MEÜ.
Mühendislik Fakültesi
Jeoloji Mühendisliği Bölümü

MÜHENDİSLER İÇİN İSTATİSTİK

YÖNTEMLER VE UYGULAMALAR

Prof. Dr. Hüseyin Çelebi

Ders Notları

Mersin 2007

Birkaç ünlü sözü

İstatistik!

Matematiğin yardımı olmadan doğa bilimlerini incelemek demek, gerçekleştirilmiyecek işe girişmek demektir.

Galileo Galilei

Her bilimin matematiğe gereksinimi vardır, ancak matematiğin hiçbirine.

Jakob Bernoulli

Matematik bilimleri, her şeyden önce, berraklığı nedeniyle hoşuma gider.

René Descartes

Hiçbir şey iyi bir teori kadar pratik olamaz.

Hermann von Helmholtz

Önsöz

"Jeolojide matematiksel ve istatistik yöntemler" ders notları kapsamında 2005/2006 öğretim yılında jeoloji, çevre, ve bilgisayar öğrencileri için hazırlanmıştır. Sınırlı yarıyıl süresi içinde mümkün olduğu kadar uygulamaya yer vermek için derin teorik işlemlere yer verilmemiştir. Bu konuya ilgi duyulduğunda kaynakçadaki özel kaynaklara başvurulabilir.

Bu ders notları yaklaşık 25 yıllık deneyimden, jeoistatistik ve istatistikte mevcut çok sayıdaki eserin incelenmesinden sonra hazırlanmıştır. Bunlardan David (1978), Akın ve Siemes (1988), Wellmer (1989), Schönwiese (1992), Anadolu Üniversitesi (2001), Arıcı (2001) ile Tüysüz ve Yaylalı (2005) en önemlileridir. Örnekler, çoğunlukla Türkiye'deki özgün çalışmalardan ve yerbilimlerinden başka dallara da örnek oluşturacak şekilde seçilmiştir.

Bu notların yazılmasını sağlayan asıl neden öğrencilerin derslerde gösterdikleri yakın ilgi ve eleştirileri olmuştur. Kendilerine, isim anmadan, çok teşekkür ederim.

Mersin, Kasım 2006 H. Çelebi

1 GİRİŞ

Sayfa

2 TEK DEĞİŞKENLİ DAĞILIMLARIN TANIMLANMASI

3 TEORİK DAĞILIMLAR

1 GİRİŞ

1.1 Genel bakış

İstatistik, belirsizlik durumundan en iyi sonucu çıkarmaya veya kararı vermeye yarayan yöntemlerin özetidir. Amaç, yanıltıcı yorumlardan kaçınmak ve ileriyi doğru kestirmektir. İstatistik, gözlemler sonucu elde edilen sayısal verileri inceler ve bunlar arasındaki bağıntıları ortaya çıkararak sonuçların grafik veya çizelgeler halinde sunulmasını sağlayan bir inceleme yöntemidir. Özet olarak **istatistik**, rastgele/tesadüfi ve tesadüf şeklindeki olayları inceleyen **bir metodik bilimdir** ve birçok bilim dalında uygulanabilmektedir. Elde edilen sonuçlardan çeşitli yorumların yapılması ile sorunlara çözüm aranır. Sonuca varmak için bazen kısıtlı bilgi ile yetinmek gerekebilir.

İstatistik, olaylarla başlar. Her olayın bir *etkisi* veya girişi, bunun da bir mekanizması, yani bir bağıntısı veya bir *fonksiyonu* bulunur. Bunların sonucu doğal olarak bir iş veya bir hizmettir. Örneğin, bir otomattan bir içecek almak şöyle gösterilebilir:

Para → **Para otomatı** → **Seçme düğmesi**

1.Giriş ***2. Etki mekanizması*** ***3. Sonuç (→ olay)***

Burada para ve seçme düğmeleri giriş büyüklükleri, verilen içecek de etki büyüklüğüdür. Etki büyüklüğü **x** ve **y** (stokastik), etki mekanizması **f(x,y)** kombinasyonu ve **w** sonucu tanımlı birkaç etkidir. Burada **olay**, **tanımlanmış** veya tanımlanmamış olabilir. Teoriler ve modeller kestirilen olayları tanımlamaya yararlar. **İstatistiksel** (kestirilemeyen) yöntemler de bunların kapsamını oluşturur.

Örneğin, bir uydunun hareketi uzayda tanımlanmış bir olaya veya gelişmeye dayanıyor. Bunun etkenleri yerçekimi, yer-uydu ağırlıkları, mesafe ve hızlarıdır. Mekanizma da çekim ve merkezkaç kuvveti yasasıdır. Bu mekanizmanın büyüklükleri ise, hesaplanabilen bir yörünge eğrisidir. Ay tutulması da bunun gibi tanımlanan bir olaydır. Ancak bulutların hareketi ve yağışlar **kestirilemeyen** veya hesaplanamayan, istatistiksel olaylardır. Zar ve yazı-tura atma da kestirilemeyen olaylardır. 1 atış temel bir tahmin, yani tanımlanmamış 1/6 ve 1/2 olasılık demektir. 2 zarın ard arda veya birden atılması 1/6.1/6 = 1/36 olasılığı ifade eder.

İstatistiğin çalışma yöntemleri,

a) Tanımlamak veya örneklemek (tanıtımsal),
b) Dağılım şekillerini incelemek (ana kütle veya örneklem özelliklerini incelemek),
c) Tahmin etmek ve kestirmek (olasılıkları araştırmak),
d) Teste tabi tutmak (hipotezler, karar kuramlarını uygulamak),
e) Analiz etmek (bağıntıları ortaya çıkarmak) ve
f) Özel yöntemlerin

uygulanmasıdır.

İstatistikte veri,

1. Saymak,
2. Ölçmek,
3. Gözlemek,
4. Anket yapmak,
5. Haritalamak ve
6. Tahmin etmek

yöntemleri ile sağlanır. Bunlardan gözlemler, istatistiğin temelini oluşturur. Planlama ve karşılaştırma istatistiğin en yaygın kullanıldığı alanlardır.

1.2 Tarihsel gelişim

Eski çağlardan beri insanlar geleceği kestirmek isterler. Gelecekte ne olacağını şimdiden bilmek, başarı ve üstünlük sağlamanın önkoşulu bilinir. Ancak geleceği bilmek mümkün değildir. Çünkü gelecek bilindiği zaman, gelecek şimdi olur ve geleceğin kendisi ortadan kalkar. Bu da doğa yasalarına, öncelikle zaman kavramına, ters düşer.

Bu engeli aşmak için insanlar büyü ve fal gibi dayanaksız yöntemlere yönelerek geleceği kestirme yollarını aramışlardır. Bu uygulamalar, güvensizliklerinden dolayı, zamanla inandırıcılıklarını kaybetmiştir. Bunların yerini gözlem ve ölçümlere dayanan basit istatistiksel hesaplamalar almıştır. Örneğin İ.Ö. Mısır'da ve Çin'de planlama, nüfus sayımı, asker ve vergi toplama işlemlerinde temel istatistiksel işlemlerden yararlar sağlamıştır.

Günümüz istatistiğinin kökleri ancak 15. yy'a kadar uzanmaktadır. Metafiziğe karşı pozitif düşüncenin üstünlük sağlaması modern istatistiğin gelişmesine de ivme kazandırmıştır. Bunun da esas kaynağı sohbet matematiği, şans oyunları (kumar), yani insanın yine ileriyi kestirme veya önceden bilme merakı, olmuştur. Bugün de istatistik bu alanların temel dayanağı olmaya devam etmektedir (loto ve toto gibi).

Tarihte istatistiği bilimsel olarak ilk irdeleyen ve kuramlara bağlamaya çalışan matematikçi İtalyan **Pacioli** (1445-1514) ve **Cardano**'dur (1501-1576). Bunlar zar atma üzerine çalışmışlardır. Ancak bugünkü istatistiğin kuramlarının temelleri **Pascal** (1623-1662) ve **Bernoulli** (1654-1705, Bernouli dağılımı) tarafından atılmıştır. Geliştirdikleri yöntemler, olasılık, şans ve risk oranlarının hesaplanmasını kolaylaştırmıştır. Bernoulli'yi olasılıkları hipotez modellerine dayandıran **Bayes** (1702-1761, Bayes teoremi) ve **Laplace** (1749-1827, Laplace teoremi), **Poisson** (1781-1840, Poisson dağılımı) ve Alman **Gauss** (1774-1855, çan eğrisi) tarafından daha ileriye götürülmüştür. 20. yy istatistikçileri arasında **Galton** (1822-1911, logaritmik dağılım), **Pearson** (1857-1936, Pearson bağıntısı) ve **Fisher** (1890-1962, varyans analizi) önemli yer tutmaktadırlar. İstatistiğin geleceği ile ilgili olarak **Tukey**, **Kendall**, **Watts** ve **Bradley** açıklamalarda bulun-maktadırlar.

İstatistik sürekli geliştirilmekte ve yaygın kullanım alanı bulmaktadır. Örneğin, istatistiğin yerbilimlerdeki adı **jeoistatistiktir**. 20. yy'ın ikinci yarısından itibaren bu alanda kullanılmaya başlamıştır. Jeoloji, madencilik, zemin etütleri, coğrafya, çevre, tarım, ormancılık ve hidroloji bu alanların sadece birkaçıdır. Jeoistatistik bugün güvenilen ve kendine özgü **bölgesel** veya **yere bağlı değişkenler** gibi teorik esasa ve **varyogram** gibi araçlara sahip bulunmaktadır.

1.3 Temel kavramlar ve tanımlar

Bir araştırmada incelenecek bireyler veya malzemenin tümünün incelenmesi gerekmez ve mümkün olmaz. Bu hem ekonomik değil, hem de yeterli örnekle elde edilen sonucu değiştirmez (bkn. büyük sayı teorisi). Bu nedenle incelenecek ana kütlenin ancak bir kısmı temsilen incelenir. Ana kütleye **popülasyon veya örnek uzayı** (sample space), temsilen incelenecek kısmına da **örneklem** denir. Bir popülasyonda verilerin tüm özellikleri ortaktır. Bulundukları yeri, ana maddeyi veya kaynağı tüm özellikleri ile temsil eder. Dolayısı ile bir örnek uzayının ancak bir temsili parçası veya örneklemi olabilir. Aynı şekilde bir örneklemin de sadece 1 **ortalama değer**i bulunur. Bu değer, tüm gözlemleri aynı oranda temsil eder. Aşağıdaki şekil örnek uzayı, örneklem ve ortalama değeri açıklamaktadır:

Örnek uzayı	Örneklem x_i	Ortalama değer $\bar{x}$
o		
ooo		
ooooo	o	
oooooo	ooo	o

Analiz aygıtlarının son 40 - 50 yılda güvenilebilir geniş kapsamlı veri üretebilmesi ve bilgisayarın uzun matematiksel işlemleri kolaylaştırması ile, örneğin yerbilimlerinde istatistik mühendislik, sosyal ve doğa bilimlerinde giderek önem kazanmakta ve kesinlik derecesi aratmaktadır. Son yıllarda istatistik, madenciliğin vazgeçilmez unsuru haline gelmiştir. Jeolojik veya bölgesel değişimler, matematiksel olarak tanımlanmakta, örnek ve sondajların en uygun aralıkları, rezerv hesaplarının kesinlik dereceleri ile tenör dağılımları, İzotropi, Anizotropi ve tabakalanma gibi özellikler ayrıntılı incelenebilmektedir.

İstatistiksel değerlendirmenin ilk adımı, seçilmiş az sayıdaki örneklerin dağılımı ve bunların bir örneklem (bir bütünlük) oluşturup oluşturmadığı özelliğinin araştırılmasıdır. Bu özelikler ilk aşamada sıklık dağılımı ve test yöntemleri ile ortaya çıkarılabilir. Buradan örneklemin dağılım fonksiyonu, ortalama değeri, tepe, ortanca, değişke, standart sapma, eğilim, tepelik gibi değişkenleri elde edilebilir. Bu değişkenler güvenirlik sınırları ile tanımlanırlar. Örneğin, ortalama değerin doğruluğu student-t sınaması yöntemiyle denetlenebilmektedir.

Mühendislikte kullanılan bu yöntemler yanında, sıkça bağıntı (korelasyon), bağınım (regresyon), kümelenme (cluster), etken (faktör) ve diskriminant analiz yöntemleri de kullanılmaktadır. Tüm yöntemlerin temeli analiz, gözlem ve ölçüm gibi değişkenlere dayanmaktadır. Bir örneklemin güvenirliği içerdiği gözlem veya veri sayısına bağlıdır. Ancak çok farklı yöntemleri bulunan örnek alma başlı başına bir konudur (sistematik ve rastlantısal örnek alma gibi). En az veya en uygun örnek sayısı ve miktarı hak-kında objektif ilke ve yöntemler bulunmamaktadır.

En az gözlem sayısı amaca göre değişir (normalde n > 20). Bu nedenle ilgili özel kaynaklara başvurmakta yarar vardır. Analizler sonucu elde edilen verilerin amaca uygun ve doğru olarak değerlendirilmeleri şarttır. Bunun için,

a) Veri çeşidinin seçilen değerlendirme yöntemine uygun olması,
b) Bağıntıların ortaya çıkarılabilmesi için yeterli ve uygun örneğin alınması,
c) Eksik veya gereksiz verinin toplanmaması ve
d) İstatistiksel homojenliğin korunması

esastır. Çoğu kez aynı maddeden birçok **özellik** araştırılır. Bunları mümkün olduğu kadar azaltmak gerekir. Bunu yaparken özellikler arası bağıntıların bozulmamasına veya ona göre inceleme yöntemi seçilmesine dikkat edilmelidir. Bu amaçla oranlar, tanımlayıcı katsayılar v.s. alınmalıdır.

Veriler matematiksel işlemlere tabi tutulabilmeleri için listelerde derlenir. Bunların sınıflara veya gruplara ayrılmaları, çalışmaları ve gözlemleri kolaylaştırır. Ortalama değer gibi bazı değişkenlerin hesaplanması ile ön bilgiler edinilebilir, önemli etkenler saptanabilir ve inceleme yöntemleri seçilebilir. Bu ön bilgilere şema, grafik ve diyagramlarla yeni veriler eklenebilir.

Bu çalışmanın amacı istatistikteki güncel durumu ve yöntemleri tanıtmak, ileride yapılacak çalışmalar için temel kaynak yaratmak ve özgün verilerle uygulama örneklerini yaymaktır. Bir istatistiksel işlemde çok sayıda değişken ve değişken işleme katılır. Ancak bunların doğru tanımı ve hesaplanması sonunda doğru istatistiksel çözüm bulunur ve yorumlanabilir. Burada bu amaçla kullanılacak temel kavram ve bunlara bağlı değişkenler tanıtılacaktır.

Çizelge 1.1. Temel kavramların notlarda ifade şekilleri.

Kavram	İşaret	Örnek 1	Örnek 2
Değişken	x, y	Zar	Isı
Özellik	X_j, Y_j (j=1,...,n)	Göz (1,...,6)	1° C
Olay	O	1 noktanın gelmesi	Her ölçümde bir ısı derecesi
Veri	x_i, y_i (i=1,...,n)	Örneğin, 1,4,2,..,6	Örneğin 15° C, 18° C v. s.

x_i, x değişkeninin verisidir.

1.3.1 Sayı, büyüklük ve ölçü birimi (skala)

İstatistikte kullanılan verilerden çoğu kez incelenen bütünlüğün (kümenin) birden fazla özelliğinin incelenmesi beklenir. Bu nedenle incelenen özellikler maddelerin belli karakter farklarıdır. Bu amaçla önce incelenen maddeden ölçüm, gözlem, tahmin ve hesaplamalarla elde edilen belli sayısal veriler elde edilir. Bu birbirinden farklı verilerin toplamı, incelenen bütünlüğün özelliklerini kapsar.

Veriler, kullanım amaçlarına göre ana hatları ile süreksiz ve sürekli veri gruplarına ayrılır. Bunlar da kendi aralarında aşağıdaki alt kısımlara ayrılır (Schönwiese, 1992, Wessel, 2000 ve Arıcı, 2001):

1. Süreksiz veriler (sayılar), örneğin bilgisayarda işlenen veriler ve bir ailedeki çocuk sayısı ve yağış süreksiz verilerdir. Bunlar aşağıdaki kısımlara ayrılır:

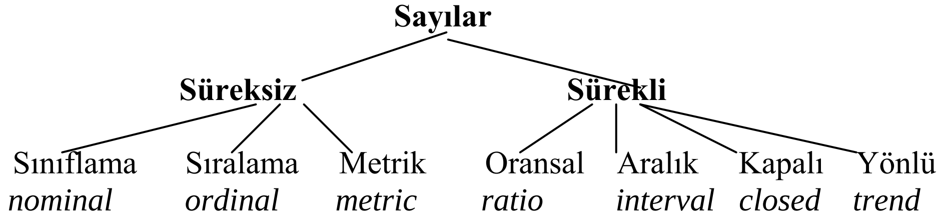
a) Sınıflama verileri (nominal ölçü birimi veya skalası), derecelendirme çeşitleri içermezler. Sadece kodlamaya ve maddeleri nicelik bakımından kısımlara ayırmaya yararlar. Nitelik bakımdan maddeleri birbirinden ayırmak veya sınıflandırmak mümkün değil. Ancak

$$mA_1 = mA_2$$

veya

$$mA_1 \neq mA_2$$

gibi sonuçlar çıkarılabilir. Örnek olarak il plaka ve cinsiyet numaraları (erkek=1, kadın=0 gibi) verilebilir.



b) Sıralama verileri (ordinal ölçü birimi), özelliklerin büyüklüğü bakımından kümeler arasında bir farklılık gözlemek mümkündür. Ancak veriler farkların büyüklüğü hakkında bilgi vermezler. Yukarıdaki

$$mA_1 = mA_2$$

veya $$mA_1 \neq mA_2$$

özellik ifadesine ek olarak burada,

$$mA_1 < mA_2$$

veya $$mA_1 > mA_2$$

gibi sonuçlar da çıkarılabilir. **Mercalli-Sieberg deprem ölçeği** ve **mohs mineral sertliği** buna örnek verilebilir. Mohs örneğinde mineraller sertliklerine göre sadece sıralanmıştır (talk-elmas gibi). Farklar, eşit sertlikleri yansıtmazlar. Örneğin, sertliği 8 olan topazın, 4 sertliğine sahip flüoritten 2 kat daha sert olduğu anlamına gelmez.

c) **Metrik veriler** (metrik ölçü birimi), eşit aralıklara ayrılan sabit bir ölçü birimi oluşturan ve doğal bir sıfırı bulunan bir ölçü birimine dayanmaktadır. Büyük öneme sahiptir. Yukarıdaki iki veri çeşidi mühendislik ve yerbilimlerinde önemli rol oynamazlar. Ancak metrik veriler hesaplamaların temelini oluşturur. Bu verilerle her türlü matematiksel işlemi yapılabilir ve özellikler arasındaki farklar saptanabilir. Örneğin, bir belirtinin diğerinin kaç katı olduğu belli bir kesinlikle bulunabilir: $mA_1 = a*mA_2$ gibi veya **1a** ve **b**'deki özelliklere ek larak,

$$mA_1 = mA_2 + b$$

gibi sonuçlar da çıkarılabilir

2. Sürekli sayılar (veriler), belli sınırlarda kesintiye uğramadan kullanıma açılan verilerdir. Sürekli sayılara ısı örnek verilebilir. Şöyle sınıflandırılırlar:

a) **Oransal veriler** (ölçü birimi), bir verinin diğer bir verinin kaç katı olduğunu gösteren verilerdir. Örneğin, K derecesi ile verilen ısı dereceleri, mutlak sıfır ile başladıkları için, oransal verilerdir. Ancak °C ile verilen ısı dereceleri aralık verileridir.

b) **Aralık veya fark verileri**, doğal sıfırın bulunmadığı ölçü biriminden elde edilen verilerdir. Bunlara örnek olarak ısı dereceleri °C ve K (-273,15° C) verilebilir. °C ile yapılan ölçümler aralık, K ile yapılan ölçümler ise, oransal verilerdir. Çünkü °C'deki 0 doğal değildir. Dolayısı ile 100° C, 10° C'ten 10 kat daha sıcak değildir. Ancak bu K için geçerlidir.

c) Kapalı verilere yüzdelikler gibi bir sabit toplam veren veriler örnek verilebilir.

d) Yönlü veriler, vektör ve fay gibi eğilim gösteren veriler bu sayılardandır. Yerçekimi ölçümleri de, yerkürenin merkezine yönelimleri nedeniyle, yönlü sayılardır.

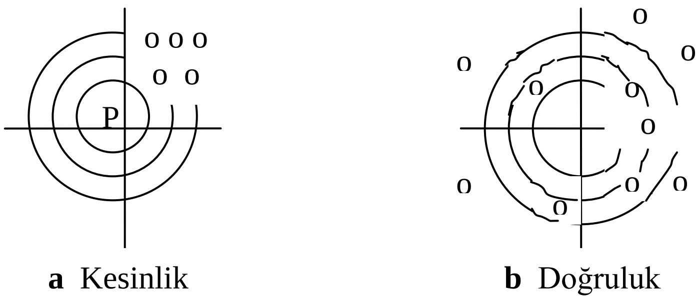
Bu temel sayıların tabii tutuldukları önemli işlemler Çizelge 1.2'deki hesaplamalarla açıklanmaktadır. Değişkenler nitel (özellik/kalitatif) veya nicel (sayısal/nicel) değişken olarak da sınıflandırılabilmektedir.

Çizelge 1.2. Sayıların birbirine çevrilmesine ilişkin örnekler.

z_i	d_i	v_i	n_i	p_i	
15,2 °C	-0,14 K	0,99	0,14	14,1	
12,4	-2,99	0,81	0,12	11,6	
13,5	-1,84	0,88	0,13	12,6	Toplam, Σz_i =107,4
17,2	1,86	1,12	0,16	16,0	
16,2	0,86	1,06	0,15	15,1	Ortalama değer c,
15,8	0,46	1,03	0,15	14,7	c= 107,4/7=15,34
17,1	1,30	1,11	0,15	15,9	

Açıklamalar: Fark sayıları : $d_i = z_i$-c, Örnek: 15,2-15,34 = - 0,14
Oransal sayılar: $v_i = z_i/c$, 15,2/15,34 = 0,99
Norm sayıları: $n_i = z_i/\Sigma z_i$, 15,2/107,4 = 14,15
Yüzdeler: $p_i = (z_i/c).100$, $(z_i/c).100:\Sigma(z_i/c)$ = 0,99.100/7 = 4,1 ($\Sigma(z_i/c)$ = 7,00 = 1)
İç içe/katlı sayılar: t_i = 1 yıl =12 ay =365 gün = 8760 saat gibi sayılar.

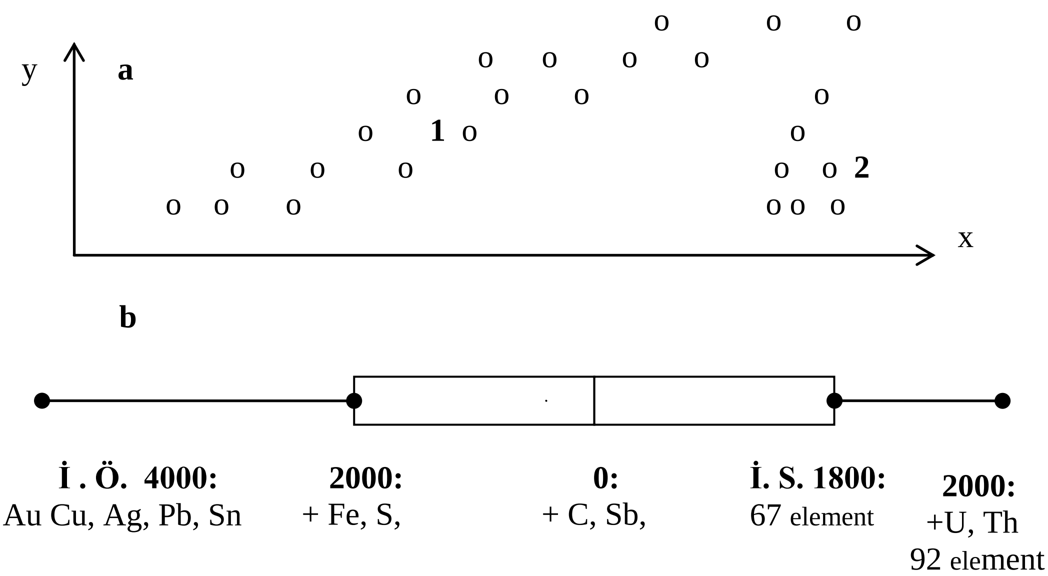
Verilerde **kesinlik derecesi** çok önemlidir ve her aşamada aranır. Yinelenen bir deneyde ölçümlerin birbirine yakınlığı kesinlik derecesini verir. Dolayısı ile kesin veriler, ölçümlerin yinelenmesinde çok az sapma gösterirler (Şekil 1.1 a). Bir kesinlik derecesi, T = 21 ± 0,5 ºC (aynı birimle oda sıcaklığı) şeklinde gösterilir. Aynı şekilde verilerin bir de gerçek değere en yakın olan **doğruluk derecesi** vardır. Bir veri kesin olabilir, ancak doğruluğu şüpheli olabilmektedir. Örneğin, saat veya yanlış ayarlanmış bir ölçü aygıtı gibi.



Şekil 1.1: Kesinlik **a** ve doğruluk **b** kavramlarının anlamı.

1.3.2 Sayıların grafiklerle sunum tarzları

Veriler değişik şekilde sunulurlar. Çizelgeler, şemalar ve grafikler halinde gösterilen verilerin sunulduğu ilk şekil çizelgelerdir. Ancak bunlar göze hitap etmezler ve incelenmeleri zordur. Bu nedenle çizelgeler öncelikle verilerin derlenmesine ve işlenmesine yararlar (bak. Çizelge 1.2.). Verilerin sade sunum şekli saçınım ve şemalarla sağlanır (Şekil 1.2.). Verilerin en iyi görüntülendiği sunum şekli grafiklerdir. Çok çeşitli olan grafikler, gözlem ve akış bakımından en uygun sunum araçlarıdır (Şekil 1.3.).



Şekil 1.2. Verilerin saçınım (scatter) **a** ve şematik çizgi-kutu diyagramda sunumu **b.** Tarihte elementlerin keşfi. **a**’daki dağılım (1 ve 2) iki ana kütleye işaret etmektedir.

Grafiklerle sunumda her değişken değeri, karşılaştırmaları ve iki veya daha çok değişken arasındaki ilişkiyi göstermek için yapılan sunumları ayırdetmek gerekir. Bunlar da ölçütlere bağlıdır. Ancak hiçbir zaman en iyi çözüm için hazır bir sunum tarzı yoktur.

Grafikler ancak istatistiksel verilerin değerini arttırdığında kullanılmalıdır. Çok grafik bir metnin okunmasını zorlaştırabilir. Yukarıda sunulan önemli grafikle sunum şekilleri yanında ilginç görülen **sap ve yaprak** (stem-and-leaf) grafik yönteminin burada belirtilmesinde yarar vardır (Çizelge 1.3). Bu yöntemle belirlenen aralıklara düşen ölçümler, büyüklük sırasına göre sıralandıklarında, bir histogram görüntüsü verir ve hem ham veri, hem de dağılım sınıflarını gösterdiğinden, daha bilgilendiricidir. Bu özellik histogramda kaybolur.

Çizelge 1.3. Steme-and-leaf grafik yöntemine bir örnek.

30	2
40	0 5 7
50	3 5 8 9
60	1 3 6 8 9 9
70	0 4 5 7 9 9
80	0 0 2
90	0 5

Veriler: {32-40-45-47-53-55-58-59-61-63-66-68-69-69-70-74-75-77-79-79-80-80-82-90-95}

1.3.3 Zaman dizileri (time serie)

Özellikle yerbilimlerinde çok sayıdaki veri zamanın bir fonksiyonu olduğu görülür. Örneğin, Çizelge 1.2'deki açıklamanın son satırındaki veriler **zaman dizisi**dir. Bu sayı çeşitleri zamanın bir fonksiyonudur. Tamamlanmamış bir olayın koşulları uzay ve zamana bağlı olarak değiştiği için x_i verileri artık sabit x*, y* ve z* koordinatlarına bağlı olmazlar. Dolayısı ile t zamanı da sabit olamaz ve x_i (x, y, z, t) şeklinde ifade edilirler. Coğrafyadan enlem, boylam ve ϕ açısı buna örnek verilebilir. Örneğin uzaydaki hareket x_i (t_i), (i=1,..., n) ve $t_{i+1}-t_i = \Delta_t = c =$ sabit, t'ye bağlı değişim, gibi. Saat başı ısı ölçümleri ve şehrin yıllık nüfus değişimi en açık zaman dizileridir. Bunların ortalamaları da Δt_i'dir. Bir caddedeki trafiğin günün saatlerine göre değişimi de bir zaman dizisidir (x(t) fonksiyonu).

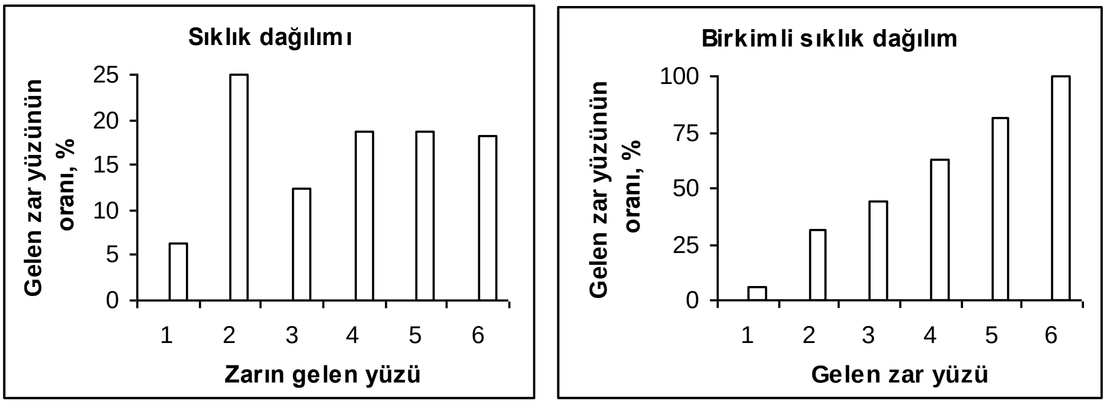
1.3.4 Sıklık dağılımı

Sıklık dağılımı bir örneklemin dağılım şeklidir ve sadece bir veri grubunun veya örneklemin özelliklerini inceleyen yöntemdir. Varılan sonuçlar teorik esaslara göre yorumlanarak çözümler aranır. Sıklık dağılımının başlıca araştırma sonuçları normal, Logaritmik ve binomial dağılım ile bunların ortalama değer, standart sapma v.s. değişkenleridir.

Aynı koşullar altında n kez yinelenen bir deneyde meydana gelen A olay sayısına A'nın H_A **mutlak sıklık dağılımı**, bunun $\frac{H_A}{n}$ oranına da A'nın h_A **göreceli sıklık dağılımı** denir. Görüldüğü gibi sıklık dağılımı klasik olasılık tanımına benzemektedir: Çünkü h_A artan n ile P_A olasılığına yaklaşmaktadır. Büyük sapmaların meydana gelme olasılığı giderek azalır (büyük sayı teorisi). Bir zar atmada atılan göz sayısı rastlantısal x_1 büyüklüğünü ifade ederken para atmada rastlantısal sayı büyüklüğü olası yazı (y) veya tura (t) x_2'dir. Örneğin,16 kez zar atmada elde edilen sayılar {2, 6, 3, 2, 5, 6, 4, 4, 2, 5, 4, 5, 3, 1, 6 ve 2} ise, bunların dağılımları gelen zar yüzüne karşılık geldiğinden, her zar yüzü belli sıklıkta yinelenmiştir. Örneğin, Zarın 1. yüzü 1 defa gelmesine karşın 2, 4 ve 5 de 3 kez atılmıştır. Buna **sıklık dağılımı** denir. Bu dağılımı tanımlıyan kuramsal fonksiyonuna da **olasılık** veya **dağılım fonksiyonu** denir. Uygulamada her zar yüzünün gelen rastlantısal değeri bir dikdörtgen ile birbirini örtmeyecek şekilde apsis üzerinde gösterilir. Çizelge 1.4 ile Şekil 1.3 buradaki zar atmanın sıklık dağılımını göstermektedir. Sınıf sayısının bulunması için birçok yöntem bulunmaktadır (bak. Örnek 2.6). Burada her zar yüzü doğrudan bir sınıfı temsil etmektedir. Bu nedenle zar atma iyi bir örnektir. Sıklık dağılımları aynı zamanda bir sunum veya gösterim şeklidir .

Çizelge 1.4. 16 kez atılan bir zarın gelen yüzlerinin sıklık dağılımı (Schönwiese, 1992).

Zar yüzü x_i	Çizgi	Listeler Sayısal H_A	Göreceli h_A	Göreceli birikimli	Sıklıklar Birikimli	Birikimli [%]
1	/	1	0,0625	0,0625	1	6,25
2	////	4	0,2500	0,3125	5	31,25
3	//	2	0,1250	0,4375	7	43,75
4	///	3	0,1875	0,6250	10	62,50
5	///	3	0,1875	0,8125	13	81,25
6	///	3	0,1875	**1,0000**	**16**	**100,00**
Σ	**16**	**16**	**1,0000**			



Şekil 1.3: Atılan bir zarın gelen yüzlerinin sıklık dağılımı (sütun diyagram, histogram).

Yukarıdaki çizelge, çizgi, kutu ve sütun diyagramlar şeklindeki sunum tarzları yanında değişik sunum şekilleri bulunmaktadır. Bunların hepsini burada vermek mümkün değil. Ancak burada bazı örneklerle yetinilecektir (bak. Tüysüz ve Yaylalı, 2005).

Bir rastlantısal x büyüklüğünün sürekliliği ancak dağılım fonksiyonunun sürekli olması, yani aralıksız olması ile mümkündür. Çok sayıdaki sürekli F dağılım fonksiyonu yanında sıklık dağılım fonksiyonları da sürekli fonksiyonlardır ve

$$F(x)=\sum_{a<x} P(x) = \int_{-\infty}^{x} f(t)dt \qquad 1.1$$

özelliğine sahiptir. $\lim_{x\to\infty}$ F(x)=1 olduğundan, her zaman,

$$\int_{-\infty}^{\infty} f(t)dt = 1 \qquad 1.2$$

değerine sahip olur. Dolayısıyla popülasyondaki herhangi bir analiz değerinin uç değerler arasındaki bir aralığa düşme olasılığı % 100'dür. Böyle bir analiz değerinin bulunduğu yerin koordinatları, olasılıkların işleyişini tanımlıyan Gauss çan fonksiyonu ile tanımlanabilir. Sıklık dağılımı, ortalama değer ve diğer istatistiksel değerlendirmeler için ön şarttır.

1.4 Sayı dizilimleri

İstatistik için çok önemli olmayan **permütasyon** (dizilim), **kombinasyon** (devşirim) ve **varyasyon**ların (çeşitleme) kullanım alanları çok yönlüdür. Sonlu miktarlarla hesaplamak ve uygun koşullarda olası düzen sayısını hem nominal, hem de metrik bulmak bu hesaplamalar sayesinde kolaylaşmaktadır. Bunların çok örneği sohbet veya oyun matematiğinden ve olasılık esaslarından gelmektedir. Sayı teorisi ile yakından ilgilidir (4 sayı düzenlenmesi gibi, bak. aşağıya). Harf-oyun kağıdı, çobanın tek başına kurt, kuzu ve lahanayı nehirden geçirmesi gibi hesaplamalar eskiden beri bilinmekte ve uygulanmaktadır. Buna toto oyunu da eklenebilir. Oyun kağıtlarında beklenen kağıdın gelmesi de buna benzer. Dolayısı ile beklenenin bulunmasının ne kadar zor olduğu (zayıf bir olasılık olduğu), oyunlarda ve hatta yaşamda, kazanmanın değil, kaybetmenin ve yanlış yapmanın kural olduğu ortaya çıkmaktadır. Eskiden çözümler özel ödevlere göre hazırlanırdı. Günümüzde sorunları temel örneklere indirgeyerek genel çözüm ve yöntemlerinin bulunmasına çalışılır.

1.4.1 Permütasyon (P, dizilim)

Permütasyon, sıranın çok önemli olduğu bir dizindir. Bir kümenin farklı n elemanı arasında kaç çeşit düzenlemenin mümkün olduğu araştırılır. Örneğin, 10 kişiden hiçbirinin yerini korumadan 10 sandalyeye oturma düzeni,

$$\begin{aligned} P(10) &= 10! \\ &= 1.2.3.4.5.6.7.8.9.10 \\ &= 3.628.800 \end{aligned}$$

dizilim, bu türden bir hesaplamadır.

3 harfle **yinelenmeden** yazılacak kelime sayısının bulunması da buna benziyor: 1. sıraya her harf gelebilir. 2. sıraya geri kalan 2 harften biri ancak gelebilir. 3. sıra için ise, seçme olanağı kalmamaktadır. Bu nedenle yazılacak kelime sayısı, 3.2.1=3! olur (=3 faktöriyel, 0! = 1).

Permütasyonun genel formülü,

$$P(n) = n! \qquad 1.3$$

şeklindedir. Hesaplanacak abc elemanlarının permütasyonu için,

$$P(3) = 1.2.3$$
$$= 6$$

geçerlidir. Yani,

$$\{\text{abc-acb-bac-bca-cab-cba}\}$$

şeklinde bir dizilim ortaya çıkar.

n elemana sahip, k kadar farklı ($k \leq n$) ve $n_1, n_2, ..., n_k$ sıklığında **yinelenen** elemanı bulunan kümeler için,

$$P(n_1, n_2, ..., n_k) = \frac{n!}{n_1!.n_2!...n_k!} \quad 1.4$$

formülü geçerlidir. Örneğin, 3, 3, 4, 4, 4, 5, 6 sayıları ile kaç tane 7 rakamlı sayı üretilebilir?

Yanıt:
$$P(7; 2, 3, 1, 1) = \frac{7!}{2!.3!.1!.1!}$$
$$= \frac{1.2.3.4.5.6.7}{(1.2).(1.2.3).1.1}$$
$$= 5040:12$$
$$= \underline{420}$$

bulunur. Ayrıntılı temel esaslar için standart kaynaklar salık verilir.

1.4.2 Varyasyon (V, çeşitleme)

Varyasyon, $k \leq n$ seçeneği ile yapılan bir özel dizindir. n elemandan n çeşitleme yerine k çeşitleme de ($k \leq n$) incelenebilir. Böyle bir çeşitlemeye k dereceden **yinelenmiyen** varyasyon denir. Genel formülü,

$$V(n, k) = \frac{n!}{(n-k)!} \quad 1.5$$

şeklindedir. İspatı permütasyon ilkelerindeki gibi, n eleman arasından seçim, n-1, n-2, ..., n-(k-1), ile sağlanır. Bu varyasyon çeşidine permütasyondaki 10 kişinin 10 sandalyeye oturması yerine 6 sandalyeye oturması örnek verilebilir:

$$V(10,6) = \frac{10!}{(10-6)!}$$

$$= \frac{10!}{4!}$$

$$= 151.200$$

olanak bulunur.

Yinelenen bir varyasyon için,

$$V_t\,(n, k) = n^k \qquad 1.6$$

geçerlidir. Yinelenen varyasyona sportoto örnek verilebilir: Oynanacak 12 maçtan beraberlik = 0, kazanma = 1 ve kaybetme = 2 ile gösterildiğinde,

$$V(3,12) = 3^{12}$$

$$= 531.441$$

oyunun gerektiği görülür.

1.4.3 **Kombinasyon** (C, devşirim)

Kombinasyonda permütasyona karşın sıra hiç önemli değil. Burada olası tüm devşirimler yerine k elemanlı kısmi miktarlar önemlidir. Buna k dereceden **yinelenmiyen kombinasyon** denir. Kombinasyon temelde tüm varyasyonları kapsıyan varyasyonun eşdeğer sınıfları olarak görülür. Kombinasyon,

$$C(n, k) = \binom{n}{k} \qquad 1.7$$

$$= \frac{n!}{k!(n-k)!}$$

ile tanımlanır. $\binom{n}{k}$ ifadesine **binom katsayıları** denir ve **n'nin k faktöriyeli** diye okunur. k yerine verilen n elemanlı kısımlar önemlidir. Uygulamalarda,

Yinelenen ve sıra dikkate alınmayan kombinasyonlar için de,

$$C(n, k) = \binom{n+k-1}{k} \quad 1.8$$

formülü kullanılır.

Örnek 1.1

n = 3 (abc) elemanın k (=2'şer) kombinasyonu.

$\mathbf{c = 3}$	$\mathbf{c_S = 6}$	$\mathbf{c_t = 6}$	$\mathbf{c_{st} = 9}$
ab, ac, bc	ab, ac, bc	ab, ac, bc	ab, ac, bc,
	ba, ca ve cb	aa, bb ve cc	ba, ca, cb,
			aa, bb ve cc

a) $$C(n, k) = \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

$$= \binom{3}{2} = \frac{3!}{2!(3-2)!}$$

$$= \frac{1.2.3}{1.2.(1)}$$

$$= \frac{6}{2}$$

$$= \underline{3}$$

b) $$C_t(n, k) = \binom{n}{k} = \binom{n+k-1}{k}$$

$$= \binom{3}{2} = \binom{3+2-1}{2}$$

$$= \binom{1.2.3.4}{2}$$

$$= \frac{24}{2}$$

$$= \underline{12}$$

seçenek bulunur.

s, sıra; **t**, tekrar demektir.

Örnek 1.2

Sayısal lotoda 2 oyunda (2 kerede) 6 bulma olasılığı. Önkoşullar: Yinelenme yok (çekilen bir sayı bir daha çekilmez) ve sıra önemli değil. Bir oyun için,

$$C(n, k) = \binom{49}{6}$$

$$= \frac{49!}{6!(49-6)!}$$

$$= \frac{49!}{6!.43!}$$

$$= \underline{13.983.816}$$

Seçenek vardır. 2 oyunda ise, P(O) = 2/13.983.816 = $1{,}43.10^{-7}$ gerçekleşme olasılığı demek-tir. Haftada bir çekiliş halinde bu, 134.460 yılda 1 gerçekleşme olasılığına karşılık gelir. Aynı zamanda 6 doğru için sadece 1 seçenek varken, 5 doğru için $\binom{6}{5}.\binom{43}{1} = 258$; 4 doğru için 13.545 ve 3 doğru için de 246.820 seçenek bulunmaktadır.

