İstatistik Nedir

İstatistik, veri toplama ve analiz etme uygulama ve çalışmasıdır.

İki ana dala ayrılır

- **Descriptive/Summary İstatistik** : Eldeki veriyi açıklar vey özetler
- **Inferential İstatistik :** Çıkarımsal istatistik, temsil edilen popülasyon hakkında sonuçlar çıkarmak için örneklemi kullanmayı içerir

İstatistik ne yapabilir?

Pratikte soruları cevaplamak için

- Türkiye'de ortalama maaş nedir?
- Bir şirketin haftada kaç müşteri sorgusu alması muhtemeldir?

Toplum genelinde uygulamalar

- Otomobil veya uçak gibi daha güvenli ürünler geliştirmek
- Hükümetin nüfusunun ihtiyaçlarını anlamasına yardımcı olmak

COVID-19 aşıları gibi bilimsel atılımları doğrular

İstatistik hangi sorulara cevap verebilir?

- Birinin bir ürünü alma olasılığı ne kadardır? İnsanlar farklı bir ödeme sistemi kullanabilselerdi, satın alma olasılıkları daha mı yüksek olurdu?
- Otelinizde kaç kişi konaklayacak? Doluluk oranını nasıl optimize edersiniz?
- Nüfusun %95'ine uyabilmesi için kaç beden kot üretilmesi gerekiyor? Her bedenden aynı sayıda mı üretilmeli?
- Hangi reklam insanların bir ürünü satın almasını sağlamada daha etkilidir?(A/B Testi)

İstatistiğin Sınırlamaları

İstatistikler; geniş, açık sorular yerine, spesifik, ölçülebilir sorular gerektirir.

- Örneğin; istatistikler bize pop müziğin Türkiyede daha popüler olup olmadığını söyleyebilir.

Ancak, ilişkilerin neden var olduğunu bulmak için istatistikleri kullanamayız

- Örneğin; insanların neden farklı müzik türlerini sevdiğini veya kadınların neden erkeklerden fazla yaşadığını

İstatistik hangi sorulara cevap veremez?

- Game of Thrones dizisi neden bu kadar popüler?: Herkese neden beğendikleri sorulabilir. Fakat yalan söyleyebilirler veya sebebini açıklamayabilirler
- Daha fazla şiddet sahnesi içeren filmlerin daha çok izleyici çekip çekmediği görülebilir. Fakat, Game of Thrones'daki şiddetin bunun sebebi olup olmadığına karar verilemez.

Veri Türleri

- Nümerik/Nicel
    1. Sürekli Veriler : Hisse Senedi Fiyatı, Günlük Rüzgar Hızı, Ürün kutusu ölçüleri ve ağırlığı

2. Ayrık Veriler : Ürün incelemelerinin sayısı, Bir günde satılan bilet sayısı, Sınıftaki öğrenci sayısı

Nümerik Verileri Görselleştirme

Sayısal veriler arasındaki ilişkiyi görselleştirmenin en yaygın yolu dağılım grafiği kullanmaktır.

Kategorik

1. Nominal Veriler : Göz Rengi gibi sıralanmamış kategorileri tanımlayan veriler.
2. Ordinal Veriler : Kategorilerin sıralandığı veriler. Likert ölçeği(Seviyorum, sevmiyorum..),

Kategorik Verileri Görselleştirme

Kategorik verileri ve bunların sayıları arasındaki ilişki görselleştirilebilir.

Descriptive/Summary İstatistik¶

Dört arkadaşa işe nasıl gittikleri sorulduğunda; %50'si işe arabayla, %25'i otobüs %25'i ise bisikletle gittiğini belirtsin. Bunlar betimsel istatistiklerdir.

Inferential İstatistik

Bir popülasyon hakkında sonuç çıkarmak için bir örneklem kullanılır.
Örneğin, 100 kişiye sosyal medya reklamlarını gördükten sonra kıyafet alıp almadıkları sorulabilir ve burdan elde edilen sonuç tüm insanların yüzde kaçının sosyal medya reklamı sonrası kıyafet aldığını anlamak için kullanılabilir.

Merkez Ölçümleri

Merkez ölçümleri neden faydalıdır?

- Bir işyerinde aylık ortalama sipariş sayısının ne olduğu sorulabilir.
- Bir evin tipik maliyeti öğrenilebilir
- En yaygın saç rengi öğrenilebilir

Ortalama, tipik, en yaygın gibi termonolijilerin tümü merkez ölçümlerinin günlük yaşamda nasıl ifade edildiğine dair örneklerdir.

