

Olasılık ve İstatistik

Fırat İsmailoğlu, PhD

İstatistik Giriş - Genel Kavramlar

Neden İstatistik?

Bir şeyi anlamak/öğrenmek için öncelikle onun hakkında **veri** toplamamız gerekir.

Örnek olarak;

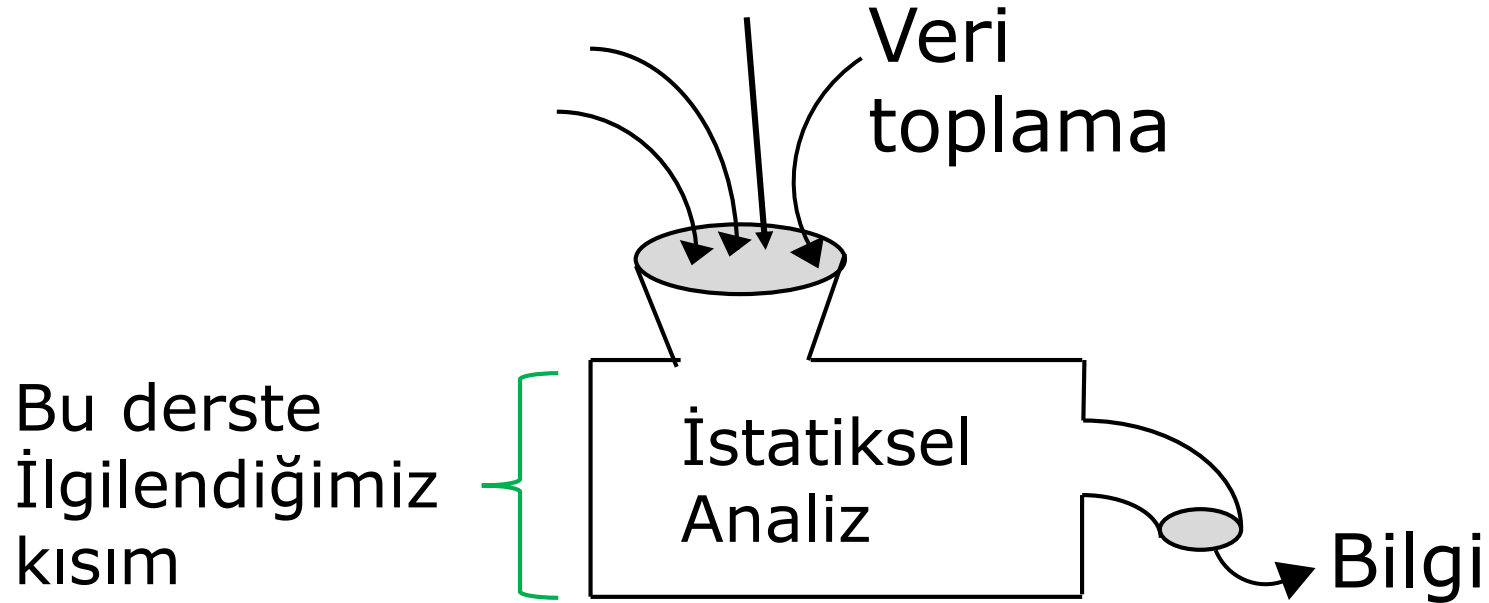
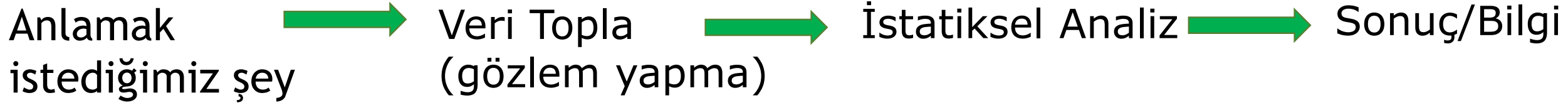
- Ekonomiye anlamak istiyorsak geçmiş yıllara ait ekonomik verileri toplarız (enflasyon değerleri, döviz kurları)
- Bir otomobilin ne kadar yakıt tükettiğini anlamak için önceki kullanımlara ait kilometre-yakıt verileri
- Bir ilacın işe yarayıp yaramadığını anlamak için bu ilacı kullanan ve kullanmayan hastalardan veri toplarız.
- Bir yolun günün hangi saatlerinde yoğun olduğunu anlamak için o yoldan daha önce geçen taşıtların verilerini toplarız.