1.4.4 Determinantlar

Yukarıda açıklanan sayı dizilimleri yanında **determinant**lara da değinmek gerekir. Determinantlar permütasyonların türevleridir (p = 1,..., n, i_1,..., i_n; n sıralı determinant). Vektör uzaylarının hesaplanmasında ve denklem köklerinin bulunmasında yararlanılır (bak mat. kaynakları). Burada bir örnekle yetinilecektir:

Bir determinant genel olarak,

$$\text{Det M} = \begin{vmatrix} a_{11} & a_{12} \ldots & a_{1n} \\ a_{21} & & \\ \ldots & & \\ a_{n1} & \ldots & a_{nn} \end{vmatrix} \qquad 1.9$$

gibi tanımlanır. Bu determinant n = 3 için,

$$\text{Det M} = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11}\begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{21}\begin{vmatrix} a_{12} & a_{13} \\ a_{32} & a_{33} \end{vmatrix} + a_{31}\begin{vmatrix} a_{12} & a_{13} \\ a_{22} & a_{23} \end{vmatrix} \qquad 1.10$$

şeklinde elemanları cinsinden yazılabilir ve

$$A = a_{11}(a_{22}a_{33}-a_{23}a_{32})-a_{21}(a_{12}a_{33}-a_{13}a_{32})+a_{31}(a_{12}a_{23}-a_{13}a_{22}) \qquad 1.11$$

eşitliğine göre de çözülür.

Böyle bir determinant 3 x 3 **matris**'li (= $A_{3,3}$), 3 (satır) x 3 (sütun; m, n) veya 3 (sütun) x 3 (satır; n, m) **düzen**'li ve a_{ij} **eleman**'lıdır (i = satır, j = sütun numarası). Sütunlar genelde x_1, x_2 veya belli yüksekliklerde alınan ölçümler gibi değişkenleri; satırlar da 1., 2..... denklemleri, yükseklikle değişen hava basıncı ile ısı gibi örnekleri içerir. Bunların sayısına göre determinantın şekli değişir (enine veya boyuna dikdörtgen gibi). Bir determinantta satır ve sütunların yerleri değiştirilebilir (a_{ij} = a_{ji}), dört işlem ve üs alma gibi cebirsel işlemlere tabi tutulabilirler.

Örnek 1.3

a) Yukarıdaki 3 basamaklı determinantın gösterdiği eşitliklerin katsayıları:

$$A = \begin{vmatrix} 1 & 6 & 4 \\ 2 & 1 & 0 \\ 5 & -3 & -4 \end{vmatrix}$$

değerlerine sahipse,

$$A = 1[1(-4) - 0(-3)] - 6[2(-4) - 0(5)] + 4[2(-3) - 1(5)]$$
$$= -4 + 48 - 44$$
$$= 0$$

elde edilir. Bu, determinantta vektör veya doğrusal bağıntının bulunmadığını gösterir. Çünkü 0, yön ve eğilim göstermez.

b) Determinanat yardımı ile 2 bilinmeyenli denklem çözümü

Verilenler:

$$5x + 7y = 19$$
$$3x - 2y = -1$$

denklemlerinin köklerinin bulunması.

$$\leftrightarrow \begin{vmatrix} 5 & 7 \\ 3 & -2 \end{vmatrix} \begin{vmatrix} x \\ y \end{vmatrix} = \begin{vmatrix} 19 \\ -1 \end{vmatrix} \rightarrow \det A = -31$$

bulunur. Bunun çözümü için 1/A determinantına gereksinim vardır. Cramer Kuralına göre A.x = b'dir (b, bir köktür). Buna göre,

$$\det A^{-1} = |A^{-1}| = \left(\frac{1}{-10-21}\right) \begin{vmatrix} -2 & -7 \\ -3 & 5 \end{vmatrix} = \begin{vmatrix} 2/31 & 7/31 \\ 3/31 & -5/31 \end{vmatrix}$$

$x = A^{-1}$.b'den,

$$x = \begin{vmatrix} 2/31 & 7/31 \\ 3/31 & -5/31 \end{vmatrix} \begin{vmatrix} 19 \\ -1 \end{vmatrix} = \begin{vmatrix} 38/31 & -7/31 \\ 57/31 & +5/31 \end{vmatrix} = \begin{vmatrix} 1 \\ 2 \end{vmatrix}$$

elde edilir. Denklemlerin kökleri: x = 1 ve y = 2'dir.

1.5 Olasılık

Olasılık, gözlenen olayların olası tüm olaylara oranı demektir (bak. 1.14). Yukarıda verilen zar atma örneği (1.3.4) para atma (yazı/tura) örneği ile değiştirilebilir (1.12). İki oyun arasındaki fark sadece olasılıkların azalmasıdır. Para atmada iki seçenek varken, zar atmada seçenekler, zarın 6 yüzünden dolayı, altıya çıkmaktadır. Bu nedenle zarda bir yüzün gelme olasılığı 1:6, para atmada ise 1:2'dir. Bu iki oyunla olasılık kavramı kısmen açıklanmaktadır. Burada olasılık tanımından amaç, olasılığı bir norma bağlamaktır.

Doğa yasalarına uyan olaylara **determinist** (belirlenimci), hiçbir kurala uymayan olaylara da **kaotik** olaylar denir. İstatistikte **rastlantısal** (stokastik) olaylar incelenir. Rastlantısal (biribirinden bağımsız) olaylar için her zaman,

$$\lim_{n\to\infty} P(O) = c = \text{sabit} \qquad 1.12$$

eşitliği geçerlidir. Yinelenen bir deneyde, örneğin para atmada, n atış sayısı arttıkça sonuç sabit bir c değerine yaklaşır*. Bununla ilgili teoriye **büyük sayılar teorisi** denir. Yazı y ve tura t ise, $n \to \infty$'a yaklaştığında,

$$H_{(y)} = H_{(t)} = \underline{0{,}5} \qquad 1.13$$

değerini alır. Bu, uygulamada,

$$\lim_{n\to\infty} P(O) = c$$

olarak sınırlanabilir demektir.

Olasılık,

$$P(O) = O/\Omega = \frac{\text{Gözlenen} \cdot \text{olaylarin.sayisi}}{\text{Olası} \cdot \text{olaylarin} \cdot \text{toplamı}} \qquad 1.14$$

olarak tanımlanır (O, olay; Ω, olay evrenidir). Buna göre bir olay, olay evreni Ω'nın bir parçasıdır veya kısmıdır. Tek elementli olaylara temel olay denir. Bunların, zar atışında olduğu gibi, aynı olasılığa sahip olması durumunda ancak sonlu Ω için bir anlam taşır. Burada, O/Ω oranı

---

* Burada gözlenen olayların, olası tüm olaylar içindeki oranı giderek küçülür. Belli bir yinelenmeden sonra gözlenen olayların önemi kalmıyor. Örneğin, yazı veya tura olasılığı 1. atışta, 1:2=0,50; 2. atışta, (1:2)(1:2)=0,25 ve 3. atışta, (1:2)(1:2)(1:2) = 0,125 v. s gibi küçülür (bak. ayrıca örnek 6.3).

olasılıktır ve olası tüm olaylar (örnek uzayı) içinden gözlenen tüm olayları ifade eder. Örneğin, 32 oyun kağıdının 10'ar dağıtıldığı bir oyunda (skat, 10-4=6) oyuncunun birine 4 asın da gelme olasılığı,

$$P(A_6)=\frac{\binom{28}{6}}{\binom{32}{10}}=\frac{376.740}{64.512.240}=0{,}00584=\%0{,}58$$

oranı gibi ender bir olaydır. Bu örnekte 376.740 sayısı **gözlenen**, 64.512.240 sayısı da **beklenen** veya olası tüm olaylardır (=örnek uzayı) .

Bir değşkenin [α, β] aralığıda bulunması olasılığı (Şekil 1.4a, beklenen sıklık dağılımı = alan),

$$P\{O|_{\alpha}^{\beta}\} = P(\alpha \le x \le \beta) \qquad 1.15$$

$$= \int_{\alpha}^{\beta} f(x)dx \qquad 1.16$$

$$= F(\beta) - F(\alpha) \qquad 1.17$$

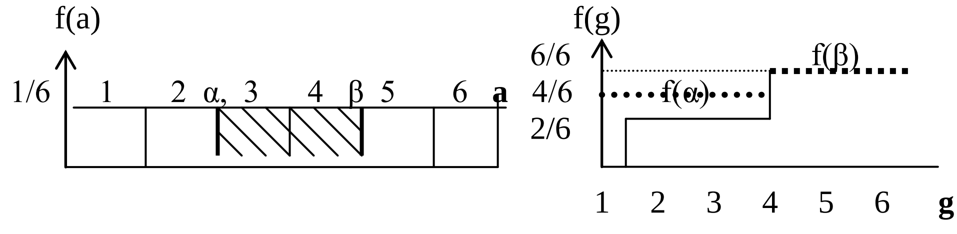
şeklinde ifade edilir. Buna **normal dağılım** denir. Normal dağılım**,** sonlu, eşit olasılıklara sahip özellikleri içeren ve istatistiksel dağılım koşullarını sağlıyan bir dağılımdır (bak. 3.2). Doğada en sık rastlanır. Bu dağılımın entegrali alındığında,

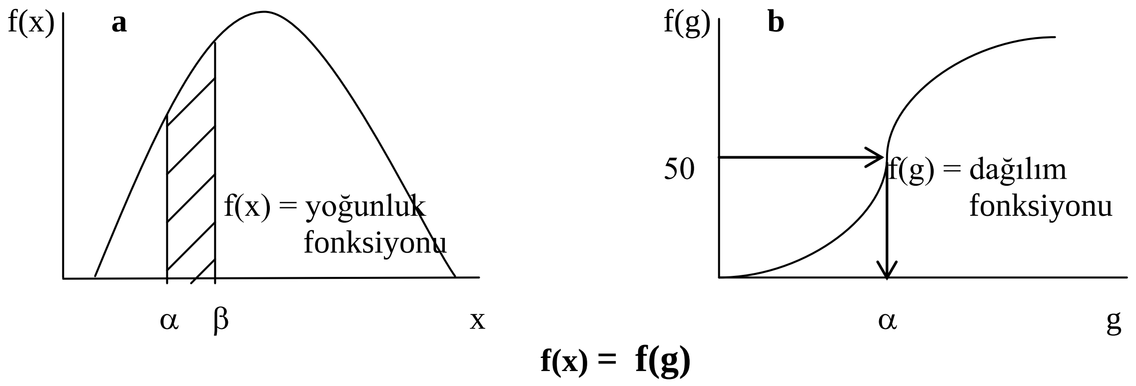
$$\int_{-\infty}^{+\infty} f(x)dx = 1 \qquad 1.18$$

olduğu görülür (eğri altında kalan alan). Bunun **birikimli** (kümülatif) dağılımı,

$$f(g)_{max} = 1 \qquad 1.19$$

da bir normal dağılımdır. Ancak bu önceki dağılımın f(x) = f(g) şekli ile bir normlanmış halidir (Şekil 1.4b). Aşağıdaki grafikte gelen zarın yüzlerinin eşit olasılıklarını, α, β da olasılık sınırlarını göstermektedir (solda; b, birikimli dağılımdır):





Şekil 1.4. Olasılık tanımı. **a**, sıklık dağılımı, **b**, birikimli sıklık dağılımı. x ve g argümanları farklıdır. x, sürekli özellik koordinatlarını (sınıf ortalarını), g ise, sınıf üst sınırlarını (özellik üst sınır değerlerini) gösterir (a'daki taralı alan olasılığı gösterir).

Olasılık, P(O) = % 100 veya 1/1 olabilir ve **mümkün** veya olasıdır. Ancak P(O) = 0, yani 0/1 olması mümkün değildir (0 olasılığa "boş" denir). Çünkü olasılık, 0<P(O)<1 arasında yer alır (istatistiksel olasılık). Buna hafif bir yağmurun 2 taştan birine 1 damlasının zamanla düşmesi (bireysel olay) örnek verilebilir. Buna karşın yoğun yağmur damlası büyük olasılıkla taşa hemen çarpar (kolektif olay).

Örnek 1.4

1 zarla 1 defada:

a) 6,
b) 5 veya 6 ve
c) Zarın para ile birlikte atılmasında 2 ve tura gelme olasılıklarının hesaplanması*:*

a) P(O) = 1:6 ≈ <u>% 17</u>,

b) P(O) = (1:6)+(1:6) = 2/6 = <u>% 33</u> ve

c) $P(O)=\frac{1}{6}.\frac{1}{2}=\frac{1}{12}=\underline{\%\ 8{,}33}$

bulunur.

Açıklama: Ortak elemanları bulunmuyan (ayrık veya bağdaşmıyan) olayların olasılıkları P(A∪B)=P(A)+P(B) olarak hesaplanır (bak. b şıkkı).

Örnek 1.5

a) 2 zarla 1 atışta 2 altı atma olasılığı: $P(O) = (1/6)\,(1/6)$

$$= (1/6)^2$$
$$= 1/36$$
$$= 0{,}02778$$
$$= \underline{\%\ 2{,}78}$$

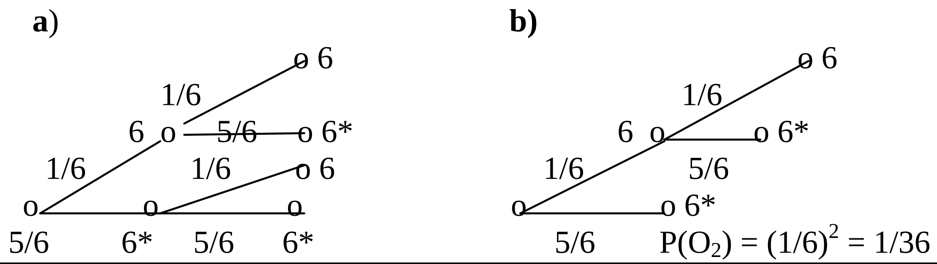
elde edilir.

b) 1 zarla 4 atışta 6 atmama olasılığı: $P(O)=(5/6)^4$

$$= \underline{\%\ 48{,}25}$$

bulunur.

Deney, zarların arka arkaya atılması ile aşağıdaki diyagramların gösterdiği gibi 2 kısma ayrılır (Şekil 1.5): Gelme ve gelmeme olasılığı.



6: 6 gelme, 6 *: 6 gelmeme olasılığı = 5/6.

Şekil 1.5: 2 kısma ayrılan zar atmanın ağaç diyagramı.

Sınırlı bir örnek sayısı da rastlantısal olayların dağılım hakkında yeterli bilgi sağlayabilir. Rastlantısal olaylarda **bağımsızlık esası** bulunmaktadır. Rastlantısal A ve B olayları,

$$P(A \cap B) = P(A)\,P(B) \qquad 1.20$$

ise, bağımsızdır. Burada,

$$P(A/B) = P(A) \quad \text{ve}$$
$$P(B/A) = P(B)$$

şeklindedir. P(A/B), B koşullarında A'nın gerçekleşmesi olasılığı demektir. Örneğin, deprem riski ve malzeme dayanıklığı gibi. Bunlara ek olarak hem o_1 hem de o_2 olayları ve o_1 veya o_2 olayları gibi olaylar birbiri ile ilişkilendirilebilir. Olasılıklar çarpma ve toplama gibi işlemlere de tabi tutulabilir. Örneğin,

$$P\{A, B\} = P(A).P(B/A) \qquad 1.21$$

veya

$$P(A).P(B) = 0,$$

mümkün olmayan olasılık gibi. Özet olarak toplam olasılıkların oranı,

$$P(B_i/A) = \frac{P(B_i)P(A/B_i)}{\sum_{j=i}^{n} (P(B_j)P(A/B_j))} \qquad 1.22$$

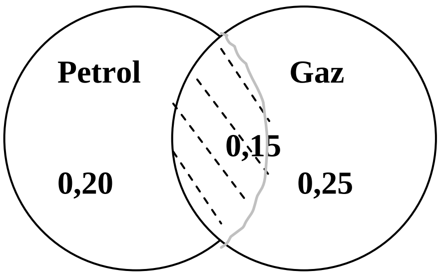
şeklindedir (Bayes teoremi, bak dipnot s. 25). Olasılığın genel toplam eşitliği ise,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \qquad 1.23$$

ile ifade edilmektedir ($\cap$: ve, $\cup$: veya okunur).

1.6 Venn Diyageamı

Mühendislikte **olasılığın rolü** büyüktür. Örneğin, **Venn diyagramı**, **olabilirlik** (possibilities) ve **olasılığı** (probablities) göstermek için kullanılan en iyi araçtır (Şekil 1.6, genel olasılık toplamı diyagramı). Doğada petrol ve gazın beraber oluştukları bilinmektedir. Bunlar karışım veya katı-sıvı-gaz fazları halinde dengede bulunurlar. Şekille petrol ve doğal gazın hangi olasılıkla dengede bulunabilecekleri açıklanmaktadır.



P(petrol) = 0,20 + 0,15 = 0,35

P(gaz) = 0,25 + 0,15 = 0,40

$P(\text{petrol} \cup \text{gaz}) = 0{,}20 + 0{,}15 + 0{,}25 = 0{,}60$

$= \text{Petrol} + \text{Petrol} \cup \text{Gaz} + \text{gaz}$

Şekil 1.6.: Petrol ve gazın bulunma olasılığını gösteren venn diyagramı.

Sadece bir toplama işlemi uygulandığında,

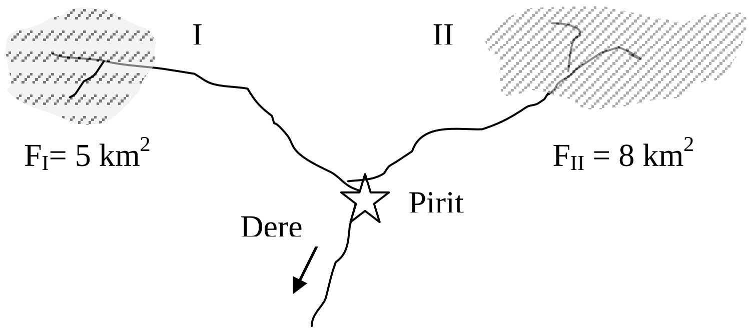
$$\begin{aligned} P(petrol \cup gaz) &= P(petrol) + P(gaz) \\ &= 0{,}35 + 0{,}40 \\ &= 0{,}75 \end{aligned}$$

bulunur. Ancak bu değer toplam değerin üstünde olduğundan, sonucun düzeltilmesi lazım ve

$$\begin{aligned} P(petrol \cup gaz) &= P(petrol) + P(gaz) - P(petrol \cap gaz) \\ &= 0{,}35 + 0{,}40 - 0{,}15 \\ &= \underline{0{,}60} \end{aligned}$$

gerçek sonucu elde edilir.

Genel olasılığa 2. örnek: Altın içeren bir pirit bloku (FeS_2) 5 ve 8 km^2 büyüklüğündeki iki bakır yatağının yer aldığı bir vadide bulunmuştur (bak. Şekil 1.7). Bloğun hangi yataktan gelme olasılığı daha büyüktür?



Şekil 1.7: Bayes teoremine bir örnek.

Genel olasılık formülü,

$$P(B_i/A) = \frac{P(B_i)P(A/B_i)}{\sum_{j=i}^{n} (P(B_j)P(A/B_j))}$$

şeklindedir (1.21; A, gerçekleşme olasılığı). Buradan “B_1 = 1., B_2 = 2. yataktan geliyor” tezleri ise,

$$P(B_1) = 5/13 = 0{,}38 \text{ ve}$$
$$P(B_2) = 8/13 = 0{,}62$$

bulunur. Ancak B_1 alanının % 60, B_2 alanın ise, % 80'inin denizaltı magmatizması ürünü olduğu saptanmıştır. Bu,

$P(A/B_1) = 0{,}60$ olasılıkla blok B_1'den,

$P(A/B_2) = 0{,}80$ olasılıkla da blok B_2'den

gelmesi ve denizaltı magmatizması ürünü olması demektir (**Bayes*** **Teoremi**). Bu olasılıklarla B_1 için koşullu olasılık,

$$P(B_1/A)=\frac{P(A/B_1)P(B_1)}{P(A/B_1)P(B_1)+P(A/B_2)P(B_2)}$$

$$=\frac{0{,}60.0{,}38}{0{,}60.0{,}38+0{,}80.0{,}62}$$

$$= 0{,}228/0{,}724$$

$$= 0{,}31$$

bulunur. Aynı şekilde 2. alan için 0,69 bulunur. Bu sonuçlarla bloğun % 69 olasılıkla 2., yani büyük alandan geldiği, tahmin edilir.

Alıştırma 1.1

Yinelenmiyen 1, 2, 3 ve 4 sayılarının kombinasyon sayısını bulunuz ve ağaç diyagramını çiziniz.

Yanıt: 27 seçenek

Alıştırma 1.2

10 kız ve 20 erkek öğrencinin bulunduğu bir sınıfın yarısı yatılıdır. Rastlantısal seçilecek 1 öğrencinin erkek veya yatılı olma olasılığı % kaçtır?

Yanıt: % 83,33

Alıştırma 1.3

Bir aygıt için alınan 3 W'lık ampüllerin % 50'si L, % 30'u M ve geri kalan da N markasına aittir. 300 saatlik kullanımdan sonra bozulan ampüllerin % 15'i L, % 10'u M ve % 5'i de N markasınındır.

1. Rastlantısal seçilen L markasına ait bir ampülün 300 saatlik kullanımdan sonra bozulma olasılığını bulunuz. Yanıt: % 7,50
2. Rastlantısal seçilen bir ampülün 200 saat içinde bozulma olasılığını hesaplayınız. Yanıt: % 11,50

---

*Thomas Bayes (1702-1761), ingiliz din ve bilim adamı. Bayes Ağı'ında kenarlar nedenden sonuca doğru yönlenmiştir. Bayes Teoremi ile bir sonuç, nedeninden sonucuna doğru ve tersine hesaplanabilmektedir.

2. TEK BOYUTLU DAĞILIMLARIN TANIMLANMASI

2. 1 Giriş

Bir örnek dizisi (örneklem) ancak yeteri kadar kapsamlı olduktan ve kesinlikle tanımlandıktan sonra istatistiksel bir anlam ifade eder ve verilerin ait olduğu örneklem özellikleri ile gelecekte beklenen olasılıklar hakkında bir düşünce yürütülebilir. **Tek boyutlu** dağılım demek, **tek** bir **değişken**le tanımlanan ana kitledir (popülasyon). Bu ana kitledeki değişkenlerin kesinlik derecesi (olasılığı) ve birimlerinin aynı olması istenir. Örnek çeşidi olaylara, yani incelemelere, bağlıdır. Genelde f(x, y) ve f(t) şeklinde ifade edilirler. Alınmış ve alınacak örnekleri ayırmak lazım. Alınacak örnekler açıklık kazanacak. Simüle edilebilir ve hazırlanabilir olmalıdır.

Örnek tanımlanması, verilerin değişken ve faktör şeklinde özetlenmesi demektir. Bunun için:

1. Ortalama ve en sık değer,
2. Saçınım, sapma değerleri (değişkeler),
3. Sıklık oranları (kısımları, yüzdelikler),
4. Dağılım şekillerini gösteren değişkenler (simetri, yassılık)
5. Eğilim/eğim ve yön, gösteren değişkenler ve
6. Dağılım fonksiyonları, uyum ve sınıflandırma

bakımından örnekler karşılaştırılmalı ve incelenmelidir.

2.2 Merkezi değerler

Bir veri dizisinde bulunan değişkenlerin bazıları bu dizinin orta kısımlarında yer alırlar ve merkezi değer olarak anılırlar. Bir örneklemin merkezi **değerler**i, **ortalama** değer, **tepe** değeri (mod) ve **ortanca**dır (medyan). 1. konuda tanımlanan kavramlar, bu ve bundan sonraki konuda veri kümelerinin değişkenlerinin ve sıklık dağılımlarının tanımlanmasında kullanılacaklar. Veriler çizelge, şema ve diyagram şeklinde düzenlenerek çeşitli ölçü sayıları bulunacaktır.

2.2.1 Ortalama değerler

Bir rastlantısal veri hem dağılım fonksiyonu, hem de sıklık dağılımı ile kesinlikle tanımlanır. Ancak istatistikte rastlantısal bir büyüklüğü yeteri kesinlikle tanımlayan bazı değişkenlerle de yetinilir. Bu değerlerin en önemlisi istatistikte en çok kullanılan **beklenti değeri** veya **ortalama değer**dir. Ortalama değer, rastlantısal x değerlerinin

$$\sum_i x_i P(x_i) \qquad 2.1$$

i üzerinden sonlu veya sonsuz toplamı veya

$$\int_{-\infty}^{\infty} t.f(t)dt \qquad 2.2$$

fonksiyonunun değeridir (f , sıklık fonksiyonudur). Ortalama değer, verilerin çan eğrisinin altında kalan alanın ağırlık merkezini oluşturur. Çok çeşitli olan ortalama değerlerin en yaygın kullanılanları aşağıya çıkarılmıştır.

a) Aritmetik ortalama ($\bar{x}$)

Aritmetik ortalama verilerin, çan eğrisinin altında kalan alanın ağırlık merkezini oluşturur. Gauss çan eğrisi için önemli olan **aritmetik ortalama** değeridir. Bu, bir örnekleme ait değerlerin toplamının toplam örnek sayısına bölümü ile elde edinilir.

1. Sınıflandırılmamış verilerin aritmetik ortalama değeri $\bar{x}$,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad 2.3$$

$$= \frac{1}{n}(x_1+x_2+...+x_n) \qquad 2.4$$

şeklinde tanımlanmaktadır. x_i, veri veya özelliktir (i =1,2,..., n). Toplam i üzerinden n'ye kadar alınır. Buna göre Çizelge 2.1'deki ölçümlerin aritmetik ortalaması $\bar{x}$,

$$\bar{x} = \frac{1}{17}106$$

$$= 6{,}235 \text{ °C}$$

bulunur.

Çizelge 2. 1: Isı ölçümleri (˚C, Schönwiese, 1992).

Ölçüm numarası n	**Ölçüm değerleri x_i ˚C**	**x_i-D (D = 6 ˚C)**	**Ortalama sapma* $\lvert x_i - \bar{x} \rvert$**	**Sapmanın karesi $(x_i - \bar{x})^2$**
1	5,6	-0,4	0,635	0,403
2	5,8	-0,2	0,435	0,189
3	**4,9**	-1,1	1,335	1,782
4	6,8	0,8	0,563	0,319
5	6,4	0,4	0,165	0,027
6	7,4	1,4	1,165	1,357
7	6,5	0,5	0,265	0,070
8	**7,7**	1,7	1,465	2,146
9	5,5	-0,5	0,735	0,540
10	5,7	-0,3	0,535	0,286
11	6,5	0,5	0,265	0,070
12	5,5	-0,5	0,735	0,540
13	6,7	0,7	0,465	0,216
14	6,9	0,9	0,665	0,442
15	6,2	0,2	0,035	0,001
16	6,0	0,0	0,235	0,055
17	5,9	-0,1	0,335	0,112
Σ	**106,00**	**4,0**	**10,035**	**8, 555**

2. Sınıflandırılmış verilerin ortalama değeri için $\bar{x}$,

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} \qquad 2.5$$

formülü kullanılır. Burada, k, sınıf (aralık) sayısı, f_i, i sınıfına düşen veri sayısı ve x_i, sınıf orta noktasıdır.

Örnek 2.1

Çizelge 2.2‘deki sınıflandırılmış değerlerin ortalaması $\bar{x}$,

$$\bar{x} = \frac{1}{17} \sum_{i=1}^{5} (106{,}75)$$

$$= 6{,}279$$

$$= \underline{6{,}28\ ˚C}$$

*Bak. yrıca 2.4.2

bulunur (değerlerin kısaltılması nedeniyle sonuç az yüksek çıkmıştır).

Çizelge 2.2: Çizelge 2.1'deki verilerin sınıflandırılması (bak. ayrıca örnek 2.6 ve konu 3).

Sınıf aralığı, frekans n, ˚C	**Sınıf ortası x_i**	**Örnek sayısı f_i**	**Örnek oranı $x_i.f_i$ [%]**	**Birikimli toplam, %**
4,9-5,4	5,15	1	5,15 (4,82)	4,82
5,5-6,0	5,75	7	40,25 (37,70)	42,52
6,1-6,6	6,35	4	25,40 (23,79)	66,31
6,7-7,2	6,95	3	20,85 (19,53)	85,84
7,3-7,8	7,55	2	15,10 (14,16)	**100,00**
Σ		**17**	**106,75 (% 100)**	

3. Sabit bir değer yardımı ile hesaplama. Burada $\bar{x}$,

$$\bar{x} = \mathrm{D} + \frac{1}{k} \sum_{i=1}^{k} (\chi_i - D) \qquad 2.6$$

şeklinde ifade edilir. (D, gelişigüzel bir değerdir). D = 6 ˚C alındığında, yukarıdaki çizelgeye göre $\bar{x}$,

$$\bar{x} = 6 + \frac{4{,}0}{17}$$

$$= 6{,}235\ ˚\mathrm{C}$$

bulunur (Çizelge 2.2).

Aritmetik ortalama çok kullanışlıdır, analiz değerlerinin irdelenmesinde en sık kullanılan merkezi eğilim ölçüsüdür. İrdelemeden kabul etmek doğru değildir. Örneğin, uç değerlerden çok etkilenir. En önemli **üstünlükleri,**

a) Kolay hesaplanması,
b) Tüm değerleri kapsaması ve
c) Başka değerlere, örneğin ağırlıklı ortalamaya, çevrilmesinin kolay olmasıdır.

Aritmetik ortalamanın **sakıncaları**,

a) En büyük ve en küçük değerlerin çok etkilemesi,

b) Gerçek bir değerin bulunmadığı bir noktada bir değer verebilmesi ve

c) Metrik sistemin şart olması.

Diğer ortalama değer çeşitleri aşağıda kısaca tanıtılmıştır (bak. ayrıca Tüysüz ve Yaylalı, 2005).

b) Ağırlıklı ortalama (x_a)

Çeşitli hesaplamalarda süre ve mesafe gibi etkenlerin önemini de hesaplamalara katmak için **ağırlıklı ortalama** hesaplanır. Örneğin yer ve çevre bilimlerinde örnek değerleri, örneğin derişimler, örneklerin arasındaki mesafe ile çarpılır, toplanır ve toplam mesafeye bölünür. Bu ortalama değerle, uç değerlerin, örneğin kalınlıkların da hesaplamada etkin olması ile, aritmetik ortalamaya oranla etkisi azaltılmış oluyor ve daha gerçekçi bir değer elde edilmiş olur.

Ağırlıklı ortalama

$$x_a = \frac{\sum_{i=1}^{n} m_i x_i}{\sum_{i=1}^{n} m_i} \qquad 2.7$$

formülü ile hesaplanır. $\sum_{i=1}^{n} m_i = 1$ durumunda sadece genel formül geçerlidir. Burada i, örnek sayısı (i=1, 2,...,n), m_i, örnekler arasındaki mesafe veya süredir ve x_i'ye göre değişen değerler alır. x_i, ölçüm değerlerini gösterir. Ağırlıklı ortalamaya en iyi örnek, ara ve bitirme sınavı notlarının farklı ağırlıkla (örneğin, % 40 ve % 60) geçme notuna katılmasıdır.

Örnek 2.2

Bir bölgeye düşen yağış miktarının hesaplanması (Schönwiese, 1992).

Tarih	Ölçüm x_i	Süre (yıl)	Faktör m_i	$x_i.m_i$
1920-1950	690	30	3	20.700
1950-1960	676	10	1	6.760
1960-1980	678	20	2	13.560
Σ	**2.044**	**60**	**6**	**41.020**

Buna göre,

$$x_a = \frac{690.3 + 676.1 + 678.2}{3 + 1 + 2}$$

$$= \frac{4102}{6}$$

$$= \underline{683{,}66} \text{ mm}$$

yağış bulunur. Bu değerlerin aritmetik ortalaması,

$$\bar{x} = 2044/3 = \underline{681,33}$$

mm'dir. 2 mm'lik yağış farkı önemli değildir.

---

c) Geometrik ortalama (x_g)

Bunlardan başka mühendislik bilimlerinde yaygın bir şekilde kullanılan **geometrik ortalama** bulunmaktadır. Bu ortalama değer öncelikle log dağılımlarında kullanılır. Burada değer çarpımının çarpılan tüm değerlerin sayısına eşit dereceden kökü alınır. Yani,

$$x_g = \sqrt[n]{\Pi_{i=1}^{n} . x_i} \qquad 2.8$$

$$= \sqrt[n]{x_1 . x_2 ... x_n}$$

şeklinde bulunur (n, örnek sayısı; Π, büyük pi = çarpım). x_g, aritmetik ortalamadan küçüktür. Geometrik ortalamanın **üstünlükleri**,

a) Tüm değerlerin kullanılması,
b) Çok açıktır ve
c) uç değerlerin etkisini azaltır.

Buna karşın **sakıncaları,**

a) Bir değerin dizide sıfır olması durumunda kullanılmaması ve
b) Elde edilen değerin dizide bulunmayan bir yere karşılık gelmesidir.

Örnek 2.3

Ölçülen çeşitli deneylerdeki yoğunluk ortalaması.

Ölçüm	1	2	3	4	5
Yoğunluk [g/cm^3]	5,4	5,6	5,3	5,7	5,5

verileri verilmiş ise, geometrik ortalama, $x_g = \sqrt[n]{x_1.x_2...x_n}$

$$= \sqrt[5]{5,4.5,6.5,3.5,7.5,5}$$

$$= 5.024,53^{1/5}$$

$$= \underline{5,50}$$

g/cm³ sonucuna varılır.

Geometrik ortalama hesaplamada kök yerine logaritma alınması, çok veri çarpımının kökünü alma işlemindeki zorluktan kaynaklanır. Bu nedenle kök alma yerine logaritma alınır. Örneğin,

$$\log x_g = \frac{1}{n}\log(x_1 + x_2 + ... + x_n) \qquad 2.9$$

$$= \frac{1}{n}(\log x_1 + \log x_2 + ... + \log x_n)$$

logaritma kuralına göre logaritmaları alınan değer toplamının anti logaritması alınarak geometrik ortalama bulunur. Örnek 2.3'teki değerlerin log toplamı 3,69 ve örnek sayısı da 5 olduğuna göre,

$$x_g = \text{antilog}\ (3,69/5)$$

$$= \text{antilog}\ 0,74$$

$$=10^{0,74}$$

$$= \underline{5,50\ [g/cm^3]}$$

bulunur.

d) Harmonik ortalama (x_h)

Özellikle doğa bilimlerinde, örneğin fizikte, kullanılan bir ortalama değer de **harmonik** değerdir. Ölçüm değerlerinin tersi (1/a) aritmetik ortalamasının tersi olarak hesaplanır ve

$$\frac{1}{x_h} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i} \qquad 2.10$$

şeklinde formüle edilir. Burada da geometrik ortalamada olduğu gibi değerlerin hepsinin sıfır-dan büyük olması gerekir. Harmonik ortalama öncelikle zaman oranlarının (hız, yol ve zaman) ortalamasının hesaplanmasında kullanılır.

Örnek. 2.4

---

80 soruluk bir test sınavında 60 soruya 50, 20 soruya da 30 dakikada yanıt veren bir öğrencinin yanıt başına ortalama yanıt süresi.

Sınav başında ortalama olarak her soruya 1 dakika süre verilmiştir. Ancak öğrenci sınavın ilk 50 dakikasında soru başına 50/60 = 5/6 dakika, geri kalan 20 dakikasında da 30/20 = 3/2 dakika kullanmıştır. Buna göre harmonik ortalama,

$$1/x_h = \frac{1}{2} \sum_{i=1}^{n} \frac{1}{x_i}$$

$$= 1/2(5/6+3/2)$$

$$= 1/2(5/6+9/6)$$

$$= 1/2.14/6$$

$$= 14/12$$

$$= 7/6$$

$$= \underline{1{,}17\ \text{dak}}.$$

elde edilir.

---

e) Değişim aralığı ortası (R_o, uç değer ortası):

Bu ortalama değerlere ek olarak **değişim aralığı ortası** da bir ortalama değer sayılmaktadır. Değişim veya yayılım aralığı R (range),

$$R = x_{max} - x_{min} \qquad 2.11$$

. olarak tanımlanmaktadır (bak. ayrıca 2.4.1). Bunun ortalaması veya ortası R_o,

$$R_o = \frac{x_{\max} - x_{\min}}{2} \qquad 2.12$$

formülü ile bulunur ve yayılım alanın ortasını verir. Bu aralık frekans veya sınıf sayısının bulunmasında kullanılır (bak. Sturges Kuralı, Örnek 2.6).

2.2.2. Ortanca (x_o, medyan)

Ortanca, bir diziyi ortalayan veya çan eğrisinin altındaki alanı eşit iki kısma ayıran değerdir. Örneklemdeki değerlerin yarısı ortancadan küçük, yarısı da büyük olur.

Ortancanın **üstünlükleri**,

a) Hesap yapmadan bulunabilmesi
b) Gerçek bir değeri temsil etmesi,
c) Uc değerleri dışlaması ve
d) Başka benzer büyüklükler yerine kullanılmamasıdır. Buna karşın,

a) Bazen zor saptanabilmesi ve
b) Çevirmelere uygun olmayışı

gibi **sakıncaları** bulunmaktadır.

Matematiksel olarak ortanca x_o,

$$x_o = \int_{-\infty}^{or} f(x)dx \qquad 2.13$$

$$= [F(x)]_{-\infty}^{x_o}$$

$$= F(x_o) - F(-\infty)$$

$$= x_o$$

$$= 0{,}5$$

veya

$$x_o = \int_{or}^{+\infty} f(x)dx$$

$$= \sum_{i=1}^{or} f(x_i) \qquad 2.14$$

$$= 0{,}5$$

olarak tanımlanmaktadır (Schönwiese, 1992 ve Arıcı, 2001). Bu değer toplam için ortanca = F(0,5) = % 50 anlamına gelir. Bu, **tek sayılı dizilerde**, büyüklük sırasına göre sıralanan ölçümleri 2'ye bölen,

$$x_o = (x_{n+1})/2 \qquad 2.15$$

orta değerine karşılık gelir (bak. Örnek 2.5). Çift sayılı dizilerde ise,

$$x_o = \frac{(x_{n-1})/2 + (x_{n+1})/2}{2} \qquad 2.15a$$

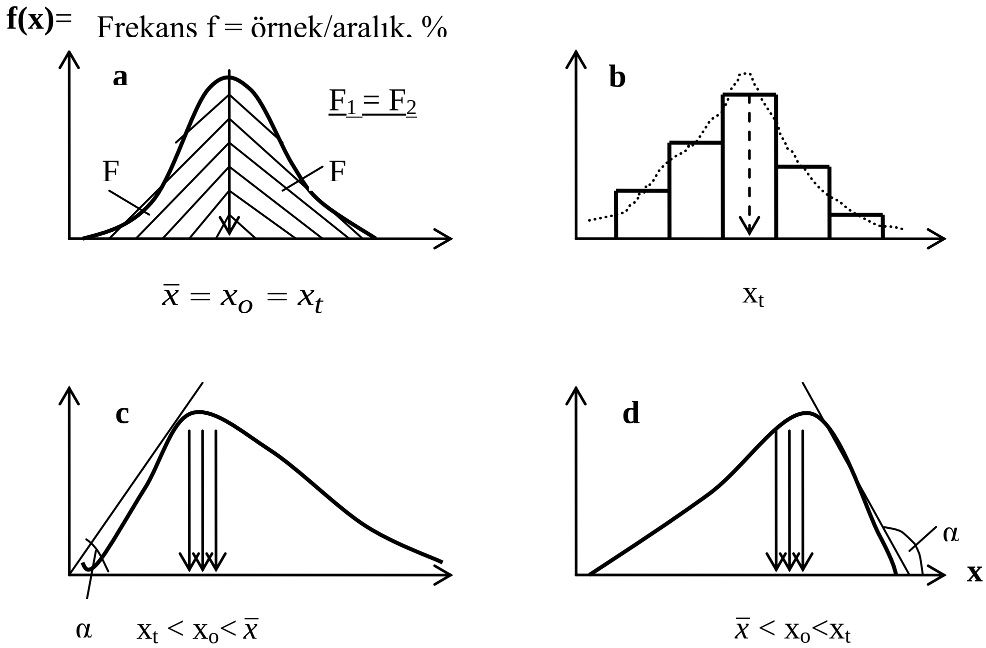
şeklinde belirlenmektedir. x_{n-1}, ortancadan küçük, x_{n+1} ortancadan büyük ortadaki iki değerdir (bak. ayrıca Şekil 2.1).

Örnek 2.5

Bir akarsuyun ölçülen debi değerleri (m^3/s):

Sıra no.	1	2	3	4	5	6	7	**8**	9	10	11	12	13	14	15

Veriler: 3,9-4,5-4,5-4,6-4,7-4,8-4,9-**5,0**-5,1-5,3-5,5-5,5-5,6-5,7-5,8



a, çan eğrisi ve ortalama değer $\bar{x}$, ortanca x_o ve tepe değeri x_t değerlerinin çakışması. F_1 ve F_2 ortanca tarafından 2'ye bölünen alanın eşit parçalarıdır. **b**, histogram, eğrinin yerleşimi ve ortanca x_o; **c**, sağa çarpık ve **d**, sola çarpık dağılımda tepe değeri-ortanca-ortalama değerin konumu. α, teğet-absis açısı. tg α > 0: düşük değer dağılım tipi (c) ve tg α < 0, yüksek değer dağılım tipi (d) dağılım grafiği (bak. Wilke, 1973).

Şekil 2.1. Çan eğrisi ve merkezi değişkenlerinin dağılıma göre konumları.

Büyüklük sırasına göre dizilen yukarıdaki değerlerin 8. değeri, yani 5,0, veri dizisinin tam ortasına düşmektedir. Burada ortanca bu değerdir. Ortanca çarpık dağılımlar için ortalama değere göre daha uygundur ve az örnek sayısı için de önemlidir. Metrik ve sıralama ölçü birimleri de kullanılabilir ve aşırı değerlerden çok etkilenir.

2.2.3 Tepe değer (x_t, mod)

Tepe değeri veya mod, dağılım fonksiyonunun sahip olduğu en yüksek frekans değeridir. Yani en çok örneğin bulunduğu, başka hiçbir sınıf tarafından aşılamayan sınıftır. Teorik olarak td,

$$t\ (mod) = f(x)_{max} \qquad 2.16$$

olarak tanımlanmaktadır. Bir normal dağılımda, simetriden dolayı,

$$\text{tepe değeri } (x_t) = \text{ortalama değer } (\bar{x}) = \text{Ortancadır } (x_o) \qquad 2.17$$

bağıntısı mevcuttur. Bu durumda eğri altındaki alan ortanca tarafından 2 eşit parçaya yarılır ($F_1 = F_2$, Şekil 2.1.). **Sağa kayan** (kuyruk sağda) veya **pozitif eğimli** (tg $\alpha > 0$) dağılımlarda ortalama değer, ortanca ve tepe değerinden büyüktür ($x_t < x_o < \bar{x}$), **sola kayan** (kuyruk solda) veya **negatif eğimli** (tg $\alpha < 0$) dağılımlarda ise, ortanca ve tepe değer'den küçüktür ($\bar{x} < x_o < x_t$). 1. durumun zayıf değerlerin dağılıma hakim olduğu anlamına gelir (fakir tip). 2. durum ise, örneğin yerbilimlerinde maden yataklarının oluşmasını sağlarlar ve yüksek değerlerin hakim olduğu bir dağılımı yansıtır (zengin cevher tipi). Çevre bilimlerinde bu, örneğin aşırı kirliliği veya bir etkenin ortalamanın üstünde artışını gösterir. Şekil 2.1 c-d çarpıklık ve eğri-teğet ilişkisini göstermektedir. Birçok tepelikli dağılımlarda, 1., 2. ve 3. tepe değeri diye sıralanır.

