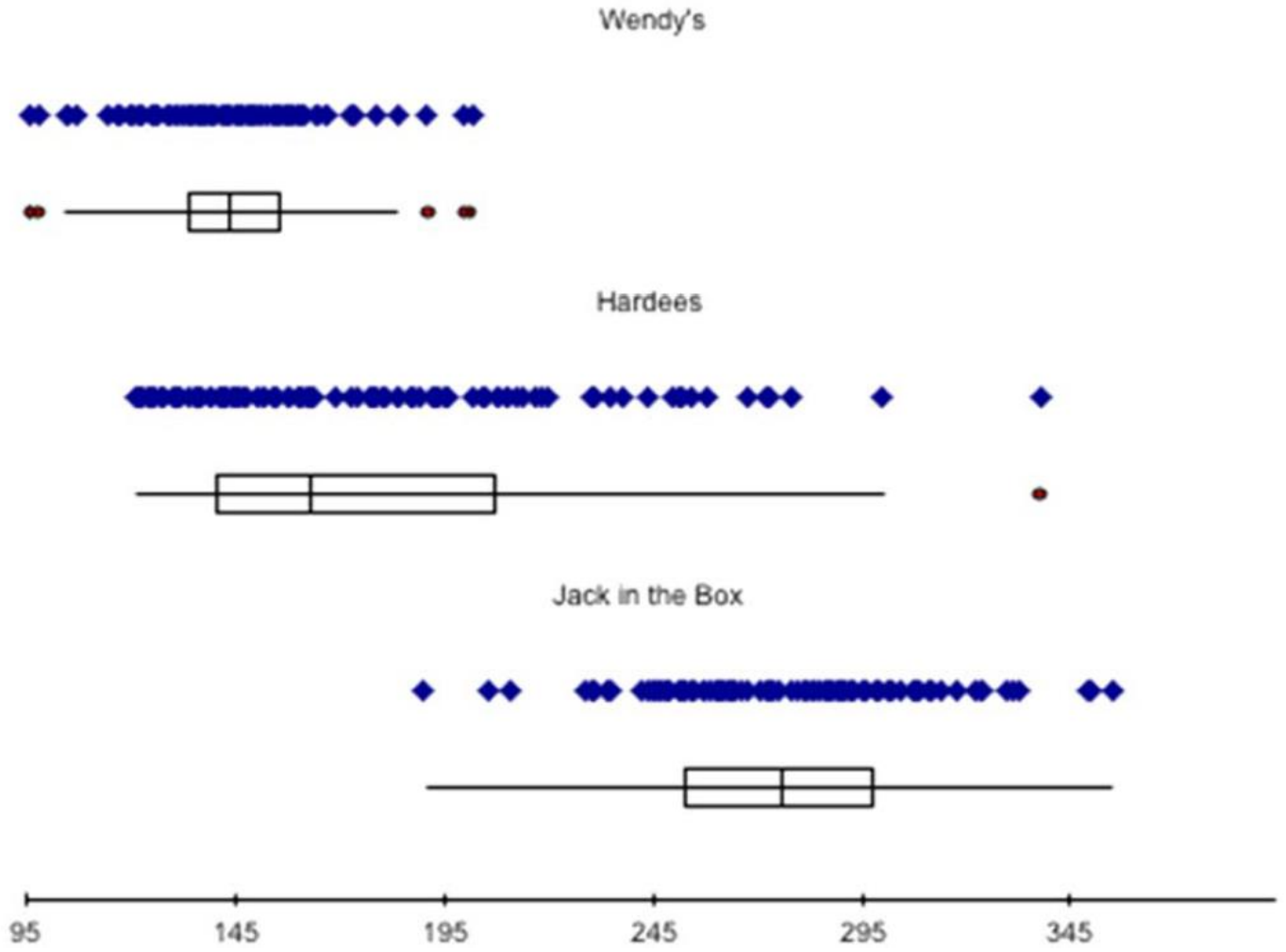
- 5 nokta tablosuna ek olarak IQR, ve değerleri hesaplanmalıdır.
- $Q_1 - 1{,}5 * \text{IQR}$ değeri 135,5 olduğundan S değerinden küçüktür. Bu yüzden çubuklar çizilirken bu değer dikkate alınmaz.
- Benzer şekilde $Q_3 + 1{,}5 * \text{IQR}$ değeri 203,5 tir ve L değerinden büyüktür. Bu nedenle çubuklar çizilirken bu değer dikkate alınmaz.



- Çubuk çizimlerinde S ve L dikkate alındığından herhangi bir aykırı değer olmayacaktır.

