Data Analitiği

Veri Analizi

“Dr. Cahit Karakuş”

Data Analysis:
Methodological Big Picture
Statistical Analysis
Probability, Inference & Estimation
Notation

Veri Analizi:
Metodolojik Büyük Resim
İstatistiksel Analiz Olasılık, Çıkarım Ve Tahmin Notasyonu

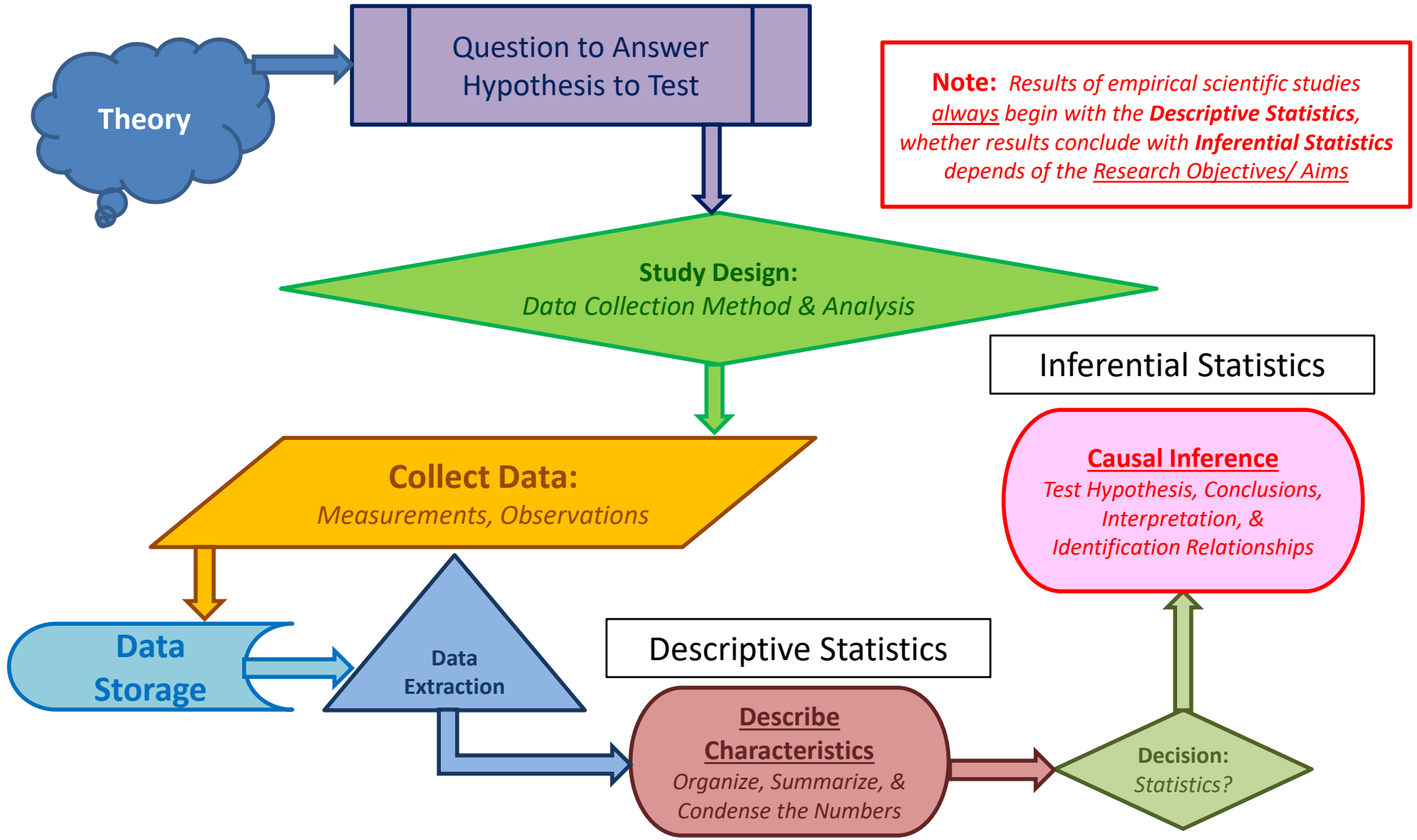
Giriş

Tanımlayıcı ve Çıkarımsal İstatistikler (Descriptive vs. Inferential Statistics)

- **Tanımlayıcı:** Medyan örneğinde olduğu gibi sahip olduğunuz ancak genelleştirilemeyen verileri tanımlar
- **Çıkarımsal (Inferential):** T-testi örneğinde olduğu gibi verilerimizin ötesinde popülasyon hakkında çıkarımlar sağlayan Makine Öğrenimi ve Tahmin için yararlanılan tekniklerdir.

Data Analysis:

In the Big Picture of Methodology



Examples of Business Questions

- **Basit (açıklayıcı) İstatistikler:** En karlı müşteriler kimlerdir?
- **Hipotez Testi:** Karlı müşterilerin şirketi için bir değer farkı var mı?
- **Segmentasyon/Sınıflandırma:** Karlı müşterilerin ortak özellikleri nelerdir?
- **Tahmin:** Bir yeni müşteri karlı bir müşteri olacak mı? Eğer öyleyse, ne kadar karlı?

Applying techniques

- Most business questions are causal: what would happen if? (e.g. I show this ad)
- But its easier to ask correlational questions, (what happened in this past when I showed this ad).
- **Supervised Learning:**
  - Classification and Regression
- **Unsupervised Learning:**
  - Clustering and Dimension reduction
- Note: Unsupervised Learning is often used inside a larger Supervised learning problem.
  - E.g. auto-encoders for image recognition neural nets.

Applying techniques

- **Supervised Learning: Classificaiton and Regression**
  - kNN (k Nearest Neighbors)
  - Naïve Bayes
  - Logistic Regression
  - Support Vector Machines
  - Random Forests

- **Unsupervised Learning:**
  - Clustering
  - Factor analysis
  - Latent Dirichlet Allocation (Doğal dil işlemede kullanılan her belgenin bir konu koleksiyonu kabul edildiği ve belgedeki her kelimenin konulardan birine karşılık geldiği en basit kabul edilen bir konu modelleme örneğidir.)

Machine Learning and Data Analytics

I. Machine learning and data analysis tasks

II. Classification
  - Classification tasks
  - Building a classifier
  - Evaluating a classifier

III. Pattern learning and clustering
  - Pattern detection
  - Pattern learning and pattern discovery
  - Clustering
    - K-means clustering

IV. Causal discovery
  - Correlation
  - Causation
  - Causal models
    - Bayesian networks
    - Markov networks

V. Simulation and modeling

VI. Practical use of machine learning and data analysis

Different Data Analysis Tasks

- **Classification**
  - Yeni bir örnek için bir kategori (yani bir sınıf) atanır.
- **Clustering**
  - Bir dizi örnekle kümeler (yani gruplar) oluşturulur.
- **Pattern detection**
  - Zamansal veya uzamsal verilerdeki düzenlilikler (yani kalıplar) tanımlanır.
- **Simulation**
  - Toplanan gözlemlere benzer veriler üretebilen matematiksel formüller tanımlanır.

Different Data Analysis Tasks

- **Classification**
- **Clustering**
- **Pattern detection**
- **Causal discovery**
- **Simulation**
- **…**

- Her görev türü, ihtiyaç duydukları veri türleri ve ürettikleri çıktı türleri ile karakterize edilir.
- Her görev türü farklı algoritmalar kullanır.

Öğrenme Yaklaşımları

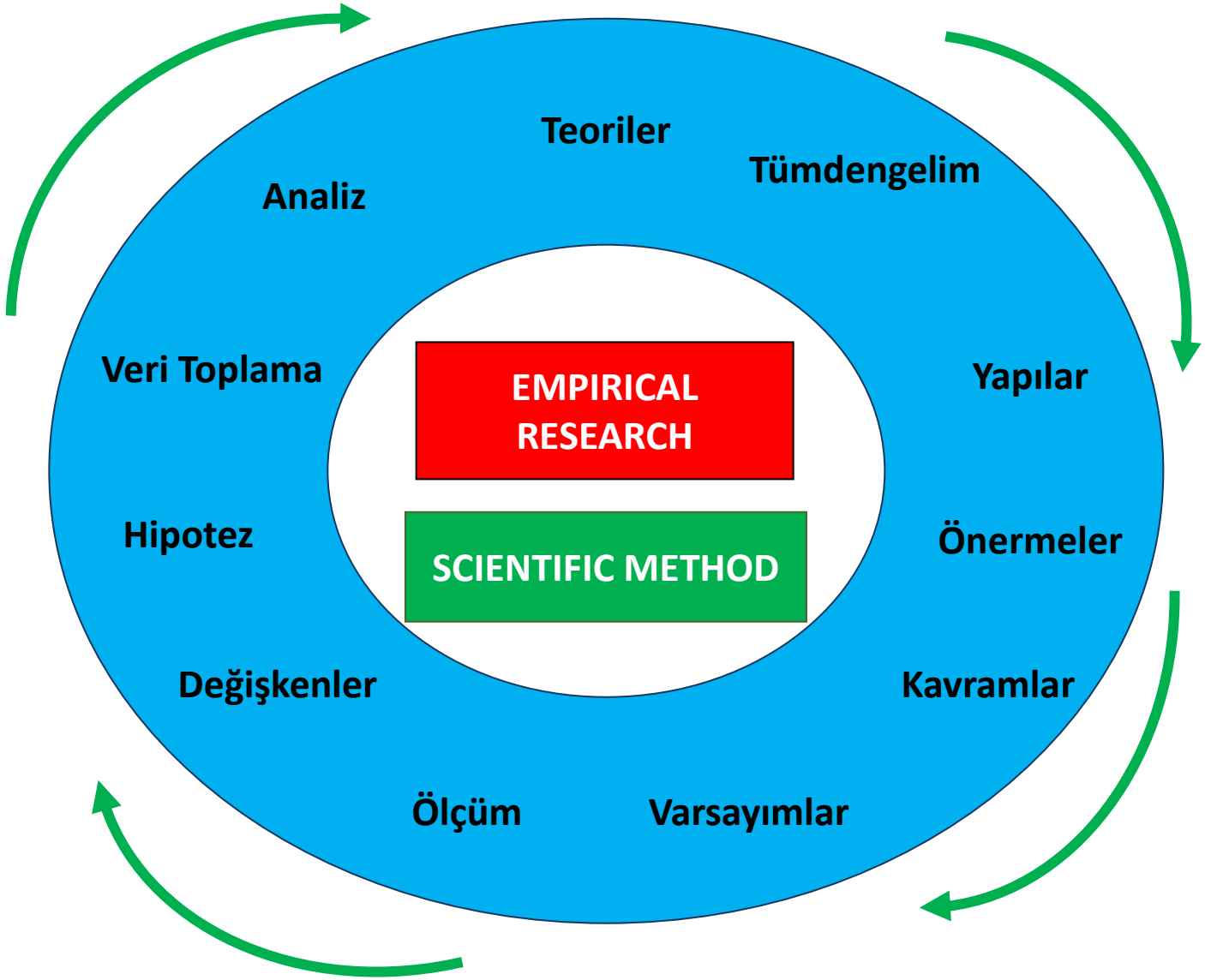
Supervised Learning

- Eğitim verileri, öğrenme sistemine yardımcı olacak bilgilerle açıklanır

Unsupervised Learning

- Eğitim verilerine, öğrenme sistemine yardımcı olmak için herhangi bir ek bilgi eklenmez.

Ampirik Araştırmanın Boyutları: Teorik olandan analitik olana bir hareket



Tümdengelim

- Tümdengelim ya da Dedüksiyon (Dedüktif) Akıl Yürütme, felsefe ve mantıkta sahip olunan genel verilerden yola çıkarak özel sonuçlar çıkarma yöntemidir. Tümdengelimin aksi tümevarım metodudur.
- Genel bir yasa ya da yargıdan yola çıktığı için adına tümdengelim denilir. Biri yasa ya da yargı olmak üzere en az iki öncül ve bu öncüllerden çıkarılması gereken bir sonuçtan oluşur. Dayandığı yasa ne kadar doğru ve kuvvetliyse o kadar geçerli bir akıl yürütmedir.
- Tümdengelim bilgimizi arttırıcı değil, elde olan bilgileri çözümleyici bir yöntemdir.
- Dayandığı kaynak ne kadar doğruysa o kadar geçerli bir akıl yürütmedir. Tümdengelim akıl yürütme biçiminde sonuç öncülleri aşmaz yani yeni bir şey öğrenmez yeni bir buluş yapmayız.

Örneğin:

- İnsan ölümlüdür, Ahmet insandır, Öyleyse Ahmet de ölümlüdür.
- Bütün metaller ısıtıldığında genleşir. Bakır bir metaldir. O hâlde, bakır ısıtıldığında genleşir.
- Tüm Trakyalılar gri gözlüdür. Einstein gri gözlüydü. O hâlde, Einstein Trakyalıdır. tüm Trakyalılar gri gözlüdür tümel önermesi de yanlıştır. Einstein rast gele gri gözlü olabilir ama bu tek özellik Einstein'ın Trakyalı olması için yeterli değildir. Aristoteles klasik mantığın içine sadece tümdengelim akıl yürütme yöntemini dâhil etmiştir ve tümdengelime önem vermiştir. Aristoteles'e göre, zihin hakikati bu yol ile arar.
- Bütün kuşlar uçabilir. Kartal bir kuştur. O hâlde, Kartal da uçabilir. Bu örnekte genelden özele, bir akıl yürütme kullanılmıştır.
- Bütün kadınlar çiçektir. Bütün çiçekler güzel kokar. O hâlde, bütün kadınlar güzel kokar. Bu örnekte genelden genele doğru bir akıl yürütme kullanılmıştır.

Tümevarım

- Tümevarım, özelden genele parçalardan bütüne doğru yapılan bir akıl yürütme biçimidir.
- Bu akıl yürütmeler zorunlu olarak geçerli değildirler.
- Olma olasılığı doğru veya geçerli akıl yürütmeleridir.
- Ayrıca bu akıl yürütme biçiminde yeni ve bilinmeyen sonuçlara ulaşabiliriz.
- Bu nedenle deneysel bilimler, olaylardan yasalara götüren bir yöntem olan tümevarım yöntemini kullanırlar.
- Bir Alman: “Tanıdığım yüzden fazla Türk arkadaşım var ve hepsi tembel insanlar. Bütün Türkler tembeldir” örneği zayıf bir tümevarım örneğidir. Çünkü Almanya’da yaşayan bütün Türkleri tanıması gerekirdi. Bu nedenle bu zayıf bir tümevarım çıkarımıdır. Diyelim ki bütün Türkleri inceledi ve gerçekten bütün Türkler tembel insanlar. Bu durumda doğru bir tümevarım akıl yürütme yapmış olur.
- Tümevarım çıkarımımızın yanlışlanma ihtimali her an vardır. Yani sonuç doğru olabilir ama yine de zorunluluk taşımaz.

Tümevarım

- Tümevarım varsayımsal bir genellemedir. Sonucun doğruluğu hiçbir zaman kesin değildir. Tümevarım akıl yürütme yönteminde öncüllerin doğruluğunu kabul etsek bile sonucun doğruluğunu kabul etmeyebiliriz. Yani bütün öncüller doğru olsa bile sonuç yanlış olabilir. Ancak yanlış olmayadabilir.

Güçlü Tümevarım Örnekleri;

- “Ali henüz bir haftalık bebektir, o halde Ali henüz emekleyemez”
- “Kapalı bir torba olduğunu düşünelim ve bu torbada otuz adet rakam var (tombala oyunu gibi) torbaya elimizi attık torbaya karıştırdık çektik 3 rakamı çıktı tekrar çektik yine 3 çıktı tekrar çektik yine 3 çıktı torbadan çektiğimiz rakam ve çekmeye devam ettik Yirminci çekişimizde de 3 rakamı çıktı 30 rakamında 3 olduğunu düşünürüz bu güçlü bir tümevarım örneğidir”

Zayıf Tümevarım Örnekleri;

- “İnsanlar piyango biletini çok defa büyük ikramiye çıkarmış Nimet Abla’dan almak isterler bu zayıf bir tümevarımlı akıl yürütme örneğidir”
- “Mehmet Adanalıdır. Adanalı insanlar kavgayı sever. O halde Mehmet kavgayı sever.”

ANALOJİ (BENZETİŞ AKIL YÜRÜTME)

- Ortak bir veya birden fazla özellikten yola çıkılarak A için verilen yargıyı B için de vermek.
- Benzerliklere bakarak sonuca varmak.
- Bilinmeyen bir olgunun bilinen bir olguyla açıklanması.
- ‘Aslan Mican geliyor” dediğimizde Mican ve Aslan arasında analoji yaparız yani Mican’ın cesur, güçlü, dayanıklı biri olduğunu söylemek isteriz.
- ‘Melek gibi bir kadındı’ dediğimizde Melek ile o kadın arasında analoji yaparız. Kadının masum, temiz ve iyiniyetli olduğunu söylemek isteriz.

HEPTENGİTMELİ AKIL YÜRÜTME

- Heptengitmeli akıl yürütmede öncül önermelerinin tümünün doğru olması durumunu en iyi açıklayan önermeye sonuç önermesi olarak ulaşılmaya çalışılır. Tümevarımlı akıl yürütmedeki gibi öncüller ne kadar doğru olursa olsun sonucun doğru olduğuna kesin olarak emin olamayız.
- “Sabah uykumuzdan uyandık dışarıdan su sesi geliyor ve penceremize de su damlaları düşüyor. Bu durumda büyük ihtimalle yağmur yağdığını düşünürüz ama bu başka bir durumda olabilir. Yine de bu en akla yakın düşüncedir”

AKIL YÜRÜTME YÖNTEMLERİ

V TÜMDENGELİM
(Dedüktif Akıl Yürütme)

- Bir şeyin doğruluğnu başka bir veya daha fazla şeye dayanarak ileri sürme, akıl yürütmedir
- Genelden özeli çıkarsar.
- Yeni bir bilgi vermez. Sadece öncülleri açık hale getirir ve tikel hakkında bilgi verir.
- Bilgimizi arttırmaz. Elde ettiğimiz bilgileri çözümlememize yardımcı olur
- Verilen bir kaç öncülden bir sonuç çıkarılır.
- Sonuç, öncülleri ne aşar ne de öncüllere yeni bir şey katar.
- Dayandığı yasa ne kadar doğru ve kuvvetliyse o kadar geçerli bir akıl yürütmedir.
- Genelden özel/özellere gider.

Λ TÜMEVARIM
(İndüktif Akıl Yürütme)

- Öncüllerden tümel veya tikel bir sonuca varılır
- Sonuç, öncüllerle sınırlı kalmayarak bizi, öncülleri aşan bir bilgiye götürür.
- Bu akıl yürütmede öncüller sonuç için bize bir dayanak sağlar ama sonucu zorunlu kılmaz
- Gözlem sayımız arttıkça, tümevarımın doğu olma olasıılığı artar.
- Sonucun doğruluğu hiçbir zaman kesin değildir
- Tümevarımda öncüllerin doğruluğunu kabul etsek bile sonucun doğruluğunu kabul etmeyebiliriz. Bütün öncüller doğru olsa bile sonuç yanlış olabilir.
- Özel/Özellerden genele gider.

= BENZETİŞ
(Analoji)

- Ortak özelliklerin sayısı ne kadar fazla ise sonucun doğru olma olasılığı artar.
- Analojide de tümevaım gibi sonucun doğruluğu hiçbir zaman kesin değildir.
- Blinen bir olgunun niteliklerinin açımlanışıyla bilinmeyeni ifade etme durumu.
- Kıyastır.
- Özelden özele gider.

Hipotez (Varsayımlar)

- Hipotez, bilimsel yöntemde olaylar arasında ilişkiler kurmak ve olayları bir nedene bağlamak üzere tasarlanan ve geçerli sayılan bir önermedir. Bilimsel bir ifadenin hipotez kabul edilebilmesi için sınanabilmesi gerekir.
- Henüz doğruluğu ispatlanmamış temel varsayımlardır.
- Hipotezler, ortaya konan problemle ilişkili olarak cevabı aranacak sorulardır.
- Varsayım veya hipotez, bilimsel yöntemde olaylar arasında ilişkiler kurmak ve olayları bir nedene bağlamak üzere tasarlanan ve geçerli sayılan bir önermedir.

Hipotez (Varsayımlar)

- İstatistiki hipotez (H0); tarafsızlık hipotezi, farksızlık hipotezi, Sıfır hipotezi. Bu kavrama göre varsayılan değişkenler arasında farklılık yoktur. Bu tür hipotez bilimsel araştırmalarda yan tutmamanın gereğidir.

- Araştırma hipotezi (H1); farklılık hipotezi, alternatif hipotez. Araştırmada daha önceki bilgilerin veya gözlemlerin doğruluğunu saptamak üzere oluşturulan, değişkenler arasında önemli ilişki olduğunu varsayan ifadelerdir. Hipotezler bilimsel yöntemlerle sınanıp geçerliliği değerlendirilir. Doğruluğu istatistiksel teknikler ile irdelenecek önermelere hipotez denir. H0 ve H1 şeklinde ifade edilir.

- H0 araştırma konusuna göre sıfır (null), yokluk, farksızlık önemsizlik veya eşitlik- hipotezi olarak bilinir.

- Hipotez problemi çözmek için yapılan araştırma ve gözlemler sonucu elde edilen bilgilerin yardımıyla kurulur.

- H1 ise H0 hipotezinin karşıt veya alternatif hipotezidir.

Aksiyom: Postulate (doğru varsayılan kanıtsız önerme)

- aksiyom sadece tanım içerirken postulatta bir şey koyutlanır.
- Aksiyoma örnek verelim: "nokta parçası olmayandır". Burda sadece bir tanım vardır, nokta şudur diye. Ya da "çizgi genişliği olmayan uzunluktur" gibi.
- postulatta ise bir şeyin varlığı koyutlanır, ileri sürülür. Örnk: "Paralel çizgiler kesişmezler" gibi. Dikkat ettiyseniz burda bir hüküm vardır “kesişmezler" diye.

- Aksiyom da postulat da tanıtlanamayacak denli kendiliğinden apaçıktırlar.

- Öklid bütün teoremlerini bunlar üzerine inşa eder. yani önce bir teorem — bir önerme yani — öne sürer, sonra da bunlar dolayısıyla isbatlar/tanıtlar.

Types of Data Analysis

Mekansal - Uzaysal veri analizi (Spatial data analysis)

Genellikle, aşağıdakiler gibi çeşitli operatörler (araçlar) ile koordinatların veya öznitelik değişkenlerinin manipülasyonlarını veya hesaplanmasını içerir:

Ölçüm
Sorgulama ve Seçim
Yeniden sınıflandırma
Arabelleğe alma
Kaplama
Ağ analizi

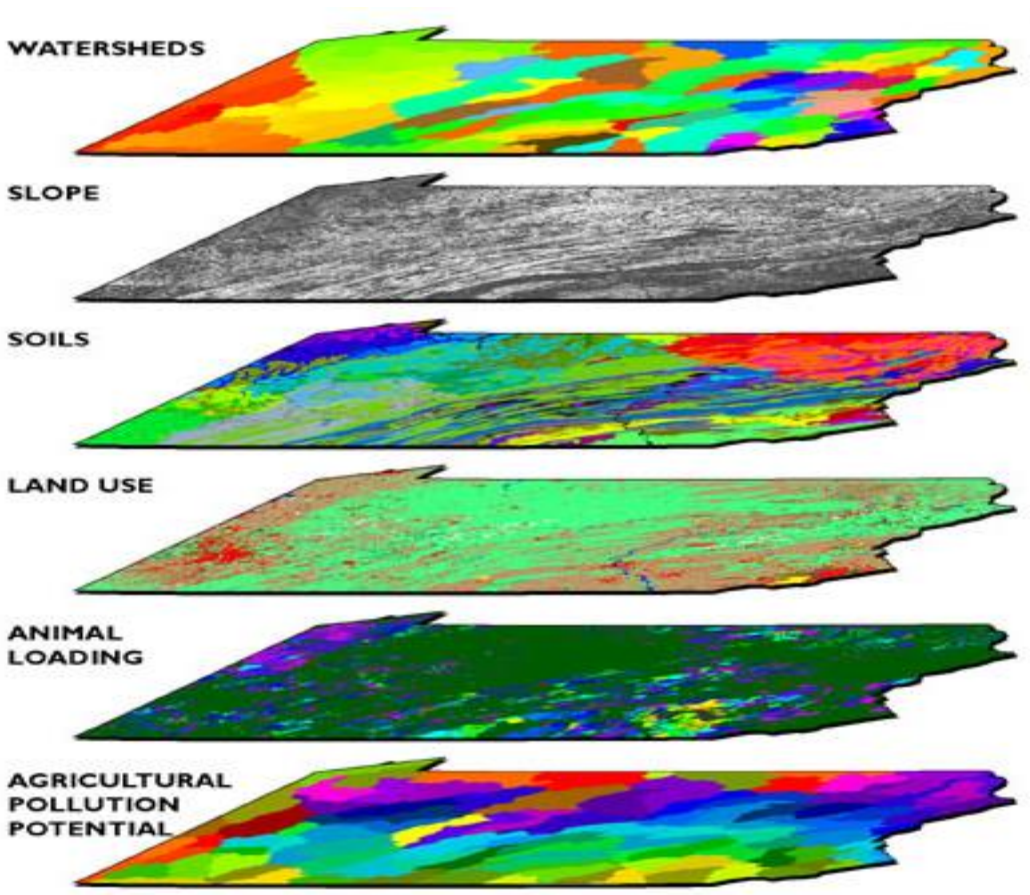
Kaplama

Farklı veri katmanlarının kombinasyonu
Hem uzamsal hem de öznitelik verileri birleştirilir

Veri katmanlarının ortak bir koordinat sistemi kullanmasını gerektirir

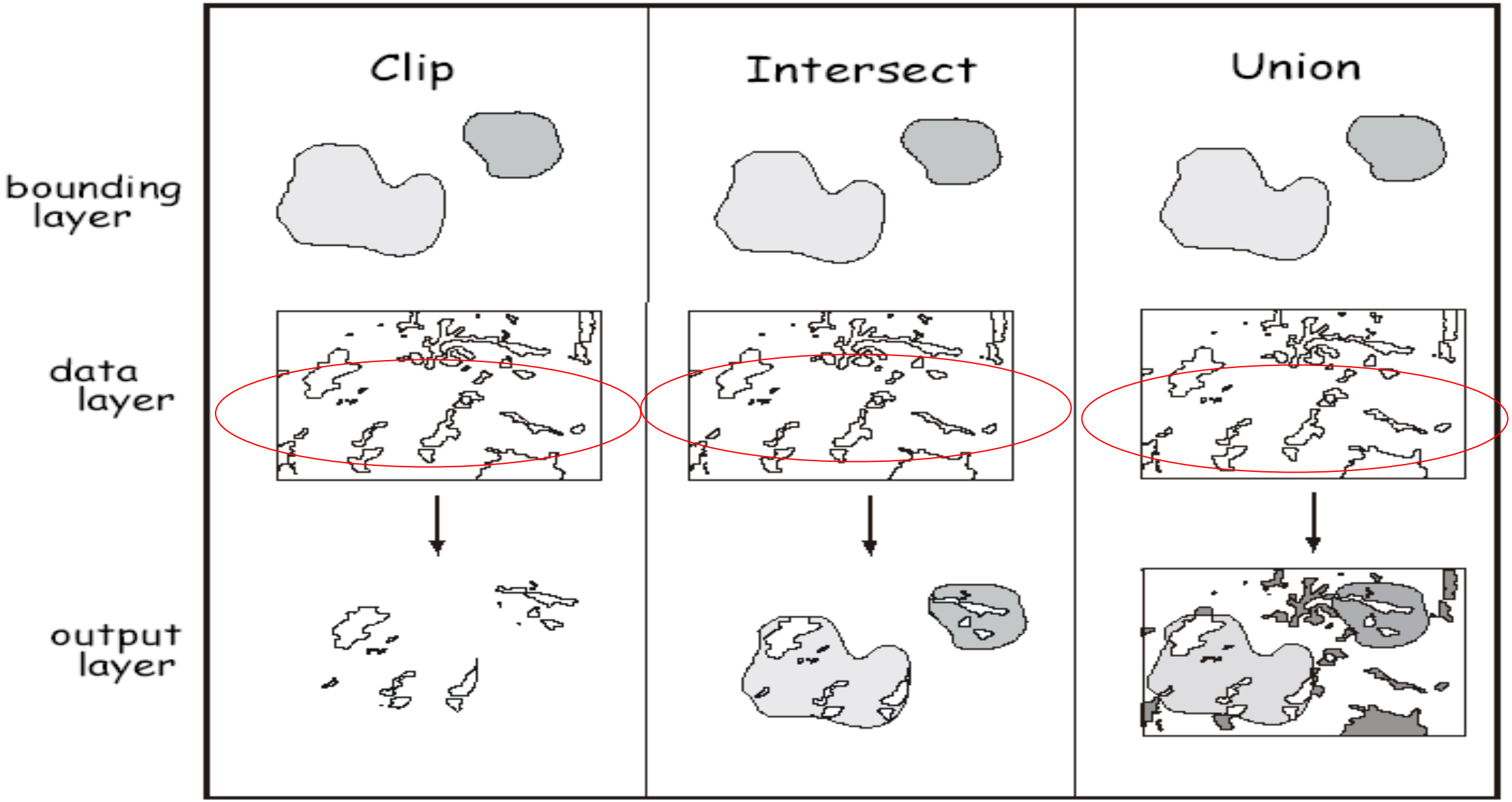
Yeni bir veri katmanı oluşturulurç

Watersheds: Su havzaları
Slope: Eğimli yerler
Soil:Arazi
Land use: Arazi kullanımı
Agricultural Pollution Potential: Arazi Kirlame Potansiyeli



Kırpma – Kesişim - Birleşim

- CLIP – Kırpma: Çerez kesici yaklaşımı. Sınırlayıcı çokgen, kırpılan ikinci katmanı tanımlar. Çıktı katmanına ne sınırlayıcı çokgen öznitelikleri ne de coğrafi (mekansal veriler) dahil edilmez.
- Kesişim: Her iki katmandaki verileri birleştirir, ancak yalnızca sınırlayıcı alan için(Sınırlayıcı çokgen ayrıca çıktı katmanını da tanımlar. Her iki katmandan gelen veriler birleştirilir. Sınırlayıcı katmanın (1. katman) dışındaki veriler atılır). Kavşak sırası önemli(A'dan B'ye veya B'den A'ya)
- Birleşim: Hem sınırlayıcı hem de veri katmanlarından tüm verileri içerir. Her katmandan alınan koordinat verilerinin birleştirilmesiyle yeni çokgenler oluşturulur.