Merkez Ölçümleri

Merkez ölçümleri neden faydalıdır?

- Bir işyerinde aylık ortalama sipariş sayısının ne olduğu sorulabilir.
- Bir evin tipik maliyeti öğrenilebilir
- En yaygın saç rengi öğrenilebilir

Ortalama, tipik, en yaygın gibi termonolijilerin tümü merkez ölçümlerinin günlük yaşamda nasıl ifade edildiğine dair örneklerdir.

Histogram sayısal verileri özetlemenin çok iyi bir yoludur ancak tanımlayıcı istatistikler de kullanılabilir.

Merkezi hesaplamanın üç yolu bulunmaktadır.

- Ortalama
- Medyan
- Mod

Ortalama

Verinin merkezini tanımlamanın en yaygın yollarından biridir.

Medyan

Verinin merkezini tanımlamanın bir diğer ölçüsü medyandır. Verilerin orta değeridir

Mod

Verideki en çok tekrar eden değerdir.

Bu aykırı değer ortalamayı kendisine doğru çekerken, medyan daha az etkilenir. Bunun nedeni, ortalama hesaplamanın tüm değerleri toplamayı gerektirmesidir. Daha büyük değerler sonucu etkiler, medyan ise sadece ortadaki değer bakar. Bu nedenler veriler simetrik olmadığında medyan kullanmak en iyi seçimdir.

Yayılım Ölçüleri

Özet istatistiğinin başka bir konusudur. Yayılım, veri noktalarının birbirinden ne kadar uzak olduğunu açıklar.

Hangi yayılım ölçüleri vardır

- **Range**
  range = maximum - minimum

Varyans

Varyans, her bir veri noktasının ortalamaya olan ortalama uzaklığını hesaplar. Varyans ne kadar büyükse veriler o kadar dağılmış demektir.

Standart Sapma
Varyansı anlamak zordur. Bu nedenle standart sapma kullanılır
Standart sapma varyansın karekökü alınarak hesaplanır.
Standart sapma sıfıra ne kadar yakınsa verilerin ortalama etrafında o kadar yakın kümlendiği anlaşılır.

Quartiles
Yayılım, verileri dört eşit parçaya bölmenin bir yolu olan çeyrekler kullanılarak da ölçülebilir.

İkinci çeyrek ortadaki değerdir ve medyana eşittir.
Boxplots
Bir boxplot kullanılarak çeyrekler görselleştirilebilir.

Interquartile Range (IQR)
Çeyrekler arası aralık olarak da tanımlanır. IQR = Q3 - Q1

Mean Absolute Deviation

- Standart sapmada uzaklıkların karesi alınır, bu nedenle daha uzun mesafeler daha uzun cezalandırılır.
- Mean Absolute Deviation'da, her uzaklık eşit ceza alır.
- Biri diğerinden daha iyi değildir ama SD MAD'den daha yaygındır.

Outliers (Aykırı Değerler)

Diğerlerinden önemli ölçüde farklı olan veri noktalarıdır. Peki öenmli bir fark olduğuna nasıl karar verilir?

Q1 - 1.5IQR < data < Q3 + 1.5IQR

Şans Nedir

İnsanlar genellikle şanstan bahsederler. Örneğin; bir satışın bitme, yarın yağmur yağma ve oyunu kazanma gibi.

Bir olayın sonucunun olma olasılığını tahmin edebilmek, birçok yönden faydalı olabilmektedir.

Peki, şansı nasıl ölçebiliriz

Bir olayın gerçekleşme olasılığı nedir?

P(Olay) = olayın gerçekleşebileceği yollar / Olası çıktıların toplam sayısı

Örneğin; bir yazı tura oyununda tura gelme olasılığı;

P(Tura) = 1 / 2 = 0.5

Olasılık her zaman 0 ile 100 arasındadır.

Bir başka örnek verirsek; Potansiyel bir müşteri ile yaklaşan bir toplantınız var ve siz de katılması için satış ekibinden birini göndermek istiyorsunuz. He rkişinin adını bir kutuya koyup, toplantıya kimin katılacağını belirlemek için rastgele birini seçeceksiniz.

Brian'ı seçme olasılığı 1/4 = 0.25'tir.