Örnek 2.5'te yinelenen <u>4,5</u> ve <u>5,5</u> değerleri dizinin tepe değerlerini oluşturmaktadır (2 tepeli dağılım). Dizide bu değerlerden daha çok sayıda yinelenen değer bulunmamaktadır.

Bu değerlerin hepsi de gözlem sayısına bağlıdır. Ne kadar çok gözlem kapsama alınırsa, o kadar güvenilir ve gözlem kümesi için o kadar tanımlayıcı olurlar. Tepe değerinin **Üstünlükleri**,

a) 2 veya daha çok tepelikli dağılım için de önemlidir,
b) Uç değerlerden etkilenmez,

c) Her ölçü birimi için kullanılır.

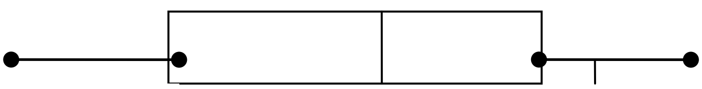
Sakıncası, her zaman kesin ölçülememesidir.

2.3 Yüzdelikler

Günlük yaşamda sıkça çeyrek, yarım ve ondalık gibi deyim ve ifadelerle karşılaşılmaktadır. Bunlar bir bütünün belli bir kısmına, ¼'üne (çeyrek) veya yarısına (yarım) karşılık gelirler. Bu terimler modern bilimdeki yüzdelere dönüştürüldüğünde **yüzdelikler** meydana gelir. Örneğin 1 Çeyrek % 25', 2 çeyrek 1 yarıma (%50) veya 3 çeyrek ¾'e (%75) karşılık gelir. Batı dillerinde bu terimler **kantil** olarak geçerler (Şekil 2.2.). Kolay hesaplanan yüzdeliklerin çok değerden hesaplandıkları nedeniyle güvenirlikleri yüksektir.

Bunlara karşılık yüzdeler sıraya dizilecek olursa, örneğin % 25 çeyrek olarak anıldığı gibi, % 10, ilk veya 1. ondalık (1. desil) olarak ifade edilir. Bunun gibi % 20, 2. onluk (2. desil, veya 1. pentil) ve 20. yüzdelik olarak da söylenebilir. Aynı şekilde % 80, 8. onluk (4. pentil) ve 4. yirmilik olarak da ifade edilebilir.

Bu değerleri hesaplamak için çeşitli formüller kullanılmaktadır. Genel olarak ortancanın hesaplanmasına benzeyen yöntemde bir temel yüzdeliğe diğer yüzdelikler eklenir. Örneğin, 1. çeyrek $y_{25} = q_1 = F(0{,}25)$, $y_{50} = q_2 = F(0{,}50)$ = ortanca olduğu anlaşılır.



En düşük, min **En yüksek, max**

% 25 **Ortanca** **% 75** **%80**

1. çeyrek (kantil) 3. çeyrek 4. beşlik (pentil), 8. onluk (desil)

Şekil 2.2: Yüzdeliklerin şematik tanımı.

2.4 Değişkenlik ölçüleri

Verilerin istatistiksel tanımlanmasında sadece ortalama değerler yeterli olmamaktadır. Ortalama değerlerle elde edilen sonuçların yanılma oranları yüksek olur. Örneğin, bir kıracı ev sahibinin verdiği 1,40 m'lik çocuklarının yüzdüğü havuzun ortalama derinliğine ne kadar inanır? Bu gibi durumlar için ortalama değerler tarafından kapsanmayan değişkenleri de incelemek gerekir. Bir veri kümesi veya popülasyon çeşitli **dağılım değişkenleriyle** de tanımlanabilir. Bu değişkenler aşağıda tanıtılan **değişim aralığı, standart sapma, değişke** ve **momentler**dir.

2.4.1 Değişim aralığı (R)

Bir veri kümesindeki veya popülasyondaki en yüksek ve en düşük değerleri arasındaki fark, **değişim** veya **yayılım** aralığı R,

$$R = x_{max} - x_{min} \qquad 2.18$$

ile tanımlanarak belirtilir. Burada sadece 2 değer, en düşük ve en yüksek değer, hesaba katılmaktadır. Dolayısı ile uç değerlerin bulunması durumunda sonuç yanıltıcı olabilir. Örneğin, bir ölçüm dizisinin veya bir madde çeşidinin içerdiği bir bileşen, bu uç değerleri arasındaki farkı değişik olabilir. Yüksek orandaki değişkenlerde bu fark nispeten küçüktür (homojen dağılım). Düşük bileşenlerde ise, büyük olabilmektedir (heterojen dağılım). Uc değerlere göre çan eğrisinin değişimini Şekil 2.3 göstermektedir.Örneğin, yukarıdaki debi ölçümü için değişim aralığı,

$$R = 5,8-3,9 = 1,9$$

m^3/s'dir. Sadece metrik ölçü birimlerinde yararlanılan değişim aralığının hesaplanması kolaydır ve verilerin dağınıklığı hakkında bilgi verir. Sadece 2 değerden hesaplandığından, içerikle ilgili yorum yapılamaz.

2.4.2 Ortalama mutlak sapma (d)

Daha gerçekçi dağılım için **o**rtalama mutlak sapma d,

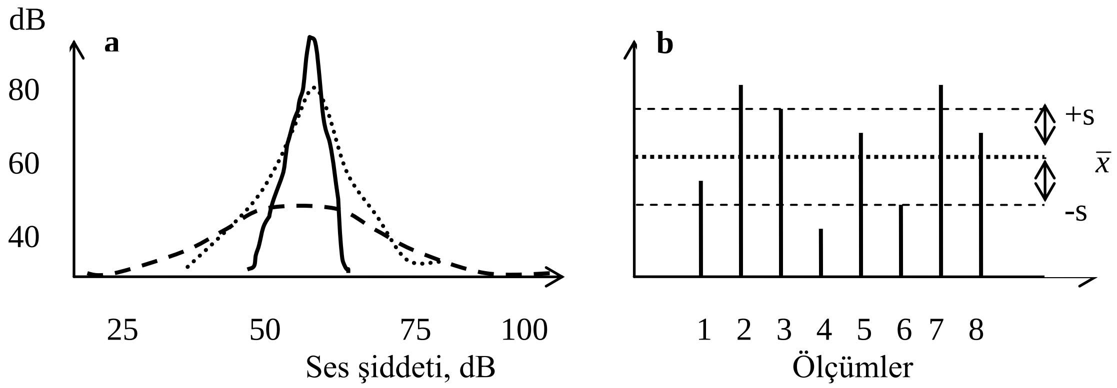
$$d = \frac{1}{n}\sum_{i=1}^{n} x' \qquad 2.19$$

$$= \frac{1}{n}\sum_{i=1}^{n} \left| x_i - \bar{x} \right|$$

hesaplanır. Bu değer, saçınımların ortalama değeridir (bak. Şekil 2.3b). Başka herhangi bir değer için bulunacak sapma değeri bu değerden küçük olamaz. Bu değer hem metrik, hem de sıralama (ordinal) veriler için anlamlıdır. Özellikle metrik veri kümelerinde aşırı eğimlerin veya çok tepeli durumların bulunması halinde hesaplanması salık verilir. Ancak saçınım ölçüsü olarak sadece metrik sayılar için kullanılabilir. Çünkü bu veri sıra sayıların farkları hakkında bir bilgi vermez.

2.4.3 Standart sapma (s)

Standart sapma, verilerin ortalama değer etrafında ne kadar yoğunlaştıklarını gösterir. Geometrik olarak, standart sapma, aritmetik ortalama değerinin üstünde ve altında kalan veri sapmalarının ortalamasıdır. Bu nedenle bir (+), bir de (-) değeri bulunur. Şekil: 2.3b standart sapmayı geometrik olarak göstermektedir.



Şekil 2.3: Verilerin değişim aralıkları **a** ve standart sapmanın geometrik anlamı **b**. +s, $\bar{x}$ 'dan büyük ($\bar{x}$ +s); -s ise, bundan küçük ($\bar{x}$ -s) ölçüm değerlerinin ortalamasıdır.

Ortalama mutlak sapmaya aynı zamanda **standart sapma** da denir. s ile gösterilir ve genelde,

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} x_i'^2} \qquad 2.20$$

$$= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad 2.21$$

$$= \sqrt{\frac{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}{n(n-1)}} \qquad 2.22$$

formülleri ile hesaplanır ($x_i' = x_i - \bar{x}$ 'dir). Ortalama değerle olan farkların toplamı 0 ettiğinden, kareleri alınır, toplanır ve karekökünün pozitif değeri standart sapma olarak kullanılır. Ortalama değer hesaplanmadan da standart sapma bulunabilir (bak. 2.22). 2.20'deki n yerine alınan n-1 teriminin sadece genel bir teorik anlamı vardır. Değerin popülasyona değil, örneklem'e ait olduğunu ifade eder ve serbestlik derecesi sayısını yansıtıyor (bak. konu 4). Bu terim aslında sınıflandırılmış verilerin standart sapması 2.22 eşitliğinin 2.5 eşitliği şekline benzetilmesi ile hesaplanabilir (bak. 2.5, $x_i = f_i.x_i$ ve $n(n-1) = \Sigma f_i$). Örnek sayısı arttıkça, n, n-1'e yaklaşır. Standart sapma s, n >1 ve s^2 >0 durumları için bir anlam taşır. 1 örnek için s^2 = +∞'dur (belirsiz). Sadece metrik ölçü birimlerinde kullanılan standart sapmadan, sonuçların yorumlanmasından nadiren yararlanılır.

2.4.4 Değişkenlik katsayısı (v)

Standart sapmanın aritmetik ortalamanın yüzdesi olarak ifade edilmesine **değişkenlik katsayısı** denir. **Göreceli standart sapma** olarak da bilinen bu değişken,

$$v = \frac{s}{\bar{x}}.100 \qquad 2.23$$

olarak bilinir (%). Bu orana göre veriler düzenli (v < % 40), düzensiz (% 40 < v < % 80) ve çok düzensiz (v > % 80) olarak sınıflandırılırlar (Wilke, 1973). Özellikle yerbilimlerinde yaklaşık örnek sayısının, örneğin, milyon t rezerv başına, saptanmasında bu değişken önemli bir ölçüt olarak kabul edilir. Ayrıca dağılımların saçınım farkını saptamada yararlanılır.

2.4.5 Değişke (s^2, σ^2, varyans)

Standart sapmanın karesine **değişke** (varyans) denir. s^2 veya sıkça σ^2 (sigma) ile gösterilir. Bir istatistiksel dağılımın esas değişkeni değişkedir ve

$$\sigma^2 = \int_{-\infty}^{+\infty} (x-\mu)^2 \, f(x)dx \qquad 2.24$$

olarak tanımlanır (σ^2, ana kütlenin değişkesidir). Buradan bir örneklemin değişkesi,

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2 \qquad 2.25$$

formülü ile hesaplanır. Verilerin ortalama değerden sapma derecesini, değişkenliğini, gösterir. Değişke, a = $\bar{x}$ için minimum değer alır (a, ortalama değerden farklı bir değerdir). Bu, saçınımların ortalama değer yakınında en az olduğunu gösterir.

Değişke ile standart sapmanın istatistikte çok büyük önemi vardır. İstatistiksel varsayımların doğruluğunda ve testlerde kullanılırlar. Varyans, bağınım ve bağıntı (korelasyon ve regresyon) analizi ile bunlara dayanan yüksek derecedeki istatistiksel değerlendirmelerde (örneğin, varyogram hesaplamalarında) vazgeçilmez bir ölçüttür. Ayrıca standart program yapımında temel teşkil ederler.

Uygulamada varyansla standart sapmayı birbirinden ayırt etmek gerekir. Aralarındaki fark, değişkenin gözlemleri tüm, standart sapmanın ise, tek tek göstermesidir. Bu yüzden değişkede bazen Σx_i toplam değerleri tüm örnek sayısına bölünürken, standart sapmada örnek sayısının 1 eksiğine bölünür. Bu nedenle σ'nın karekökü, standart sapmadan daima küçüktür. σ ve s değerleri bir güven derecesi içerirler. Bu da örneğin, student-t testi ile denetlenebilir.

Değişke, kısımlarına ayrılabilir. Çünkü bir örneklemin varyansı, örneklemdeki grupların varyanslarının toplamına eşittir. Örneğin,

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{I} x_i^2 + \sum_{j=1}^{J} x_j^2 + \ldots + \sum_{k=n-k}^{K} x_k^2\right) \qquad 2.26$$

gibi. Bu özellikten, varyans analizinde ve jeoistatistikte yayılım değişkesi gibi çeşitli değişkelerin hesaplanmasında yararlanılır.

2.4.6 Katışık değişke (s_{xy}, kovaryans)

Örneklem değişkesi s^2'nin,

$$s_x{}^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \qquad 2.27$$

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})}{n-1}$$

olduğu yukarıda gösterilmişti (2.25). Veri kümesinin x değerlerinin çift bulunması halinde, x_i ve y_i değerleri bu kümenin özellikleri olurlar (yerin derinlik ve basıncı veya ısı ile genleşme gibi). Yukarıda yazılan eşitlik y için de yazılabilir ve

$$s_y^{\,2} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} \qquad 2.28$$

$$= \frac{\sum_{i=1}^{n}(y_i - \bar{y})(y_i - \bar{y})}{n-1}$$

şekli x ve y için,

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \qquad 2.29$$

haline dönüştürülebilir. Bu, veri kümesine ait örneklemin x ve y özelliklerinin **katışık değişkesidir** (=kovaryans). Eşitlik, x ve y değişkenlerinin birbirine göre nasıl değiştiklerini gösterir. Bağıntı ve bağınım analizinde çok kullanılır. Bir örneklemin katışık değişkesinin değişkenlerin çarpımına oranı değişkenler arasındaki bağıntı katsayısını,

$$r = \frac{s_{xy}}{s_x . s_y} \qquad 2.30$$

verir. $r = |1|$ için bağıntı en iyi, $r = 0$ ise, bağıntı yok demektir. $-1<r<0$ durumunda değişkenler uyumsuz (x ve y arasında zıt gelişme), $0<r<1$ durumunda ise, uyumludur (x ve y birbirine bağlı; bak. konu 8).

2.5 Momentler

İstatistikte dağılımların çoğu simetrik değildir. Bazen değerler ortancanın sağında veya solunda yoğunlaşır (Şekil 2.1 c ve d). Bu dağılım özellikleri yukarıda açıklanan çarpıklık ve değişken sıralaması gibi yöntemler yanında **3. moment**; dağılımın standart çan eğrisinden yüksek veya basık (yassı) olduğu da **4. moment** veya **merkezi değerle*** saptanabilmektedir.

Bir moment genel olarak $$m_k = \frac{1}{n}\sum_{i=1}^{n}(x_i)^k \qquad 2.31$$

---

* **Merkezi değer** demek, ortalama değeri $\bar{x} = 0$ olan değer demektir. Bir transformasyon yoluyla, örneğin, $x_1, \ldots, x_n$ değerleri, $y_i = x_i - \bar{x}$ (i = 1,.., n) şeklinde merkezileştirilirler. Bu arada değişke σ^2 değişmez.

denklemiyle ifade edilmektedir. Bir dağılımda 4 büyüklüğün, **merkezi moment**lerin, yani ortalama değer ($\bar{x}$), değişke (s^2), kayma (m_3, çarpıklık) ve basıklığın (m_4, yassılık) incelenmesi yararlı olur.

2.5.1 Kayma (g, çarpıklık, asimetri; ing. skewness)

Bir normal dağılımda ortalama değer, ortanca ve tepe değerin eşit oldukları yukarıda belirtilmişti. Bu değerlerin karşılaştırılması ile ortalama değere göre kayma saptanabilir. Ancak kayma durumunda bunların oranı değişir.

Kesin bir kayma değeri ancak **3. moment m_3**'ün hesaplanması ile elde edilebilir. Bu,

$$m_3 = \frac{1}{n}\sum_{i=}^{n}(x_i - \bar{x})^3 \qquad 2.32$$

olarak tanımlanmaktadır. Buradan kayma, standart sapmaya bölünerek standartlaştırılır. Yani,

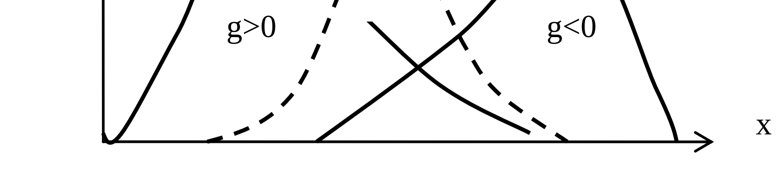
$$g = \frac{m_3}{s^3} \qquad 2.33$$

$$= \frac{1}{n}\sum_{i=}^{n}(x_i - \bar{x})^3 / s^3$$

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{n.s^3}$$

eşitliği ile birimsiz hale getirilir.

g = 0 için normal dağılım, g > 0 için pozitif ve g < 0 için de negatif eğimli (sağa kayma) bir dağılım mevcut demektir (Şekil 2.4). Kayma, esasında bir logaritmik dağılımın (dinamik dengenin) belirtisidir.



Şekil: 2.4. Farklı kaymalara örnek dağılım çeşitleri (bak. ayrıca Şekil 2.1).

2.5.2 Basıklık (e, sivrilik; ing. excess)

Basıklık veya sivrilik tanımı için de **4. moment m_4** (exess, kurtosis) kullanılır. Bu moment tepelik durumunu belirtmeye yarar ve

$$m_4 = \frac{1}{n}\sum_{i=}^{n}(x_i - \bar{x})^4 \qquad 2.34$$

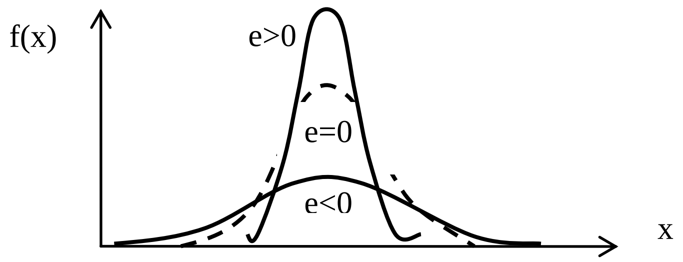
şeklinde tanımlanmaktadır. Buradan basıklık,

$$e = \frac{m_4}{s^4} - 3 \qquad 2.35$$

$$= \frac{1}{n}\sum_{i=}^{n}(x_i - \bar{x})^4 / s^4 - 3$$

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{n.s^4} - 3$$

formülünden çıkarılır. Bir normal dağılımda e = 3'tür (eşitlikte 0 alınmıştır). Bu durumda normal çan eğrisinden basık olan dağılımlarda e < 0, sivri olanlarda ise, e > 0'dır (Şekil 2.5). Bu dağılımların ortalama değeri eşit, ancak standart sapmaları farklıdır.



Şekil 2.5: Farklı basıklığa örnek dağılım çeşitleri. Standart sapmaları aynı olan bu dağılımların e < 0 olanlarına **platikurtik**, e > 0 olanlarına da **leptokurtik** denir.

Logaritmik dağılımların incelenmesi, metrik değerlerin logaritması alınarak aynı yöntemlerle gerçekleştirilir.

Örnek 2.6

Sıklık dağılım değişkenlerinin hesaplanması ve dağılım grafiğinin çizilmesi (Çizelge 2.1, n = 17, ısı ölçümleri, ˚C).

1. *Değişkenlerin bulunması*

Aritmetik ortalama $\bar{x} = \sum_{i=1}^{17} x_i \,/\, n$

$= 106/17$

$= \underline{6,235}$ ˚C

Standart sapma $\mathbf{s} = \sqrt{\dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$

$= \sqrt{\dfrac{8,555}{17-1}}$

$= \sqrt{0,53}$

$= \underline{0,73}$ ˚C

Buna göre aritmetik ortalama değeri $y = \bar{x} \pm 0,73 =$ aralığını kapsıyor veya y= $\bar{x} \pm 0,73$ aralığı için geçerlidir.

Değişke $\mathbf{s^2} = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$

$= \dfrac{8,555}{17-1}$

$= \underline{0,53}$ ˚C ^{2}

Kayma $\mathbf{g} = \dfrac{m_3}{s^3} = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{n.s^3}$

$= \dfrac{1,626}{17.0,73^3}$

$= \underline{0,25}$ ˚C

Dağılım pozitif eğimlidir.

Basıklık $\mathbf{e} = \frac{m_4}{s^4} - 3$

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{n.s^4} - 3$$

$$= \frac{10{,}86}{4{,}82} - 3$$

$= 2{,}57-3$

$= \underline{-0{,}43}$ ˚C

bulunur. Buna göre pozitif eğimli (sağa kayma), standart çan eğrisinden basık bir eğri veya dağılım bulunmaktadır.

Değişkenlik katsayısı $\mathbf{v} = s.100 / \bar{x}$

$= 0{,}73.100/6{,}235$

$= \%\ \underline{11{,}7}$

bulunur. Bu oranla veriler çok düzenli bir dağılıma sahiptir.

2. Dağılım grafiğinin çizimi:

1. Adım: Değişim aralığının bulunması:

Değişim aralığı $R = x_{max}-x_{min}$

$= 7{,}7- 4{,}9$

$= \underline{2{,}8}$ ˚C

2. Adım: sınıf sayısının hesaplanması (Sturges Kuralı, bak. dip not s. 47):

$k = 1+3{,}32 \log n$
$= 1+3{,}32.1{,}23$
$= 1+4{,}08$
$= 5{,}1$
$= \underline{5}$

sınıf veya grup bulunur.

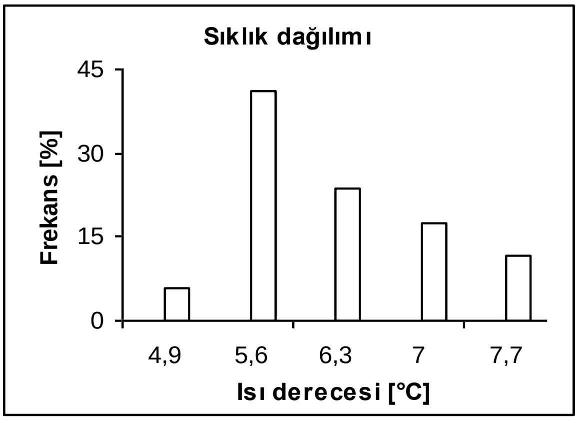
3. Adım: Frekans veya sınıf sınır değerlerinin bulunması: Frekans f,

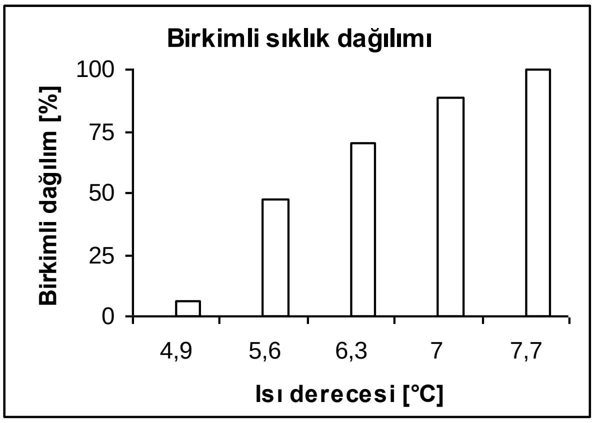
$$\begin{aligned} f &= R/k \\ &= 2{,}8/5 \\ &= 0{,}56 \\ &\approx \underline{0{,}6}\ ^\circ C \end{aligned}$$

elde edilir.

4. Adım: Sıklık dağılım çizelgesi:

Frekans [˚C] f	**Mutlak sıklık** f_i	**Göreceli sıklık [%]** h_i	**Birikimli sıklık [%]** $\Sigma\, h_i$
4,90 – 5,49	1	5,88	5,88
5,50 – 6,09	7	41,18	47,06
6,10 – 6,69	4	23,53	70,59
6,70 – 7,29	3	17,65	88,24
7,30 – 7,89	2	11,76	**100,00**
Σ =	**17**	**100,00**	





* Sıklık dağılımlarında sınıf sayısının saptanması için bir genel kural bulunmamaktadır. Ancak bazı ilkeler önerilmektedir: Örneğin sınıf sayısının örnek sayısına göre değişimi gibi. Ayrıca bir sıklık dağılımındaki sınıf sayısı örnek sayısının kare kökünden az olması istenir. Bir dağılımdaki sınıf aralıkları eşit olmayabilir, örneğin büyüyebilir veya küçülebilir. Sınıf sayısının bulunmasında yararlanılan en yaygın kural Sturges Kuralı'dır. Buna göre bir sıklık dağılımında en iyi sınıf sayısı $k = 1+\log 2^{-1}.\log n = 1+3{,}32.\log n$ şeklindeki formüldür. Buradan da frekans (sınıf aralığı) f hesaplanır ($f = R/k = x_{max} - x_{min} / (1+3{,}32 \log n)$).

3. TEORİK DAĞILIMLAR

3.1 Giriş

İlk 2 konuda işlenen örnek tanımlamaları sonuçları bakımından belirsizlikler taşımaktadır. Çünkü bu örneklemler sonlu bir veri dizisini kapsıyor. Bu nedenle incelenen olaylar (işlev veya mekanizma) sadece kısmen işlemlere tabi tutulabilmektedir. Dolayısı ile deneysel sıklık dağılımlarının görünümleri örnek kapsamının genişlemesi ile değişir.

Sözkonusu olay genel olarak, yani örneklem rastlantılarının etkisi dışında, istatistiksel olarak incelenmek istenirse, örneklemin geldiği popülasyonun özelliklerinin incelenmesi gerekir. Bunlar genelde bilinmedikleri için istatistikte çeşitli kuramsal popülasyon dağılımlarının bulunması için değişik yollar bulunarak deneysel dağılımlara uyarlanmıştır.

Uygulamada birçok kuramsal dağılım şekli bulunmaktadır. Burada bunların ancak önemli olanları üzerinde durulacaktır. Herhangi bir veriyi inceliyebilmek maksadiyle kuramsal dağılımların sadece standart şekli, yani **olasılık sıklık dağılımı** f(x) yardımıyla, tanımlana-caktır. Bir görsel dağı-lımın bir kuramsal dağılıma uyarlanması demek, görsel dağılıma en yakın kuramsal dağılımın bulunması demektir. Görsel dağılımın kuramsal dağılıma uyumu için çevirme işlemlerinin yapılması, uyum derecesinin (anlamlılık) saptanması ve sınanması şarttır.

İşlenecek kuramsal yöntemler ilk önce sadece tek boyutlu durumlar için uygulanacaktır. Bunların çok boyutlu durumlar için uygulanması sıkça sorunlar doğurur. Hesaplanacak değişkenlerin örneklem ve ana kütle (popülasyon) ilişkisinin ayırdedilmesi için kullanılan değişik karakterler aşağıya çıkarılmıştır:

Değişken	Örneklem	Ana kütle
Ortalama değer	$\bar{x}$	μ
Ortanca	x_o	μ_o
Tepe değeri	x_t	μ_t
Standart sapma	s	σ
Değişke	s^2	σ^2
Kayma	g	γ
Basıklık	e	η
Kapsam	n	ν

3.2 Normal dağılım (ND)

Doğadaki dağılımların büyük çoğunluğu **normal dağılım** göstermektedir. Buradaki normal dağılımın anlamı sadece çok sayıdaki işleve uygulanabilirliğini ifade eder. Tüm dağılım çeşitleri oluştukları yerin mevcut koşullarına göre gelişirler. Örneğin, bakkal ve manavların bir şehrin her tarafına yaklaşık aynı sıklıkta dağılmalarına karşın, fotoğraf dükkanlarının sadece şehir merkezinde yoğunlaştıkları görülür. Bu gözlem sonucunun, yani hipotezin, genelleşmesi için açıklanması ve kuramsal bir toplu sistemde ya doğrulanması, ya da reddedilmesi gerekir.

Normal dağılımın en önemli özelliği, yinelenen verilerin ortalama değer etrafında yoğunlaşmasıdır. Uclara, yani büyük ve küçük değerlere, doğru veriler kendiliğinden sınıflanarak seyrelir. Örneğin, bir yerleşim birimindeki erkeklerin boy ortalaması 1,70 m ise, 1,68-1,64 m boylu erkeklerin sayısı 1,64-1,60 m boylulardan; 1,72-1,76 m boyluların sayısı da 1,76-1,80 m boylulardan daha çok olduğu aşağıdaki çizelgeden görülmektedir. Aynı şey çalışanların aylık geliri ve yumurta ağırlığı için de geçerlidir. Normal dağılımın doğadaki en iyi örnekleri volkan konileri lavlarının,; kumul ile kum saatleri kumlarının konik dağılımı; petrol kuyuları veriminin zamanla değişimi ve uygarlıkların yükseliş ve batışlarında da görülmektedir. Normal dağılımın nedeni, oluşum sırasında birbirini etkiliyen bağımsız birçok etkenin bir araya gelmesidir.

Sınıf	**Sıklık**
160,50-164,49	5
164,50-168,49	57
168,50-172,49	162
172,50-176,49	64
176,50-180,49	7

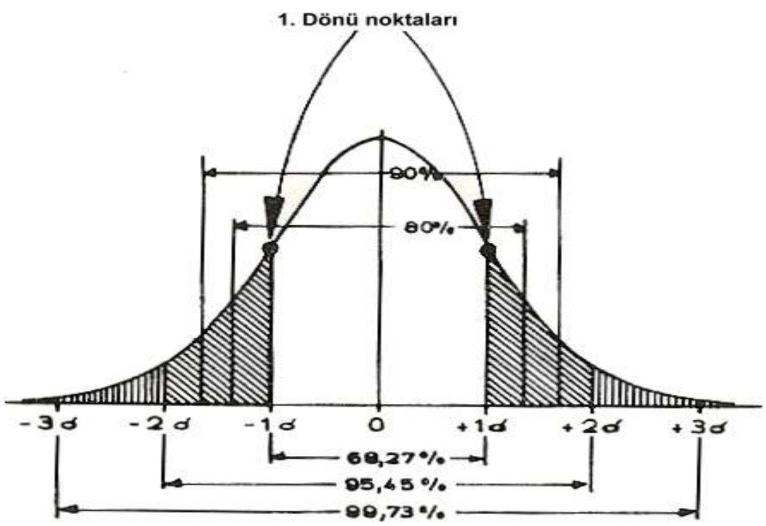
Böyle bir dağılımda örneklerin tek tek incelenmesi oldukça zordur. Dağılım üzerinde denetimi sağlamak amacıyla ve bilgi kaybına neden olmadan, yukarıda da anlatıldığı gibi, veriler veya ölçümler gruplara (sınıflara) ayrılır (bak.örnek. 2.6.). Bu sınıflardan bir dağılımın eğilimini gösteren sütun diyagramlar elde edilir. Elde edilen sütun dağılımına uyarlanan bir eğri dağılım fonksiyonunu meydana getirir (Şekil 2.1b ve 2.6). Bu dağılım fonksiyonu (probablity density fuction) sürekli ve simetrik ($\gamma = 0$) normal sıklık dağılımınını* gösterir ve

* Gauss dağılımı veya çana benzediği için çan eğrisi de denir. C. F. Gauss (1777-1855) sıklık dağılımı (distribution function) ilk formüle eden alman matematikçidir.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad 3.1$$

eşitliği ile tanımlanmaktadır ($-\infty<x<+\infty$; $-\infty<\mu<+\infty$; $\sigma>0$). Eşitlikten birbirine benziyen (maksimum μ, simetrik dönü noktaları $\mu-\sigma$ ve $\mu+\sigma$) çok sayıda sıklık dağılımı fonksiyonunun bulunduğu anlaşılmaktadır.

Normal dağılım ortalama değerlerin en sık ve en olası oldukları her yerde dağılım modeli olarak beklenebilir. Ortalama değerin solundaki ve sağındaki sapmalar eşit olasılığa sahiptir. Artan sapma değeri ile azalarak daha az olası duruma giderler. Bu özelliği ile normal dağılım istatistiğin temelini oluşturan bir öneme sahiptir. Hata hesaplamaları (bak. 4. konu), standart sapma gibi formüller ve yöntemler normal dağılımın esaslarına dayanmaktadır. Bu nedenle incelenen her dağılımın normal dağılımı gerektirip gerektirmediği araştırılır (**parametrik**, dağılıma bağlı veya **non parametrik**, dağılıma bağlı olmıyan, yöntemler). Bütün istatistiksel analizlerdeki ilk ana kural örneklemdeki normal dağılımın gerekliliğidir. Bu kural iyi ve güvenilir sonuçlar için kaçınılmaz bir zorunluluktur. Bir veri kümesindeki normal dağılımın varlığı ise, histogramlarda görülebilir. Bu durumda dağılım eğrisi bir çan şeklindedir.



Şekil 3.1. Standart sapma ile eğrinin alan içeriği arasındaki ilişki (Wellmer, 1989).

Bir normal dağılıma sahip f(x) eğrisinin dönü noktaları absise yansıtıldığında bu eksen üzerinde μ ortalama değerinden itibaren tam σ'ya karşılık gelen değerler elde edilir. Bir örneklemin standart sapması bu oluşuma bağlıdır (dağılıma bağlı formül, Şekil 3.1). Buna karşın yüzdelikler gibi değişkenler dağılıma bağlı değildir. f(x) fonksiyonunun dönü noktalarındaki teğetler x eksenini μ±2σ'da keserler.

Bir örneğin dağılım fonksiyonunun μ±σ aralığına (1. dönü noktaları) düşme olasılığı % 68,27, μ±2σ aralığına % 95,45 ve μ±3σ aralığına da % 99,73'tür. Bu oranlar aynı zamanda verilen sınırlar arasındaki alanın eğri altında kalan toplam alana oranına karşılık gelir (Şekil 3.1).

μ = 0 ve σ =1 durumu için f(x) dağılım fonksiyonu,

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \qquad 3.2$$

şeklini alır. Böylece fonksiyon değişkenlerden bağımsız hale gelir ve z = $\frac{x-\mu}{\sigma}$ transformasyonu ile

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \qquad 3.3$$

şekline dönüştürülmüş olur. Bu durumdaki bir fonksiyon **standardize normal dağılım** denir. Bu şekildeki fonksiyonlar (z dağılımı) birimsiz hale geldikleri için değişik yön-temlerle daha kolay hesaplanırlar. İstatistikteki sınama (test) ve tahmin için oldukça önemlidir. Yüzdeliklerin hesaplanması entegral gerektirdiğinden, zordur. Ancak bu, çizelgelerden yarar-lanarak, aşılabilir. Bazı kaynaklar z dağılımı çizelgelerini de vermektedir (bak. Schönwiese, 1992 ve David, 1977).

z'ye **standart değer*** denir. Bir dağılımda her X_i değerine karşılık bir x_i sapma değeri belirlendiği gibi, bir z_i değeri de belirlenebilir. Standart değer,

$$z_i = x_i/s \qquad 3.4$$

* Merkezileştirilmiş ve standart sapması s = 1 olan veriler **standardize** edilmiş demektir. Standardizasyon $z_i = \frac{x_i - \bar{x}}{s}$ (i = 1,...,n) gibi bir transformasyonla mümkündür. Standardize veriler birimlere bağlı değil. Dolayısı ile değişik birimli verilerin matematiksel işlenmesi mümkün olur.

eşitliği ile tanımlanır ve bununla sapma değerleri standart değerlere çevrilir. Örneğin, $x_i = X_i - \bar{x}$ değeri için $z_i = x_i/s$ standart değeri bulunur.

Örnek 3.1

Ortalama değeri $\bar{x} = 5$ ve s=2 olan bir dağılımdaki $X_i = 7$ ölçümü hangi standart değere karşılık gelir?

Çözüm: $z_i = x_i / s$
$= (X_i - \bar{x}) / 2$
$= (7 - 5) / 2$
$= 2 / 2$
$= \underline{+1}$ bulunur (1. dönü noktası =1standart sapma).

Burada $x_i = \bar{x} + s$ değerine karşılık geldiği için 1 değeri, **tolerans ölçüsü,** elde edilmiştir. Yani ortalama değerden bir standart sapma kadar büyük değerlerin standart değeri 1'dir. Bunun gibi 2s kadar büyük olan değerlerin standart değerleri 2, 3s kadar büyük olanların da 3'tür. Aynı sonuçlar standart sapmadan küçük -1, -2 ve -3 değerleri için de geçerlidir. **Ortalama değere eşit ($x_i = \bar{x}$) bir değerin standart değeri ise, 0'dır.** Örneğin,

$$\begin{aligned} z_i &= x_i/s \\ &= [x_i - \bar{x}]/s \\ &= [\bar{x} - \bar{x}]/s \\ &= 0/s \\ &= 0 \end{aligned}$$

Bu nedenle standart çan eğrisi 0'a (y ekseni) göre simetrik verilir (bak. Şekil 3.2).

Görüldüğü gibi standart çan eğrisinin değişkenleri normal çan eğrisinin değişkenlerinden farklı olmaktadır. Standardize edilmemiş çan eğrisinin dönü noktaları $\mu\pm\sigma$, $\mu\pm2\sigma$ ve $\mu\pm3\sigma$ iken, standart çan eğrisinde ±1, ±2 ve ±3 olmaktadır. Ortalama değerin standart değeri 0'dır. Bu nedenle standart çan ağrisi ölçümler yerine bu değerlerle tanımlanır ve bu sayede normal çan eğrisinden ayırdedilir.

Sıklık dağılımı fonksiyonu,

$$F(x \le \chi) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} dy \qquad 3.5$$

şekliyle doğrusal ve entegrale benziyen dağılım toplamını verir (birikimli toplam, Şekil 3.2). Bu dağılım **olasılık kağıdında** özel bir dağılımla bir doğruya dönüştürülür ve hesaplamaların denetiminde kullanılır. Olasılık kağıdının absisi doğrusal, ordinatı ise, 3.5 eşitliğine göre düzenlenmiştir. Bu yaklaşık doğrunun başka özelliklerinden biri de F(-1) = 0,84 ve F(+1) = 0,16 olmasıdır. Buradan 0,84-0,16 = 0,68 elde edilir. Bu, eğrinin altındaki alanın μ-σ ve μ+σ sınırları arasında kalan kısmıdır.

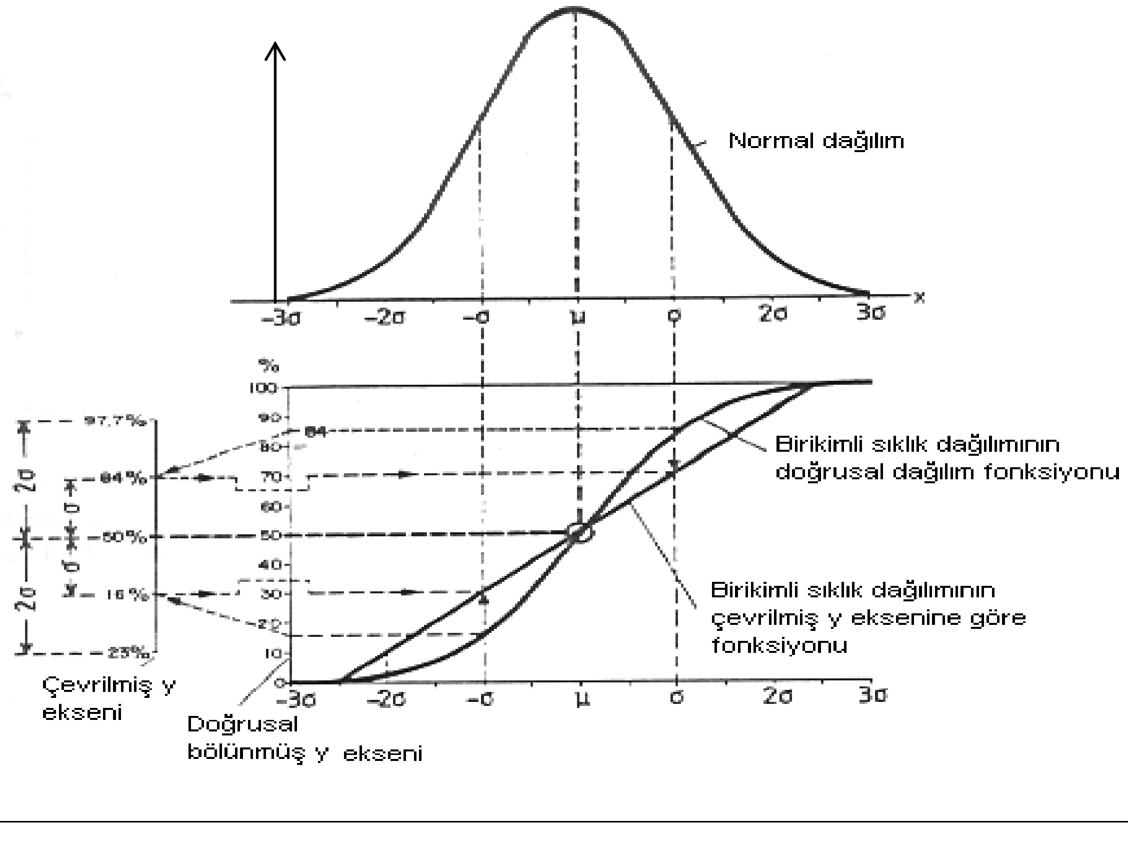
Yukarıda da belirtildiği gibi normal dağılım simetrik olduğundan, birçok özel duruma sahiptir. Örneğin,

Ortalama değer μ = ortanca = tepe değeri (1. moment).

Değişke σ^2 (2. moment),

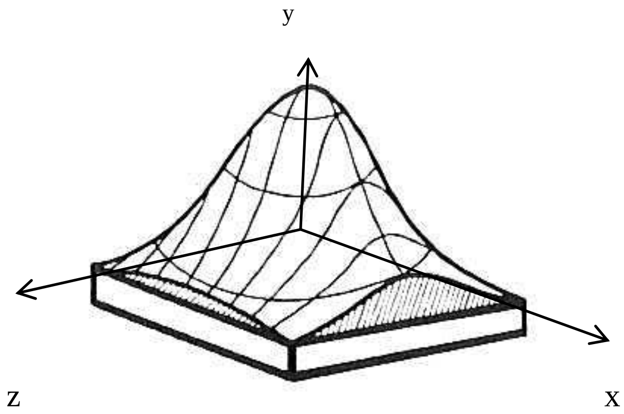
Kayma g = 0 (3. moment) ve

Basıklık e =3'tür (4. moment) gibi.



Şekil 3.2. Gauss çan eğrisinin bazı değişkenleri ve olasılık kağıdına dönüştürülmesi (Wellmer, 1989). Dağılım doğrusuna "Hazan Doğrusu" denir.

μ değişkeni olasılık sıklık dağılımında değerlerin absis üzerinde sağa veya sola kayarak kaymayı (g, kayma), σ da ordinat boyunca ortalama değer μ etrafında yoğunlaşıp seyrelerek basıklığı (e, sivri/yassılık) meydana getirir.



Şekil 3.3. İdeal simetrik, 2 boyutlu dağılım yüzeyi perspektifi. (Sachs, 1984, değiştirilmiştir).

