

Moleküler Biyolojide Kullanılan İstatistiksel Yöntemler



Prof. Dr. Ercan ARICAN
Moleküler Biyoloji ve Genetik Bölümü

VERİLERİN TOPLANMASI, ÖZETLENMESİ ve SUNULMASI

Araştıracının, konusu ile ilgili verilerini doğru olarak toplaması, özetlemesi, tanıtıcı değerleri hesaplaması, araştırmada dikkate alınan faktörlere göre analiz etmesi, eğer araştırmada birden fazla özellik dikkate alınmış ise bu özellikler arasındaki ilişkiyi araştırması, analizler sonucunda bulunan sonuçları değerlendirmesi, yorumlaması ve genelleştirmesi İSTATİSTİK metotlar olarak bilinir.

Bir araştırma yürütülürken üzerinde durulan topluluğun bütün bireylerinden veri toplamak çok zor, hatta olanaksızdır.

Çünkü araştırma için ayrılan zaman ve maddi olanaklar bütün bireylerden veri toplanmasını sınırlar.

Bu sınırlayıcı faktörleri aşmak mümkün olsa da gene gereksiz yere para ve zaman harcanmış olur. Bu sebeple veriler sınırlı sayıdaki bireylerden toplanır.

Bu sınırlı sayıda bireylerin oluşturduğu gruba örnek denir. Örneğe giren özellikteki bireylerin tümünün oluşturduğu topluluğa (kitleye, yığına) ise populasyon denir.

Biyolojik anlamda populasyon, canlıların topluluğudur. Örneğin, öğrenciler, kuşlar, balıklar biyolojik anlamda populasyonu oluşturur.

İstatistik anlamda populasyon ise üzerinde çalışılan özelliktir. Örneğin, öğrencilerin ağırlıkları, boyları, kanlarında ölçülen potasyum, bakır, çinko değerleri gibi özellikler bir populasyondur.

Populasyondan hesaplanan değerlere parametre, parametrelere karşılık gelen ve örnekten hesaplanan değerlere de istatistik adı verilir.

Populasyondan hesaplanan parametreler Grek (Yunan) harfleri, örnekten hesaplanan istatistikler ise Latin harfleri ile gösterilir.

Örneğin, herhangi bir üniversitedeki öğrencilerin tamamı popülasyonu oluşturur.

Bu öğrencilerin ağırlıklarına ait ortalama hesaplanır ise bu bir parametredir ve (μ ile gösterilir).

Eğer bu öğrencilerden üniversitedeki öğrencileri temsil eden belirli bir sayıda öğrenci seçilir ise bu bir örnektir ve örnekten hesaplanan ortalama popülasyondan hesaplanana karşılık gelen bir istatistiktir (bu ortalamanın bir tahminidir).

Araştırmacı üzerinde durduğu özellikleri istediği harfle tanımlayarak, bunların ortalamalarını üstüne çizgi koyarak (μ veya $\bar{X}$ gibi) gösterebilir.

Bunlara karşılık gelen parametreler ise μ_A ve μ_X ile gösterilir.

Örnekteki birey sayısına "örnek genişliği" denir. Bir araştırmada örnek genişliği (öğrenci sayısı) 500'dür denirse, bunun anlamı öğrencilerinden 500 tanesine anket uygulanmıştır yani örnekte 500 öğrenci vardır demektir.

Bir örneği oluşturan bireyler konu gereği, öğrenciler, hastalar, deney hayvanları vs. olabilir.

Bir araştırmacı bir bölgedeki insanların kanlarındaki bakır ve çinko miktarları ile ilgili bir araştırma yapmayı düşünüyor olsun.

Bu arařtırıcının söz konusu bölgedeki bütün insanlardan kan alınması çok zor ve hatta olanaksızdır.

Bu durumda bölgedeki insanları temsil edecek belirli sayıda kiři seçer. Belirlenen insanlar örneęi oluşturur.

Örnekten elde edilen deęerlerin güvenilir olması ve populasyon için genelleřtirilebilecek sonuçlar vermesi için arařtırıcının örneęini seçerken dikkat etmesi gereken önemli noktalar vardır.

Bunlar ařaęıdaki řekilde sıralanabilir:

- Araştırmacı örneğe girecek bireyleri seçerken tarafsız (yansız) olmalıdır. Yani örneği oluşturacak bireyleri tamamen tesadüfi (rastgele) olarak seçmelidir.
- Araştırmacı örneğini alacağı popülasyonu iyi tanımalıdır ve popülasyonu temsil edebilecek yeterli sayıda bireyi örneğine almalıdır.

Örnekteki birey sayısı arttıkça, örnekten tahmin edilen değer (hesaplanan istatistik) parametre değerine daha yakın olacak ve yapılan bu tahminin güvenilirliği artacaktır.

İstatistikte, örneğin herhangi bir özelliğini belirten rakamlara gözlem değeri denir.

Gözlem değerleri, eğer saymak sureti ile elde edilmişler ise kesikli gözlem değeri, tartmak, ölçmek veya herhangi bir analiz metodu kullanılarak elde edilmişler ise sürekli gözlem değeri denir.

Sürekli gözlem değerleri tarif aralığında her değeri alabilirler. Örneğin, öğrencilerin ağırlıkları tartmak, boyları ölçmekle elde edilir ki bunlar sürekli gözlem değerleridir.

Fakat bir ailedeki çocuk sayısı, okunan gazete sayısı, bir işletmede çalışan eleman sayısı vs. saymak sureti ile elde edilir. Bunlar da kesikli gözlem değerleridir.

Eğer bir öğrencinin ağırlığı ölçülür ve 45.0 kg olarak kaydedilir ise bu ağırlık 44.5 ile 45.4 arasında her değeri alabilir.

Fakat bir ailedeki çocuk sayısı, örneğin, ya 2'dir veya 3'dür. Bu iki değer arasında herhangi bir değer olması söz konusu değildir.

Verilerin Özetlenmesi

Araştırmacı örneğini oluşturduktan sonra, örneğe giren bireylerden rakamlarını toplar.

Rakamlar toplandıktan sonra, değerlendirmenin ilk aşaması verilerin özetlenmesidir.

Örneğin, 80 öğrencinin istirahat halinde dakikada kalp atışları sayısı aşağıdaki gibi belirlenmiş olsun.

87 98 81 83 96 89 81 60 67 102
83 69 77 62 77 82 68 76 72 80
85 79 58 81 99 93 73 75 68 70
75 105 80 79 76 77 72 84 81 55
89 70 81 48 71 67 80 75 77 80
88 61 81 75 73 81 63 83 82 87
82 83 82 71 65 62 82 87 69 74
50 74 66 77 93 74 78 52 85 73

Eğer araştırmacı yazdığı bir rapor veya bir makalede bu verileri yukarıdaki şekilde verirse okuyucuya 80 rakamı vermekten başka bir şey yapmamış olur.

Okuyucu eğer nabız sayısı 80'den fazla olan kişilerin sayısını merak ederse rakamları tek tek kontrol ederek nabız sayısı 80'den fazla olan bireyleri sayması gerekir.

Eğer veriler özetlenirse ileride açıklanacak olan eklemeli frekans tablolarından bu kolayca hesaplanabilir.

Verilerin özetlenmesi frekans dağılım tablosu düzenleyerek veya grafikler çizilerek yapılabilir.

Frekans Dağılım Tablosu

Frekans dağılım tablosu, aynı veya birbirine yakın değere veya özelliğe sahip olan gözlem değerlerini aynı sınıfa koyarak ve her sınıftaki birey sayılarını belirterek hazırlanan tablolardır.

Frekans dağılım tablolarının hazırlanmasında güçlüklerle karşılaşılması için işlem basamakları aşağıda sıralanmıştır:

1. Frekans dağılım tablosu hazırlamanın ilk adımı frekans dağılım tablosunda kaç sınıfın olacağına karar vermektir.

Genellikle frekans dağılım tablolarında 8-10 sınıf yapılır.

Sınıf sayısının 8'den az ve 15'den fazla olmamasına dikkate edilir.

Eğer sınıf sayısı 15'i geçerse özetlemeden beklenen amaca ulaşılmamış olur.

En uygun sınıf sayısının belirlenmesinde aşağıda verilen Sturges formülü de kullanılabilir:

$$\text{Sınıf sayısı} = 1 + 3.3 \log (n)$$

Formülde n örnek genişliğidir.

80 öğrenciden oluşan örnek için hazırlanacak frekans dağılım tablosunda Sturges formülüne göre en uygun sınıf sayısı:

$$1 + 3.3 \log (80) \cong 7' \text{ dir.}$$

Sınıf sayısı en az 8 olması istendiğinden ele alınan örnekte de 8 sınıf öngörülmüştür.

Örnek genişliği 200 olsaydı bu sayı

$$1 + 3.3 \log (200) \cong 9$$

olarak bulunacaktı.

2. Frekans dağılım tablosunda kaç sınıfın olacağına karar verildikten sonra yapılması gereken ikinci işlem basamağı, sınıf aralığının belirlenmesidir.

Sınıf genişliği (sınıf aralığı) ard arda iki sınıfın sınırları arasındaki farktır ve yaklaşık olarak aşağıdaki eşitlik kullanılarak hesaplanır:

$$\text{Yaklaşık Sınıf Aralığı} = \frac{\text{Değişim Genişliği}}{\text{Sınıf Sayısı}}$$

Değişim Genişliği = En büyük gözlem değeri - En küçük gözlem değeri

İlk adımda yukarıdaki örnek için frekans dağılım tablosunda 8 sınıfın yapılmasına karar verilmişti.

Bu durumda sınıf aralığı yaklaşık olarak aşağıdaki şekilde hesaplanır:

$$\text{Yaklaşık sınıf aralığı} = \frac{105 - 48}{8} = 7.125$$

Yaklaşık sınıf aralığı 7.125 olarak hesaplandıktan sonra araştırmacı sınıf aralığını, işlemleri kolay yürütebilmek için yuvarlaklaştırmalıdır.

3. Frekans dağılım tablosunda sınıf sayısı kararlaştırılıp sınıf aralığı hesaplandıktan sonra sınıf sınırlarının belirlenmesi gerekir.

Her sınıf alt ve üst sınırları ile belirtilir.

İşlemlerin kolay yürütülmesi için ilk sınıfın alt sınırı örnekteki gözlem değerlerinin en küçüğü olarak seçilir (örneğimizde birinci sınıfın alt sınırı 48).

Diğer sınıfların alt sınırları ise ilk sınıfın alt sınırına sınıf aralığı eklenerek hesaplanır.

Sınıf sınırlarının duyarlılığı gözlem değerlerinininki ile aynı olmalıdır.

Ele alınan örnekte 48.0 veya 48.99 gibi bir sınır belirlenmez.

Bir sınıfın alt sınırı ile bir sonraki sınıfın üst sınırı aynı olmamalıdır.

Yani sınıf sınırları çakışmamalıdır ve birbirini takip eden iki sınıftan küçük olanın üst sınırı ile bundan sonrakinin alt sınırı arasında boşluk bırakılmamalıdır.

Örnekteki en büyük gözlem değeri en son sınıfa girecek şekilde sınıflar oluşturulur.

4. Frekans dağılım tablosunda sınıfların alt ve üst sınırları hesaplandıktan sonra her sınıf için "sınıf değeri" hesaplanır.

Sınıf değeri bir sınıfın alt sınırı ile üst sınırının orta noktasıdır ve söz konusu sınıfı en iyi temsil eden değerdir.

5. Frekans dağılım tablosundaki sınıflar düzenlendikten sonra sınıfların frekansları belirlenir.

Örnekteki gözlem değerlerinin frekans dağılım tablosundaki sınıflardan hangisine dahil olduğu tek tek sırasıyla kontrol edilerek hangi sınıfa dahil olduğu kararlaştırılır ve böylece sınıfların frekansları önce işaretle sonra da rakamla belirlenir.

6. Frekans dağılım tablosundaki sınıfların frekansları mutlak olarak belirlendikten sonra "Nisbi (%)" frekanslar hesaplanır.

Örneğimiz de 1. sınıfa ait % frekans $3 \cdot 100 / 80 = 3.75$ ve 2. sınıfa ait % frekans $4 \cdot 100 / 80 = 5.0$ olarak bulunmuştur.

Öğrencilerin dakikadaki kalp atışları ile ilgili örnek için yukarıda açıklanan bilgiler ışığında hazırlanan frekans dağılım tablosu verilmiştir.

80 öğrencinin dakikadaki kalp atış sayısı için düzenlenen frekans dağılım tablosu

Sınıflar	Sınıf Değeri	İşaretle Frekans	Sayı ile Frekans (Mutlak frekans)	% (Nisbi) Frekans
48-54	51.0	///	3	3.75
55-61	58.0	////	4	5.00
62-68	65.0	///// ////	9	11.25
69-75	72.0	///// ///// ///// ///	18	22.50
76-82	79.0	///// ///// ///// ///// ///// /	26	32.50
83-89	86.0	///// ///// ///	13	16.25
90-96	93.0	///	3	3.75
97-103	100.0	///	3	3.75
104-110	107.0	/	1	1.25

Eklemeli Frekans Dağılım Tablosu

Frekans dağılım tablosu düzenlendikten sonra belirli bir değerden daha büyük veya daha küçük olan gözlem değerlerinin mutlak ve/veya nisbi miktarını belirtmek için eklemeli frekans dağılım tabloları düzenlenir.

Eklemeli frekans dağılım tabloları "...den daha az" veya "...den daha fazla" eklemeli frekans tabloları olarak düzenlenebilir.

Böylece belirli bir değerden daha küçük veya daha büyük gözlem değerlerinin (bireylerin) mutlak ve nisbi miktarları kolayca bulunabilir.

Eklemeli frekans dağılım tabloları düzenlenirken sınıfların "Gerçek Sınırları" kullanılır.

Gerçek sınır, bir sınıfın üst sınırı ile bir sonraki sınıfın alt sınırının orta noktasıdır.

Bu orta nokta önceki sınıfın üst gerçek sınırı, aynı zamanda sonraki sınıfın alt gerçek sınırıdır.

Bir sınıf üst gerçek sınırında biter ve bu noktada sonraki sınıf başlar.

Bu nokta söz konusu sınıfın alt gerçek sınırıdır.

ÖRNEK: 80 öğrencinin dakikada kalp atış sayısı için düzenlenen frekans dağılım tablosunda gerçek sınırların hesaplanması aşağıda açıklanmıştır:

Aşağıdaki frekans dağılım tablosunda birinci sınıfın gerçek sınırı hesaplanırken, eğer birinci sınıftan önce bir sınıf daha olsa idi bu sınıfın üst sınırı 47 olacağından, birinci sınıfın alt gerçek sınırı $(47+48)/2=47.5$ olarak bulunur.