Peki, farklı zamanlarda iki toplantı yaparsak ne olur?

Her toplantı için dört ekip üyesinden birini rastgele seçebiliriz. İlk toplantı için seçilen kişinin, ikinci toplantı için seçilme şansını etkilenmez.

Örenğin ilk toplantı için Brain seçilmişse, Brain'ın öğleden sonraki toplantı için seçilme şansı yine %25'tir.

Örnek yerine geri yerleştirildiğinden ve tekrar seçilebildiğinden buna **değiştirmeli örnekleme (sampling with replacement)** denir.

Bu bağımsız olasılığın bir örneğidir.

Bağımsız Olasılık

İkinci olayın olasılığı ilk olayın sonucuna bağlı olarak değişmiyorsa iki olay bağımsızdır.

Koşullu Olasılık (Conditional Probability)

Bir olayın sonucunun başka bir olayı etkilemesi durumudur.

Daha önce çıkardığımız kişiyi tekrar yerine koymadığımız için buna değiştirmeden örnekleme (sampling without replacement) denir.

Bağımlı Olasılık

İlk olayın sonucunun ikinci olayın sonucunu etkilediği durumdur.

Örneğim ilk seçimde Claire çekilseydi, ikinci çekilişte Claire'in seçilme olasılığı %0'dır. Eğer başka biri ilk seçilirse Claire'in ikinci seçilme olasılığı %33.3'tür.

Koşullu olasılık, bağımlı olayların olasılığını hesaplamak için kullanılır.

- Bir olayın olasılığı diğerinin sonucuna bağlıdır. Örneğin, bir önceki tren göz önüne alındığında, bir trenin zamanında varma olasılığı.

Ayrık Dağılımlar

Standart altı yüzlü bir zarın atıldığını düşünelim. Alto olası sonuç vardır ve her birinin gerçekleşme şansı altıda birdir.

Bu daha önceki senaryoya benzemektedir ve burada sayıların yerini isimler almaktadır. Tıpkı zarın atılması gibi, her sonucun veya ismin seçilme şansı eşittir.ir

Olasılık Dağılımı

Bir senaryodaki her olası sonucun olasılığını açıklar.

Bir dağılımın ortalaması olan beklenen değeri de bulunabilir.

Bu, her değeri olasılığıyla (bu durumda altıda biri) çarpıp toplayarak hesaplanır.

Olasılık Dağılımları Neden Önemlidir?

- Riski ölçmeye ve karar alma sürecini bilgilendirmeyi sağlar.
- Hipotez testlerinde sonuçların şans eseri çıkıp çıkmadığını anlamak için

Olasılık Dağılımlarının Görselleştirilmesi

Histogram kullanarak olasılık dağılımları görselleştirilebilir. Burada her çubuk bir sonucu temsil eder ve her çubuğun yüksekliği bu sonucun olasılığını temsil eder.

Olasılık = Alan

Olasılık dağılımının alanlarını bularak farklı sonuçların olasılıkları hesaplanabilir.

Sürekli Dağılımlar

Şu durumlarda sürekli dağılımlar kullanılır

Belediye otobüsü her 12 dakikada bir geliyor. yani rastgele bir saatte gelirseniz bir süre bekleyebilirsiniz. Otobüs tam durağa gelirken gelinirse sıfır dakikadan, otobüs kalkarken gelinirse 12 dakikaya kadar beklenebilir.

Bu olay bir olasılık dağılımı ile modellenebilir. Durakta, beklenebilecek sonsuz sayıda dakika vardır (5 dk., 1.5 dk, 1.53 dk. vb.). Bu nedenle sayım ve aralık verileriyle yapıldığı gibi bireysel bloklar oluşturulamaz.

Bunun yerine olasılığı temsil etmek için sürekli bir çizgi kullanılır. 0'dan 12'ye kadar herhangi bir süre bekleme olasılığı olduğu aynı olduğundan çizgi düzdür. Buna sürekli düzgün dağılım adı verilir (Continuous Uniform Disttibutions).

Bimodal Distributions

Sürekli dağılımlar, bazı değerlerin diğerlerinden daha yüksek olasılığa sahip olduğu tekdüze olmayan biçimler alabilir.

Normal Dağılım

Binomial (Binom) Distribution

Binom dağılımı, bir dizi bağımsız denemedeki başarı sayısının olasılığını tanımlar.