Clip
Intersect
Union
bounding layer
data layer
output layer

Veri Analiz teknikleri

Veri Erişimi

- Veri ve Veri Kaynakları;
  - Veri haritası oluştur
    - Veri kaynakları
    - Veri Tabloları
    - Tablo İçerikleri
    - Veri sözlüğü
  - Bilgi İşlem Birimi ile işbirliği
  - Veri ve Veri teknolojileri

Veriye Erişim

- **Veri Kaynakları**
- Veri tabanları (Oracle, SQL, Access, Excel, Sybase, DB2, AS400 vb.)
- Veri tabanı kayıtlarından üretilmiş dosyalar (genelde metin formatında *.txt)
- Sistem raporları (Excel, print dosyaları, txt dosyalar)
- Dış kaynaklı veriler çok farklı formatta gelebilir

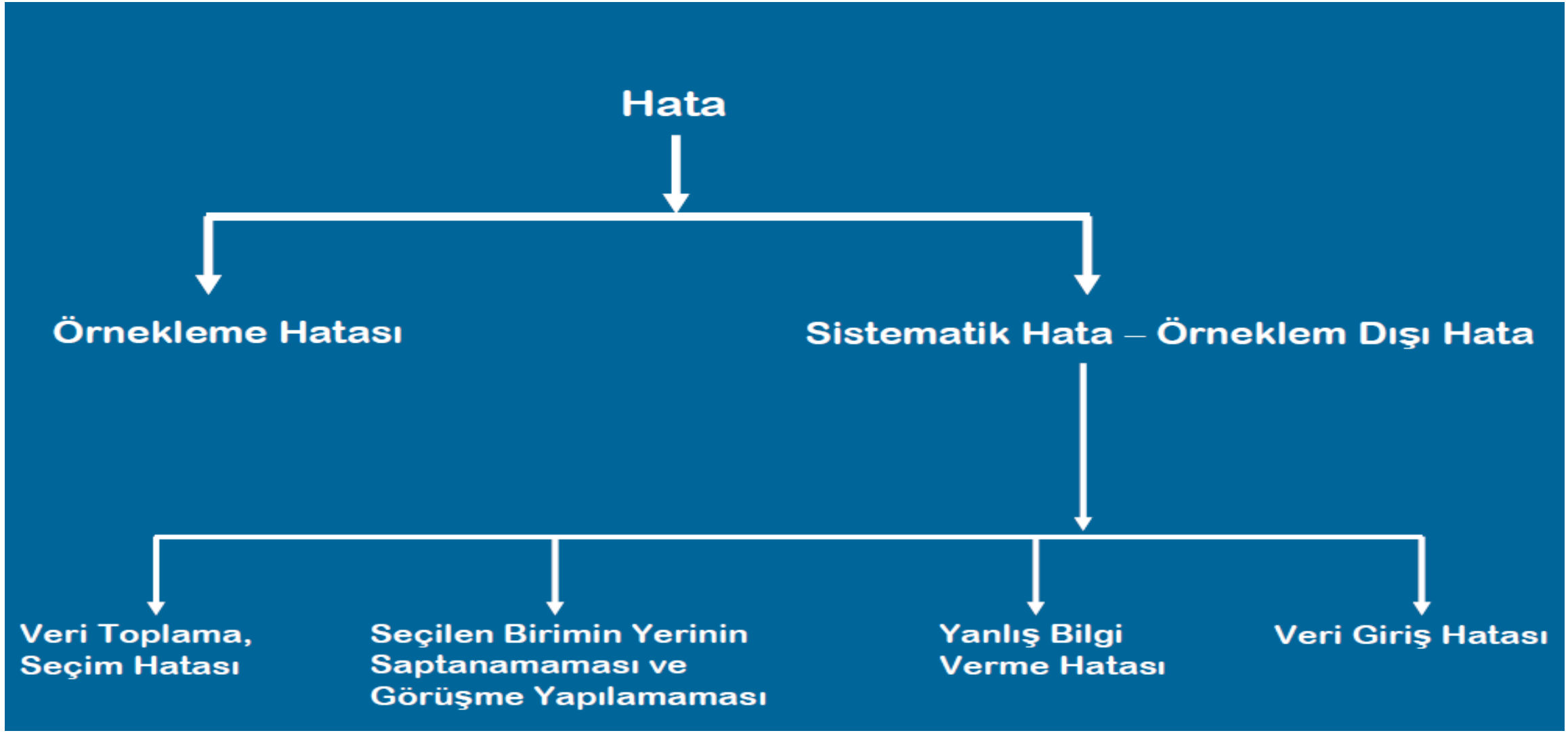
Veri Toplama

Sinyalleri ölçerseniz, yönetirsiniz

- Bilgi toplama kaynaklarına çevresel etkiler ve riskler
- Donanım kalitesi: kalibre, onay, yetkilendirme, verim, kayıp, ısısal davranış
- Veri toparlama, transferi, sınıflandırma ve saklama

Veri Doğrulama

- Yalnız analiz için gerekli olan tüm veri mi?
- Sayısal alanlarda sadece sayısal karakterler mi var?
- Mükerrer kayıt var mı?
- Geçersiz veri var mı?
- Veri alanları ile veri tutarlı mı?
- .....


Hata
Örnekleme Hatası
Sistematik Hata – Örneklem Dışı Hata
Veri Toplama, Seçim Hatası
Seçilen Birimin Yerinin Saptanamaması ve Görüşme Yapılamaması
Yanlış Bilgi Verme Hatası
Veri Giriş Hatası

Verilerin düzenlenmesi

Veri Analizi

- Analiz: Bütünü oluşturan bileşenlerin bütün içerisindeki davranışlarını incelemektir.
- Analiz, bir madde içerisindeki bileşiklerin hepsini veya bir kaçının miktarını ve neler olduğunu ortaya koyma. Çözümleme, Tahlil...

- ... verilerin etkili analizine ilişkin disiplinler arası bir çalışma;
- ... Sürekli akan verinin büyük miktarlarından faydalı bilgileri ayıklamak için kullanmak;
- Mevcut veri tabanlarından arzu edilen bilgileri veya ilginç kalıpları çıkarmak;
- Verilerden anlam çıkarma tekniği.
- Analiz, yorumlama ve sunum
- Matematiksel model ve algoritma
- Analiz, yorumlama ve iyileştirme için katılım verisi stratejilerini kullanma

Sinyal Analizi

Sinyallerin sayısal veri haline dönüştürülmesi aşamasında örnekleme,

- Değerlendirme
- Risk analizi
- Sapma analizi
- Regresyon analizi, verileri denklemleştirme
- Korelasyon
- Olasılık
- Kestirim, öngörü
- İstatistiksel analiz
- İşaret işleme

Niçin veri analizi yapılır?

- forecast outcomes: tahmin sonuçları

Verilerin analizinin amacı, kullanışlı ve yararlı bilgiler elde etmektir.
Verilerin nitel veya nicelik bakımından analizinde şunlar yapılabilir:

- veriyi tanımlamak ve özetlemek
- değişkenler arasındaki ilişkileri belirlemek
- değişkenleri karşılaştırmak
- değişkenler arasındaki farkı saptamak
- Çıktılardan tahmin yapabilmek, yorumlamak

Veri Analizi Yetenekleri

- Ön işleme (eleme!)
  - –Ölçekleme ve ortalama
  - - Enterpolasyon ve yok etme
  - –Kırpma ve eşikleme
  - –Veri bölümlerini çıkarma
  - –Düzleştirme ve filtreleme
- Sayısal ve matematiksel işlemleri uygulama (kırın!)
  - –Korelasyon, temel istatistikler ve eğri uydurma
  - –Fourier analizi ve filtreleme
  - –Matris analizi
  - –1-Boyutlu tepe, vadi ve sıfır bulma
  - –Diferansiyel denklem çözücüler

Veri analiz yöntemleri

- **Betimsel (Açıklayıcı) Analiz:** En basit ve herkes tarafından kolaylıkla anlaşılabilir veri analizi türüdür. Analiz için kullanılan verilerden “Yaş aralığı” ve “Nicelik” gibi sonuçların hızlı ve kolay bir şekilde ortaya çıkmasını sağlar.
- **Keşif Analizi:** Analiz sürecinde kullanılan veriler arasındaki doğrudan ya da dolaylı ilişkileri anlamak için keşif analizinden yararlanılır.
- **Çıkarımsal Analiz:** Küçük miktarda veri kullanarak, daha büyük miktardaki gruplar hakkında yorum yapabilmek ya da kararlar alabilmek için çıkarımsal analiz kullanılır.
- **Tahmin Analizi:** Bir grup ya da olaydaki verileri kullanarak başka bir grup ya da olay hakkında yorum yapabilmek için tahmin analizi kullanılır.

Gelişmiş Veri Analiz Yöntemleri

- Eğri Uydurma
- Filtre tasarımı
- İstatistik
- İletişim
- Optimizasyon (Eniyileme)
- Wavelet
- Spline
- Görüntü işleme
- Sembolik matematik
- Kontrol sistemi tasarımı
- Kısmi diferansiyel denklemler
- Nöral ağlar
- Sinyal işleme
- Bulanık mantık

Veri Analiz Teknikleri

- Bilgisayar Destekli Denetim Araç ve Teknikleri
  - Sürekli Denetim ve Gözetim
- Denetim, Kontrol
- Tutarlılık
- Suistimal Araştırma ve Önleme
- Raporlama
- Risk Yönetim Teknikleri
- İlişki Analizi
- Metin (Text) Tarama Teknikleri
- Kimlik Çözümleme Teknikleri

Verileri denklemleştirme, Korelasyon

- Korelasyon, bir değişkenin değeri değişirken diğer bir değişken bununla lineer(doğrusal) lineer olmayan ilişkili olarak değişiyorsa korelasyon vardır.
- Varsayım, kestirim, olasılık
- Sebep sonuç ilişkisi
- Test
- Ret bölgeleri

Teknik Analiz

- Tarihsel fiyat hareketlerini analiz etmek ve mevcut ticaret ortamını göz önüne almak potansiyel fiyat hareketini belirlemeye izin verir.
- Aklıma gelen eski bir ifade var - o tarihin tekrar tekrar etme eğiliminin var olmasıdır.
- Aslında bunların hepsi teknik analiz konusu ile ilgilidir. Fiyat, tepeden tırmanışa (direnç) veya olumsuzluğa (destek) kadar kıramadığı bazı alanları gösterdiğinde, tüccarlar mutlaka not edip ticaret stratejilerini bu bilgilere dayanarak ayarlarlar.
- Teknik analiz çalışmaları yalnızca belirli seviyeleri değil, bazı tarihi fiyat hareket modelleri ile teknik tüccarlar ve analistler, bu kalıpların gelecekte tekrarlanacağına ve fiyatın aynı şekilde davrandığına inanıyor. Yatırımcılar, bu fiyat modellerini tekrarlama konseptine dayanan ticaret fikirlerini yaratırlar.

İstatistiksel Analiz

- Statistical Analysis
  - Sampling Distributions and z-scores
  - Hypothesis Test
  - Estimation

İstatistik

- İstatistik, yeni bilgiler elde etmek için verileri toplama, görüntüleme, analiz etme ve yorumlama sanatıdır.
- İstatistik, gerçekliğin matematiksel olarak ele alınmasıdır.
- Üç tür yalan vardır: yalanlar, lanet olası yalanlar ve istatistikler.
- İstatistik, sayısal verileri etkin bir şekilde kullanma bilimidir.
- Verilerin toplanması, analizi ve yorumlanması dahil bunun tüm yönleriyle ilgilenir.

İstatistik

- **Tanımlayıcı istatistikler**
  - Tablo ve Grafik Gösterimler
  - Karakteristik Ölçüler
  - Temel Bileşen Analizi
- **Tümevarım İstatistikler**
  - Parametre Tahmini (nokta ve aralık tahmini, tahmin ediciler bulma)
  - Hipotez Testi(parametre testi, uygunluk testi, bağımlılık testi)
  - Model Seçimi (bilgi kriterleri, minimum açıklama uzunluğu)

Tümevarımsal İstatistikler

- Tümevarımsal istatistikler (veya tümevarımsal akıl yürütme), daha büyük bir popülasyondan örnek almak ve bu verileri aşağıdakiler için kullanmakla ilgilenen bir istatistik dalıdır:

  - Sonuca varmak,
  - Karar vermek,
  - Tahmin,
  - Gelecekteki davranışı tahmin etmek.

- Örneğin, yeni, toplu olarak pazarlanan bir gıda ürününün, tüketicilerin ilk algılarına göre başarılı olursa ne olacağını bilmek isteyebilirsiniz. Ürünü satın alan ve tatan her tüketiciyle iletişime geçmeniz mümkün olmadığından, 100 tüketici için bir tat testi yapabilir ve tahminlerinizi bu küçük örneğe dayandırabilirsiniz.

Çıkarımsal ve Tümevarımsal İstatistikler

- Tümevarımsal (Inductive) istatistikler ve çıkarımsal (Inferential) istatistikler aslında aynı şeyin iki adıdır. Aslında, birçok yazar bu iki terimi birbirinin yerine kullanır. Örneğin:
- Çıkarımsal istatistikler aynı zamanda tümevarımlı akıl yürütme veya tümevarımlı istatistikler olarak da adlandırılır.
- Tümevarımsal istatistikte olasılık teorisi, veriyi oluşturan süreç hakkında çıkarımlar yapmak için uygulanır.
- Bununla birlikte, iki terim arasında çok ince bir fark var. “Tümevarımsal” adı, belirli bilgilere dayalı genel sonuçlara varma süreci olan tümevarımlı akıl yürütme teriminden gelir. Bu nedenle, Tümevarım istatistik, belirli bilgi parçalarına dayalı genel sonuçlara varmanın mantıksal sürecidir; üretilen verilerin (istatistiklerin) aksine, çıkarımsal istatistiklerin arkasındaki temel süreçtir. Başka bir deyişle, çıkarımsal istatistik dalı (tahmin ve hipotez testini içerir) tümevarımlı akıl yürütmeyi kullanır.

- Başka bir deyişle, "çıkarımsal", küçük veri örneğinizle yapmak isteyebileceğiniz tüm tahmin, riskten korunma ve tahminleri kapsayan geniş bir fırçadır. Öte yandan, "Tümevarım", bu geniş fırçanın daha küçük bir parçasıdır - spesifik verilere dayalı genel sonuçlar çıkarmak için kullanılan kısım.

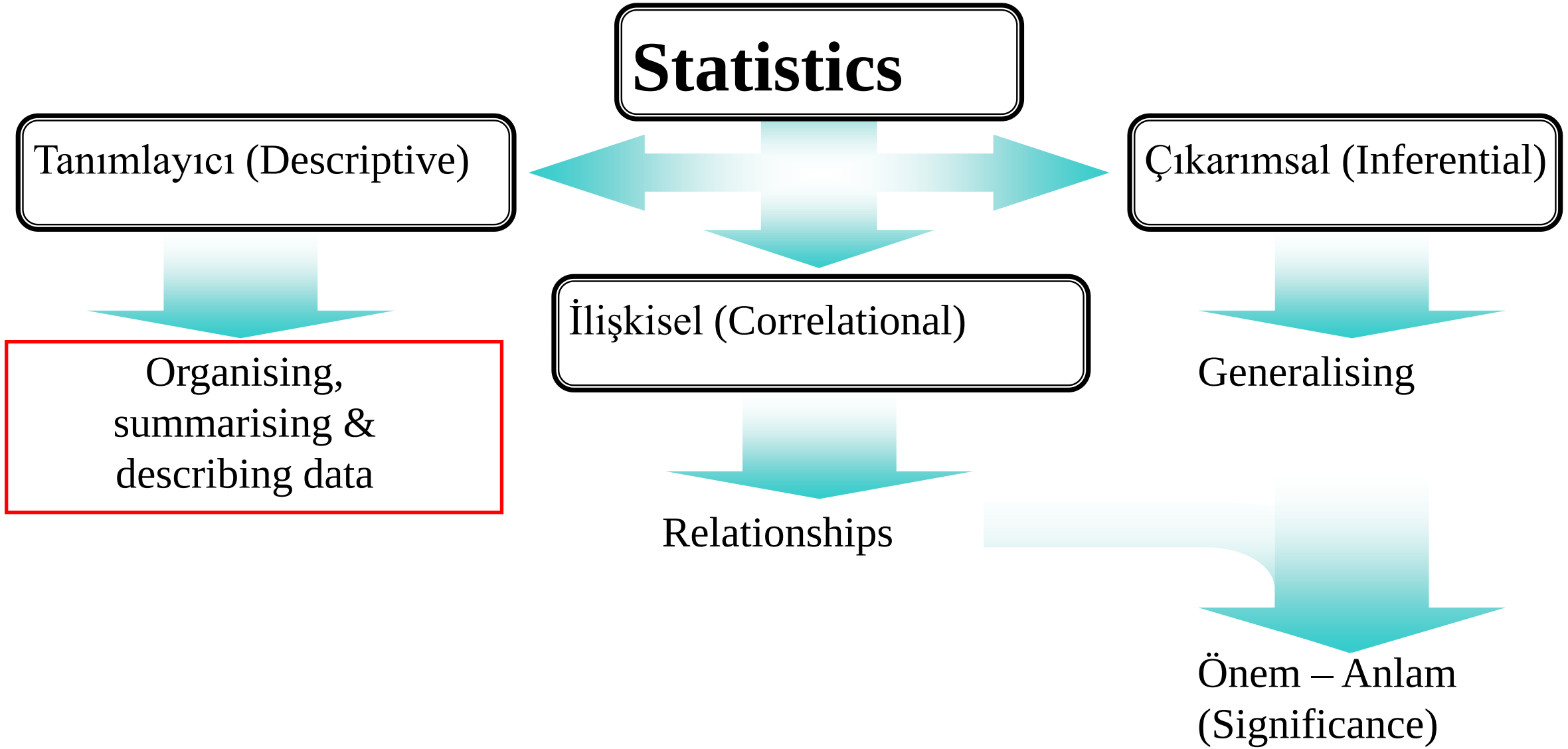
İstatistiksel veri analizi

Tarım, tıp, fizik, biyoloji, kimya vb. Gibi çeşitli alanlardaki araştırmalar, “gözlemlerin” toplanmasını gerektirir. Gözlemler neredeyse her zaman rastgele hatalara tabidir. Bu nedenle, verileri toplamak ve analiz etmek için istatistiksel yöntemler kullanılmalıdır.

İstatistiksel veri analizi kullanarak bir problemi incelemek genellikle dört temel adımı içerir: Sorunu tanımlama. Verilerin toplanması. Verilerin analizi. Sonuç ve Öneriler.

Defining the problem: Sorunla ilgili doğru verileri elde etmek için sorunun tam olarak tanımlanması zorunludur. Sorunun net bir tanımı olmadan veri toplamak son derece zordur.

Deneysel tasarımın üç temel ilkesi şunlardır: Randomizasyon. Çoğaltma. Engelleme.


Statistics
Tanımlayıcı (Descriptive)
İlişkisel (Correlational)
Çıkarımsal (Inferential)
Organising, summarising & describing data
Relationships
Generalising
Önem – Anlam (Significance)

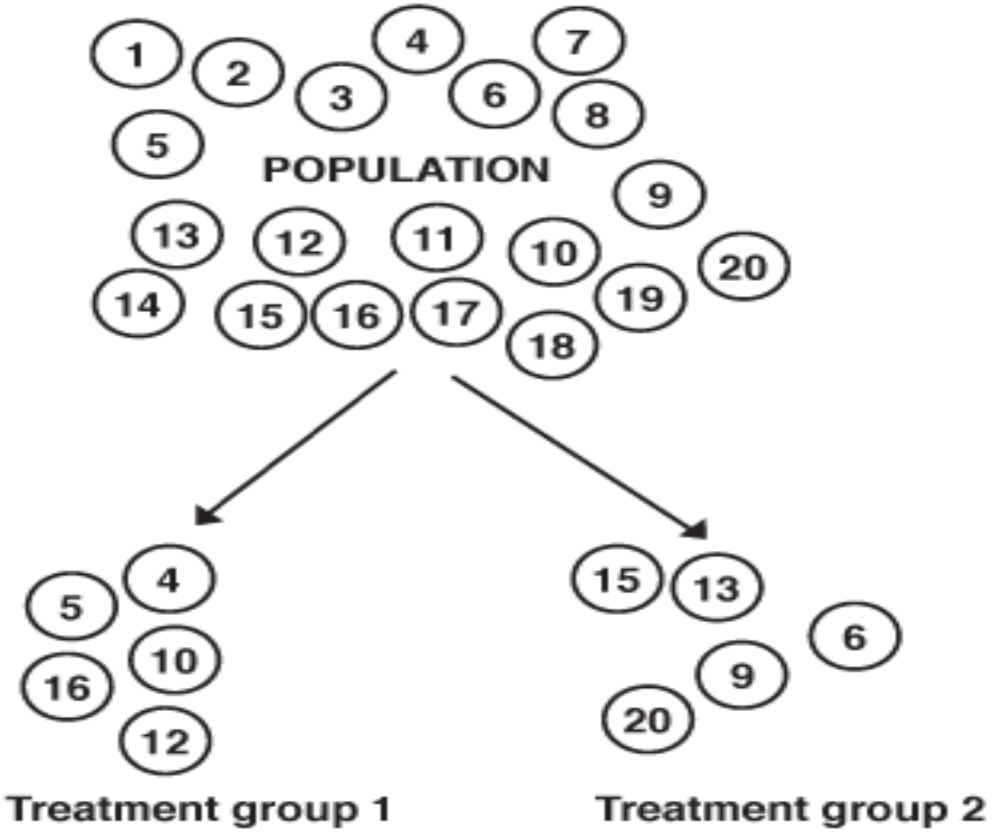
Statistics

Üç alan:

1) Tanımlayıcı –Descriptive: Merkezi eğilim, dağılım ölçüleri
2) İlişkisel – Relational- Correlational: Tek değişkenli, iki değişkenli veya çok değişkenli istatistikler
3) Çıkarımsal – Inferential: Ortalamaların farkı, istatistiksel anlamlılık testleri

Randomization

- Randomizasyon, deneysel tasarımda istatistiksel yöntemlerin kullanımının temelini oluşturan temel taşıdır.
- Randomizasyon ile, hem deneysel materyalin tahsisinin hem de deneyin bireysel çalışmalarının veya denemelerinin gerçekleştirileceği sıranın rastgele belirlendiği kastedilir.



Verileri Tanımlamak ve Sunmak

- Üç önemli kriter—doğruluk, kısalık ve anlaşılabilirlik
- Araştırmacılar verilerini her zaman verileri en doğru şekilde temsil edecek şekilde sunmalıdır.
- Sayısal veriler sayısal olarak sınıflandırılabilir (örnek: yüzdeler, (ortalamalar) veya grafik (grafikler) yöntemi)

Statistical terms

- Popülasyon: Komple bireyler, nesneler veya ölçümler seti
- Örnek: Bir popülasyonun alt kümesi
- Değişken: Farklı değerler alabilen bir özellik
- Veriler: Toplanan sayılar veya ölçümler
- Bir parametre, bir nüfusun bir özelliğidir: Tüm Türklerin ortalama yüksekliği.
- İstatistik, örneğin bir özelliğidir: Bir Türk erkeğinin ortalama boyu.

Başlıca merkez ve değişkenlik ölçüleri

- İstatistik: Sayısallaştırılmış bir örneği karakterize eden ölçümler.
- Parametre: Veri yığınını karakterize eden ölçümler.
- Örnek: Veri yığının sınıflandırılmış alt grubu
- Veri yığını: Belli bir özelliğe ilişkin veri değerleri topluluğu
- Varyans : Ortalamadan ayrılışların kareleri toplamının aritmetik ortalamasıdır.
- Ortalamanın Standart Hatası : Örnek ortalamalarının örnekleme dağılışının standart sapmasına ortalamanın standart hatası denir.
- Mod (Tepe) değeri: Bir dizide en çok tekrarlanan adeti mod (Tepe) değerini verir. Birden fazla sayının en çok tekrarlanma adeti aynı ise bu sayılar alınır.
- Medyan (Ortanca): Dizinin terimleri büyükten küçüğe ya da küçükten büyüğe sıralandığında baştan ve sondan eşit uzaklıktaki sayıya medya (ortanca) denir. Dizinin tam ortasındaki sayı medyandır. Dizinin terim sayısı çift ise ortadaki iki terimin aritmetik ortalaması alınır.
- Açıklık (Aralık)= En büyük değer – En küçük değer
- Üst çevrek açıklığı= Ortanca değerin üst kısmını ifade eder.
- Alt Çeyrek Açıklığı= Ortanca değerin alt kısmını ifade eder.
- Ortanca=(alt çeyrek açıklığı + Üst çeyrek açıklığı)/2

Örnek boyutu seçimi

Neden plan yapmak isteyelim?