Bu nokta birinci sınıfın başladığı gerçek sınırdır.

Aynı şekilde ikinci sınıfın gerçek sınırı $(54+55)=54.5$ 'tir.

Bu ikinci sınıfın alt gerçek sınırı, birinci sınıfın ise üst gerçek sınırıdır.

Diğer sınıflar için de bu şekilde hesaplanabileceği gibi birinci sınıfın gerçek sınırına sınıf aralığı eklenerek diğer sınıfların gerçek sınırları hesaplanabilir.

Sınıflar	Alt gerçek sınır	Frekans
48-54	47.5	3
55-61	54.5	4
62-68	61.5	9
69-75	68.5	18
76-82	75.5	26
83-89	82.5	13
90-96	89.5	3
97-103	96.5	3
104-110	103.5	1
	110.5	0

Gerçek sınırlar hesaplandıktan sonra "...den daha az" ve "...den daha fazla" eklemeli frekans dağılım tabloları düzenlenebilir, "...den daha az" frekans dağılım tablosu düzenlenirken esas şudur:

47.5'den daha az değere sahip gözlem değeri sayısı 0'dır. 54.5'den daha az değere sahip gözlem değeri sayısı 3, yani önceki sınıfın frekansı kadardır.

61.5'den daha az değere sahip gözlem değerlerinin sayısı, bu sınıftan önceki sınıfların frekansları toplamı kadar, yani 7'dir ve diğer sınıflar için de "...den daha az" eklemeli frekansları aynı şekilde bulunur.

Yani frekanslar yukarıdan aşağıya eklenerek bulunur, "...den daha fazla" eklemeli frekanslarının bulunması için yapılan işlemde "..den daha az" eklemeli frekansları için yapılan toplama işleminin tam tersidir.

110.5'den daha fazla değere sahip frekansların toplamı O'dır.

103.5'den daha fazla değere sahip gözlem değerlerinin frekansı l'dir ve diğer sınıflar içinde aynı işlem tekrarlanır.

Bu sefer frekanslar aşağıdan yukarıya eklenerek tablo tamamlanır. Bu şekilde bulunan eklemeli frekanslar sınıfların gerçek sınırları ile birlikte Tablo'da verilmiştir.

Alt gerçek sınır	Frekans	...den daha az eklemeli frekansları		...den daha fazla eklemeli frekansları	
		Mutlak	Nisbi(%)	Mutlak	Nisbi(%)
47.5	3	0	0.00	80	100.00
54.5	4	3	3.75	77	96.25
61.5	9	7	8.75	73	91.25
68.5	18	16	20.00	64	80.00
75.5	26	34	42.50	46	57.50
82.5	13	60	75.00	20	25.00
89.5	3	73	91.25	7	8.75
96.5	3	76	95.00	4	5.00
103.5	1	79	98.75	1	1.25
110.5	0	80	100.00	0	0.00

Eklemele frekanslar bulunduktan sonra belirli bir değerdan, örneğın dakıkada kalp atışı 75.5'den daha fazla olanların örnekte %57.5'ini oluşturduğı veya 61.5'den daha az olanların da %8.75 olduğı kolayca görülebilir.

Açık Uçlu Frekans Dağılım Tabloları

Bazı durumlarda araştırmacı örneğinden verileri elde ettiği zaman belirli bir değerden çok küçük veya çok büyük olan çok az sayıda gözlem değeri olduğunu görebilir.

Bu gibi durumlarda açık uçlu frekans dağılım tabloları düzenlenir.

Açık uçlu frekans dağılım tablolarının iki ucu açık olabileceği gibi tek ucu da açık olabilir.

Örneğin bir araştırmacı uyguladığı anket sonucunda kilosunu muhafaza için rejim yaptığını belirten 200 kişi tespit etmiş olsun.

Bunların yaş gruplarına dağılımı incelendiğinde rejim uygulayanlar arasında, 19 yaşından küçük ve 54 yaşından büyük bireyler az olduğu için iki ucu açık frekans dağılım tablosu oluşturulmuştur.

Yaş Grupları	Birey sayısı (f)
<25	9
25-29	47
30-34	57
35-39	37
40-44	27
45-49	9
50-54	7
>54	7

Kalitatif Sınıflı Frekans Dağılım Tabloları

Eczacılar odası son 3 yıl içinde Eczacılık fakültesinden mezun olanlardan rastgele seçtiği 200 eczacıyla yaptığı anket sonucunda çalışma şekillerini tespit etmiş olsun.

Ankette eczacıların kendi hesaplarına açtıkları eczanede (A), bir başkasıyla ortak eczanede (B), özel sektörde eczacı olarak (C), devlet hizmetinde eczacı olarak (D) veya eczacılık dışında başka alanda (E) çalışıp çalışmadıkları sorularak bilgi toplanmıştır.

Çalışma şekilleri	Eczacı sayısı
A (kendi eczanesi)	90
B (ortak eczane)	30
C (özel sektör)	40
D (devlet hizmeti)	30
E (başka alanlar)	10

Kesikli Gözlem Değerleri İçin Frekans Dağılım Tabloları

Bazı durumlarda araştırmacı üzerinde çalıştığı örnekten dar sınırlar arasında değişen tam sayılı gözlem değerleri elde eder.

Bu durumda frekans dağılım tablosunu nasıl düzenleyeceği topladığı tam sayılı değerlerin arasındaki farklılığa bağlıdır.

ÖRNEK 1: Bir araştırmacı bir ilaç şirketinde ilaç tanıtıcı olarak çalışan kişilerin bir günde dolaştıkları yer sayısını araştırmış ve 50 kişinin dolaştığı yer sayısını aşağıdaki gibi belirlemiş olsun.

6	6	5	7	8	8	5	6	7	6
7	7	5	6	6	6	6	7	6	7
4	6	6	6	6	6	7	5	7	6
6	5	7	6	7	4	7	8	7	7
5	7	7	5	6	6	5	6	4	6

Bu durumda araştırmacı frekans dağılım tablosundaki sınıfları 4'den 8'e kadar belirlemesi yeterlidir. Çünkü topladığı gözlem değerleri arasında büyük farklılık yoktur.

Sınıflar (gidilen yer sayısı)	Frekans (pazarlamacı sayısı)
4	3
5	8
6	21
7	15
8	3

Sürekli Gözlem Değerleri İçin Frekans Dağılım Tablosu

Genellikle biyolojik araştırmalarda üzerinde durulan özellikler sürekli dağılım gösterirler.

Bu özellikleri belirlemek üzere tespit edilen gözlem değerleri de sürekli dir.

Konunun özelliği, rakamların büyüklüğü ve kullanılan ölçü aletinin duyarlılığına bağlı olarak sürekli olan bu gözlem değerleri virgülden sonra belirli sayıda hane yürütülerek kaydedilirler.

Örneğin bir doğumevinde doğan bebeklerin doğum ağırlıkları 50 gr duyarlı bir terazide tespit ediliyor olsun.

Bu doğumevinde bir ayda doğan 120 erkek bebeğin ağırlıkları aşağıdaki gibi olsun.

3.45	3.40	3.15	3.70	3.20	3.50	3.10	3.20	3.65	3.10
3.20	3.25	3.20	3.30	3.35	3.60	3.30	3.60	3.60	3.30
3.20	3.55	3.50	3.95	3.30	3.50	3.15	3.60	3.30	3.25
3.60	3.50	3.40	3.25	3.30	3.10	3.40	3.60	3.50	3.50
3.80	3.60	3.80	3.50	3.85	3.45	3.70	3.40	3.10	3.20
3.20	3.35	3.10	3.10	3.45	3.30	3.30	3.45	3.30	3.70
3.60	3.35	3.40	3.50	3.40	3.35	2.85	3.10	3.30	3.85
3.40	3.25	3.45	3.45	3.65	3.55	3.85	3.20	3.75	3.65
3.35	3.80	3.50	3.20	3.20	3.30	3.40	3.00	3.40	3.20
3.75	3.40	3.40	3.45	3.40	3.20	3.25	3.50	3.30	3.20
3.35	3.30	3.35	3.10	3.60	3.45	3.55	3.40	3.35	3.30
3.00	3.40	3.50	3.65	3.55	3.60	3.65	3.10	3.40	3.40

Verilerin incelenmesinden görüleceği gibi en küçük değer 2.85 ve en büyük değer de 3.95'dir.

Eğer 10 sınıf yapılması istenirse;

$$\text{Sınıf aralığı} = \frac{3.95 - 2.85}{10} = 0.11 \cong 0.1$$

Sınıf aralığının 0.1 kg alınması hem diğer sınıfların sınırlarının belirlenmesinde hem de ileride yapılacak hesapların kolaylığı yönünden yerinde bir karardır.

Daha önce belirlenen ilkeler doğrultusunda 120 bebeğin doğum ağırlığına ilişkin frekans dağılımı aşağıdaki Tablo'da ki gibidir.

Daha önce de belirtildiği gibi birbirini takip eden iki sınıfın üst ve alt sınırları çakışmamak ve aralarında boşluk olmamalıdır.

Sınıflar	Sınıf değeri	f_i	% Frekans
2.85 - 2.94	2.895	1	0.83
2.95 - 3.04	2.995	2	1.67
3.05 - 3.14	3.095	9	7.50
3.15 - 3.24	3.195	15	12.50
3.25 - 3.34	3.295	19	15.83
3.35 - 3.44	3.395	24	20.00
3.45 - 3.54	3.495	19	15.83
3.55 - 3.64	3.595	14	11.67
3.65 - 3.74	3.695	8	6.67
3.75 - 3.84	3.795	5	4.17
3.85 - 3.94	3.895	3	2.50
3.95 - 4.04	3.995	1	0.83

Grafikler

Frekans dağılım tablosu düzenlendikten sonra grafikler çizilebilir.

Eğer frekans dağılım tablosu sürekli gözlem değerleri için düzenlenmişse bu durumda araştırmacı "POLİGON" veya "HISTOGRAM" çizilebilir.

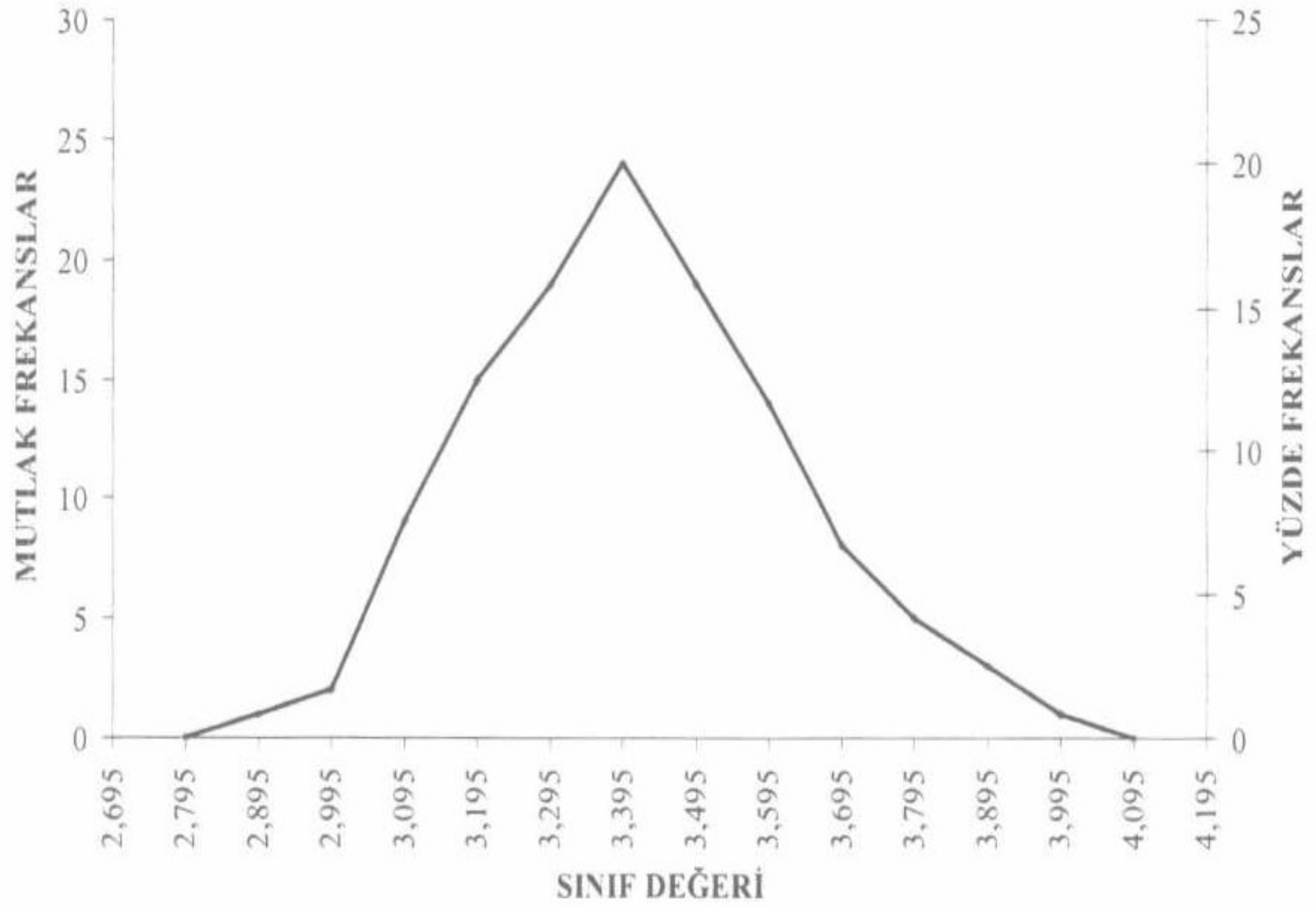
Fakat düzenlenen frekans dağılım tablosu kesikli (tam) gözlem değerleri kullanılarak hazırlanmışsa bu durumda sadece "ÇUBUK DİYAGRAM" çizilebilir.

Poligon

Frekans dağılım tablosu hazırlandıktan sonra eğer araştırmacı isterse poligon grafiği çizebilir.

Frekans poligonu çizilirken X eksenine "sınıf değerleri" ve Y eksenine de her sınıfa ait mutlak veya nisbi (%) "frekans" konularak grafik çizilir.

120 bebeğin doğum ağırlığı için düzenlenen frekans dağılım tablosundan çizilen frekans poligonu Şekil'deki gibidir.



Frekans poligonu çizilirken, poligon frekans dağılım tablosundaki birinci sınıftan önce olduğu varsayılan sınıfın, sınıf değerinden başlar, en son sınıftan sonra mevcut olduğu varsayılan sınıfın sınıf değerinde biter.

