

ASTROİSTATİSTİK

11. KONU

Hazırlayan: Doç. Dr. Tolgahan KILIÇOĞLU

11. NORMAL DAĞILIM

Önceki konularda da değindiğimiz gibi doğada karşımıza çıkan birçok olgu bir normal dağılım (Gauss eğrisi veya çan eğrisi) sergilemektedir. Aşağıda, genellikle normal dağılıma benzer dağılım sergileyen durumlara bazı örnekler verilmektedir:

- İnsanların boylarının uzunlukları (kadın-erkek ayrı olarak ele alındığında),
- İnsanların ağırlıkları (kadın-erkek ayrı olarak ele alındığında),
- Bir sınav sonucunda alınan notlar,
- IQ testi sonuçları,
- Bir tarladaki biberlerin uzunlukları,
- Bir masanın boyunun farklı insanlar tarafından ölçümü,
- Bir kavanozun içinde kaç adet bilye olduğunun birçok insan tarafından tahmin edilmesi,
- Farklı insanlar tarafından kullanılan aynı model bulaşık makinesinin ilk defa kaç gün sonra arıza verdiği,
- Bir firmaya ait 1 litrelik paketli portakal sularının ağırlıkları,
- Bir yıldızın birçok kere gözlenmesi sonucunda elde edilen akı değerleri,
- Bir yıldızın tayfındaki soğurma çizgileri,
- Hayatınızda geçirdiğiniz günleri 1'den (berbattan) 10'a (müthişe) doğru numaralandırdığınızda oluşan dağılım,
ve daha bunlar gibi daha niceleri...

Bir gözlem veya deney yapıldığında verilerin normal dağılıma uyduğuna kanaat getirilirse veriler üzerinde olasılık hesapları ve istatistiksel çıkarımlar yapmak oldukça kolaylaşır. Bunun temel nedeni normal dağılımın bir matematiksel ifadesinin olması ve bu sayede hesaplamaların nicel olarak rahatlıkla yapılabilmesidir.

Bir veri normal dağılımla temsil edilebiliyorsa o veriye ilişkin sadece iki parametreyi bilmek yeterli olmaktadır: ortalama (\bar{x} veya μ) ve standart sapma (s veya σ). Öyle ki bu iki değer oluşturabileceği yalnız bir Gauss eğrisi vardır. Gauss eğrilerinin hepsi birbirinin kopyasıdır. Farklı veriler için (farklı ortalama ve standart sapma değerleri için) oluşturulmuş olan Gauss eğrileri x ve y eksenlerinde gerekli miktarlarda şikıştırılır/esnetilirse hepsinin birebir aynı şekilde olduğu görülecektir.

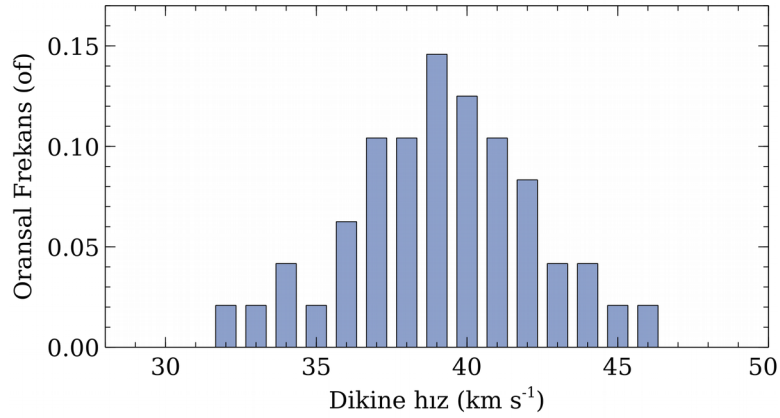
Normal dağılım simetriktir, yani çarpıklığı $\zeta=0$ dır. Normal dağılımın basıklığı da $b=0$ dır. Bir verinin çarpıklığının ve basıklığının hesap edilmesiyle onun normal dağılıma ne kadar benzediği ortaya konulabilir.