Neden İstatistik?

Toplanan veriyi anlamak, bu veriden sonuçlar çıkarmak içinse istatistiğe ihtiyaç duyarız. Böylece herhangi bir şeyi anlamamanın yolu istatistikten geçer.

Kısaca; bir şeyi anlamak için veriye ihtiyacımız vardır; veriyi analiz etmek için ise istatistiğe ihtiyacımız vardır.



İstatiksel Analiz

Tanımlayıcı İstatistik (Descriptive Statistics)

Elimizdeki veriyi (örneği) tanımlar, özetlememizi sağlar.

Tanımlayıcı istatistik iki şeyi ölçer:

1. Merkezi eğilim (Central tendency): Ortalama, medyan, mode..
2. Dağınıklık (varyasyon): standart sapma, kurtosis

Çıkarımsal İstatistik (Inferential Statistics)

Elimizdeki verinin toplandığı popülasyon ile ilgili (büyük resim ile ilgili) çıkarım yapmamızı sağlar. Genel bir yargıya varmamızı, elimizdeki verinin ötesine geçmemizi sağlar.

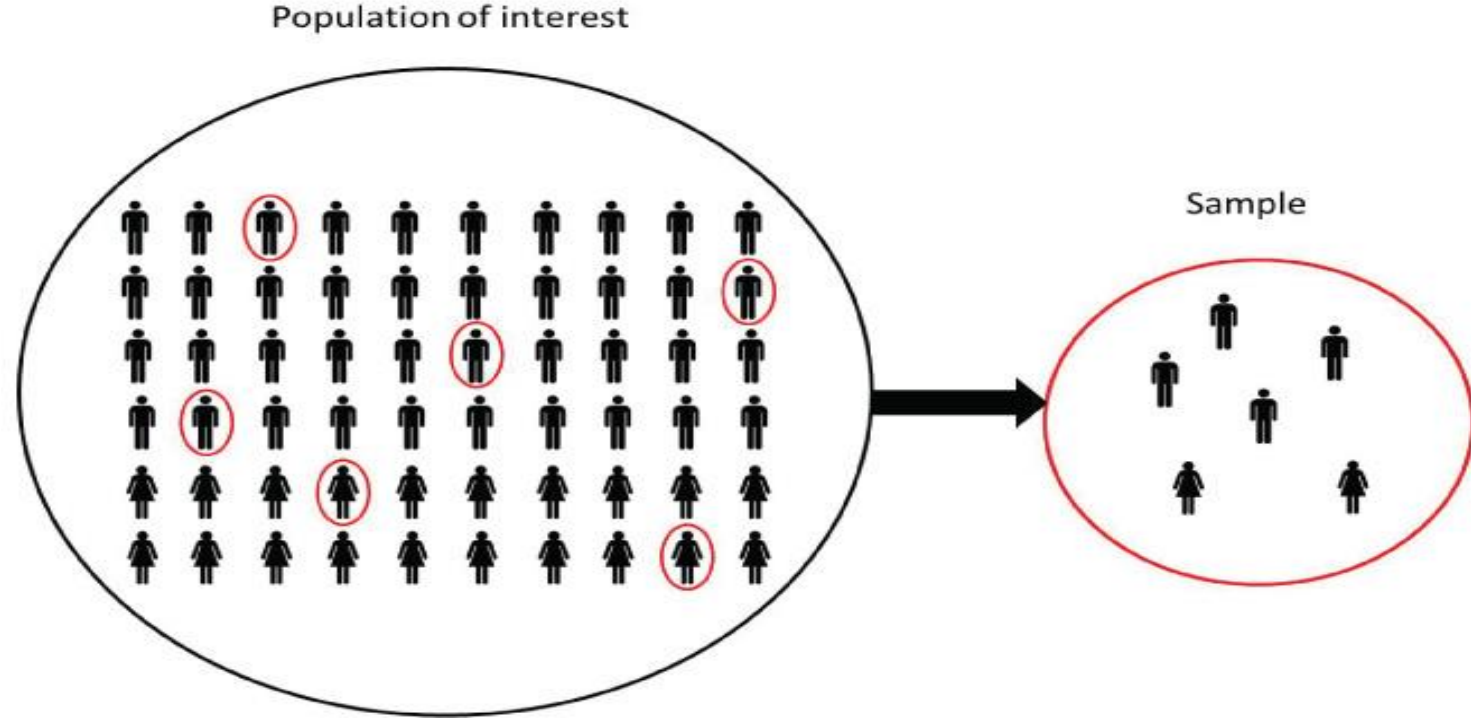


Tanımlayıcı İstatistik **VS** Çıkarımsal İstatistik

Bu iki istatistik türünün farkını anlamak için öncelikle poplasyon- örnek kavramlarını bilmeliyiz.

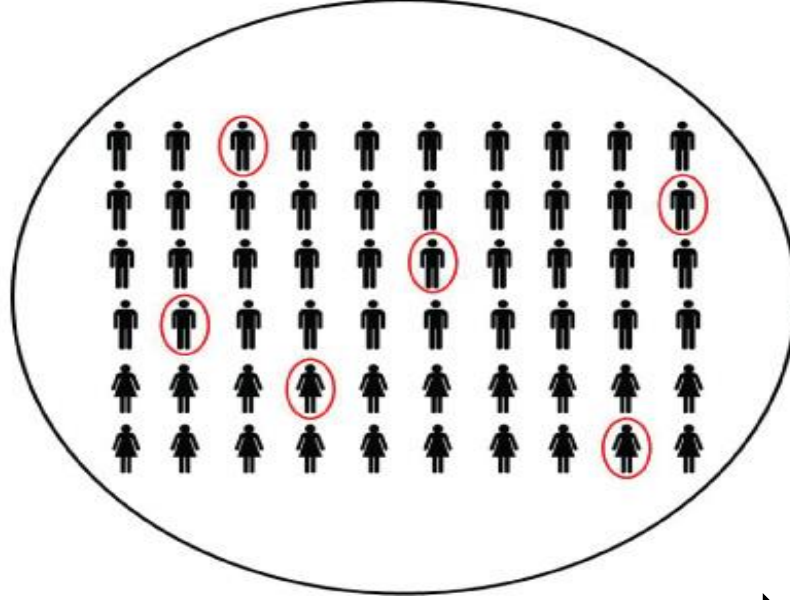
Populasyon : İlgilendiğimiz, çıkarım yapmak istediğimiz, elemanların tamamının kümesini kast eder.

Örnek Küme (örneklem, numune, sample): Populasyondan alınan herhangi bir alt kümedir; elimizdeki veri setidir.

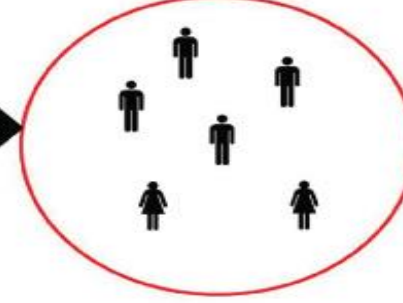


Tanımlayıcı İstatistik VS Çıkarımsal İstatistik

Populasyon



Örnek küme (sample)