Yukarıda açıklandığı şekilde eklemeli frekans poligonları da çizilebilir. 120 bebeğin doğum ağırlığı için düzenlenen mutlak ve nisbi eklemeli frekans poligonları aşağıdaki Şekil 'deki gibidir.

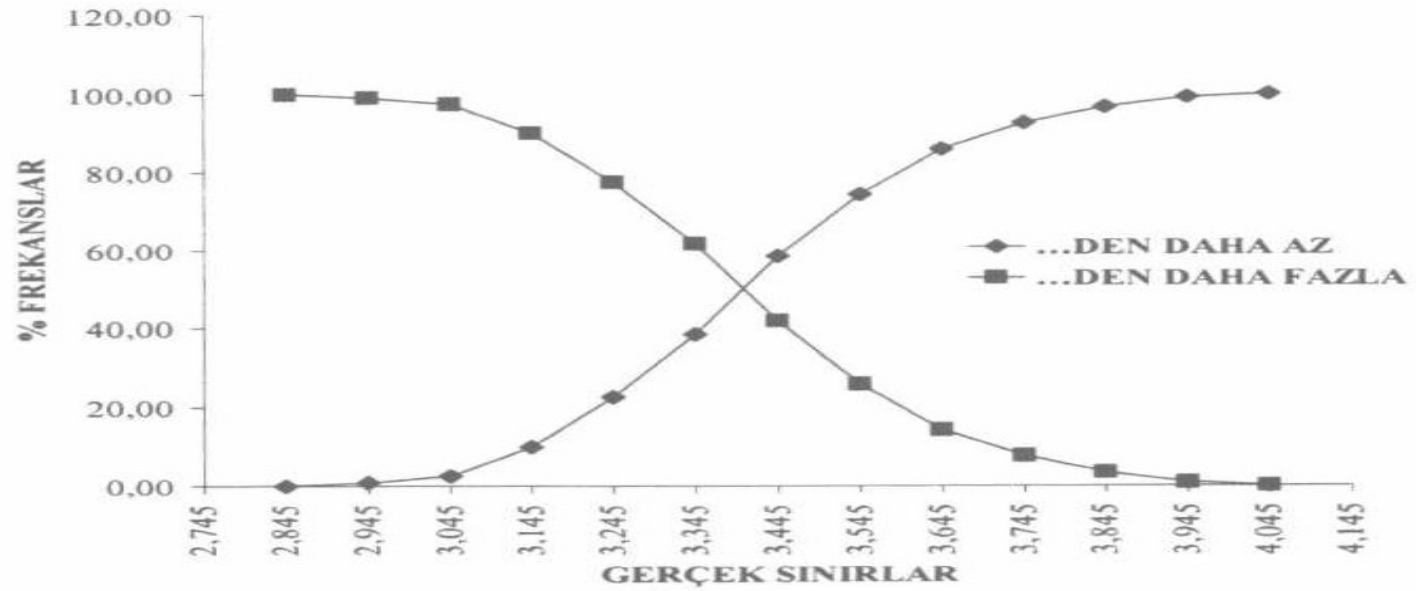
Bu grafikte, belirli bir doğum ağırlığından daha az veya daha çok doğum ağırlığına sahip olan bebeklerin sayısı kolayca görülebilir.

Şekil'den de görüldüğü gibi, nisbi "...den daha az" ve "...den daha fazla" eklemeli poligonları %50 noktasında birbirini keser.

a.



b.

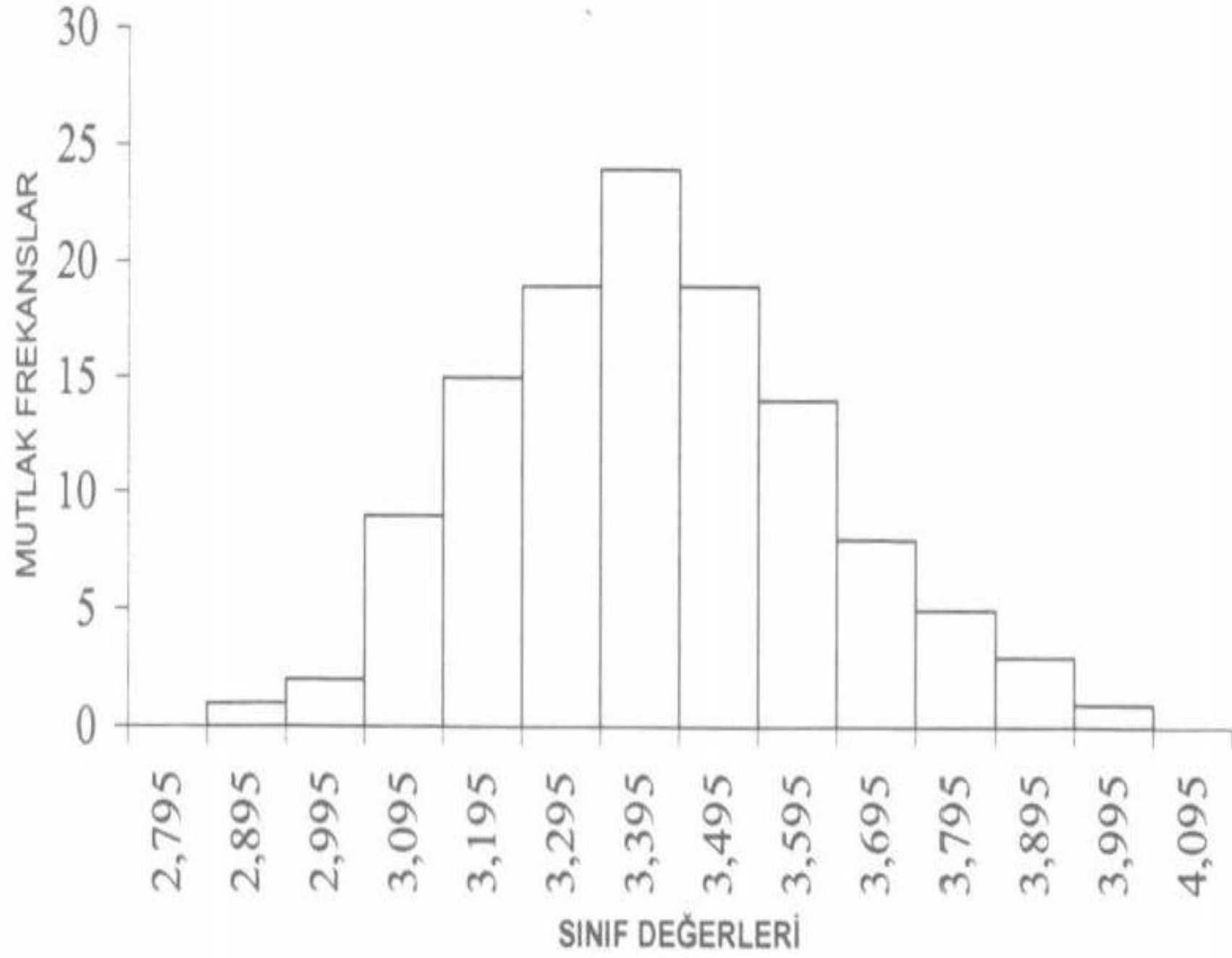


Histogram

Düzenlenen bir frekans dağılım tablosu için histogram da çizilebilir.

Histogram çizilirken X-eksenine gerçek sınırlar ve Yeksenine her sınıfa ait frekanslar konur ve dikdörtgenler oluşturulur.

Dikdörtgenlerin orta noktaları Şekil'de verilen histogramda da görüldüğü üzere sınıf değerlerine karşılık gelir.

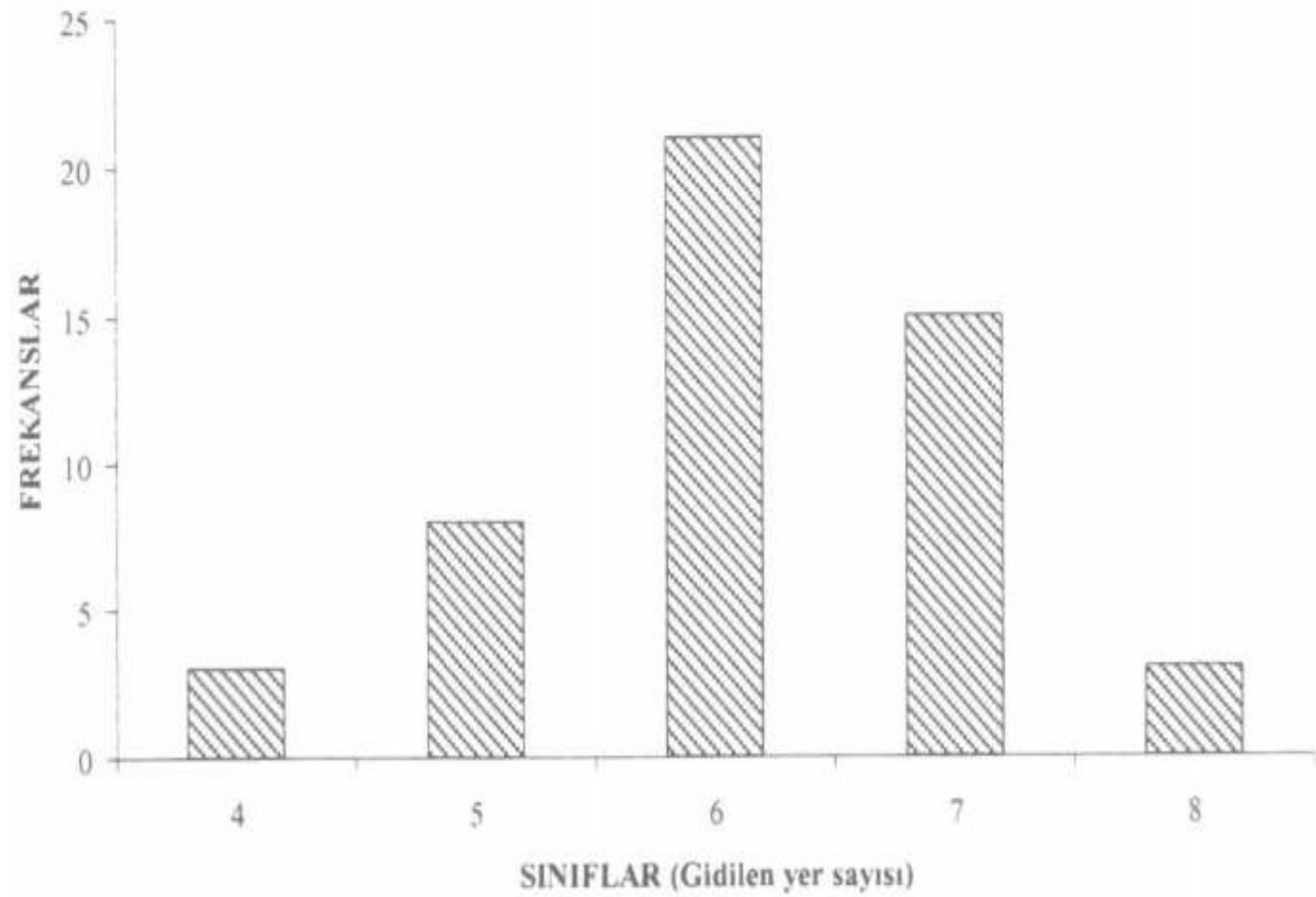


Çubuk Diyagram

Eğer düzenlenen frekans dağılım tablosu tamsayı gözlem değerleri içinse, bu frekans dağılım tabloları için çubuk diyagram çizilir.

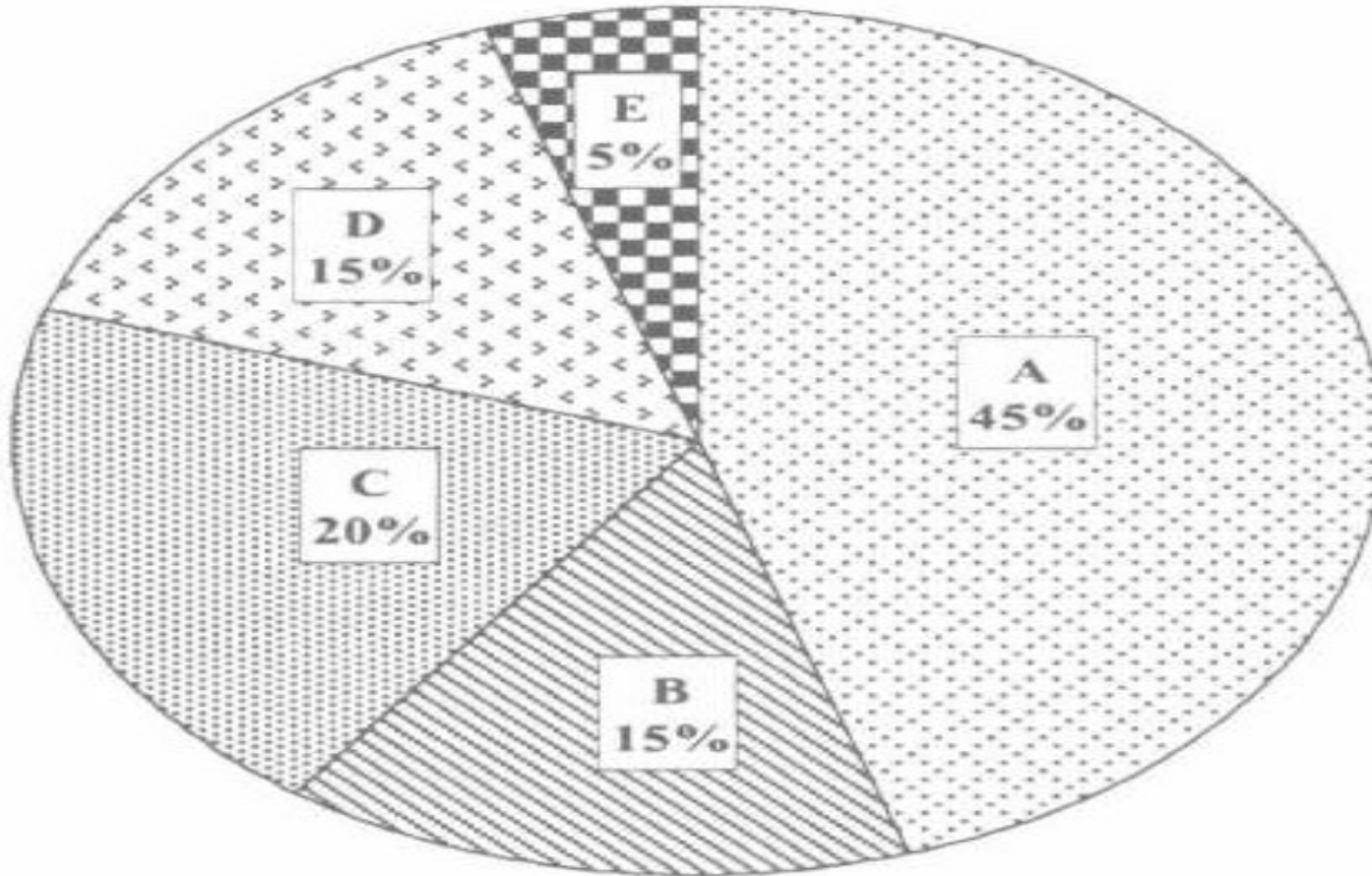
X-eksenine sınıflar ve Y-eksenine sınıflara ait frekanslar konur.