Şimdi, bu bölümde kullanmak üzere bir örnek veri sunalım. Çizelge 11.1'de Hyades (Öküz) Açık Yıldız Kümesi'ne üye 46 yıldızın dikine hız (v_r) değerleri km s^{-1} biriminde verilmektedir (frekans dağılımını kolaylaştırmak amacıyla veriler tam sayılara yuvarlanmış olup gerçek dikine hız verileri daha hassastır).

Çizelge 11.1 Hyades açık yıldız kümesine üye 46 yıldızın dikine hızları (km s⁻¹)

39	38	33	39	41	40	43	37	36	42	41	38
40	43	40	42	37	39	38	32	41	38	35	39
36	41	37	39	46	40	45	42	39	36	40	37
37	44	34	44	42	38	41	39	40	34		

Şekil 11.1'de Çizelge 11.1'de verilen dikine hız değerlerinin frekans dağılımı bir histogram üstünde gösterilmektedir.



Şekil 11.1 Hyades açık yıldız kümesine üye 46 yıldızın dikine hızlarının frekans dağılımı

Soru 11.1 Çizelge 11.1'deki verilerin bir popülasyondan mı yoksa bir örneklemden mi geldiğini tespit ediniz. Buna göre verilerin ortalama ve standart sapma değerlerini hesaplayınız.

İlgili ifadeler kullanıldığında yukarıdaki sorunun cevaplarının $\bar{x}=39.2 \text{ km s}^{-1}$ ve $s=3.1$ olduğu görülür. Başka bir deyişle, Hyades Kümesi'nin dikine hızının $39.2 \pm 3.1 \text{ km s}^{-1}$ olduğu söylenebilir. Elde ettiğimiz bu değerleri konu ilerledikçe kullanacağız.

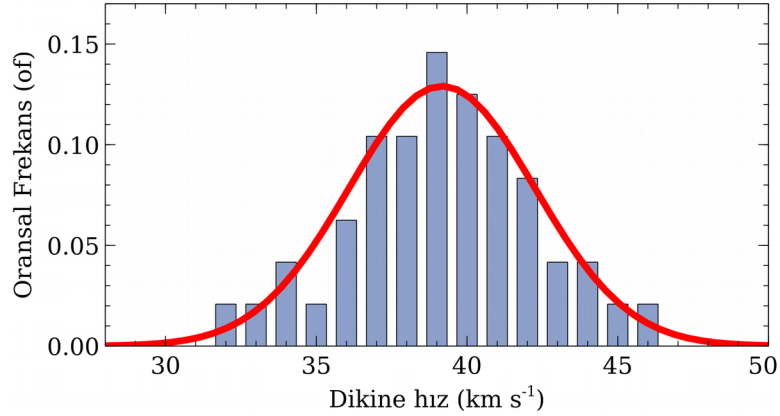
11.1 Normal Dağılımın Matematiksel İfadesi

Normal dağılımın (Gauss eğrisinin) matematiksel ifadesi aşağıdaki şekildedir;