1. Örnek boyutları ne kadar büyük olursa, ortamalardaki farklılıkları saptamak veya bulmak o kadar kolay olur.
2. Örneklem boyutu ne kadar büyükse, "maliyet" o kadar yüksek olur ve pratik olarak önemsiz farklılıkların istatistiksel olarak anlamlı bulunma olasılığı o kadar yüksektir.

Veri Analizi

Data Analysis:
Types of Statistics

- Tanımlayıcı istatistikler: Örnek için değişken değerlerin/puanların özetlenmesi ve organizasyonu
- Çıkarımsal istatistikÖrnek İstatistikten Popülasyon Parametresine Yapılan Çıkarımlar. Nedensellik Tahmin Edebilir veya Nedensel Çıkarım YapabilirDeneysel (Bağımsız) Değişkenin Sonuç (Bağımlı) Değişkeni üzerindeki etkisini izole edin

Veri Analizi:
Tanımlayıcı İstatistikler

- Tanımlayıcı İstatistikler, araştırmacıların ilgilenilen değişkenleri tanımlayabilmeleri veya iletebilmeleri için bir örneklemdeki puanları düzenlemek ve özetlemek için kullanılan prosedürlerdir.
- Not: Tanımlayıcı İstatistikler yalnızca örneklem için geçerlidir: verilerin popülasyondaki gerçekliği ne kadar doğru yansıtabileceği hakkında hiçbir şey söylemez.
- Tüm popülasyondaki ilişkiler hakkında bir şeyler “çıkarmak” için Örnek İstatistikleri kullanır: örneğin popülasyonu temsil ettiğini varsayar.
- Tanımlayıcı İstatistikler tek değişkeni özetler.
- Ortalama, Medyan, Mod, Aralık, Frekans Dağılımı, Varyans ve Standart Sapma Tanımlayıcı İstatistiklerdir: Tek Değişkenler

Veri Analizi:
Çıkarımsal İstatistik

- *Çıkarımsal İstatistikler, aynı popülasyondan alınan başka bir örnekle bir örnekten aynı sonuçları bulma olasılığını test etmek için tasarlanmış prosedürlerdir: aslında, popülasyondan olası tüm örnekler test edildiğinde örnek sonuçlarının elde edilip edilmeyeceğini matematiksel olarak test eder.*
- *Sonuçlar için bir açıklama olarak şansı dışlama girişimleri: bu sonuçlar, popülasyonda var olan ve sadece rastgele veya tesadüfen olmayan gerçek ilişkileri yansıtır.*
- *İstatistikleri kullanarak bir ilişkiyi tanımlamadan veya değerlendirmeden önce, araştırma sorunuzun ele alınabilmesi için çalışmanızı tasarlamanız gerekir.*
- *Bu Metodolojidir: teorinin Veri Toplama Yöntemleri ve Veri Analizi ile buluştuğu yer.*

Data Analysis:
Statistics Notation

Harfler Büyük ya da Küçük Olması

- Genel olarak, büyük harfler popülasyon özelliklerine (yani parametrelere) atıfta bulunur; ve küçük harfler örnek özniteliklere (yani istatistiklere) atıfta bulunur.
- Örneğin,P bir yığın oranını ifade eder; ve p, örnek bir orana göre.
- X, bir dizi popülasyon elemanına atıfta bulunur; ve x, bir dizi örnek elemana.
- N, popülasyon büyüklüğünü ifade eder; ve n, örnek boyutu için.

Yunan ve Roma Harfleri

- Büyük harfler gibi, Yunan harfleri de yığın özelliklerini ifade eder.
- Bununla birlikte, örnek karşılıkları genellikle Roma harfleridir.
- Örneğin,μ bir popülasyon ortalamasını ifade eder;ve x, bir örnek ortalamaya.
- σ bir popülasyonun standart sapmasını ifade eder;ve s, bir örneğin standart sapmasına.

Data Analysis:

Statistics Notation

Yığın Parametreleri

Geleneksel olarak, belirli semboller belirli popülasyon parametrelerini temsil eder.

Gösterim:

- μ bir popülasyon ortalamasını ifade eder.
- σ bir popülasyonun standart sapmasını ifade eder.
- σ^2, bir popülasyonun varyansını ifade eder.
- P, belirli bir özniteliğe sahip popülasyon öğelerinin oranını ifade eder.
- Q, belirli bir niteliğe sahip olmayan popülasyon öğelerinin oranını ifade eder, dolayısıyla Q = 1 - P.
- ρ, bir popülasyondaki tüm öğelere dayanan popülasyon korelasyon katsayısıdır.
- N, bir popülasyondaki eleman sayısıdır.

Örnek almalı İstatistik

Geleneksel olarak, belirli semboller belirli örnek istatistikleri temsil eder.

Gösterim:

- x bir örnek ortalamayı ifade eder.
- s, bir örneğin standart sapmasını ifade eder.
- s^2, bir örneğin varyansını ifade eder.
- p, belirli bir özniteliğe sahip örnek öğelerin oranını ifade eder.
- q, belirli bir niteliğe sahip olmayan örnek öğelerin oranını ifade eder, dolayısıyla q = 1 - p.
- r, bir örnekteki tüm öğelere dayanan örnek korelasyon katsayısıdır.
- n, bir örnekteki eleman sayısıdır.

Data Analysis:

Summation/ Sigma Notation

Summation Notation is shorthand that relies on Greek alphabet and mathematical symbols to indicate how to process values: *aka formulae*.

- Σ = summation
- X = Variable

What do each of these mean?

- ΣX
  - Add up the values of X
- $\Sigma X + 2$ versus $\Sigma(X + 2)$
  - Add up the values of X and add 2 to the Sum,
  - Add 2 to each value of X and then Sum the values
- ΣX^2 versus $(\Sigma X)^2$
  - Square each value of X and then Sum
  - Sum the values of X and then Square the Sum
- $\Sigma(X + 2)^2$ versus $\Sigma(X^2 + 2)$
  - Add 2 to each value of X, square the value, then Sum the squared values
  - Square each value of X, add 2 to the value, then Sum the values

Data Analysis:

Summation/ Sigma Notation

Σ : summation
X : Independent Variable, typically
Y: Dependent Variable, typically
N= Size of the Population
n= Size of the Sample
$\leq \geq \neq =$: Equalities or Inequalities
$\pm \times \div + -$: Mathematical Operators
α: alpha, refers to constant/ intercept
μ: mu, sample mean
β: beta coefficient/ standardized
δ: sigma, sample standard deviation
$\delta^{2:}$ sigma squared, sample variance

Data Analysis:

Frekans dağılımları

- Verileri topladıktan sonra, bir araştırmacının ilk görevi, sonuçlara genel bir bakış için verileri düzenlemek, özetlemek, yoğunlaştırmak ve basitleştirmektir.
- Frekans Dağılımları, verileri organize etmek, özetlemek, yoğunlaştırmak ve basitleştirmek için geleneksel yöntemdir.

Data Analysis:

Frekans dağılımları

- Bir Frekans Dağılımı en az iki sütundan oluşur:
  - ölçüm ölçeğinde (X) bir listeleme kategorisi ve
  - sıklık (f) için başka bir kategori.
- X sütununda, değerler en yüksekten en düşüğe listelenir: değerlerin hiçbirini atlamayın.
- Sıklık sütunu, her X değeri için hesaplamaları içerir: veri kümesinde her X değerinin ne sıklıkta gerçekleştiği. Bu çeteleler, her X değeri için frekanslardır.
- Frekansların toplamı N'ye eşit olmalıdır.

Data Analysis:

Frekans dağılımları

- Her kategori için oran (p) için üçüncü bir sütun kullanılabilir: p = f/N.
- p sütununun toplamı 1,00'e eşit olmalıdır.
- Her X değerine karşılık gelen dağılımın yüzdesini görüntülemek için genellikle dördüncü bir sütun eklenir.
- Yüzde, p'nin 100 ile çarpılmasıyla bulunur.
- Yüzde sütununun toplamı %100'dür.

Data Analysis:
Frekans dağılımları

- Düzenli veya Normal Frekans Dağılımı
  - Tüm bireysel kategoriler (X değerleri) listelenir
- Bir puan kümesi geniş bir değer aralığını kapsadığında, tüm X değerlerinin bir listesi oldukça uzun olacaktır: verilerin "basit" bir sunumu olamayacak kadar uzun.
- Çok sayıda ve çeşitli X değerlerinin olduğu bir durumda, Gruplandırılmış Frekans Dağılımı kullanılır.

Data Analysis:

Frekans dağılımları

- Gruplandırılmış Frekans Dağılımı: X sütunu, bireysel değerler yerine Sınıf Aralıkları adı verilen puan gruplarını listeler.
- Sınıf Aralıklarının tümü aynı genişliğe sahiptir: tipik olarak 2, 5, 10 vb. gibi basit bir sayı.
- Her Sınıf Aralığı, Aralık Genişliğinin katı olan bir değerle başlar.
- Aralık Genişliği, dağılımın yaklaşık 10 aralığı olacak şekilde seçilir.

Data Analysis: *Grouped Frequency Distribution*

- Choosing a width of 15 **Class Intervals** produces the following **Frequency Distribution**.

- **Age** is typically displayed as **Grouped Frequency Distribution**:
  - For Example:
    - 45 to 54 Years
    - 55 to 64 Years

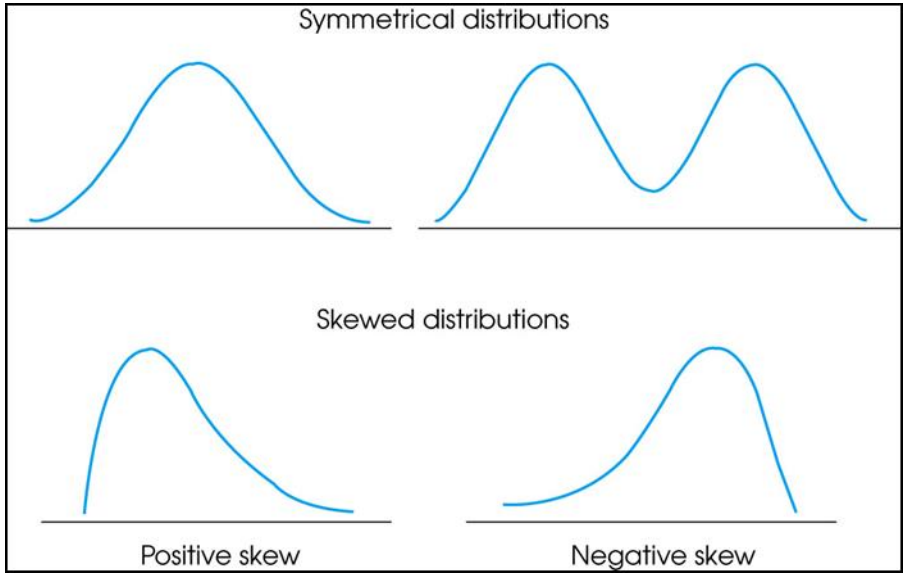
Class Interval	Frequency	Relative Frequency
100 to <115	2	0.025
115 to <130	10	0.127
130 to <145	21	0.266
145 to <160	15	0.190
160 to <175	15	0.190
175 to <190	8	0.101
190 to <205	3	0.038
205 to <220	1	0.013
220 to <235	2	0.025
235 to <250	2	0.025
	79	1.000

- Günümüzün Bilgisayar Teknolojisi, verilerin otomatik açıklayıcı raporlamasına sahiptir.Veri Ambarı'nın ortaya çıkışı, ulusal anketlerden veya gözetim sistemlerinden alınan verileri otomatik işleme, rutin raporlama işlevi ve görsel grafik çıktıları olan ürünlere dönüştürmüştür.

Data Analysis:

Pozitif ve Negatif Çarpık Dağılımlar

- Pozitif Çarpık: Puanlar, kuyruk sağa doğru sivrilirken dağılımın sol tarafında birikme eğilimindedir.
- Negatif Çarpık: Skorlar sağ tarafta birikme eğilimindedir ve kuyruklar sola doğru yığılır.



Data Analysis:
Yüzdelikler, Yüzdelik Sıraları ve İnterpolasyon
Percentiles, Percentile Ranks, & Interpolation

- Yüzdelikler ve Yüzdelik dereceler şunları açıklar: bireysel puanların bir dağılım içindeki göreli konumu: örneğin, bebek ağırlığının 90. yüzdelik dilimi
- Belirli bir X değeri için Yüzdelik Sıralaması, bu X değerine eşit veya daha düşük puanlara sahip bireylerin yüzdesidir.
- Sıralaması ile tanımlanan bir X değeri, Yüzdelik birimidir.

Data Analysis:
X to z and z to X

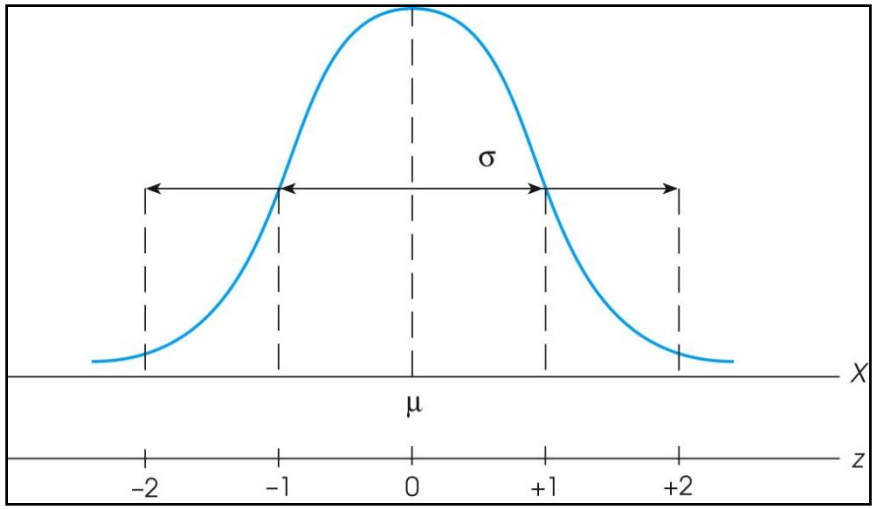
- Temel z puanı tanımı, çoğu z puanı dönüşümünü tamamlamak için genellikle yeterlidir. Ancak tanım, herhangi bir X değeri için z puanını hesaplamak için bir formül oluşturmak üzere matematiksel gösterimde yazılabilir.

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

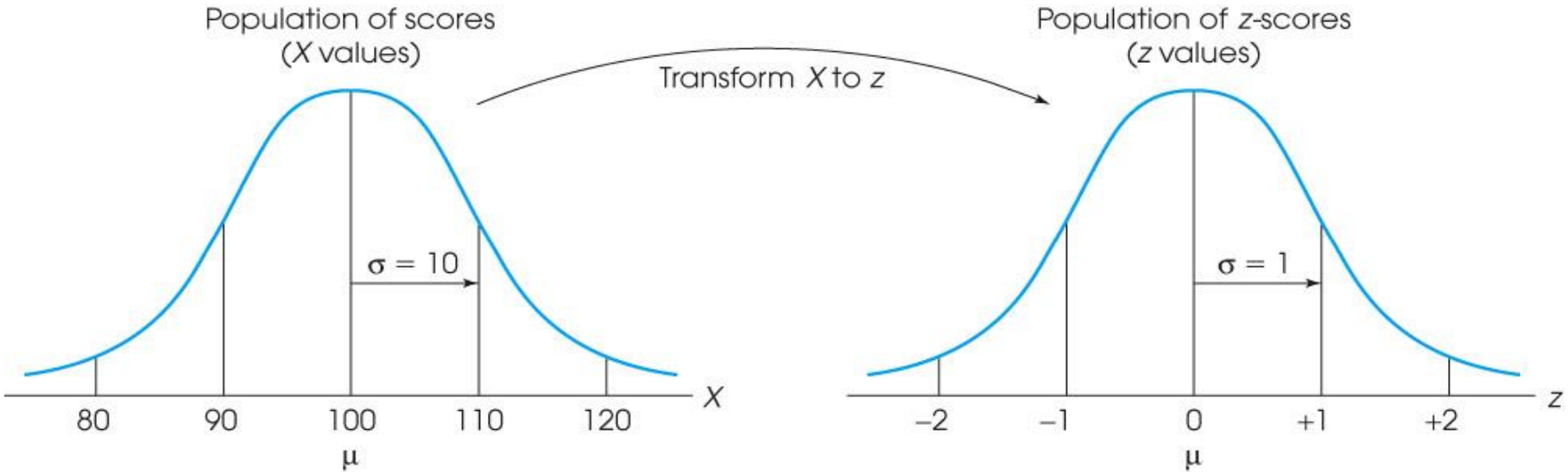
- Ek olarak, formüldeki terimler, herhangi bir özel z puanına karşılık gelen X'in değerini hesaplamak için bir denklem oluşturmak üzere yeniden gruplandırılabilir.

$$X = \mu + z\sigma$$

Bir popülasyon dağılımındaki z-skor değerleri ile konumlar arasındaki ilişki.

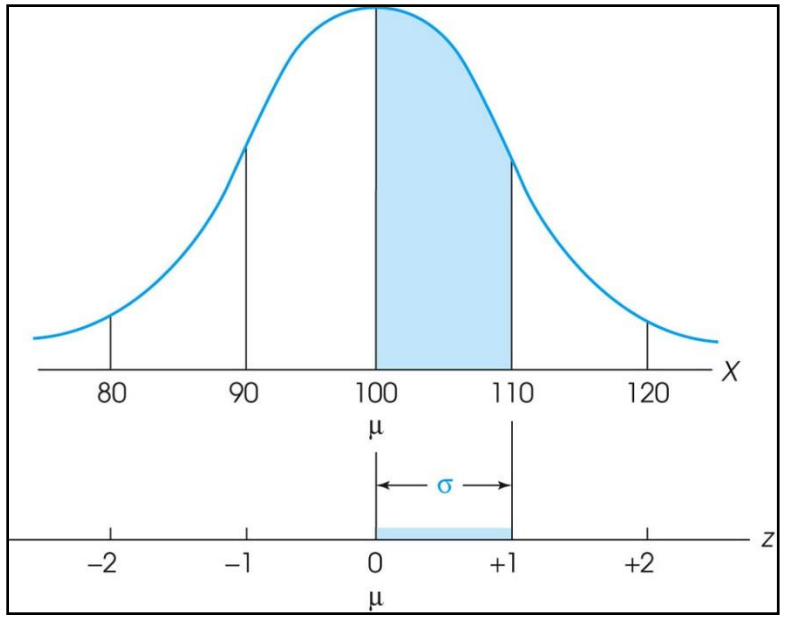


Tüm bir puan popülasyonu z puanlarına dönüştürülür. Dönüşüm popülasyonun şeklini değiştirmez, ancak ortalama 0 değerine dönüştürülür ve standart sapma 1 değerine dönüştürülür.

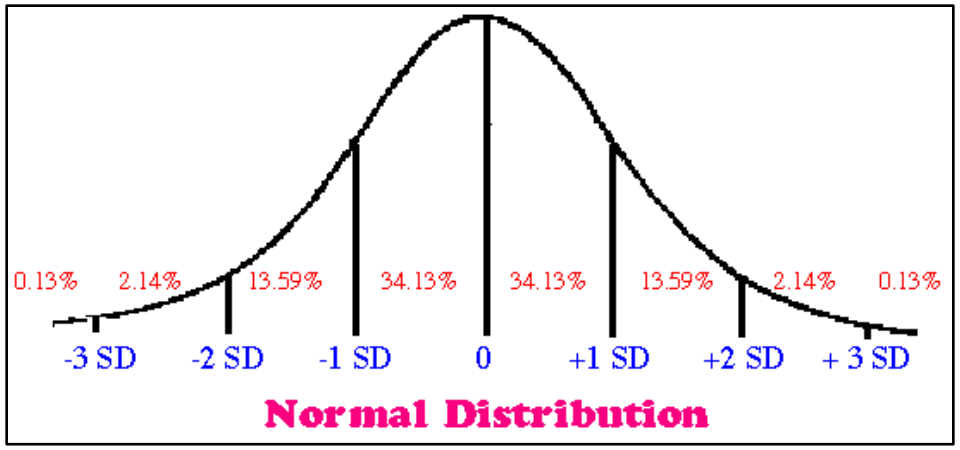


Following a *z*-score transformation, the *X*-axis is relabeled in *z*-score units.

The distance that is equivalent to 1 standard deviation on the *X*-axis (σ = 10 points in this example) corresponds to 1 point on the *z*-score scale



Why are z-scores important? Because if you know the distribution of your scores, you can test hypothesis, and make predictions.



Data Analysis: *Characteristics of z Scores*

- Z scores tell you the number of standard deviation units a score is above or below the mean
- The mean of the z score distribution = 0
- The SD of the z score distribution = 1
- The shape of the z score distribution will be exactly the same as the shape of the original distribution
- $\Sigma z = 0$
- $\Sigma z^2 = SS = N$
- $\sigma^2 = 1 = (\sigma z^2/N)$

Data Analysis:

Sources of Error in Probabilistic Reasoning

- The Power of the Particular
- Inability to Combine Probabilities
- Inverting Conditional Probabilities
- Failure to Utilize sample Size information
- The Gambler's Fallacy
- Illusory Correlations & Confirmation Bias
- A Tendency to Try to Explain Random Events
- Misunderstanding Statistical Regression
- The Conjunction Fallacy

Data Analysis:
Characteristics of the Normal Distribution

- It is **ALWAYS** unimodal & symmetric
- The height of the curve is maximum at μ
- For every point on one side of mean, there is an exactly corresponding point on the other side
- The curve drops as you move away from the mean
- Tails are asymptotic to zero
- The points of inflection always occur at one SD above and below the mean.

Data Analysis:
The Distribution of Sample Means

- A distribution of the means from all possible samples of size n
- The larger the n, the less variability there will be
- The sample means will cluster around the population mean
- The distribution will be normal if the distribution of the population is normal
- Even if the population is not normally distributed, the distribution of sample means will be normal when n > 30

Data Analysis:
Properties of the Distribution of Sample Means

- The mean of the distribution = μ
- The standard deviation of the distribution = $\sigma/\sqrt{n}$
- The mean of the distribution of sample means is called the *Expected Value of the Mean*
- The standard deviation of the distribution of sample means is called the *Standard Error of the Mean* (σ_M)
- Z scores for sample means can be calculated just as we did for individual scores. $Z = M-\mu/\sigma_M$

Data Analysis:
What is a Sampling Distribution?

- It is the distribution of a statistic from all possible samples of size n
- If a statistic is unbiased, the mean of the sampling distribution for that statistic will be equal to the population value for that statistic.

Data Analysis:

Re-Introduction to Hypothesis Testing

- We use a sample to estimate the likelihood that our hunch about a population is correct.
- In an experiment, we see if the difference between the means of our groups is so great that they would be unlikely to have been drawn from the same population by chance.

Methodology:
Formulating Hypotheses

- The Null Hypothesis (H_0)
  - Differences between means are due only to chance fluctuation
- Alternative Hypotheses (H_a)
- Criteria for rejecting a null hypothesis
  - Level of Significance (Alpha Level)
    - Traditional levels are .05 or .01
  - Region of distribution of sample means defined by alpha level is known as the "critical region"
  - No hypothesis is ever "proven"; we just fail to reject null
  - When the null is retained, alternatives are also retained.

z ratio= Obtained Difference Between Means / Difference due to chance/error:
the basis for most of the hypothesis tests

Data Analysis: *Errors in Hypothesis Testing*

- **Type I Errors**
  - You reject a null hypothesis when you shouldn't
  - You conclude that you have an effect when you really do not
  - The alpha level determines the probability of a Type I Error (hence, called an "alpha error")

- **Type II Errors**
  - Failure to reject a false null hypothesis
  - Sometimes called a "Beta" Error.

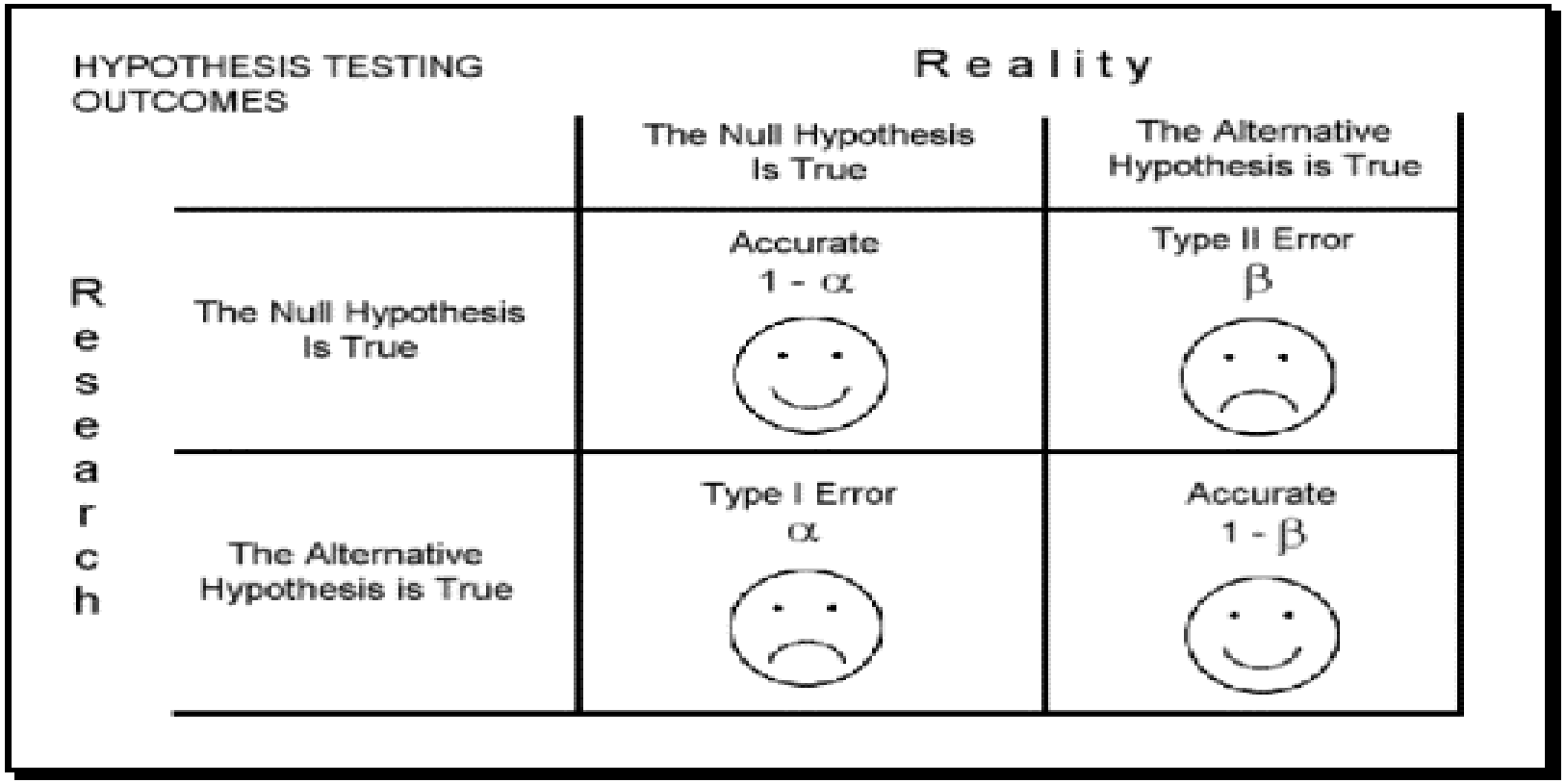
Definition of Type I and Type II errors

Bazen kararlarımız doğru olur bazen de yanlış olur. Sırasıyla Tip I ve Tip II hataları olarak adlandıracağımız iki olası hata vardır.