Çubuk diyagram çizildikten sonra çubuklar çizgi ile birleştirilmez çünkü tamsayı gözlem değerleri söz konusu iki değer arasında her hangi bir değeri alamaz.



Çember (Pasta Dilimi) Grafik

Pasta dilimi şeklindeki grafikte bir bütünün kısımları pastanın dilimleri şeklinde gösterilir.



TANITICI İSTATİSTİKLER

Araştırmacı örneğini oluşturacak bireyleri seçip, verilerini topladıktan sonra, bu örneğinin özelliklerini tanımlayacak değerleri hesaplar.

Bu hesaplanan değerlere tanıtıcı istatistikler adı verilir.

Bir örnek için hesaplanan tanıtıcı istatistikler merkezi eğilim ölçüleri (ortalamalar) ve değişim ölçüleri olmak üzere iki grupta toplanır.

Merkezi Eğilim Ölçüleri (Ortalamalar)

Ortalamalar, merkezi eğilim ölçüleri olarak da adlandırılır.

Çünkü gözlem değerleri ortalama etrafında toplanma eğilimi gösterilirler.

Merkezi eğilim ölçüleri, bir örnekteki bireyleri temsil eden tipik değerlerdir.

Merkezi eğilim ölçüleri aşağıda sıralandığı gibi 5 gruba ayrılır:

1. Aritmetik Ortalama,
2. Ortanca değer (Median),
3. Tepe Değeri (Mode),
4. Geometrik Ortalama,
5. Harmonik Ortalama.

Eğer sadece "ortalama" denirse bunun ile kastedilen "aritmetik ortalama"dır.

Aritmetik Ortalama

Aritmetik ortalama, n tane gözlem değerinin toplamının n 'e bölümüdür ve aşağıdaki şekilde gösterilir:

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} \quad \dots (2.1)$$

Eşitlikte, x_i , örnekteki i inci gözlem değeri, n , toplam gözlem sayısıdır.

ÖRNEK 1: Bir sınıftan tesadüfi olarak seçilen 10 öğrencinin ağırlıkları aşağıdaki gibi bulunmuş olsun.

49 51 55 50 57 60 55 50 62 56

Bunların aritmetik ortalaması;

$$\bar{X} = \frac{49 + 51 + 55 + 50 + 57 + 60 + 55 + 50 + 62 + 56}{10} = \frac{545}{10} = 54.5$$

ÖRNEK 2: Verilerin özetlenmesi açıklanırken 80 öğrencinin kalp atışları örnek olarak verilmişti.

Araştırmacı bu verilerini özetledikten ve grafik ile sunduktan sonra dakikadaki ortalama kalp atışı sayısını hesaplamak isteyebilir.

Bu durumda, aritmetik ortalama, 80 değer toplamının gözlem sayısına bölümü ile aşağıdaki şekilde bulunabilir:

$$\bar{X} = \frac{87 + 98 + 81 + 83 + \dots + 50 + 74 + 66}{80}$$

$$\bar{X} = \frac{6146}{80} = 76.8$$

Frekans Dağılım Tablosundan Ortalamanın Hesaplanması

Eğer herhangi bir örnekte verilerin frekansları farklı ise aritmetik ortalamanın hesaplanmasında bu dikkate alınır.

Örnekte gözlenenler ve frekansları aşağıdaki gibi olsun:

X_i	6	8	10	12	14
f_i	1	2	4	3	1

Burada aritmetik ortalama:

$$\bar{X} = \frac{1 \times 6 + 2 \times 8 + 4 \times 10 + 3 \times 12 + 1 \times 14}{1 + 2 + 4 + 3 + 1} \cong 10.2' \text{ dir.}$$

Araştırmacı örneğinden elde ettiği değerler için frekans dağılım tablosu düzenlemiş ise aritmetik ortalamayı benzer şekilde frekans dağılım tablosundan hesaplayabilir.

Bu amaçla aşağıdaki eşitlik kullanılır:

$$\overline{X} = \frac{\sum_{i=1}^k f_i X_i}{\sum_{i=1}^k f_i}$$

Eşitlikte;

f_i ; i . sınıfın frekansı,

k , frekans dağılım tablosundaki sınıf sayısı,

X_i , i . sınıfın sınıf değeridir.

Frekans dağılım tablosundan ortalama hesaplanırken, her bir sınıf değerinin aritmetik ortalamaya katkısı, söz konusu sınıfın frekansı ile tartılır.

ÖRNEK 3: 120 bebeğin doğum ağırlığı ile ilgili örnek için araştırmacı frekans dağılım tablosunu düzenlemişse ortalamayı hesaplayabilir.

Sınıf Değeri (X_i)	Frekans (f_i)	$X_i f_i$
2.895	1	2.895
2.995	2	5.990
3.095	9	27.855
3.195	15	47.925
3.295	19	62.605
3.395	24	81.480
3.495	19	66.405
3.595	14	50.330
3.695	8	29.560
3.795	5	18.975
3.895	3	11.685
3.995	1	3.995

$$\bar{X} = \frac{2.895 + 5.990 + 27.855 + \dots + 11.685 + 3.995}{1 + 2 + 9 + \dots + 3 + 1}$$

$$\bar{X} = \frac{409.7}{120} = 3.414 \text{ kg}$$

Örnekten elde edilen veriler tek tek toplanarak hesaplanan ortalama ile frekans dağılım tablosundan hesaplanan ortalama arasında küçük bir farklılık olacaktır.

Bunun sebebi frekans dağılım tablosu düzenlenirken verilerin özetlenmiş olmasıdır.

Eğer veriler tek tek toplanarak aritmetik ortalama hesaplanırsa 3.403 kg olarak bulunur.

Aritmetik Ortalamanın Özellikleri

1. Bir örnekteki gözlem değerlerinin örnek ortalamasından farkları alınır ve toplanırsa bu sapmaların toplamı sıfır olarak bulunur.

Yani, $\sum (X_i - \bar{X}) = 0$ 'dır.

2. Bir örnekteki gözlem değerlerinin hepsine bir A sabiti eklenirse, elde edilen değerlerin ortalaması, önceki ortalamadan A kadar fazla olacaktır. Yani;

$$y_i = x_i + A \text{ ise}$$

$$\bar{y} = \bar{x} + A \text{ olarak bulunur.}$$

Aritmetik Ortalamanın Özellikleri

3. Bir örnekteki gözlem değerlerinden hepsinden bir A sabiti çıkarılırsa, elde edilen değerlerin ortalaması, önceki ortalamadan A kadar az olacaktır. Yani:

$$y_i = x_i - A \text{ ise}$$

$$\bar{y} = \bar{x} - A \text{ olarak bulunur.}$$

4. Bir örnekteki gözlem değerlerinin hepsi c gibi bir sabit ile çarpılırsa, elde edilen değerlerin ortalaması önceki değerlerin ortalamasının c katı kadardır:

$$y_i = cx_i$$

$$\bar{y} = c\bar{x}$$

Aritmetik Ortalamanın Özellikleri

5. Bir örnekteki gözlem değerlerinin hepsi c gibi bir sabite bölünürse, elde edilen değerlerin ortalaması önceki değerlerin ortalamasının c ile bölümüne eşit olacaktır:

$$y_i = \frac{x_i}{c}$$

$$\bar{y} = \frac{\bar{x}}{c}$$

Kısa Yoldan Aritmetik Ortalamanın Hesaplanması

Bir frekans dağılım tablosu düzenlendikten sonra aritmetik ortalamaya ait özelliklerden de yararlanarak aritmetik ortalama kısa yoldan hesaplanabilir.

Kısa yoldan aritmetik ortalama hesaplanırken frekans dağılım tablosunda her sınıfa ait sınıf değerleri işlem yapması kolay "b" değerlerine aşağıdaki şekilde çevrilir:

$$b_i = \frac{X_i - A}{c}$$

Yani ortalamanın 3. ve 5. özellikleri birlikte uygulanır, A ve c özel olarak seçilir.

Böylece yeni değişken tam sayılardan oluşur.

X_i : i. sınıfın sınıf değeri,

A: frekans dağılım tablosunda ortalandaki bir sınıfın sınıf değeri,

C: sınıf aralığıdır.

A değerinin hangi sınıf değeri olacağı hakkında her hangi bir kural yoktur.

Fakat mümkün olduğu kadar ortadaki bir sınıf değerinin seçilmesi işlem yapılması daha kolay olan "b" değerlerinin bulunmasını sağlar.

Ortalamanın özelliklerine göre b_i 'lerin ortalaması $\bar{b} = \frac{\bar{X} - A}{c}$ dir.

x_i 'lerin ortalaması, $\bar{b}$, A ve c bilindiğine göre $\bar{X} = c\bar{b} + A$ şeklinde bulunur.

Frekans dağılım tablosundaki her sınıf için “b” değerleri hesaplandıktan sonra “b” değerlerine ait aritmetik ortalama aşağıdaki şekilde hesaplanır:

$$\bar{b} = \frac{\sum_{i=1}^k f_i b_i}{\sum_{i=1}^k f_i}$$

Ortanca Değer (Median)

Örnekten elde edilen gözlem değerleri büyüklüklerine göre sıralandığı zaman en ortadaki gözlem değeri ortanca değerdir (median).

Tüm gözlem değerlerinin yarısı bu değerden küçük diğer yarısı bu değerden büyüktür.

Eğer üzerinde çalışılan örnekte tek sayıda gözlem değeri varsa ortanca değer

$$\frac{(n+1)}{2}$$

ci gözlem değeridir.

Örnekteki gözlem değeri sayısının çift olduğu zaman ise ortanca değer,

$$\frac{n}{2} \text{ci}$$

gözlem değeri ile $(\frac{n}{2} + 1)$ ci gözlem değerinin ortalamasıdır.

Örnek 1: On beş bireylik bir örnekte elde edilen ölçümler aşağıdaki gibi bulunmuştur.

7 14 18 18 26 16 8 28 12 18 9 45 20 15 8

$n=15$

Bu örnek için ortanca değeri hesaplayabilmek için ilk olarak gözlem değerlerinin büyüklüklerine göre (küçükten büyüğe veya büyükten küçüğe doğru) aşağıdaki gibi sıraya konulması gerekir.

7 8 8 9 12 14 15 16 18 18 18 20 26 28 45

Gözlem değerleri büyüklüklerine göre sıraya konduktan sonra $\frac{(n+1)}{2} = 8$. gözlem değeri yani 16 bu örnek için ortanca değerdir.

Örnek 2: Eğer üzerinde çalışılan örnekte çift sayıdaki gözlem değeri büyüklüklerine göre aşağıdaki şekilde sıralanarak verilmiş ise;

7 8 8 9 15 18 25 29

$n=8$

Bu örnek için ortanca değer, 4'üncü (yani $\frac{n}{2}$ ci) ile 5'inci (yani $(\frac{n}{2} + 1)$ ci) gözlem değerinin ortalamasıdır. Bunun değeri de $\frac{9 + 15}{2} = 12$ 'dir.

Frekans Dağılım Tablosundan Ortanca Değerin Hesaplanması

Frekans dağılım tablosunda ortanca değer $n/2$ 'ci gözlem değerinin bulunduğu sınıftadır.

Ortanca değer hesaplanırken birinci sınıftan başlayarak frekanslar toplanır ve $n/2$ 'ci gözlem değerinin dahil olduğu sınıf ortanca değer bulundugu sınıftır.

Ortanca değerin frekans dağılım tablosundan hesaplanması 120 bebeğin doğum ağırlığı için düzenlenen frekans dağılım tablosu kullanılarak açıklanacaktır.

Bu örnekte doğum ağırlıkları büyüklük sırasına konduğu zaman değerlerin 60'ından küçük, 60'ından büyük olan değer ortancadır.

Daha önce verilmiş olan frekans dağılım tablosunda "...den daha az« eklemeli frekanslarına bakıldığı zaman 60. değer 7'inci sınıfta olduğunu görülür.

Bu sınıfın sınırlarını ve eklemeli frekanslarını aşağıdaki gibi göstermek mümkündür.



Sınıf aralığında 24 gözlem değerinin eşit aralıklar ile dağıldığını kabul ederek 60'ncı gözlem değerini (yani ortancayı) bulmak için aşağıdaki şekilde bir orantı kurulur.

Burada ortanca değer 46. gözlem değerinden, yani 3.345 kg'dan sonra 14 frekansın kapladığı sınıf aralığı kadar daha fazladır.

Aşağıda kurulan orantıda 14 frekansın 0.0583 birimlik sınıf aralığına dağılmış olduğu bulunmuştur.

Bu değer 46. gözlem değerinin bulunduğu sınıfın gerçek sınırına eklenerek ortanca hesaplanır.

0.1 sınıf aralığında
x

24 frekans
14 frekans

$$X = \frac{(0.1)(14)}{24} \cong 0.0583$$

ve Ortanca değer=3.345+0.0583=3.4033 olarak bulunur.

Frekans dağılım tablosundan ortanca değer hesaplamak için aşağıdaki formül de kullanılabilir.

Bu eşitlik yukarıda açıklanan çözümün formül şeklinde ifadesidir.

$$\text{Ortanca Değer} = L_1 + \frac{(n : 2) - (\sum f)_1}{f_{OD}} c$$

L_1 ; ortanca değerin bulunduğu sınıfın alt gerçek sınırı,

$(\sum f)$; ortanca değerin bulunduğu sınıftan önceki sınıfların frekansları toplamı,

f_{OD} , ortanca değerin bulunduğu sınıfın frekansı,

n ; toplam gözlem sayısı,

c ; sınıf aralığıdır.

$$\text{Ortanca deęer} = 3.345 + \frac{120 / 2 - 46}{24} (0.1) = 3.4033$$

Eęer 120 deęer sıraya konur 60'inci ve 61'inci deęerlerin ortalaması alınırsa 3.400 olarak bulunur.

Aradaki önemli olmayan fark, frekans dağılım tablosunda verilerin özetlenmesinden ileri gelir.

Frekans dağılım tablosu düzenlendikten sonra araştırmacı 120 bebekten en az doğum ağırlığına sahip %25'nin hangi deęerden daha az ağırlığa sahip olduğunu (yani ilk çeyrek deęeri) hesaplamak isteyebilir.