Aynı madeni parayı birden fazla kez havaya atıp sonuçları kaydedebiliriz, örneğin burada yazı gelirse 1, tura gelirse 0 olarak gösterilir.

- Binom dağılımı, bir dizi bağımsız olaydaki başarı sayısının olasılığını tanımlar. Örneğin, bir dizi yazı-tura atışında belirli sayıda tura gelme olasılığını söyleyebilir.
- Sayılabilir bir sonuçla çalışıldığı için, ayrık bir dağılımdır.
- Binom dağılımındaki iki parametre n ve p olarak tanımlanabilir.
  n : gerçekleştirilen etkinlik sayısı
  p : başarı olasılığı

Expected Value = n * p

scipy.stats.pmf(k, n, p, loc=0)

olasılık kütle fonksiyonu (PMF) hesaplamak için kullanılır. PMF, bir dağılımın belirli bir olasılıkta kesikli bir değer alma ihtimalini hesaplar.

Normal Distribution¶

Sürekli bir olasılık dağılımıdır. Diğer olasılık dağılımlarından daha fazla gerçek dünya durumu için geçerlidir.

Çan eğrisi şeklindedir.

- Simetriktir. Dolayısıyla sol taraf sağ tarafın ayna görüntüsüdür.
- Herhangi bir olasılık dağılımı gibi, eğrinin altındaki alan 1'e eşittir.
- Uçların öyle görünse bile olasılık hiçbir zaman sıfıra ulaşmaz.

Normal Dağılım Neden Önemlidir

- Birçok gerçek dünya verisi normal dağılıma çok benzemektedir.
- Hipotez testinde, bir örneğin ortalamasını temsil ettiği popülasyonla karşılaştırmak gibi birçok istatistiksel testi gerçekleştirmek için verilerin normal bir dağılıma uyması gerekir.
- Uçların öyle görünse bile olasılık hiçbir zaman sıfıra ulaşmaz.

norm.ppf kullanılarak da yüzde hesaplanabilir

scipy.stats.norm.ppf(q, loc=0, scale=1) fonksiyonu, normal dağılımın ters kümülatif dağılım fonksiyonunu (percent-point function, PPF) hesaplar. PPF, kümülatif dağılım fonksiyonunun (CDF) tersidir. Yani, verilen bir olasılığa (q) karşılık gelen kritik değeri (x) hesaplar. Başka bir deyişle, belirli bir olasılığa karşılık gelen kesim noktasını verir.

Skewness (Çarpıklık)

Veri dağılımını yorumlarken verilerin sona erdiği yönü tanımlayan çarpıklık terimi kullanılır.

kuyruk daha büyük pozitif değerlerin olduğu yerde sağda olduğundan dağılım pozitif çarpık veya sağa çarpıktır. Tersine negatif çarpık veya sola çarpık dağılım sağda zirve yapar ve sola doğru sona erer.

Kurtosis (Basıklık)

Bir dağılım basıklığıyla da yorumlanabilir. Kurtosis, dağılımdaki aşırı değerlerin oluşumunu açıklamanın bir yoludur.

Pozitif Basıklık (Leptokurtic) : Grafikte kırmızıyla gösterilmiştir. Ortalama etrafında, büyük bir tepe noktası ve daha küçük standart sapma ile karakterize edilir.

- Mesokurtic Basıklık, çizimde mavi olarak gösterilen normal dağılımdır.
- Negatif Basıklık (Platykurtic) : Çizimde yeşil renkle gösterilmiştir. Daha düşük zirveye ve daha büyük standart sapmaya sahip bir dağılımdır.

The Central Limit Theorem¶

Ortalama gibi bir özet istatistiğinin dağılımına örneklem dağılımı denir.

Bu dağılım, özellikle, örnek ortalamasının bir örnekleme dağılımıdır.

Bir istatistiğin örnekleme dağılım, örneklem büyüklüğü arttıkça normal dağılıma yaklaşır.

Merkezi limit teoreminin yalnızca örnekler rastgele alındığında ve bağımsız olduğunda geçerli olduğunu belirtmek önemlidir.

Genel olarak merkezi limit teoreminin uygulanabilmesi için örneklem büyüklüğünün en az 30 olması önemlidir.

MLT diğer özet istatistikler için de geçerlidir.

Poisson Distribution¶

Poisson süreci, belirli bir zaman dilimindeki ortalama olay sayısının bilindiği, ancak olaylar arasındaki zaman veya boşluğun rastgele olduğu bir süreçtir.