Tip I hatası, doğru olduğunda sıfır hipotezini reddetme hatasıdır. Tip I hata yapma olasılığı genellikle α ile gösterilir.

Tip II hatası, yanlış olduğunda boş hipotezin kabul edilmesidir. Tip II hata yapma olasılığı genellikle β ile gösterilir.

Type I and Type II errors



Temel İstatistik Terimleri
Hipotez Testi

İhtiyaç duyulan örnek büyüklüğünü belirlemek için önce bazı temel istatistik terimlerini tanımlamamız gerekir.

Boş hipotez Ho, hipotez testi şeklindeki istatistiksel kanıt aksini gösterene kadar doğru olduğu varsayılan bir hipotezdir.

Belirli bir sıfır hipotezini formüle ederken, her zaman alternatif bir hipotez Ha formüle ediyoruz, bu hipotez, gözlenen veri değerlerinin sıfır hipotezi altında yeterince olası olmaması durumunda kabul edilecektir.

Data Analysis: *Statistical Power*

How sensitive is a test to detecting real effects?

- A powerful test decreases the chances of making a Type II Error
- Ways of Increasing Power:
  - Increase sample size
  - Make alpha level less conservative
  - Use one-tailed versus a two-tailed test

Data Analysis:
Assumptions of Parametric Hypothesis Tests
(z, t, ANOVA)

- *Random Sampling* or *Random Assignment* was used
- Independent Observations
- Variability is not changed by experimental treatment: *homogeneity of variance*
- Distribution of Sample Means is normal

Data Analysis: *Measuring Effect Size*

- Statistical significance alone does not imply a substantial effect; just one larger than chance
- Cohen's *d* is the most common technique for assessing effect size
- Cohen's *d* = Difference between the means divided by the population standard deviation: *d > .8 means a large effect!*

Data Analysis: t Statistic

- Since we *usually do not know* the population variance, we must use the sample variance to estimate the standard error
  - Do you Remember? $S^2 = SS/n-1 = SS/df$

- Estimated Standard Error = $S_M = \sqrt{S^2/n}$
  - $t = M - \mu_0/S_M$

Data Analysis:
The Distribution of the t Statistic vs. the Normal Curve

- t is only normally distributed when n is very large. **Why?**
  - The more statistics you have in a formula, the more sources of sampling fluctuation you will have.
  - M is the only statistic in the z formula, so z will be normal whenever the distribution of sample means is normal
  - In "t" you have things fluctuating in both the numerator and the denominator
  - Thus, there are as many different t distributions as there are possible sample sizes.
    - You have to know the degrees of freedom (df) to know which distribution of t to use in a problem.
  - All t distributions are unimodal and symmetrical around zero.

Data Analysis:
Comparing Differences btw Means with t Tests

- There are two kinds of t tests:
  1. t Tests for Independent Samples
     - Also known as a "Between-Subjects" Design
     - Two totally different groups of subjects are compared; randomly assigned if an experiment
  2. t Tests for related Samples
     - Also known as a "Repeated Measures" or "Within-Subjects" or "Paired Samples" or "Matched Groups" Design
     - A group of subjects is compared to themselves in a different condition
     - Each individual in one sample is matched to a specific individual in the other sample

Data Analysis:
PROS & CONS of Independent Sample Designs

PROS

- Independent Samples have *no carryover effects*
- Independent Samples do not suffer from *fatigue or practice effects*
- You do not have to worry about attrition or retaining of subjects: people don not need to show up more than once
- Demand characteristics may be stronger in repeated measure studies than in independent designs
- Since more individuals participate in studies with Independent Samples, the results may be more *generalizeable*

CONS

- Usually requires more subjects (larger n)
- The effect of a variable cannot be assessed for each individual, but only for groups as a whole
- There will be more individual differences between groups, resulting in more variability

Data Collection Methods:

PROS & CONS of Paired-Sample Designs

PROS

- Requires fewer subjects
- Reduces variability/more statistically efficient
- Good for measuring changes over time
- Eliminates problems caused by individual differences
- Effects of variables can be assessed for each individual

CONS

- Carryover Effects
  - 2nd measure influenced by 1st measure
- Progressive Error: Fatigue, Practice Effects
  - Counterbalancing is a way of controlling carryover and practice effects
- Getting people to show up more than once
- Demand characteristics may be stronger

Data Analysis:

What is really going on with t Tests?

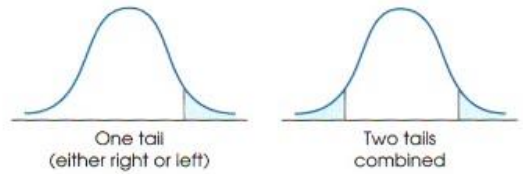
- Essentially the difference between the means of the two groups is being compared to the estimated standard error.
- t = difference between group means/estimated standard error
- t = variability due to chance + independent variable/variability due to chance alone
- The t distribution is the sampling distribution of differences between sample means.
  - comparing obtained difference to standard error of differences

Underlying Assumptions of t Tests

- Observations are independent of each other: except btw paired scores in paired designs
- Homogeneity of Variance
- Samples drawn from a normally distributed population
- At least interval level numerical data

TABLE B.2 THE *t* DISTRIBUTION

Table entries are values of t corresponding to proportions in one tail or in two tails combined.



	PROPORTION IN ONE TAIL					
	0.25	0.10	0.05	0.025	0.01	0.005
	PROPORTION IN TWO TAILS COMBINED					
df	0.50	0.20	0.10	0.05	0.02	0.01
1	1.000	3.078	6.314	12.706	31.821	63.657
2	0.816	1.886	2.920	4.303	6.965	9.925
3	0.765	1.638	2.353	3.182	4.541	5.841
4	0.741	1.533	2.132	2.776	3.747	4.604
5	0.727	1.476	2.015	2.571	3.365	4.032
6	0.718	1.440	1.943	2.447	3.143	3.707
7	0.711	1.415	1.895	2.365	2.998	3.499
8	0.706	1.397	1.860	2.306	2.896	3.355
9	0.703	1.383	1.833	2.262	2.821	3.250
10	0.700	1.372	1.812	2.228	2.764	3.169
11	0.697	1.363	1.796	2.201	2.718	3.106
12	0.695	1.356	1.782	2.179	2.681	3.055
13	0.694	1.350	1.771	2.160	2.650	3.012
14	0.692	1.345	1.761	2.145	2.624	2.977
15	0.691	1.341	1.753	2.131	2.602	2.947
16	0.690	1.337	1.746	2.120	2.583	2.921
17	0.689	1.333	1.740	2.110	2.567	2.898
18	0.688	1.330	1.734	2.101	2.552	2.878
19	0.688	1.328	1.729	2.093	2.539	2.861
20	0.687	1.325	1.725	2.086	2.528	2.845
21	0.686	1.323	1.721	2.080	2.518	2.831
22	0.686	1.321	1.717	2.074	2.508	2.819
23	0.685	1.319	1.714	2.069	2.500	2.807
24	0.685	1.318	1.711	2.064	2.492	2.797
25	0.684	1.316	1.708	2.060	2.485	2.787
26	0.684	1.315	1.706	2.056	2.479	2.779
27	0.684	1.314	1.703	2.052	2.473	2.771
28	0.683	1.313	1.701	2.048	2.467	2.763
29	0.683	1.311	1.699	2.045	2.462	2.756
30	0.683	1.310	1.697	2.042	2.457	2.750
40	0.681	1.303	1.684	2.021	2.423	2.704
60	0.679	1.296	1.671	2.000	2.390	2.660
120	0.677	1.289	1.658	1.980	2.358	2.617
∞	0.674	1.282	1.645	1.960	2.326	2.576

ble III of R. A. Fisher and F. Yates, *Statistical Tables for Biological, Agricultural and Medical Research*, 6th ed. London: Longman Group Ltd., 74 (previously published by Oliver and Boyd Ltd., Edinburgh). Adapted and reprinted with permission of the Addison Wesley Longman blishing Co.

Data Analysis:
Analysis of Variance (ANOVA)

- Use when comparing the differences between means from more than two groups
- The independent variable is known as a “Factor”
- The different conditions of this variable are known as “levels”
- Can be used with independent groups
  - Completely randomized single factor ANOVA
- Can be used with paired groups
  - Repeated measures ANOVA

Data Analysis: *F Ratio & ANOVA*

F Ratio & ANOVA

- F = variance between groups/variance within groups
- F = Treatment Effect + Differences due to chance/Differences due to chance
- F = Variance among sample means/variance due to chance or error
- The denominator of the F Ratio is known as the "error term"

Evaluation of F Ratio

- Obtained F is compared with a critical value
- If you get a significant F, all it tells you is that at least one of the means is different from one of the others
- To figure out exactly where the differences are, you must use Multiple Comparison Tests

Data Analysis: *Multiple Comparison Tests*

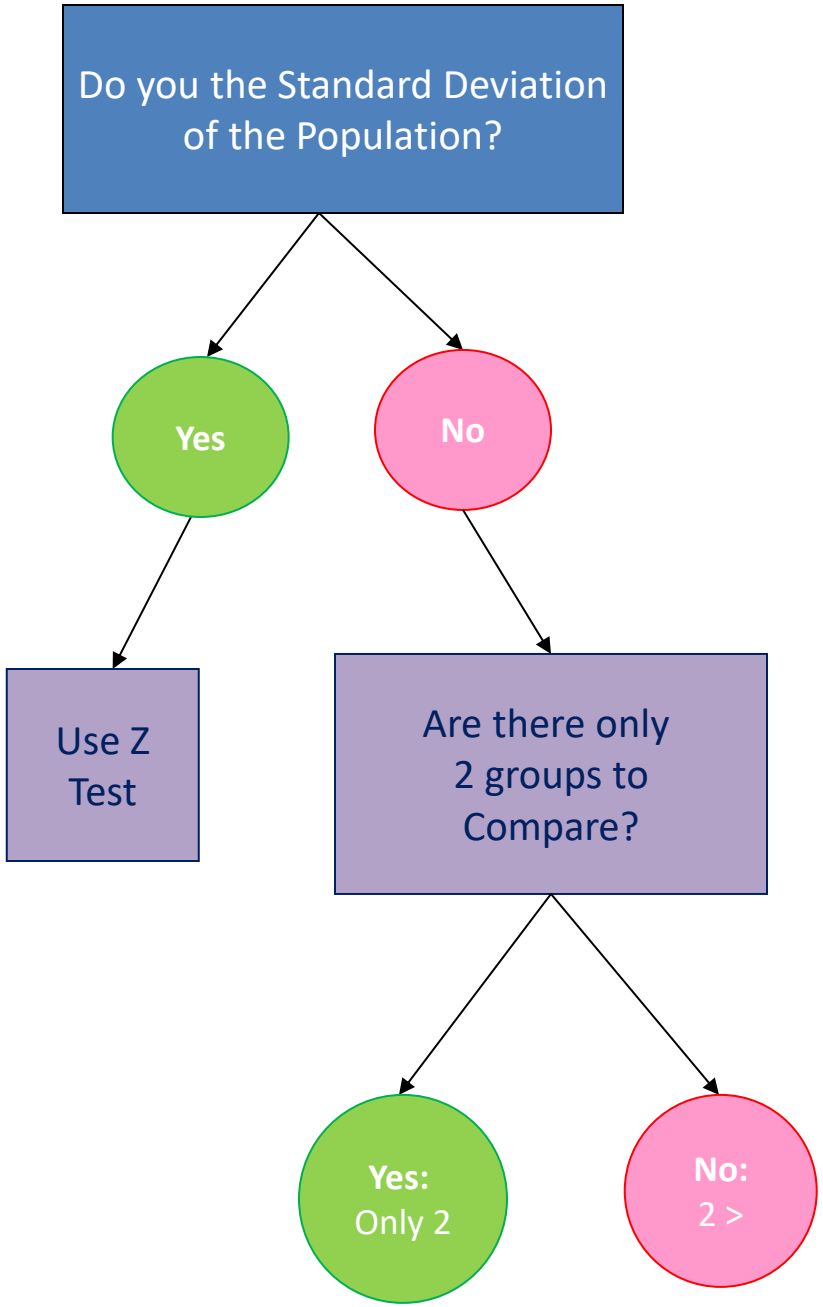
- The issue of "Experiment-wise Error"
  - Results from an accumulation of "per comparison errors"
- Planned Comparisons
  - Can be done with t tests (must be few in number)
- Unplanned Comparisons (Post Hoc tests)
  - Protect against experiment-wise error
  - Examples:
    - Tukey's HSD Test
    - The Scheffe Test
    - Fisher's LSD Test
    - Newman-Keuls Test

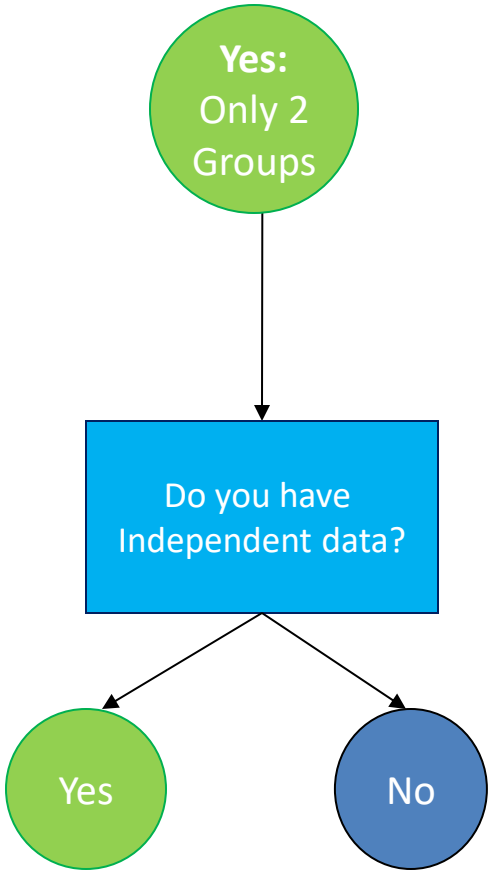
Data Analysis:
Measuring Effect Size in ANOVA

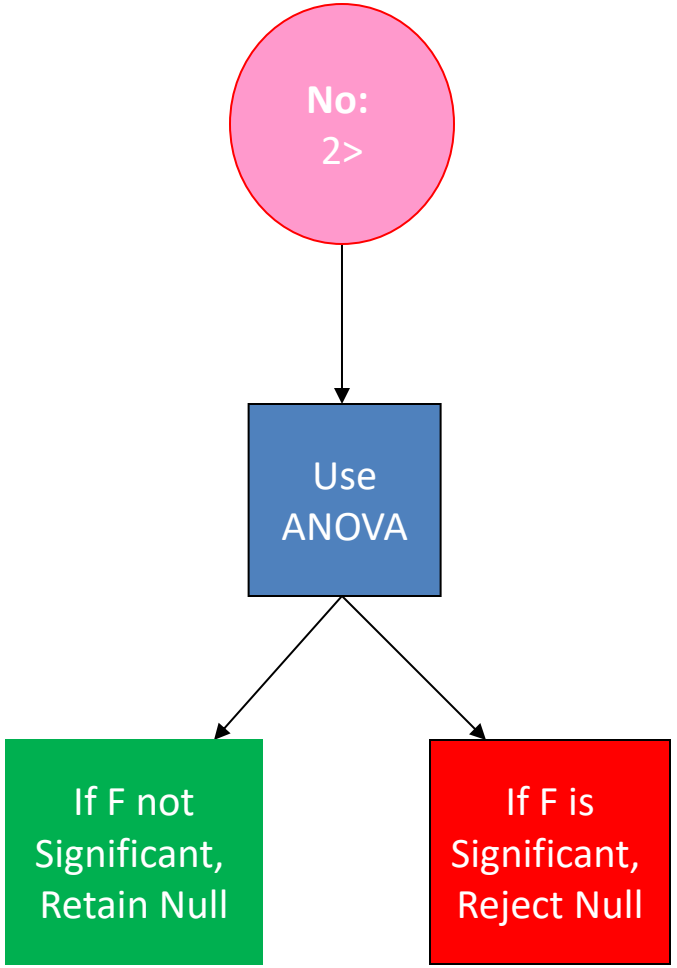
- Most common technique is "r^2"
  - Tells you what percent of the variance is due to the treatment
  - r^2 = SS between groups/SS total

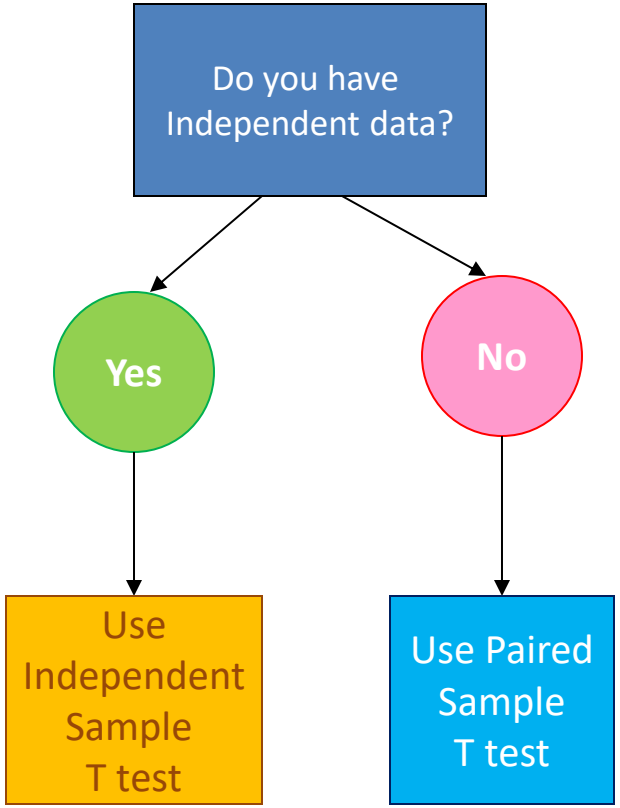
Single Factor ANOVA: *One-Way ANOVA*

- Can be Independent Measures
- Can be Repeated Measures


Do you the Standard Deviation of the Population?
Yes
No
Use Z Test
Are there only 2 groups to Compare?
Yes: Only 2
No: 2 >


Yes:
Only 2
Groups
Do you have
Independent data?
Yes
No

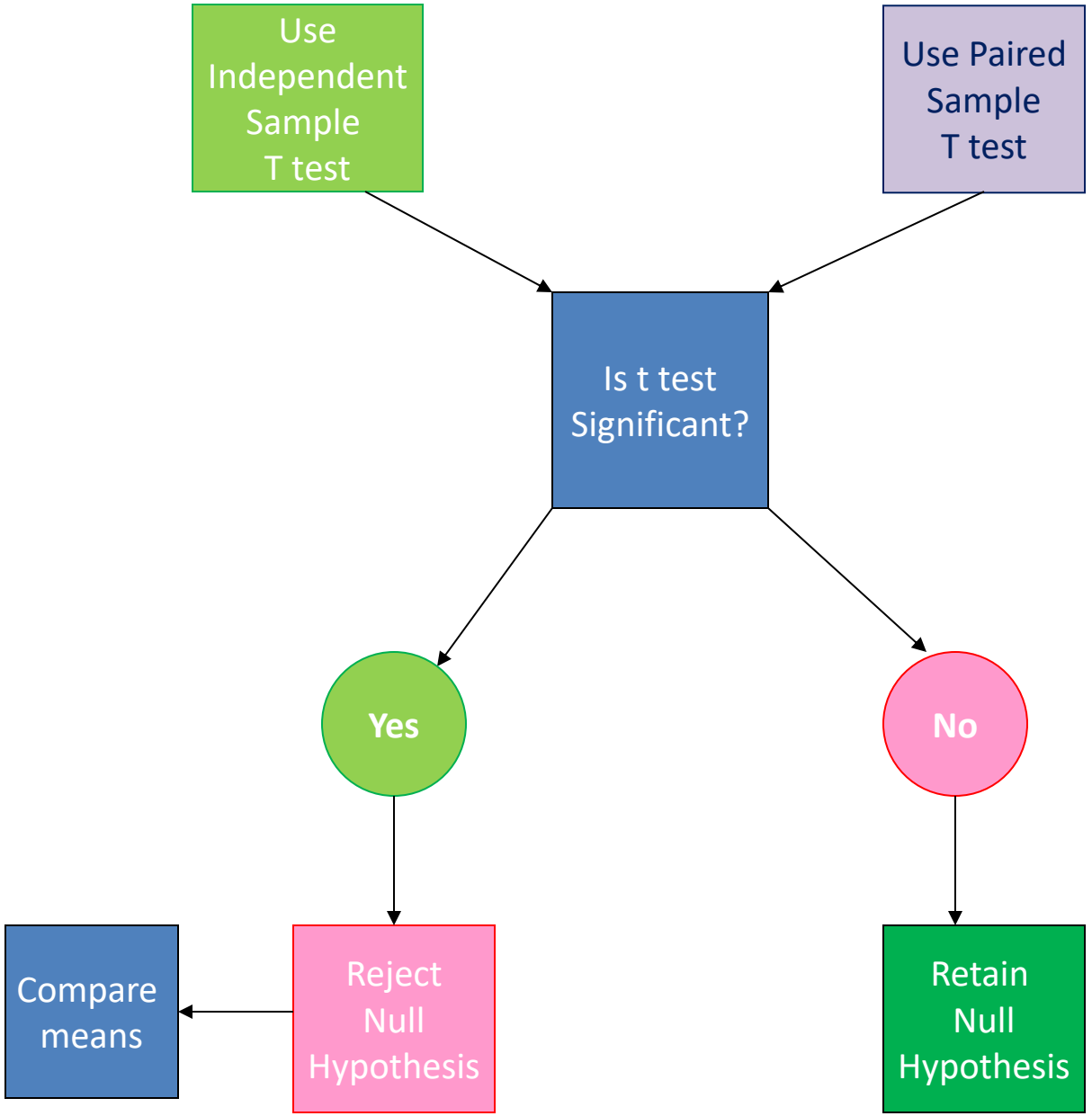

No:
2>
Use
ANOVA
If F not
Significant,
Retain Null
If F is
Significant,
Reject Null



If F not Significant, Retain Null

If F is Significant, Reject Null

Compare Means With Multiple Comparison Tests


Use Independent Sample T test
Use Paired Sample T test
Is t test Significant?
Yes
No
Compare means
Reject Null Hypothesis
Retain Null Hypothesis

Measures of Variability

- Measures of Variability—descriptive statistics that convey information about the spread or variability of a set of data
- Variance—a numerical index of the variability in a set of data
- Standard deviation—a measure of variability that is equal to the square root of the variance
- Range—the difference between the highest and lowest scores in a distribution

Relational Statistics

- 1) univariate – study of one variable for a subpopulation (ex: age of murderers)
- 2) bivariate – study of relationship between two variables (ex: correlation)
- 3) multivariate – study of relationship between three or more variables (ex: multiple correlation)

The Uses of Correlation

- Predicting one variable from another
- Validation of Tests
  - Are test scores correlated with what they say they measure?
- Assessing Reliability
  - Consistency over time, across raters, etc
- Hypothesis Testing

Correlation Coefficients

- Can range from -1.0 to +1.0
- The DIRECTION of a relationship is indicated by the sign of the coefficient (i.e., positive vs. negative)
- The STRENGTH of the relationship is indicated by how closely the number approaches -1.0 or +1.0
- The size of the correlation coefficient indicates the degree to which the points on a scatterplot approximate a straight line
  - As correlations increase, standard error of estimate gets smaller & prediction becomes more accurate
- The closer the correlation coefficient is to zero, the weaker the relationship between the variables.

Types of Correlation Coefficients

- The Pearson r
  - Most common correlation
  - Use with scale data (interval & ratio)
  - Only detects linear relationships
  - The coefficient of determination (r^2) measures proportion of variability in one variable accounted for by the other variable.
    - Used to measure "effect size" in ANOVA
- The Spearman Correlation
  - Use with ordinal level data
  - Can assess correlations that are not linear
- The Point-Biserial Correlation
  - Use when one variable is scale data but other variable is nominal/categorical

Correlation

- Measure of the strength of some relationship between two variables, but not causality.
- Correlations can be positive, negative, or zero.
- Strength of relationship depends on coefficient.

Correlations

Correlation	Strength
0.8 to 1.0	Very strong
0.6 to 0.8	Strong
0.4 to 0.6	Moderate
0.2 to 0.4	Weak
0.0 to 0.2	Very weak

Inferential Tests

- Refer to a variety of tests for inferential purposes.
  - 1. difference of means – to test hypotheses, most common is Z-test.
  - 2) statistical significance – most common are t-test and chi-square (used for less than interval data)

Problems with Interpreting Pearson's r

- Cannot draw cause-effect conclusions
- Restriction of range
  - Correlations can be misleading if you do not have the full range of scores
- The problem of outliers
  - Extreme outliers can disrupt correlations, especially with a small n.

Introduction to Regression

- In any scatterplot, there is a line that provides the “best fit” for the data
  - This line identifies the “central tendency” of the data and it can be used to make predictions in the following form:
    - $Y = bx + a$
    - “b” is the slope of the line, and a is the Y intercept (the value of Y when X = 0)
- The statistical technique for finding the best fitting line is called “linear regression,” or “regression”
- What defines whether a line is the best fit or not?
  - The “least squares solution” (finding the line with the smallest summed squared deviations between the line and data points)
- The Standard Error of Estimate
  - Measure of “average error;” tells you the precision of your predictions
  - As correlations increase, standard error of estimate gets smaller

Simple Regression

- Discovers the regression line that provides the best possible prediction (line of best fit)
- Tells you if the predictor variable is a significant predictor
- Tells you exactly how much of the variance the predictor variable accounts for

Multiple Regression

- Gives you an equation that tells you how well multiple variables predict a target variable in combination with each other.

Nonparametric Statistics

- Used when the assumptions for a parametric test have not been met:
  - Data not on an interval or ratio scale
  - Observations not drawn from a normally distributed population
  - Variance in groups being compared is not homogeneous
  - Chi-Square test is the most commonly used when nominal level data is collected

Tanımlayıcı İstatistik

Tanımlayıcı İstatistikler: Karakteristik Ölçüler

Fikir: Belirli bir örneği birkaç karakteristik ölçü ile açıklayın ve böylece verileri özetleyin.

- **Yerelleştirme ölçüleri,** bir örneğin veri noktalarının bir özniteliğin etki alanında bulunduğu, genellikle tek bir sayı ile açıklanır.
- **Dağılım ölçümleri**, veri noktalarının bir yerelleştirme parametresi etrafında ne kadar değiştiğini açıklar ve böylece bu parametrenin verilerin yerelleştirmesini ne kadar iyi yakaladığını gösterir.
- **Şekil ölçüleri,** veri noktalarının bir referans dağılımına göre dağılımının şeklini tanımlar. En yaygın referans dağılımı normal dağılımdır (Gauss).

Measures of Central Tendency

- The mean—average; most common and useful measure of central tendency; impacted by extreme scores
- The median—middle score of a distribution; less affected by extreme scores (“outliers”)
- The mode—most frequent score

Central limit theorem

- A quantity produced by the cumulative effect of many independent variables will be approximately Gaussian.
  - **human heights - combined effects of many environmental and genetic factors**
  - **weight is non-Gaussian as single factor of how much we eat dominates all others**

- The Gaussian distribution has some important properties which we will consider in a later lecture.