Türkiye'deki tüm seçmenler



Her ilden rastgele seçilmiş 100 kişinin oluşturduğu topluluk

Belirli bir otobandan geçen tüm arabalar



Bu otobandan saat 14.00-15.00 arası geçen arabalar

Bir fabrikanın ürettiği tüm mallar



Fabrikanın ürettiği mallar içinden rastgele seçilen 5000 ürün

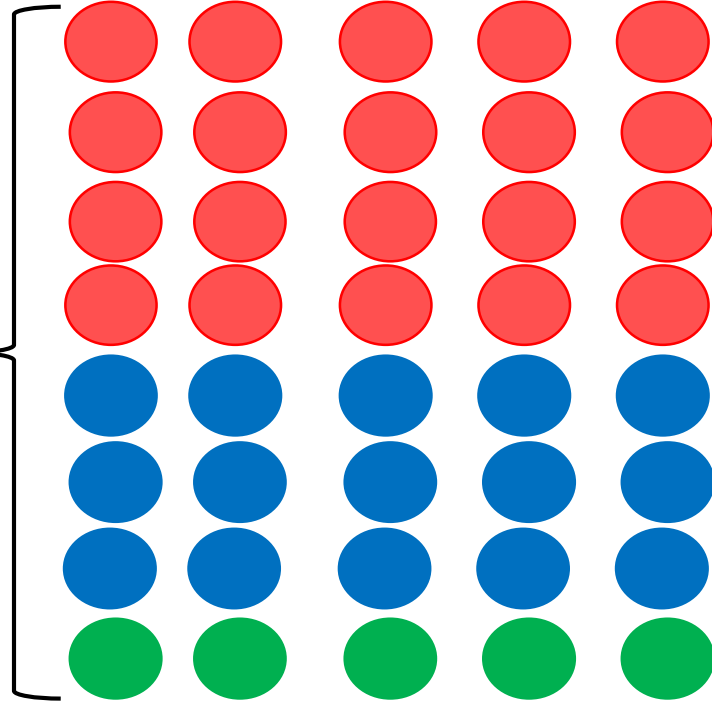
Populasyondan Örnek Nasıl Seçilmelidir?

Bir populasyondan bir örnek (sample) seçilerken iki şeye dikkat edilir:

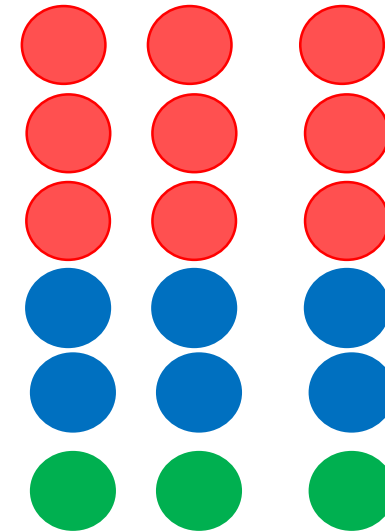
1. Seçim *rastgele* (random) yapılmalıdır. Bu, populasyondaki her bir elemanın eşit seçim şansının olması demektir.
2. Örnek, elde edildiği populusyona benzemelidir. Yani populusyonu temsil edebilmelidir (representative olmalı).

Populusyon

Populusyonda
sırasıyla en çok
kırmızı, mavi
ve yeşil toplar
var.



Örnek küme (sample)



Örnekte de
sırasıyla en çok
kırmızı, mavi
ve yeşil toplar
var.

Tanımlayıcı İstatistik **VS** Çıkarımsal İstatistik

Tanımlayıcı istatistik, örnek kümenin (elimizdeki verinin) merkezinin ne olduğunu (en çok hangi elemanlar gözlemlendiğini) ve bu elemanların küme içinde nasıl dağıldığını anlatır.

Tanımlayıcı istatistik içerisinde şu kavramları göreceğiz:

- ortalama, medyan, mod;
- varyasyon, standart sapma.

Çıkarımsal istatistik ise örneğin alındığı populasyon hakkında çıkarım yapmamızı sağlar.

Çıkarımsal istatistik içerisinde şu kavramları göreceğiz:

- hipotez testi,
- güven aralığı,
- regresyon.

ör. Tanımlayıcı istatistik: Bu verideki (elimizdeki örnekte) x değişkenin ortalaması 44.3 tür.

Çıkarımsal istatistik: x değişkeninin bu verinin alındığı populasyondaki ortalaması (%95 kesinlikle) $[37.8, 51.27]$ aralığındadır.