Poisson süreçleri günlük hayatta çok yaygındır.

- Bir hayvan barınağından her hafta sahiplenilen hayvan sayısı poisson sürecidir.
- Saat başına bir restorana gelen kişi sayısı.
- Günlük web sitesi ziyaret sayısı.

Poisson dağılımı, belirli bir zaman diliminde bazı olayların meydaha gelme olasılığını açıklar.

- Haftada en az beş hayvanın bir hayvan barınağından sahiplenilme olasılığı.
- Bir restorana saatte 12 kişinin gelme olasılığı.
- Bir web sitesinin bir günde 200'den az ziyaret edilme olasılığı.

Poisson dağılım Lambda(λ) ile tanımlanır.

λ = zaman periyodu başına ortalama olay sayısı

- Restoran örneğinde bu değer, saat başına ortalama müşteri sayısı olan 20'dir.
- Bu değer aynı zamanda dağılımın beklenen değeridir.

Olaylar sayıldığı için kesikli bir dağılımdır ve 20 bir saat içinde ziyaret etmesi en muhtemel müşteri sayısıdır.

Lambda dağılımın şeklini değiştirir.

Tıpkı diğer dağılımlarda olduğu gibi, çok sayıda örnek varsa ve her birinin ortalaması hesaplanırsa, Poisson Dağılımı olarak örnek ortalamalarının dağılımı normal dağılıma benzer.

Diğer Olasılık Dağılımları

Exponential Distribution (Üstel Dağılım)

Poisson olayları arasında belirli bir zaman geçme olasılığını temsil eden dağılımdır.

- Evlat edinmeler arasında 1 günden fazla zaman geçme olasılığı
- Restorana gelişler arasında 10 dakikadan az zaman geçme olasılığı
- Depremler arasında 6-8 ay geçme olasılığı

Üstel dağılım, Poisson dağılımında olduğu gibi oranı temsil eden aynı lambda değerini kullanır. Bu bağlamda lambda ve oranın aynı değer anlamına gelir.
Ayrıca Poisson dağılımının aksine, zamanı temsil ettiği için süreklidir.

Üstel Dağılımın Beklenen Değeri

- Lambda, sıklığı oran veya olay sayısı cinsinden ölçen Poisson dağılımının beklenen değeridir.
- Üstel dağılım, sıklığı olaylar arasındaki süre açısından ölçer. Üstel dağılımın beklenen değeri, 1 / λ ile hesaplanabilir.

Student's) t-distribution

t-dağılım, küçük örneklem büyüklükleriyle uğraşırken veya popülasyon standart sapması bilinmediğinde istatistikte yaygın olarak kullanılır.

Şekli normal dağılıma benzer, ancak tam olarak aynı değildir.

- Mavi ile gösterilen normal dağılım ile turuncu ile gösterilen bir serbestlik dereceli t-dağılımını karşılaştırırsak, t-dağılımının kuyruklarının daha kalın olduğunu görürüz.
- Bu, t-dağılımında gözlemlerin ortalamadan daha uzak düşme olasılığının daha yüksek olduğu anlamına gelir.

Degrees of freedom (Serbestlik Derecesi)

Serbestlik derecesi (df), istatistikte temel bir kavramdır ve istatistiksel bir parametreyi tahmin etmek için kullanılabilen bağımsız değer veya bilgi parçalarının sayısını ifade eder. Daha basit bir ifadeyle, bir hesaplamada veriler tarafından empoze edilen herhangi bir kısıtlamayı ihlal etmeden serbestçe değişebilen değer sayısını temsil eder.

Serbestlik dereceleri t-dağılımının kuyruklarının ne kadar geniş ve düz olacağını etkiler.
Daha düşük serbestlik dereceleri (küçük df): Dağılımın daha ağır (daha geniş) kuyrukları vardır ve daha yayılmıştır. Bu, normal dağılıma kıyasla daha yüksek uç değerler (aykırı değerler) olasılığı olduğu anlamına gelir. Örneklem boyutları küçük olduğunda daha fazla belirsizliğe neden olur.
Daha yüksek serbestlik dereceleri (büyük df): Serbestlik dereceleri arttıkça t-dağılım normal dağılıma (çan eğrisi) yaklaşır. Serbestlik dereceleri yaklaşık 30 veya daha fazlasına ulaştığında, t-dağılım normal dağılımdan neredeyse ayırt edilemez hale gelir.