Alıştırma 3.1

{12-14-36-22-37-29-30-38-41-49-36-45-62-53-58-67-69-56-55-50} analiz sonuçlarının (n=20),

a) Aritmetik (42,95), geometrik ve harmonik ortalamalarını,
b) Standart sapmasını (16,40), değişke, kayma ve basıklığını,
c) Tepe değerini (36) ve ortancasını (40) hesaplayınız,
d) Sıklık dağılımını çiziniz ve
e) Birikimli dağılımını olasılık kağıdında değişkenleri ile gösteriniz (parantezde yanıtlar).

3.3 Birikimli normal dağılım

Bir dağılımın kesin sağlanması ancak matematiksel sınamalarla (testlerle), örneğin χ^2 (ki kare, bak. 4.3. veya Kolmogoroff-Smirnow) sınaması ile, sağlanabilir. Burada bu konulara geniş yer ayırmak bu notların kapsamını aşacağından, uygulamadaki yaklaşık grafik denetimi ile yetinilecektir.

Grafikle çözüm piyasada satılan **olasılık kağıdı** ile yapılır (Şekil 3.2). Olasılık ağı veya kağıdı çan eğrisinin veya bir normal dağılımın bir doğru şeklini alacağı bir x-y diyagramıdır. y ekseni buna göre bölünmüştür. Ancak x ekseni doğrusaldır. Gerektiğinde logaritmik de yapılabilir. Bu ağla normal dağılımların doğruluğu sağlanır. Bu nedenle birikimli sıklık dağılımı ölçüm sonuçlarının düzenlenmesinden sonra hesaplanır. Bunun için bu ağın absisine özellik sınıfı üst sınırları x_i, ordinatına da bunların birikimli sıklığı Σh_i kaydedilir (Örnek 2.6 ve 3.2). Sınıfların göreceli değerleri, en küçük değerden başlamak üzere, her sınıf için ayrı ayrı toplanır. Buna göre birikimli sıklık, özellik sınıflarının üst sınır değerlerinden küçük veya bunlara eşit değerlerin toplamıdır. Bunun dağılım fonksiyonuna **birikimli sıklık dağılım fonksiyonu** denir (bak. eşitlik 3.5). Bu dağılımın fonksiyonu normal dağılım halinde bir doğru verir (Hazan Doğrusu). Aksi durumlarda fonksiyon kırık çizgi şeklinde görülür. Doğrunun çizimi sırasında başlangıç ve bitiş kısımları büyük önem taşımaz. Ancak % 50 civarının bir doğru çıkması belirleyicidir. Grafikten ayrıca şu değerler okunabilir:

a) % 50 değerinden absise çizilen paralelin doğruyu kestiği noktadan absise inilen dik, doğrunun absisi kestiği nokta ortancayı veya aritmetik ortalamayı,

b) % 16 ve % 84 noktalarından aynı şekilde absise çizilen paralellerin doğruyu kestiği noktadan absise inilen dikler de absis üzerinde $\mu \pm \sigma$ noktalarını, yani standart sapmayı veya **tolerans ölçüsü** 1'yi, verir (Şekil 3.2). $\bar{x}$ sadece bu aralık için geçerlidir. Bu iki sınır arasında kalan alan tüm analiz değerlerinin % 68,26'sına karşılık gelir.

Böylece elde edilen bir birikimli sıklık dağılım doğrusuna ortanca noktasından bakıldığında, standart sapmanın, σ veya s'nin, birkaç kat dilimlerinden oluştuğu görülür. Doğrunun eğimi artan standart sapma ile azalır ve uc kısımlar çoğu kez uyumda sorun yaratır. Ancak bu araçla gerçek dağılımların grafiklerle karşılaştırılması mümkün olmakta ve önemli değişkenleri doğrudan okunabilmektedir.

Dağılımın doğru vermemesi durumunda heterojen bir örneklemin, başka dağılımların mevcut olabileceği veya değerlerin belli ölçüde kaydırılması gerektiği üzerinde durmak gerekir. Böyle bir durumda genelde logaritmik dağılım söz konusu olur.

Birikimli dağılım doğrusunun şeklinden örneklem hakkında önemli bilgiler edinilebilir. Örneğin, birkaç parçadan oluşan bir kırık **hazan doğrusu**, saf olmıyan, birkaç popülasyona ait bir örnekleme işarettir. Bu durumda her popülasyonun ayrı ayrı çözülmesi ve incelenmesi gerekir. Bu da ancak büyük örnek sayısı için bir anlam taşır. Bir örneklemde birden fazla ve açıkça farklı dağılımların bulunması, bu değerlerin bir ana kütleye ait olmadığını ve bunun için de bir ortalama değerin hesaplanamıyacağı anlamına gelir.

Ordinatın doğrusal bölünmesi durumunda fonksiyon, entegral işareti şeklinde **bir eğri** verir (Şekil 3.2 ve 3.4b). Bu eğrinin kıvrım derecesi standart sapmanın büyüklüğüne bağlıdır. Doğru, ortanca'ya (% 50) göre simetriktir. 0 ve % 100 noktaları, y eksenine asimtot olarak yaklaştıklarından, sonsuza yönelirler ($\pm \infty$). Uygulamada bu uc değerler önemli görülmez ve çok ender durumlarda, örneğin, uc değerlerin elimine dilmesi durumunda, yararlanılır.

3.4 Logaritmik normal dağılım (logND)

Asimetrik bir normal dağılım, özellikle örneklemin bir pozitif eğime sahip olması durumunda, logaritmik normal dağılım şeklinde görülür. Logaritmik dağılıma **Galton dağılımı** da denir. Bu dağılım, özellikle geniş yayılım aralıklarına sahip, birkaç on veya yüz kat değişen, dağılımlara uygulanır. Ancak bu yöntemle büyük ve küçük değerlerin göreceli değişimleri aynı derecede açıkça gösterilebilir.

Bir normal dağılımın basıklık (e<0, platikurtik) ve sivrilik (e>0, leptokurtik) özellikleri bazen örnek sayısının arttırılması ile giderilebilmesine karşın, kayma etkilenmemektedir. Bu, basıklığın sıkça rastlantısal bir örneklem etkeni olduğunu, buna karşın kaymanın anlamlı bir etken olduğunu gösterir. Bu durumlar uygulamada ancak logaritmik dağılım modelleri ile açıklanabilmektedir. Logaritmik normal **olasılık** ve **yoğunluk fonksiyonları**,

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\ln x-\mu}{\sigma}\right)^2} \tag{3.6}$$

$$F(x)= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \frac{1}{y} e^{-\frac{1}{2}\left(\frac{\ln y-\mu)}{\sigma}\right)^2} dy \tag{3.7}$$

şeklindedir (x>0). Logaritmik dağılımın **en belirgin işareti**, sınıflara ayrılmış birikimli bir örneklemin yüzde değerlerinin **olasılık kağıdında** ortalama değerin sağındaki değerlerin solundakilere oranla daha **yatay** bir durum göstermesidir. Absisleri logaritmik bölünmüş olasılık kağıtları piyasada satılmaktadır. Buradaki logaritmik dağılımın, doğal logaritma (ln) olduğu unutulmamalıdır. Bazen 10 tabanına göre logaritma, doğal logaritmaya göre daha iyi yaklaşık değerler verebilir. Bu durumda yukarıdaki denklemlerde sadece ln yerine log'un yazılması yeterlidir. Buna ek olarak μ ve σ yerine z yazılarak, yukarıda olduğu gibi, denklemler genelleştirilebilir. Logaritmik dağılım değişkenlerinin hesaplanması daha karışıktır* (Wellmer, 1989).

Uygulamada bir logaritmik dağılımın beklenmesi durumunda ölçüm veya analiz değerlerinin logaritmaları alınır. Çevrilen bu değerlerden ancak ortalama ve standart sapma değerleri hesaplanır, ilgili çizelgeler, örneğin z dağılım çizelgesi, kullanır ve karşılaştırmalar gerçekleştirilir. Logaritmik normal dağılımın standart sapması temel değerlerin geometrik ortalamaya oranıdır, normal dağılımdaki gibi sapmaların ortalama değeri değildir.

Değişken	**Formül**	
Ortalama değer	$\mu_L = e^{\mu+\frac{\sigma}{2}}$	3.10
Ortanca	$\mu_{oL} = e^{\mu}$	3.11
Tepe değeri	$\mu_{tL} = e^{\mu-\sigma^2}$	3.12
Değişke	$\sigma^2 = e^{(2\mu+\sigma^2)(e^{\sigma^2}-1)}$	3.13
Kayma	$\gamma = (e^{\sigma^2}+2)\sqrt{e^{\sigma^2}+1}$	3.14
Basıklık	$\eta = 0$	3.15

Logaritmik dağılımın önemi bilim dallarına göre değişmektedir. Örneğin, yerbilimlerinde çok önemlidir. Bu nedenle Ahrens (1954) ve Rodionov'a (1964) magmatik kayaçlardaki logaritmik normal dağılımı jeokimyanın temel yasası olarak tanımlamaktadırlar. Yukarıda da değinildiği gibi, düşük veri normal dağılımları ağırlıkta olmaktadır. Jeolojide, örneğin, hafif şiddetteki depremler, kuvvetlilerden daha sık meydana gelmektedir. Rüzgar hızı ve yağış dağılımı da buna benzer ve genellikle logaritmik dağılım sunarlar. Aslında bu dağılımlar derişime de bağlıdır. Örneğin, organların bileşen dağılımı, vücuttaki bileşen dağılımını da belirlemektedir.

Örnek 3.2

İçme suyunda normal ve logaritmik normal (log normal) demir (Fe) dağılımı.

* log a.ln 10 = ln a (ln 10 = 2,3026); ln a.log e = log a (log e = 0,4343) veya $\log_{10}x = \log_e x/\log_e 10$ = lnx/ln10 = lnx/2,3026'dır.

Veriler

Derişim: ppm [g/t] Fe : {11-13-14-2-15-17-50-35-6-24-12-25-18-3-7-40-9-4-5-4-1}

log ppm Fe: {1,04-1,08-1,15-0,30-1,15-1,23-1,70-1,54-0,78-1,38-1,08-1,40-1,26-0,48-0,78-1,60-0,95-0,60-0,70-0,60-0,00}

Çözüm*:* Değerlerin dağılım sınırları R= x_{max}-x_{min}

= 50-1

= <u>49</u> ppm Fe'den

sınıf aralığı, f =R /i , (i=1+1/log2.log n)

=R/1+3,322.log n

=49/(1+3,322 log 21)

=49/(1+3,322.1,32)

= 49/5,39

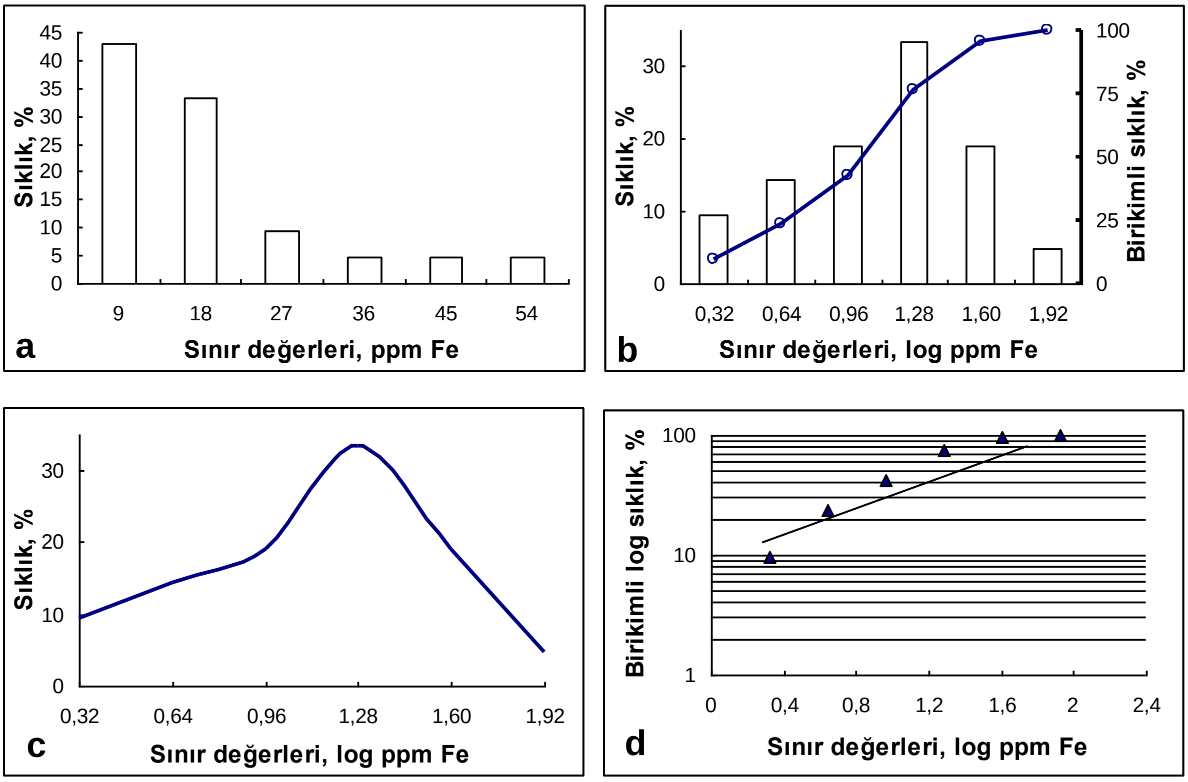
=9,09 ppm

≅<u>9 ppm</u>

bulunur (bak. aşağıdaki çizelge ve şekil). Aynı şekilde $f_{log\ Fe}$ = 0,32 elde edilir.

Normal dağılım				**Logaritmik dağılım**			
Frekans [%]	Sıklık		Birikimli sık. [%]	Frekans [log %]	Sıklık		Birikimli sık. [%]
f	**f_i**	**hi**	**∑hi**	**f**	**fi**	**hi**	**∑hi**
<9	9	0,43	42,86	<0,32	2	9,52	9,52
9-18	7	0,33	76,19	0,32-0,64	3	14,29	23,61
18-27	2	0,09	85,72	0,64-0,96	4	19,05	42,85
27-36	1	0,05	90,48	0,96-1,28	7	33,33	76,19
36-45	1	0,05	95,24	1,28-1,60	4	19,05	95,23
>45	1	0,05	100,00	>1,60	1	4,76	100,00
Toplam	**21**	**1,00**	**100,00**		**21**	**100,00**	**100,00**

Sayısal dağılım kuvvetli pozitif eğimliyle sağa kaymakta ve bir düşük değerli veri tipini göstermektedir (Şekil 3.2a). Aynı değerlerin birikimli log dağılımları entegral ve doğru şekilleri ile normal dağılımı pekiştirmektedir (Şekil 3.2b ve d). Birbirlerine çok yakın tepe değeri (1,12), ortanca 1,08 ve ortalama değeri (1,00) de bu savı desteklemektedir. Yassılık (kurtosis) e<3 (1,34) olduğundan, dağılım normal dağılımdan daha yassıdır. Buna göre Fe, birkaç kaynaktan gelmektedir (kayaç, sanayi v. s.).



Örnek 3.2 değerlerinin sıklık dağılım şekilleri: **a,** Göreceli sıklık (x_t=4, x_o=12 ve $\bar{x}$=15 ppm Fe'dir), **b**, logaritmik ve birikimli logaritmik sıklık dağılımı (bak. d), **c,** Sıklık dağılımının (b) çizgisel görünümü ve **d**, log birikimli dağılım.

3.5 Binom dağılımı (BD)

Neden doğal dağılımların normal veya logaritmik normal dağılım gösterdiklerine veya bu dağılım şekillerine uyduklarına ilişkin herhangi bir bilimsel esas bulunmamaktadır. Ancak deneyimler bu dağılımlarla verilerin incelenmesi yararlı sonuçlar verdiği görülmüştür. Bunların dışındaki dağılım modelleri de, örneğin, binom* veya Bernuli dağılımı da önemli rol oynarlar. Dolayısı ile burada kuramsal esasları incelenmiyecektir. Binom ve Poisson dağılımları süreksiz, normal, logaritmik, student-t, F ve ki kare dağılımları ise, sürekli dağılımlardır.

* Latince *bi*, iki; *nom*, isim demektir.

Binom dağılımı, bir olayın olasılığını n bağımsız rastlantısal denemede kesinlikle k=x kadar olası-lıkla gerçekleşeceğini tanımlar. Yani bir zarla 6 atma şansı her zaman için 1/6'dır. Daha önce 6'nın atılıp atılmaması hiç önemli değil. Binom dağılımında **başarı**,

$$p = 1/6,$$

ve **başarısızlık,** $q = 5/6$

dikkate alınır. Buradan, $p+q = 1/6+5/6$

$$=1$$

elde edilir.

Genel olarak binom dağılımı,

$$P_B = \binom{n}{x}.p^x.q^{n-x} \qquad 3.8$$

şeklinde formüle edilmektedir (Wellmer, 1989). Burada,

x, başarı (gerçekleşme) olasılığı; n, deneme sayısıdır (olası olasılık). x ve n'in kombinasyonu,

$$\binom{n}{x} = \frac{n(n-1)(n-2)\ldots(n-x+1)}{1.2.3\ldots x} \qquad 3.9$$

$$= \frac{n!}{x!\,(n-x)!} \qquad 3.10$$

demektir. Bu son terime göre binom katsayıları hesaplanır. Dağılım n ve p parametrelerine sahiptir. q kendi başına bir parametre değildir (q=1-p).

Bunun birikimli dağılım fonksiyonu,

$$F_{n,\,p}(x \leq \chi) = \sum_{k=1}^{|x|} \binom{n}{k} p^k.q^{n-k} \qquad 3.11$$

şeklindedir.

Görüldüğü gibi binom dağılımında veriler başarı ve başarısızlık olarak 2 grup halinde ele alınmak-tadır. Örneğin, bir malın veya hizmetin kalitesi iyi (başarı) ve kötü (başarısızlık) olarak incelenebilir. Doğada bazı dağılımlar bu modele uymaktadır. Örnek olarak jeolojiden

lateritik boksitlerle lateritik olmayan boksitler (Al yatakları) verilebilir. Lateritik boksitlerde Hf/Zr oranı binomial dağılım modeline uymaktadır. Dağılım tipinden boksitin lateritik olup olmadığı anlaşılabilir (Schroll, 1976). Ancak veriler birbirinden bağımsız olacaklar.

Binom dağılımının değişkenlerinin hesaplanmasında kullanılan eşitlikler aşağıda verilmiştir (Schönwiese, 1992):

Ortalama değer	$\mu = np$	3.12
Ortanca	$\mu_o = np$, eğer np tam sayı ise, yoksa ondan sonraki sayı	3.13
Tepe değeri	$\mu_t = (n+1)p$ ve $(n+1)q$	3.14
Değişke	$\sigma^2 = npq$	3.15
Kayma	$\gamma = (q-p)/\sigma = (q-p)/\sqrt{npq}$	3.16
Basıklık	$\eta = (1-6pq)/\sigma^2 = (1-6pq)/npq$	3.17

Binom dağılımı daha çok rastlantısal olayların kombiasyonunda kullanılır. Sınama ve kuramsal uygulanması enderdir. Yerbilimlerinde düzensiz verilerin incelenmesinde ve uygun örnek büyüklüğünün tane boyuna göre seçilmesinde yararlanılır. p=0,5 için simetrik bir şekil alan binom dağılımı, normal dağılıma çok iyi yaklaşır. Bu durumda dağılım eşitliği,

$$P_{B\ (p=0,5)} = \binom{n}{x} / 2^n \qquad 3.18$$

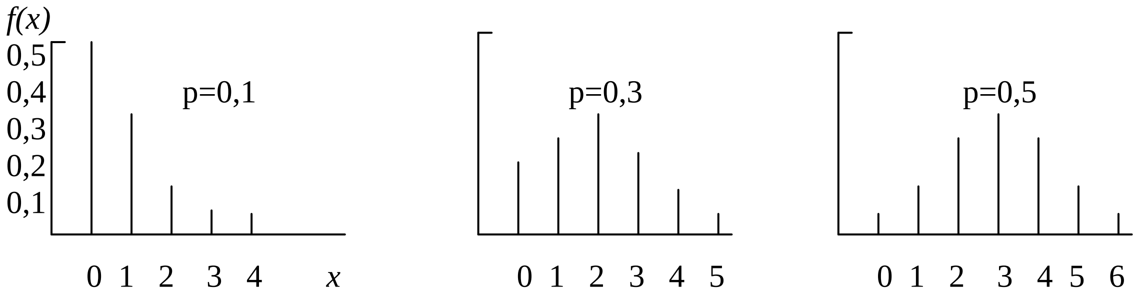
normlanmış olur (ayrıntı için bak. örneğin, Arıcı, 2001, Schönwiese, 1992 ve Wellmer, 1989). n ve p değişkenlerine bağlı olarak binom dağılımının nasıl değiştiğini Şekil 3.4 göstermektedir. Örneğin,

$$x=0,$$

$$P_{B\ n,p}(x)=P_{B\ 6;\ 0,5}$$

$$=\binom{6}{0}.0,5^0.0,5^{6-0}$$

$$= 0,0156\text{'dır.}$$



Şekil 3.4. p ve n değişkenlerine göre binom dağılımının değişimi. p=0,5 için dağılım normale yaklaşır.

Örnek 3.3

Bir yerdeki depremin meydana gelme yıllık olasılığı % 10’dur. Bunun 10 yılda,

a) 3 kez gerçekleşme olasılığı ve

b) En çok 3 kez gerçekleşme olasılığı nedir?

Çözüm. a) x = 3,

n = 10 ve

p = % 10 = 0,10’dur. Buna göre,

$$P_B = \binom{n}{x}.p^x.q^{n-x}$$

$$= \binom{10}{3}\ 0{,}10^3.0{,}9^{10-3}$$

$$= \frac{10!}{3!(10-3)!}.0{,}10^3.0{,}90^7$$

$$= \frac{3.628.800}{6.5040}.0{,}001.0{,}4783$$

$$= \frac{3.628.800}{30.240}.47{,}83.10^{-5}$$

$$= 120.47{,}83.10^{-5}$$

$$= 5739{,}6.10^{-5}$$

$$= 0{,}057$$

$$\approx \underline{\%\ 6}$$

bulunur.

b) k=0, 1, 2 ve 3;
   n=10 ve
   p=0,10'dur.

Buna göre,

$$P_B = \sum_{k=1}^{|x|} \binom{n}{k} p^k . q^{n-k}$$

$$= \binom{10}{0} 0,10^0 \ . \ 0,9^{10} + \binom{10}{1} 0,1^1 \ . \ 0,9^9 + \binom{10}{2} 0,1^2 \ . \ 0,9^8 + \binom{10}{3} 0,1^3 \ . \ 0,9^7$$

$$= 0,349 + 0,387 + 0,194 + 0,057$$

$$= 0,987$$

$$\approx \underline{\%\ 99}$$

çıkar. Bu, % 99 olasılıkla gelecek 10 yılda en çok 3 deprem olabilir demektir veya hiç olmaz.

Alıştırma 3.2

Bir tedavi yöntemi hastalarda % 60 yan etkisi görülmiyen olumlu sonucu vermektedir. Buna karşın hastaların % 20'inde ağır yan etki bırakmaktadır. Buna göre 100 hastanın tedavisinde,

1. 50'den fazla hasta üzerinde başarı olasılığı nedir? (Y: % 97,29)
2. 20 hastada ağır yan etki bırakma olasılığını bulunuz. (Y: % 55,95)
3. 37'den az hastada görülmiyen yan etki olasılığı nedir? (Y: % 99,99)

3. 6 Poisson dağılımı (PD)

Poisson* dağılımı, binom dağılımının n = ∞ ve p = 0 durumunda ortaya çıkan bir dağılımdır. n=100 ve p=0,05'ten itibaren, yani gerçekleşme olsılığı oldukça düşük olaylarda, binom dağılımı yerine poisson dağılımı kullanılır. Çünkü binom dağılımında n>10 ve küçük p değerleri için hesaplamalar oldukça zorlaşır. Poisson dağılımından bilimsel araştırma, planlama

*S. D. Poisson (1781-1840), fransız matematikçisine atfen poisson [puason] dağılımı.

ve bazı ender dağılım ile elektrik alanında impuls incelenmesinde yaygın bir şekilde yararlanılmaktadır. Yer bilimlerinde bu tür dağılımlara, deprem, taşkınlık, fırtına ve volkan püskürmesi gibi, sık sık rastlanır. Poisson dağılımının en önemli üstünlüğü olasılık sıklık dağılım ve dağılım fonksiyonlarının basit olmasıdır. Bunlar,

$$f(x) = \frac{e^{-\lambda}.\lambda^{x}}{x!} \quad (x>0) \qquad 3.19$$

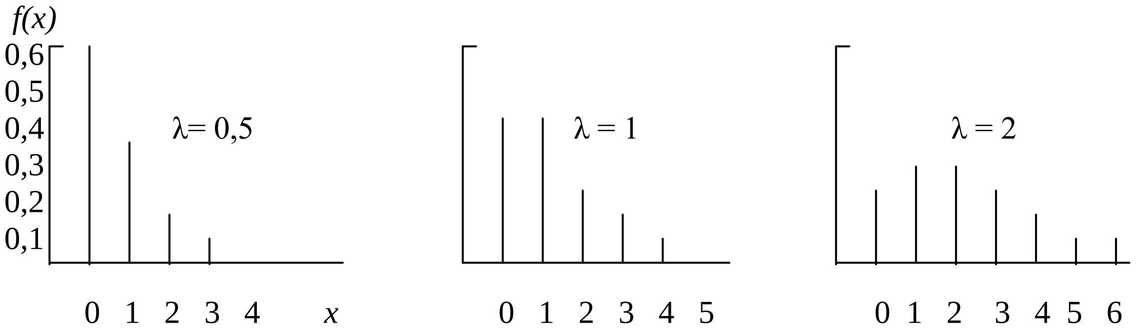
ve

$$F(x \leq \chi) = \sum_{k=1}^{|x|} \frac{e^{-\lambda}.\lambda^{k}}{k!} \qquad 3.20$$

eşitlikleri ile ifade edilirler. Bu dağılımda hem ortalama değer, hem de değişke olan λ, tek para-metredir. Poisson dağılımı değişkenlerinin hesaplanmasında kullanılan eşitlikler şunlardır:

Ortalama değer	$\mu = \lambda = np$ (λ: gerçekleşen, x: beklenen olasılıktır)	3.21
Ortanca	$\mu_o = \mu$, μ tam sayı ise; yoksa eğime göre, ondan sonraki tam sayı	3.22
Tepe değeri	$\mu_t = \mu$ ve $(\mu-1)$, μ tam sayı ise, yoksa $z< \mu$ koşulu ile en büyük sayı	3.23
Değişke	$\sigma^2 = \mu$	3.24
Kayma	$\gamma = 1/\sqrt{\mu}$	3.25
Basıklık	$\eta = 1/\mu$	3.26

Poisson dağılımı, f(x)/f(x+1) oranının x'e bağlı olarak bir doğru vermesi ile saptanır. Şekil 3.5'te λ = 0,5, 1 ve 2 için sıklık dağılımları verilmiştir. Büyük ortalama değerler için dağılımlar eşdeğer duruma gelir (normal dağılıma yaklaşır).



Şekil 3.5: Poisson dağılımında ortalama değer λ'ya göre dağılımın değişimi.

Örnek 3.4

Bir olayın, örneğin, sel baskınının, yıllık gerçekleşme olasılığı 0,03'tür (% 3). Böyle bir olayın gelecek 100 yılda 5 kez gerçekleşme olasılığı nedir?

Çözüm: p = 0,03 ve n = 100;

$$\lambda = np\text{'den}, = 0{,}03.100 = \underline{3}$$

bulunur (bak. 3.21). Buna göre,

$$P_p = \frac{e^{-\lambda}.\lambda^{x}}{x!} = \frac{e^{-3}.3^{5}}{5!} = \frac{3^{5}}{e^{3}.5!} = \frac{243}{2{,}718^{3}.(1.2.3.4.5)} = \frac{243}{20{,}08.120} = \frac{243}{2409{,}6} = 0{,}1008 = 0{,}101 \approx \underline{\%\ 10{,}1}$$

elde edilir.

Alıştırma 3.3

Mersin'deki bir sigorta şirketine sigortalanan minibüslerin salı günlerindeki ortalama kaza sayısı 5 olduğuna göre,

a) 7 minibüsün kaza yapma olasılığı % kaçtır (Y: %10,44)
b) Hiçbirinin kaza yapmama olasılığı nedir? (Y: % 0,7)

3.7 Diğer dağılım şekilleri

Yukarıda incelenen dağılımların yanında özel durumlar için, sınama ve kuramsal denetim amaçlı çok sayıda dağılım şekli bulunmaktadır. **Student-t**, **Fisher F**, **ki kare** (χ^2) ve **üstel dağılımlar** bunların bazılarıdır. Bunlar burada kuramsal dağılım akbul edildiklerinden, sık kullanılanları aşağıda **sınama yöntemlerinden** önce (bak. konu 4-6) kısaca tanıtılacaktır.

3.7.1 Student-t dağılımı (tD)

Normal dağılım gibi sürekli bir dağılımdır. Tahmin ve sınama kuramlarında önemli rol oynıyan ve örnek sayısının 1 eksiği (n-1) olan serbestlik derecesi (F) diye adlandırılan bir tek değişkene sahiptir. tD dağılımının eşitliği oldukça karışıktır. Ancak örnek sayısının artması ile, n>30'dan itibaren, normal dağılıma yaklaşması nedeniyle de çok yararlanılmaktadır (Şekil 3.6). Özellikle tez sınamalarında vazgeçilmez kuramsal üstünlükler sağlamaktadır. Student-t dağılımının olasılık sık-lık dağılımı,

$$f_{(t)} = \Gamma \frac{((sd+1)/2)}{\sqrt{\pi . sd}\,\Gamma\left(sd/2\right)} (1+\frac{t^2}{sd})^{-\frac{sd+1}{2}} \qquad 3.27$$

şeklinde formüle edilmektedir. Burada Γ (gamma),

$$\Gamma = \lim_{n=\infty} \frac{n!\,n^{x-1}}{x(x+1)(x+2)\ldots(x+n-1)} \qquad 3.28$$

alınır. Bunun dağılım fonksiyonu da,

$$F_{(t)} = \int_{-\infty}^{t} f_{td}(y)dy \qquad 3.29$$

şeklindedir. Bu dağılımın F = 4 için,

$\bar{x}$ = tepe

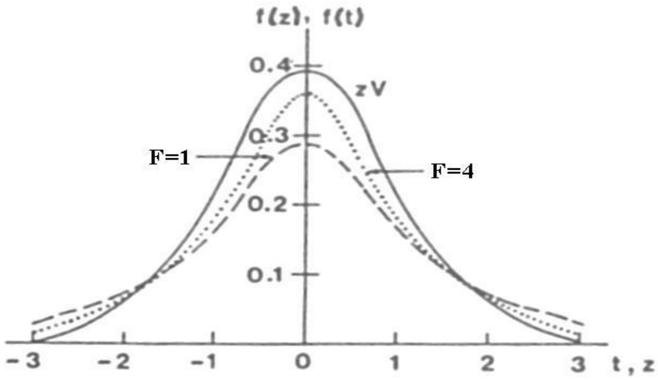
= ortanca

= 0,

σ^2 = F/F-2 ve

γ = 0'dır,.

Bu karışık eşitliği nedeniyle tD, öncelikle az sayıda örnekten oluşan normal dağılımların ince-lenmesinde, örneğin, tez sınamasında (ortalama değerlerinin karşılaştırılması) ve geçerlilik sınır-larının bulunmasında kullanılır (tD dağılımı ek 3.1'de verilmiştir).



Şekil 3.6. Serbestlik derecesi F'ye göre değişen student-t dağılımın olasılık yoğunluk fonksiyonu (z dağılımı kalın çizilmiştir).

Bir örneklemin ölçüleri olan ortalama değerler, tepe değeri, ortanca ve standart sapma ile değişke güven sınırları, yani değerin belli bir kesinlik derecesi ile **geçerli** olduğu **aralık**, **l** (= level; güven aralığı, bak. konu 4) arasında bulunan değerlerdir. Örneğin, az sayıdaki örneklerin ortalama değeri için güven sınırlarının bulunmasında tD dağılımından yararlanılır. Bunun için kullanılan genel eşitlik,

$$\bar{x} \pm \frac{t.s}{\sqrt{n}} \qquad 3.30$$

ile ifade edilir. Ortanca x_o, standart sapma s ve değişke s^2 için de, kesinlik derecelerine bağlı olarak, benzer eşitlikler kullanılmaktadır.

Örnek 3.5

Alıştırma 3.1'de ortalama değeri $\bar{x}$,

n=20,

$\bar{x}$ =42,95 ve

s=16,40'dır.

Buna göre % 95 olasılıkla güven sınırları kolaylıkla Çizelge 3.1'de F=19'a karşılık gelen t=2,093 yardımı ile hesaplanabilir: Buna göre güven aralığı l,

$$l = 42{,}95 \pm 16{,}40.\frac{2{,}093}{\sqrt{20}}$$

$$= 42,95 \pm 4,69$$

bulunur. Buradan l'nin,

$$42,95-4,69<l<42,95+4,69$$

$$38,26<l<47,64$$

güven aralığını kapsadığı görülür.

Araştırmalarda, iki örneklemin ayırdedilmesi, bunların ikisinin aynı ana kütleye ait olup olmadı-ğının araştırılması oldukça önemlidir. Aynı şekilde bir temel varsayımın deneyimlere dayanan A sonuçları ve kuramsal olasılık dağılımı B ile sınanması da büyük önem taşır. Bunun için istatistik çeşitli sınama (test) yöntemlerinin kullanımını hizmete sunmaktadır. **Student-t sınaması** da bu yöntemlerden biridir.

Student-t sınaması ile 2 ana kütle ortalamaları μ_1 ve μ_2 karşılaştırılması amacı ile yapılır. Bu amaçla örneklemlerin ortalama değerleri x_1 ve x_2'den yararlanılarak aynı ana kütleye ait olup olmadıkları sağlanır (H_o: $\bar{x}_1 = \bar{x}_2$ varsayımı savunulur, bak. ayrıca konu 6). Bunun için Çizelge 3.1'den yararlanılarak örneklerin eşit olması durumunda,

$$t = (\bar{x}_1 - \bar{x}_2)\sqrt{\frac{n-1}{s_1^2 + s_2^2}} \qquad 3.31$$

eşitliğinden (F=2n-2); eşit olmaması durumunda da,

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\hat{s}}\sqrt{\frac{n_1 . n_2}{n_1 + n_2}} \qquad 3.32$$

eşitliğinden t değerleri hesaplanır. Ancak burada önce **tahmini standart sapma** $\hat{s}$'nin hesaplanması lazım: $\hat{s}^2$,

$$\hat{s}^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2} \qquad 3.33$$

değerine eşittir. Kuramsal dağılımlarda $\hat{s}^2 = \sigma^2$'ye eşittir (n, örnek sayısını gösterir).

Örnek 3.6

İki komşu semtte elektrik kesintilerinin ortalama $\bar{x}$ süreleri,

Semt	Ölçüm sayısı n	$\bar{x}$ [h]	s	s^2
A	166	2,31	±0,95	0,90
B	224	1,63	±0,70	0,49

şeklindedir. Savlar,

H₀: Kesintiler aynı etkenden kaynaklanıyor,
H₁: Kesintiler farklı etkenden kaynaklanıyorlar.

demektedir.

Çözüm: Eşitlik 3.32 ve 3.33'ten,

$$\hat{s}^2 = \frac{165.0{,}90 + 223.0{,}49}{166 + 224 - 2} = 0{,}67\,,$$

$$\hat{s} = 0{,}82 \text{ ve}$$

$$t = \frac{2{,}31 - 1{,}63}{0{,}82}\sqrt{\frac{166.224}{166 + 224}} = 8{,}10$$

bulunur. Çizelge 3.1'den,

n>100 ve

P = % 99'dan

$$t_{99,\, n>100} = 2{,}60$$

okunur. t_{den} (8,1) > t_{kur} (2,60)* olduğundan, kesintiler (sapmalar) **aynı etkenden kaynaklanmıyor** (H_o : red). Olasılık P= % 99 alındığından, fark anlamlı olabilir. Bu nedenle $t_{95}<t<t_{99}$ için sapmalar "**muhtemelen anlamlıdır**" denir.

* t_{den} : t deneysel (hesaplanan değer); t_{kur} : t kuramsal (çizelge değeri) demektir.

Çizelge 3.1. Studenet-t dağılımı (sınaması) değerleri. P, olasılık [%], F, serbestlik derecesi (bak. ayrıca Ek 9.2).

F	P_{95}	P_{99}	F	P_{95}	P_{99}	F	P_{95}	P_{99}
1	12,71	63,66	13	2,16	3,01	30	2,04	2,75
2	4,30	9,92	14	2,14	2,98	40	2,02	2,70
3	3,18	5,84	15	2,13	2,95	50	2,01	2,67
4	2,78	4,60	16	2,12	2,92	60	2,00	2,66
5	2,57	4,03	17	2,11	2,90	70	2,00	2,65
6	2,45	3,71	18	2,10	2,88	80	1,99	2,64
7	2,36	3,50	19	2,09	2,86	90	1,99	2,64
8	2,31	3,36	20	2,09	2,84	100	1,99	2,63
9	2,26	3,25	22	2,07	2,82	150	1,98	2,61
10	2,23	3,17	24	2,06	2,80	200	1,97	2,60
11	2,20	3,11	26	2,06	2,78	300	1,97	2,60
12	2,18	3,06	28	2,05	2,76	400	1,96	2,58

Alıştırma 3.4

Aşağıdaki A ve B mıknatıs grupları süseptibilitelerinin [Am^2kg^{-1}] ortalama değerlerini karşılaştırarak student-t yöntemiyle farklı olup olmadıklarını gösteriniz (P = % 95; $\mathbf{H_o}$: $\bar{x}_1 = \bar{x}_2$, $\mathbf{H_1}$: $x_1 \neq x_2$).

A grubu: {87,4 93,4 96,8 86,1 96,4} n_A=5.

B grubu: {106,2 102,2 105,7 93,4 95,0 97,0}, n_B=6. (Y: t=2 < 5 ve farklı)

Bunlara ek olarak n gözlem sayısından elde edilen $\bar{x}_{den}$ değerleri $\bar{\bar{x}}_{kur}$ değeri ile karşılaştırılır.

Bunun için,

$$t = \frac{|\,\bar{\bar{x}} - \bar{x}_{1,2}\,|}{s}\sqrt{n-1} \qquad 3.34$$

formülü kullanılır ve aynı şekilde Çizelge 3.1'den yararlanarak değerlendirilir.

3.7.2 Fisher (F) dağılımı (FD)

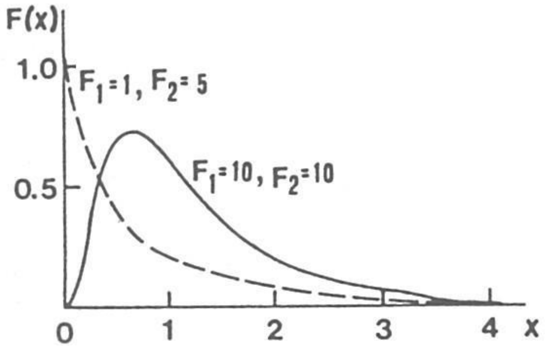
Yukarıda anlatılan yöntemlerin yanında F dağılımı da istatistikte tahmin ve sınama yöntemlerinde büyük önem taşır. Bu yöntem de student-t dağılımı gibi sürekli bir dağılımdır ve serbestlik dereceleri nedeniyle, n_1 ve n_2, 2 değişkene sahiptir. Şimdiye kadar incelenen yöntemlerin en karmaşığıdır. Bu dağılımın olasılık yoğunluk fonksiyonu,

$$f(x) = \frac{\Gamma\left[\frac{sd_1 + sd_2}{2}\right]\left[\frac{sd_1}{sd_2}\right]^{\frac{sd_1}{sd2}}}{\Gamma(sd_1/2)\Gamma(sd_2/2)} x^{\frac{sd_1-2}{2}}\left[1+\frac{sd_1}{sd_2}x\right]^{-\frac{sd_1+sd_2}{2}} \qquad 3.35$$

olarak tanımlanmaktadır (x>0 için). Gamma fonksiyonu tD'daki değeriyle aynıdır. Bunun dağılım fonksiyonu,

$$F(x)= \int_0^x f(y)_{FD}\,dy \qquad 3.36$$

ile ifade edilir. Bu dağılımın temel tanımına göre, χ^2 (ki kare) dağılımına göre dağılmış ve bağımsız u ve v rastlantısal değişkenleri (u/v).(F_2/F_1) şeklinde FD'na uyarlar. Yüzdeliklerin hesaplanmasında ve çeşitli sınamalarda yararlanılan F dağılımının ortalama değeri $\mu=F_2/(F_2-2)$'dir. Diğer değişkenleri serbestlik derecelerine bağlı ve daha karmaşıktır. F değişkeninin olasılık yoğunluk fonksiyonun grfikleri Şekil 3.7'de görülmektedir (F değerleri için bak. Çizelge 3.2).



Şekil 3.7. Serbestlik derecesi F kombinasyonlarına göre F dağılımı fonksiyonu.

Değişke veya standart sapmaların karşılaştırılması ile veri kümeleri arasındaki benzerlik veya ayrıcalıkların saptanmasında F dağılımından veya **F sınamasından** (F testi) yararlanılır. Bunun için dağılımların normal dağıldıkları ve aynı değişkelere sahip oldukları tezi savunulur. Bu da örnek değişkesi Q = büyük $\hat{s}^2$ değeri/ küçük $\hat{s}^2$ değeri ile sınanır. Bu oranın büyüklüğüne göre örneklemlerin farklı ana kütlelerden gelme olasılığı artar. Kesin bir sonuç ancak Çizelge 3.2'den F_1 ve F_2 değerlerine göre okunan F değerinin Q değeri ile karşılaştırılmasından elde edilir. Değişkeler arasındaki fark ancak hesaplanan Q_{den} değerinin çizelgeden okunan F_{kur} değerinden büyük çıkması halinde anlamlıdır (bak. Örnek 3.7).

Örnek 3.7

Örnek 3,6'daki verilerin değişkelerinin farklılıklarını % 95 olasılıkla karşılaştırılması:

$$\hat{s}_A^2 = 0{,}90,$$

$$\hat{s}_B^2 = 0{,}49 \text{ 'dan,}$$

$$Q_{den} = \frac{\dfrac{\hat{s}_A^2}{F_1}}{\dfrac{\hat{s}_B^2}{F_2}}$$

$$= \frac{\dfrac{0{,}90}{165}}{\dfrac{0{,}49}{223}}$$

$$= \frac{0{,}0055}{0{,}0022}$$

$$=2{,}50$$

bulunur. Çizelge 3.2'den, P=%95 olasılık derecesine ve

F_1=165,

F_2=223'e

göre F_{kur}, $F_{165;\ 223;\ 0,95} = 1$

alınarak

Q_{den} (2,50)>F_{kur} (1,00)

olduğu ve $\mathbf{H_o}$'ın reddedildiği görülür. Bu fark anlamlıdır (bak. 3.7.2).