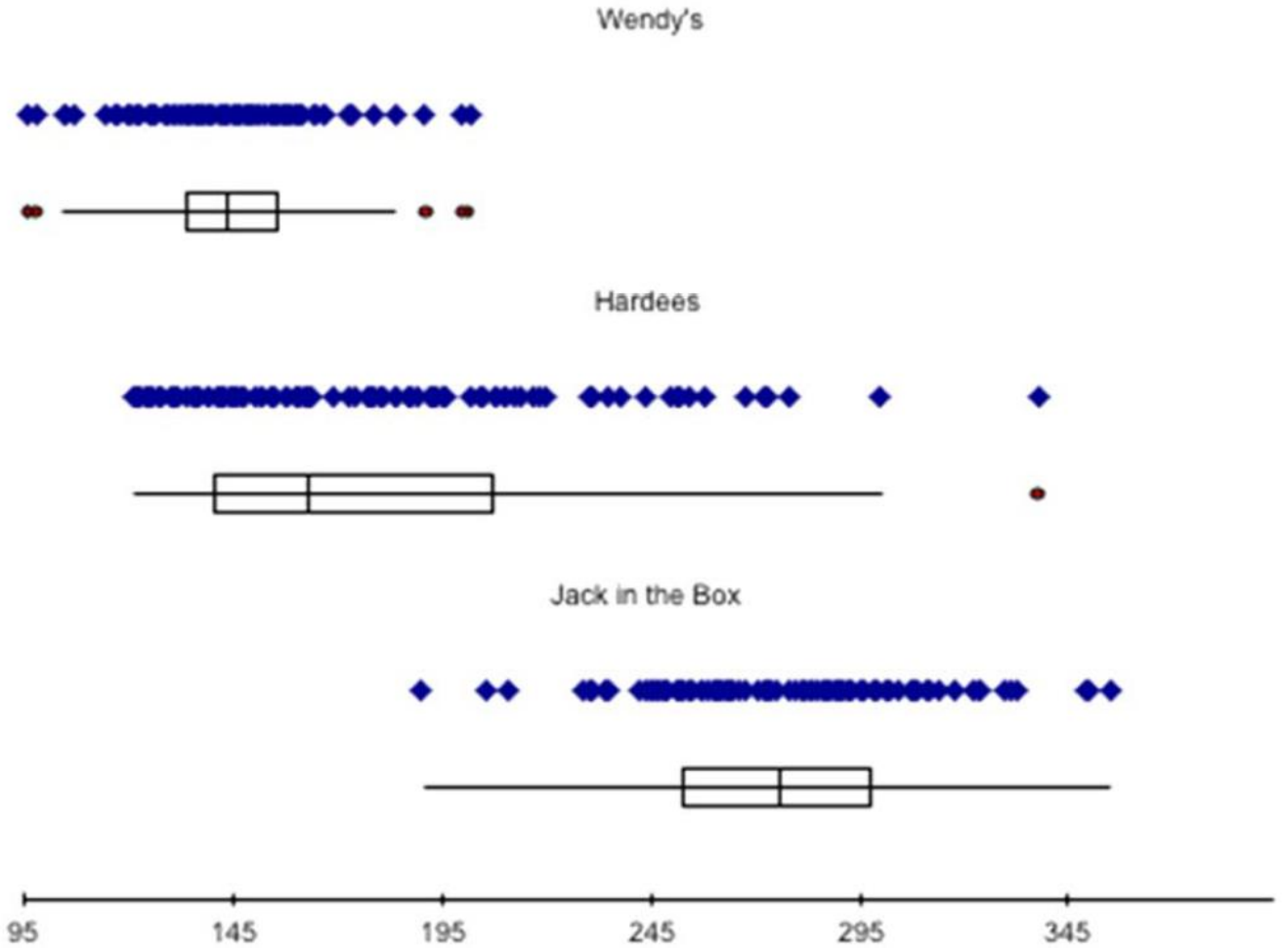
Örnek

- Amerikadaki 3 büyük fast food şirketinin ortalama servis sürelerinin karşılaştırılması için kutu grafiğinden yararlanılmıştır.