Çoğu istatistiksel testte, t-dağılımındaki serbestlik dereceleri genellikle örneklem büyüklüğüne bağlıdır. Örneğin, tek örneklemli bir t-testinde, serbestlik dereceleri n−1'dir, burada n örneklem büyüklüğüdür.

Küçük örneklem büyüklükleri (küçük df): Küçük bir örneklem büyüklüğünüz olduğunda, t-dağılımının kuyrukları daha ağırdır çünkü popülasyon ortalamasını tahmin etmede daha fazla belirsizlik vardır.
Büyük örneklem büyüklükleri (büyük df): Daha büyük örneklerde, t-dağılımının normal dağılıma yaklaşması, popülasyon ortalamasının tahmininin daha kesin hale gelmesidir.

Serbestlik dereceleri hipotez testindeki kritik değerleri de etkiler.
Daha düşük df (daha küçük örneklem boyutu): Güven aralıkları veya önem testleri (t-testleri gibi) için kritik değerler daha büyük olacaktır. Bunun nedeni, daha az veri noktasıyla, ek belirsizliği hesaba katmak için daha geniş bir hata payına ihtiyaç duymanızdır.
Daha yüksek df (daha büyük örneklem boyutu): Örneklem boyutu arttıkça kritik değerler küçülür, bu da güven aralıklarının daha dar hale geldiği ve hipotez testlerinin sıfır hipotezini reddetmek için daha az uç değer gerektirdiği anlamına gelir. Bu, daha büyük örneklerin daha büyük kesinliğini yansıtır.

Log-normal distribution

- Log-normal dağılımı izleyen değişkenlerin logaritması normal dağılım gösterir.
- Bu da normal dağılımdan farklı olarak çarpık dağılımlara neden olur.
- Satranç oyunlarının uzunluğu, yetişkinlerde kan basıncı ve 2003 SARS salgınında hastaneye yatış sayısı gibi bu dağılımı izleyen çok sayıda gerçek dünya örneği vardır.

Hypothesis Testing

Hipotez testi, popülasyonları karşılaştırmak için kullanılan bir grup teori, yöntem ve tekniktir.

Neden hipotez testleri hakkında bilgi sahibi olmak gerekir?

- İlk olarak, birçok sektörde rutin olarak kullanılmaktadır. Örneğin, bir şirketin ürün fiyatını artırmanın geliri artıracağı veya bir web sitesinin adını değiştirmenin trafiği artırabileceği yönünde bir teorisi olabilir. Bir ilacın belirli sağlık koşullarının tedavisinde etkili olup olmadığını analiz etmek için hipotez testi kullanılabilir.

- Hipotez testlerinde, her zaman popülasyonlar arasında hiçbir fark olmadığı varsayımıyla başlanır. Bu, teste herhangi bir önyargı ekleme riskini azaltmak için yapılır.
- Buna sıfır(null) hipotezi denir.

Bir örnek vermek gerekirse;

Sıfır hipotezi, C vitamini takviyesi alan ve almayan kadınlar arasında cinsiyete göre doğum oranında bir fark olmadığı şeklindedir.

Daha sonra alternatif bir hipotez oluşturulur ve bu hipotez tipik olarak iki şekilde olabilir. C vitamini takviyesi alan ve almayan kadınlar arasında erkek ve kadın doğumları arasında bir fark olduğu söylenebilir. Ya da farkın yönü, örneğin C vitamini takviyesi alan nüfusun takviye almayanlara göre daha fazla kadın doğumuna sahip olduğunu belirtilebilir.

Hipotez Testi İş Akışı

- Popülasyon tanımlanmalıdır.
  Aralarındaki farkı analiz etmek istediğimiz popülasyona karar verirlir. Bu durumda C vitamini takviyesi kullanan veya kullanmayan yetişkin kadınlar popülasyondur.

- Null ve Alternatif hipotez belirlenir.
  Ardından, her iki popülasyonda da doğumların erkek veya kız olma olasılığının eşit olduğu veya C vitamini takviyesi alan kadınlarda bebeklerin kız olma olasılığının daha yüksek olduğu şeklinde boş ve alternatif hipotezler geliştirilir.
- Örnek veriler toplanır veya bunlara erişim sağlanır.
- Veriler üzerinde istatistiksel testler gerçekleştirilir
- Sonuçlar örneklemin temsil ettiği nüfus hakkında sonuçlar çıkarmak için kullanılır.