Ortanca değerin bulunmasına benzer şekilde bu değeri de hesaplanabilir.

120 bebeğin %25'i 30 bebek demektir.

Yani, 120 bebeğe ait doğum ağırlığı küçükten büyüğe doğru sıralandığı zaman bebeklerin %25'i 30'uncu sıradaki bebekten daha az doğum ağırlığına sahiptir.

Daha önce verilen frekans dağılım tablosunda 30. bebek frekansı 19 olan sınıftadır.

$$\text{İlk çeyrek değeri} = 3.245 + \frac{120 / 4 - 27}{19} (0.1) = 3.261$$

120 bebekten en fazla doğum ağırlığına sahip %25'nin hangi değerin üzerinde (yani, üçüncü çeyrek değeri) olduğunu bulmak için de küçükten büyüğe doğru sıralanmış olan doğum ağırlıklarından 90.'sını (120'nin %75'i 90'dır) hesaplamak gerekir.

Daha önce verilen Tablo'daki frekans dağılım tablosunda 90. gözlem değeri frekansı 14 olan sınıftadır.

Yine (2.4) numaralı eşitliğe benzer şekilde söz konusu doğum ağırlığı:

$$\text{Üçüncü çeyrek (\%75'lik) değer} = 3.545 + \frac{120(3/4) - 89}{14}(0.1) = 3.552$$

Ortanca Değerin Kullanıldığı Durumlar

Bazı durumlarda bir örnek için aritmetik ortalama iyi bir tanıtıcı değer olmayabilir.

Bu durumlarda ortanca değer kullanılması gerekir.

Ortanca değer tercih edildiği durumlar aşağıdaki şekilde sıralanabilir:

* Üzerinde çalışılan örnekte gözlem değerlerinde büyük çarpıklık varsa ortanca değer kullanılır.

Çünkü aritmetik ortalama sapan değerlerden etkileneceği için söz konusu örnek için iyi bir tanıtıcı istatistik olamaz.

Örneğin her hangi bir bölgede 6 eczanenin günlük satışları 25 35 55 49 33 500 milyon TL olsun.

Aritmetik ortalama 116.2 milyon TL ($697/6=116.2$) olarak bulunur ki bu yanıltıcıdır.

Çünkü ortalama 500'den etkilenmiştir.

Bu ve bu gibi sapan değerlerin bulunduğu örneklerde ortanca değer kullanılır. Bu örnek için ortanca değer 42 milyon TL'dir ve örneği daha iyi tanımlar.

- Bazı durumlarda araştırmacı bir veya iki ucu açık frekans dağılım tablosu düzenlemek durumunda kalabilir.

Bir veya iki ucu açık frekans dağılım tablolarında aritmetik ortalama hesaplanamaz.

Bu gibi durumlarda da söz konusu frekans dağılım tablosu için ortanca değer hesaplanabilir.

Örneğin araştırmacı 195 eczanenin aylık net gelirleri için düzenlenen ve Tablo'da verilen iki ucu açık bir frekans dağılım tablosu üzerinde çalışıyor ise ancak ortanca değeri hesaplayabilir.

Aylık gelir (milyon TL)	Frekans
<10	20
10-24	38
25-39	46
40-54	31
55-69	18
70-84	12
85-99	18
>99	12

$$\text{Ortanca deęer} = 24.5 + \frac{(195)/2 - 58}{46}(15) \cong 37.38 \text{ olarak bulunur.}$$

- Örneğe seçilmiş olan tüm ünitelerden veri elde etmek için uzun zaman gerekiyorsa bu gibi durumlarda da ortanca değer hesaplanır.

Örneğin, bir çeşit kanser için geliştirilen ilaçla tedavi edilmekte olan 8 hastanın, tedaviye başladıktan sonraki yaşam süreleri aşağıdaki gibi olsun:

Hasta no	6	7	3	2	1	4	5	8
Yaşam süresi (ay)	16	16	18	22	26	Yaşıyor	Yaşıyor	Yaşıyor

Görüldüğü gibi araştırmaya başladıktan 26 ay sonra hastalardan üçü yaşamaktadır.

Bu durumda aritmetik ortalama hesaplanamaz, ancak sıraya konmuş bu sonuçlardan 4. ve 5. sıradakilerin ortalaması olan ortanca değer (burada 24 ay) hesaplanabilir.

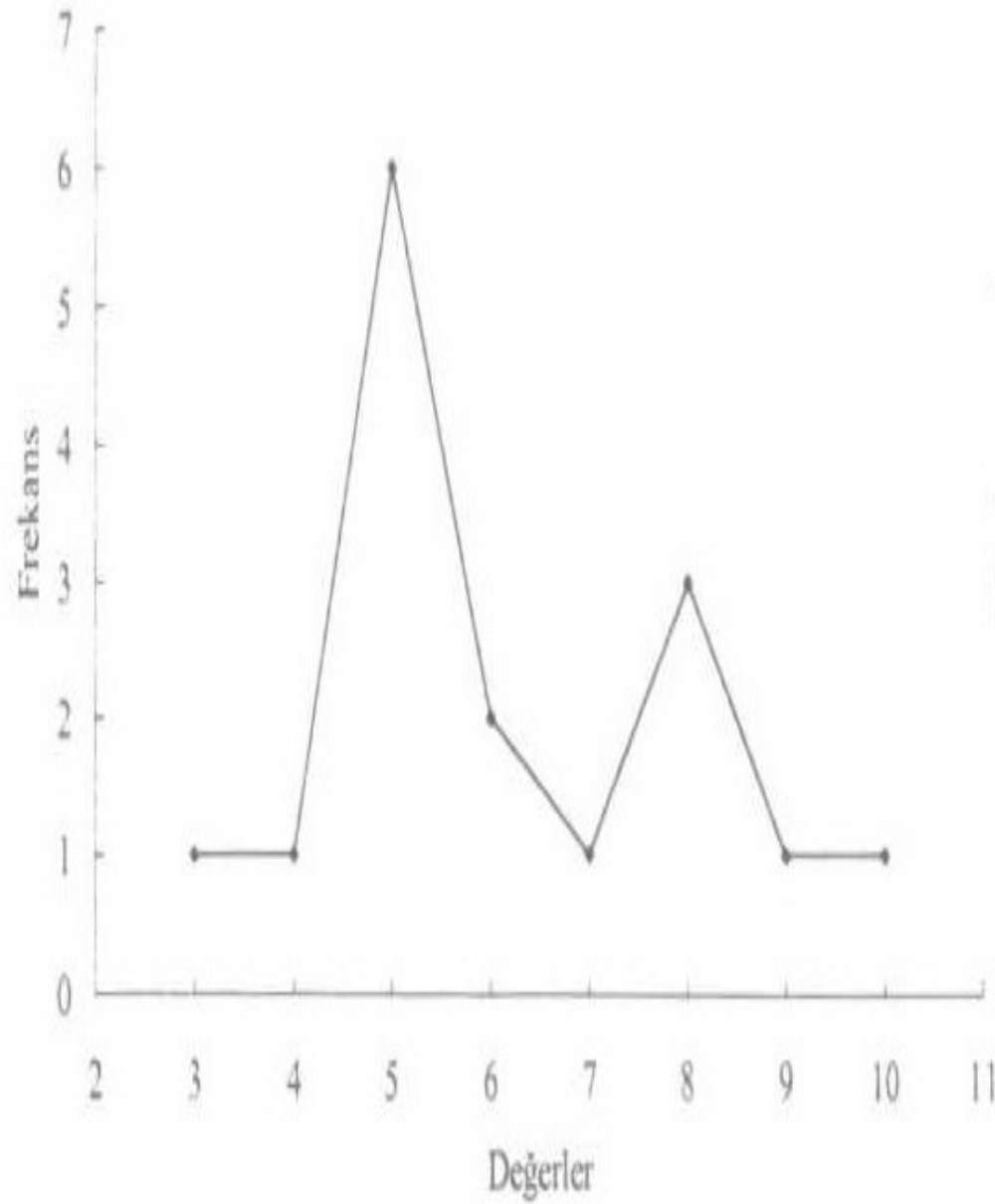
Tepe Değeri (Mode)

Bir örnekte frekansı en yüksek yani en çok tekrarlanan gözlem değeri tepe değeridir.

Bir örneği oluşturan değerler;

5, 6, 7, 5, 4, 8, 8, 8, 5, 10, 5, 5, 6, 3, 9 ve 5

ise bu örnekte iki tepe değeri vardır ve bunlar 5 ve 8'dir.



Yandaki şekilde görüldüğü gibi 5 ve 8 bu örnekteki tepe değerleridir.

Frekans dağılım tablosu düzenlendiği zaman bir örneğin tepe değeri, frekans dağılım tablosunda frekansı en yüksek olan sınıfın sınırları içindedir.

Tepe değerinin frekans dağılım tablosundan hesaplanması için aşağıda verilen eşitlik kullanılır:

$$\text{Tepe Değeri} = L_1 + \frac{d_1}{d_1 + d_2} c$$

Eşitlikte,

L_1 , tepe değerinin bulunduğu sınıfın alt gerçek sınırı,

d_1 , tepe değerinin bulunduğu sınıfın frekansı ile bir önceki sınıfın frekansı arasındaki fark,

d_2 , tepe değerinin bulunduğu sınıfın frekansı ile bir sonraki sınıfın frekansı arasındaki fark,

c ise sınıf aralığıdır.

ÖRNEK: 120 bebeğin doğum ağırlığı ile ilgili örnek için düzenlenen frekans dağılım tablosunda tepe değeri, frekansı 24 olan sınıfın sınırları içindedir.

Bu sınıfın alt gerçek sınırı 3.345, tepe değerinin bulunduğu sınıftan bir önceki sınıfın frekansı 19 ve bir sonraki sınıfın frekansı 19'dur.

$$\text{Tepe Değeri} = 3.345 + \frac{(24 - 19)}{(24 - 19) + (24 - 19)} (0.1) = 3.395$$

Görüldüğü gibi tepe değerinin bulunduğu sınıftan önce ve sonraki sınıfların frekansları eşit ise tepe değeri sınıf değerine eşittir.

Frekans dağılım tablosunda eğer ardışık iki sınıfın frekansı birbirine eşit ve en yüksek ise bu durumda ilk sınıfın alt gerçek sınırı alınır ve sınıf aralığı iki katına çıkarılır, yani bu iki sınıf tek bir sınıf gibi kabul edilerek tepe değeri hesaplanır.

Fakat ardışık olmayan iki sınıfın frekansı en yüksek ise bu durumda ayrı ayrı iki tepe değeri hesaplanır.

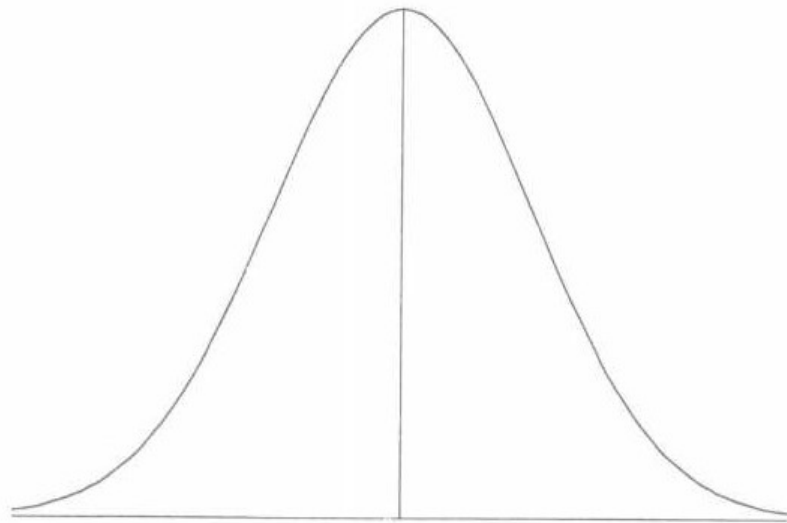
Aritmetik Ortalama, Ortanca ve Tepe Değeri Arasındaki İlişki

Aritmetik ortalama, ortanca ve tepe değerinin sıralanışı, örnekten elde edilen gözlem değerlerinin dağılımına bağlıdır.

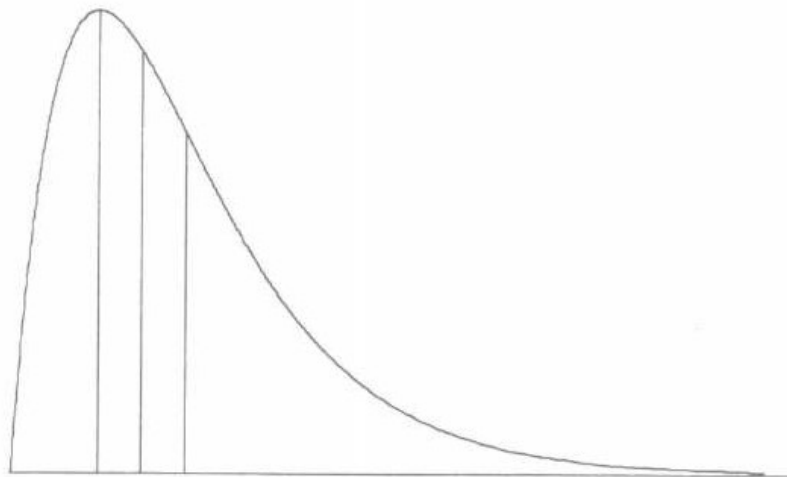
Eğer dağılım simetrik (Şekil 1) ise her üç ortalama da aynı değere sahiptir.

Şekil 2'de gösterildiği gibi dağılım sola yatık (çarpık) bir dağılım ise en küçük aritmetik ortalama, en büyük ise tepe değeridir.

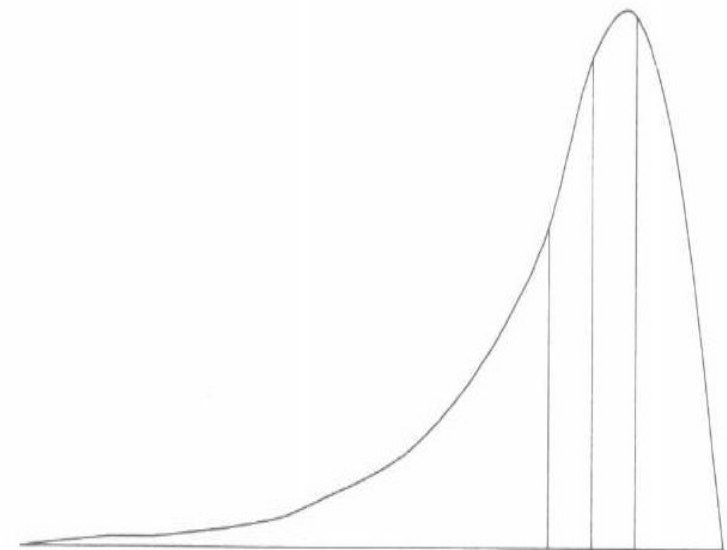
Dağılım sağa yatık (çarpık) (Şekil 3) olduğu zaman ise en büyük aritmetik ortalama, en küçük tepe değeridir.