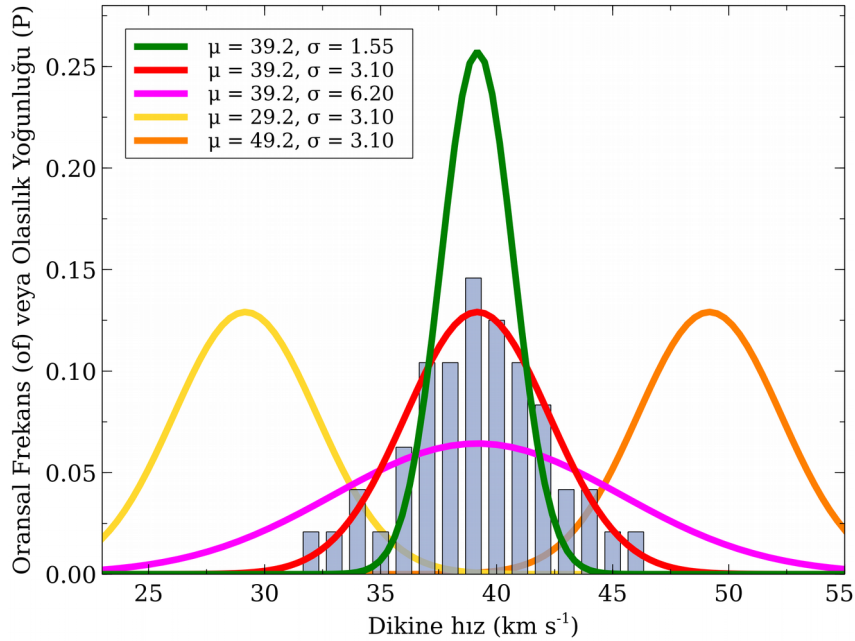
$$N(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

olarak ifade edilir. Burada π ve e matematiksel sabitler, σ ve μ ise yine standart sapma ve ortalamadır.

Konunun başında verdiğimiz örnek dikine hız dağılımı için $\bar{x}=39.2 \text{ km s}^{-1}$ ve $s=3.1$ olduğunu hesaplamıştık. Şekil 11.1'de verdiğimiz dikine hız dağılımının üzerine hesapladığımız bu ortalama ve standart sapma değerleri yukarıdaki ifadede yerine konularak ($\mu \approx \bar{x}$ ve $\sigma \approx s$ alınacaktır) bir normal dağılım fonksiyonu çizdirildiğinde Şekil 11.2 elde edilir. Bu şekilden görüldüğü gibi dikine hız verisi normal dağılımla oldukça iyi temsil edilebilmektedir.



Şekil 11.2 Dikine hız frekans dağılımı (mavi) üzerinde normal olasılık dağılımı (kırmızı)



Şekil 11.3 μ ve σ değerlerinin değiştirilmesinin normal dağılım (Gauss eğrisi) üzerindeki etkisi

Şekil 11.3'te çeşitli μ ve σ değerleri için çizdirilen normal dağılımların aynı veri üzerinde bir karşılaştırması verilmektedir. Görüldüğü gibi μ değeri değiştiğinde Gauss eğrisi sadece sağa veya sola kaymakta, σ değeri arttığında eğri genişlemekte ve tepe değeri aşağı inmekte, σ değeri azaldığında ise eğri daralmakta ve tepe değeri daha yukarı

çıkılmaktadır. Burada, σ nın verinin duyarlılığını temsil ettiğini söylemek yanlış olmaz. σ değeri küçük olan bir veri eğer bir ölçümden geliyorsa bu ölçümün daha az saçılma gösterdiğini işaret eder.

Tüm normal dağılımların altında kalan alan 1'e eşittir. Çünkü bir normal dağılım bir işlemin sonucunda oluşabilecek tüm olası durumları içermektedir.

11.2 Standart Normal Dağılım ve z Puanı

Bu konunun başında tüm Gauss eğrilerinin şekilsel olarak bakıldığında birbirinin aynısı olduğunu söylemiştik. Bu durumda, normal dağılım gösteren bir verideki değerler uygun sayılarla toplanır ve çarpılırsa başka bir normal dağılım gösteren veri ile aynı değerlere sahip olur.

Bu noktada farklı veriler için farklı normal dağılımlar tanımlamak yerine bir **standart normal dağılım** tanımlamak ve bize verilen değerleri bu dağılıma uydurmak daha akla yatkındır. Standart normal dağılımın ortalaması $\mu=0$ ve standart sapması $\sigma=1$ dir. Bir verinin standart sapmasının 1'e ortalamasının ise 0'a çekilebilmesi için o verideki her değerden ortalama değer çıkarılması ve çıkan sonucun standart sapmaya bölünmesi gerekir:

$$z = \frac{x - \mu}{\sigma}$$

Bu işlem sonucu elde edilen değere **z puanı** denir. Bir başka ifade ile, **z puanı verideki bir değer ortalama değerden kaç standart sapma katı kadar uzak olduğunu verir**. İfadede pay ve paydanın birimleri aynı olduğundan z puanı normalize edilmiş birimsiz bir değerdir (benzer bir bölme işlemi çarpıklık ve basıklık için de yaptığımızı hatırlayın). Böylece artık verinin yıldızların dikine hızını içermesi ile insanların boy veya kilolarını içermesi arasında hiçbir fark kalmaz. Her iki durumda da z puanları hesaplanıp grafiğe aktarıldığında aynı standart normal dağılım karşımıza çıkar.

Şimdi Çizelge 11.1'de sunduğumuz dikine hız verisinin normal dağılımla temsil edildiğini varsayarak 45 km s^{-1} değerinin z puanını örnek olarak hesaplayalım (σ ve μ değerlerini daha önce bulmuştuk):

$$z = \frac{x - \mu}{\sigma} = \frac{45 - 39.2}{3.1} = +1.87$$

Bu değer ne ifade eder? Bu değer, 45 km s^{-1} değerinin ortalama değerden $+1.87 \sigma$ kat (pozitif yönde) uzakta olduğunu söylemektedir. z puanının değeri ele alınan değer olasılığı hakkında da bilgi verir. Şimdi bunun nasıl olduğunu başlıklar altında görelim.

z = 0 durumu: Eğer ele alınan değer z puanı sıfıra eşitse bu değer ortalama değere eşit olması gerekir. Bu değer aynı zamanda **en olası değer** veya **beklenen değer** olarak da adlandırılır.

0 < |z| < 1 durumu: Eğer ele alınan değerin z puanı negatif veya pozitif yönde 0 ile 1 arasında ise, bu bize değerin ortalamaya göre $\pm 1\sigma$ aralığında olduğunu söyler. Daha önceki konularda (örneğin bkz. Bölüm 4.5) da söz ettiğimiz gibi popülasyondan rastgele seçilen bir değerlerin yaklaşık %68'i $\pm 1\sigma$ aralığında yer alır. Bu nedenle ele alınan değerin z değeri bu aralıkta ise değerle karşılaşma (veya olayın gerçekleşmesi) olasılığı oldukça yüksektir. Bu türden değerler **yüksek olasılıklı değerler** olarak adlandırılabilir.

1 < |z| < 2 durumu: Eğer ele alınan değerin z puanı negatif veya pozitif yönde 1 ile 2 arasında bulunuyorsa böyle bir değerle karşılaşma olasılığının nispeten düşük ama yine de olanaklı olduğu söylenebilir. Çünkü rastgele seçilen bir değer bu aralıkta karşımıza çıkma olasılığı %27'dir. z puanı bu aralıkta olan değerleri **düşük olasılıklı değerler** olarak adlandırmak yanlış olmaz.

3 < |z| durumu: Eğer ele alınan değerin z puanı negatif veya pozitif yönde 3'ten büyükse böyle bir değerle karşılaşma olasılığımız son derece düşüktür. Öyle ki rastgele seçilen bir değerin bu aralıkta kalması olasılığı sadece %5'tir. z puanı bu aralıkta olan değerler **nadir/sıradışı/beklenmedik değerler** olarak adlandırılabilir.

Şimdi dikine hız örneğine geri dönersek, 45 km s^{-1} lik hız değeri için $z = +1.87$ hesaplamıştık. Dikine hızların dağılımının normal dağılım olduğu varsayımıyla, böyle bir değerin karşımıza çıkma olasılığının düşük olduğu söylenebilir. Başka bir deyişle, örnekte ele alınan Hyades Kümesi'nden rastgele bir yıldız seçildiğinde dikine hız değeri muhtemelen 45 km s^{-1} olmayacaktır.