Ne kadar veriye ihtiyaç var?

Peki kaç doğumun cinsiyeti kaydedilmelidir?
Merkezi limit teoremi uygulanırsa, örneklem büyüklüğü arttıkça erkek ve kadın doğumlarının ortalama sayısı popülasyon ortalamalarına yaklaşır. Ancak, büyük örnekler toplamak çok fazla zaman ve kaynak gerektirebilir!
Yaygın bir yaklaşım, örneklemlerin ne kadar büyük olduğunu bulmak için benzer hipotez testleri üzerine hakemli araştırmalara bakmaktır. Bu daha sonra bir ölçüt olarak kullanılabilir.

Independent and dependent variables

- Bağımsız değişken, diğer verilerden etkilenmesi beklenen veriyi tanımlar. C Vitamini Takviyesi
- Bağımlı Değişken, Bağımsız değişken veya değişkenler tarafından etkilenen değişken. Doğum cinsiyet oranı

Design of Experiments

- Veriler genellikle belirli bir soruyu yanıtlamayı amaçlayan bir çalışmanın sonucu olarak oluşturulur. Ancak, verilerin nasıl oluşturulduğuna ve çalışmanın nasıl tasarlandığına bağlı olarak verilerin farklı şekilde analiz edilmesi ve yorumlanması gerekir.
- Deneyler genellikle "Tedavinin tepki üzerindeki etkisi nedir?" biçimindeki bir soruyu yanıtlamayı amaçlar. Bu ortamda, tedavi açıklayıcı veya bağımsız değişkeni ifade eder ve tepki tepkiyi veya bağımlı değişkeni ifade eder. Örneğin, bir reklamın satın alınan ürün sayısı üzerindeki etkisi nedir? Bu durumda, tedavi bir reklamdır ve tepki satın alınan ürün sayısıdır.

Controlled experiments

- Kontrollü bir deneyde, katılımcılar rastgele tedavi grubuna veya kontrol grubuna atanır. Tedavi grubu tedaviyi alır ve kontrol grubu almaz.
- Bununen iyi örneklerinden biri A/B testidir. Örneğin, tedavi grubu bir reklam görecek, kontrol grubu ise görmeyecektir. Bu farkın dışında, grupların karşılaştırılabilir olması gerekir; böylece bir reklam görmenin insanların daha fazla satın almasına neden olup olmadığını belirlenebilir.
- Gruplar karşılaştırılabilir değilse, bu durum kafa karıştırıcılığa veya önyargıya yol açabilir. Tedavi grubundaki katılımcıların ortalama yaşı 25 ve kontrol grubundaki katılımcıların ortalama yaşı 50 ise, daha genç kişilerin daha fazla satın alma olasılığı daha yüksekse yaş potansiyel bir karıştırıcı olabilir ve bu da deneyi tedaviye doğru önyargılı hale getirecektir.

Deneylerin Altın Standardı

- Altın standart veya ideal deney, belirli araçları kullanarak mümkün olduğunca fazla önyargıyı ortadan kaldıracaktır.
- Kontrollü deneylerde önyargıyı ortadan kaldırmaya yardımcı olan ilk araç, randomize kontrollü bir deneme kullanmaktır. Randomize kontrollü bir denemede, katılımcılar tedavi veya kontrol grubuna rastgele atanır ve atamaları şanstan başka bir şeye dayanmaz. Bu tür rastgele atama, grupların karşılaştırılabilir olduğundan emin olmaya yardımcı olur.
- İkinci yol, tedaviye benzeyen ancak hiçbir etkisi olmayan bir şey olan plasebo kullanmaktır. Bu şekilde, katılımcılar tedavi veya kontrol grubunda olup olmadıklarını bilmezler. Bu, tedavinin etkisinin tedavinin kendisinden kaynaklandığını, tedaviyi alma fikrinden kaynaklanmadığını garanti eder. Bu, bir ilacın etkinliğini test eden klinik çalışmalarda yaygındır. Kontrol grubuna yine bir hap verilir, ancak bu, yanıt üzerinde minimum etkisi olan bir şeker hapıdır.
- Çift kör bir deneyde, tedaviyi uygulayan veya deneyi yürüten kişi, gerçek tedaviyi mi yoksa plaseboyu mu uyguladığını bilmez. Bu, yanıttaki önyargıya ve sonuçların analizine karşı koruma sağlar. Bu farklı araçların hepsi aynı prensibe dayanır: deneyinize önyargının sızması için daha az fırsat varsa, tedavinin yanıtı etkileyip etkilemediğine dair daha güvenilir bir sonuca varabilirsiniz.