Neyi Ölçeceğiz?

Veri toplarken öncelikle neyi ölçeceğimize karar vermemiz gerekir. Ölçeceğimiz şey, anlamak istediğimiz şey hakkında ayırt edici, bilgi verici olmalıdır.

Örnek olarak,

- Belirli bir yolun ne kadar yoğun olduğunu anlamak istiyorsak, o yoldan gün içinde geçen araba sayısını ölçebiliriz.
- Bir semtteki evlerin en çok hangi renkte olduğunu anlamak istiyorsak, o semtteki evlerin rengini ölçebiliriz.
- Bir sınıftaki öğrencilerin çalışkanlık düzeylerini anlamak istiyorsak, her bir öğrencinin günde toplam kaç saat çalıştığını ya da derslerde aldıkları notları ölçebiliriz.

Ölçtüğümüz şeye *değişken (variable)* diyeceğiz. Burada önemli olan değişkenin populasyonun farklı elemanları arasında değişkenlik göstermesi; böylece populasyonun elemanlarını karakterize etmesidir!

Örneğin, bir sınıfın tüm öğrencilerinin istatistik dersinden AA alması durumunda bu ders notu değişken olamaz, öğrencileri karakterize edemez. Bu durumda kendimize başka değişken bulmamız gerekir.



Değişken Türleri (Neden - Sonuç İlişkisine Göre)

Değişkenleri neden - sonuç ilişkisine göre ikiye ayıracamız: bağımsız değişkenler (independent variables) ve bağımlı değişkenler (dependent variables).

Kısaca:

Bağımsız değişken = neden

Bağımlı değişken = sonuç

ör. Kafeinin iştaha etki edip etmediğini anlamak istiyorsunuz.

Alınan kafein miktarı bağımsız değişken, açlığınızın miktarı bağımlı değişken.

ör. Evin büyüklüğü, evin merkeze uzaklığı, evin yaşı : bağımsız değişkenler

Evin fiyatı: bağımlı değişken

ör. Sınava kaç saat çalışıldığı: bağımsız değişken

Sınavdan alınan not: bağımlı değişken

ör. Yaş : bağımsız değişken

Göz etrafındaki çizgi miktarı: bağımlı değişken

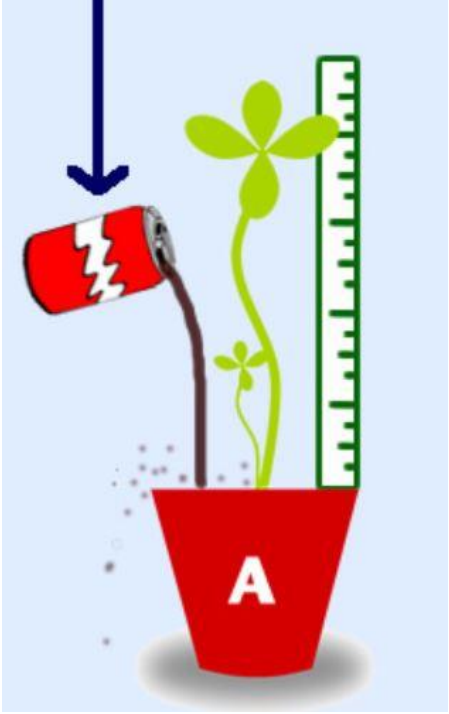


Bağımsız Değişken - Bağımlı Değişken

$$Y = 3X - 4$$

Bağımlı
değişken

Bağımsız
değişken



Kola



Meyve suyu



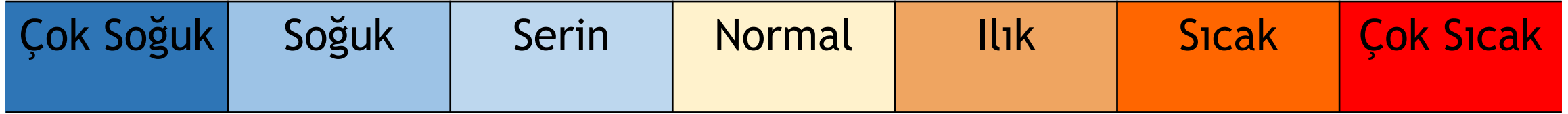
Su

Diyelimki cesitli sivilarin bir bitkinin büyümesine olan etkilerini ölçmek istiyoruz. Bu deneyde sıvı türü bağımsız değişken, bitkinin boyu bağımlı değişken olur.