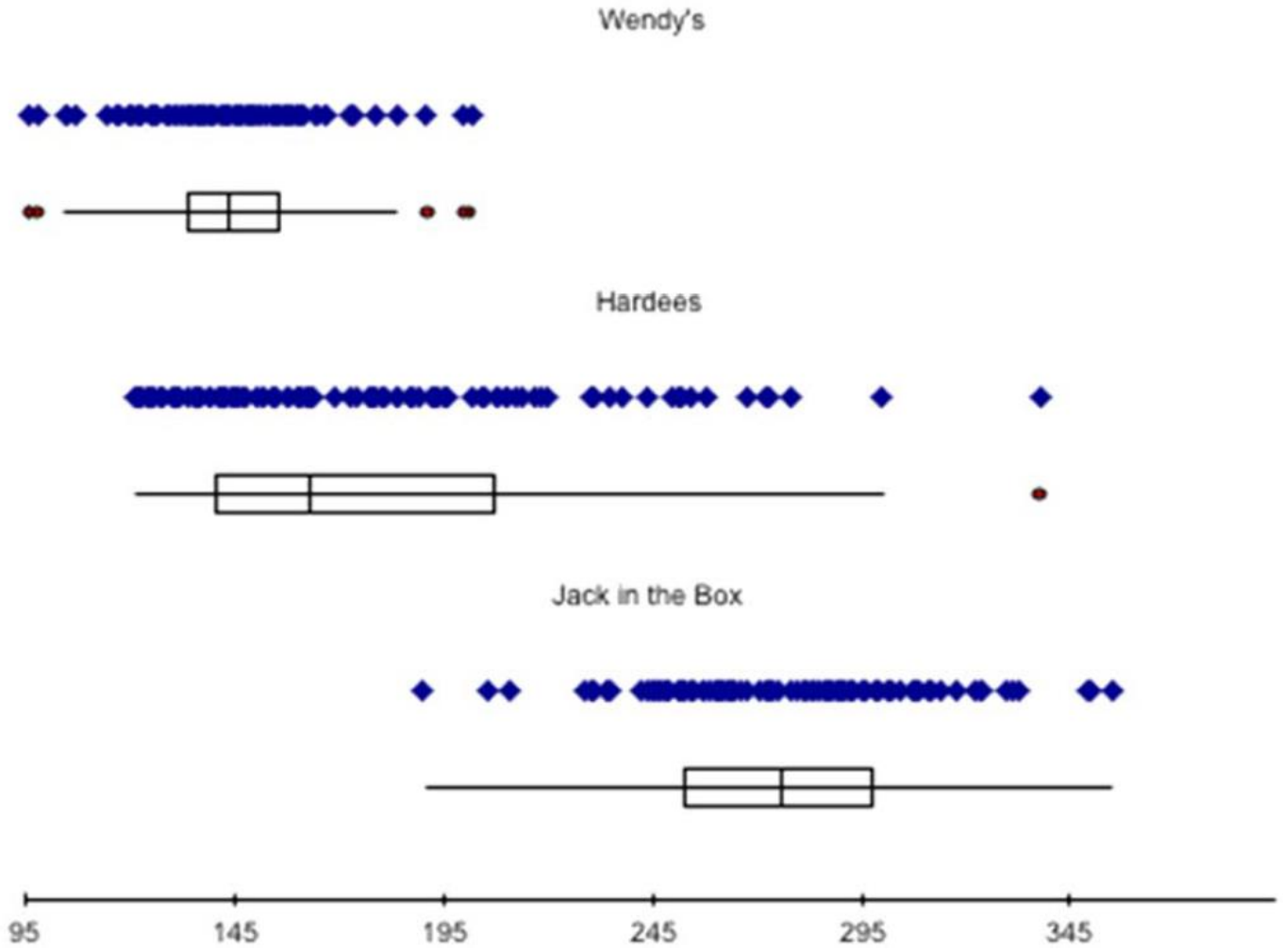
Örnek

- Grafiğe göre öncelikle Wendy’s firmasının ek kısa bekleme sürelerine sahip olduğu görülmektedir. (100 ila 180 saniye arasında)
- Wendy’s aynı zamanda servis süresi söz konusu olduğunda en düşük değişkenliğe sahiptir. Çünkü hem kutu dar hem de çubuklar kısadır.
- Bununla birlikte Wendy’s firmasında en az 5 müşteri servisinin beklenenden çok uzun veya çok kısa olduğu çubuklar sonrasında görünen aykırı değerlerden anlaşılmaktadır.