- The central limit theorem can be proved mathematically and empirically.

Central value

- Give information concerning the average or typical score of a number of scores
  - mean
  - median
  - mode

Central value: The Mean

- The Mean is a measure of *central value*
  - What most people mean by “average”
  - Sum of a set of numbers divided by the number of numbers in the set

$$\frac{1+2+3+4+5+6+7+8+9+10}{10} = \frac{55}{10} = 5.5$$

Central value: The Mean

Arithmetic average:

Sample

$$\overline{X} = \frac{\Sigma x}{n}$$

Population

$$\mu = \frac{\Sigma x}{N}$$

$$X = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]$$

$$\sum X \,/\, n = 5.5$$

Central value: The Median

- Middlemost or most central item in the set of ordered numbers; it separates the distribution into two equal halves
- If *odd n*, middle value of sequence
  - if $X = [1,2,4,6,9,10,12,14,17]$
  - then 9 is the median
- If *even n*, average of 2 middle values
  - if $X = [1,2,4,6,9,10,11,12,14,17]$
  - then 9.5 is the median; i.e., (9+10)/2

- Median is not affected by extreme values

Yerelleştirme Ölçüleri: Medyan

Medyan – Ortanca:
Sınıflandırılmamış serilerde ortanca hesaplanırken, veriler öncelikle küçükten büyüğe doğru sıralanırlar. n çift ise n/2 değeri ile n/2+1 değerlerin ortalaması, n tek ise (n+1)/2 değeri ortancadır.
Alt çeyrek, ortalamanın altındaki değerlerdir. Üst çeyrek ortalamanın üstündeki değerlerdir. Alt çeyrek ve üst çeyrek değerlerinden medyan bulunabilir:
Alt çeyrek değerlerinden maksimum olanı < Medyan < Üst çeyrek değerlerinden minumum olanı

Median $\tilde{x}$

The median minimizes the sum of absolute differences:

$$\sum_{i=1}^{n} |x_i - \tilde{x}| = \min . \qquad \text{and thus it is} \qquad \sum_{i=1}^{n} \operatorname{sgn}(x_i - \tilde{x}) = 0$$

If $x = (x_{(1)}, \ldots, x_{(n)})$ is a sorted data set, the median is defined as

$$\tilde{x} = \begin{cases} x_{(\frac{n+1}{2})}, & \text{if } n \text{ is odd,} \\ \frac{1}{2}\left(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\right), & \text{if } n \text{ is even.} \end{cases}$$

The median is applicable to ordinal and metric attributes.
(For non-metric attributes either $x_{(\frac{n}{2})}$ or $x_{(\frac{n}{2}+1)}$ needs to be chosen for even n.)

n is odd: 5, 28, 8, 10, 9

Ordered data 5, 8, 9, 10, 28

i =(5+1)/2=3

Median is 3rd value which is 9.

n is even: 19, 20, 17, 27, 6, 21

Ordered data 6, 17, 19, 20, 21, 27

i=(6+1)/2=3.5

Median is halfway between the 3rd and 4th values, which is 19.5.

Central value: The Mode

- The mode is the most frequently occurring number in a distribution
  - if $X = [1,2,4,7,7,7,8,10,12,14,17]$
  - then 7 is the mode
- Easy to see in a simple frequency distribution
- Possible to have no modes or more than one mode
  - *bimodal* and *multimodal*
- Don't have to be exactly equal frequency
  - *major mode, minor mode*
- Mode is not affected by extreme values

Yerelleştirme Ölçüleri: Mod

Mod, örnekte en sık görülen özellik değeridir. Benzersiz olması gerekmez, çünkü birkaç değer aynı frekansa sahip olabilir. En genel ölçüdür çünkü tüm ölçek türleri için geçerlidir.

Mod (Tepe) değeri:
Mod dağılım kümesinde olasılığı en yüksek değerdir. Dizide en çok tekrarlanma sayısıdır. Dağılımda en yüksek olasılık değeri birden fazla nokta ile temsil ediliyorsa, mod bu değerlerin hepsine karşılık geldiğinden sonuç tek anlamlı olmaktan çıkar. Böylesi durumlarda dağılımın bimodal, trimodal ya da multimodal olduğundan söz edilir. Frekans, bir saniyedeki titreşim sayısı, sıklık sayısıdır.

Örnek:
"Bu cümlenin her bir kelimesindeki harfleri sayın ve modu verin." Cümlenin içindeki ilk 10 harflerin sayılarını

B – 2 U – 2 C – 1 Ü – 1 M – 3 L – 3 E – 8 N – 5 İ – 7 H – 2
R – 4 K – 2 S – 2 D – 2 A – 2

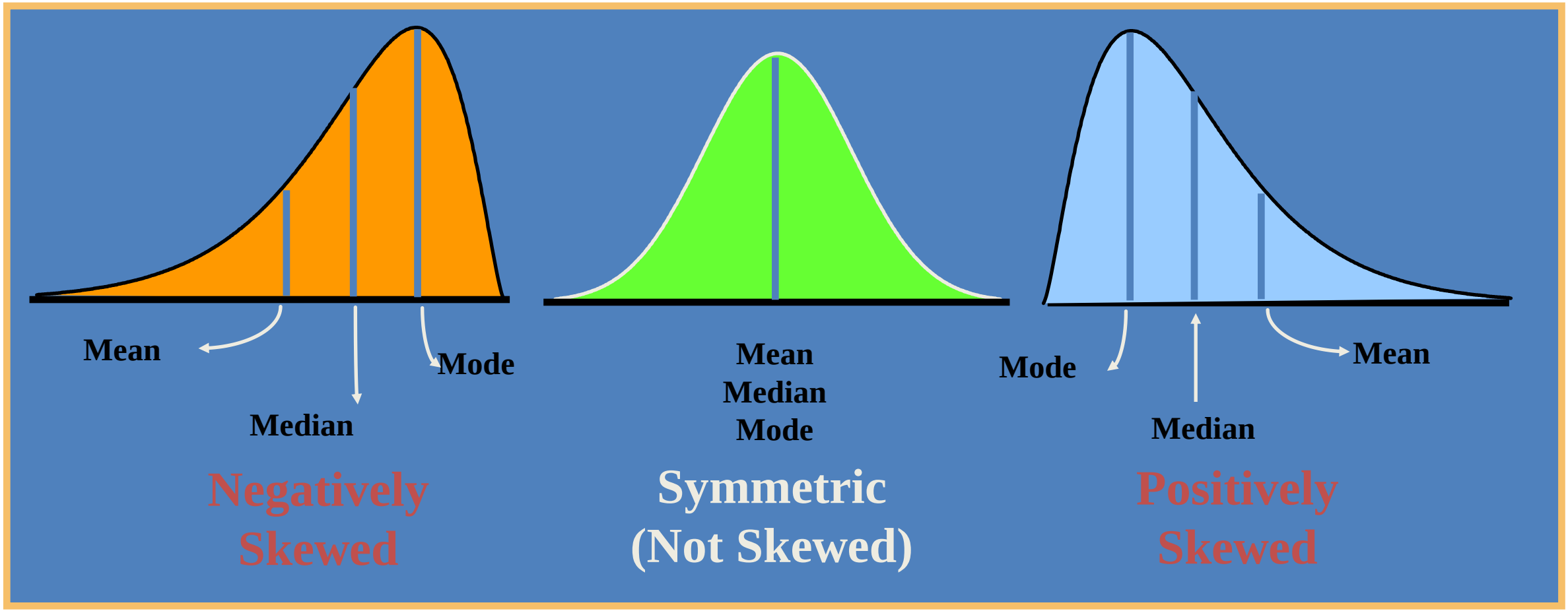
2 2 1 1 3 3 8 5 7 2 4 2 2 2 2

Verileri tararken, modun 2 olduğunu görüyoruz, çünkü ikişer defa tekrarlanan çok fazla harf olduğunu görüyoruz. Mod=2

When to Use What

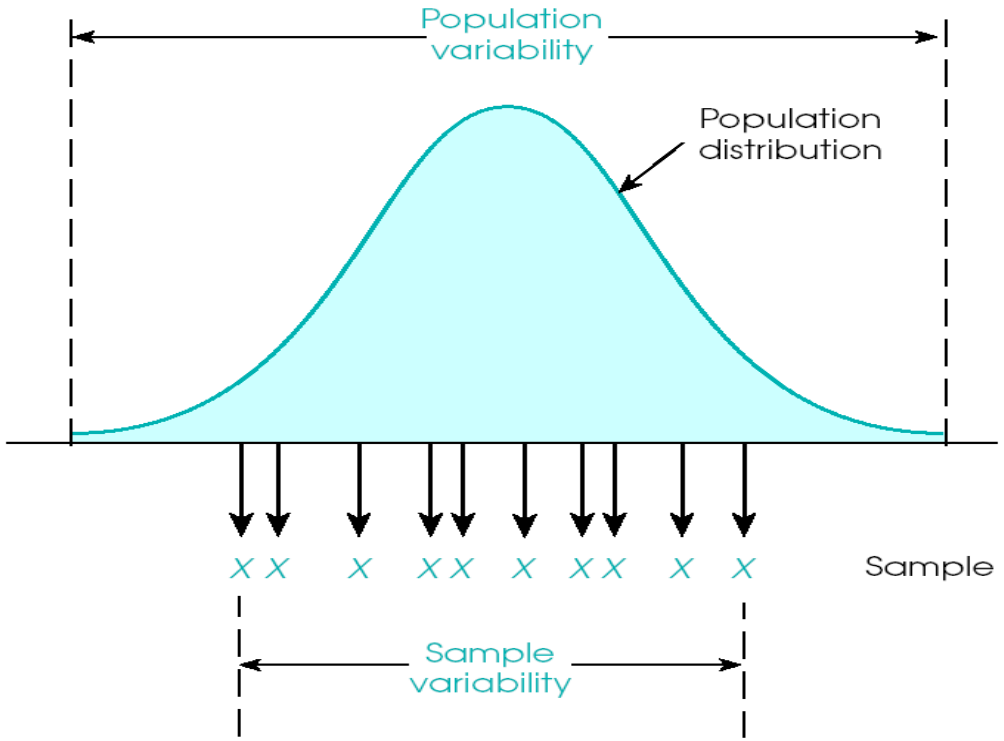
- Mean is a great measure. But, there are time when its usage is inappropriate or impossible.
  - Nominal data: Mode
  - The distribution is bimodal: Mode
  - You have ordinal data: Median or mode
  - Are a few extreme scores: Median

Mean, Median, Mode



Dispersion

- Dispersion
  - How tightly clustered or how variable the values are in a data set.
- Example
  - Data set 1: [0,25,50,75,100]
  - Data set 2: [48,49,50,51,52]
  - Both have a mean of 50, but data set 1 clearly has greater *Variability* than data set 2.



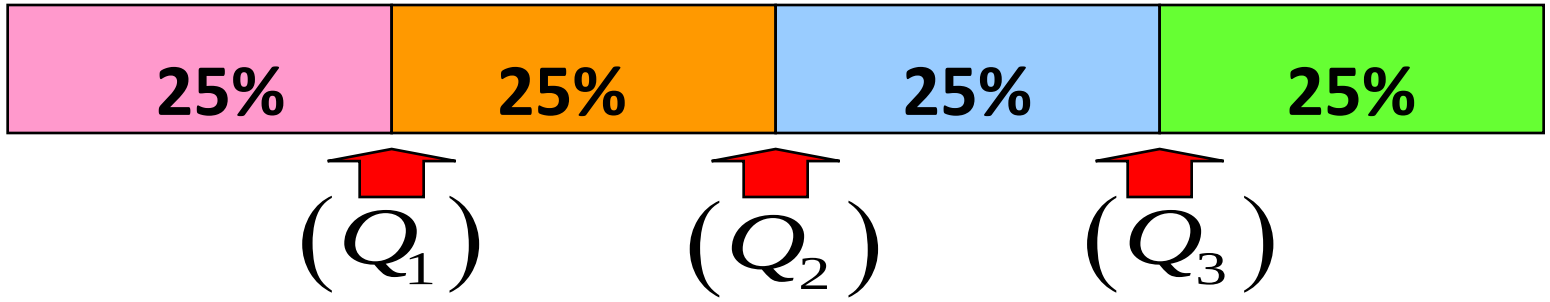
Dispersion: The Range

- The *Range* is one measure of dispersion
  - The range is the difference between the maximum and minimum values in a set
- Example
  - Data set 1: [1,25,50,75,100]; R: 100-1 +1 = 100
  - Data set 2: [48,49,50,51,52]; R: 52-48 + 1= 5
  - *The range ignores how data are distributed and only takes the extreme scores into account*

$$-RANGE = (X_{largest} - X_{smallest}) + 1$$

Quartiles

- Split Ordered Data into 4 Quarters



- Q_1 first quartile
- Q_2 second quartile= Median
- Q_3 third quartile

Dispersion: Interquartile Range

- Difference between third & first quartiles
  - Interquartile Range = $Q_3 - Q_1$
- Spread in middle 50%
- Not affected by extreme values

Yerelleştirme Ölçüleri: Example

Data {1,3,7,3,2,3,6,7}

- Mode : 3

Data {1,3,7,3,2,3,6,7,1,1}

- Mode : 1 and 3

Data {1,3,7,0,2,-3, 6,5,-1}

- Mode : No mode

Suppose the age in years of the first 10 subjects enrolled in your study are:

34, 24, 56, 52, 21, 44, 64, 44, 42, 46

Then the mean age of this group is 42.7 years

To find the median, first order the data:

21, 24, 34, 42, 44, 44, 46, 52, 56, 64

The median is (44+44)/2 = 44 years

Suppose the next patient enrolls and her age is 97 years.

How does the mean, median and mode change?

Ordered data:

21, 24, 34, 42, 44, 44, 46, 52, 56, 64, 97

Mean is 47,6	↑	42,7
Median is 44	⟷	44
Mode is 44	⟷	44

Karşılaştırma: Ortalama değer ile Medyan

Comparison of Mean and Median

- Mean is sensitive to "outliers" (a few very large or small values), so sometimes mean does not reflect the quantity desired.

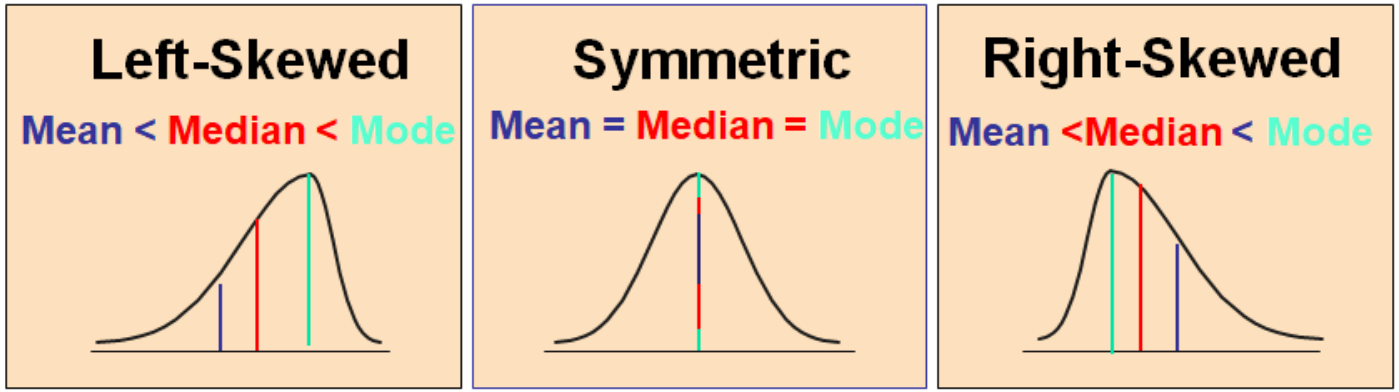
$20, 21, 22, 23, 24, 25, 26, 90 \rightarrow \bar{x} = 31{,}38$

87.5% of observations

- Median is "resistant" to outliers

Median = 23.5

- Mean is attractive mathematically.
- 50% of sample is above the median, 50% of sample is below the median.



Yerelleştirme Ölçüleri: Aritmetrik Ortalama

- **Arithmetic Mean** $\bar{x}$

  The arithmetic mean minimizes the sum of squared differences:

  $$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \min . \qquad \text{and thus it is} \qquad \sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - n\bar{x} = 0$$

  The arithmetic mean is defined as

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

  The arithmetic mean is only applicable to metric attributes.

Aritmetik ortalama en yaygın yerelleştirme ölçüsü olsa da, aşağıdaki koşullarda medyan tercih edilir:

- birkaç örnek vaka var,
- dağılım asimetriktir ve / veya
- aykırı değerlerin mevcut olması beklenir.

Aritmetik Ortalama

- Aritmetik ortalamayı bulmak için, tüm yanıt değerleri toplanır ve toplam, toplam yanıt sayısına bölünür. Toplam yanıt veya gözlem sayısına N denir.

Mean number of library visits

Data set	15, 3, 12, 0, 24, 3
Sum of all values	15 + 3 + 12 + 0 + 24 + 3 = 57
Total number of responses	*N* = 6
Mean	Divide the sum of values by *N* to find *M*: 57/6 = **9.5**

Age distribution of seven children attending to a children clinic is given below

$$\{1,3,6,7,2,3,5\}$$

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{\sum_{i=1}^{7} X_i}{7} = \frac{1+3+6+7+2+3+5}{7} = \frac{27}{7}$$

$$\overline{X} = 3{,}9 \text{ years}$$

Dağılım Ölçümleri: Aralık ve Çeyrekler Arası Aralık

Kafası dondurucuda ve ayakları fırında olan bir adam ortalama olarak oldukça rahattır.

- Bir veri kümesinin aralığı, maksimum ve minimum değer arasındaki farktır.

$$R = x_{\max} - x_{\min} = \max_{i=1}^{n} x_i - \min_{i=1}^{n} x_i$$

- Çeyrekler arası aralık: Bir veri setinin p-niceliği, tüm örnek değerlerin p'nin bir kısmının bu değerden daha küçük olacağı bir değerdir.

(The median is the $\frac{1}{2}$-quantile.)

The p-interquantile range, $0 < p < \frac{1}{2}$, is the difference between the $(1-p)$-quantile and the p-quantile.

The most common is the *interquartile range* ($p = \frac{1}{4}$).

Dağılım Ölçüleri: Ortalama Mutlak Sapma

Ortalama mutlak sapma, numune değerlerinin medyandan veya aritmetik ortalamadan mutlak sapmalarının ortalamasıdır.

- Average Absolute Deviation from the **Median**

$$d_{\tilde{x}} = \frac{1}{n}\sum_{i=1}^{n} |x_i - \tilde{x}|$$

- Average Absolute Deviation from the **Arithmetic Mean**

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

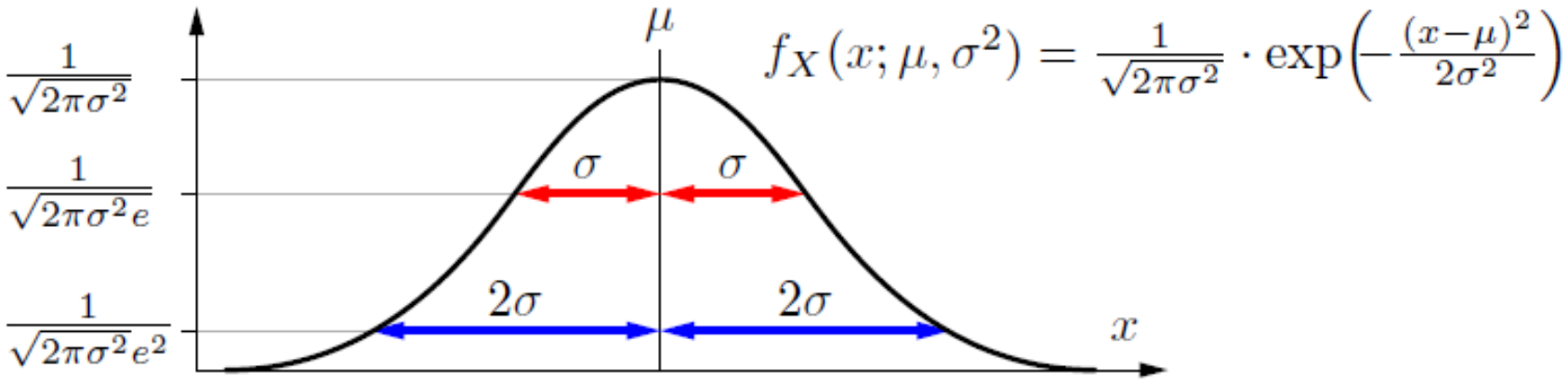
- It is always $d_{\tilde{x}} \leq d_{\bar{x}}$, since the median minimizes the sum of absolute deviations

Dağılım Ölçüleri: Varyans ve Standart Sapma

- Varyansı ortalama kare sapma olarak tanımlamak doğal olacaktır:
- $v^2 = \frac{1}{n}\sum_{i=1}^{n}(\mathrm{x}_i - \bar{\mathrm{x}})^2$
- Bununla birlikte, tümevarımsal istatistikler bunun daha iyi tanımlandığını göstermektedir:
- $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\mathrm{x}_i - \bar{\mathrm{x}})^2$
- Standart sapma, varyansın karekökü̈dür.
- Her bir verinin rtalama değeri ile farklarının karelerinin toplamı standart sapmayı verir.

Dağılım Ölçüleri: Varyans ve Standart Sapma

- Özel Durum: Normal / Gauss Dağılımı: Varyans / standart sapma, modun yüksekliği ve eğrinin genişliği hakkında bilgi sağlar.



μ:	expected value,	estimated by mean value $\bar{x}$
σ^2:	variance,	estimated by (empirical) variance s^2
σ:	standard deviation,	estimated by (empirical) standard deviation s

Dağılım Ölçüleri: Varyans ve Standart Sapma

- Aşağıdaki dönüşümden kaynaklanan formülü kullanarak varyansı hesaplamanın genellikle daha uygun olduğuna dikkat edin:

$$
\begin{aligned}
s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i^2-2x_i\bar{x}+\bar{x}^2\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2-2\bar{x}\sum_{i=1}^{n}x_i+\sum_{i=1}^{n}\bar{x}^2\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2-2n\bar{x}^2+n\bar{x}^2\right) = \frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2-n\bar{x}^2\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2-\frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2\right)
\end{aligned}
$$

Variance and standard deviation

Variance:

$$s^2 = \frac{\sum (X - \overline{X})^2}{n-1} = \frac{ss}{n-1} = \frac{SS}{df}$$

- ***deviation***
- ***squared-deviation***
- ***'Sum of Squares' = SS***
- ***degrees of freedom***

Standard Deviation of sample:

$$s = \sqrt{\frac{\sum (X - \overline{X})^2}{n-1}} = \sqrt{\frac{ss}{n-1}} = \sqrt{\frac{SS}{df}}$$

Standard Deviation for whole population:

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}}$$

Standard Deviation and Variance

- How much do scores deviate from the mean?
  - ***deviation*** = $X - \mu$

X	X-μ
1	
0	
6	
1	

$\mu = 2 \qquad \Sigma \qquad (X - \mu) = 0!$

- Why not just add these all up and take the mean?

Standard Deviation and Variance

- Solve the problem by squaring the deviations!

X	X-μ	$(X-\mu)^2$
1	-1	1
0	-2	4
6	+4	16
1	-1	1

$\mu = 2$

Variance =

$$\sigma^2 = \frac{\sum (X - u)^2}{N}$$

Sample variance and standard deviation

- Correct for problem by adjusting formula

$$s^2 = \frac{\sum (X - M)^2}{n - 1}$$

- *Different symbol:* $s2$ vs. $\sigma 2$
- *Different denominator:* n-1 vs. N
- *n-1 = "**degrees of freedom**"*
- Everything else is the same
- Interpretation is the same

Dağılım Ölçüleri: Varyans ve Standart Sapma

Örnek:

Step 1 x	Step 3 $(x-\bar{x})$	Step 4 $(x-\bar{x})^2$
6	1	1
3	-2	4
8	3	9
5	0	0
3	-2	4
25	0	18

Step 2 $$\bar{x} = \frac{\sum x}{n} = \frac{25}{5} = 5$$

Step 5 $$s^2 = \frac{\sum (x-\bar{x})^2}{n-1} = \frac{18}{4} = 4.5$$

$$s = \sqrt{s^2} = \sqrt{4.5} = 2.12$$

x	$(x-\bar{x})$	$(x-\bar{x})^2$
1	-4	16
3	-2	4
5	0	0
6	1	1
10	5	25
25	0	18

$$\bar{x} = \frac{\sum x}{n} = \frac{25}{5} = 5$$

$$s^2 = \frac{\sum (x-\bar{x})^2}{n-1} = \frac{46}{4} = 11.5$$

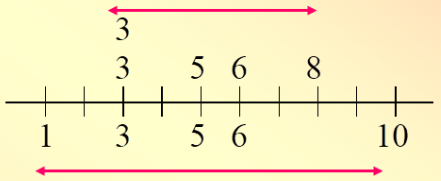
$$s = \sqrt{s^2} = \sqrt{11.5} = 3.39$$

NOTE: The sum of the deviation, $\sum_{i=1}^{n} (x_i - \bar{x})$, is always zero.

The last set of data is more dispersed than the previous set, and therefore its variance is larger.