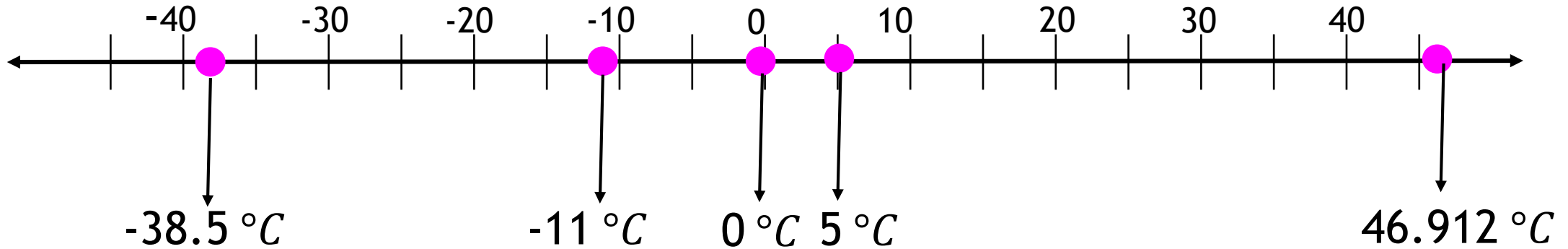
Değişken Türleri (Ölçüm Seviyesine Göre)

Diyelimki havanın sıcaklığı bizim değişkenimiz. Bu sıcaklığı iki şekilde ölçebiliriz:

1. **kategorik** olarak: çok soğuk, soğuk, serin, normal, ılık, sıcak, çok sıcak

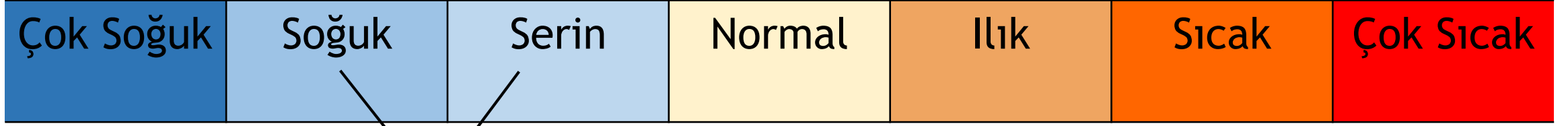


2. **sayısal** olarak: -38.5°C , -11°C , 0°C , 5°C , 46.912°C ,



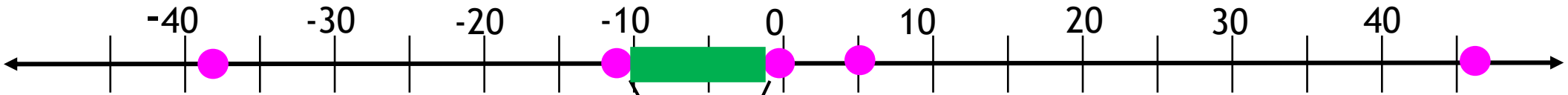
Değişken Türleri (Ölçüm Seviyesine Göre)

Kategorik ve sayısal değişkenler arasındaki temel fark: kategorik değişkende herhangi iki değişken arası bir başka değişken olacağı garanti değildir.



Soğuk ve Serin değişkenleri arasında bir başka değişken yok. (Soğuktan serine atlıyor)

Sayısal türdeki değişkenlerde ise iki değişken arası her zaman bir başka değişken bulunur. Bu yüzden sayısal değişkenin bir adı da sürekli (continious) değişkendir.