A.O=T.D= O.D.



TD OD AO



AD OD TD

Yukarıdaki şekillerden de görüldüğü gibi dağılımın şeklinden en çok etkilenen aritmetik ortalamadır.

Aritmetik ortalama sapan değerler yönünde kayar.

Ortanca değer, gözlem değerleri büyüklüklerine göre sıralandığı zaman en ortadaki değer olduğu için dağılımın şeklinden etkilenmez.

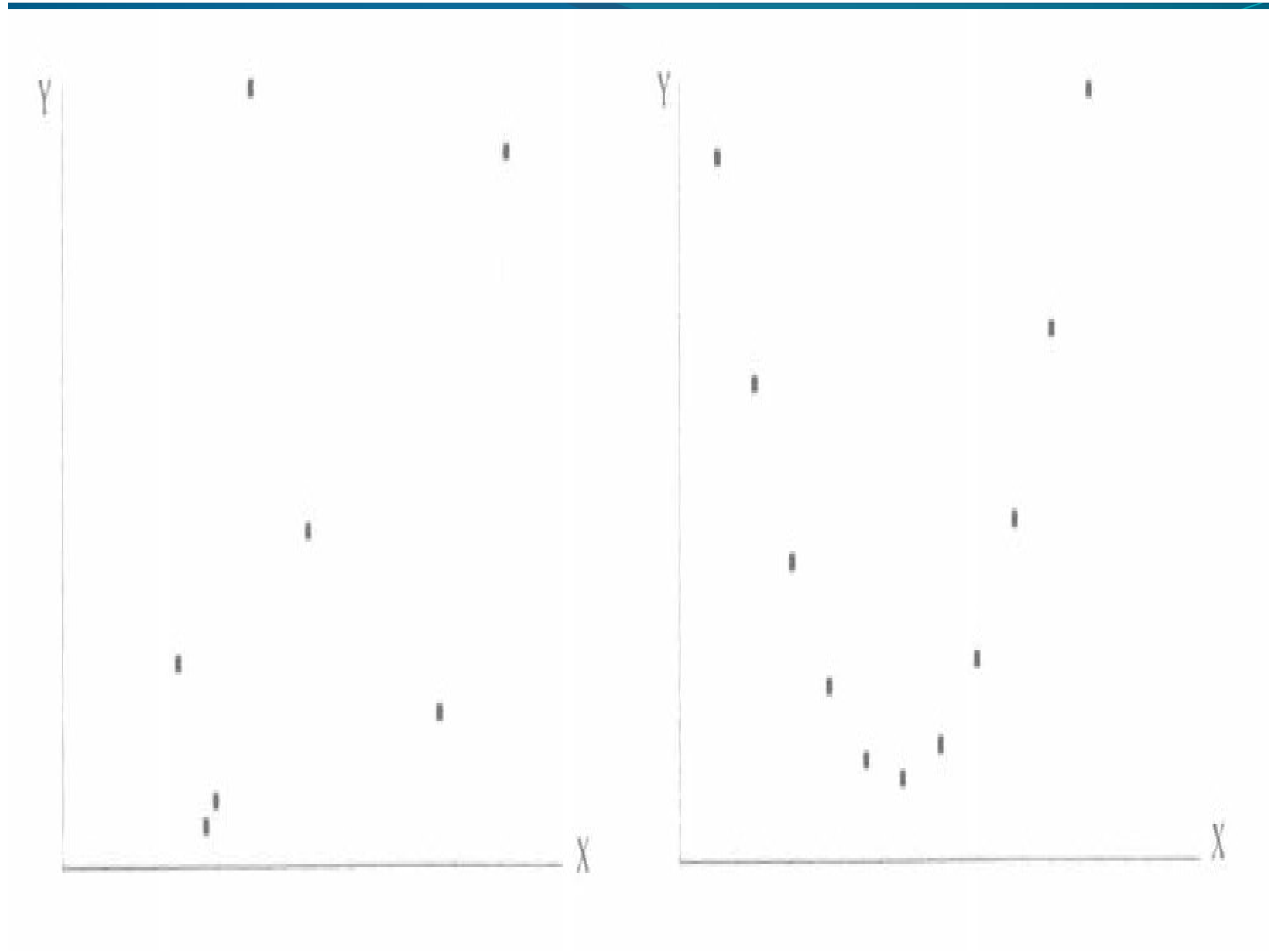
Dağılım simetrik olduğu zaman her üçü de örneğin özelliklerini aynı derecede tanımlarken, dağılım simetriden uzaklaştıkça ortanca değer daha güvenilir bir tanıtıcı istatistik haline gelir.

Korelasyon Katsayısı

Bir örnekten iki özelliğe ait toplanan verilerin koordinat sisteminde noktalar halinde gösterilmesi araştırmacıya iki özellik arasında bir ilişki olup olmadığı, eğer varsa ilişkinin şekli hakkında bir ön bilgi verecektir.

Eğer veriler koordinat ekseninde işaretlendiği zaman aşağıdaki şekil'de verilen grafiklerde görüldüğü gibi bir dağılım gösteriyorsa iki özellik arasındaki ilişkinin derecesini belirten korelasyon katsayısı (r) 0 veya 0'a çok yakın bir değerdir.

Bu iki özellik arasında doğrusal bir ilişki olmadığını gösterir.

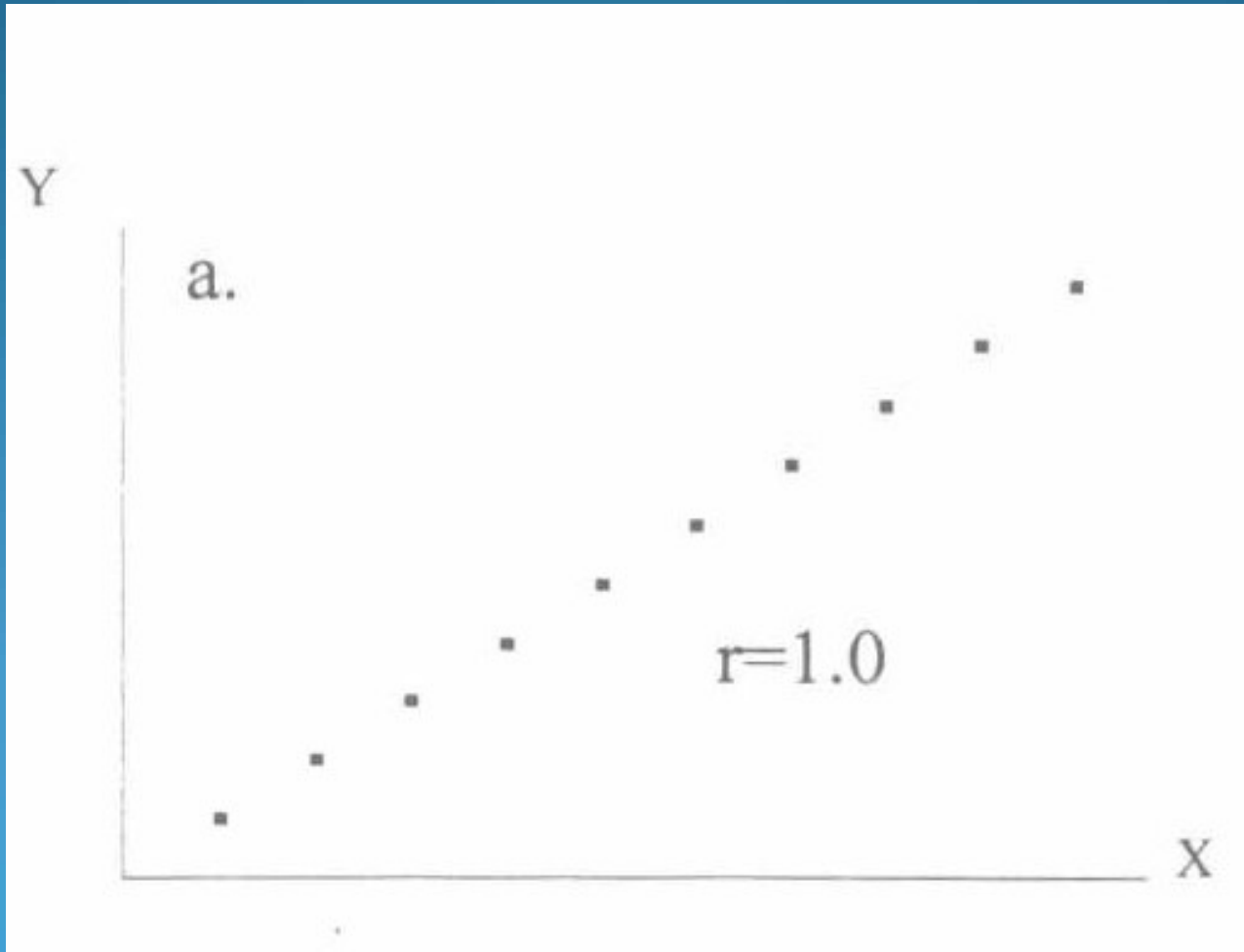


İki özelliğe ait veriler koordinat sisteminde işaretlendiği zaman aşağıdaki grafiklerde gösterildiği gibi noktalar bir doğru üzerinde sıralanabilir.

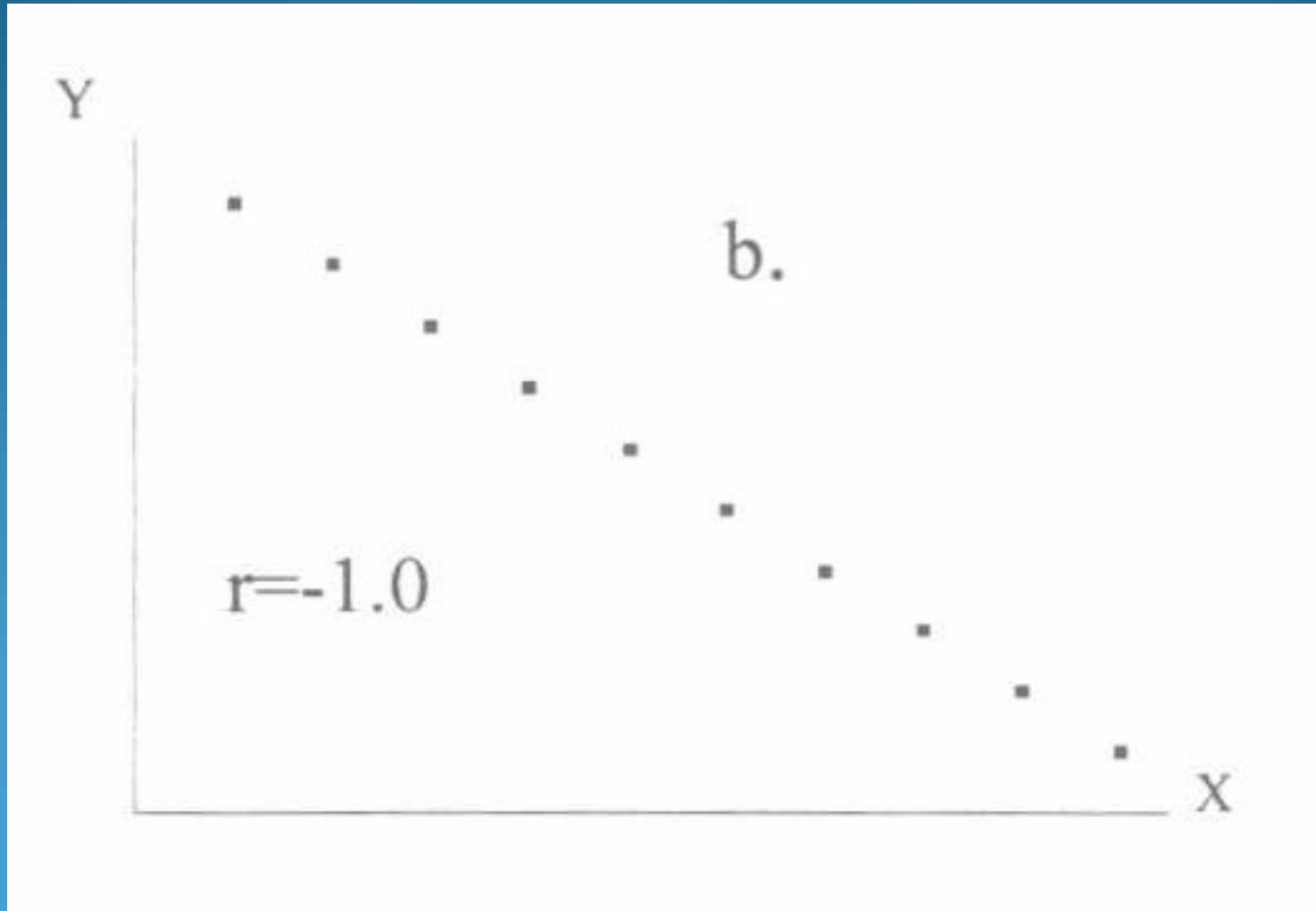
Bu, iki özellik arasında tam bir ilişki olduğunu gösterir, yani her X değerine karşılık gelen bir tek Y değeri vardır.

Tam ilişki olduğu zaman korelasyon katsayısı 1.0 olarak bulunur. İlişkinin yönünü ise korelasyon katsayısının işareti belirtir.