Alıştırma 3.5

Alıştırma 3.4'teki değerlerin değişkelerini F sınaması ile (P=%95) karşılaştırınız.

(Y: F=9,36 ve fark anlamlı değil, Q < F).

Çizelge 3.2: P = 0,95 olasılığı için F- Snedecor sınaması değerleri (P = 0,99 için bak. Ek 7.2 ve 7.3).

Küçük s^2 – Değerleri için Serbestlik Derecesi sayısı	Büyük s^2 değerleri için serbestlik derecesi sayısı									
	1	**2**	**3**	**4**	**5**	**7**	**10**	**20**	**50**	**100**
1	161	200	216	225	230	237	242	248	252	254
2	18,5	19,0	19,2	19,3	19,3	19,4	19,4	19,4	19,5	19,5
3	10,1	9,6	9,3	9,1	9,0	8,9	8,8	8,7	8,6	8,5
4	7,7	6,9	6,6	6,4	6,3	6,1	6,0	5,8	5,7	5,6
5	6,6	5,8	5,4	5,2	5,1	4,9	4,7	4,6	4,4	4,4
7	5,6	4,7	4,4	4,1	4,0	3,8	3,6	3,4	3,3	3,2
10	5,0	4,1	3,7	3,5	3,3	3,1	3,0	2,8	2,6	2,5
14	4,6	3,7	3,3	3,1	3,0	2,8	2,6	2,4	2,2	2,1
20	4,4	3,5	3,1	2,9	2,7	2,5	2,4	2,1	2,0	1,8
50	4,0	3,2	2,8	2,6	2,4	2,2	2,0	1,8	1,6	1,4
100	3,8	3,0	2,6	2,4	2,2	2,0	1,8	1,6	1,4	1,0

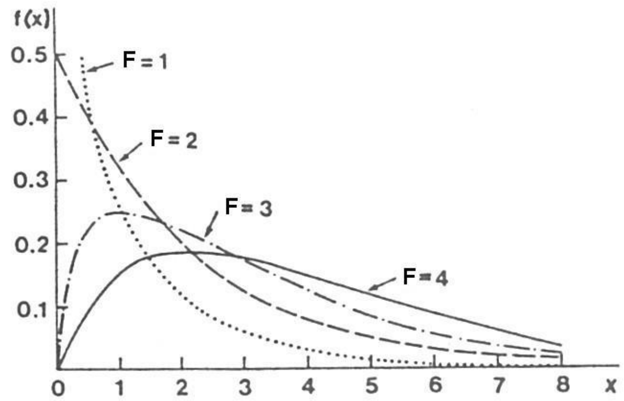
3.7.3 Ki kare (χ^2) dağılımı

Ki kare sınama yöntemi ana kütle özelliklerinin dağılımları hakkında hiçbir koşul gerektirmiyen bir evrensel sınama yöntemidir (Şekil 3.8). Ancak bu yöntem sadece sınıflandırılmamış değişkenlere uygulanmak-tadır. Bundan farklı göstergelerin olması durumunda bir sınıflandırma ile gerekli sınıflandırma tipi yapılarak sorun aşılır.

Ki kare yöntemi,

a) Bir gözleme dayanan veya deneysel dağılımın bir kuramsal dağılıma, örneğin, normal dağılıma, uyumu ve

b) Bir örneklemin özellikleri arasında bir ilişkinin bulunup bulunmadığının incelenmesi için kullanılır.



Şekil 3.8. χ^2 olasılık dağılımının F serbestlik derecelerine göre olasılık yoğunluk fonksiyonu.

Bu çok kullanışlı yöntem, aşağıda **bir örnekle** açıklanacaktır:

Örnek 3.8

a) şıkkına örnek olarak bir bilgisayarcının araba, cep telefonu ve internet kullanımı incelemesi verilebilir: 3 araç çeşidinin de normal dağılmış olmasına karşın bilgisayarcı, araç çeşitlerinin aynı sıklıkla dağılmadığını savunmaktadır. Buna göre,

1. H_o varsayımı: 3 araç türü de aynı sıklıkta dağılmıştır.
   H_1 varsayımı: 3 araç türü aynı sıklıkta dağılmamıştır.

2. Anlamlılık düzeyi $\alpha = 0{,}05$ verilmiştir.
   Bilgisayarcı rastlantısal olarak 600 aracı ele alır ve örneklemin,

180 arabayı,

186 cep telefonunu ve

234 internet bağlantısını

kapsadığını görür. Örneklemin araç dağılımı H_o'da savunulduğu gibi eşit dağılmış mıdır?

H_o varsayımına göre 3 araç çeşidinin oranları eşit olması lazımdır. Ancak incelenen araçlar bir örneklem olduğundan, ana kütleyi oluşturan araç çeşitleri ile aynı dağılımı göstermesi olanaksızdır. Farkların da H_o'ı kabul edecek kadar büyük olmaması gerekir. H_o varsayımına göre örneklemde her araç çeşidinden 200 adet bulunması beklenir. Buna göre araba için,

186-200= 14

tane araba farkı H_o varsayımını reddetmeye yeter mi?

Çözüm için her özelliğin **gözlenen sıklığı g_i** ile **beklenen sıklığı b_i** (frekanslar) arasındaki farkların bulunması gerekir (bak. aşağıdaki veriler). Artan örnek sayısına bağlı olarak, H_o varsayımı reddedilecek şekilde, farkların büyüme olasılıkları da artar. Bu nedenle farklar, b_i'ye bölünmek suretiyle, standardize edilir. Böylece sadece bir değer farkı değil, tüm değerlerin farkları işlem için önem kazanır. Bu arada birbirini sıfırlıyan ve yanlış sonuca neden olacak pozitif ve negatif farklar ortaya çıkar. Bu nedenle farkların karelerinin alınması ve toplanması gerekir. Bu işlemler sonunda ki kare sınamasının sınama büyüklüğü χ^2,

$$\chi^2 = \sum_{i=1}^{k} \frac{(g_i - b_i)^2}{b_i} \qquad 3.37$$

elde edilir (bak. aşağıdaki veri çizelgesi). Veriler:

	g_i	b_i	g_i^2	g_i^2/b_i
	180	200	32.400	162,00
	186	200	34.596	172,98
	234	200	54.756	273,78
$\Sigma_{i=1}^{k}$	**k = 600**	**600**		**608,76**
			$\Sigma_{i=1}^{k}$ **k** =	-600,00
			χ^2 =	8,76

Ki karenin **büyük değeri** örneklemin gözlem veya deneysel sonuçlara dayanan dağılımının, H_o ile ifade edilen şeklinin, kuramsal dağılımdan uzaklaştığının bir belirtisidir. Bu durumda H_o varsayımı reddedilir. Buradaki örnek için hesaplanan χ^2,

$$\chi^2_{den} = 8{,}76$$

değerine karşılık gelecek kritik çizelge değerini okuyabilmek için serbestlik derecesi F'ye gereksinim vardır. Bu örnekte F=3-1=2'dir.

Buna göre sınama değeri $\chi2_{kur}$,

$$\chi2_{2;0,05}=5{,}99$$

olarak okunur (Çizelge 3.3). $\chi 2_{den}$= 8,76

>5,99

olduğundan, fark α =0,05 için anlamlıdır ve kesinti süreleri aynı oranda dağıldıklarına ilişkin H_o varsayımı reddedilir. Buna göre kesintiler aynı sıklıkta dağılmamıştır.

Ancak α = 0,01 alındığında χ^2 çizelgesinden F = 2 için,

$\chi 2_{kur}$ = 9,21

bulunur. $\chi 2_{den}$ (8,76) < $\chi 2_{kur}$ (9,21)

çıktığından, H_o varsayımı kabul edilir. Anlamlılık düzeyi α = 0,01 değeri, yapılan her 100 deneyin ancak 1'nde böyle bir sonuca rastlamanın umulduğu anlamına gelir. Bu da çok düşük bir olasılıktır.

Bu dağılımlarla sadece yukarıda verilen örneklerdeki değişkenler karşılaştırılmıyor. Ortanca, tepe değeri, kayma ve basıklık da sınanabilir. 6. konuda da burada anlatılanlara ek ayrıntılı bilgiler verilecektir. Bu açıklanan çeşitlerine ek olarak **Weibull** ve **Kolmogorov-Smirnov** gibi özel dağılımlar sayılabilir. Gereksinim duyanların bu konudaki özel istatistik kaynaklarına başvurmaları salık verilir (bak. kaynakça).

Çizelge 3.3. Bazı kritik ki kare değerleri. H_o varsayımının doğru olması halinde, ki karenin çizelge değerine eşit veya bundan büyük eğer alma olasılıkları

Anlam düzeyi	**P** (%) →	**0,99**	**0,95**	**0,90**	**0,70**	**0,50**	**0,30**	**0,10**	**0,05**	**0,01**	**0,001**
Serbestlik derecesi F↓	**1**	0,00016	0,0039	0,016	0,15	0,46	1,07	2,71	3,84	6,64	10,83
	2	0,02	0,10	0,21	0,71	1,39	2,41	4,60	5,99	9,21	13,82
	3	0,12	0,35	0,58	1,42	2,37	3,66	6,25	7,28	11,34	16,27
	4	0,30	0,71	1,06	2,20	3,36	4,88	7,78	9,49	13,28	18,46

Açıklama: 3 sınıf bulunduğundan, ancak 2 sınıf için beklenen sıklıklar serbest seçilebilir. 3. sınıf için sıklıklar, verilen örnek sayılarından, e_3=n-e_1-e_2, hesaplanabilir. 3. sınıf artık serbest seçilemiyeceğinden, 2 serbestlik derecesi bulunmaktadır.

Alıştırma 3.6

Bir işletmede 4 tür üretim hatasına rastlanmaktadır. Bu hataların aynı oranda yapılıp yapılmadığını saptamak için bir araştırma yapılmıştır. Hatalı üretilen üründen rastgele 120 örnek alınarak incelenmiş ve hataları 32; 29; 27 ve 32 olarak kaydedilmiştir. Bunun sonucu olarak,

H_o: Hata türleri normal dağılmıştır.
H_1: Hata türleri normal dağılmamıştır.

(Y.: % 90 olasılıkla H_o kabul,)

EKLER

Ek 3.1. Normal dağılımda ortalama değer $\bar{x}$ ve belli z değerleri arasında bulunabilecek olayların oranı (% F*).

z	F*	z	F*(%)	z	F*(%)	z	F*(%)	z	F*(%)
0,0	0,00	1,0	34,13	2,0	47,72	3,0	49,865	4,0	49,997
0,1	3,98	1,1	36,43	2,1	48,21	3,1	49,903	5,0	49,99997
0,2	7,93	1,2	38,49	2,2	48,61	3,2	49,931	6,0	49,9999997
0,3	11,79	1,3	40,32	2,3	48,93	3,3	49,952		
0,4	15,54	1,4	41,92	2,4	49,18	3,4	49,966		
0,5	19,15	1,5	43,32	2,5	49,38	3,5	49,977		
0,6	22,57	1,6	44,52	2,6	49,53	3,6	49,984		
0,7	25,80	1,7	45,54	2,7	49,65	3,7	49,989		
0,8	28,81	1,8	46,41	2,8	49,74	3,8	49,993		
0,9	31,59	1,9	47,13	2,9	48,81	3,9	49,997		

z=(x- $\bar{x}$)/σ. F değeri, z=0 ve z=(x- $\bar{x}$)/σ aralığında eğri altındaki toplam alana karşılık gelir. Bu, $\bar{x}$ ve x sınırları arasında değişkenin % olarak gerçekleşme sıklığı demektir.

Örnek: $\bar{x}$ = 4, σ = 1 ve x = 1 için z = 3’ten → F* = % 49,865 ve F = % 0,135 bulunur. Tüm dağılım eğrisi için F* = % 99,73 ve 0 F = 0,27 geçerlidir.

Ek 3.2. Standart normal dağılımın birikimli olasılık değerleri (z değerleri, Davis, 1973).

Ortalama değerin standart sapması (z)	Birikimli olasılık	Ortalama değerin standart sapması (z)	Birikimli olasılık	Ortalama değerin standart sapması (z)	Birikimli olasılık
-3,0	0,0014	-0,9	0,1841	+1,2	0,8849
-2,9	0,0019	-0,8	0,2119	+1,3	0,9032
-2,8	0,0026	-0,7	0,2420	+1,4	0,9192
-2,7	0,0035	-0,6	0,2743	+1,5	0,9332
-2,6	0,0047	-0,5	0,3085	+1,6	0,9452
-2,5	0,0062	-0,4	0,3446	+1,7	0,9554
-2,4	0,0082	-0,3	0,3821	+1,8	0,9641
-2,3	0,0107	-0,2	0,4207	+1,9	0,9713
-2,2	0,0139	-0,1	0,4602	+2,0	0,9773
-2,1	0,0179	0,0	0,5000	+2,1	0,9821
-2,0	0,0228	+0,1	0,5398	+2,2	0,9861
-1,9	0,0287	+0,2	0,5793	+2,3	0,9893
-1,8	0,0359	+0,3	0,6179	+2,4	0,9416
-1,7	0,0446	+0,4	0,6554	+2,5	0,9938
-1,6	0,0548	+0,5	0,6915	+2,6	0,9953
-1,5	0,0668	+0,6	0,7257	+2,7	0,9965
-1,4	0,0808	+0,7	0,7580	+2,8	0,9974
-1,3	0,0968	+0,8	0,7881	+2,9	0,9981
-1,2	0,1151	+0,9	0,8159	+3,0	0,9987
-1,1	0,1357	+1,0	0,8413		
-1,0	0,1587	+1,1	0,8643		

4 İSTATİKSEL KESTİRİM (TAHMİN) YÖNTEMLERİ

4.1 Genel

İstatistiksel kestirim veya tahmin yöntemleri, örneklemin bilinen ölçütlerinden bilinmiyen ana kütle (popülasyon) ölçütleri hakkında bilgi edinmeye yarıyan yöntemlerdir. Ana hatları ile 2 kısma ayrılır:

1. Ana kütle değişkenlerinin örneklemin moment gibi değişkenleri yardımı ve kestirim yöntemleri ile kestirmek. Buna **nokta kestirimi** denir ve sadece değişkennin kendisi hesaplanır. Örneklemin veya ana kütlenin (örnek evreninin) dağılım değişkenleri de hesaplanabileceğinden, **dağılım kestirimi** de denir.

2. Örneklemlerin değişkenlerinden ana kütlenin değişkenlerinin belli bir olasılıkla kestirilmesine de **aralık** veya **güven sınırları** kestirimi denir.

4.2 Nokta kestirimi

Nokta kestirimi, örneklemin ölçütlerinden yararlanarak ana kütlenin özelliklerini bulma yöntemidir. **Değişken yönteminde**, örneklemin değişkenleri, bilinmiyen ana kütlenin değişkenlerine eşit sayılır. **En iyi kestirim yönteminde** ise (maximum likelihood), örneklemin n bağımsız verisinden çıkacak sonuçların olasılığını bilinmiyen ana kütle değişkenlerinin fonksiyonu olarak verilir: $L(\theta)=f(x_1, x_2,...x_n; \theta)$ gibi.

Örneklemin hesaplanan değişkenlerinden ana kütlenin ölçütleri kestirilebilir. Böylece bir örneklem bir ana kütleye uyarlanabilir. Bir örneklem, örneğin, ana kütleden daha az örnek kapsadığından, her zaman hata içerir. Dolayısı ile hiçbir zaman ana kütleyi kesin temsil edemez. Örneklemin büyüklüğü hata oranını etkilediği gibi aynı ana kütlenin bağımsız örneklemlerinin yinelenmesi de, örneğin, farklı $\bar{x}$ değeri verebilir.

Bu nedenle örneklemlerin değerleri ile ana kütlenin değerlerinin bulunması ancak bir kestirim fonksiyonu ile mümkündür. x, örneklemin değişkeni ve n de örnek kapsamı ise, bu fonksiyon,

$$f(x, n) = \hat{x} \qquad 4.1$$

ana kütlenin α gerçek değerinin kestirimidir (estimation).

Bir kestirimin belli matematiksel koşulları yerine getirmesi gerekir. Bunlar,

1. Kestirimin **beklenen değere** (ana kütle değerine) **yaklaşması** (unbaised): Yani deneyin aynı örnek kapsamı ile yinelenmesi halinde ortalama değer ana kütlenin gerçek değişken değerine yaklaşması,

2. Kestirimin **tutarlı** olması (consistent): Artan n örnek veya deney sayısı ile kestirim, ana kütlenin değişken değerlerine yaklaşması,

3. Kestirimin **etkili** olması (efficient): Aynı örnek sayısı için olası en küçük değişkeye sahip olması ve

4. **Bilgilendirici** olması lazım (sufficient): Başka hiçbir değişken kestirilen değişkenden daha çok bilgi verememeli.

4.3 Aralık kestirimi

Aralık kestiriminde, örneğin, örneklemin ortalama değeri için,

$$\mu = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad 4.2$$

eşitliğinden yararlanılır. Örneklem $\bar{x}$ ile ana kütle μ ortalama değerleri arasında her zaman için bir hata payı (ε) bulunur ve bu,

$$\bar{x} = \mu + \varepsilon \qquad 4.3$$

şeklinde ifade edilir. Bu nedenle hata payı ε için ancak belli bir olasılıkla bir aralığın belirtilmesi bir anlam taşır. Ana kütlenin gerçek değerini bulabilmek için ya örnek kapsamını oldukça büyük seçmek veya çok sayıda aynı kapsamlı örneklemi almak gerekir. Burada aynı kapsamlı birçok örneklem tercih edilip ortalama değer ayrı ayrı hesaplandığında bu ortalama değerlerin eşit olmadığı ve tüm ortalamaların ortalama değeri $\bar{\bar{x}}$ etrafında $\sigma_{\bar{x}}$ standart sapması ile (ortalama değerin standart hatası) normal dağıldıkları görülür. Kestirim fonksiyonunun ana kitlenin gerçek değişkenine yaklaşması beklentisi nedeniyle,

$$\mu = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad 4.4$$

ve $\bar{x} = \mu$

olur. Değişkesi için de ,

$$\sigma^2 = s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu)^2 \qquad 4.5$$

$$\sigma^2 = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad (N \neq n) \qquad 4.6$$

eşitliği uygulanır. Burada örneklemin bilinen $\bar{x}$ ortalama değerinden ana kütlenin μ ortalama değeri kestirilebilmektedir. Aynı şekil s^2 değerinden de σ^2 kestirilebilir. Örneklem için rastlantısal değişken X= $\bar{x}$ =μ alındığında,

$$P(a \leq X \leq b) = 0{,}95 \qquad 4.7$$

yazılabilir. X=Z, $-z=\dfrac{a-\mu}{\sigma_{\bar{x}}}$, $z=\dfrac{b-\mu}{\sigma_{\bar{x}}}$ ve 0,95 = 1-α alındığında (standardize edildiğinde),

$$P\left(\frac{a-\mu}{\sigma_{\bar{x}}} \leq Z \leq \frac{b-\mu}{\sigma_{\bar{x}}}\right) = 1-\alpha \qquad 4.8$$

$$P(-z \leq Z \leq z) = 1-\alpha \qquad 4.9$$

elde edilir (Şekil 4.1). z değişkeninin normal eğri alanları çizelgesinden okunacak $z_{\alpha/2}$ değerleri arasında herhangi bir değeri alma olasılığı,

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1-\alpha \qquad 4.10$$

şeklinde güven aralığına dönüştürelebilir. Bu eşitlikte z= $\bar{X}$ -μ / $\sigma_{\bar{x}}$

yazılırsa,

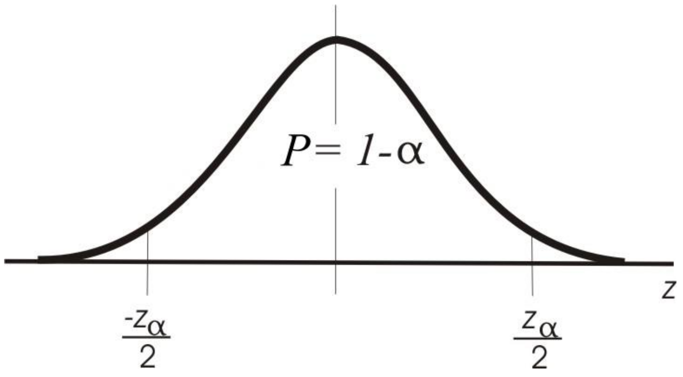
$$P\left(-z_{\alpha/2} < \frac{\bar{X}-\mu}{\sigma_{\bar{x}}} < +z_{\alpha/2}\right) = 1-\alpha \qquad 4.11$$

ve
$$P(-z_{\alpha/2}.\sigma_{x/2} < \bar{X}-\mu < +z_{\alpha/2}.\sigma_{x/2}) = 1-\alpha \qquad 4.12$$

yazılabilir. Buna göre ana kütle ortalamasının güven aralığı,

$$P(\bar{X} - z_{\alpha/2}.\sigma_{x/2} < \mu < \bar{X} + z_{\alpha/2}.\sigma_{x/2}) = 1-\alpha \qquad 4.13$$

şeklini alır.



Şekil 4.1. Aralık kestirimi. Bir değerin normal dağılımın kenarlarından birine düşme olasılığı α/2'dır.

Örneğin, t çizelgesinden (Çizelge 4.1) $z_{\alpha/2}$ =1,96 seçildiğinde,

$$\frac{a-\mu}{\sigma_{\bar{x}}} = -1{,}96 \qquad 4.14$$

$$\frac{b-\mu}{\sigma_{\bar{x}}} = 1{,}96 \qquad 4.15$$

$$a = \mu - 1{,}96\,\sigma_{\bar{x}} \qquad 4.16$$

$$b = \mu + 1{,}96\,\sigma_{\bar{x}} \qquad 4.17$$

bulunur. Bu sonuç, i'ninci örneklemin ortalama değeri $\bar{x}_i = \mu_i$'nin % 95 olasılıkla,

$$\mu - 1{,}96\,\sigma_{\bar{x}} \le \bar{x}_i \le \mu + 1{,}96\,\sigma_{\bar{x}} \qquad 4.18$$

aralığında olduğunu gösterir. Bu, ana kitle ortalama değeri μ'nün bilinmesi halinde örneklemin ortalama değeri $\bar{x}_i$ hakkında bir yargıya varılabilir demektir. Ancak **asıl hedef** $\bar{x}_i$ yardımı ile **μ hakkında bir yargıya varmak**tır. Örneğin,

$$\mu - 1{,}96\,\sigma_{\bar{x}} \le \bar{x}_i \qquad \bar{x}_i \le \mu + 1{,}96\,\sigma_{\bar{x}} \qquad 4.19$$

$$-\bar{x}_i - 1{,}96\,\sigma_{\bar{x}} \le -\mu \qquad -\mu \le 1{,}96\,\sigma_{\bar{x}} - \bar{x}_i \qquad 4.20$$

$$\bar{x}_i + 1{,}96\,\sigma_{\bar{x}} \ge \mu \qquad \mu \ge -1{,}96\,\sigma_{\bar{x}} + \bar{x}_i \qquad 4.21$$

olduğu ortaya çıkmaktadır. Buradan da güven aralığının genel eşitliği,

$$\bar{x}_i - z_\alpha \sigma_{\bar{x}} \leq \mu \leq \bar{x}_i + z_\alpha \sigma_{\bar{x}} \qquad 4.22$$

bulunur.

Böylece anlamlılık derecesi α'nın verilmesi ve $\bar{x}_i = \bar{x}$ ile $\sigma_{\bar{x}}$ değerinin bilinmesi durumunda ana kütlenin ortalama değeri μ'yü içeren aralık hakkında bir yargıya varılabilir. Bu aralık **güven aralığıdır**. $\sigma_{\bar{x}}$ 'in bilinmemesi koşullarında,

$$\sigma_{\bar{x}} = \sigma/\sqrt{n} \qquad 4.23$$

$$\approx s/\sqrt{n}$$

alınır. Genel olarak,

$$\bar{x}_i - z_\alpha . s/\sqrt{n} \leq \mu \leq \bar{x}_i + z_\alpha . s/\sqrt{n} \qquad 4.24$$

eşitliği geçerlidir. Çizelge 4.1'de kullanılacak bazı önemli değişkenler özetlenmiştir.

Çizelge 4.1. P olasılığının (kesinlik derecesi) veya α anlam düzeyinin (yanılma olasılığı) verilmesi durumunda güven sınırlarının kestiriminde kullanılacak önemli z ve t değerleri ile F serbestlik dereceleri (n-1).

P [%]	**α (1-P)**	**z**	**t = F_{10}**	**F_{15}**	**F_{20}**	**F_{25}**	**F_{50}**	**F_{100}**
80	0,20	1,282	1,37	1,34	1,32	1,32	1,30	1,29
90	0,10	1,645	1,81	1,75	1,72	1,71	1,68	1,66
95	0,05	1,960	2,23	2,13	2,09	2,06	2,01	1,98
99	0,01	2,576	3,17	2,95	2,84	2,79	2,69	2,62
99,9	0,001	3,290	4,59	4,07	3,85	3,72	3,50	3,39

Örnek 4.1

Bir maddenin özgül ağırlık ölçümleri {2,20; 2,25; 2,25; 2,30; 2,30; 2,30; 2,35} g/cm^3 olarak verilmiştir. Örneklemin ortalama değerinin P = % 95 olasılıkla güven aralığı kaçtır?

Çözüm:

$$\bar{x} = \frac{1}{7}\sum_{n=1}^{7}(2{,}20 + 2.2{,}25 + 3.2{,}30 + 2{,}35)$$

$$= 2{,}28 \text{ g/cm}^3$$

bulunur.

Kesinlik derecesi $P = 0{,}95$'ten,

Yanılma payı
$$\alpha = 1-P = 1-0{,}95 = 0{,}05$$
elde edilir.

Serbestlik derecesi
$$F = n-1 = 7-1 = 6\text{'dır.}$$
Buna göre,

F = 6 için t testi çizelgesi (Çizelge 3.1) $t_{\alpha/2} = t_{0,025} = 2{,}45$

değeri bulunur. Eşitlik 4.24'ü kullanarak ($t_{\alpha/2} = z_{\alpha/2}$ ve $\alpha = \sigma$),

$$2{,}28-2{,}45.(0{,}05/\sqrt{7}) < \mu < 2{,}28+2{,}45.(0{,}05/\sqrt{7})$$
$$2{,}28-0{,}046 < \mu < 2{,}28+0{,}046$$
$$2{,}234 < \mu < 2{,}326$$
$$\mu = 2{,}28 \pm 0{,}05$$

sonucuna varılır.

Alıştırma 4.1

Çizelge 2.1'deki verilerin ortalama değerinin % 95 olasılıkla güven aralığını bulunuz. Aynı örneklem aynı koşullarda n=170 ornek kapsasaydı güven aralığı ne olurdu?

(Y: 6,24±0,38 ve 6,24±0,13)

EKLER

Ek 4.1 Önemli bazı değişkenler için istatistiksel güven sınırları

Değişken	**n**	**Güven sınırı (±ℓ)** **P= % 95**	**P= % 99**
Ortalama değer ($\bar{x}$)	n ≥ 30	$1{,}96\ .s/\sqrt{n-1}$	$2{,}58\ .s/\sqrt{n-1}$
Ortanca (x_o)	n ≥ 30	$2{,}46\ .s/\sqrt{n-1}$	$3{,}23\ .s/\sqrt{n-1}$
Değişke (s^2)	n ≥ 100	$2{,}77\ .s/\sqrt{n-1}$	$3{,}65\ .s/\sqrt{n-1}$
Standart sapma (s)	n ≥ 100	$1{,}39\ .s/\sqrt{n-1}$	$1{,}82\ .s/\sqrt{n-1}$

Örnek

n=30 için $\bar{x} \pm 1{,}96.s/\sqrt{n-1}$ sınırları % 95 güvenle/olasılıkla ortalama değeri kapsar. Küçük örnek sayısı için ortalama değerin güven sınırları Student-t-değerleri vasıtası ile hesap-lanabilir: Örneğin, $x \pm t.s/\sqrt{n-1}$ gibi.

5 HATA HESAPLAMALARI

5. 1 Genel bakış

4. konuda örneklem değişkenlerinin hiçbir zaman gerçek değerleri olan ana kütle değerlerine eşit olamıyacağı ve bunların ancak belli olasılıklarla değişik aralıklarda bulunabileceği incelenmişti. Bir normal dağılımda örneğin, örnek sayısının % 95'i % 95 olasılıkla $\pm 2\sigma$ aralığında bulunur. Bu sınırlar (güven aralığı) dışında kalan uc kısımlar bu konuda incelenecektir.

Fiziksel ve kimyasal ölçüm ilkelerinin görevi g ölçü büyüklüğünün nicel ve nitel özelliklerini saptamaktır. Bunun için x, y, z ve t'ye göre aynı koşullarda ölçüm yapılır ve ölçü birimleri (ÖB) karşılaştırılır. $G = g(x, y, z, t) = z$.ÖB'dir. Burada z, özellik; n, örnek kapsamını ifade eder. z, sınırlı kesinliktedir ve zor saptanır. Buradaki değişimin ölçümü ile **hata hesaplanır**.

Hatalar,

1. Sistematik hata,
2. Rastlantısal hata

olmak üzere 2'ye ayrılır.

1. Sistematik hata, hedef değeri yönünde sapmadır, kolay giderilir. Bu hatalar tanımlandıktan sonra anlamlı dağılımdan sapmaları ile saptanır ve giderilebilirler (deneysel hata dağılım yasası). Bunlar,

- aygıt hatası,
- el ve kullanım hatası,
- değerlendirme hatası ve
- yorum hatası

gibi hatalardır. Kural olarak bunlar yuvarlak rakama veya en yakın tama tamlanırlar.

2. **Rastlantısal hatalar** ise, hedef yönündeki sapmadan ayırdedilemezler. Dolayısı ile zor giderilirler. Her zaman ve tesadüfen meydana gelebilirler. Bunlar örneğin, ölçüm sırasında gerilim düşmesinden kaynaklanan yanlış ölçüm, kaplardan parçacıkların bulaşması gibi hatalardır. Sonuçların esas kesinliğini bu tür hatalar belirler.

Rastlantısal hatalar, sistematik hataların ortadan kalkmasından sonra geride kalan hataları veya hata kısmını oluşturur. Düzenli dağılımlar için belirleyicidir ve ölçüm değişimlerinin dağılımı belli bir yasaya uyarlar. Bu yasaya, **hata dağılım yasası** denir ve rastlantısal dağılımı tanımlar: Buna göre, örneğin, v ölçüm büyüklüklerinin çeşitleme olanakları ise, çok sayıdaki x_i ölçümlerinin dağılımı, $n=\infty$ ve $v<\infty$ için **binom**, $v=\infty$ için de **normal dağılım** gösterir. Bu dağılımlarda eğri altında kalan alanın toplamı 1'e eşittir**.** x_i normal dağılım ölçümlerinde sadece rastlantısal hatalar araştırılır. Bun-ların içinde okuma hataları gibi sistematik hatalar da olabilir. Gauss hata hesaplama yöntemi uygulanır. Bunun için örneğin, ortalama değerler karşılaştırılır. Buradaki zorluk, sonuçların yaklaşık değer vermesidir. Elde edilen sonuçlar normal ve binom dağılımları ile karşılaştırılır.

5. 2. Hata kestirimi

Hata tahminleri veya kestirimleri ancak ölçüm dizisinde x_i (i=1, 2,...,n) ölçümlerinin aynı koşullarda ve belli birimlerle sağlanması durumunda mümkündür. Burada x_i'nin dağılım yasalarına uygun olup olmadığı denetlenir. Hesaplamalar için $\bar{x}$ en iyi değer olarak alınır. Bu değerden sapmalar **hata** olarak tanımlanır. Bunlar aşağıdaki şakilde incelenemektedir (bak. ayrıca Schönwiese, 1992):

a) Ortalama hata, bir tek ölçümün hatası veya sapma değeri $x_i' = x_i - \bar{x}$ 'dir. Bunların ortalama değeri d,

$$d = \frac{1}{n}\sum_{i=1}^{n}|x_i'| \qquad 5.1$$

eşitliği ile hesaplanır.

b) Standart sapma s de bir çeşit hatadır ve

$$s=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}x_i'^2} \quad (s>0) \qquad 5.2$$

bağıntısı ile hesaplanır. **Standart hata** anlamına gelir. En iyi veya ortalama değerin sapması veya hatasıdır. Burada,

c) Mutlak standart hata (Δx<d) önemlidir:

$$\Delta x = \sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n} x_i'^2} \qquad 5.3$$

$$= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} x_i'^2} \;/\; \sqrt{n} \qquad 5.4$$

$$= \frac{s}{\sqrt{n}} \qquad 5.5$$

bulunur ve varılan ölçüm kesinliğini belirler. Ölçüm sonucu (en iyi değer) Δx'teki gibi kesin verilmez. Bunun ölçü sonucuna (en iyi değere) göre neye oranlandığını göstermek için

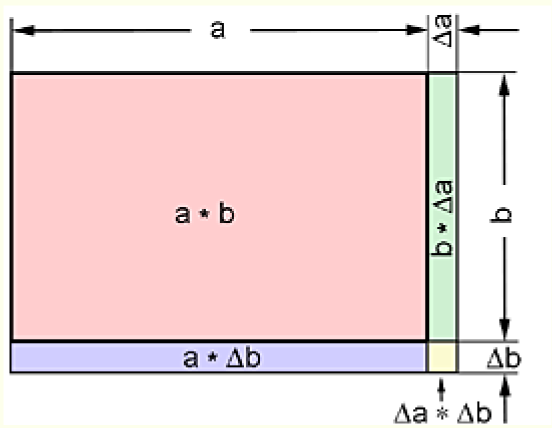
$$\frac{\Delta x}{\bar{x}}.100 = \pm\delta x \ [\%] \qquad 5.6$$

şeklinde verilir. Bu sonuç, en iyi değerin,

d) Göreceli standart hatasıdır (bak. aynı zamanda değişkenlik katsayısı v).

y = f(t, x, z, ...) gibi bir fonksiyonun hata payını hesaplamak oldukça karmaşıktır. Bu durumda her t, x, z değeri için ayrı ayrı hatalar hesaplanarak toplam fonksiyona etkisi bulunur. Bu işleme **hata ekleme** veya **hata ilerlemesi** denir (ayrıntılar için bak. Schönwiese, 1992).

Örnek 5.1 Hata türleri ve hesaplanması



Ölçüm hataları dikkate alınarak dikdörtgen alanının hesaplanması

$$(a+\Delta a)*(b+\Delta b) = \underbrace{a*b}_{\text{beklenen deger}} + \underbrace{a*\Delta b + b*\Delta a}_{\text{mutlak hata}} + \underbrace{\Delta a*\Delta b}_{\text{ihmal edilebiler hata.kismi}}$$

a) Beklenen değer A [Burada: Toplam alan]	**A = a * b**
b) Mutlak hata ΔA	
-> Gerçek değer (efektif, varılan sonuç) -> Beklenen değer (olması gereken/planlanan) -> Mutlak hata=gerçek değer - beklenen değer **-> Mutlak hata ΔA**	A* = (a + Δa) * (b + Δb) A = a * b ΔA = (a + Δa) * (b + Δb) - a * b **ΔA = a * Δb + b * Δa**
c) Göreceli hata h_g = $\dfrac{\text{mutlak.hata}}{\text{beklenen.deger}}$	$h_g = \dfrac{\Delta A}{A} = \dfrac{a*\Delta b + b*\Delta a}{a*b} = \dfrac{\Delta b}{b} + \dfrac{\Delta a}{a}$

Örnek 5.2

Bir ölçüm dizisi aşağıdaki sonuçları vermiştir (bak. alttaki Çizelge, $x_i' = x_i - \bar{x}$). Standart hata türlerini hesaplayınız.

Ölçüm No.	X_i	$\lvert x_i' \rvert$	x'^2
1	2,6	0,125	0,0156
2	2,8	0,075	0,0056
3	2,7	0,025	0,0006
4	2,7	0.025	0,0006
5	3,0	0,275	0,0756
6	2,7	0,025	0,0006
7	2,4	0,325	0,1056
8	2,8	0,075	0,0056
Σ n = 8	**21,8**	**0,850**	**0,1548**

Çözüm: Aritmetik ortalama $\bar{x}$,

$$\bar{x} = 21{,}8/8$$
$$= 2{,}725$$

eder.

a) Sapma değeri / **ortalama hata** d,

$$d = \pm\, 0{,}85 / 8$$
$$= \pm\, 0{,}10625$$
$$\approx \pm\, \underline{\%\ 11}$$

bulunur.

b) Standart hata s,

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} x'^{2}_{i}}$$

$$= \sqrt{0{,}1548/7}$$

$$= \underline{\pm\, 0{,}1487}$$

sonucunu verir.

c) Mutlak standart hata Δx ,

$$\Delta x = \pm \sqrt{\frac{1}{n(n-1)}\sum x'^{2}_{i}}$$

$$= \pm \sqrt{0{,}1548/8.7}$$

$$= \pm \sqrt{0{,}00276}$$

$$= \pm\, 0{,}0526$$

$$= \pm\, \underline{0{,}05\text{'tir.}}$$

Bulunan ortalama değer $\bar{x} = 2{,}72 \pm 0{,}05$ şeklinde gösterilir. Bu, $\bar{x}$'nın +2,77 ile -2,67 sınırları içinde geçerli demektir ve ± % 5 yanılma payını (1-P=α) verir. Bu gibi durumlarda yapılacak kritik, yorum anlamındadır.

d) Göreceli standart hata δx ise,

$$\delta x = \frac{\Delta x}{\bar{x}}.100$$

$$= \frac{0{,}05.100}{2{,}725}$$

$$= \underline{1{,}83\ [\%]}$$

bulunur. Bu, aynı zamanda düşük değişkenlik katsayısı v demektir ve örneklerin çok düzenli (v< % 20) dağıldığına işaret etmektedir.

Alıştırma 5.1

Bir fabrikada üretilen bir ürünün üretim süresinin belirlenmesi için 100 deney yapılmıştır. Bu deneylerin ortalama üretim süresi 7,30 saattir. Standart sapması 1,2 saat olduğuna göre, ortalama değerin standart hatasını bulunuz. (Y.: % 0,19)

6 SINAMA (TEST) YÖNTEMLERİ

6.1 Sınama yöntemlerinin ilkeleri

Sınama yöntemleri ile bir farkın hangi büyüklükten itibaren belli bir güvenirlikle rastlantısal olmadığı veya anlamlı görmek gerektiği araştırılır. Tüm sınamalar **varsayım** veya **sav**larla başlar. Varsayım veya sav, bir kuramsal sistemde kabul veya reddedilen, deneysel araştırma sonuç ve gözlemlerinin açıklanmasına yarıyan bir **öngörü**dür (= **hipotez, tez**). Sınama yöntem-lerinin uygulanması önemli bir görevdir. İşlenmesi çok kuramsaldır. Ancak burada öncelikle uygulamaya yönelik işlenecektir.

Sınama uygulamalarında 2 varsayım bulunur:

a) H_o, sıfır varsayımı (Null-hypothesis),
b) 1. veya 2. seçenek varsayımlar (alternative hipothesis), H_1 ve H_2 gibi.

Bu varsayım veya savlar karşılaştırılır. Biri kabul edilirse, diğeri reddedilir. **H_o varsayımı rast-lantısal,** sıfır veya bir hiçtir. Bununla 2 örneklem arasında bir fark olmadığı, iki örneklemin de aynı ana kütleye ait olduğu varsayılır. Ancak **karşıt H_1 veya H_2 savları** H_o'ın tersini savunur ve H_o'ın yanlış olduğu gösterilir. H_1 ve H_2 rastlantısal değil, belirleyici veya anlamlıdır. **Anlamlılık düzeyi, P** ile tanımlıdır. Burada P= olasılık veya **güven derecesi** = 1-α; α, **anlam düzeyi** veya yanılma olasılığıdır (bak. Şekil 4.1 ve Çizelge 4.1).

İstatistikte kesinlik yerine olasılık bulunduğunda,

H_1 kabul edilirse, H_o reddedilir → H_o (olumlu, istenen sınama sonucu)
H_o kabul görürse, H_1 reddedilir → H_1 (olumsuz sınama sonucu)

İki varsayımın olasılıkları toplamının, yani $H_o + A_1 + A_2 = 1$ etmesi gerekir. Buradan savların tüm olasılıkları kapsaması gerektiği ortaya çıkar.

Sıfır varsayımı: $H_o=\{a=b\}$ savını savunur. Karşıt savlar, H_1: $\{a\neq b\}$ ve H_1: $\{a<b\}$ ile H_2 : $\{a>b\}$ gibi savları savunur.

6.2 Sınama yöntemi ilkelerinin belirlenmesi

Sıfır ve seçenek varsayımlarının içeriğinden yöntem düzlemine geçerken varsayımı kavramını iyi tanımlamak gerekir. Konuyu bir örnekle açıklamak daha kolay olacaktır:

Örnek 6.1

Bir çiftlikten şimdiye kadar ortalama 9,5 birim/ha verim alınmaktadır. Çiftlik için on görülen bir sulama yönteminin ekonomik olabilmesi için en az 11 birim/ha verim gerektirmektedir. Standart sapması s = 5 birim/ha olan n= 30 örneğin ortalama değeri $\overline{x}$ = 13 birim/ha'dır. Başka örneklemeler farklı ortalama değerleri verdiğinden, ha başına 13 birim verimin rastlantı sonucu 11 birim/ha'dan büyük olduğuna işaret etmektedir.

Sınama yöntemi sorusu: Hangi büyüklükten itibaren fark belli bir güvenirlikle rastlantısal değil, anlamlıdır?

Çözüm: $\mathbf{H_o}$: a ve $\overline{x}$'in farklı değerleri rastlantıdır (fark anlamlı değil). Ana kütle için μ=a geçerlidir.

$\mathbf{H_1}$**:** Sulama ile elde edilen verim öngörülen verimden büyük ve anlamlıdır. Ana kütle için a<μ geçerlidir.

Amaç: İçerik düzeyinde sulama yöntemi uygulama amacına varmak için H_o'ın yanlış olduğu gösterilir.

Tüm sınama yöntemleri konuyu sonunda H_o kabul de edilse, red de edilse, α yanılma payı ile, reddetmeyi ön plana çıkararak işlerler.

Böylece sınama yönteminin şematik ilkesi, içerik ve yöntem düzeyi tamamlanmış olur.

Temel hedef, hesaplanan bir değerin rastlantısal olamıyacağını göstermektir.
Kabul sınaması $\mathbf{H_o}$: Fark yoktur.
Belli bir istatistiksel olasılıkla H_o reddedilir: Böylece amaca varılmış olur.
H_o kabul edilir: Amaca varılmamıştır.

Yukarıdaki örnekte a=11 birim/ha, $\bar{x}$=13 birim/ha olarak verilmiştir. Burada $\bar{x}$-a farkının belli bir kesinlik derecesi ile sadece tesadüfen sıfırdan farklı olup olmadığına karar vermek gerekir. Bunun için ± ($\bar{x}$-a)'nın verilen sınırı aşma olasılığı hesaplanır.