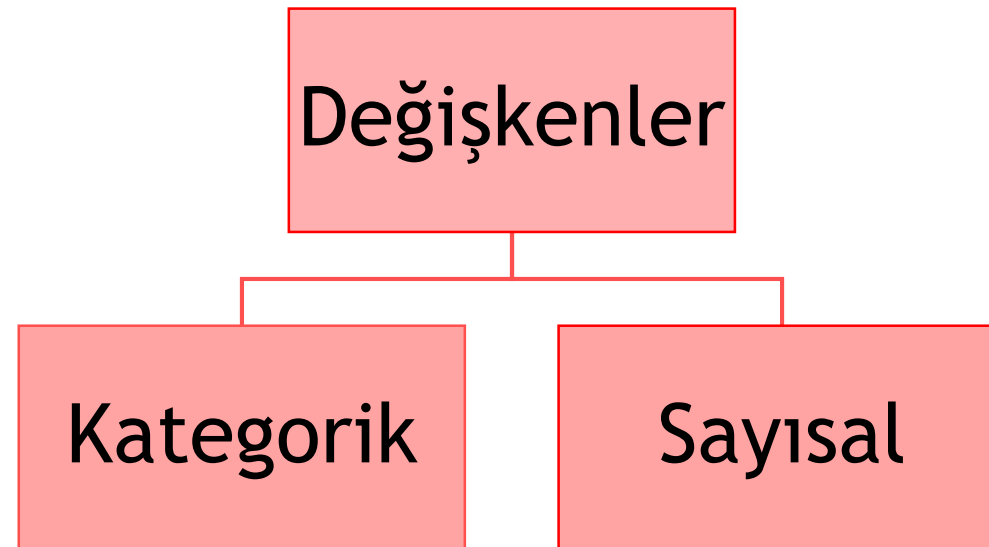
-11 ile 0 değişkenleri arasında sonsuz değişken var.

Değişken Türleri (Ölçüm Seviyesine Göre)

Sayısal türdeki değişken daha değerlidir, çünkü yapılan ölçüm çok daha fazla kesinlik bildirir. **ör:** Hava 34.12°C . Bu yüzden sayısal veriye daha kaliteli diyebiliriz.

Kategorik olarak ölçülmüş değişken ise daha az bilgi verir: Hava ılık (fakat kime göre, neye göre??)

Şu halde değişkenleri, neden sonuç ilişkisine ilaveten ölçüm seviyelerine göre de iki ana grup altında inceleyebiliriz: kategorik, sayısal.



Kategorik Değişken

Kategorik değişkenleri de kendi içinde iki grupta inceleyebiliriz:
Sembolik (Nominal) ve Sıralı (Ordinal).

Sembolik Kategorik Değişken

Sembolik değişkenler birer isimdir. Bu değişkenleri sıralayamayız, aritmetik işlem yapamayız.

ör. Renkler: Bir torbadaki topların rengi değişebilir: mavi, kırmızı, yeşil.

Cinsiyet: Bir sınıftaki öğrencilerin cinsiyeti değişebilir: erkek, kız.

Hasta şikayeti: Hastaların şikayeti değişebilir: mide bulantısı, ateş, tansiyon.

(anlamsızdır).

Sembolik kategorik değişkenler ile yapabileceğimiz tek işlem: iki değişken aynı mı değil mi diye bakmaktır. Yani burada tanımlı tek operasyon: $=$ ve \neq