Hesaplamamızı şimdi 40 km s^{-1} değeri için yapalım;

$$z = \frac{x - \mu}{\sigma} = \frac{40 - 39.2}{3.1} = +0.26$$

Burada z değeri 0 ile 1 arasında (hatta sifıra oldukça yakın) olduğundan Hyades Kümesi'nden rastgele bir yıldız seçildiğinde dikine hız değerinin 40 km s^{-1} olma olasılığının oldukça yüksek olduğu söylenebilir.

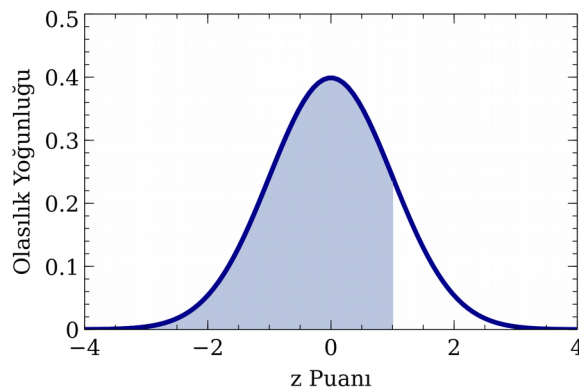
Buraya kadar olasılıkları hep nitel olarak verdik; yani düşük olasılık, yüksek olasılık vb. Gerçekte z puanı kullanılarak bir değerle karşılaşma olasılığının nicel değeri de elde edilebilmektedir. Şimdi bu konuya bir göz atalım.

11.3 z Puanı Kullanılarak Olasılığın Belirlenmesi

Bir rassal x değişkeninin z puanı hesaplandığında **z-çizelgesi** (Çizelge 11.2) olarak adlandırılan bir çizelge yardımıyla iki z puanı aralığı veya belirli bir z puanından büyük/küçük olma durumları için de olasılık belirlenebilmektedir.

Çizelgedeki değerlerin anlamına geçmeden önce bir z puanına karşılık gelen alan değerinin nasıl bulunduğunu açıklayalım. Öncelikle aranan z puanının noktadan önceki ve sonraki bir hanesi alınır ve tablonun ilk sütununda yeri tespit edilir. Daha sonra noktadan sonraki ikinci basamağın yeri ilk satırda bulunur. Sütunda ve satırda tespit edilen bu iki değer kesimi z değerine karşılık gelen bir alan vermektedir (bu alanın ne olduğunu birazdan açıklayacağız). Örneğin, z değeri -1.55 ise ilk kolonda -1.5 değeri bulunur, ilk satırda ise 0.05 değeri bulunur. Bu iki değer kesiminin tabloda 0.606 olduğu görülmektedir. Peki bu ne anlama gelir?

Çizelgede bir z değerine karşılık gelen alan, z değerinin sol tarafına doğru ilerlendiğinde normal dağılım eğrisinin altında kalan alandır. Şekil 11.4'te $z=1.00$ değeri için z-çizelgesi okunan 0.8413 değerinin karşılık geldiği alan gösterilmektedir.



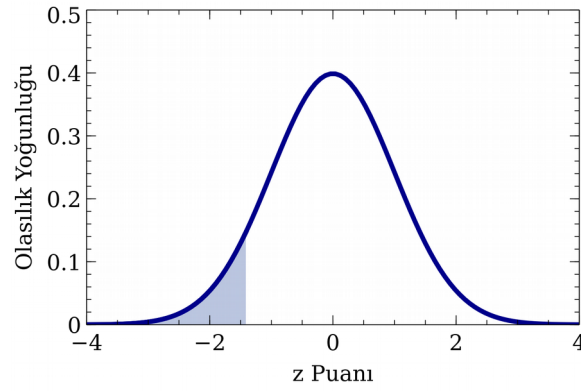
Şekil 11.4 z-çizelgesinden $z=1.00$ değeri için okunan alan değerinin normal dağılım üzerinde gösterimi

Örnek 11.1 Hyades kümesinden rastgele bir yıldız seçildiğinde dikine hızının 35 km s^{-1} den daha küçük olması olasılığı nedir (kümenin dikine hız dağılımının normal dağılıma sahip olduğunu kabul ediniz)?