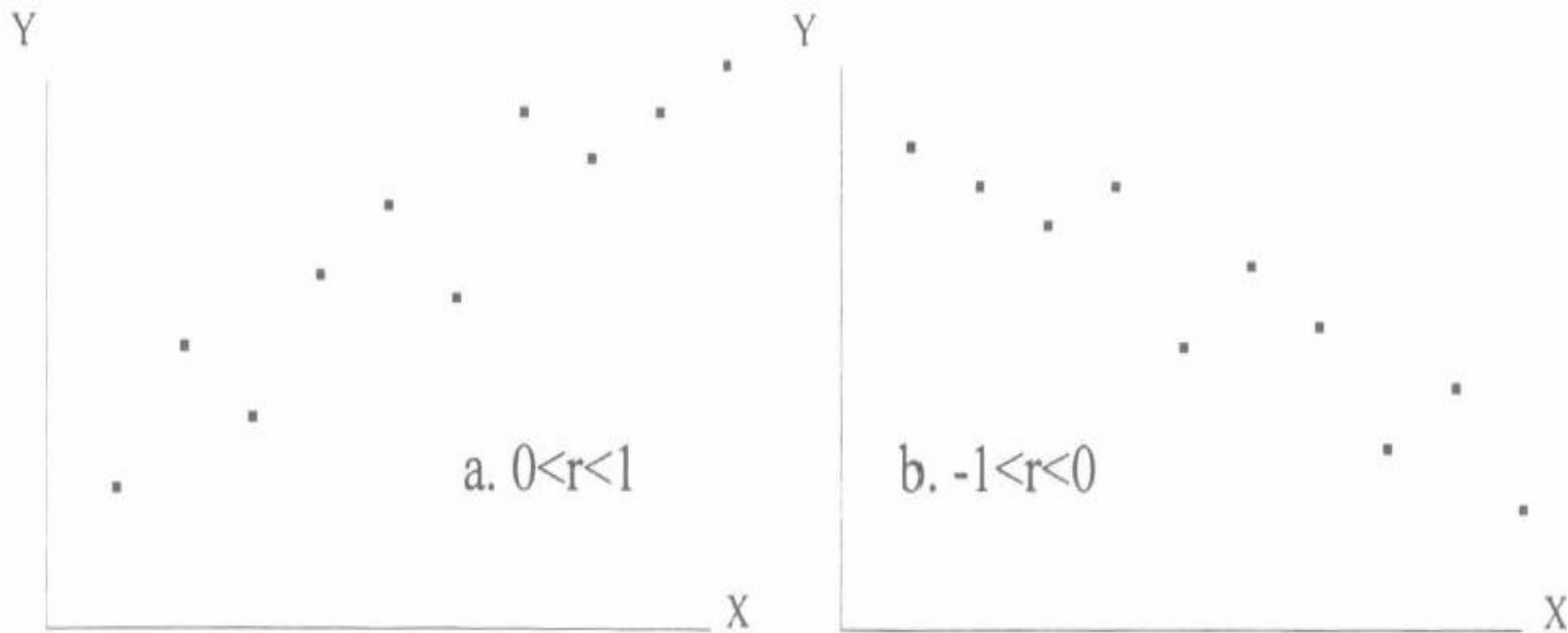
Eğer iki özellik arasındaki ilişki şekildeki gibi artan bir ilişki ise yani X arttıkça Y de artıyorsa korelasyon pozitif işaretlidir, yani $r=1.0$ olarak bulunur.



Fakat iki özellik arasındaki ilişki Şekil'de verildiği gibi negatif (ters, azalan) bir ilişki ise korelasyon katsayısı negatif işaretlidir, yani $r=-1.0$ olarak bulunur.



Veriler koordinat sisteminde işaretlendiği zaman



Şekil'de görüldüğü gibi iki özellik arasında tam bir ilişki yoktur, yani noktalar bir doğru üzerinde dizilmemişlerdir.

Bu durumda, eğer şekil a'da olduğu gibi iki özellik arasında tam olmayan pozitif bir ilişki varsa korelasyon katsayısı 0 ile +1 arasında bir değer alır.

İki özellik arasında negatif tam olmayan bir ilişki söz konusu olduğu zaman ise (Şekil b) korelasyon katsayısı -1 ile 0 arasındadır.

Yukarıda da açıklandığı gibi korelasyon katsayısı -1 ile +1 arasında değişir.

İlişki tam ise 1 değerini alır. Korelasyon katsayısının işareti ise ilişkinin yönünü gösterir.

Korelasyon katsayısının eşiti ise popülasyonda aşağıdaki gibidir:

Popülasyonda :

$$\rho = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum (x_i - \mu_x)^2 \sum (y_i - \mu_y)^2}}$$

Örnekte:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Ve kısaca aşağıdaki gibi gösterilir.

$$r = \frac{\sum d_x d_y}{\sqrt{\sum d_x^2 \sum d_y^2}}$$

Eşitlikte, $\sum d_x d_y$, çarpımlar toplamı, $\sum d_x^2$, X değişkenine ait kareler toplamı, $\sum d_y^2$, Y değişkenine ait kareler toplamıdır.

HİPOTEZ KONTROLLERİ

Bütün bilim dallarında yapılan araştırmaların amacı elde edilen örneklerden yararlanarak üzerinde çalışılan populasyonun bilinmeyen parametreleri hakkında sonuç çıkarmaktır.

Sonuç çıkarma iki şekilde yapılabilir:

populasyonu temsil edecek parametrelerin tahmini veya populasyon parametreleri ile ilgili hipotezlerin test edilmesi.

Hipotez kontrolleri, bilimsel metotların bir sonucu olarak ortaya çıkar. Bir bilim adamı üzerinde çalıştığı konuyu gözler, verilerini toplar, hipotezini kurar ve oluşturduğu hipotezini topladığı gözlemlerine karşı test eder.

Eğer yaptığı kontrol sonucunda gözlemleri ile hipotezi arasında bir uyum yoksa oluşturmuş olduğu hipotezi reddeder.

Yapılan kontrol sonucunda gözlemler ile hipotez uyum içinde ise kurmuş olduğu hipotezin geçerli olduğuna veya örnekten elde edilen sonuçların hipotezle belirtilen populasyondan çekilen örneklerden hesaplanabilen sonuçlarla uyum içinde olduğuna karar verir.

Verdiği bu kararlarda yanılma olasılıklarını hesaplar.

Hipotez kontrolleri, kurulan hipotezlerin gerçeklere karşı test edilebildiği bütün alanlarda yapılabilir.

Örneğin, bir ilaç fabrikasında, üretilen vitamin-mineral kombinasyonu drajelerde ortalama olarak 20 mg B₁ vitamini olması gerektiğini bilen bir yönetici gerçekten drajelerin ortalama olarak bu miktarda vitamin içerip içermediğini araştırmak isteyebilir.

Bunun için n-drajelik bir örnek alır.

Bu drajelerin B₁ vitamini miktarını belirler.

Daha sonra elde ettiği verilere dayanarak drajelerin ortalama olarak 20 mg B₁ vitamini içerip içermediğine karar verir.

Yani ortalama B₁ vitamini miktarının 20 mg'a eşit ($\mu_x = 20\text{mg}$) olduğu hipotezini kontrol ederek karar verir.

Ayrıca yönetici yanlış karar verme ihtimalinin ne olduğunu da bilmek ister.

Yapılan bir araştırma sonucunda bir hipotez kurulduğu zaman, bu hipotez yapılan kontrol sonucunda kabul veya reddedilir.

Bu karara varılabilmesi bir dizi istatistik işlemi gerektirir. Bu işlemleri 5 grupta toplamak mümkündür.

1. Hipotezlerin kurulması,
2. Yapılacak kontrolün çift veya tek taraflı kontrol olup olmadığına karar verilmesi,
3. Test istatistiğinin seçilmesi,
4. I. tip hata olasılığının belirlenmesi,
5. Gerekirse testin gücünü hesaplanması ve güç eğrilerinin belirlenmesi.

Hipotezler

Hipotez kontrolünün birinci aşaması, araştırmacının hipotezlerini kurmasıdır.

Hipotez kontrolünde iki tip hipotez vardır:

1. Kontrol hipotezi (H_0)

2. Karşıt (alternatif) hipotez (H_1 , birçok yabancı kitapta alternatif hipotez olarak geçer ve H_A ile gösterilir.)

Kontrol hipotezi adından da anlaşılacağı gibi kontrol edilebilir olmalı, bir karşıtı olmalı ve bir anlam taşımalıdır.

Yapılan kontrol sonucunda bu hipotez reddedilirse karşıt hipotez kabul edilir.

İstatistik hipotezler üzerinde çalışılan örnekten elde edilen verilerin durumları ile ilgilenir. Bir araştırmacının $X_1, X_2, x_3 \dots x_n$ olmak üzere verileri varsa bu verilerin bir dağılımı söz konusudur.

Eğer araştırmacı bu dağılımda bir bölge belirlerse X değerlerinin bu bölgeye dahil olma ihtimalini hesaplayabilir.

Kurulan hipotezler varsayılan dağılımın bir veya daha fazla parametresini kapsayabilir.

Bir parametreyi kapsarsa "basit hipotez", birden fazla parametreyi kapsarsa "bileşik hipotez" denir.

Bazı yazarlar basit hipotezleri (h) ile bileşik hipotezleri (H) ile gösterirler.

Ancak bu konuda yazarlar arasında tam bir birlik olduğunu söylemek mümkün değildir.

Test ve karşıt hipotezin basit ve bileşik olmasına göre hipotez testleri aşağıdaki gibi sıralanabilir:

1. Basit bir hipotezi basit bir karşıt hipoteze karşı test edilir. Test ortalama ile ilgili olarak ise,

$$H_0: \mu = a$$

$$H_1: \mu = b$$

şeklinde olabilir.

2. Basit bir hipotezin bileşik bir hipoteze karşı testi. Bu test aşağıdaki şekilde olabilir.

$$H_0: \mu = a$$

$$H_1: \mu \neq a$$

3. Bileşik bir hipotezin bileşik bir hipotezle test edilmesi. Bu test de aşağıdaki şekilde olabilir:

$$H_0: \mu \geq a$$

$$H_1: \mu < a$$

Hipotez testleri karşıt hipotezin kapsadığı yöne göre iki veya tek taraflı olabilir.

$$H_0: \mu = a$$

veya

$$H_0: \mu = a$$

$$H_1: \mu < a$$

$$H_1: \mu > a$$

testleri tek taraflıdır. Buna karşılık aşağıdaki test iki taraflıdır:

$$H_0: \mu = a$$

$$H_1: \mu \neq a$$

Test İstatistiği

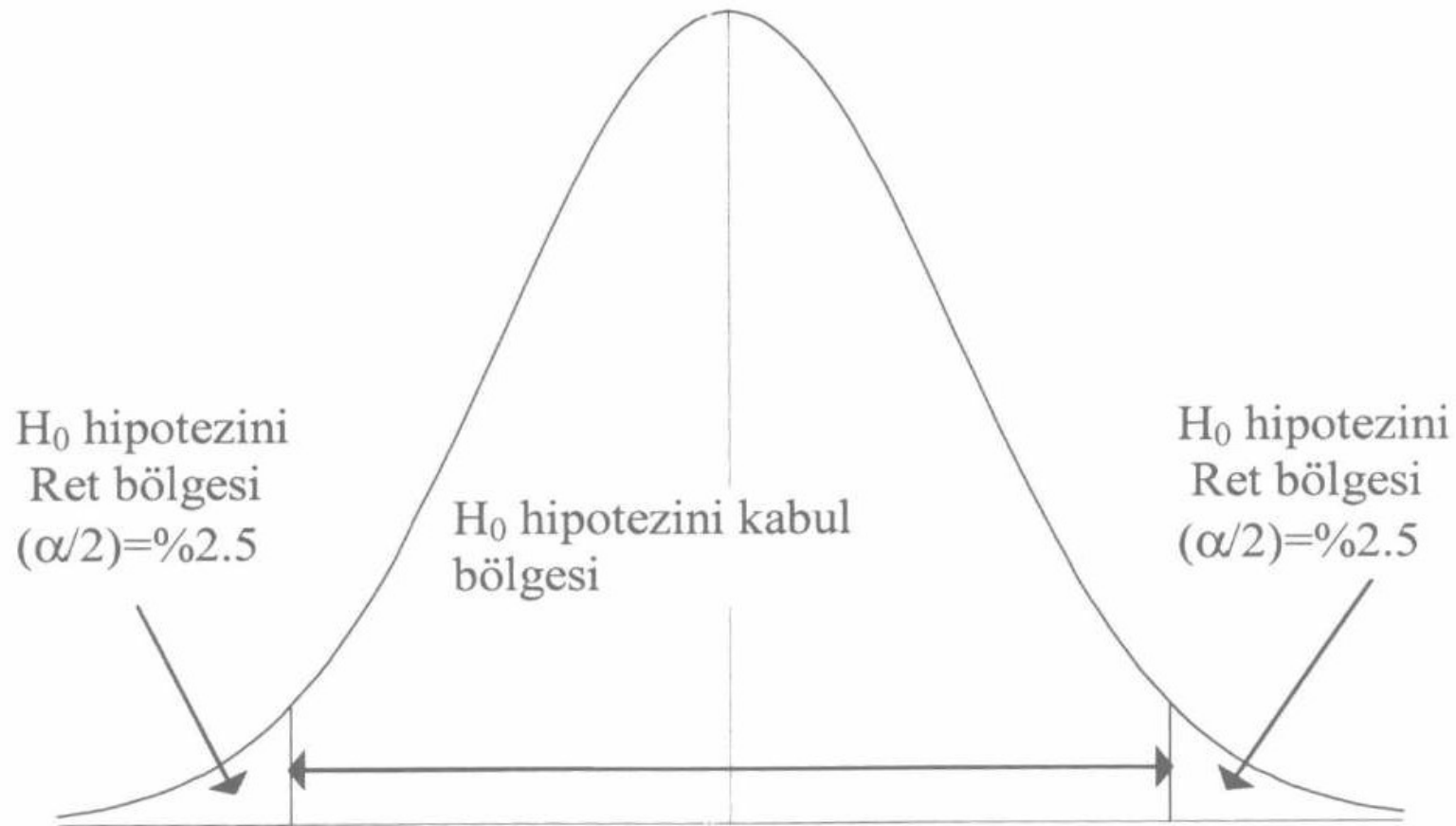
Hipotezler kurulduktan sonra araştıracının test istatistiğini belirlemesi gerekir.

Çünkü hipotez kontrolü, örnekten hesaplanan istatistiğin, söz konusu populasyondan aynı genişlikteki örneklerden elde edilecek örnekleme dağılımına dahil olma ihtimalini hesaplamak ve bu sonuca göre karar vermektir.

Test istatistiği örnekten elde edilen verilerin bir fonksiyonudur.

Araştıracının amacı ve elde edilen verilerin doğrultusunda hipotez kontrolünde kullanılacak test istatistiği belirlenir.

Kullanılacak test istatistiğinin doğru olarak belirlenmesi yapılan hipotez kontrolünün geçerli olması için önemlidir.



SAPAN DEĞERLER (OUTLIERS)

Araştırma yapan herhangi bir kimse topladığı rakamları yakından incelediğinde bazılarının incelediği gruptan olup olmadığından veya doğru toplanıp toplanmadığından kuşkulananabilir.

Bu gibi değerlere İngilizcede "OUTLIER" denir.

Bunun Türkçe karşılığı olarak "sapan değer" ve "kuşkulu gözlem" denmektedir. Her ikisi de uygun karşılıktır.

Ancak toplanmış rakamlar arasında bir veya birkaçından araştırmacı kuşkulaniyorsa bunları incelemeye almalıdır.

Eğer veriler bir analiz sonucu elde edilmişse ve numuneleri halen elde varsa kuşkulanılan rakamlara ait analizleri tekrarlamak en doğrusudur.