ör. mavi $=$ mavi, mavi \neq kırmızı



Kategorik Değişken

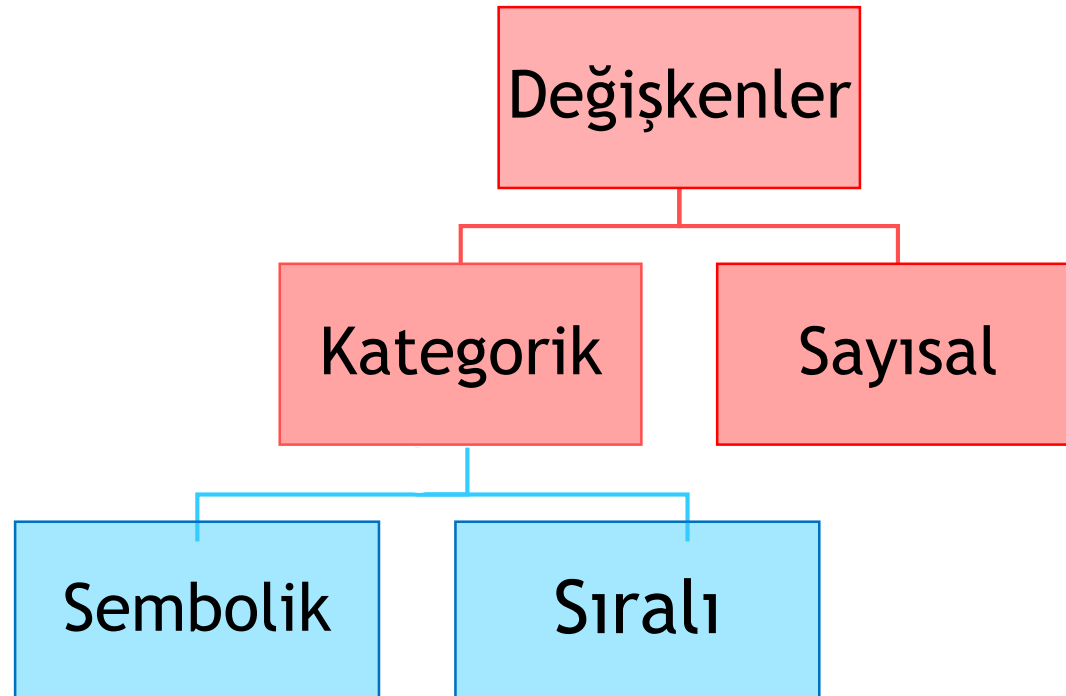
Sıralı Kategorik Değişken

Sıralı kategorik değişkende ise sembolik olandan farklı olarak değişkenleri sıralayabiliriz. Bu bize sıralama $\leq, \geq, >, <$ operasyonlarını kullanırır.

ör. çok soğuk $<$ soğuk $<$ serin $<$ normal $<$ ılık, $<$ sıcak $<$ çok sıcak

ör. fakir $<$ orta sınıf $<$ zengin

ör. xs $<$ s $<$ m $<$ l $<$ xl $<$ xxl

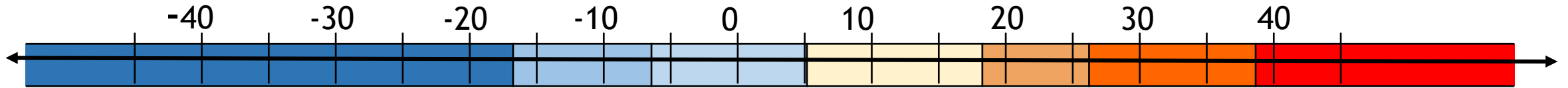


Sayısal (Sürekli) Değişken

Sayısal değişkenler genelde bir ölçüm aleti kullanılarak elde edilen değişkenlerdir. Örneğin, kilo, boy, büyük tansiyon, kandaki demir miktarı, hava sıcaklığı...

Sayısal değişkenler de sıralamaya ek olarak bütün aritmetik işlemleri yapabiliriz. Yani iki sayısal değişkeni toplayabilir, çıkartabilir, birbirine bölebilir yada bu değişkenleri çarpabiliriz. Örneğin $27 \text{ kg} + 41 \text{ kg} = 68 \text{ kg}$

Not: Sayısal değişkenler, sıralı kategorik değişkene dönüştürülebilir:



Çok Soğuk	Soğuk	Serin	Normal	Ilık	Sıcak	Çok Sıcak
$(-\infty, -15]$	$(-15, -5]$	$(-5, 7]$	$(7, 20]$	$(20, 27]$	$(27, 40]$	$(40, \infty)$