Sonsuz sayıda çok ve n eşit sayıdaki örneklemlerin $\bar{x}$ ortalama değerlerinin ana kütlenin μ ortalama değeri etrafında

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}} \qquad 6.1$$

satandart sapması ile normal dağıldıkları yukarıda gösterilmişti (n, örnek sayısı). Her örneklemin ortalaması $\bar{x}$ 'ten a değeri çıkarıldığında $\bar{x}$-a farkları μ-a=0 etrafında

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}} \qquad 6.2$$

standart sapması ile **normal** dağılmış olacaklar (H_o'a göre μ-a=0). σ^2 bilinmediğinde, s^2 ile kestirilir. Bu durumda,

$$\frac{\bar{x}-a}{\frac{\sigma}{\sqrt{n}}} \approx \text{N (0;1)} \qquad 6.3$$

farkları,

$$\frac{\mu-a}{\frac{s}{\sqrt{n}}} \approx \text{t (1-α, n-1)} \qquad 6.4$$

etrafında **standart normal** dağılacaklar.

Bir örneklemin mevcut $\bar{x}$ ortalama değeri güven sınırları ile saptanan,

$$\frac{\bar{x}-a}{\frac{s}{\sqrt{n}}} \qquad 6.5$$

aralığına,

$$\frac{\mu-a}{\frac{s}{\sqrt{n}}} = 0 \qquad 6.6$$

farkı ile düşerse, fark rastlantısaldır ve H_o kabul edilir.

Ancak, $$\frac{\bar{x}-a}{\frac{s}{\sqrt{n}}} \qquad 6.7$$

bu aralığa düşmezse, $\bar{x}$ a'dan anlamlı bir şekilde farklı demektir (anlam düzeyi=α) ve H_o reddedilir. H_o'ın sağlanması için **sınama büyüklüğü d**,

$$d=\frac{\bar{x}-a}{\frac{s}{\sqrt{n}}} \qquad 6.8$$

hesaplanır ve standart normal dağılım değeri d_{kur} ile karşılaştırılır.

Tek taraflı sorgulama*	**Çift taraflı sorgulama**
$H_o : \mu=a \qquad H_1: \mu>a$	$H_o : \mu=a \qquad H_1: \mu \neq a$
$P(X\leq x) =0,95$	$P(\|X\|\leq x)=0,95$
$f(z)=0,95$	$f(z)=0,95$
$z=1,65$	$z=1,96\ (=d_{kur})$

Buradan sınama büyüklüğü d_{den},

$$d_{den}=\frac{\bar{x}-a}{\frac{s}{\sqrt{n}}} \qquad 6,9$$

$$=\frac{13-11}{\frac{5}{\sqrt{30}}}$$

$$=\underline{2,19}$$

$$\underline{>1,65}$$

bulunur. Hesaplanan bu sınama büyüklüğü **deneysel d_{den} değeri**dir. Burada d_{den}'nin, kuramsal değer d_{kur}= 1,65'ten büyük olması, farkın rastlantısal görülemiyeceği anlamına gelir. Dolayısı ile H_o reddedilir.

* **Tek taraflı sınama**, alternatif savın a<b gibi bir yön belirttiği; **çift taraflı sınama** ise, a≠b gibi yönün belirtilmediği sınama demektir. Anlam düzeyi $\alpha_{tek} = 1/2\ \alpha_{çift}$ taraflı sınamadır.

Örnek sayısı n≤30 durumları için t sınaması uygulanır: Buna göre,

	Tek taraflı sorgulama	Çift taraflı sorgulama
n=30 ve α=0,05 için,	d_{kur} =1,699	d_{kur} =2,045

elde edilir. Sonuç: ortalama değerler (13-11) arasındaki fark anlamlıdır ($d_{den} > d_{kur}$), sulama verimi anlamlı bir şekilde arttıracak ve ekonomiktir.

6.3 Sınama çalışmaları şemaları

Bir sınama aşağıdaki şemaya göre sonuçlandırılır:

a) Verilerin sıralanması: a=11 birim/ha,
n=30,
s=5 ve
$\bar{x}$ =13 birim/ha.

1. H_o ve H_1'in formüle edilmesi

H_o: $\bar{x}$ ve a arasındaki farklar tümüyle rastlantısaldır (ana kütle: $\bar{x}$ -a=0)
H_1: $\bar{x}$ ve a arasındaki fark tümüyle rastlantısal değildir (ana kütle: $\bar{x}$ >0).

Tek taraflı sorgulama

2. Yanılma olasılığı α'nın saptanması, α=0,05

3. Sınama dağılımının saptanması: t dağılımı (n≤30)

b) H_o'ın gerçek olup olmadığına ilişkin sınama büyüklüğünün saptanması:

4. d_{kur}'un hesaplanması (bak. ayrıca 6.9), $d_{kur} = \dfrac{\bar{x} - a}{\frac{s}{\sqrt{n}}}$

5. d_{den} ve d_{kur}'un karşılaştırılması: $d_{den} > d_{kur}$ ise, H_o red edilir (d_{kur}, F'ye göre çizelgeden okunur),

$d_{den} \leq d_{kur}$ ise, H_o kabul edilir.

Örneklem ölçütlerine göre, d_{den} =2,19> d_{kur} =1,699, H_o red edilir.

6. Sonuçların içerik olarak yorumlanması.

c) **Ana kütlelerin eşit kapsam ve değişkeye sahip ortalama değerleri için sınama**

Örnek 6.2

Normal dağılımlı iki ana kütlenin μ_a ve μ_k ortalamalarının eşitliğinin sınanması (ortalama buğday verimi, a, Adana; k, Konya):

a) $\sigma_a^2 = \sigma_k^2 = \sigma^2$,:

b) $\sigma_a^2 \neq \sigma_k^2$

Verilenler:

I.	**II.**
Adana	Konya
$\bar{x}_a$ = 6,26 birim/ha	$\bar{x}_k$ = 6,52 birim/ha
s_a^2 = 0,19 birim/ha	s_k^2 = 0,20 birim/ha
n_a = 22 örnek	n_k = 22 örnek

Adana ve Konya'nın buğday verimleri farklı mıdır?

1. H_o ve H_1'in formüle edilmesi:

H_o: $\mu_a = \mu_k$,

H_1: $\mu_a \neq \mu_k$ (iki taraflı sorgulama)

2. Yanılma olasılığı α'nın saptanması: α=0,05

3. Dağılım çeşidinin belirlenmesi: Normal dağılım (2n=44>30)

Serbestlik derecelerinin bulunması: F=2n-2=44-2=42

H_o'ın gerçek olup olmadığına dair sınama büyüklüğü d'nin bulunması

4. d_{den}'in bulunması (çizelgeden)

5. d_{kur} ve d'nin karşılaştırılması: $d_{kur} = 1{,}96 > d_{den} = 1{,}9527 \rightarrow H_o$ kabul.

6. Kapsamın irdelenmesi ve içeriğin yorumu.

Adana ve Konya örneği için çeşitli çözümler:

H_o	**H_1**	**α**	**d_{den}**	**d_{kur}**	**Karar**
$\mu_a = \mu_k$	$\mu_a \neq \mu_k$	0,05	1,9527	1,96	H_o kabul
$\mu_a = \mu_k$	$\mu_a < \mu_k$	0,05	1,9527	1,65	H_o red
$\mu_a = \mu_k$	$\mu_a < \mu_k$	0,01	1,9527	2,33	H_o kabul

6.4 Sınama yöntemlerinin genel sorunları

Sınama yöntemleri parametrik (parametric) ve parametrik olmıyan (nonparametric) yöntemlere ayrılırlar. t sınamsı örneğin, bir parametrik sınamadır. Çünkü ana kütlede normal dağılım geçerliliğini şart koşuyor. Parametrik olmıyan sınamalar dağılıma bağlı değildir. Parametrik sınamalar kadar kesin değiller. Bunlar, ki kare, Friedman, Mannn-Witney sınaması gibi sınama yöntemleridir.

Aşağıdaki kurallara dayanarak α = 0,05 veya 0,01 gibi düşük oranlarda tutuluyor. H_o'ın red ve kabul edilmesi α'nın seçimine büyük oranda bağlıdır. Her zaman için yanlış karar olasılığı bulunmaktadır. Çünkü sadece olası 4 karar düşünülebilir:

1. İki ortalama değerin farkı **saf rastlantısaldır**. Sınama sonucunda **H_o kabul** edilirse, **doğru karar** verilmiş olur.

2. İki ortalama değerin farkı **saf rastlantısaldır**. Sınama sonucunda **H_o kabul** edilirse, **karar yanlış** ve hata yapılmış olur. Buna 1. tür hata denir. Bu hata, α'nın çok küçük seçilmesi ile düşük tutulabilir. Çünkü t_α veya z_α yüksektir.

3. İki ortalama değerin farkı **rastlantısal değildir**. Sınama sonucunda **H_o reddedilirse**, verilen **karar doğru** olur.

4. İki ortalama değerin farkı **rastlantısal değildir**. Sınama sonucunda **H_o kabul** edilirse, verilen **karar yanlış** olur. 2. tür hata* yapılmış olur. Bu hata, α'nın çok büyük seçilmesi ile azaltılabilir. Çünkü t_α veya z_α küçüktür.

H_o hakkında karar	H_o doğru	H_o yanlış
Red	1. tür hata (2.)	Doğru karar (3.)
Kabul	Doğru karar (1.)	2. tür hata (4.)

Baştan itibaren farkın rastlantısal olup olmadığı bilinmediği için, her zaman bir hata yapılır. Bunun da mümkün olduğu kadar küçük tutulması istenir. 2. ve 4. noktaların karşılaştırılmasından bir hatanın küçük tutulması ancak diğerinin büyük tutulması ile mümkün olduğu anlaşılmaktadır. Bu iki hata riski karşı karşıya gelir ve ikisinin birden küçültülmesi mümkün olmamaktadır. Onun için baştan itibaren hangisinin küçük tutulması gerektiğine karar verilmelidir. Hangi riske girildiğinin bilinmesi için α'nın önceden belirlenmesi gerekir.

Örnek 6. 3

Yazı tura atmada tura gelme olasılığı rastlantısaldır ve P(O)=1/2= % 50 turadır. Yinelenen rastlantısal (stokastik) olaylarda,

2. kestirme için $P(O)=(1/2)^2 = 1/4 = \% \; 25$
3. kestirme için $P(O)=(1/2)^3 = 1/8 = \% \; 12{,}5$
4. kestirme için $P(O)=(1/2)^4 = 1/16 = \% \; 06$
5. kestirme için $P(O)=(1/2)^5 = 1/32 = \% \; 03$

geçerlidir (bak. ayrıca 1.5). Yanılma olasılığı α = % 5 verildiğinden, P = (1-α) = % 95 anlamlılık düzeyi bulunur. 4. atışta gelen tura rastlantısal kabul edilirken (H_o kabul), 5. kerede gelen tura rastlantısal kabul edilmez (H_1 kabul). Çünkü P(O) = % 3 < α = % 5 verilmiştir (güven sınırı P'nin dışına taşıyor) .

* Doğru sıfır varsayımının reddedilmesine **1.**; yanlış sıfır varsayımının kabul edilmesine de **2. tür hat** denir.

Alıştırma 6.1

Aşağıda belli ölçütleri verilen iki örneklemin aynı ana kütleye ait olup olmadıklarını bulunuz.

Ölçütler: $\bar{x} = 5{,}7$ $n_x = 30$ $s^2_x = 2{,}5$

$\bar{y} = 7{,}1$ $n_y = 20$ $s^2_y = 2{,}9$ $\alpha = 0{,}05$, P= % 95.

(Y.: Fark % 95 olasılıkla anlamlıdır, H_o red)

6.5 Önemli sınama yöntemlerinin uygulanması

6.5.1. Ortalama değerlerin karşılaştırılması

Ortalama değerlerin karşılaştırılmasında Ek 3.1'deki çizelgeden yararlanılarak Student-t sınaması uygulanır. Bu sınama yönteminin uygulanması için gereken önkoşul örneklemlerin en azından yaklaşık bir normal dağılım göstermesi ve standart sapmalarının anlamlı bir fark gösterme-meleridir. İki durum ayırdedilir:

a) Örnek sayılar $n_1 = n_2 = n$ olan deneysel 2 örneklemin ortalama değerlerinin karşılaştırılması. Burada önce sınama değeri t,

$$t = (\bar{x}_1 - \bar{x}_2)\sqrt{\frac{n-1}{s_1^2 + s_2^2}} \qquad 6.10$$

eşitliği ile hesaplanır ve serbestlik derecesi F=2n-2 eşitliği ile bulunur.

b) Örnek sayılar $n_1 \neq n_2$ olan deneysel 2 örneklemin ortalama değerlerinin karşılaştırılması. Burada önce $\hat{s}^2$ değerinin,

$$\hat{s}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \qquad 6.11$$

eşitliğine göre bulunması gerekir (bak. 3. konu). Böylece sınama değeri t,

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\hat{s}}\sqrt{\frac{n_1 . n_2}{n_1 + n_2}} \qquad 6.12$$

hesaplanabilir duruma getirilmiş olur.

Hesaplanan mutlak t değeri çizelgedeki t_{99} değerinden büyük çıkarsa, ortalama değerin sapması "anlamlıdır" denir. Bu, ortalama değerin % 99 kesinlikle örneklemin esas alındığı dağılıma uymadığı anlamına gelir. Eğer bulunan t değeri t_{95} ile t_{99} arasında ise, % 95 kesinlikle "muhtemelen anlamlıdır" denir.

6.5.2 Standart sapmaların karşılaştırılması

2 standart sapmanın karşılaştırılması için Snedecor-F sınaması (F sınaması, bak konu 3) uygulanır. Bu sınamada da örneklemlerin en azından yaklaşık normal dağılıma uymaları beklenir. F sınamasında, varsayıma göre, standart sapmaların eşit olduğu savunulur (H_o) ve Q oranı,

$$Q=\text{büyük } \hat{s}^2 \text{ değeri}/\text{küçük } \hat{s}^2 \text{ değeri}$$

ile sınanır. Bu oran ne kadar büyük olursa, örneklemlerin de o kadar farklı değişkeleri olan ana kütlelere ait olma olasılıkları bulunur. Kesin bir sonuç için **F sınaması** çizelgesinden (Çizelge 3.2) n_1-1 ve n_2-2 serbestlik derecelerine göre F değeri okunur. Q değerinin çizelge değerinden büyük olması durumunda örneklemlerin değişkeleri arasındaki fark anlamlı demektir (bak. 3. konu).

6.5.3 Dağılımların karşılaştırılması

t ve F sınamalarından farklı olan ki kare (χ^2) sınaması, dağılımları karşılaştırır; dağılım çeşidi hakkında başka bir koşul ileri sürmez. χ^2 yönteminin uygulanması evrenseldir, ancak alışık t ve F sınamalarına karşın kritik bakış açısı ister (bak. 3. konu).

Koordinatlardan absis, örneklemlerin 5'er, hatta 10'ar değerini alacak şekilde k (I_1, I_2,...I_k). aralıklarına bölünür. Her I_j aralığı için 1. ve 2. örneklemlerin I_j aralıklarına düşen g_j (gözlenen) ve b_j (beklenen) örnek değerleri belirlenir. Buradan χ^2 değerleri,

$$\chi^2_{den} = \frac{\Sigma(g_i - b_j)^2}{b_j} \qquad 6.13$$

hesaplanır. χ^2_{den} değeri ne kadar büyük çıkarsa, dağılımların da o kadar farklı olduğu anlaşılır. Buna bağlı olarak farkın anlamlılığı da artar. Önceki sınamalarda olduğu gibi tahminlerde bulunabilmek için çizelgelerden yararlanılır (bak. Çizelge 3.3). Çizelge, F serbestlik derecelerine göre, F = (k-1) (I-1)-h, hesaplanan χ^2_{den} ‘e uyan χ^2_{kur} değerinin bulunmasına yarar.

Tanıma göre saf ana kütler için aşağıdaki kurallar geçerlidir (P %):

$\chi^2_{den} > \chi^2_{kur}$ P_{99} = Dağılımın farklılığı anlamlıdır,.

$\chi^2_{den} > \chi^2_{kur}$ P_{95} = Dağılımın farklılığı muhtemelen anlamlıdır

$\chi^2_{den} \approx \chi^2_{kur}$ P_{95} = Büyük olasılıkla çok az örnek kapsamı

$\chi^2_{den} < \chi^2_{kur}$ P_{95} = Farkın olmadığına ilişkin H_o varsayımı çürütülemiyor.

Örnek 6.4

Bir bilgisayar firmasının satışları aylara göre aşağıda verilmiştir. H_o: Satışlar normal dağılım göstermektedir. H_1: Satışlar normal dağılım göstermemektedir.

Ay	Mayıs	Haziran	Temmuz	Ağustos
Satış	8	25	45	22
% oranı	15,9	34,1	34,1	15,9

$$\chi^2_{den} = \frac{\sum_{i=1}^{4} (g_i - b_i)^2}{b_i}$$

$$= \frac{(8-15{,}9)^2}{15{,}9} + \frac{(25-34{,}1)^2}{34{,}1} + \frac{(45-34{,}1)^2}{34{,}1} + \frac{(22-15{,}9)^2}{15{,}9}$$

$$= \underline{12{,}18}$$

bulunur. Serbestlik derecesi, yalnız 1 satır bulunduğundan, F = 4-1 = 3’tür. Bununla çizelgeden (Çizelge 3.3) kritik χ^2_{kur} değeri 11,34-16,27 arasında bir değer olarak okunur. Böylece satışların bir normal dağılım olduğunu savunan H_o hipotezi kabul edilmiş olur ($\chi^2_{den} > \chi^2_{kur}$: 12,18>11,34).

Alıştırma 6.2

Uykusuzluk çeken 10’ar 2 grup öğrenci üzerinde bir uyku ilacı denenmiştir. Deneme sonucu aşağıda verilmiştir. Buna göre,

a) H_o varsayımını ifade ediniz,

b) t sınaması ile sınayınız.

Veriler:

Hasta s.	1	2	3	4	5	6	7	8	9	10
A	1,9	0,8	1,1	0,1	-0,1	4,4	5,5	1,6	4,6	3,4
B	0,7	-1,6	-0,2	-1,2	-0,1	3,4	3,7	0,8	0,0	2,0
Fark	**1,2**	**2,4**	**1,3**	**1,3**	**0,0**	**1,0**	**1,8**	**0,8**	**4,6**	**1,4**

(Y.: a) İki ilacın etkisi aynıdır. B) $t_{9;\,0,01}$=3,25, H_o reddediliyor, A, B’den etkilidir).

7 VARYANS ANALİZİ

7.1 Temel ilkeler

Basit bir varyans analizi* (VA) bir ana kütle değişkenlerini karşılaştırır, yani ana kütlenin veya örneklemin örnekten örneğe değişen standart sapmasını inceler. Ana kütlenin saf olması durumunda örneklemler arasındaki farklılık karakteristik değildir. Aksine karışık bir ana kütlede örneklemler arası değişke örneklemlerin değişkesindan anlamlı bir şekilde büyük olur. Bu farklılık, örnklemin birine veya tümüne farklı olayların veya aynı şiddetteki bir olayın farklı etki ettiğini gösterir.

Varyans analizinden birkaç ana kütlenin karşılaştırılmasında yararlanılır. Saf ana kütlelerin sınanmasında, grup farklarının sınırlandırılmasında, örneğin, tıpta tedavi yöntemlerinin, yer-bilim-lerinde kayaç türlerinin veya fasiyes tiplerinin ayırdedilmesinde, uygulanır. Varyans analizi, **önemli etkenleri önemsizlerden ayırmayı sağlar**. Bu amaçla R. A. Fisher (1890-1962) tarafından deneyleri planlamak ve değerlendirmek için bulunmuş ve geliştirilmiştir. Bir tek bağımsız özelliğin incelenmesine dayanan varyans analizine **tek etkenli veya basit**, 2 veya daha fazla bağımsız özelliğin incelenmesi durumunda ise, **çok etkenli** varyans analizi denir. Yöntemin esasını <u>toplam varyansla kısmi varyansların karşılaştırılması</u> teşkil eder. Buna göre inceleme 1 etkene indirgenir. Yani bir ana kütlede bir özelliğin değişkenliği bağımsız birkaç etkenden kaynaklanıp kaynaklanmadığı araştırılır. Bu durumda,

$$\sigma_t^2 = \sigma_1^2 + \sigma_2^2 + ... + \sigma_n^2 \qquad 7.1$$

eşitliği geçerli olur (t, toplam; n, örnek sayısı). Eşitliğin sağlanması, incelenen örneklemlerin bir ana kütleden oluştuğunu, aksi taktirde kümenin heterojen olduğunu, yani 1'den fazla ana kütleden meydana geldiğini gösterir. Bu eşitlik standart sapma s için geçerli olamaz, yani,

$$s_t \neq s_1 + s_2 + ... + s_n\text{'dir.}$$

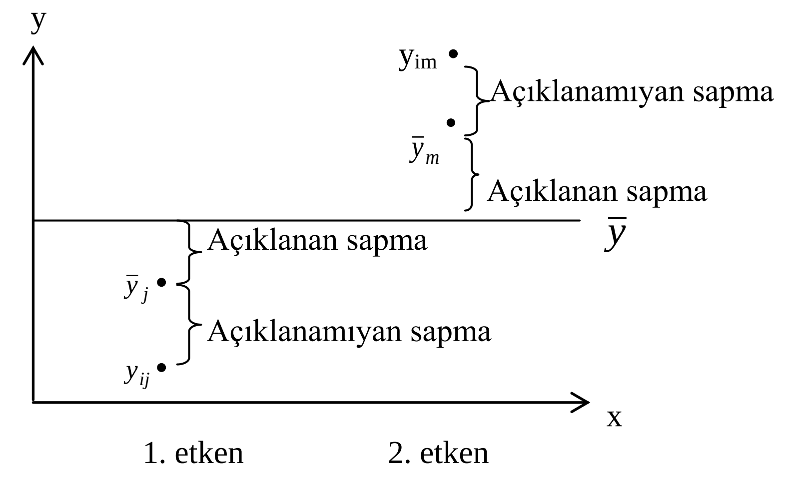
* Şimdiye kadar varyans yerine kullanılan "değişke" burada yerleşmiş "varyans analizi" yerine kullanılamaz. Değişke, standart sapmanın karesi olduğundan, yerine göre biri diğeri yerine kullanılmıştır.

Bulunan değişke farkları F sınamasına tabi tutulurlar. Karşılaştırmak için gerekli sınama değerleri ilgili tablolardan okunarak farklılıkların karakteristik olup olmadıkları tesbit edilir. Farklılığın tam belirlenemediği durumlarda **derece** veya **rang** yönteminden yararlanılır. Bu yöntemde mevcut analiz değerleri k sınıflarına ayrılır (k=1+2+3+...+m). Her sınıfın derişimi toplam yüzde ile çarpılarak dereceler elde edilir. Bu dereceler de çeşitli matematiksel işlemlere tabi tutularak farkların ortaya çıkarılması sağlanır.

Varyans analizinde değişkenler ana kütleler arasında araştırılır. Anlamlı olmıyan fark durumunda 1, aksi durumda en az 2 ana kütlenin bulunduğu ortaya çıkar. Bunun nedenleri oldukça farklı olmaktadır. Karar karşılaştırma sonunda verilir. İncelenen ana kütlelerin örnek sayısı eşit veya farklı sayıda olabilir.

Varyans analizinin geometrik anlamı Şekil 7.1.'de gösterilmiştir. Burada, bulunan bağımsız ortalama değer $\bar{y}$'dir. Bir özelliğin etkin olması durumunda da ortalama değerler $\bar{y}_j$ veya $\bar{y}_m$'dir. Bu beklenen ortalama değerlerden sapan $(y_{ij} - \bar{y}_j)$ farkları rastlantısal dış etkenlerden kaynaklandıklarından, açıklanamazlar. Böylece 7.1'de gösterilen toplam sapma veya saçınım,

açıklanan sapma ile açıklanamıyan sapmanın toplamı olduğu veya toplam sapmanın açıklanan ve açıklanamıyan sapma bileşenlerine (elemanlarına) ayrıldığı görülür (Backhaus ve diğ, 1990). Bunun ispatı Ek 7.1'de verilmiştir.



Açıklama: i, örnek sayısı; **j**, grup sayısıdır. Örneğin, y_{ij}, j grubunun i'ninci değeri y demektir.

Şekil 7.1 Varyans analizinin geometrik ilkesi, açıklanan ve açıklanamıyan sapmalar.

Varyans analizinin Şekil 7.1'deki geometrik anlamı Çizelge 7.1'de matematiksel olarak özetlenmiştir.

Çizelge 7.1. Sapma kareleri ve hesaplama eşitlikleri (ispat için bak Ek 7.2).

Toplam sapma		**Açıklanan sapma**		**Açıklanamıyan sapma**
Açıklanan sapmaların karelerinin toplamı	=	gruplar arası sapmaların karelerinin toplamı	+	gruplar içi sapmaların karelerinin toplamı
KT$_{\mathbf{t}(\text{oplam})}$*		**KT$_{\mathbf{a}(\text{rası})}$**		**KT$_{\mathbf{i}(\text{çi})}$**
$\sum_{j=1}^{m}\sum_{i=1}^{n}(y_{ij}-\bar{y})^2$	=	$\sum_{j=1}^{m} n_j(\bar{y}_j - \bar{y})^2$	+	$\sum_{j=1}^{m}\sum_{i=1}^{n_j}(y_{ij}-\bar{y}_j)^2$

* **KT**: **K**areler **t**oplamı; i, örnek sayısı (i=1,2,..,n); j, grup sayısı (j=1,2,...,m); $\bar{y}_j$, grup; $\bar{y}$, genel ortalama değeridir.

Örneklem ortalamaları $\bar{y}_j$ ile toplam ortalama $\bar{y}$ arasındaki farklara ($\bar{y}_j - \bar{y} = e_i$) **etken** (effect) denir. Bu fark, dolayısı ile kareleri toplamı, ne kadar küçükse, grup veya örneklemler birbirine o kadar benziyor demektir. Buradan "örneklemlerin aynı ana kütleye ait olabilmeleri için etkenlerin kareleri toplamının sıfıra yaklaşması gerektiği" sonucu ortaya çıkar. Bu da **en küçük kareler yöntemi**nin geçerli olduğunu gösterir. Yani,

$$\sum_{i=1}^{n}\frac{e_i^2}{n} \to \min \qquad 7.2$$

olması gerekir. Bu, aynı zamanda hataların ve etkenlerin ortalama değerlerinin 0 olmasını gerektirir.

7.2 Tek değişkenli varyans analizi

Tek değişkenli varyans analizinde 1 tek öezelliğin değerleri, örneğin, ağırlıklar g olarak, karşılaştırılır. Varyans analizi ilke olarak F sınamasından başka bir şey değildir. Başka bir deyişle varyans analizi, F sınamasını kullanan karmaşık bir varsayım yöntemidir. Bu varsayımda, tek etkenli varyans analizinde (ANOVA=**an**alysis **o**f **va**riance), sıfır ve karşıt varsayımları,

H_o: Ortalamalar eşittir. Örneğin, $\mu_1 = \mu_2 = ... \mu_n$ 7.3

H_1: En az biri farklıdır. Öreğin, $\bar{y}_i \neq \bar{y}_j$

şeklinde ifade edilirler. Bununla grupların homojen olduğu savunulmaktadır (H_0, sıfır hipotezi = ortalama değerler arasında fark yoktur). Tersine seçenek varsayımına (H_1) göre belki de bir ana kütleye ait değildirler (= ortalama değerler arasında fark vardır). H_o'ın kabul veya reddedilmesi için farkın anlamlılığı F sınaması ile denetlenir.

Varyans analizi ile bir örneklemde birçok etkenin etkisi birden araştırılır. Grup sayısı normal koşullarda 10'u, örnek sayısı da 50'yı geçmez. j, örneklem (grup) sayısı olarak verilmiştir. Bu durumda j örneklemlerinin örnek sayıları n_1, n_2, ...,n_m olur. Buna göre,

x_i değerlerinin genel sayısı $n = n_1 + n_2 + ... + n_m$, 7.4

x_i değerlerinin genel toplamı $\Sigma x_i = x_1 + x_2 + ... + x_n$ 7.5

eder. Buradan,

Genel ortalama $\bar{x} = \Sigma x_i / n$, 7.6

Serbestlik derecesi $F = n-1$, 7.7

Toplam karelerin ortalama değeri $KO_t = KT_t / (n-1)$ 7.8

bulunur (KO_t, **t**oplam **k**areler **o**rtalaması; n-1, serbestlik derecesidir).

7.3 Sapmaların hesaplanması

Varyans analizinde sapmaların hesaplanması ve bunlarla ilgili işlemler yöntemin temelini oluşturur. Bunun için işleme girecek ölçümler ve gruplar ilk önce adlandırılır veya numaralandırılır. Örneğin, j gruplarının ortalama değerleri $\bar{x}_1$, $\bar{x}_2$,..., $\bar{x}_m$ ise,

1. **T**üm **k**arelerin **t**oplamı,

$$KT_t = (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + ... + (x_n - \bar{x})^2 , \quad 7.9$$

2. Kümeler **a**rası **k**areler **t**oplamı,

$$KT_a = n_1\,(\bar{x}_1-\bar{x})^2 + n_2\,(\bar{x}_2-\bar{x})^2 +\ldots+n_j\,(\bar{x}_j-\bar{x})^2 \qquad 7.10$$

ve

3. Kümeler **i**çi **k**areler **t**oplamı,

$$KT_i = KT_t - KT_a \qquad 7.11$$

olarak bulunur.

Değişke, ortalama karelerin sapması olarak tanımlandığından,

$$KO = \frac{KT}{gözlem.sayisı - 1} \qquad 7.12$$

eşitliğinden hesaplanır (Çizelge 7.1). Paydadaki sayı, serbestlik derecesi, gözlem sayısının 1 eksiğidir. Sapmaların hesaplandığı değerlerden ortalama değerin kendisi de hesaplandığından, değerlerin biri hiçbir zaman serbest olamaz ve toplam serbestlik derecesi,

$$F_t = n-1 \qquad 7.13$$

değerini alır. Aynı şekilde her gruptaki örneklerin de biri hiçbir zaman serbest olamıyacağından, gruplar arası serbestlik derecesi de,

$$F_a = m-1 \qquad 7.14$$

olur. Genel kareler toplamı KT_t, KT_a ve KT_i kısımlarına ayrıldığı gibi serbestlik dereceleri de kısımlara ayrılabilir. Buna göre, m sayıdaki grupların her birinin içindeki örneklerden ancak n-1 tanesi serbest hareket edebileceğinden, gruplar içi serbestlik derecesi,

$$F_i = n-1-(m-1)$$
$$= n-m \qquad 7.15$$

olur. Kareler toplamının (KT_t, KT_a ve KT_i) ortalama sapma değerlerinin hesaplamasında kullanılan formüller Çizelge 7.2'de toplanmıştır.

Çizelge 7.2. Kareler toplamı ortalamalarının hesaplanması.

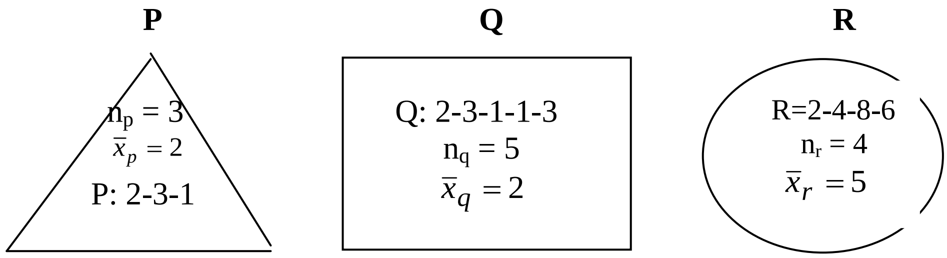
Toplam ortalama değişke,	$KO_t = \dfrac{KT_t}{n-1}$	7.16
Gruplar arasındaki ortalama değişke,	$KO_a = \dfrac{KT_a}{m-1}$	7.17
Gruplar içi ortalama değişke	$KO_i = \dfrac{KT_i}{n-m}$	7.18

KO_a ve KO_i'ye F testi uygulanarak anlamlı farklar bulunur (bak. Örnek 7.1).

Örnek 7.1

Aşağıda P, Q ve R şekilleri ile tanımlanan örneklemlerin aynı ana kütleye ait olup olmadıklarının gösterilmesi.

Verilen deney sonuçları:



Toplam grup ve örnek sayısı **m** = 3, $\mathbf{n_t}$ = 12

İstatistik yöntemlerinde hesaplamaları düzgün yürütmek ve sağlama olanağı vermek için örnekler kullanılacak formüllere göre düzenlenirler. Bir varyans analizinde verilerin hesaplamaya hazırlanmasına **faktöryel dizayn** (factorell desing*)* denir (bak. Çizelge 7.3.). Burada amaca uygun bir çizelge yapılarak işlemler yapılır. Bu şekil çözüm, Örnek 7.1 için Çizelge 7.3'te verilmiştir.

Çizelge 7.3. Örnek 7.1'in çözümü: Devamı için bak. Çizelge 7.4.

Grup/ Örneklem	**KT_t** $\sum_{j=1}^{m}\sum_{i=1}^{n}(y_{ij}-\bar{y})^2$	**KT_a** $\sum_{j=1}^{m} n_j(\bar{y}_j-\bar{y})^2$	**KT_i** $\sum_{j=1}^{m}\sum_{i=1}^{n_j}(y_{ij}-\bar{y}_j)^2$	
1	$(2-3)^2 = 1$ $(3-3)^2 = 0$ $(1-3)^2 = 4$	$(2-3)^2 = 1$ $(2-3)^2 = 1$ $(2-3)^2 = 1$	$(2-2)^2 = 0$ $(3-2)^2 = 1$ $(1-2)^2 = 1$	
2	$(2-3)^2 = 1$ $(3-3)^2 = 0$ $(1-3)^2 = 4$ $(1-3)^2 = 4$ $(3-3)^2 = 0$	$(2-3)^2 = 1$ $(2-3)^2 = 1$ $(2-3)^2 = 1$ $(2-3)^2 = 1$ $(2-3)^2 = 1$	$(2-2)^2 = 0$ $(3-2)^2 = 1$ $(1-2)^2 = 1$ $(1-2)^2 = 1$ $(3-2)^2 = 1$	
3	$(2-3)^2 = 1$ $(4-3)^2 = 1$ $(8-3)^2 =25$ $(6-3)^2 = 9$	$(5-3)^2 = 4$ $(5-3)^2 = 4$ $(5-3)^2 = 4$ $(5-3)^2 = 4$	$(2-5)^2 = 9$ $(4-5)^2 = 1$ $(8-5)^2 = 9$ $(6-5)^2 = 1$	
Toplam	**50**	**24**	**26**	

2 grup veya küme ile 9 örnek serbestlik derecelerine göre kuramsal değişkeler oranı,

$$F_{den}=F_{2;\,9;\,0,95}=F_{kur} \qquad 7.19$$

F sınaması çizelgesinden 4,3 olarak okunur (Ek 7.2). Buna göre farklılık anlamlı değildir. Çünkü bulunan deneysel 4,1 ($=F_{den}=F_{2;9;0,95}$) ile çizelgeden okunan kuramsal F_{kur} = 4,3 aynı grup ve örnek serbestlik derecelerine aittir. H_o varsayımının reddedilmesi için,

$$\mathbf{F_{2;9;0,95} < F_{kur}} \qquad 7.20$$

olması lazımdır. Dolayısı ile buradaki örneklemler homojendir ve bir ana kütleye aittirler. Bu sonuç, ortalama değerlerin karşılaştırılmasından da anlaşılmaktadır ($x_o=x_p=2$, $x_q=5$). Bu sonuçla H_o varsayımı kabul, H_1 reddedilmiştir.

İşlemin sağlanması: KT_t= ?$KT_a + KT_i$

$\rightarrow$ 50 = 24 + 26

ile çözüm sağlanmıştır.

Çizelge 7.4. Varyans analizinde kullanılan eşitliklerin özeti. Sayılar Örnek 7.1'deki sonuçlardır.

Değişke Kaynağı		Kareler toplamı	Serbeslik derecesi	Karelerin ortalaması	$F_{2;9;0,95}$ (F_{den})*
Toplam	**KT_t**	$\sum_{j=1}^{m}\sum_{i=1}^{n}(y_{ij}-\bar{y})^2 = 50$	$F_t = n-1 = 11$	$KO_T = \frac{KT_t}{n-1} = 4{,}54$	--
Gruplar arası	**KT_a**	$\sum_{j=1}^{m} n_j(\bar{y}_j - \bar{y})^2 = 24$	$F_a = m-1 = 2$	$KO_a = \frac{KT_a}{m-1} = 12{,}00$	--
Gruplar içi	**KT_i**	$\sum_{j=1}^{m}\sum_{i=1}^{n_j}(y_{ij}-\bar{y}_j)^2 = 26$	$F_i = n-m = 9$	$KO_i = \frac{KT_i}{n-m} = 2{,}90$	$\frac{KO_a}{KOi}$ =12/2,9=**4,1**

7.4 Çok değişkenli varyans analizi

İncelenen veri kümeleri birden fazla özellikleri bakımından incelenmek istendiğinde, çok değişkenli varyans analizi uygulanır. Bunun uygulanışı ilke olarak tek değişkenli varyans analizi ile aynıdır. Ancak daha karmaşık hesaplama yöntemlerinin izlenmesi gerekir. Bu nedenle örneğin çoğu yerde determinantlar kullanılarak satır, tabaka ve sütunlar (3 değişkenli) ayrı ayrı ele alınır. Burada bu yöntemlerin işlenmesi bu notların kapsamını aşar. Bu nedenle kaynakçadaki Tüysüz ve Yaylalı (2005), Schönwiese (1992), Backhaus ve diğ. (1990) ve Sachs (1984) eserlerine baş vurmaları önerilir.

Alıştırma 7.1

Aşağıdaki çizelgede verilen değerler için bir varyans analizi veya F sınaması hesaplayınız.

	Örneklemler			
i \ j	1	2	3	
1	3	4	8	
2	7	2	4	
3		7	6	
4		3		
n_i =	2	4	3	n_i = 9
$\bar{x}_i$ =	5	4	6	
Σx_i =	10	16	18	$\Sigma\,\Sigma x_i$= 44

Yanıt: H_o kabul; $F_{2;\,6;\,0,95}$ =0,69<5,14

* $F_{2;9;0,95}$: 2 grup, 9 serbestlik dereces ve % 95 anlam düzeyi (kesinlik derecesi) ile demektir.

EKLER

Ek 7.1 7.1 eşitliğinin ispatı:

$$\sum_{j=1}^{m}\sum_{i=1}^{n}(y_{ij}-\bar{y})^2=\sum_{j=1}^{m}\sum_{i=1}^{n}(y_{ij}\underbrace{-\bar{y}_j+\bar{y}_j}_{=0}-\bar{y})^2 \qquad 7.21$$

basite indirgemek için,

$$KT_t=\sum_{j=1}^{m}\sum_{i=1}^{n}(\underbrace{y_{ij}-\bar{y}_j}_{v_{ij}}+\underbrace{\bar{y}_j-\bar{y}}_{zj})^2 \qquad 7.22$$

$$=\sum_{j=1}^{m}\sum_{i=1}^{n}(v_{ij}+z_j)^2 \qquad 7.23$$

$$=\sum_{j=1}^{m}\sum_{i=1}^{n}(v_{ij}^2+2v_{ij}z_j+z_j^2) \qquad 7.24$$

$$=\sum_{j=1}^{m}\sum_{i=1}^{n}v_{ij}^2+\sum_{j=1}^{m}\sum_{i=1}^{n}2v_{ij}z_j+\sum_{j=1}^{m}\sum_{i=1}^{n}z_j^2 \qquad 7.25$$

$$=\sum_{j=1}^{m}\sum_{i=1}^{n}v_{ij}^2+2\sum_{j=1}^{m}(z_j\sum_{i=1}^{n}v_{ij})+n_j\sum_{j=1}^{m}z_j^2 \qquad 7.26$$

k_{ij} = y_{ij}- $\bar{y}_j$ bireysel hata olduğundan ve hataların toplamının 0 olması koşulu bulunduğundan,

$$=\sum_{j=1}^{m}\sum_{i=1}^{n}v_{ij}^2+2\sum_{j=1}^{m}(z_j\underbrace{\sum_{i=1}^{n}v_{ij}}_{=0})+n_j\sum_{j=1}^{m}z_j^2 \qquad 7.27$$

Buna göre eşitlik,

$$=\sum_{j=1}^{m}\sum_{i=1}^{n}v_{ij}^2+n_j\sum_{j=1}^{m}z_j^2 \qquad 7.28$$

$$=n_j\sum_{j=1}^{m}z_j^2+\sum_{j=1}^{m}\sum_{i=1}^{n}v_{ij}^2 \qquad 7.29$$

k_{ji} ve l_j'nin değerlerinin yerlerini konulması ile,

$$KT_t=n_j\sum_{j=1}^{m}(\bar{y}_j-\bar{y})^2+\sum_{j=1}^{m}\sum_{i=1}^{n_j}(y_{ij}-\bar{y}_j)^2 \qquad 7.30$$

$$=\underline{KT_a+KT_i} \qquad 7.31$$

bulunur.

Ek 7.2 F sınaması çizelgesi (P_{95}, α=0,05, Backhaus ve diğ., 1990).

i \ j	1	2	3	4	5	6	7	8	9	10
1	161,4	199,5	215,7	224,6	230,2	234,0	236,8	238,9	240,5	241,9
2	18,5	19,0	19,1	19,2	19,3	19,3	19,3	19,3	19,3	19,4
3	10,1	9,5	9,2	9,1	9,0	8,9	8,8	8,8	8,8	8,7
4	7,7	6,9	6,5	6,3	6,2	6,1	6,0	6,0	6,0	5,5
5	6,6	5,7	5,4	5,1	5,0	4,9	4,8	4,8	4,7	4,7
6	5,9	5,1	4,7	4,5	4,3	4,2	4,2	4,1	4,1	4,0
7	5,5	4,7	4,3	4,1	3,9	3,8	3,7	3,7	3,6	3,6
8	5,3	4,4	4,0	3,8	3,6	3,5	3,5	3,4	3,3	3,3
9	5,1	4,2	3,8	3,6	3,4	3,3	3,2	3,2	3,1	3,1
10	4,9	4,1	3,7	3,4	3,3	3,2	3,1	3,0	3,0	2,9
11	4,8	3,9	3,5	3,3	3,2	3,0	3,0	2,9	2,9	2,8
12	4,7	3,8	3,4	3,2	3,1	3,0	2,9	2,8	2,8	2,7
13	4,6	3,8	3,4	3,1	3,0	2,9	2,8	2,7	2,7	2,6
14	4,6	3,7	3,3	3,1	2,9	2,8	2,7	2,7	2,6	2,6
15	4,5	3,6	3,2	3,0	2,9	2,7	2,7	2,6	2,5	2,5
16	4,4	3,6	3,2	3,0	2,8	2,7	2,6	2,5	2,5	2,4
17	4,4	3,5	3,2	2,9	2,8	2,7	2,6	2,5	2,4	2,4
18	4,4	3,5	3,1	2,9	2,7	2,6	2,5	2,5	2,4	2,4
19	4,3	3,5	3,1	2,9	2,7	1,6	2,5	2,4	2,4	2,3
20	4,3	3,4	3,1	2,8	2,7	2,6	2,5	2,4	2,3	2,3
21	4,3	3,4	3,0	2,8	2,6	2,5	2,4	2,4	2,3	2,3
22	4,3	3,4	3,0	2,8	2,6	2,5	2,4	2,4	2,3	2,3
23	4,2	3,4	3,0	2,8	2,6	2,5	2,4	2,3	2,3	2,2
24	4,2	3,4	3,0	2,7	2,6	2,5	2,4	2,3	2,3	2,2
25	4,2	3,3	2,9	2,7	2,6	2,4	2,4	2,3	2,2	2,2
26	4,2	3,3	2,9	2,7	2,5	2,4	2,3	2,3	2,2	2,2
27	4,2	3,3	2,9	2,7	2,5	2,4	2,3	2,3	2,2	2,2
28	4,2	3,3	2,9	2,7	2,5	2,4	2,3	2,2	2,2	2,1
29	4,1	3,3	2,9	2,7	2,5	2,4	2,3	2,2	2,2	2,1
30	4,1	3,3	2,9	2,6	2,5	2,4	2,3	2,2	2,2	2,1
40	4,0	3,2	2,8	2,6	2,4	2,3	2,2	2,1	2,1	2,0
60	4,0	3,1	2,7	2,5	2,3	2,2	2,1	2,1	2,0	1,9
120	3,9	3,0	2,6	2,4	2,2	2,1	2,0	2,0	1,9	1,9
∞	3,8	3,0	2,6	2,3	2,2	2,1	2,0	1,9	1,8	1,8

i: Açıklanabilen değişkenin serbestlik derecesi (n-1), **j**: Paydanın serbestlik derecesi (n-m).