First sample $s^2=4.5$

Second sample $s^2=11.5$



Measures of Dispersion (Variance and Standard Deviation)

Variance:

$$var(X) = E[(X-\mu)^2] = \begin{cases} \sum_x (x-\mu)^2 \, f(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} (x-\mu)^2 \, f(x) \, dx & \text{if } X \text{ is continuous} \end{cases}$$

Standard Deviation:

$$\begin{aligned} \sigma^2 = var(X) &= E[(X-\mu)^2] = E[X^2 - 2\mu X + \mu^2] \\ &= E[X^2] - 2\mu E[X] + \mu^2 = E[X^2] - 2\mu^2 + \mu^2 \\ &= E[X^2] - (E[X])^2 \end{aligned}$$

Measures of Dispersion (Variance and Standard Deviation)

Variance:

$$var(X) = E[(X-\mu)^2] = \begin{cases} \sum_x (x-\mu)^2 \, f(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} (x-\mu)^2 \, f(x) \, dx & \text{if } X \text{ is continuous} \end{cases}$$

Standard Deviation:

$$\begin{aligned} \sigma^2 = var(X) &= E[(X-\mu)^2] = E[X^2 - 2\mu X + \mu^2] \\ &= E[X^2] - 2\mu E[X] + \mu^2 = E[X^2] - 2\mu^2 + \mu^2 \\ &= E[X^2] - (E[X])^2 \end{aligned}$$

Sample Variance & Standard Deviation:

$$\hat{\sigma}^2 = \sum_x (x-\hat{\mu})^2 \hat{f}(x) = \sum_x (x-\hat{\mu})^2 \left(\frac{\sum_{i=1}^n I(S_i = x)}{n} \right) = \frac{\sum_{i=1}^n (S_i - \hat{\mu})^2}{n}$$

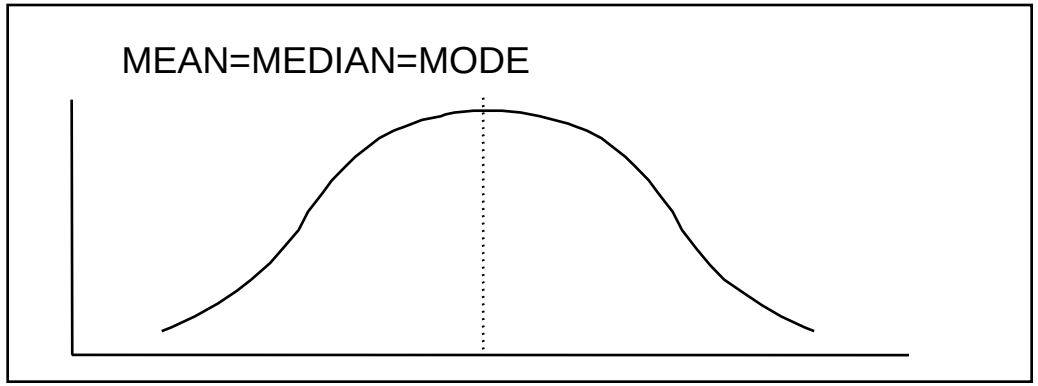
Dispersion: Standard Deviation

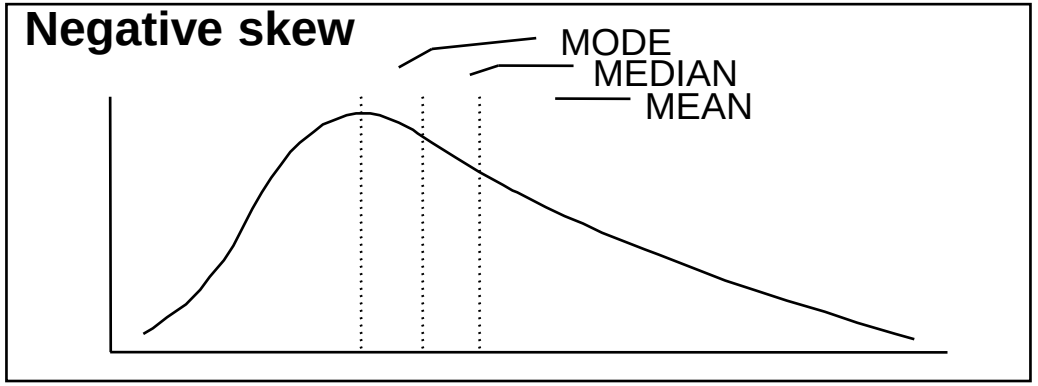
- let $X = [3, 4, 5, 6, 7]$
- $\overline{X} = 5$
- $(X - \overline{X}) = [-2, -1, 0, 1, 2]$
  - ⇧ subtract x from each number in X
- $(X - \overline{X})^2 = [4, 1, 0, 1, 4]$
  - ⇧ squared deviations from the mean
- $\Sigma (X - \overline{X})^2 = 10$
  - ⇧ sum of squared deviations from the mean (SS)
- $\Sigma (X - \overline{X})^2 / n-1 = 10/5 = 2.5$
  - ⇧ average squared deviation from the mean
- $\Sigma (X - \overline{X})^2 / n-1 = \sqrt{2.5} = 1.58$
  - ⇧ square root of averaged squared deviation

$$s = \sqrt{\frac{\sum (X - \overline{X})^2}{n-1}}$$

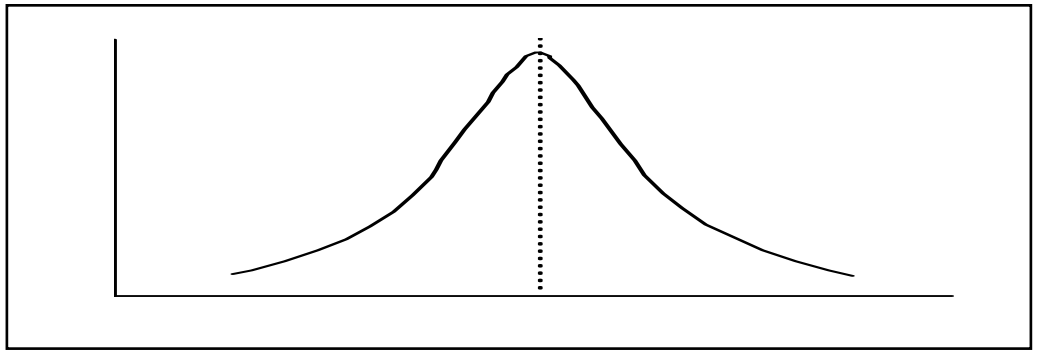
Symmetry

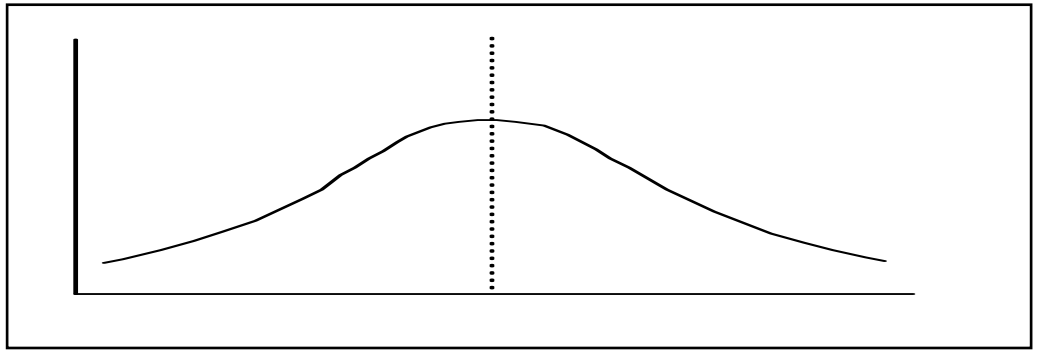
Skew - asymmetry





Kurtosis - peakedness or flatness



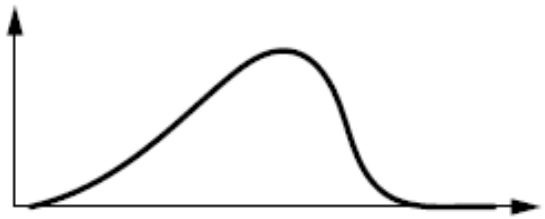


Şekil Ölçüleri: Çarpıklık (Skewness)

- Eğiklik α3 (veya kısaca çarpıklık), bir dağılımın simetrik bir dağılımdan farklı olup olmadığını ve ne kadar farklı olduğunu ölçer. Endeks 3'ü açıklayan ortalama ile ilgili 3. andan itibaren hesaplanır.

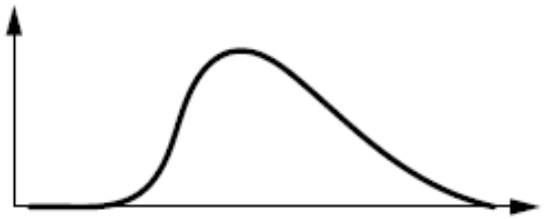
$$\alpha_3 = \frac{1}{n \cdot v^3} \sum_{i=1}^{n} (x_i - \bar{x})^3 = \frac{1}{n} \sum_{i=1}^{n} z_i^3$$

$$\text{where} \quad z_i = \frac{x_i - \bar{x}}{v} \quad \text{and} \quad v^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$



$\alpha_3 < 0$: right steep

$\alpha_3 = 0$: symmetric



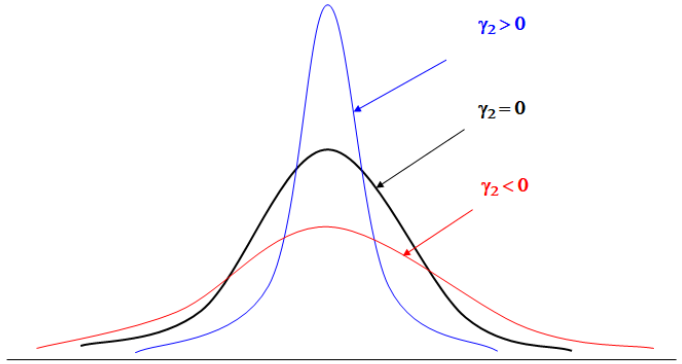
$\alpha_3 > 0$: left steep

Şekil Ölçüleri: Basıklık (Kurtosis)

- Eğiklik α3 (veya kısaca çarpıklık), bir dağılımın simetrik bir dağılımdan farklı olup olmadığını ve ne kadar farklı olduğunu ölçer. Endeks 3'ü açıklayan ortalama ile ilgili 3. andan itibaren hesaplanır.

$$\alpha_4 = \frac{1}{n \cdot v^4} \sum_{i=1}^{n} (x_i - \bar{x})^4 = \frac{1}{n} \sum_{i=1}^{n} z_i^4$$

$$\text{where} \quad z_i = \frac{x_i - \bar{x}}{v} \quad \text{and} \quad v^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

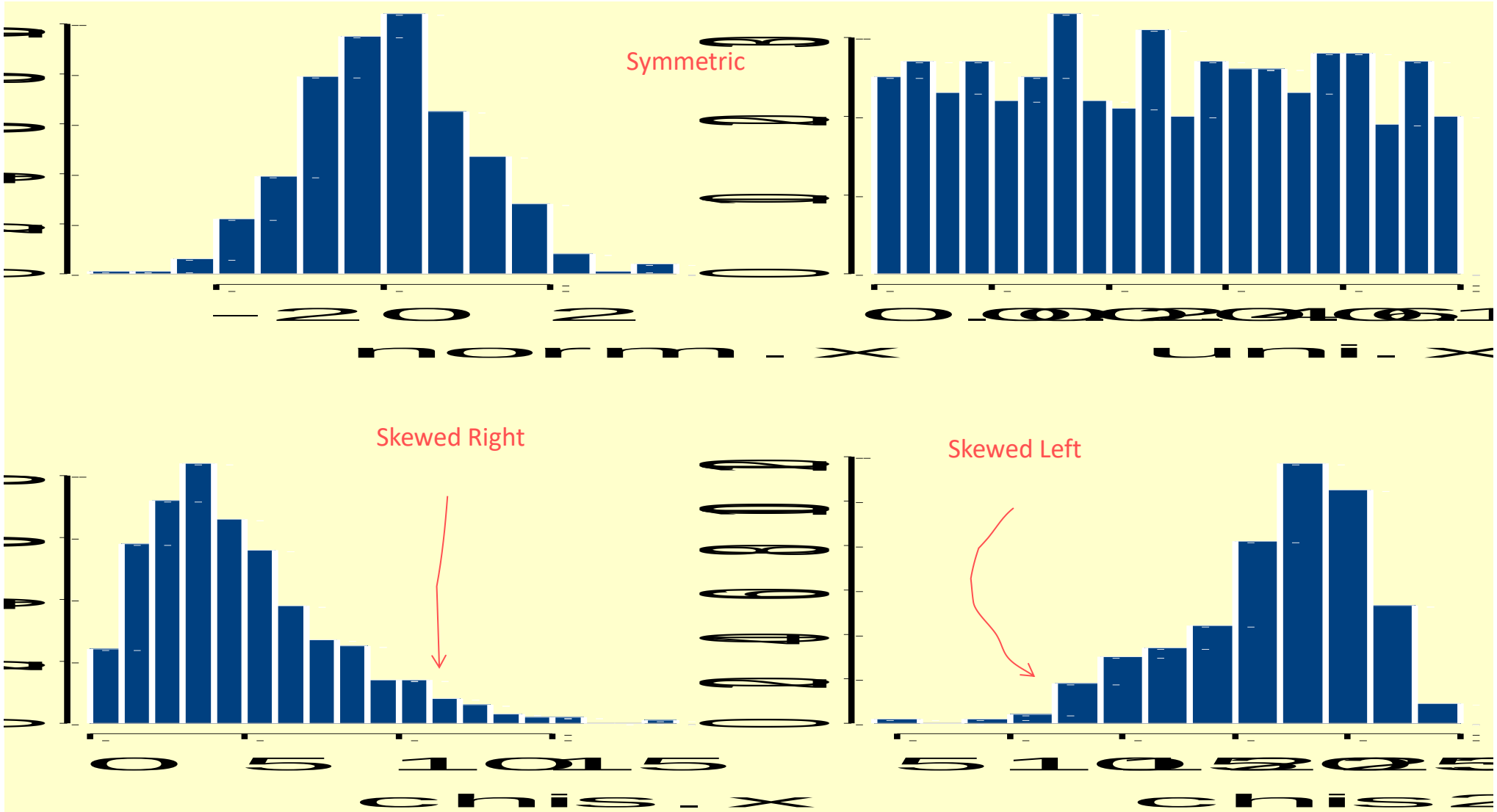


$\alpha_4 < 3$: leptokurtic $\quad \alpha_4 = 3$: Gaussian $\quad \alpha_4 > 3$: platikurtic

Symmetrical vs. Skewed Frequency Distributions

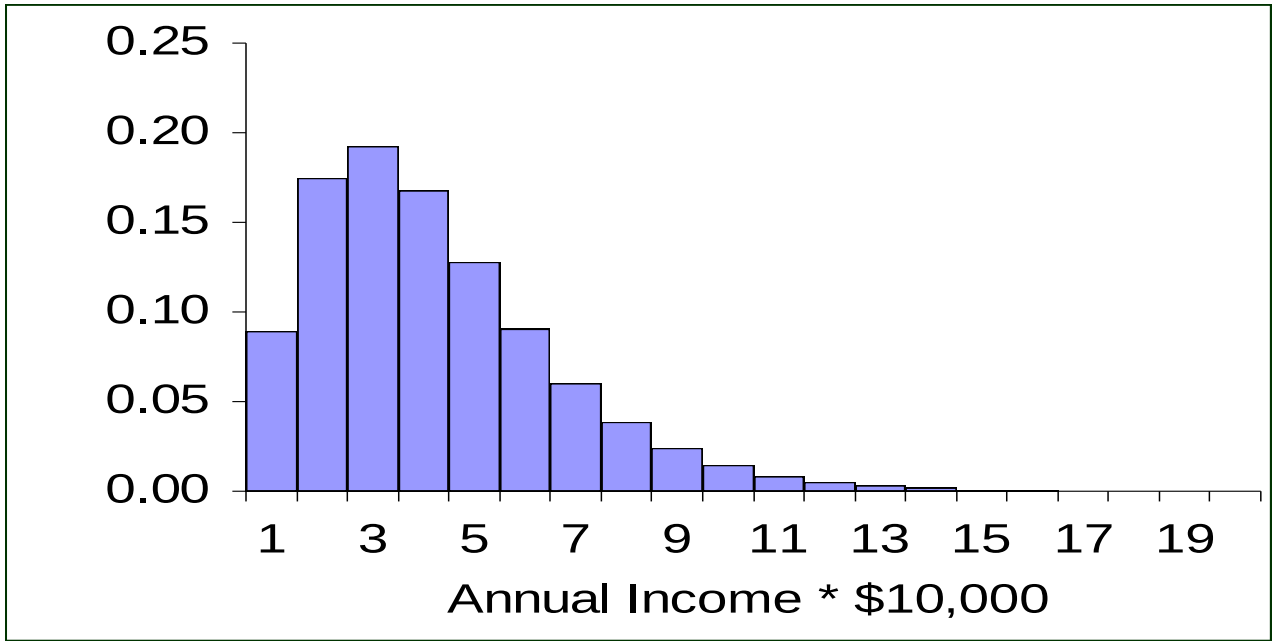
- Symmetrical distribution
  - Approximately equal numbers of observations above and below the middle
- Skewed distribution
  - One side is more spread out that the other, like a tail
  - Direction of the skew
    - Positive or negative (right or left)
    - Side with the fewer scores
    - Side that looks like a tail

Symmetrical vs. Skewed



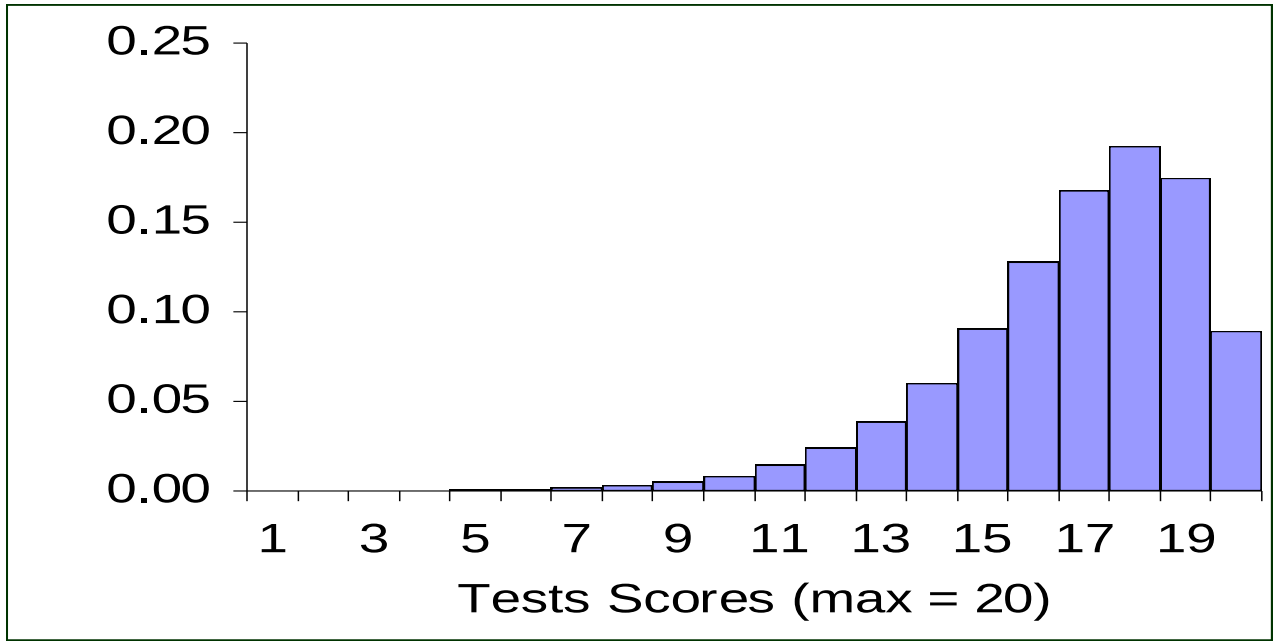
Skewed Frequency Distributions

- Positively skewed
  - AKA Skewed right
  - Tail trails to the right
  - ****** The skew describes the skinny end ******



Skewed Frequency Distributions

- Negatively skewed
  - Skewed left
  - Tail trails to the left



Symmetry: Skew

- The third `moment' of the distribution
- Skewness is a measure of the asymmetry of the probability distribution. Roughly speaking, a distribution has positive skew (right-skewed) if the right (higher value) tail is longer and negative skew (left-skewed) if the left (lower value) tail is longer (confusing the two is a common error).

$$g_1 = \frac{\sqrt{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{3/2}}$$

Symmetry: Kurtosis

- The fourth `moment' of the distribution

- A high kurtosis distribution has a sharper "peak" and fatter "tails", while a low kurtosis distribution has a more rounded peak with wider "shoulders".

$$g_2 = \frac{n\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} - 3$$

Accuracy (again!)

- *Accuracy:* **the closeness of the measurements to the "actual" or "real" value of the physical quantity.**

  - **Statistically this is estimated using the standard error of the mean**

Standard error of the mean

The mean of a sample is an estimate of the true (population) mean.

$$\bar{x} \approx \mu$$

The extent to which this estimate differs from the true mean is given by the *standard error of the mean*

$$SE\ (\bar{x}) = \frac{s}{\sqrt{N}}$$

s = standard deviation of the sample mean and describes the extent to which any single measurement is liable to differ from the mean

The standard error depends on the standard deviation and the number of measurements

$$\frac{1}{\sqrt{N}}$$

Often it is not possible to reduce the standard deviation significantly (which is limited instrument precision) so repeated measurements (high N) may improve the resolution.

Precision (again!)

- ***Precision:* is used to indicate the closeness with which the measurements agree with one another.**
  - Statistically the precision is estimated by the standard deviation of the mean

- **The assessment of the possible error in any measured quantity is of fundamental importance in science.**
  - Precision is related to random errors that can be dealt with using statistics
  - Accuracy is related to systematic errors and are difficult to deal with using statistics

Weighted average

A set of measurements of the same quantity, each given with a known error

$$x_1 \pm s_1$$
$$x_2 \pm s_2$$
$$x_3 \pm s_3$$
$$x_4 \pm s_4$$
$$\ldots\ldots\ldots$$

The mean value is calculated by "weighting" each of the measurements (x-values) according to its error.

$$\overline{x_{tot}} = \frac{\sum x_i / s_i^2}{\sum 1 / s_i^2}$$

with a standard deviation given by

$$s_{tot} = \sqrt{\frac{1}{\sum 1 / s_i^2}}$$

Veri Kümelerinin Momentleri

The k-th **moment** of a dataset is defined as

$$m'_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k.$$

The first moment is the **mean** $m'_1 = \bar{x}$ of the data set.
Using the moments of a data set the **variance** s^2 can also be written as

$$s^2 = \frac{1}{n-1}\left(m'_2 - \frac{1}{n}m'^2_1\right) \qquad \text{and also} \qquad v^2 = \frac{1}{n}m'_2 - \frac{1}{n^2}m'^2_1.$$

The k-th **moment about the mean** is defined as

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

It is $m_1 = 0$ and $m_2 = v^2$ (i.e., the **average squared deviation**).
The **skewness** is $\alpha_3 = \frac{m_3}{m_2^{3/2}}$ and the **kurtosis** is $\alpha_4 = \frac{m_4}{m_2^2}$.

Çok Boyutlu Karakteristik Ölçüler

Karakteristik ölçüleri vektörlere aktarın.

- Aritmetik ortalama: Çok boyutlu veriler için aritmetik ortalama, veri noktalarının vektör ortalamasıdır. İki boyut için

$$\overline{(\mathrm{x}, y)} = \frac{1}{n}\sum_{i=1}^{n}(\mathrm{x}_i, y_i) = (\,\bar{\mathrm{x}}, \bar{y}\,)$$

- Aritmetik ortalama için birkaç boyuta geçiş, yalnızca tek tek boyutların aritmetik ortalamalarını tek bir vektörde birleştirir.
- Diğer tanımlar benzer şekilde aktarılır.
- Bununla birlikte, bazen, ikinci dereceden doğası nedeniyle adaptasyon gerektiren varyansa gelince, transfer yeni niceliklere yol açar.

Farklılık, Çelişki: Vektör Ürünleri

- Genel Fikir: Dağılım ölçülerini vektörlere aktarılır.
- Varyans için, farkın ortalamaya karesi genelleştirilmelidir.
- Prensipte, her iki vektör ürünü de bir genelleme için kullanılabilir.
- Ancak ikincisi, dağıtım hakkında daha fazla bilgi verir:
  - özniteliklerin (doğrusal) bağımlılığının bir ölçüsü,
  - dağılımın yöne bağlılığının bir açıklaması.

Inner Product
Scalar Product

$$\vec{v}^\top \vec{v} \qquad \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_m \end{pmatrix}$$

$$(v_1, v_2, \ldots, v_m) \quad \sum_{i=1}^{m} v_i^2$$

Outer Product
Matrix Product

$$\vec{v}\vec{v}^\top \qquad \begin{pmatrix} v_1, & v_2, & \ldots, & v_m \end{pmatrix}$$

$$\begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_m \end{pmatrix} \quad \begin{pmatrix} v_1^2 & v_1 v_2 & \cdots & v_1 v_m \\ v_1 v_2 & v_2^2 & \cdots & v_2 v_m \\ \vdots & & \ddots & \vdots \\ v_1 v_m & v_2 v_m & \cdots & v_m^2 \end{pmatrix}$$

Kovaryans matrisi

Covariance Matrix

Compute variance formula with vectors (square: outer product $\vec{v}\vec{v}^{\top}$):

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\begin{pmatrix} x_i \\ y_i \end{pmatrix} - \begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix}\right)\left(\begin{pmatrix} x_i \\ y_i \end{pmatrix} - \begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix}\right)^{\top} = \begin{pmatrix} s_x^2 & s_{xy} \\ s_{yx} & s_y^2 \end{pmatrix}$$

where s_x^2 and s_y^2 are variances and s_{xy} is the covariance of x and y:

$$s_x^2 = s_{xx} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)$$

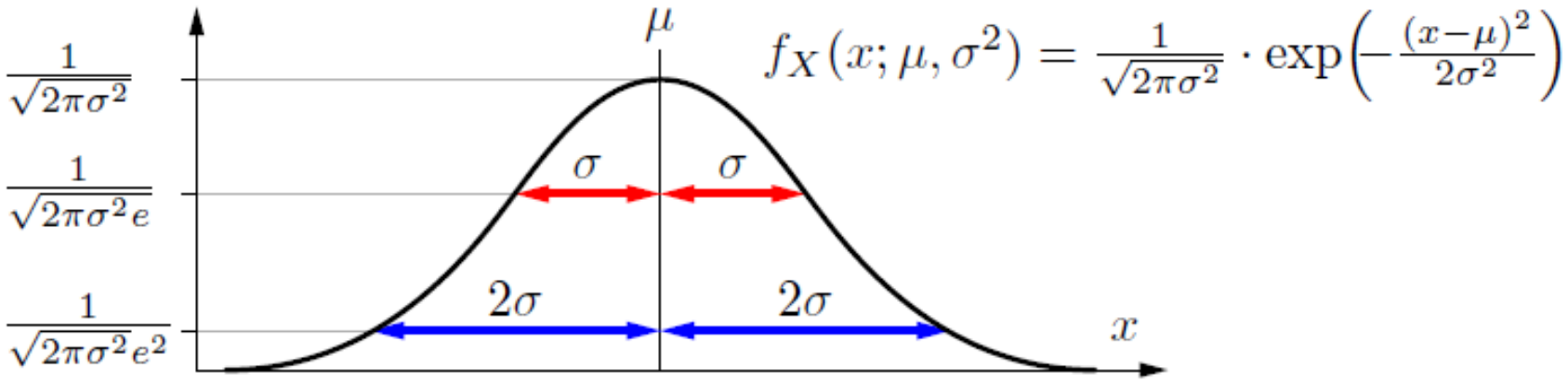
$$s_y^2 = s_{yy} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)$$

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

(Using $n-1$ instead of n is called Bessel's correction, after Friedrich Wilhelm Bessel, 1784–1846.)

Dağılım Ölçüleri: Varyans ve Standart Sapma

- Özel Durum: Normal / Gauss Dağılımı: Varyans / standart sapma, modun yüksekliği ve eğrinin genişliği hakkında bilgi sağlar.



μ:	expected value,	estimated by mean value $\bar{x}$
σ^2:	variance,	estimated by (empirical) variance s^2
σ:	standard deviation,	estimated by (empirical) standard deviation s

Çok Değişkenli Normal Dağılım

- A **univariate normal distribution** has the density function

$$f_X(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

μ:	expected value,	estimated by mean value $\bar{x}$,
σ^2:	variance,	estimated by (empirical) variance s^2,
σ:	standard deviation,	estimated by (empirical) standard deviation s.

- A **multivariate normal distribution** has the density function

$$f_{\vec{X}}(\vec{x};\vec{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^m|\boldsymbol{\Sigma}|}} \cdot \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu})^\top \boldsymbol{\Sigma}^{-1}(\vec{x}-\vec{\mu})\right)$$

m:	size of the vector $\vec{x}$	(it is m-dimensional),
$\vec{\mu}$:	expected value vector,	estimated by mean value vector $\bar{\vec{x}}$,
$\boldsymbol{\Sigma}$:	covariance matrix,	estimated by (empirical) covariance matrix $\mathbf{S}$,
$\lvert\boldsymbol{\Sigma}\rvert$:	determinant of the covariance matrix $\boldsymbol{\Sigma}$.	