Yine aynı sonuçlar elde edilebilir.

Ancak numunelerin numaralarının karıştığından kuşkulanabilir.

Bazen toplanmış olan veriler arasından kuşkulanılanların tekrarlanma imkanı olmayabilir.

Bu gibi gözlemleri dikkate almayın demek de doğru olmaz.

Çünkü herkesin şüphe derecesi aynı değildir.

Ayrıca araştıracının eğilimi de kuşkulanma derecesini etkiler.

Araştırmacılar sapan değerler problemi ile çok eski tarihlerden beri ilgilenmişlerdir.

Bu konu ile ilgili birçok araştırma ve teorik çalışmalar yapılmıştır.

Bütün çalışmalarda amaç, eğer kuşkulanılan gözlemin ele alınan gruba dahil olma olasılığı çok düşük ise (kabul edilen α olasılığından az ise) bu rakamı değerlendirmelerde dikkate almamaktır.

Eğer ele alınan gruba dahil olma olasılığı önceden kabul edilen olasılıktan fazla ise şüphelenilen rakam değerlendirme dışı bırakılmamalıdır.

Şüphelenilen rakam birden fazla da olabilir, bu sıralanmış verilerde büyük tarafta veya her iki tarafta olabilir veya çok değişkenli olarak ele alınan verilerde sapan değer problemi ortaya çıkabilir.

Bunlar ve burada sayılmayan çok çeşitli durumları ele alan çalışmalar yapılmıştır.

Kİ-KARE (χ^2) DAĞILIMI ve bu DAĞILIM ile İLGİLİ TESTLER

Ki-Kare (χ^2) Dağılımı: İstatistikte önemli diğer bir sürekli dağılım ki-kare dağılımıdır.

Ki-kare değişkeni 0 ile sonsuz arasında dağılım gösteren tek taraflı bir değişkendir.

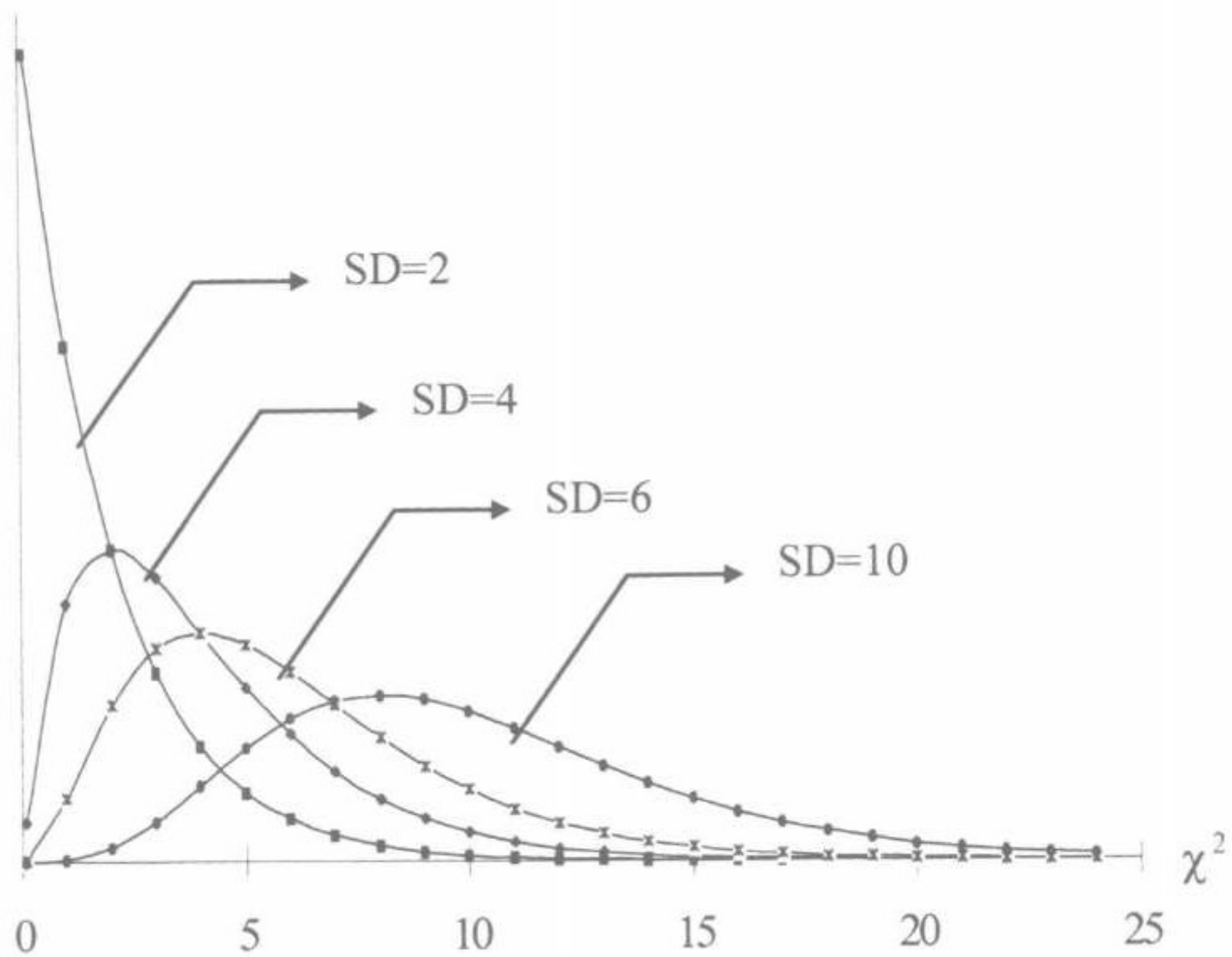
Ortalaması μ ve standart sapması σ olan normal dağılan bir popülasyondan çekilen her bir X değeri için $X - \mu / \sigma$ şeklinde hesaplanacak r tane bağımsız Z -değerinin karelerinin toplamının gösterdiği dağılım ki-kare (χ^2) dağılımı olarak adlandırılır.

$$Z_1^2 + Z_2^2 + Z_3^2 + Z_4^2 + \dots + Z_r^2 = \sum_{i=1}^r Z_i^2$$

Bu dağılımın fonksiyonu aşağıdaki gibidir:

$$f(\chi^2) = \frac{1}{2^{\frac{r}{2}} \left(\frac{r}{2}\right)!} (\chi^2)^{\frac{(r-2)}{2}} e^{-\frac{\chi^2}{2}} \dots$$

$$\sum_{i=1}^k \frac{(\text{Gözlenenfrekans}(f) - \text{Beklenenfrekans}(f'))^2}{\text{Beklenenfrekans}(f')} = \chi^2$$



Her serbestlik derecesi için farklı bir X^2 -dağılımı vardır.

Ki-Kare Kontrolleri

Bir örnekten elde edilen veriler belirtilen özelliklerine göre sınıflandırılabilir ve bir araştırmacı öne sürülen bir hipoteze göre her sınıfta olması beklenen frekans ile gözlenen frekanslar arasında uyum olup olmadığını kontrol etmek isteyebilir. Ki-kare kontrolleri 3 grup altında incelenebilir:

1. Homojenlik Kontrolleri
2. Dağılımlara Uyum Kontrolleri
3. Bağımsızlık Kontrolleri

Homojenlik Kontrolleri

ÖRNEK 1: Bir eczaneye bir hafta içinde gelen müşterilerin günlere göre dağılımı aşağıdaki gibi bulunmuştur.

Söz konusu eczaneye gelen müşterilerin sayısı günlere göre değişmekte midir?

Günler	Gözlenen frekans	Beklenen frekans	$(f-f')^2/f'$
1	22	26	0.165
2	29	26	0.346
3	31	26	0.962
4	21	26	0.962
5	27	26	0.038
6	26	26	0.000
Toplam	156	156	$\chi^2=2.923$

H_0 : Eczaneye bir hafta boyunca gelen müşterilerin günlere dağılımı homojendir.

Günler arasında eczaneye gelen müşteri sayısı bakımından fark önemli değildir.

H1: Eczaneye bir hafta boyunca gelen müşterilerin günlere dağılımı homojen değildir.

Günler arasında eczaneye gelen müşteri sayısı bakımından fark önemlidir.

Burada hangi hipotezi kabul edeceğine karar vermek için araştırmacının X^2 -değerini hesaplaması gerekir.

Bunun için de haftanın her günü için beklenen müşteri sayısının bulunması lazımdır.

Kontrol hipotezi ile araştırmacı müşteri sayısı bakımından haftanın günlerinin homojen olduğunu öne sürdüğüne göre, toplam müşteri sayısı (156) haftanın günlerine eşit olarak dağılmalıdır.

Yani her gün için beklenen frekans $156/6=26$ 'dır.

$$\begin{aligned}\chi^2 &= \frac{(22-26)^2}{26} + \frac{(29-26)^2}{26} + \frac{(31-26)^2}{26} + \\ &\frac{(21-26)^2}{26} + \frac{(27-26)^2}{26} + \frac{(26-26)^2}{26} = 0.615 + 0.346 + 0.962 + 0.962 + 0 \\ &= 2.923 \text{ olarak bulunur.}\end{aligned}$$

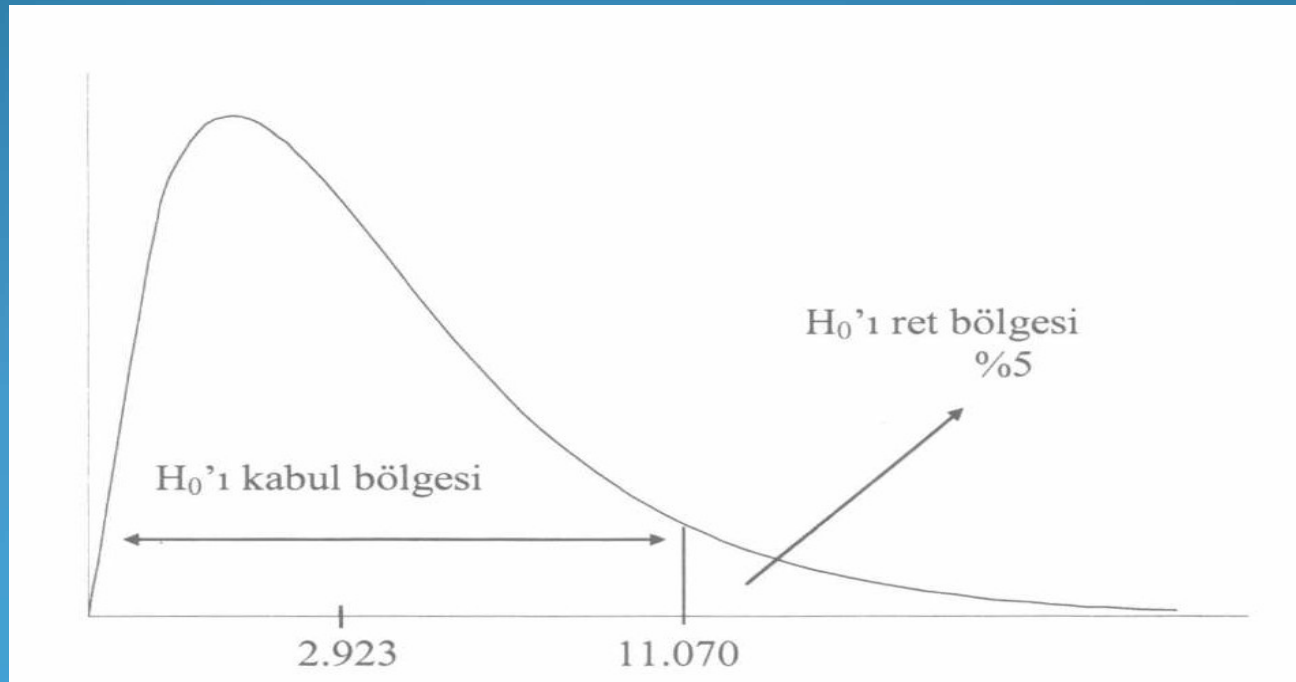
Bu örnekte serbestlik derecesi, $SD=6-1$ (gün sayısı-1)=5'tir.

5 serbestlik dereceli χ^2 -dağılımında %5'lik alanın başladığı χ^2 -değeri 11.070'tir

Hesaplanan ki-kare deęerinin 5 serbestlik dereceli ki-kare daęılımına dahil olma ihtimali %5'den büyüktür.

Yani hesaplanan X^2 -deęeri kabul bölgesindedir.

Böylece kontrol hipotezi reddedilemez, yani eczaneye bir hafta boyunca gelen müşterilerin günlere daęılımı homojendir ve günler arasında eczaneye gelen hasta sayısı bakımından fark önemli deęildir.



ÖRNEK 2: Bir kliniğe bir yıl içinde gelen 240 hastanın aylara göre dağılımı aşağıdaki gibi bulunmuştur.

Kliniğe gelen hastaların aylara göre dağılımı homojen midir?

H_0 : Hastaneye bir yıl boyunca gelen hastaların aylara göre dağılımı homojendir.

Aylar arasında hastaneye gelen hasta sayısı bakımından fark önemli değildir.

H_1 : Hastaneye bir yıl boyunca gelen hastaların aylara göre dağılımı homojen değildir.

Aylar arasında hastaneye gelen hasta sayısı bakımından fark önemlidir.

Aylar	Gözlenen hasta sayısı (f)	Beklenen hasta sayısı ($\hat{f}$)	$(f-\hat{f})^2/\hat{f}$
1	32	20	7.2
2	30	20	5.0
3	24	20	0.8
4	22	20	0.2
5	19	20	0.05
6	14	20	1.8
7	12	20	3.2
8	12	20	3.2
9	14	20	1.8
10	18	20	0.2
11	22	20	0.2
12	21	20	0.05
Toplam	240	240	$\chi^2=23.7$

Burada hangi hipotezi kabul edeceğine karar vermek için araştırmacının χ^2 -değerini hesaplaması gerekir.

Bunun için de her ay beklenen hasta sayısının bulunması lazımdır.

Kontrol hipotezi ile araştırıcı hasta sayısı bakımından ayların homojen olduğunu öne sürdüğüne göre, her ay için beklenen frekans $240/12=20$ 'dir.

$$\chi^2 = \frac{(32 - 20)^2}{20} + \frac{(30 - 20)^2}{20} + \dots + \frac{(21 - 20)^2}{20} = 7.2 + 5.0 + \dots + 0.05 \\ = 23.7$$

Serbestlik dereceli χ^2 -dağılımında %5'lik alanın başladığı χ^2 -değeri 19.675'tir.

Hesaplanan ki-kare değerinin 11 serbestlik dereceli ki-kare dağılımına dahil olma ihtimali %5'den küçüktür.

Bu durumda kontrol hipotezi reddedilir.