Ek 7.3 F sınamsı değerleri (P_{99}, α=0,01, Backhaus ve diğ., 1990.).

i \ J	1	2	3	4	5	6	7	8	9	10
1	4052	4999,50	5403	5625	5764	5859	5928	5982	6022	6056
2	98,50	99,00	99,17	99,25	99,30	99,33	99,36	99,37	99,39	99,40
3	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,35	27,23
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55
5	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05
6	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87
7	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62
8	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81
9	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26
10	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85
11	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30
13	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10
14	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69
17	8,40	6,11	5,18	4,67	4,34	4,10	3,93	3,79	3,68	3,59
18	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51
19	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37
21	8,02	5,78	4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31
22	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26
23	7,88	5,66	4,76	4,26	3,94	3,71	3,54	3,41	3,30	3,21
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17
25	7,77	5,57	4,68	4,18	3,85	3,63	3,46	3,32	3,22	3,13
26	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,18	3,09
27	7,68	5,49	4,60	4,11	3,78	3,56	3,39	3,26	3,15	3,06
28	7,64	5,45	4,57	4,07	3,75	3,53	3,36	3,23	3,12	3,03
29	7,60	5,42	4,54	4,04	3,73	3,50	3,33	3,20	3,09	3,00
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80
60	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63
120	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56	2,47
∞	6,63	4,61	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32

i: Açıklanabilen değişkenin serbestlik derecesi (n-1), **j**: Paydanın serbestlik derecesi (n-m).

8 BAĞINTI (KRELASYON) ANALİZİ

8.1 Genel bakış

Bir anakitlede veya örneklemde birden fazla özelliğin veya değişkenin incelenmesi ile özelliklerden birinin diğerine bağlı olarak değişip değişmediği de araştırılabilir. Bu değişim örneğin, bir matriks şeklinde satır ve sütunlarla ifade edilebilir. Örneklemin bir insan çevresi olması halinde, bu çevre bireylerinin x_i boyları ile y_i kilolarının karşılaştırılması gibi.

Bu şekilde yapılan karşılaştırmalara **bağıntı** (korelasyon), inceleme yöntemine de bağıntı analiz yöntemi denir ve en çok kullanılan istatistiksel analiz yöntemlerindendir. Uygulanması oldukça esnektir. Değişken arasındaki ilişkilerin saptanmasında, yorumlanmasında, değişken değerlerinin tahmin edilmesinde ve beklenen gelişmelerin önceden kestirilmesinde önemli rol oynar. Örneğin, bir malın fiyatının satışına etkisi veya bir kare alanının değişen a kenar uzunluğu ile değişmesi ($F = a^2$) v. s. gibi.

Varyans analizi ve χ^2 yöntemleri ile de bağıntıların olup olmadığı araştırılır. Ancak bu yöntemler birer fonksiyon değildir. Bunlara karşın bağıntı ve **bağınım** (regresyon) analiz yöntemlerinde birkaç veri kümesi arasında bağıntıların varlığı gösterilebidiği gibi, **fonksiyon ilişkileri** de araştırılabilmektedir.

Doğada birçok değişken, örneğin, yoğunlukla hacim, ısı ile basınç veya element derişimleri, birbirine bağımlıdır. Korelasyon, kelime anlamı olarak, ilişki, bağıntı demektir. Yöntem, bir **x değişkeninin** (bağımsız değişken) alabildiği değerler dizisinin oluşturduğu dizi ile bir **y değişkeninin** (bağlı değişken) alabildiği değerlerin oluşturduğu bir bağıntının olup olmadığını araştırır ve bağıntı varsa, bunun derecesini saptar. Bağlı ve bağımsız değişkenler yer değiştirebilir. Değişkenler arasındaki bağıntı, **bağıntı katsayısı r** ile gösterilir ve bağıntı analiz yönteminde bağıntının derecesini belirler. Bağınım analizinde ise, sadece fonksiyonal bağımlılık incelenir. Bağıntı ve bağınım birbirlerine sıkı sıkıya bağlıdırlar. Bağınım analizinde hem bağımlı, hem de bağımsız değişkenin **metrik ölçüde** olması gerekir. Bunun dışındaki ölçüler (biner veya nominal) istisnadır.

8.2 Bağıntı analizi (BA) ve çeşitleri

Bağıntı analizi, değişkenler arasındaki ilişkileri inceler. Değişkenler, ana kütlelerin, öncelikle derişim, çekim kuvveti, ısı ve basınç gibi özellikleri olmaktadır. Bu özellikler arasındaki bağıntılar çok çeşitlidir. Bir tahmine işaret ederler ve değişkenler arasındaki neden ve sonuç ilişkisine dayanırlar. Doğrusal bağıntı, yukarıda da belirtildiği gibi, en yaygın gözlenen bağıntı şeklidir, doğada ve günlük yaşamda örneklerine sık rastlanır. Bunların bazıları aşağıya çıkarılmıştır (Şekil 8.1). Bunun yanında üslü fonksiyonların da çeşitli tipleri sıkça görülür. Gözlemler arasındaki ilişkinin tanımlanması için önce sıklık dağılımlarının homojenliğinin bilinmesi gerekir. Çünkü bunlar değişik özelliklere sahip ana kütlelere ait olabilirler.

Bir iki boyutlu bağıntı genelde aşağıdaki özellikleri taşır:

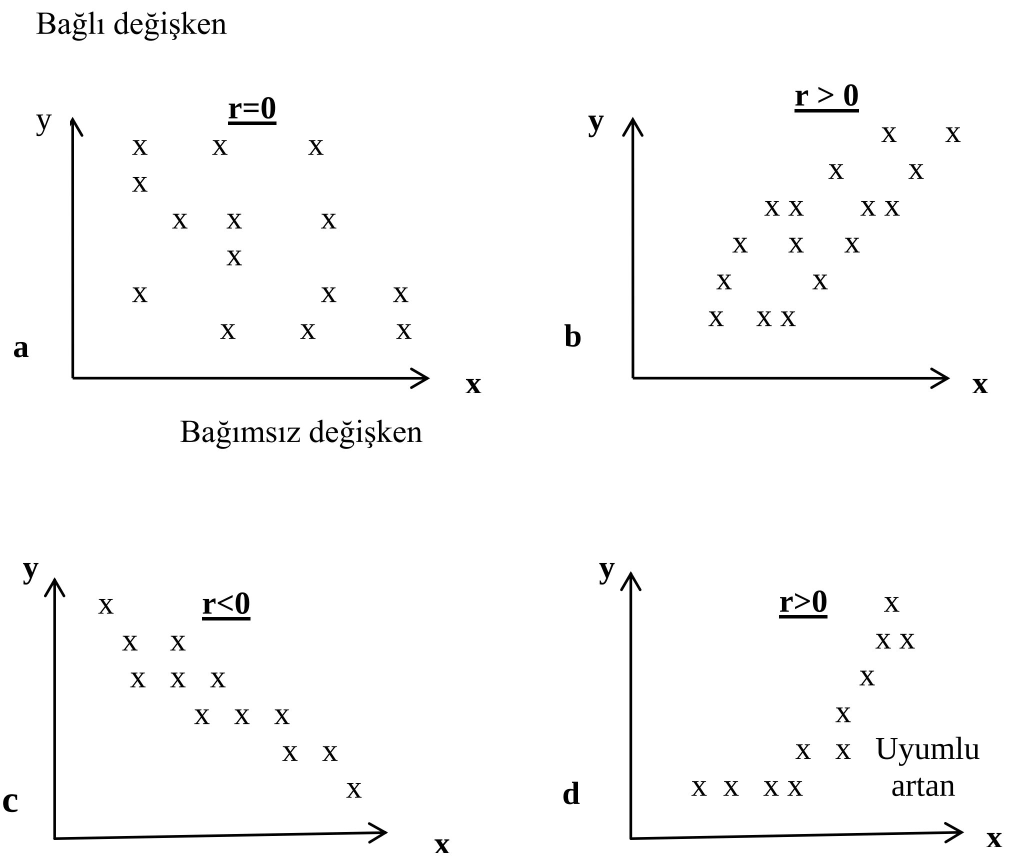
1.İki değişken istatistiksel olarak birbirine bağlı değildir (Şekil 8.1a). Bu durumda veriler birbirinden **bağımsız** veya ilişkisizdir ($r = 0$). Değişkenlerden birinin sabit, diğerinin değişmesi durumunda da bağıntı olmaz.

2. İki değişken tamamen veya kısmen birbirine bağlıdır (Şekil 8.1 b ve c). Bağlılık doğrusaldır (doğrusal bağıntı). Bu bağıntıda değişkenler eşit oranlarda artarlar (Şekil 8.3), ancak bağınım doğrusunun eğimi,

a) **pozitif** olabilir ($r > 0$). Bu durumda değişkenlerden biri artarken diğeri de artar demektir (uyumlu veya pozitif bağıntı). $r = +1$ değeri en **uyumlu** bağıntıyı gösterir ve tüm noktalar 1 doğru üzerindedir.

b) **negatif** olabilir ($r < 0$). Bu durumda değişkenlerden biri artarken diğeri azalır demektir (uyumsuz veya negatif bağıntı). $r = -1$ en **uyumsuz** bağıntıyı gösterir.

3. Bağınım doğrusal değildir (doğrusal olmayan bağıntı). Burada artan veya azalan bir üssel bağıntı mevcuttur (Şekil 8.1 d). Bu tür bağıntılarda değişkenlerden biri bir ölçü birimi artarken, diğeri değişik ölçülerde artar veya azalır.



a, Bağıntı yok, r = 0; **b**, uyumlu doğrusal bağıntı, r > 0, **C**, uyumsuz doğrusal bağıntı, r < 0 ve **d**, üssel uyumlu bağıntı, r > 0.

Şekil 8.1. Önemli bağıntı çeşitleri.

8.2.1 Bağıntı katsayısı (r, BK)

Bağıntı derecesini belirliyen bağıntı katsayısı r, ancak,

$$-1 < r < +1 \qquad 8.1$$

değerlerini alabilir. En iyi bağıntı değeri r,

$$r = |1|$$

olduğu zamandır (bak. yukarıya). Bunun arasındaki değerler zayıf, ancak anlamlı olabilirler.

Bir bağıntı verilerinin bağımsız ve normal dağılmış olması lazımdır. Bu, bağıntının ön koşuludur. İki değişken arasındaki doğrusal bağıntı derecesi r, matematiksel olarak,

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \qquad 8.2$$

$$= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{x})^2}}$$

$$= \frac{\sum_{i=1}^{n}x_i y_i - \frac{1}{n}(\sum_{i=1}^{n}x_i)(\sum_{i=1}^{n}y_i)}{\sqrt{[\sum_{i=1}^{n}x^2 - \frac{1}{n}(\sum_{i=1}^{n}x)^2][\sum_{i=1}^{n}y^2 - \frac{1}{n}(\sum_{i=1}^{n}y)^2]}}$$

şeklinde ifade edilmektedir (n-1 terimleri kısaltılmıştır). Görüldüğü gibi r, **katışık değişke** (kovaryans) s_{xy} ile değişkenlerin standart sapmaları olan s_x ve s_y'den veya bunların ham değerlerinden hesaplanmaktadır. Öneminden dolayı katışık değişke aşağıda açık şekliyle verilmesinde yarar vardır:

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}x'.y'$$

$$= \frac{n\sum_{i=1}^{n}x_i y_i - \sum_{i=1}^{n}x_i . \sum_{i=1}^{n}y_i}{n(n-1)} \qquad 8.3$$

gibi.

Görüldüğü gibi katışık değişke, çarpanları eşit olmıyan değişkedir, yani $(x_i-\bar{x})$ $(x_i-\bar{x})$ yerine $(x_i-\bar{x})$ $(y_i-\bar{y})$ şeklidir.

Bağıntı analizinde az sayıdaki örnek için yüksek r istenirken, çok sayıdaki örnek için düşük r değerleri de anlamlı sonuç verebilir. Yani r'nin belirlilik sınırları artan örnek sayısı ile azalır. Örnek sayısına veya serbestlik derecesine bağlı olarak hesaplanmış belli bir güvenirlik derecesini, P=% 90, 95, 99 gibi, ifade eden r değerleri hazır çizelgelerden alınabilir (Ek 8.1).

Bağıntı katsayısı r bir örneklemin tüm örneklerine ait olduğundan, örneklerin sadece tümü için geçerlidir. Ayrıca karşılaştırılacak veya ortalamaları alınacak iki örneklemin bağıntı katsayısının yaklaşık aynı sayıdaki örneğe ait ve aynı birimde olmaları gerekir. Çünkü belirlilik örnek sayısına göre çabuk değişir. Ancak bu ve buna benzer durumlar gözönüne alınarak bir bağıntı analizinin yapılacak yorumu, gerçekçi ve objektif olabilir.

8.2.2 Belirlilik katsayısı (r^2)

Bir veri kümesindeki toplam sapma veya değişke açıklanabilen ve açıklanamıyan değişkelerden meydana gelir. Buna göre toplam değişke veya sapma,

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y})^2 \qquad 8.4$$

Toplam sapma = Açıklanan sapma + açıklanamıyan sapma

terimlerinden oluşur (Şekil 8.2). Buradan belirlilik katsayısı r^2, **açıklanabilen** saçınımın**, toplam saçınıma oranı** olarak,

$$r^2=\frac{Açiklanan.sapma}{Toplam.sapma}$$

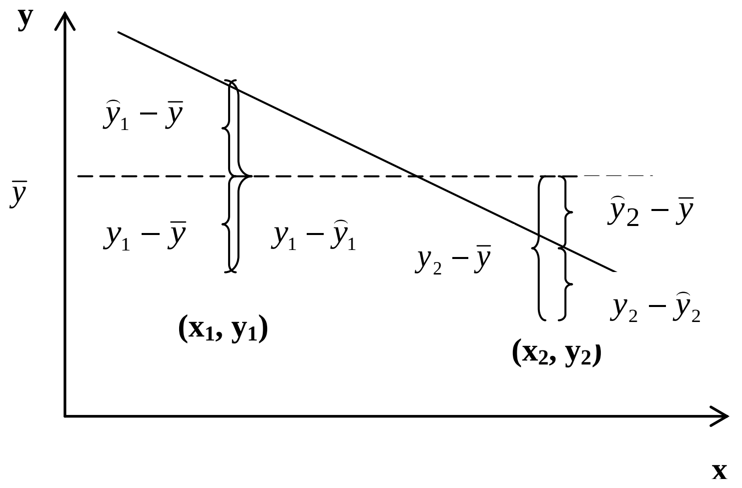
$$r^2=\frac{\sum_{i=1}^{n}(\hat{y}-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \qquad 8.5$$

ifade edilir. Açıklanan sapma oranı ne kadar yüksekse, r^2 de o kadar büyük olur. Dolayısı ile r^2 en çok 1 değerini alır ve saçınımların tümünün açıklandığı anlamına gelir. Aksine r^2=0'dir.

Belirlilik katsayısı r^2,

$$r^2=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \qquad 8.6$$

şeklinde, açıklanamıyan sapmanın toplam sapmaya oranının en yüksek değeri olan 1'den çıkarılması ile de hesaplanabilir. Buradaki çevirme işlemi hesaplamayı kolaylaştırmaktadır. 1 ile olan **fark**, toplam değişkenin açıklanan, yani bağınım doğrusu tarafından kapsanan, **sistematik** etkenlerden kaynaklanan belirlilik kısmını, **geri kalan** kısmı da açıklanamıyan, **rastlantısal** etkenlerden kaynaklandığını gösterir.



y_i = Bağlı değişkenin i gözlem değeri, i = 1, 2,...,n

x_i = Bağımsız değişkenin i gözlem değeri,

$\hat{y}$ = Bağlı değişkenin okunan (kestirilen) değeri,

$\bar{y}$ = Bağlı değişkenin ortalama değeri

Şekil 8.2. Sapmaların geometrik olarak elemanlarına ayrılması.

r^2, istatistiksel açıdan r'den daha anlamlıdır (% r^2 = r^2.100). Açıklanan saçınımın toplam saçınıma oranını, yani **gerçek değerden spma** oranını, ifade eder. Logaritmik veya yüksek dereceli fonksiyon özellikleri için daha önemlidir.

Hesaplanan bir r değeri çizelgelerdeki (Örnek 8.2) veya grafikten (Wellmer, 1989) okunan değerlerle karşılaştırılarak anlamlılığı saptanır. Bu en kolay yoldur. Bulunan bir bağıntı katsayısının anlamlılığı t veya F sınaması ile de saptanabilir. Örnek sayısının 50'den az (n<50) olması durumunda t sınamasında belirlilik,

$$t = \frac{r_{yx}}{\sqrt{(1-r^2)}}\sqrt{n-2} \qquad 8.7$$

eşitliği ile sağlanır (n-2, serbestlik derecesi). Burada verilen anlam düzeyi için hesaplanan deneysel t_{den} değerinin kuramsal t_{kur}'dan büyük ($t_{den}>t_{kur}$) olmsı gerekir (H_o= red, H_1 = kabul, bak. Örnek 8.1).

Varyans analizinde olduğu gibi bağınım analizinde de belirlilik derecesinin bulunmasında kullanılacak değişkenler kolay ve sistematik hesaplama amacı ile **faktöriyel dizayn** edilir (Çizelge 8.5). Basit bir hesaplama için gerekli değişkenler Çizelge 8.1'de derlenmiştir:

Örnek 8.1.

Altın (g Au) satışının (talebinin) fiyatına (YTL/g) etkisi* (bak. Şekil 8.5).

Verilerin hazırlanması

1	**2**		**3**	**4**	**5**
Satış No,	**Değerler**		**$(x_i-\bar{x})^2$**	**$(y_i-\bar{y})^2$**	**$(x_i-\bar{x})(y_i-\bar{y})$**
i	**x_i, (g Au)**	**y_i, (YTL/g Au)**	**s^2_x**	**s^2_y**	**s_{xy}**
1.	30	47	82,81	88,36	85,54
2.	25	41	16,81	11,56	13,94
3.	21	37	0,01	0,36	-0,06
4.	25	46	16,81	70,56	34,44
8.	11	27	98,01	112,56	104,94
6.	16	32	24,01	31,36	27,44
7.	11	27	98,01	112,36	104,94
8.	17	35	15,21	6,76	10,14
9.	27	44	37,21	40,96	39,04
10.	26	40	26,01	5,76	12,24
$\sum$ n = 10	**209,00**	**376,00**	**414,90**	**480,40**	**432,60**

Çizelge 8.1'deki verilerin değişkenleri:

$\bar{x}$ = 209:10 $\qquad$ $\bar{y}$ = 376:10

= 20,90 g Au (bağımsız değişken) $\qquad$ = 37,60 YTL/g Au (bağlı değişken)

$s^2_x = 414{,}90:(10-1)$ $\qquad$ $s^2_y = 480{,}40:(10-1)$

$= 46{,}10$ $\qquad$ $= 53{,}38$

$s_x = \sqrt{46{,}10}$ $\qquad$ $s_y = \sqrt{53{,}38}$

= 6,79 g Au $\qquad$ = 7,31 YTL/g

* Hindistan'da örneğin, düğün dönemlerindeki altın talebi fiyatların artmasına neden oluyor (güz, 2007)

s_{xy} = 432,60:9 $\quad$ $s_{yx} = s_{xy}$

= 48,07

$\mathbf{r}_{xy} = s_{xy}:(s_x.s_y)$

= 48,07:6,79.7,31

= 48,07:49,63

= **0,969**

bulunur. Buna göre belirlilik katsayısı $0,969^2$ =0,939 eder. Buradan,

$$t = \frac{r.\sqrt{n-2}}{\sqrt{1-r^2}}$$

$$= \frac{0,969.\sqrt{10-2}}{\sqrt{1-0,939}}$$

$$= \frac{0,969.\sqrt{8}}{\sqrt{0,061}}$$

= 0,969.2,828:0,247

= 2,740:0,247

= 11,09

elde edilir.

Çizelge 3.1'de, F=8 (n-m) satırı ve P_{99} sütunundan t_{kur}=3,36 okunur (t_{kur}>3,36: red bölgesi). $t_{den}=t_{8;0,99}$ (11,09)>t_{kur} (3,36)'dır. Böylece H_o varsayımı reddedilmiştir (y, x'e bağlıdır). Yani bu değer % 99 olasılıkla (güvenle) anlamlıdır veya x ve y arasında geçerli bir bağıntı vardır. Sistematik ve rastlantısal hata oranlarının saptanması için bağınım doğrusu denkleminin bulunması gerekir (bak. 8.3.2 ve Örnek 8.2). t hesaplamasına gerek bırakmıyan anlamlı minimum r değerleri Ek 8.1'de verilmiştir.

8.3 Bağınım (regresyon) analizi (BaA)

Bağıntı ve bağınım analiz yöntemleri çok önemli bir araştırma yöntemidir ve çok yararlı sonuçlar elde edilebilir. Örneğin, benzer ilişkiler, çiftlerin değişim ve davranışları ile saklı bazı özelliklerin keşfedilmesi hakkında önemli bilgiler sağlanabilir. Bağınım doğrusunun eğimi de değişimin değerlendirilmesinde kullanılabilir ve bağıntılar incelenebilir.

8.3.1 Bağınım doğrusu (BD)

Bir bağıntı analizinde genelde bağlı değişkenle bağımsız değişken arasında bir **doğrusal ilişkinin** bulunduğu kabul edilir. Doğrusal ilişki demek, değişkenlerin sabit oranda değişmesi demektir (Şekil 8.3). Yani,

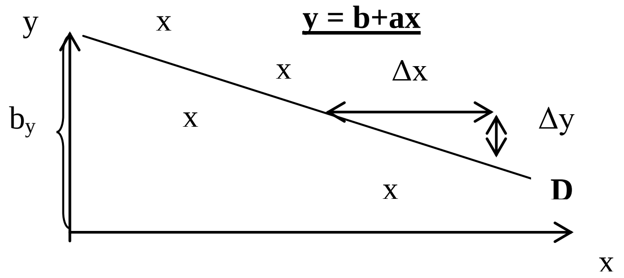
$$\frac{\Delta y}{\Delta x} = sabit \qquad 8.8$$

ifadesi geçerlidir. Bu sabite, bağınım doğrusunun absisle yaptığı açının eğimidir.

Bağıntı analizi, bir örneklemde bağlı ve bağımsız değişken arasında bir ilişkiyi saptamak mecburiyetindedir. Bu eşitlik,

$$y = b + ax \qquad 8.9$$

şeklinde bir doğru denklemidir. Burada **y**, bağlı değişken; **x**, bağımsız değişken; **a**, **bağınım doğrusunun eğimi** (a = tg α) ve **b**, doğrunun y ekseni üzerinde kestiği parça, **intersept**tir. a'ya **bağınım katsayısı** da denir.



x: Bağımsız değişken
y: Bağlı değişken
b: D doğrusunun y ekseni üzerinde kestiği parça, intersept
Δy/ Δx : D bağınım (regresyon) doğrusunun eğimi, a = tg α

Şekil 8.3. Bağınımın geometrik anlamı.

Bağınım analizinin amaçları,

1. Verilen eşitliğin katsayılarını bulmak (Örnek 8.2) ve
2. Örneklemde saptanan bağıntının örnek evreni için de geçerli olup olmadığını incelemektir.

Bağınım (regresyon), iki değişken arasındaki bağlılığın şeklini ifade eden bir terimdir. Yani bir örneklemin 2 veri özelliği veya bağımsız değerleriyle çalışılır (bak. Şekil 8.3). Burada regresyonun öztürkçe karşılığı **bağınım** kavramı kullanılacaktır. Bu bağlılığı gösteren eğriye de bağınım eğrisi denir. Bağınım, geometrik bir ifadeye sahip olması nedeniyle, özel önem taşır.

8.3.2 Kalıntı değerler (e_i)

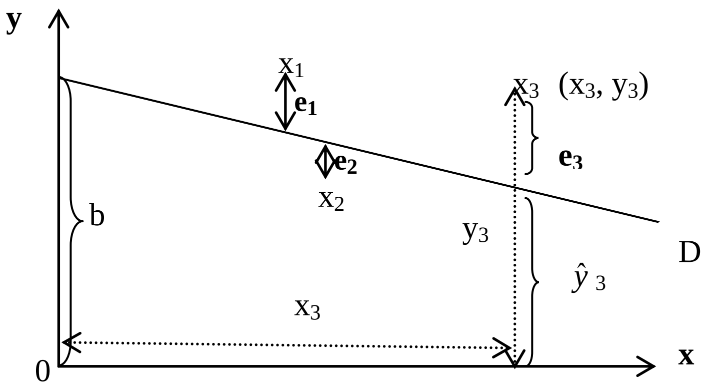
Bağıntı analizi ile bağınım analizi birbirine sıkı sıkıya bağlıdır. Bağınım, bağıntı analizi ile elde edilen nokta bulutuna en uygun doğrunun uyarlanmasıdır. Bu işlemde her nokta uyarlanan doğruya belli mesafede olur. Ölçüm değerlerinin bağınım doğrusuna uzaklıklarına **kalıntı** (residual) değerler denir. Bunlar Şekil 8.4'te e_i ile gösterilmiştir. Değerlerin bağınım doğrusuna olan uzaklıklarının karelerinin toplamı $0<\Sigma e_i^2<1$ arasında değişir. Başka hiçbir doğruya olan uzaklıklar bağınım doğrusuna olan uzaklıkların toplamından küçük olamaz. Bir dağılım ne kadar uyumlu ise, kalıntı değerler o kadar küçük olur. Bunların büyüklüğünden bağınım doğrusunun noktalara uyumu belirlenir. En iyi uyumda, yani noktaların doğru üzerinde bulunması durumunda, e_i^2 değerlerinin toplamı min veya 0 olur. Bu duruma **en küçük kareler toplamı**, yönteme de en küçük kareler yöntemi denir. Tüm hesaplamalarda e_i^2 değerlerinin toplamının mümkün olduğu kadar küçük olması istenir.

Bağınım analizinin ojektif fonksiyonu,

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left[y_i - (b + a x_i) \right]^2 \rightarrow \min \qquad 8.10$$

fonksiyonudur. Eşitlikte,

e_i = i'ninci gözlemin kalıntı değeri, i = 1, 2, ..., n
y_i = Bağlı değişken ($\hat{y}_i$, doğru üzerinde okunan, kestirilen değerdir)
b = Bağınım doğrusunun sabitesi, intersept
a = Bağınım sabitesi, D doğrusunun eğimi, a = tg α
x_i = Bağımsız değişken
n = Gözlem sayısı



e_3 = Bağlı özellik kalıntı (residüal) değeri, $e_i = y_i - \hat{y}_i$

x_3 = Bağımsız özellik büyüklüğü

y_3 = Bağlı özellik büyüklüğü

$\hat{y}$ = Bağınım doğrusuna göre x_3'ün okunan (kestirilen) y değeri

Şekil: 8.4. Kalıntı değerlerinin geometrik açıklanması.

Bu koşulu Şekil 8.3 için sağlıyan doğru,

$$\hat{y}_i = b - a\,x_i \qquad 8.11$$

bağınım doğrusudur. Ancak bu doğru veya eşitlik vasıtası ile x'e bağlı ve D doğrusu üzerindeki $\hat{y}_i$ değerleri kestirilebilir (bak. Örnek 8.2).

8.3.3 Bağınım doğrusunun değişkenlerinin hesaplanması

Bağınım doğrularının çeşitli değişkenlerinin hesaplanması için standart sapma, değişke ve katışıksız değişke gibi temel verilere gereksinim vardır. Bunlar burada yinelenmiyecektir, ancak ilişkileri üzerinde kısaca durulmasında yarar vardır.

Bağınım doğrusu değerlerin sapmaları konusunda bilgi vermezler. Ancak bir bağınım doğrusu % 95 kesinlikle $y = \pm 3s_y$ aralığına düşer (s, standart sapma). Doğrunun **eğimi**nden (a) değişkenlerin birbirine ne kadar bağlı olduğu bulunabilir. Örneğin, a ile r arasında şu ilişkiler bulunmaktadır:

$a \leq 0$ için, $r \leq 0$,

$a > 0$ için, $r > 0$ ve

$a = s_y/s_x$ için de $r = 1$

değerini alır. **r**, aynı zamanda bağınım doğrusunun **apsisle** yaptığı açının **cos**'dür. Şekil 8.5'te bağımsız ve bağlı değişkenlerin yer değiştirmesi halinde ortaya çıkacak durumları göstermektedir. Buna göre alt simgeler xy, **önce bağlı değişken** gelecek şekeilde bağlı değişkene göre yer değiştirir. Örneğin, y=b+ax eşitliği için a_{xy} şeklinde verilir. Dolayısı ile **D** bağınım doğrusunun eğimi,

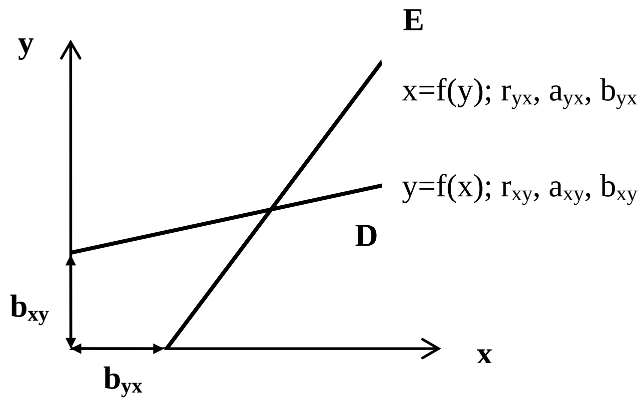
$$a_{xy} = s_{xy} / s_y^2 \qquad 8.12$$

$$= \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum y_i}{n\sum y_i^2 - (\sum_{i=1}^{n} y_i)^2} \qquad 8.13$$

değerine sahiptir ($s_{xy}=s_{yx}$'dir). Buna göre a'nın r cinsindeki değeri,

$$a_{xy} = r\,\frac{s_x}{s_y} \qquad 8.14$$

şeklini alır.



Şekil 8.5: Bağlı değişkenin değişimi.

Bunun gibi **kesim noktası** da r cinsinden hesaplanabilir. Örneğin,

$$b_{xy} = \bar{y} - a_{xy}\bar{x} \quad 8.15$$

$$= \bar{y} - r\frac{s_y}{s_x}\bar{x}$$

gibi. Doğru denklemindeki sabiteleri **bağlı ve bağımsız değişkenler cinsinden hesaplamak** mümkündür. Bağınım D ve E doğrularının y ve x eksenlerine göre eşitlikleri sırası ile,

$$y = f(x) \quad 8.16$$

$$= b+ax \quad 8.17$$

veya

$$x = f(y) \quad 8.18$$

$$= b+ay \quad 8.19$$

şeklindedir (bak. Şekil 8.5). Bu durumda bağıntı katsayısı r ve katışık değişke değişmez. Değişkenlerin değerleri aynı olan 2 doğru çakışır. Bilindiği gibi koşut doğruların eğimleri eşit, dik doğruların eğimleri ise, birbirinin negatif tersidir ($a_{xy}=-1/a_{yx}$). Bu işlemlerde tüm formüller ordinata uyarlanır. Buna göre,

$$r_{xy} = r_{yx} \quad 8.20$$

ve

$$a_{xy} . a_{yx} = r^2 \quad 8.21$$

olur. Yani |r|= 1 için,

$$a_{xy} \leq 1/a_{yx} \quad 8.22$$

değerinde olur.

Benzer şekilde yukarıdaki x=b+ay E doğrusunun kesme noktası,

$$b_{yx} = \bar{x} - a_{yx}\bar{y} \quad 8.23$$

$$= \bar{x} - r\frac{s_x}{s_y}\bar{y} \quad 8.24$$

şeklini alır (bak. ayrıca 8.15).

E doğrusunun r cinsinden eğimi,

$$a_{yx} = r.s_y/s_x \quad 8.25$$

olduğu görülür.

Dolayısı ile y eşitliği (8.16 eya 8.17),

$$y = b+(r.s_x/s_y)x \qquad 8.26$$

ve

$$= \bar{y} + r.\frac{s_x}{s_y}(x_i - \bar{x}) \qquad 8.27$$

genel bağınım doğrusuna çevrilebilir. Bu eşitlikteki x_i değerleri vasıtası ile, örneğin E doğrusu üzerindeki $\hat{y}_i$ değerleri, önceden yaklaşık kestirilebilir.

8.4 Sonuçların sağlanması

Bağıntı ve bağınım analizlerinde sonuçların sağlanması önemli bir yer tutar. Yapılan uzun ve karmaşık işlemler sıkça hatalara neden olur. Bunları önlemek için bir çizelge hazırlanır (faktöriyel dizayn) ve hesaplanan sonuçlar buraya aktarılarak,

$$\mathbf{KT_t = KT_b + KT_s} \qquad 8.28$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y} - \bar{y})^2 + \sum_{i=1}^{n}(\hat{y} - y_i)^2$$

formülünün sağlanması gerçekleştirilir. Çizelge 8.1 buna bir örnek oluşturmaktadır.

Çizelge 8.1: Örnek 8.1 değerleri için değişkelerin hesaplanması (ANOVA: analayses of variance) faktöriyel dizaynı.

1. Değişim kaynağı	**2. Kareler toplamı**	**3. Serbestlik dercesi**	**4. Kareler toplamı ortalaması**	**5. t sınaması sonucu**
2. Toplam değişim	KT_t	n-1	KO_t	
3. Doğrusal bağınım*	KT_b	m-1	KO_b	
4. Toplam sapma	KT_s	n-2	KO_s	$\mathbf{KO_b/KO_s = t_{den}}$

***KT**, **k**areler **t**oplamı, KO, **k**areler **o**rtalaması; b, bağınım; s, sapma ve t, toplam demektir

Doğrusal bağınımın ortalamasının toplam sapma ortalamasına oranı, $KO_b/KO_s = t_{den}$ sonucunu verir. Bu değer t veya F dağılım çizelgesindeki teorik değerlerle karşılaştırılır (bak. Çizelge 3.1). Bunun için önce bir kesinlik derecesi, örneğin, P=%99, seçilir. Bundan sonra F serbestlik derecesinin (n-m) satırına karşı gelen P sütununun (olasılık derecesi) kuramsal t_{kur} değeri okunur ($t_{F;\,P;}$) ve t_{den} ile karşılaştırılır. $t_{den} > t_{kur}$ durumunda, H_o kabul, $t_{den} < t_{kur}$ olduğunda da red edilmiş olur (bak. Örnek 8.2). Aynı şey determinantlarla da yapılabilir.

Örnek 8.2

Örnek 8.1'deki dağılımın bağınım katsayıları ve doğru denklemi (devam):

a) Doğru denklemi,

$$y = b + a\,x \quad \text{ise,}$$

$$a_{xy} = s_{xy}/s_y^2 \qquad\qquad b_{xy} = \bar{y} - a_{xy}\bar{x}$$

$$= \frac{48{,}07}{(6{,}79)^2} \qquad\qquad = 37{,}60 - 1{,}042.20{,}90$$

$$= 1{,}042 \qquad\qquad = 15{,}83 \text{ g Au}$$

bulunur. Böylece, y = b+ax doğrusunun denklemi (Şekil 8.4),

$$y = \underline{15{,}83 + 1{,}04\,x}$$

veya

$$\underline{\text{YTL/g Au} = 15{,}83 + 1{,}04 \text{ g Au}}$$

bulunur (Şekil 8.6).

Eksenlerin değişmesi durumunda E doğrusu için,

$$\underline{x = -12{,}94 + 0{,}90\,y}$$

$$\underline{\text{g Au} = -12{,}94 + 0{,}90 \text{ YTL/g Au}}$$

doğru eşitliği bulunur (a_{yx}=48,07/7,31^2)=0,90; b_{yx} =20,90-0,90.37,60=20,90-33,84=-12,94).

Aynı şekilde $r^2 = r_{xy}.r_{yx}$ ve $\hat{x}_i = \bar{x} + r\dfrac{s_x}{s_y}(y_i - \bar{y})$ olduğu görülür ($\hat{x}$, doğru üzerinde okunan değerdir).

b) Değişkelerin hesaplanması

Bağınım değerlerinin hesaplanması (bak. Örnek 8.1, faktöryel dizayn).

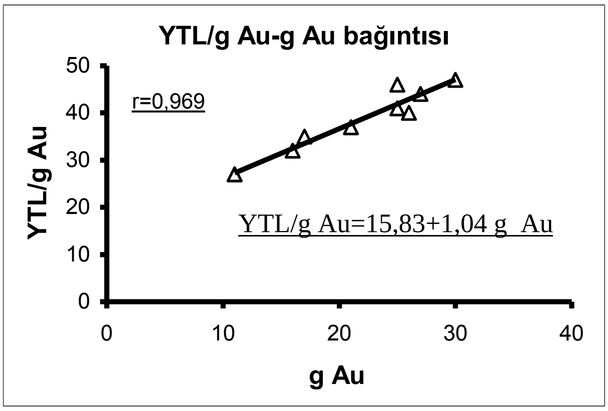
i	x_i	y_i	$\hat{y}_i$ =15,8+1,04x	$\bar{y}$	KT$_b$ = $\sum_{i=1}^{n}(\hat{y}-\bar{y})^2$	KT$_s$ = $\sum_{i=1}^{n}(\hat{y}-y_i)^2$	KT$_t$ = $\sum_{i=1}^{n}(y_i-\bar{y})^2$
1	30	47	47,12	37,60	90,63	0,01	88,36
2	25	41	41,94	37,60	18,83	0,88	11,56
3	21	37	37,77	37,60	0,03	0,59	0,36
4	25	46	41,94	37,60	18,83	16,49	70,56
5	11	27	27,34	37,60	105,33	0,11	112,36
6	16	32	32,55	37,60	25,48	0,30	31,36
7	11	27	27,34	37,60	105,33	0,11	112,36
8	17	35	33,60	37,60	16,04	1,97	6,76
9	27	44	44,03	37,60	41,28	0,00	40,96
10	26	40	42,98	37,60	28,97	8,89	5,76
Σn=10	**209,00**	**376,00**	**376,59**	**376,00**	**450,74**	**29,37**	**480,40***

*KT$_t$ ile KT$_b$+KT$_s$ arasındaki 0,29'luk fark sayıların tamlanmasından kaynaklanıyor.

Değişkelerin hesaplanması (ANOVA -analayses of variance-, bak. Örnek 8.1 ve aşağıya).

1. Değişim kaynağı	**2. Kareler toplamı**	**3. Serbestlik derecesi**	**4. Kareler toplamı ortalaması**	**5. t/F sınaması t_{den}**
2. Toplam değişim	KT$_t$ = 480	n-1 = 9	KO$_t$ = 53	--
3. Doğrusal bağınım*	KT$_b$ = 451	m-1 = 1	KO$_b$ = 451	**KO$_b$/KO$_s$ =**
4. Toplam sapma	KT$_s$ ≈ 29	n-2 = 8	KO$_s$ ≈ 3,6	= 451/3,6 ≈ **125**

*KT, **k**areler **t**oplamı, KO, **k**areler **o**rtalaması; b, bağınım; s, sapma ve t, toplam demektir. m=2 element (Cu, Co).



Altın talebi ile fiyat bağıntısı.

c) Bağıntı katsayısı r'nin sağlanması

$KO_b/KO_s = t_{(8;\ 0{,}99)} = 125$ sonucu t dağılım çizelgesindeki değerlerle karşılaştırılır (bak.Çizelge 3.1). Bunun için önce bir kesinlik derecesi, örneğin, P = % 99, seçilir. t_{kur} çizelgesinin F=8 satırı (n-m) ile P_{99} sütunundan $t_{(8;\ 0{,}99)} = 3{,}36$ okunur. $t_{8;0{,}99}$ (125) > t_{kur} (3,36) olduğundan, bağınım katsayısı % 99 olasılıkla anlamlıdır, yani bulunan r değeri incelenen örnek sayısına göre doğrudur. Bu da yukarıdaki saptamayı kuvvetlendirmektedir.

Bunun yanında,

$$r^2 = \frac{Aciklanabilen.\deg iske}{Toplam.\deg iske}$$

$$= \sqrt{\frac{(\hat{y}-\bar{y})^2}{s_y^{\,2}}}$$

$$= \sqrt{\frac{KT_b}{KT_t}}$$

$$= \sqrt{\frac{451{,}39}{480{,}40}}$$

$$= 0{,}969$$

olarak örnek 8.1'deki değere eşit bir değer bulunur.

Toplam değişke ortalaması, KO_t, yaklaşık 53; açıklanan değişke kısmı, KO_b, 50 ve açıklanamıyan değişke kısmı da, KO_s, 3,6 ≈ 4'tür (Örnek 8.2; 3 değişken de n-1'e bölünür). Yaklaşık rakamlarla toplam değişke 53 ≈ 50+4 alınabilir. Buna göre açıklanan değişke kısmı (50:53) açıklanamıyan değişkeden (4:53) oldukça büyüktür ve toplam değişkenin % 94'üne karşılık gelmektedir. Bu, bağınım modeli (doğrusu) tarafından kapsanan değişke kısmı demektir. Dolayısı ile bağınım katsayısı r, oldukça güçlüdür. Buna yanında kalıntı değişkesi $s^2_{xy} = 29/8 = 3{,}6$ tutmakta ve açıklanamıyan değişkenin çok küçük olduğunu pekiştirmektedir.