İki Değişkenli Tanımlayıcı İstatistikler
(Bivariate Descriptive Statistics)

- Birden fazla değişkenle ilgili veri topladıysanız, aralarında ilişki olup olmadığını keşfetmek için iki değişkenli veya çok değişkenli açıklayıcı istatistikleri kullanabilirsiniz.

- İki değişkenli analizde, birlikte değişip değişmediklerini görmek için aynı anda iki değişkenin sıklığını ve değişkenliğini incelersiniz. Daha fazla istatistiksel test yapmadan önce iki değişkenin merkezi eğilimini de karşılaştırabilirsiniz.

- Çok değişkenli analiz, iki değişkenli analizle aynıdır, ancak ikiden fazla değişken içerir.

Tablo Gösterimleri: Frekans Tablosu

- Given data set: $\vec{x} = (3, 4, 3, 2, 5, 3, 1, 2, 4, 3, 3, 4, 4, 1, 5, 2, 2, 3, 5, 3, 2, 4, 3, 2, 3)$

a_k	h_k	r_k	$\sum_{i=1}^{k} h_i$	$\sum_{i=1}^{k} r_i$
1	2	$\frac{2}{25} = 0.08$	2	$\frac{2}{25} = 0.08$
2	6	$\frac{6}{25} = 0.24$	8	$\frac{8}{25} = 0.32$
3	9	$\frac{9}{25} = 0.36$	17	$\frac{17}{25} = 0.68$
4	5	$\frac{5}{25} = 0.20$	22	$\frac{22}{25} = 0.88$
5	3	$\frac{3}{25} = 0.12$	25	$\frac{25}{25} = 1.00$

- **Absolute Frequency** h_k (frequency of an attribute value a_k in the sample).
- **Relative Frequency** $r_k = \frac{h_k}{n}$, where n is the sample size (here $n = 25$).
- **Cumulated Absolute/Relative Frequency** $\sum_{i=1}^{k} h_i$ and $\sum_{i=1}^{k} r_i$.

Absolute Frequency: Mutlak Frekans

Relative Frequency: Göreceli frekans

Cumulated Relative Frequency: Kümülatif Bağıl Frekans

Tablo Gösterimleri: Beklenmedik Durum Tabloları

- İki veya daha fazla öznitelik için sıklık tablolarına beklenmedik durum tabloları denir.
- Değer kombinasyonlarının mutlak veya göreli frekansını içerirler.
- Bir olasılık tablosu ayrıca marjinal frekansları, yani bireysel niteliklerin değerlerinin frekanslarını da içerebilir.
- Daha yüksek sayıda boyut (≥ 4) için beklenmedik durum tablolarının okunması zor olabilir.

	a_1	a_2	a_3	a_4	$\sum$
b_1	8	3	5	2	18
b_2	2	6	1	3	12
b_3	4	1	2	7	14
$\sum$	14	10	8	12	44

Görselleştirme

Summarizing data

- Frequency distribution, Set of categories with numerical counts
- Tables
  - Simplest way to summarize data
  - Data are presented as absolute numbers or percentages
- Charts and graphs
  - Visual representation of data
  - Data are presented as absolute numbers or percentages

Key messages

- Use the right graph for the right data
  - Tables – can display a large amount of data
  - Graphs/charts – visual, easier to detect patterns
  - Label the components of your graphic
- Interpreting data *a*dds meaning by making connections and comparisons to program
- Service data are good at tracking progress & identifying concerns – do not show causality

Common types of data visualizations?

Common types of data visualizations are:

- Charts
- Graphs
- Tables
- Maps
- Histograms

İyi Veri Görselleştirme

Kişi sayısı, konum, sıcaklık vb gibi çeşitli miktarlarla ilgili bilgileri anlamak ve aktarmak kolaydır.

Tablo Şekilli Gösterimler: Frekans Tablosu

Verilen veri kümesi: $\vec{x}$ = (3, 4, 3, 2, 5, 3, 1, 2, 4, 3, 3, 4, 4, 1, 5, 2, 2, 3, 5, 3, 2, 4, 3, 2, 3)

a_k	h_k	r_k	$\sum_{i=1}^{k} h_i$	$\sum_{i=1}^{k} r_i$
1	2	$\frac{2}{25} = 0.08$	2	$\frac{2}{25} = 0.08$
2	6	$\frac{6}{25} = 0.24$	8	$\frac{8}{25} = 0.32$
3	9	$\frac{9}{25} = 0.36$	17	$\frac{17}{25} = 0.68$
4	5	$\frac{5}{25} = 0.20$	22	$\frac{22}{25} = 0.88$
5	3	$\frac{3}{25} = 0.12$	25	$\frac{25}{25} = 1.00$

- Mutlak Frekans h_k(örnekte bir öznitelik a_k değerinin frekansı).
- Bağıl Frekans, $r_k = \frac{h_k}{n}$, burada n örneklem boyutudur (n = 25).
- Kümülatif Mutlak / Bağıl Frekans, $\sum_{i=1}^{k} h_i$ ve $\sum_{i=1}^{k} r_i$

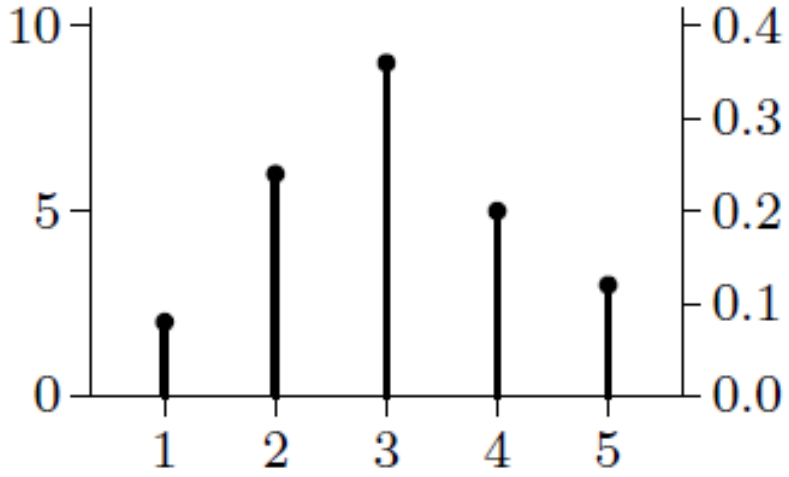
Tablo Şeklindeki Gösterimler: Olasılık Tabloları

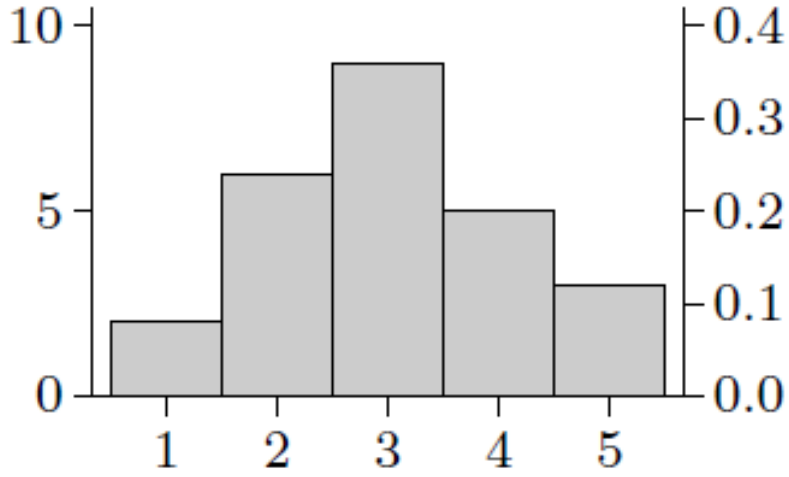
- İki veya daha fazla öznitelik için frekans tabloları, ihtimal durum tabloları olarak adlandırılır.
- Değer kombinasyonlarının mutlak veya göreceli sıklığını içerirler.
- Bir beklenmedik durum tablosu ayrıca marjinal frekansları, yani bireysel özniteliklerin değerlerinin frekanslarını da içerebilir.
- Daha yüksek sayıda boyut (≥ 4) için beklenmedik durum tablolarının okunması zor olabilir.

	a_1	a_2	a_3	a_4	$\sum$
b_1	8	3	5	2	18
b_2	2	6	1	3	12
b_3	4	1	2	7	14
$\sum$	14	10	8	12	44

Grafik Gösterimler: Kutup ve Çubuk Grafik

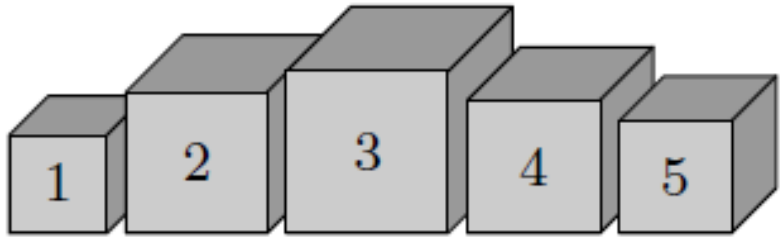
- Örneğin, öznitelik değerlerinin frekansları olabilecek sayılar, kutupların / çubukların uzunlukları (sol) veya çubukların yüksekliği (sağ) ile temsil edilir.
- Çubuk grafikler, mutlak frekansları görüntülemenin en sık kullanılan ve en anlaşılır yoludur.
- Dikey ölçek 0'da başlamazsa (frekanslar veya diğer mutlak sayılar için) yanlış bir izlenim ortaya çıkabilir.





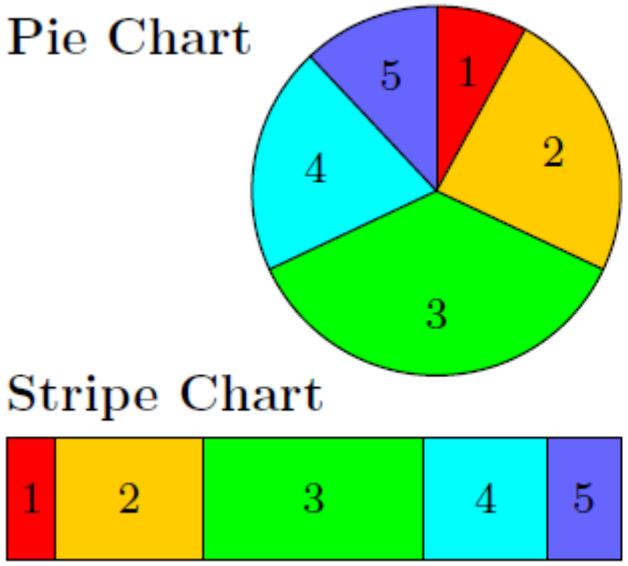
Alan ve Hacim Grafikleri

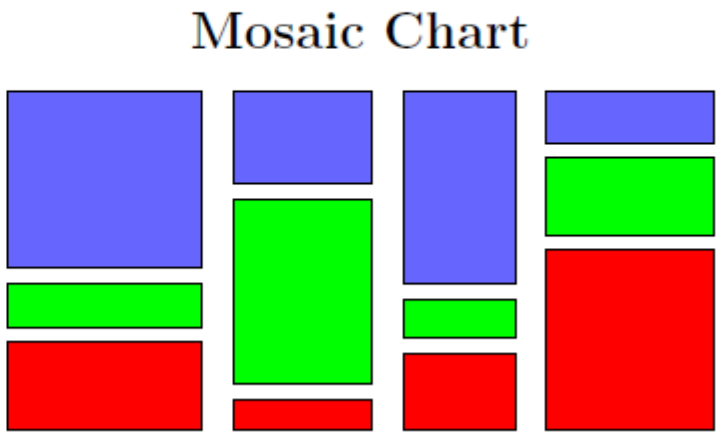
- Sayılar, alanlar veya hacimler gibi uzunluklardan farklı olarak geometrik miktarlarla da gösterilebilir.
- Alan ve hacim grafikleri genellikle çubuk grafiklerden daha az anlaşılırdır, çünkü insanlar alanların göreceli boyutunu ve özellikle hacimleri uzunluklardan ziyade karşılaştırmak ve değerlendirmek konusunda daha fazla zorlanırlar. (istisna: gösterilen sayılar alanları veya hacimleri tanımlar)
- Bazen iki veya üç boyutlu bir nesnenin yüksekliği bir sayıyı temsil etmek için kullanılır. Diyagram daha sonra yanıltıcı bir izlenim uyandırır.



Pasta ve Şerit Grafikler

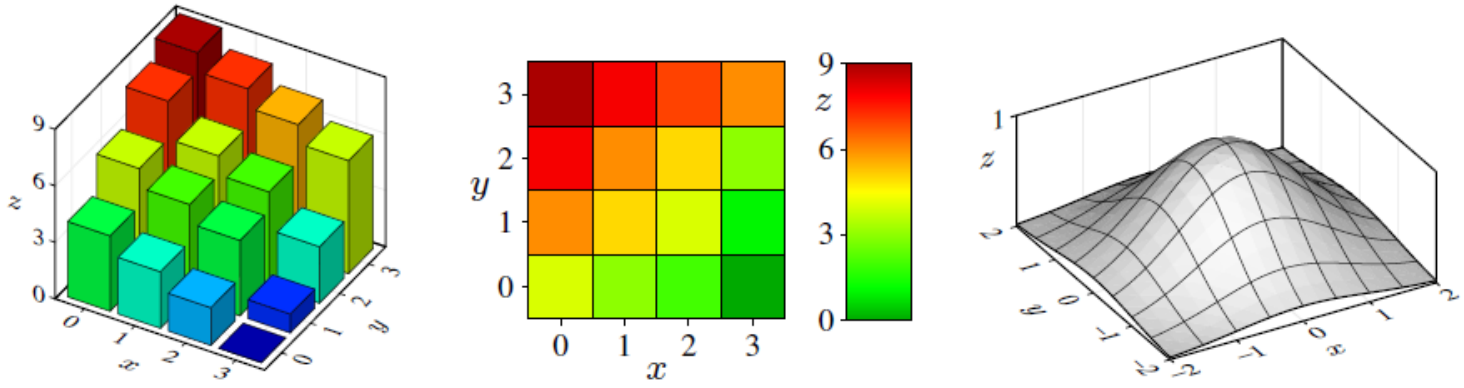
- Bağıl sayılar, bir şeridin açıları veya bölümleri ile temsil edilebilir.
- Mozaik grafikler, acil durum tablolarını görüntülemek için kullanılabilir.
- İkiden fazla özellik mümkündür, ancak bu durumda ayırma mesafeleri ve rengi, görselleştirmeyi anlaşılır kılmak için desteklemelidir.





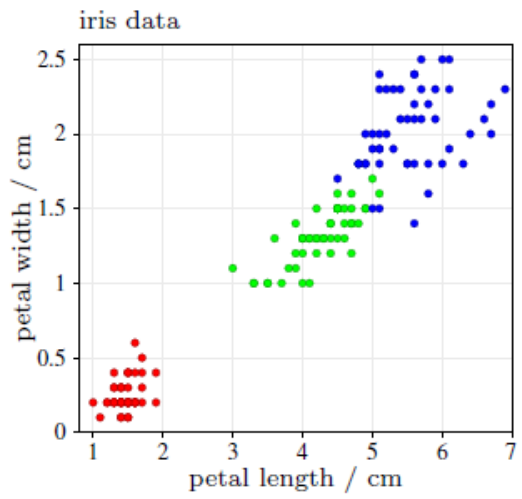
3 Boyutlu Diyagramlar

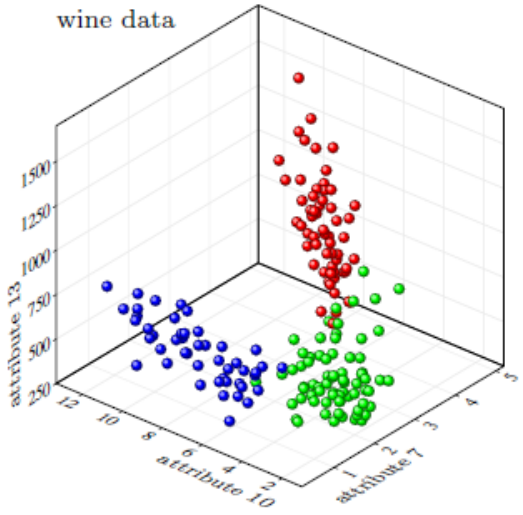
- Olasılık durum tablolarını görüntülemek için 3-boyutlu çubuk grafikler kullanılabilir (3. boyut, değer çifti frekansını temsil eder).
- 3. uzaysal boyut bir renk skalası ile değiştirilebilir. Bu tür bir grafik bazen ısı haritası olarak adlandırılır. (3-boyutlu bir çubuk grafikte renk, z değerini de kodlayabilir (fazlalık olarak).)
- Yüzey grafikleri, çizgi grafiklerin 3-boyutlu analoglarıdır.



Dağılım Grafikleri

- Dağılım grafikleri, 2 veya 3 boyutlu metrik veri kümelerini görüntülemek için kullanılır.
- Örnek değerler, bir noktanın koordinatlarıdır (yani, sayılar uzunluklarla temsil edilir).
- Dağılım grafikleri, bağımlılığı kontrol etmek için basit araçlar sağlar.





Data Analysis: *Graphing Frequency Distribution*

- **Graph of Frequency Distribution**:

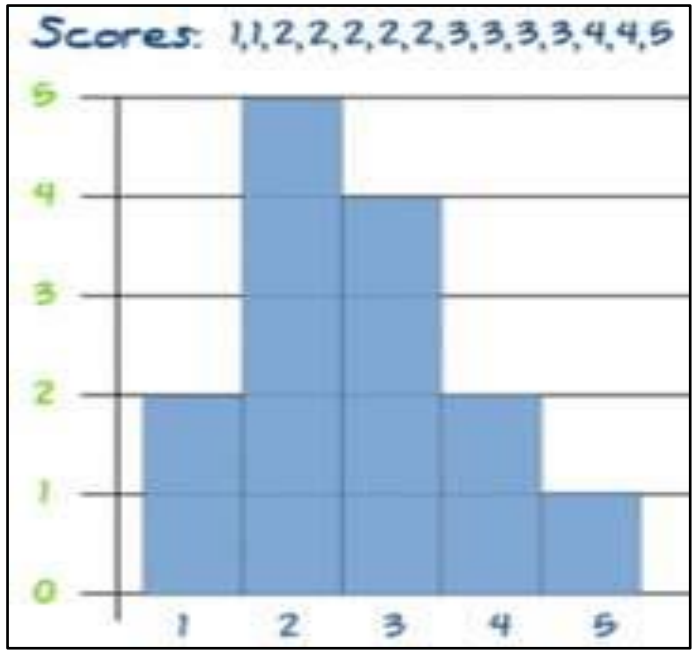
  - *The score categories (**X** values) are listed on the X axis and the frequencies (Number of categories of **X** values) are listed on the Y axis.*

  - When the score categories are numerical scores measured at interval or ratio level, the graph should be either a **Histogram** or a **Polygon**.

Data Analysis: *Histograms*

- In a **Histogram**, a bar/column is centered above each score (or Class Interval) so that the height of the bar/column corresponds to the *frequency* of the ***X*** values and the width of the bar/column extends to that adjacent bars/columns touch one another.

Histogram of Scores



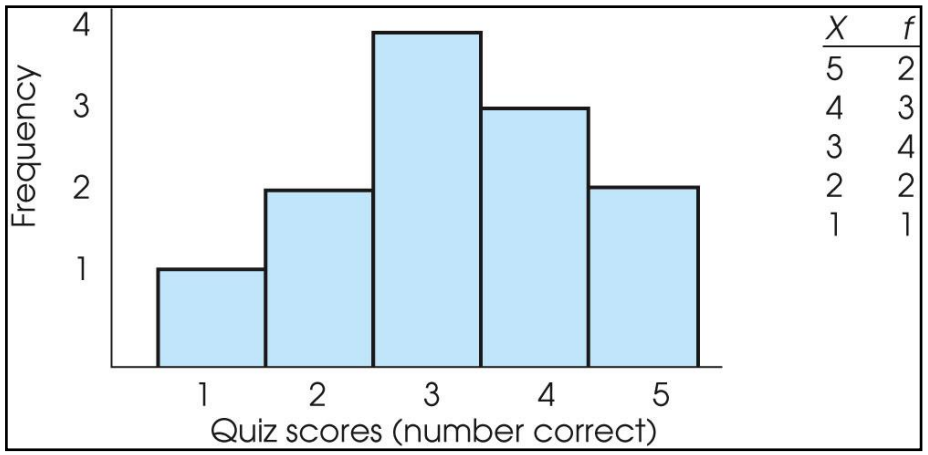
You will probably **never** have to draw a Histogram by hand beyond a class exercise.

Data Management and Analytical Software have automated reporting routines

Data Analysis:
Histograms

Also see **Age Distribution of Martians** examples from Sampling PowerPoint

Regular or Normal Frequency Distribution

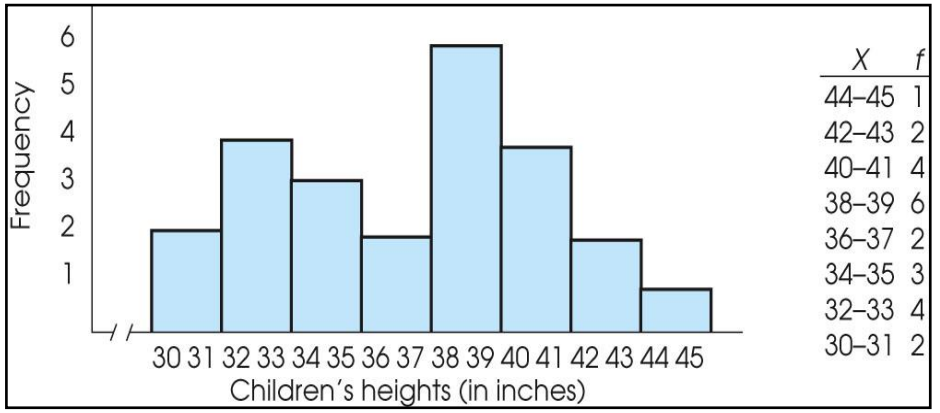


Table

X	f
5	2
4	3
3	4
2	2
1	1

A frequency distribution histogram: *same set of quiz scores as a table and in a histogram.*

Grouped Frequency Distribution



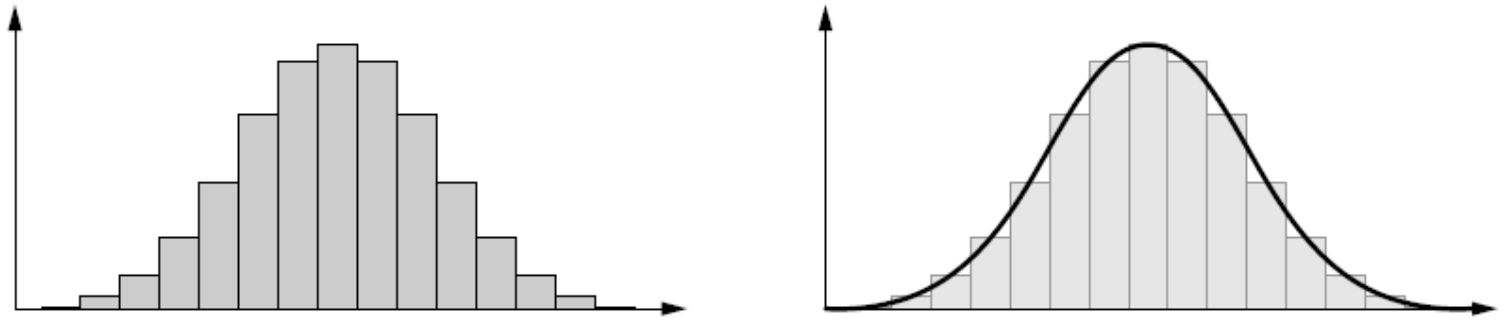
Table

X	f
44–45	1
42–43	2
40–41	4
38–39	6
36–37	2
34–35	3
32–33	4
30–31	2

A frequency distribution histogram for grouped data: *same set of children's as a table and in a histogram.*

Histogramlar

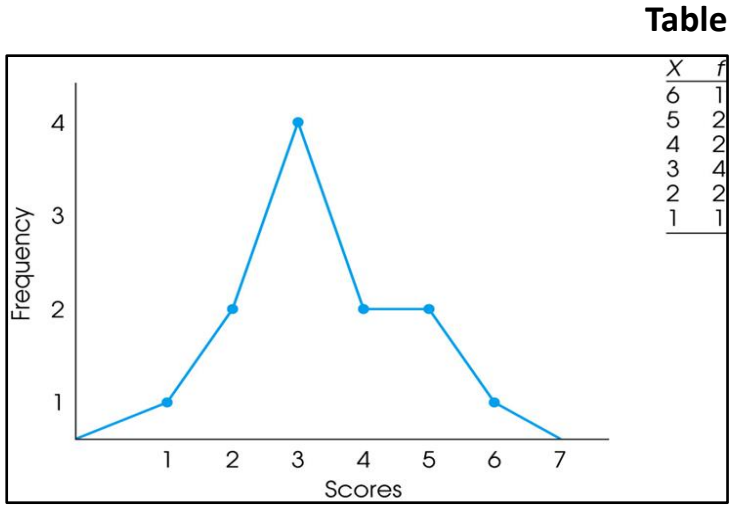
- Sezgisel: Histogramlar, metrik veriler için sıklık çubuk grafiklerdir.
- Ancak: Çok fazla farklı değer olduğundan, uygun bir temsile ulaşmak için değerlerin gruplanması gerekir.
- En yaygın yaklaşım: eşit büyüklükte aralıklar (sözde bölmeler) oluşturun ve her aralıktaki örnek değerlerin sıklığını sayın.
- Dikkat: Bölmelerin boyutuna ve konumuna bağlı olarak histogram önemli ölçüde farklı görünebilir.
- Eskizlerde genellikle bir histogramın yalnızca kaba bir taslağı çizilir:

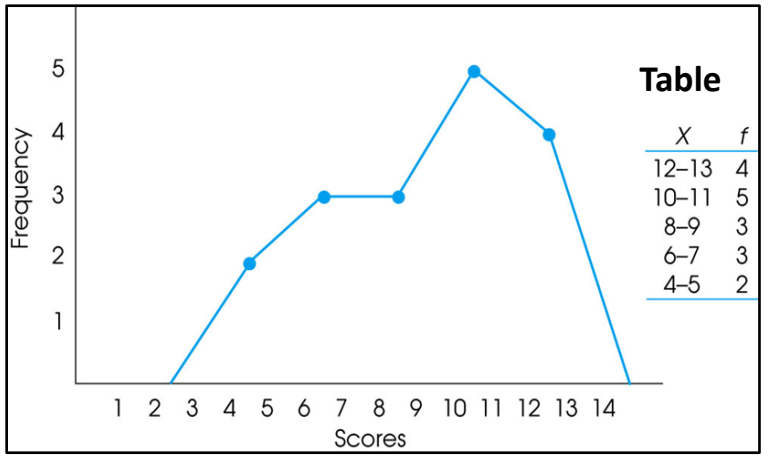


Data Analysis: *Polygons & Plots*

- **Polygon/ Plots**: a dot or point is centered above each score so that the height of the dot corresponds to the frequency.
  - Then straight lines connect those dots/ points
  - The graph is centered to a zero frequency by drawing additional lines at each end
- These descriptions are bit hard to visualize, but you see histograms and plots all the time: *visualizations of data*

Frequency Distribution Polygon: *same set of data as a table and in a polygon.*

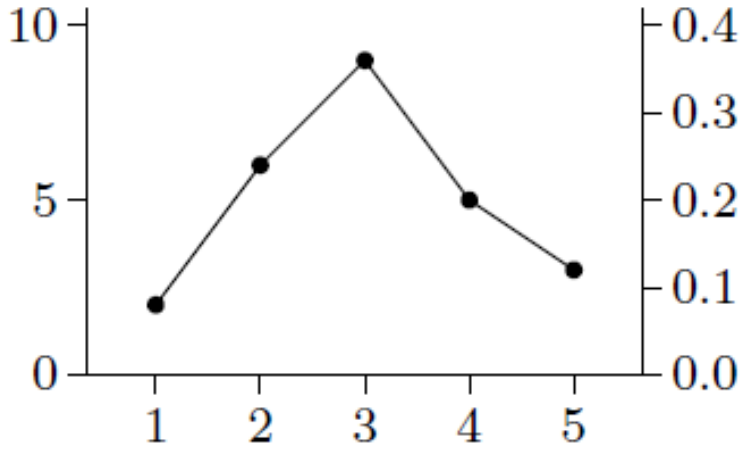


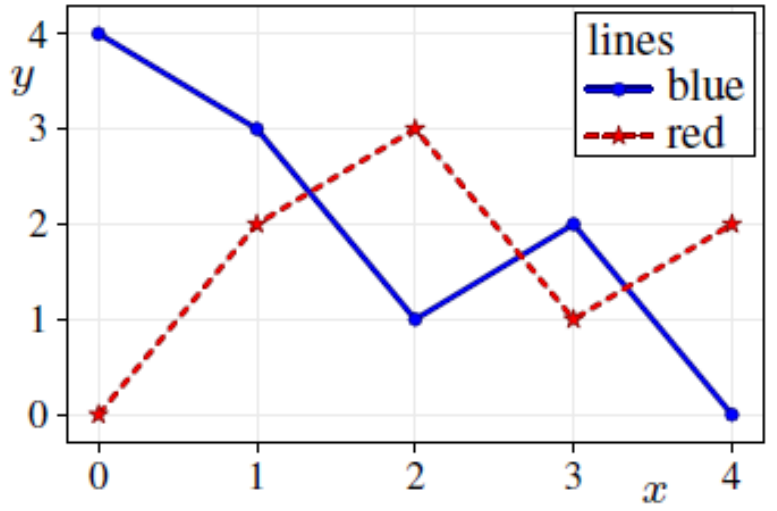


Frequency Distribution Polygon for Grouped Data: *same set of data as a grouped table and in a polygon.*

Frekans Poligonu ve Çizgi Grafik

- Frekans poligonu: Bir kutup grafiğinin kutuplarının uçları çizgilerle birbirine bağlanır. (Sayılar hala uzunluklarla temsil edilmektedir.)
- Yatay eksendeki öznitelik değerleri sıralanmamışsa, kutupların uçlarını bağlamak anlamsızdır.
- Zaman serilerini görüntülemek için sıklıkla çizgi grafikler kullanılır.