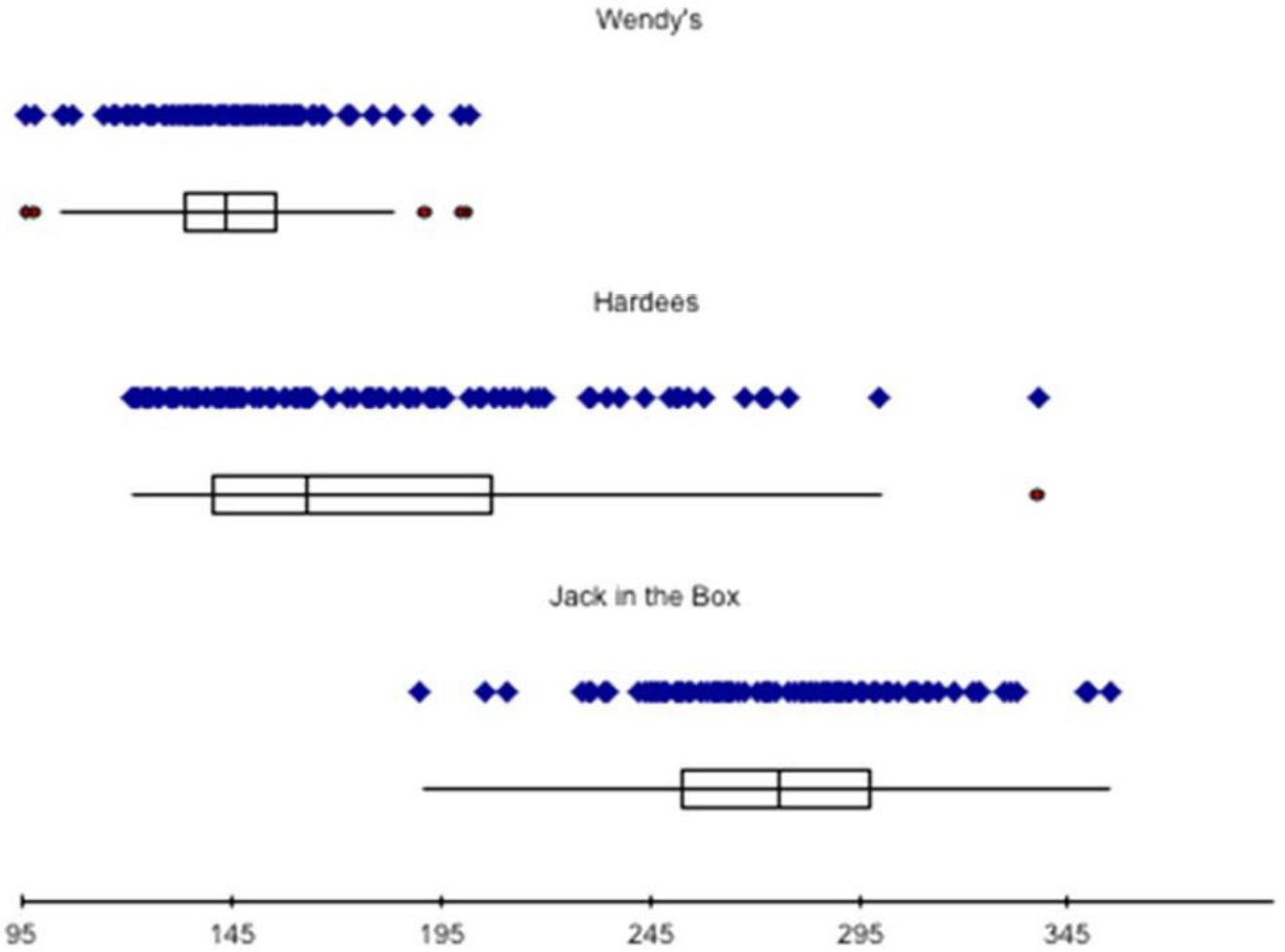
Örnek

- Hardee’s yüksek değişkenlikle müşterilerine 120 ila 300 saniye aralığında hizmet vermekte, fakat bu müşterilerin büyük kısmına 200 saniyenin altında hizmet sunmaktadır.
- Sağa çarpık bir dağılıma sahip Hardees servis süreleri simetrik bir veri dağılımına sahip değildir.



Örnek

- Simetrik bir dağılıma sahip Jack in the Box 200 ila 350 saniye servis süreleri ile müşterilerini en çok bekleten firmadır.



Göreceli Durum Ölçütleri

Z Skoru

- Örneklem içerisindeki verilerin örneklem ortalamasından ne kadar uzakta olduğunun oransal olarak gösterebilmek adına hesaplanan, gözlemlenen değişkenin standardize edilmiş halidir. Aşağıdaki formülle hesaplanabilir.

$$z = \frac{x - \bar{x}}{s}$$

- Burada hesaplanan z değeri ile herhangi bir değerin ortalamadan kaç standart sapma ile uzaklaştığını belirleyebiliriz.

Örnek

- Örneğin İstatistik dersine ait final sınavının ortalamasının 60 ve standart sapmanın 5 olduğu belirlenmiş ise, 45 alan bir öğrenci için z skorunu hesaplayıp yorumlayalım.

$$z = \frac{x - \bar{x}}{s} = \frac{45 - 60}{5} = -3$$

- Z değerinin – 3 olarak ortaya çıkması, 45 notu almış olan öğrencinin ortalamadan 3 standart sapma daha aşağıda not aldığı şeklinde yorumlanabilir. İstatistik dağılımlarının en önemlisi olan normal dağılım olasılık değerlerinin kolaca hesaplanabilmesi için z skorlarından faydalanılır.

Kaynaklar

- Bilişim Teknolojileri için İşletme İstatistiği ders notları, Dr. Öğr. Üyesi Halil İbrahim Cebeci
- İstatistik ders notları, Prof. Dr. Mehmet Aksaraylı