Hyades kümesinin dikine hız dağılımı için ortalama ve standart sapma değerlerini $\bar{x}=39.2 \text{ km s}^{-1}$ ve $s=3.1$ olarak elde etmiştik. Verilerin normal dağılım sergilediğini kabul ettiğimiz için bu değerleri μ ve σ değerleri olarak kabul edeceğiz. Öncelikle 35 km s^{-1} değerinin z puanını bulalım.

$$z = \frac{x - \mu}{\sigma} = \frac{35 - 39.2}{3.1} = -1.35$$

olduğu görülür. Şimdi sorunun cevabı için aradığımız alanı bir şekil üzerinde gösterelim:



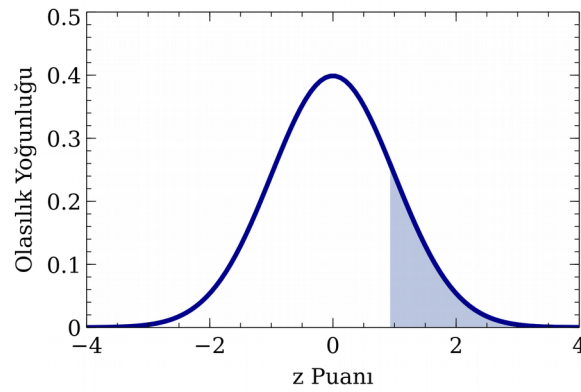
Şekilden de görüldüğü gibi tek yapmamız gereken z-çizelgesinden $z = -1.35$ 'e karşılık gelen değeri okumaktır. Çizelgede sol sütunda -1.3'e üst satırda ise 0.05'e karşılık gelen değerin 0.0885 olduğu görülmektedir. Yani sorunun cevabı olarak Hyades Kümesi'nden rastgele seçilen bir yıldızın dikine hızının 35 km s^{-1} den daha küçük olması olasılığı %8.85 dir.

Örnek 11.2 Hyades kümesinden rastgele bir yıldız seçildiğinde dikine hızının 42 km s^{-1} den daha büyük olması olasılığı nedir (kümenin dikine hız dağılımının normal dağılıma sahip olduğunu kabul ediniz)?

45 km s^{-1} değerinin z puanı:

$$z = \frac{x - \mu}{\sigma} = \frac{42 - 39.2}{3.1} = 0.90$$

Aradığımız alanı bir şekil üzerinde gösterelim:



Aradığımız alanın z-çizelgesinde doğrudan bulunamayacağı görülmektedir. Çünkü z-çizelgesi verilen z değerinden geriye doğru eğrinin altında kalan alanı vermektedir. Bizim ise ileriye doğru eğrinin altında kalan alana ihtiyacımız vardır. Bu noktada önemli bir bilgiye sahip olduğumuzu unutmamalıyız: normal eğrisinin altında kalan alan 1'e eşittir. Bu durumda z puanına karşılık gelen alanı çizelgeden okuduktan sonra eğer bu değeri 1'den çıkarırsak aradığımız alanı elde etmiş oluruz. Çizelgede $z = 0.90$ değerine 0.8159 değeri karşılık gelmektedir. Bu durumda aradığımız alan $1 - 0.8159 = 0.1841$ olarak elde edilir. Yani Hyades kümesinden rastgele bir yıldız seçildiğinde dikine hızının 42 km s^{-1} den daha büyük olması olasılığı %18.41'dir.

Œimdi ğrendiđiniz bilgileri kullanarak aŒađıdaki soruyu özmeye alıŒın.

Soru 11.1 Hyades kümesinden rastgele bir yıldız seçildiđinde dikine hızının 36 ile 39 km s⁻¹ arasında olması olasılıđı nedir (kümenin dikine hız dađılımının normal dađılıma sahip olduđunu kabul ediniz)?