8.5 Çok değişkenli bağıntılar

Doğrusal olmayan korelasyonların denklemi parabol formülleri ile hesaplanır. Her bağınım gerçek olmayabilir. Bazen görünür veya saklı bir bağınım da mevcut olabilir. Bağıntı analizi; element çiftleri, grupları ve mineraller arasındaki ilişkiler hakkında önemli ipuçları verebilir. Bu sayede komşu örnekler, cevherleşmeler ve element grupları karşılaştırılabilir. Eğer çok sayıda değişken varsa (birkaç element gibi) çoklu bağıntı (multiple correlation) ve bağınım analizi de yapılabilir. Böylece bir örneğin elementlerin birine veya birkaçına bağlı olduğu ortaya çıkarılabilir. Örneğin,

$$y = b_0+b_1x_1+b_2x_2 \qquad 8.29$$

doğrusal bağıntı eşitliği gibi.

Bağıntı ve bağınım, **faktör, diskriminant ile cluster analizi** gibi analiz yöntemleri ve diğer çok değişkenli istatistik yöntemlerinin temelini teşkil eder. Ancak bunları bağıntı analizi kadar detaylı incelemek, bu notların kapsamını aşar. Bunlara yeri geldikçe değinilecektir. İlgi duyanların bu konularla ilgili kaynaklara başvurmaları önerilir. Sonuç olarak bağıntı ve bağınım analizleri ile ancak olası sonuçlar belirlenir veya bunlara işaret edilir.

Alıştırma 8.1

Bir şirketin ilk 5 yıllık karı aşağıda verilmiştir. Sabit giderleri 500.000 $/yıl olan şirketin gelecek 3 yılın yaklaşık karını ve değişim fonksiyonunu bulunuz.

Verilenler:

Yıllar	1	2	3	4	5
Kar (10^6 $)	1,52	1,35	1,53	2,17	3,60

Yanıt: Yaklaşık % 15.50 ortalama yıllık artış.

Alıştırma 8.2

Bir radyoaktif elementin (izotop) N miktarının aşağıda verildiği şekilde dönüştüğü saptanmıştır. Yarılanma süresi $t_{1/2}$’yi ve dönüşüm eşitliğini bulunuz.

N [g]	1.0000	.9797	.8795	.7736	.2770	.0768
t [gün]	0	10	50	100	500	1000

Açıklama: Genel yarılanma eşitliği,

$$N = N_0 \, e^{-\lambda t}$$

şeklindedir. Bu eşitlik,

$$\underline{\ln N = \ln N_0 - \lambda t}$$

şeklindeki bir doğru eşitliğine dönüştürülebilir. Burada N_0, t=0 başlangıç miktarıdır. $N_{1/2}$ de $t_{1/2}$ yarılanma süresi sonundaki miktardır. Buna göre,

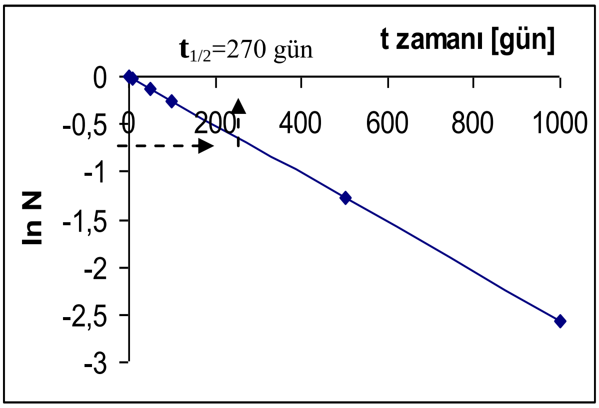
$$N_0 = 1 \text{ ise,}$$

$$N_{1/2} = 0{,}5\text{'tir.}$$

Burada yarılanma süresi aşağıdaki grafikten okunur veya yukarıdaki doğru denkleminden hesaplanır. Bunun için ordinat üzerindeki

$$\ln N_{1/2} = \ln 0{,}5$$
$$= 0{,}6931$$

noktasından absise çizilen koşutun grafik doğrusunu kestiği noktadan absise indirilen dik, absisi $t_{½} = 270$ gün noktasında keser. Bu, yarılanma süresidir.



Yukarıdaki verilerden t zamanı doğrusal absis değerleri (x); miktarın da (N) doğal logaritması ordinat değerleri (y) olarak alındığında, yarılanma fonksiyonu ortaya çıkar. Buna göre fonksiyon değerleri şöyledir:

x	0	10	50	100	500	1000
y	1,0000	-0,0250	-0,1284	-0,2567	-1,2837	-2,5666

Radyometrik yaş tayini veya yarılanma süresinin bulunması:

Yukarıdaki

$$\underline{\ln N_{1/2} = \ln N_0 - \lambda t_{1/2}}$$

eşitliğinden,

b=0 (orijinden geçen doğru) ve

a=-λ

$\underline{= - 0,02567}$

bulunur (bak. 8.12 ve 8.24). Buradan doğru denklemi,

$$\ln 0,5 = \ln 1 - (-0,002567)t_{1/2}$$

çıkar. Değerler yerine konulduklarında,

$$0,6931=0+0,002567t_{1/2},$$

$$t_{1/2}=\frac{0,6931}{0,002567}$$

$$=\underline{270,00}$$

gün elde edilir. Bu süre, başlangıçtaki radyoaktif N_o kütlesinin yarısının dönüştüğü süredir.

EKLER

Ek 8.1 Anlamlı en küçük |r| değerleri (Schroll, 1976).

Serbestlik Derecesi F (n-2)	**Olasılık P** p =95 %	p =99 %	**Serbestlik Derecesi F** (n-2)	**Olasılık P** p =95 %	p =99 %
1	0,997	1,000	21	0,413	0,526
2	0,950	0,990	22	0,404	0,515
3	0,878	0,959	23	0,396	0,505
4	0,811	0,917	24	0,388	0,496
5	0,754	0,874	25	0,381	0,487
6	0,707	0,834	26	0,374	0,478
7	0,666	0,798	27	0,367	0,470
8	0,632	0,765	28	0,361	0,463
9	0,602	0,735	29	0,355	0,456
10	0,576	0,708	30	0,349	0,449
11	0,553	0,684	35	0,325	0,418
12	0,532	0,661	40	0,304	0,393
13	0,514	0,641	45	0,288	0,372
14	0,497	0,623	50	0,273	0,354
15	0,482	0,606	60	0,250	0,325
16	0,468	0,590	70	0,232	0,302
17	0,456	0,575	80	0,217	0,283
18	0,444	0,561	90	0,205	0,267
19	0,433	0,549	100	0,195	0,254
20	0,423	0,537	125	0,174	0,228
			150	0,159	0,208
			200	0,138	0,181
			300	0,113	0,148

Örnek: Örnek sayısı n=10 çift olan bir doğrusal bağıntının % 95 güvenle bağıntı katsayısı |r| > 0,632'dir.

9 ZAMAN DİZİLERİ

9.1 Genel bilgiler

Zaman dizileri, n örnekten oluşan bir bütünün veri bütünlüğüdür. Belli t_i zaman aralıklarında oluşması ve zamana bağlılığı esastır. Genelde,

$$\Delta t = t_{(i+1)} - t_i \qquad 9.1$$

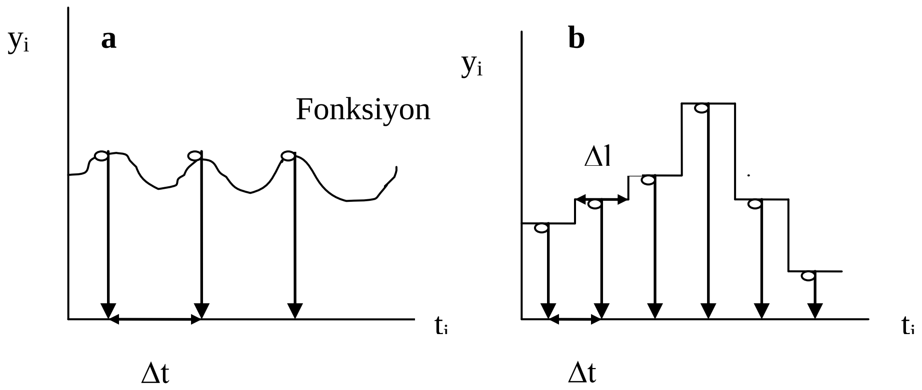
gibi sabit bir zamanın süresi ile incelenir ve tanımlanır. Öncelikle coğrafi şekillerin veya sistemlerin uzun zaman gözlenmesi sonucu ortaya çıkarlar. Gözlemler ancak belli zamanlarda gerçekleştikleri için zaman dizileri **saklı** veya gizli verilerden oluşur. Zaman aralıkları kural olarak eşit seçilirler. Buna karşın gözlemlerin sürekli yapıldığı zaman dizileri de bulunmaktadır. Bunlara **sürekli zaman dizileri** denir.

Zaman dizilerinin en önemli özelliği, zaman fonksiyonu y(t)'ye karşın, y zamana bağlı bir sürekli fonksiyon (sabite) olmalarıdır. Bu nedenle zaman aralıklarının kesin ölçümü zaman dizilerinin incelenmesinde büyük rol oynar. Geçmiş değerlerinden gelecek değerlerinin okunabildiği ve akışının kestirilebildiği zaman dizilerine **belirlenimci** (determinist) zaman dizileri denir. Pratik uygulamada bu tür zaman dizilerine ender rastlanır. Ancak zaman dizilerinin çoğunluğu **rastlantısal**dır (stokastik). Her hangi bir zaman aralığında ortalama değeri veya değişkesi değişmiyen, yani bir yönelimin de olmadığı, zaman dizilerine **durağan** (stationer) zaman dizileri denir. Durağanlık, zaman dizilerinin genel özelliğidir.

Zaman dizilerinin analizinde ilk adım dizinin şeklini çıkarmaktır. Bu yolla periyotlar veya mevsimsel değişimlerle uc değerler tanınabilir. Şekil 9.1 zaman dizilerini geometrik olarak göstermektedir.

a) veriler t_i başlarında (Δt = zaman adımı, periyot, o = veri), örneğin, saat başı pip sesi, nüfus sayımı, motor ve ısı ölçümleri gibi,

b) veriler t_i ortalarında (ortalama değer olarak), örneğin, fırtınalı geçen günler, mevsimler, don, güneşli veya yağışlı dönemler, günler ve saatlerin aylık veya yıllık ortalamaları gibi.



Şekil 9.1. Zaman dizilerinin geometrik gösterimi.

Matematiksel olarak bir zaman dizisi,

$$y_i(t_i),\ t_{i+1}-t_i = \Delta t = \text{sabit} \qquad 9.2$$

ile ifade edilir (i = 1, ..., n). 3 çeşit zaman dizisi ayırdedilmektedir:

a) Eşit zaman Δt aralıklı diziler, saat başı ölçümler gibi gerçek değerler,
b) Δt birleşim aralıklarında (periyotlarında) verilerin meydana geldiği ve gerçekleştiği diziler,
c) Δt zaman aralıklarında ortalama değer olarak görünen sabit zaman dizileri, örneğin, n değerleri gibi.

Dizi uzunluğu L, L = n. Δt, ile hesaplanır. Başlangıç için her zaman dizisi a, b veya c gibi bir değişkenye sayılabilir. Örneğin, 30 yıllık ortalamalar, n = 360 dizisi olarak,

$$n_+ = n/12 = 360/12 = 30$$

aylık ortalamalara ayrılabilir. Aralık (interval) ortalaması,

$$\bar{x}_j = \frac{1}{r-q+1}\sum_{i=q}^{n}(x_i) \qquad 9.3$$

i=1,...,n

j = 1,...,N<n

q= 1,.., n

r = q+1,..., N ve

r-q=sabittir.

Zaman aralıkları, saniye (s), dakika (min), saat (h), gün (d) ve saire zaman birimleridir. Zaman dizileri çok yönlü ve çeşitlidir. Burada hepsini incelemek ve zaman özelliklerinin tümünü bulmak mümkün değil. Bunun için burada geleneksel yöntemlerle sınırlı kalarak zamanın ancak değişim özellikleri incelenecektir. Bu inceleme yöntemleri yönelim, düzenli ve düzensiz değişimlere ayrılabilmektedir. Değişimlerine göre şu zaman dizileri ayırdedilmektedir:

a) **Periyodik** zaman dizileri. Bunlar kesin zaman dilimleri ile, T periyoduyla, sinüssoydal, yinelenirler, $y_i(t_i) = y_i(t_i+T)$ gibi. Yerbilimlerinde gel-git olayı buna örnek verilebilir.

b) **Düzenli** zaman dizileri. Bunlar ritmik, hemen hemen periyodiktir. $y_i(t_i) = y_i(t_i+\bar{T})$ gibi.

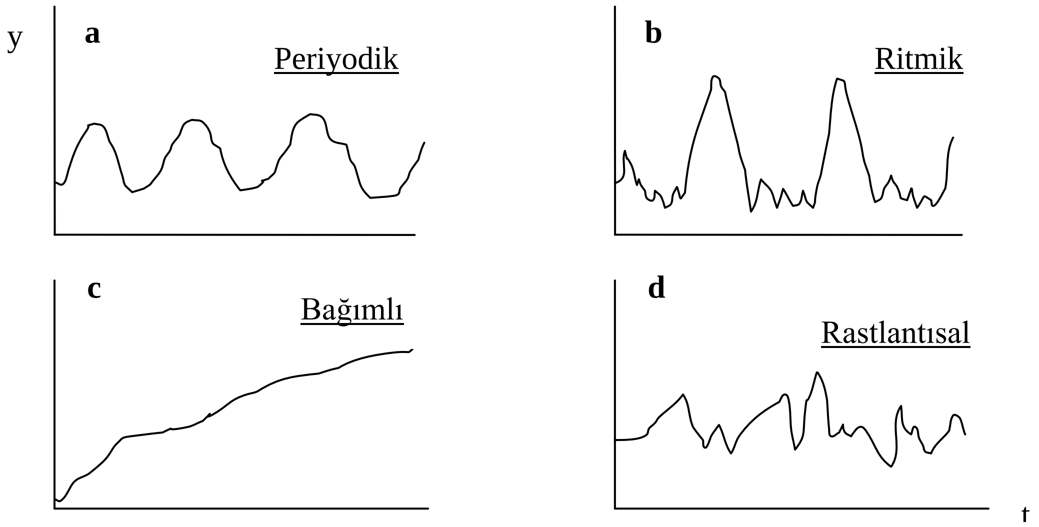
c) **Dönüşümlü** zaman dizileri. Bunlar bağımlıdır, eğilim gösterirler ve

d) **Rastlantısal** (stokastik) zaman dizileri (Şekil 9.2).

Zaman dizileri ile öncelikle zaman özellikleri incelenir. Karışık tipleri çoktur. Sınıflandırmak oldukça zordur. Çünkü iç periyotler veya ritimler saklı olabilir. Dizinin uzunluğu L, $L = n.\Delta t$ ile bulunur. $T = 1,2,...,T_k$ ise, **harmonik zaman dizisi**, sözkonusudur (k = yıl, ay, gün v.s.). Zaman dizilerinde de homojenlik esastır (örneğin, L periyodunun değişimi bağımlılığı etkilememeli). Uc değerler kesinlikle dikkate alınmalıdır.

Bir zaman dizisinin esas yapısı çeşitli kısımlardan oluşmaktadır. Örneğin, bunlar dizinin özgün değerinden oluşur. Bu,

özgün değeri = mevsimsel değeri + yönelim değeri + kalıntı değeri demektir. Dolayısı ile dizinini mevsimsel etkenden arındırılmış değeri = yönelim değeri + kalıntı değeri olur.



Şekil 9.2. Zaman dizilerinin başlıca çeşitleri.

Zaman dizileri analizlerinin amacı, esas elemanlardan temel diziyi ortaya çıkarmak, yani **filtrelemek**tir. Bir zaman dizisinin kısımları arasındaki **yönelim**in genel yönelimle uyumlu olması gerekir. Buna karşın **mevsimsel etken**le yönelim elemanlarının uyumlu olmaması, yönelim her hesaplamada aynı değeri vermesi ve yıllık **özgün değer** toplamının yönelim ve **artık değer** toplamlarının aynı olması lazımdır.

9.2 Filtreleme

Belli frkans aralıklarındaki değişimleri ön plana çıkarma veya ortadan kaldırma yöntemlerine **filtreleme** denir. Doğrusal filtreleme yönteminde kısa süreli değişimler zayıflatılır veya ortadan kaldırılır. Burada bir $|x_t|$ zaman dizisi doğrusal bir operatörle, örneğin,

$$y_i = \sum_{i=-s}^{s} a_i x_{t+1}\,, \quad \sum_{i=-s}^{s} a_i = 1 \qquad 9.4$$

ile, düzgünleştirilmiş $|y_t|$ gibi başka bir zaman dizisine dönüştürülür. y_t değerlerine t noktasındaki düzgünleştirilmiş ortalama değer denir. Bu değerlerin destek alanı [t-q, t+s] terimidir. a_i katsayıları hesaplamalarda kullanılan ağırlıklardır. Doğrusal filtrelemeler çoğu zaman simetriktir, yani, s=q ve $a_i = a_{i-1}$'dir.

En basit filtreleme yöntemi ağırlıksız düzgünleştirilmiş ortalama değerdir. Bu simetrik filtrelemede tüm değerler aynı ağırlıkla, yani,

$$a_i = \frac{1}{2q+1} \qquad 9.5$$

katsayısı ile hesaba katılırlar. Diğer düzgünleştirme ortalama değerlerinin katsayıları binom filtreleme yönteminden, örneğin,

$$\left(\frac{1}{2}+\frac{1}{2}\right)^{2q} \qquad 9.6$$

teriminden, ortaya çıkarlar. q = 1 için,

$$a_{-1} = a_{+1} \qquad 9.7$$
$$=1/4 \text{ ve}$$
$$a_0 = 1/2\text{'dir.}$$

q = 2 için,

$$a_{-2} = a_{+2} \qquad 9.8$$
$$=1/16,$$
$$a_{-1} = a_{+1}$$
$$=1/4 \text{ ve}$$
$$a_0 = 1/2\text{'dir.}$$

Bunlardan başka kullanılabilen filtreler, q=1 için,

$$a_{-1} = a_{+1} \qquad 9.9$$
$$=1/8$$
$$a_0 = 3/4,$$

ve q=2 için,

$$a_{-2} = a_{+2} \qquad 9.10$$
$$= 1/24,$$

$$a_{-1} = a_{+1} = 7/24$$

$$a_0 = 1/3,$$

$$a_{-2} = a_{+2} = 1/8,$$

$$a_{-1} = a_{+1} = a_0 = 1/4$$

olarak bulunur.

Doğrusal filtreleme yöntemi zaman dizilerinin sınırlarını geçebileceklerinde, ilk q ve son s'te sorunlar çıkar. Bunu aşmak için, örneğin $|x_t|$ dizisi uzatılır. Polinomların ilk ve son değerleri genelde saptıklarından, bunlar, diğer değerlerden elde edilirler.

Bundan başka filtreleme olanağı bu ilk ve son zaman dizisi değerlerini tek taraflı filtre ile uyarlanır. Bunun için,

$$y_t = \sum_{i=0}^{s} a_i x_{t+1} \quad \text{ve} \quad \sum_{i=-q}^{0} a_i x_{t+i} = 1 \qquad 9.11$$

şeklinde bir operatörle ağırlıklı düzgünleştirilmiş uyum sağlanır. Ağırlıksız uyum için dizinin ilk ve son değerleri tahmin edilirler (qy_t değerleri $2q+x_t$ değerleri vasıtası ile, bak Barsch ve Billwitz, 1990).

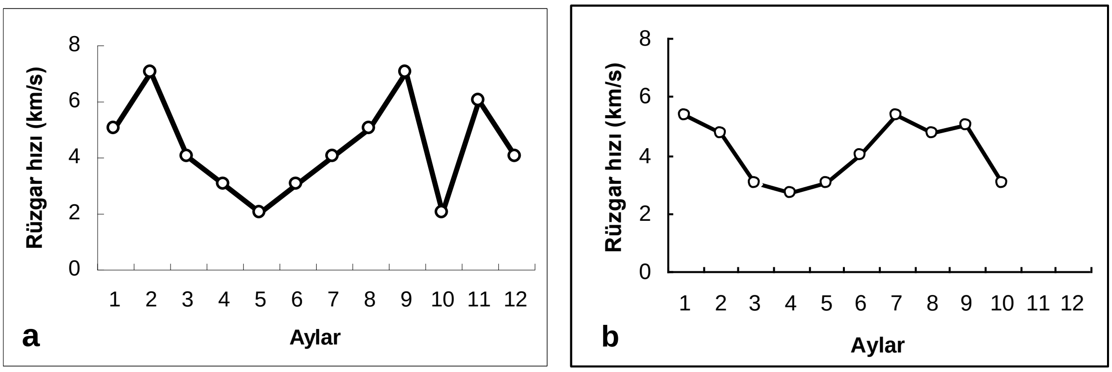
Örnek 9.1

Bir ortalama değerin n = 3 aralığı kullanılarak düzgünleştirilmesi (bak. alttaki şekil).

Çözüm:

Özgün aralıklar: 5 7 4 3 2 3 4 5 7 2 6 4

Taşınan aralık: **5,3 4,7 3,0 2,7 3,0 4,0 5,3 4,7 5,0 3,0**



a, özgün veriler; **b**, piklerin giderildiği düzgünleştirilmiş veriler

Örnek 9.1'e ait ortalama değerin n = 3 aralığı kullanılarak düzgünleştirilmesi.

9.3 Mevsimsel arındırma

Zaman dizilerinin bazıları bir periyot içinde düzenli iniş-çıkışlar gösterir. Örneğin, uzun süreli yıllık üretim, yağış ve ısı ölçümleri belirgin bir şekilde böyle düzenli özellikler gösterir. Bu periyotluklar yönelim veya şimdiye kadar fark edilmiyen başka özellikleri örtebilir. Bu nedenle bazı durumlarda bu etkenin giderilmesi gerekir. Bu da ancak düzgünleştirilmiş ortalama değer-le mümkündür. **Mevsimsel etkenden arındırılmış değer = yönelim değeri + artık değer** olduğundan, periyot uzunluğu p tek sayılı ise (p=2k+1), düzgünleştirilmiş ortalama değer y_t,

$$y_t = \frac{1}{p}\sum_{i=-k}^{k} x_{t+1} \qquad 9.12$$

eşitliği ile hesaplanır. y_t çift sayılı ise (p=2k), mevsimsel arındırma,

$$y_t = \frac{1}{p}\left[\frac{1}{2}x_{t-k} + \sum_{i=-k+1}^{k-1} x_{t+i} + \frac{1}{2}x_{t+k}\right] \qquad 9.13$$

eşitliği ile gerçekleştirilir. Mevsimsel etkenin kestirimi zaman dizisinin gelişmesine bağlıdır. Tüm zaman boyunca mevsimsel etken aynı kaldığında, toplam (additif) demektir ve x_t-y_t farkından kestirilir. Etken değişen ortalama değere göre değiştiğinde de x_t/y_t oranından bulunur.

Basit durumlarda bir mevsimsel arındırma sadece,

$$y_t=x_t-x_{t-p} \qquad 9.14$$

farkı ile de yapılabilir.

9.4 Yönelim saptama

Yönelim, bir zaman dizisinin uzun süreli gelişmesini gösterir. Gözlem süresinden kısa periyotlu (dalda boyunda) artma ve azalmaları içermez. Bu nedenle yönelimler uzun süreli değişimler olarak tanımlanırlar.

Bir yönelim ölçülmeden önce mevsimsel dalgalanmaların temizlenmesi istenir. Yönelimleri göstermeye yarıyan en uygun yöntem doğrusal veya doğrusal olmıyan bağınım modellerinin diziye uyarlanmasıdır. Bunun için zaman dizisinin x_t değerlerinin t zamanları bağımsız, x_t değerlerinin kendileri de bağımlı değişken olarak ele alınırlar.

8. konuda gösterilen bağnım modelleri yanında lojistik fonksiyon,

$$x_t = \frac{\delta}{1+ae^{-bt}} \qquad 9.15$$

ile **Gompertz fonksiyonu**, $\ln xt = \delta + ae^{-bc}$ 9.16

çok kullanışlı olmaktadır. Burada a, b ve δ bağınım değişkenleridir ve b pozitiftir.

İki doğru da artan t ile **doygunluk sınırı** denilen asimtot değerine yaklaşırlar. Lojistik eğri için bu değer δ, Gompertz eğrisi için de e^{δ}'dir. Bu değerlerin başta saptanamaması veya kestirile-memesi durumunda uyumun doğrusal olmıyan yöntemlerle tahmin edilmesi gerekir.

Zaman dizisi değerleri x_t ile bağınım doğrusu değerleri arasındaki farklara artık (residual) değerler denir. Bu değerler gözlem uzayındaki değişimleri gösterir.

Zaman dizilerine doğruların uyarlanması bir küresel kestirimdir. Bundan amaç, dizinin zaman akışını en yüksek güvenirlikle kestirmektir. Küreselleşmenin diziye uygun olmaması durumun-da gözlem aralığı [τ_o, τ...] veya [τ_1, τ_2...] gibi parçalara ayrılarak irdelenir. f(t, a_1,,..., a_k) gibi uygun bir bağınım modelinin seçilmesi ile bağınım değişkenleri a_1,..., k_1 hesaplanır.

Örneğin kısmi yönelim (sequentiell trend analysis) analizlerinde doğrusal yönelim esas alınır. Dizinin her hangi bir aralığında, $[\tau_i, \tau_{i+1}]$ gibi, $|x_t|$ dizisi,

$$y_{(t)} = a_1 + a_2(t - \tau) \qquad 9.17$$

yönelim fonksiyonu ile tanımlanması gerekir. Bu eşitlikte yönelim fonksiyonu değeri $y_{(\tau 1)}$ alt aralık sonu ile sabittir. $y_{(\tau 1)} = a_1$ geçerli olduğundan, 1. bağınım katsayısı saptanmış olmaktadır. 2. katsayı,

$$\sum_j \left[x_{tj} - y_{(tj)}\right]^2 = \sum_j \left[x_{tj} - a_1 - a_2(t_j - \tau_i)\right]^2 = \min \qquad 9.18$$

eşitliği ile bulunur (optimizasyon kriteri, Barsch ve Billwitz, 1990). Burada da en küçük kareler toplamının geçerli olduğu görülmektedir. Toplam, dizinin $[\tau_i, \tau_{i+1}]$ aralıklarının τ_j zamanlarını kapsar. Yukarıdaki toplamdan a_2 için,

$$\sum_j \left[x_{tj} - a_1 - a_2(t_j - \tau_i)\right]\,(t_j - \tau_i) = 0$$

eşitliğinden,

$$a_2 = \frac{\sum (x_{tj} - a_j)(t_j - \tau_i)}{\sum (t_j - \tau_i)} \qquad 9.19$$

bulunur. Böylece süreklilik nedeni ile a_1 değeri yönelim fonksiyonun kendisinden sonraki aralığın $y_{(\tau_i+1)} = a_1 + a_2(\tau_i + 1 - \tau_i)$ dizi değeri ile özdeş olur.

Bağınım fonksiyonunun zaman dizisine uyumu gereksiz görüldüğü veya arzu edilmediği zaman yönelimin düzleştirilmesi uygun olur. Bunun için sıkça doğrusal filtreler kullanılır. Bir düzgünleştirmenin yeterli olmadığı yerlerde birkaç kez filtreleme kullanılır. Örneğin,

$$y_t = \sum_{i=-q}^{q} a_i x_{t+i} \quad \text{ve} \qquad 9.20$$

$$z_t = \sum_{j=-s}^{s} b_j y_{t+1} \qquad 9.21$$

filtreleri gibi. İkisinin arka arkaya gelmesi durumunda,

$$z_t = \sum_{j=-s}^{s} b_j \sum_{i=-q}^{q} a_i x_{t+j+i} \qquad 9.22$$

$$= \sum_{k=(-q+s)}^{q+s} c_k x_{t+k}$$

şeklini alır. Toplam ağırlık,

$$ck = \sum_{j=-s}^{s} b_j a_{k-1} \qquad 9.23$$

değerine sahiptir.

Bazen bir zaman dizisinin yönelimin arındırılması istenir. Bu durumda ya yönelim saptanarak değerleri dizinin değerlerinden çıkarılır, ya da,

$$y_t = x_{t+1} - x_t \qquad 9.24$$

farkları oluşturularak yönelim elimine edilinceye kadar diziden çıkarılır.

Örnek: 9.2.

Bir şirketin ilk 5 yıllık karı aşağıda verilmiştir. Gelecek 5 yılın yaklaşık karını ve değişim fonksiyonunu bulunuz.

Verilenler:

Yıllar	1	2	3	4	5
Kar (10^6 \$)	1,52	1,35	1,53	2,17	3,60

Çözüm:

a) 3. yıla göre 1 ve 2 ile 4 ve 5'in ortalama oranları birleştirilerek uzatıldığında gelecek 5 yılın değerleri doğru üzerinde yaklaşık okunur (ba. Alttaki şekil). 1. ve 2. yılın ortalaması yaklaşık, (1,52+1,35)/2 = 1,44 mil. $ ve 4. ile 8. yılın ortalaması da yaklaşık (2,17+3,60)/2 = 2,89 mil. $ olarak bulunur. Grafikte bu iki noktadan geçen doğru üzerinde diğer değerler yak-laşık okunur. Örneğin, 7. yıl için 4,05 mil. $ gibi.

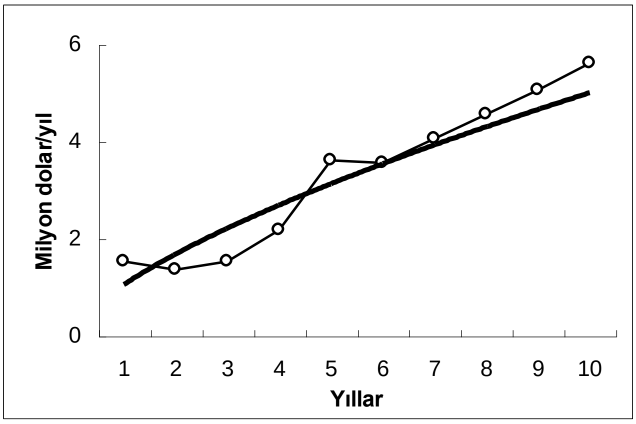
b) 2. yöntem en küçük kareler yöntemidir. Buna göre bu noktalara uyan doğru veya eğri denklemi bulunarak 5. yıldan sonraki yıllar için eşitliği sağlıyan değerler okunur. Buna göre y = mx + n doğrusunda,

$$m_y = s_{yx}:s^2_x = 0{,}496 \text{ ve}$$

$$n_y = \bar{y} - m_y x = 0{,}549\text{'dan}$$

$$y_{1\text{-}5} = 0{,}549 + 0{,}496\ x$$

bulunur. Buna göre örneğin, 8. yıl için y_8 = 0,496.8 + 0,546 = 3,968 + 0,546 = 4,514 mil. $ kar okunur (bak. aşağıdaki şekil).



Yanıt: Yaklaşık % 10 ortalama yıllık artış.

Örnek 9.2'ye ait yönelim analizi ile gelecekteki değerlerin kestirimi.

9.5 Zaman dizilerinin korelasyonu (otokorelasyon)

Otokorelasyon, korelasyon fonksiyonunda olduğu gibi, $y_i = y_{i(t_i)}$ şeklinde bir bağıntı sunan fonksiyonlardır. Otokorelasyonda bağıntı katsayısı r_A,

$$r_A = \frac{\sum_{i=1}^{n-\tau}(x_i - \bar{x})(y_{i+\tau} - \bar{y}_\tau)}{\sqrt{\sum_{i=1+\tau}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n-\tau}(y_{i+\tau} - \bar{y}_\tau)^2}} \qquad 9.25$$

$$= \frac{s_A}{s_1.s_2} \qquad 9.26$$

eşitliği ile bulunur.

Burada,

$-1<r_A<+1$,

$i= 1,..., n$ ve

$\tau = 0,1,...,M<n$'dir.

$\tau = 0$, zaman kaydırıması yok,

$\tau = 1$ için x_i sonraki x_{i+1} ile korele edilmiyor demektir.

Zaman kaydırma, $k = M.\Delta t$ olarak hesaplanır.

Örnek 9.3

Zaman kaydırma ile otokorelasyon

I.	x_1	x_2	x_3	x_4	x_5	x_6	x_7 .... $\tau = 0.\Delta t$
II.		x_1	x_2	x_3	x_4	x_5	x_6 .... $\tau = 1.\Delta t$
III.			x_1	x_2	x_3	x_4	x_5 .... $\tau = 2.\Delta t$
IV.				x_1	x_2	x_3	x_4 .... $\tau = 3.\Delta t$

Bir zaman dizisi başka bir zaman dizisi ile zaman kaydırmadan da korele edilebilir.

Örnek 9.4

1950-60 ısı ortalaması (n=11) otokorelasyonu.

Yıl	x_i	$x_{i+\tau}$		
	1.	**2.**	**3.**	**4.**
1950	8,8 (ºC)	8,4	8,1	8,3
1951	8,4	8,1	8,3	7,3
1952	8,1	8,3	7,3	7,3
1953	8,3	7,3	7,3	6,5
1954	7,3	7,3	6,5	8,2
1955	7,3	6,5	8,2	8,5
1956	6,5	8,2	8,5	8,2
1957	8,2	8,5	8,2	8,4
1958	8,5	8,2	8,4	
1959	8,2	8,4		
1960	8,4			

Çözüm:

1. adım: $\tau = 0$

1. sütun ve 1 korele edilir. $r_A(1,1) = 1$, $s_A(1,1) = 0,44 = s_2$

2. adım: $\tau = 1$

1. ve 2. sütunlar korele edilir. $r_A(1,2) = 0,36$, $s_A = (1,2) = 0,16$

3. adım: $\tau = 2$

1. ve 3. sütun korele edilir. $r_A(1,3) = 0,20$, $s_A(1,3) = 0,00$ v. s. şeklinde devam edilir.

9.6 Harmonik analiz

Harmonik analiz, bir sürekli, kesintisiz ve sonsuz periyodik zaman fonksiyonudur. x(t)= x(t+T) fonksiyonunun sinüs-kosinüs dizisidir ve $T=n_i.P$ periyoduna sahiptir. Bu eşitlikte n_i, bir doğal sayıdır. Analiz veya çözüm için fourier dizisi kullanılır. Fourier dizisi,

$$x(t) = \frac{B_o}{2} + \sum_{i=1}^{n} A_i \sin\omega\, t + \sum_{i=1}^{n} B_i \cos\omega\, t \qquad 9.27$$

$$\omega = 2\pi /T = 2\pi\, f,$$

$$A_i = \frac{2}{T}\int_0^T x(t)\sin i\,\omega\, t\, dt \text{ ve}$$

$$B_i = \frac{2}{T}\int_0^T x(t)\cos i\,\omega t\, dt \qquad \text{Fourier katsayıları} \qquad 9.28$$

şeklindedir. Bu çözümlerdeki zorluklar şunlardır:

- Veriler sonludur,
- $x_i(t)$ şeklinde ortaya çıkarlar,
- Kesin periyodik olaylara ender rastlanır.

Örnek: 9.5

Bir ana periyodun harmonik kısmi salınımlara ayrılması ($H_0=T_0$).

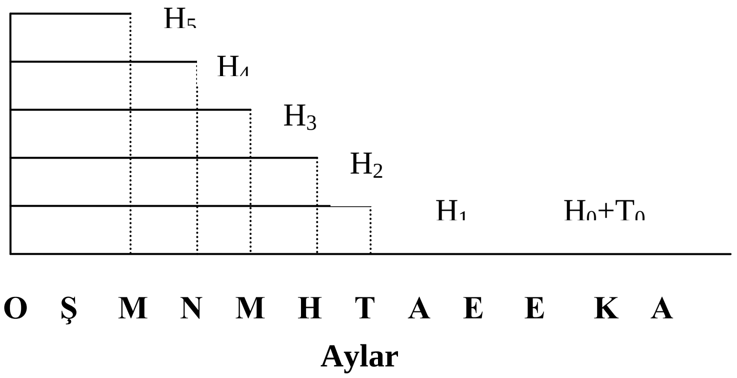
P yılı = 12.Δt

ΔT=1 ay, N=12, N/2=6 kısım salınım veya titreşim.

P_1= 12, P_2= 12/2=6 ay

P_{min}=2 ay

P_{max}= 12 ay.



Örnek 9.5'e ait ana periyodun harmonik kısmi salınımlara ayrılması (H_0=T_0).

Harmonik analiz yöntemi, fourier (Furiye) dizisi fonksiyonlarının geliştirilmiş değişik bir şeklidir. Kullanımı bakımından saklı zaman dizilerini oluşturur. Analizden saklı periyotların bulunmasında yararlanılır. Bulunan periyodik özellikler "düzenli dönem" olarak nitelendirilir.

Analizlerde önkoşul, dizinin duraylılığıdır. Bunun için önce, dizideki diğer belirtileri örtme-meleri amacı ile, yönelim ve mevsimsel değişimler ortadan kaldırılır. Kolaylık sağlaması için t zamanının 1, 2, ..., n gibi değerler aldığı ve n'nin çift olduğu kabul edilir.

Basitleştirilmiş bu düşünceden çıkarak bir $|x_t|$ zaman dizisinin bir belirlenimci, ω frekanslı sinüs şekilli elemanı bulunduğu ve bir rastlantısal hata olduğu görülür. Buna göre diziye,

$$X_t = \mu + \alpha \cos\omega t + \beta \sin\omega t \qquad 9.29$$

eşitliği uyarlanabilir. Burada α, β ve ω en küçük kareler yöntemine göre dizi verilerinden hesap-lanırlar. Eğer $\omega_p = 2\pi.p/n$ (p = 1, 2, ..., n/2) ise, p ≠ için,

$$\mu = 1/n \, \Sigma \, x_t = \bar{x} \, ,$$

$$\alpha = 2/n \, \Sigma \, x_t \cos(\omega_p t) \text{ ve}$$

$$\beta = 2/n \, \Sigma \, x_t \sin(\omega_p t)$$

olarak hesaplanabilir. p = 2 ise, her hangi bir t için ($\sin\omega_p t = 0$ ilişkisinden),

$$\mu = \bar{x} \text{ ve} \tag{9.30}$$

$$\alpha = 1/n \, \Sigma \, x_t \cos \pi \, t \tag{9.31}$$

$$= 1/n \, \Sigma \, (-1)^t \, x_t$$

elde edilir.

Buna göre her hangi bir zaman dizisi için dizi değerlerinin toplam ortalama değerden sapmalarının sinüs şeklindeki elemanlarının toplamı ve ω_p frekansı olarak gösterilebilir. Bu, açıkça bir Fourier dizisidir (harmonik analiz) ve

$$x_t = a_0 + \sum_{p=1}^{n/2-1} \left[a_p \cos \omega_p t + b_p \sin \omega_p t \right] + a_{n/2} \cos \pi t \tag{9.32}$$

ile ifade edilir (t= 1, 2,..., n). $a_o = \bar{x}$, a_p ve b_p ω_p frekansları için özdeştir ve yukarıda hesaplanan α ve β değişkenlerine karşılıktır. Fourier analizinde n zaman dizisi değerleri x_t, a_o, a_n/2, a_p ve b_p ω_p (p=1,2,...,n/2-1) yardımı ile tanımlanırlar.

Kaynakça

1. Ahrens, L. H, 1954: The lognormal distribution of the elements (1+2). **Geochim. et Cosmochim. Acta**, 5/6, 49-73/121-131

2. Anadolu Üniversitesi (yayınlıyan), 2001: İstatistik. 4. basım, **Anadolu Üniversitesi yayınları** 175, Eskişehir, 312 s.

3. Arıcı, H., 2001. İstatistik. **Metaksan Basımevi**, 13. baskı, İstanbul, 269 s.

4. Backhaus, K., Erichson, E., Plinke, W. ve Weiber, R., 1990: Multivariate Analysenmethoden. **Springer Verl**., 6. Basım, Berlin-Heidelberg-New York, 411 s.

5. Barsch, H. ve Billwitz, K., 1990: Geowissenschaftliche Arbeitsmethoden. **Verl. Harri Deutsch**, Thun ve Frankfurt am Main, 256 s.

6. David, M., 1977: Geostatistical Ore Reserve Estimation II**, Elsevier Sc. Publ. Comp**., 3. baskı, Amsterdam, 364 s.

7. Davis, J. C., 2002: Statistics and data analysis in geology. **John Wiley and Sons. Inc**., 3. basım, New York-Chichester- Brisborn-Toronto, 638 s.

8. Ergün Bülbül, S., 2001: Çözümlü istatistik. **Alfa yayınları 1006**, İstanbul, 591 s.

9. Jischa, M. F., 2000: Die Dynamik des technischen Wandels. **Bergbau 9**, 401-405.

10. Kara, İ., 2000: Olasılık. **Bilim Teknik Yayınevi**, 4. basım, Ankara, 296 s.

11. Leonhart, M., 2002: Statistik für Psychologen. http://**www.psychologie.uni-freiburg.de**, 750 s.

12. Lepeltier, C. (1969): A simplified statistical treatment of geochemical data by graphical representation. **Econ Geol. 64**, 538-550.

13. Marsal, D., 1990: Yerbilimciler için istatistik. **Hacettepe Üniversitesi, Müh. Fak. yayını**, Ankara, 168 s.

14. Perillo, G. M. E. Ve Marone, F., 1986: Determination of optimal numbers of class intervalls using maximum entropy. **Math. Geol. 18**, 4, 401-407.

15. Reinhardt, F. ve Heinrich, S., 1994: dtv-Atlas zur Mathematik. Cilt 1 ve 2., **dtv Verlag**, Münih, 498 s.

16. Rodionov, D. A, 1962: Bestimmung des Durchschnittsgehaltes und der Streuung einer Lognormalverteilung von Komponenten in Gesteinen und Erzen (Rusça). Geokh. 624, **Geochem. 728**.

17. Sachs, L., 1984: Angewandte Statistik. **Springer Verl.,** 6. basım, Berlin-Heidelberg, 528 s.

18. Schönwiese, Ch.-D., 1992: Praktische Statistik für Meteorologen und Geowissenschaftler. **Borntraeger**, 2. basım, Berlin, Stuttgart, 231 s.

19. Schroll, E., 1976: Analytische Geochemie. **Ferdinand Enke Verl**, Stuttgart, 374 S.

20. Sinclair, A., 1976: Aplication of probablity graphs in mineral exploration. The Association of Exploration Geoche-mists (yayınlıyan). **Richmond Printers Ltd**., Vencouver B. C., 95 s.

21. Tüysüz, N. ve Yaylalı, G., 2005: Jeoistatistik. **KTÜ yayını 220**, Trabzon, 382 s.

22. Wellmer, F.-W., 1989: Rechnen für Lagerstaettenkundler und Rohstoffwirtschaftler. **Ellen Pilger**, Clausthal-Zellerfeld, 462 S.

23. Wessel, P., 2001: Geological data analysis. http//**www.higp.hawaii.edu**./cecily/courses, 325 s.

24. Wilke, A., 1973: Verfahren zur Probenahme aus Erzen (Konzentraten) und und ähnlichen Rohstoffen. Analyse der Metalle 3., **Springer Ver**l., 3. basım, Berlin-Heidelberg, 82-170.