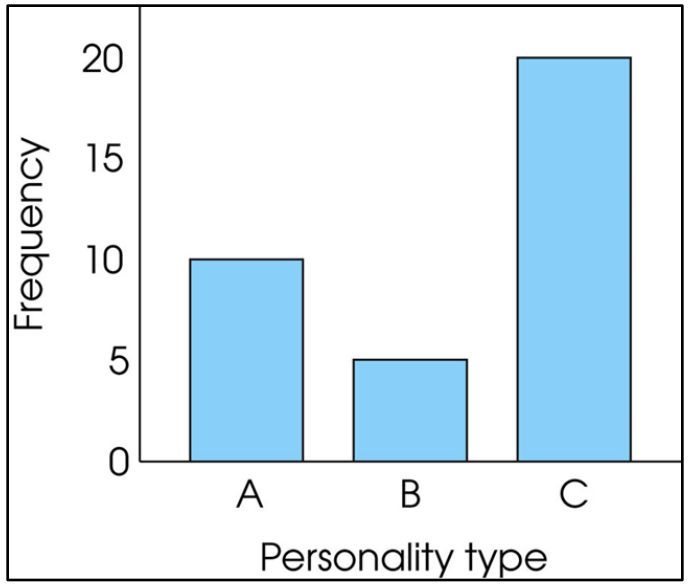


Data Analysis: *Bar Graphs*

- **Bar Graph** are appropriate when the score categories (X values) are measurements at nominal or ordinal level
- A **Bar Graph** is just like a **Histogram** except that there are gaps or spaces between adjacent bars/ columns.

Personality Type Bar Graph

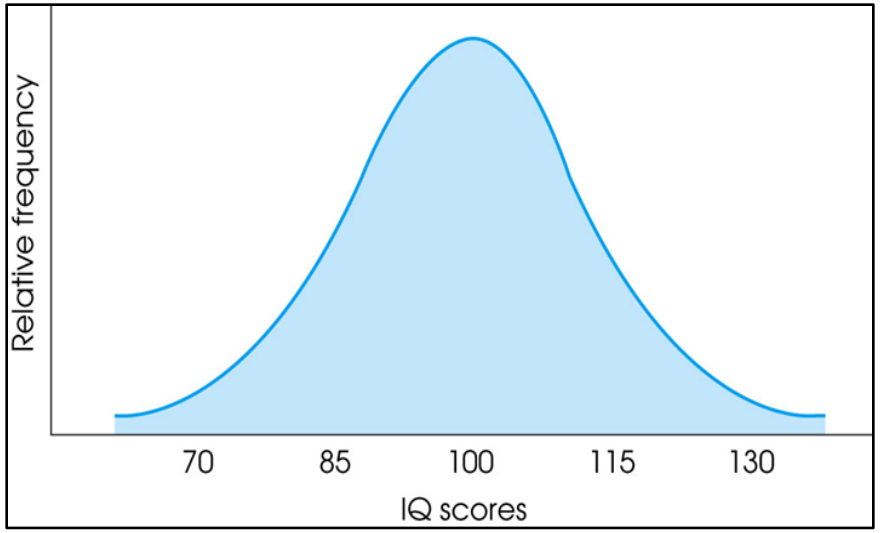
- A Bar Graph showing the distribution of personality types in a sample of college students.
- Because personality type is a discrete variable measured on a nominal scale, the graph is drawn with space between the bars.



Data Analysis: *Smooth Curve*

- The conventional display of a distribution of *interval* or *ratio level* scores is a **Smooth Curve:** *not jagged Histogram or Polygon*

- The Smooth Curve emphasizes the shape of the distribution: *not the exact frequency for each category*



The population distribution of IQ scores: *an example of a* ***Normal Distribution****.*

Data Analysis:

Frequency Distributions, Graphs, Plots & Histograms

- *Graphs, Plots* & *Histograms* of *Frequency Distributions* are useful because they show the entire set of scores.
  - These *info-grpahics* quickly allow you to see the highest score, the lowest score, and where the scores are centered.

- These data visualizations also show how the scores are clustered together or scattered apart.

Data Analysis:

Frequency Distributions, Graphs, Plots & Histograms

- A graph shows the **shape** of the distribution.
  - A distribution is **Symmetrical** if the left side of the graph is (roughly) a mirror image of the right side.
  -
- A familiar example of a Symmetrical Distribution is the bell-shaped normal distribution: *the bell curve*.

- Distributions are **skewed** when scores pile up on one side of the distribution: *leaving a "tail" of a few extreme values on the other side*

Frequency Distributions

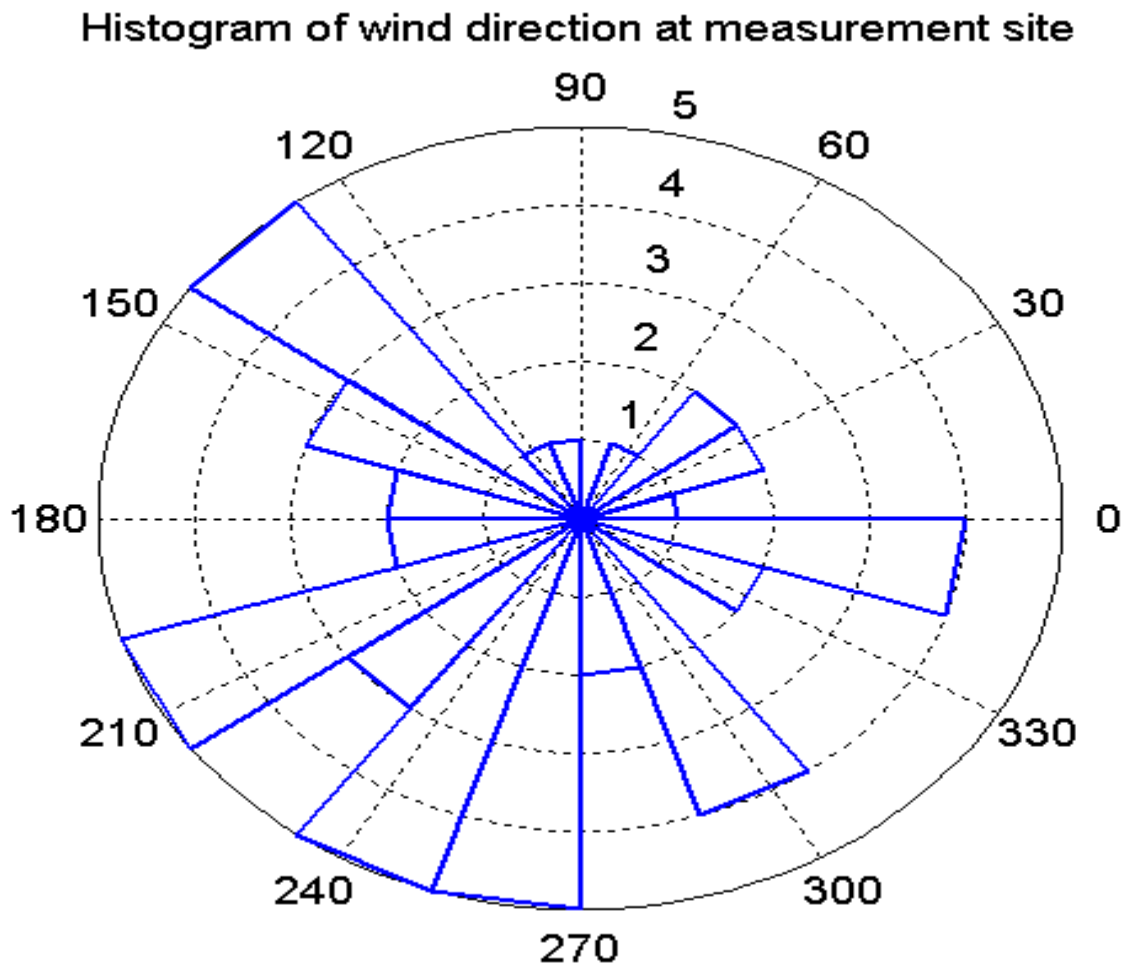
A table that summarizes raw data by showing the number of scores that fall within each of the categories

Pounds Lost by Dieter	Frequency
0	5
1-5	10
6-10	15
Over 10	5

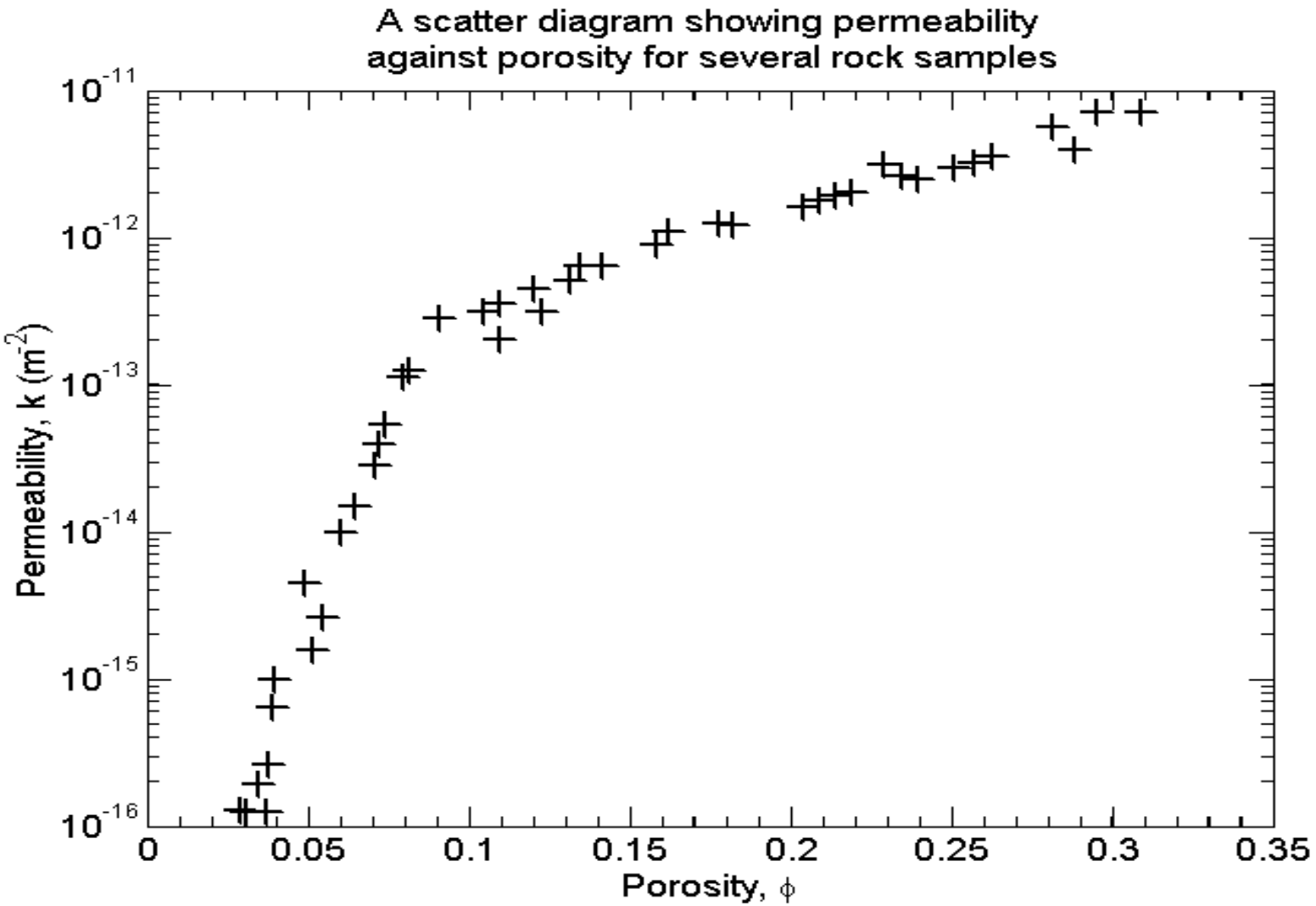
Frequency Histograms and Polygons

- Sometimes information in frequency distributions is more easily grasped when it is presented graphically
- Histogram is used when horizontal (x-axis) variable is measured on an interval or ratio scale (bars on graph touch each other)
- If data is nominal or ordinal, the bars do not touch each other and it is a bar graph

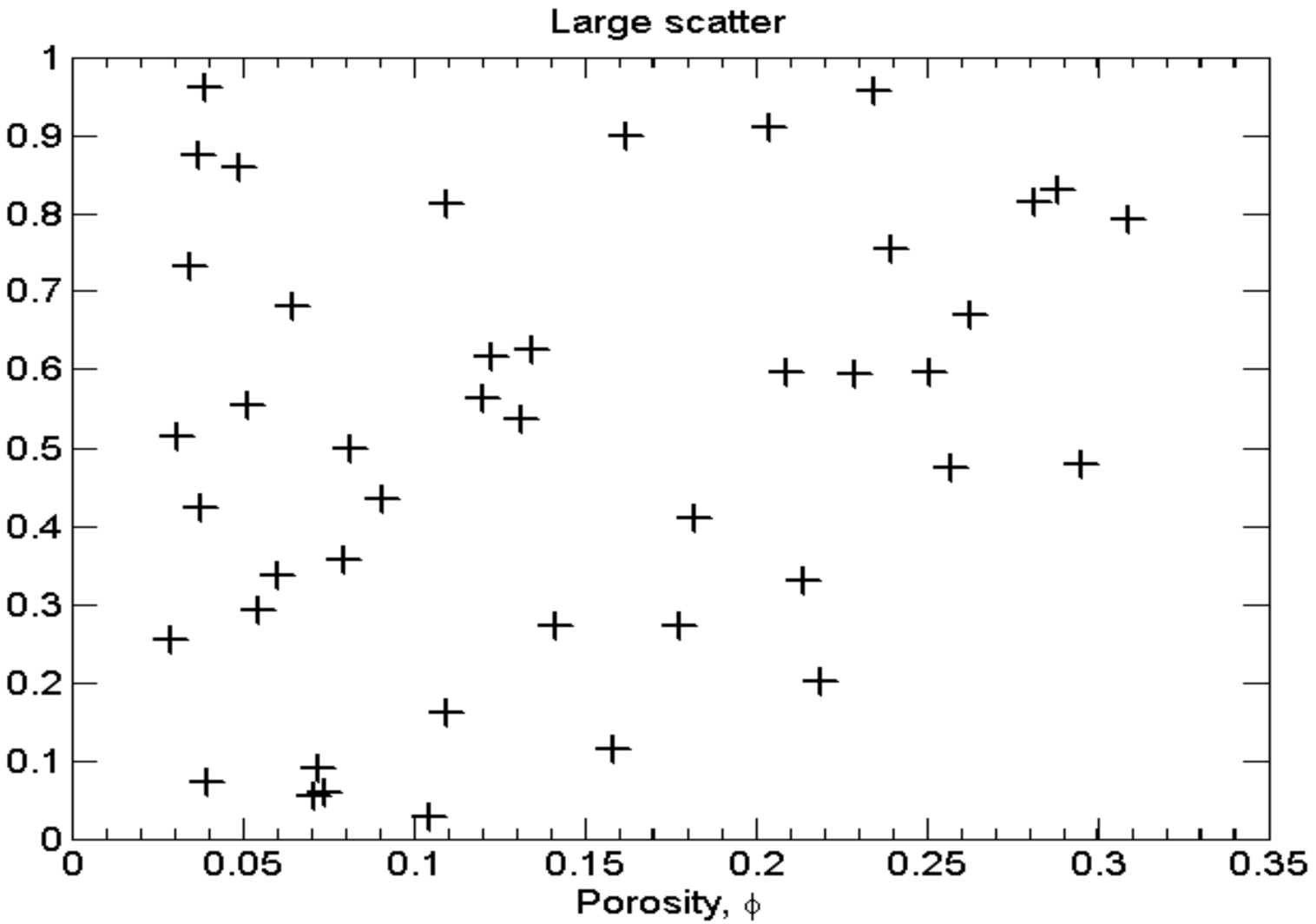
Graphing data – rose diagram



Graphing data – scatter diagram



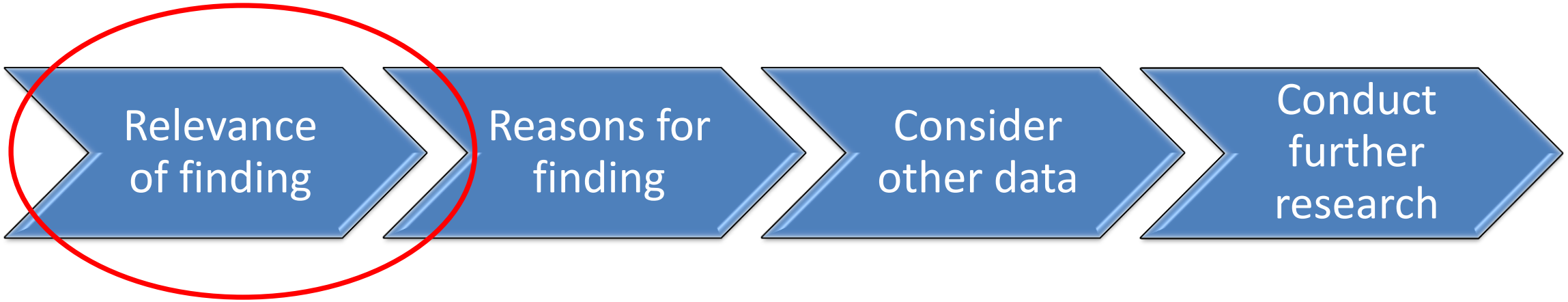
Graphing data – scatter diagram



Veri Yorumlama
(Interpreting data)

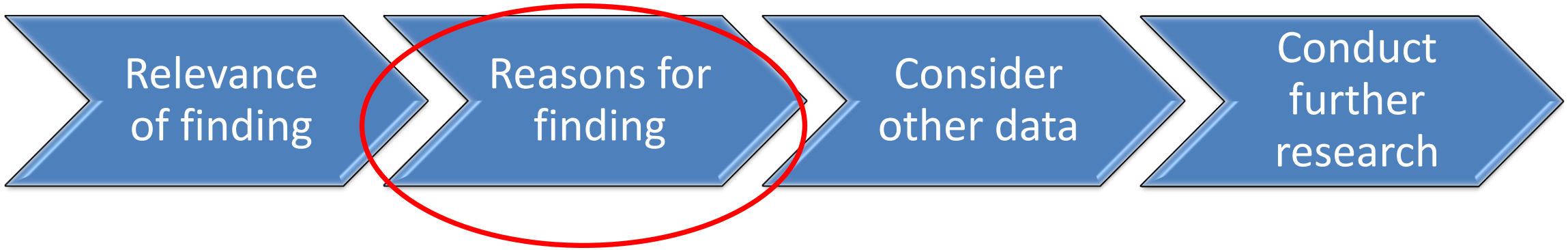
Yorum – Bulgunun uygunluğu

- Bağlantılar ve karşılaştırmalar yaparak ve neden ve sonuçları keşfederek bilgiye anlam katmak
- Gösterge hedefi karşılıyor mu?
- Hedeften ne kadar uzakta?(Diğer zaman dilimleri, diğer tesislerle) nasıl karşılaştırılır?
- Verilerde aşırı inişler ve çıkışlar var mı?



Yorum – Olası nedenler?

- Uzman görüşüne ek
- Program veya hedef kitle hakkında bilgisi olan diğerleri



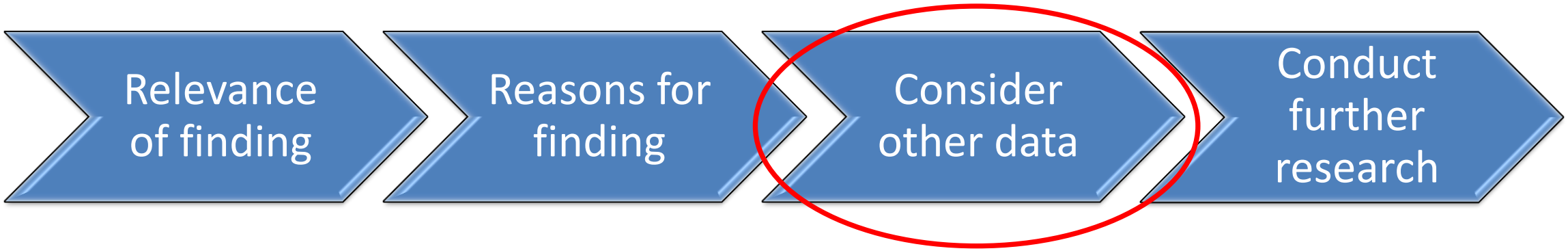
Yorumlama – Diğer verilerin dikkate alınması

Soruları netleştirmek için rutin servis verilerini kullanılır.

Oranlar hesaplanır.

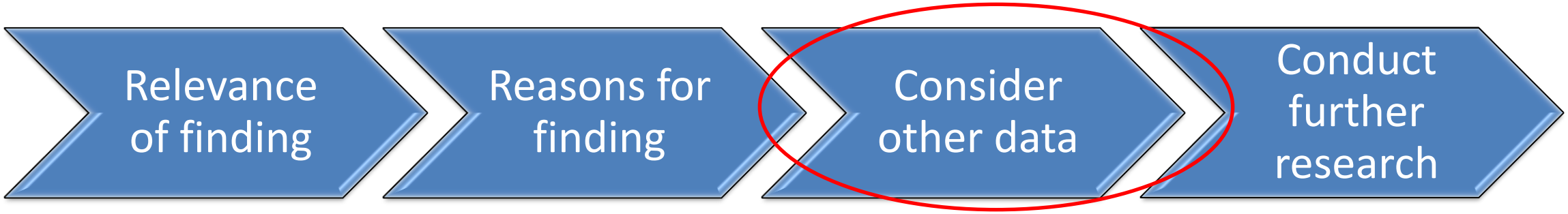
Verilerini yüküne göre gözden geçirilir, vb.

Diğer veri kaynakları kullanılır.



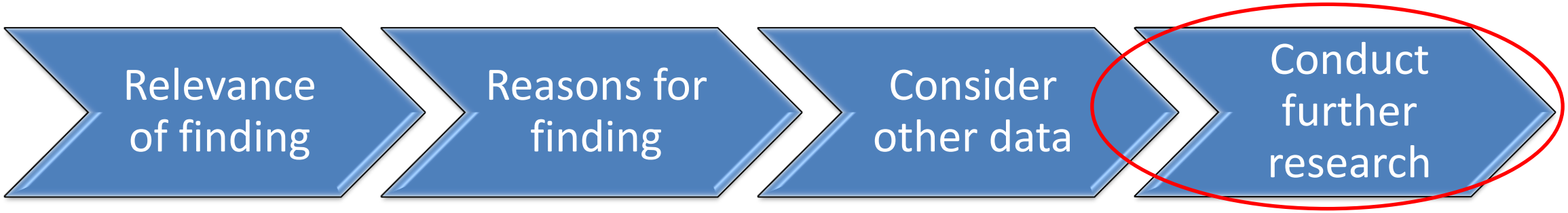
Yorumlama – Diğer veri kaynakları

- Durum analizleri
- Anketler
- Performans iyileştirme verileri



Yorumlama – Daha fazla araştırma

- Veri boşluğunu doldurmak amacıyla daha fazla araştırma yapılır.
- Metodoloji, sorulan sorulara ve mevcut kaynaklara bağlıdır



Sonuç

İstatistiklerle Nasıl Yalan Söylenir?

- Genellikle bir kutbun veya çubuk grafiğin dikey ekseni sıfırdan değil, daha yüksek bir değerden başlar. Böyle bir durumda, tasvir edilen değerlerin oranına ilişkin aktarılan izlenim tamamen yanlıştır. Bu efekt, ciro, hız vb artışlarla övünmek için kullanılır.

- Bir direğin, çubuğun veya çizgi grafiğin sıfır çizgisinin konumuna bağlı olarak tamamen farklı izlenimler aktarılabilir.

- Diyagramı estetik açıdan daha çekici kılmak için direklerin ve çubukların yerini sıklıkla nesneler (eskizleri) alır. Bununla birlikte, nesneler 2 veya hatta 3 boyutlu olarak algılanır ve bu nedenle sayısal oranların tamamen farklı bir izlenimini verir.
- Diyagramda, varillerin alanları sayısal değeri temsil eder. Bununla birlikte, veriler 3 boyutlu çizildiğinde, sayısal oranların yanlış bir izlenimi aktarılır.

Usage Notes

- A lot of slides are adopted from the presentations and documents published on internet by experts who know the subject very well.
- I would like to thank who prepared slides and documents.
- Also, these slides are made publicly available on the web for anyone to use
- If you choose to use them, I ask that you alert me of any mistakes which were made and allow me the option of incorporating such changes (with an acknowledgment) in my set of slides.

Sincerely,

Dr. Cahit Karakuş

cahitkarakus@gmail.com