Observational studies

- Gözlemsel bir çalışmada, katılımcılar gruplara rastgele atanmazlar. Bunun yerine, katılımcılar genellikle önceden var olan özelliklere göre kendilerini atarlar. Bu, kontrollü bir deney için elverişli olmayan soruları yanıtlamak için yararlıdır.
- Sigara içmenin kanser üzerindeki etkisini incelemek istiyorsanız, insanları sigara içmeye zorlayamazsınız. Benzer şekilde, geçmiş satın alma davranışının birinin bir ürünü satın alıp almayacağını nasıl etkilediğini incelemek istiyorsanız, insanları belirli geçmiş satın alma davranışlarına sahip olmaya zorlayamazsınız.
- Atama rastgele olmadığından, grupların her açıdan karşılaştırılabilir olacağını garantilemenin bir yolu yoktur, bu nedenle gözlemsel çalışmalar nedensellik kuramaz, yalnızca ilişki kurabilir. Tedavinin etkileri, belirli kişileri kontrol grubuna ve belirli kişileri tedavi grubuna sokan faktörler tarafından karıştırılabilir. Ancak, ilişki hakkındaki sonuçların güvenilirliğini güçlendirmeye yardımcı olabilecek karıştırıcıları kontrol etmenin yolları vardır.

Longitudinal vs. cross-sectional studies (Uzunlamasına ve kesitsel çalışmalar)

- Uzunlamasına bir çalışmada, aynı katılımcılar, tedavinin yanıt üzerindeki etkisini incelemek için bir süre boyunca takip edilir.
- Kesitsel bir çalışmada, veriler zaman içinde tek bir anlık görüntüden toplanır.
- Yaşın boy üzerindeki etkisini araştırmak isterseniz, kesitsel bir çalışma farklı yaşlardaki insanların boylarını ölçer ve bunları karşılaştırır. Ancak, sonuçlar doğum yılı ve yaşam tarzı tarafından karıştırılacaktır çünkü her neslin daha uzun olması mümkündür.
- Uzunlamasına bir çalışmada, aynı kişilerin boyları hayatlarının farklı noktalarında kaydedilir, böylece karıştırıcı faktör ortadan kalkar.
- Uzunlamasına çalışmaların daha pahalı olduğunu ve gerçekleştirilmesinin daha uzun sürdüğünü, kesitsel çalışmaların ise daha ucuz, daha hızlı ve daha rahat olduğunu belirtmek önemlidir.

Experiments

Deneyler, bir popülasyon hakkında sonuçlar çıkarmak için örnek veriler üzerinde istatistiksel testler yapmayı içeren hipotez testinin bir alt kümesidir.

Bu sadece akademi ve araştırma için geçerli değildir; deneyler, özellikle ürün içgörüleri elde etmek ve ticari performansta iyileştirmeler sağlamak için endüstride de gerçekleştirilir.

Deneyler genellikle "Uygulamanın yanıt üzerindeki etkisi nedir?" şeklinde bir soruyu yanıtlamayı amaçlar; burada uygulama bağımsız değişkeni, yanıt ise bağımlı değişkeni ifade eder.

Advertising as a treatment

Bir deney örneği olarak, bir reklamın satın alınan ürün sayısı üzerinde ne gibi bir etkisi olduğu bilinmek istenebilir.

Bu durumda, uygulama bir reklam, yanıt ise satın alınan ürün sayısıdır.

Controlled experiments

Yaygın bir deney türü, katılımcıların rastgele bir şekilde uygulama grubuna ya da kontrol grubuna atandığı kontrollü bir deneydir.

Örnekte, uygulama grubu bir reklam görecek, kontrol grubu ise görmeyecektir. Bu fark dışında, gruplar karşılaştırılabilir olmalıdır, böylece bir reklam görmenin insanların daha fazla satın almasına neden olup olmadığı belirlenebilir.

Gruplar karşılaştırılabilir değilse, sonuçlara dayanarak yanlış sonuçlar çıkarılabilir. Uygulama grubundaki katılımcıların yaş ortalaması 25 ve kontrol grubundaki katılımcıların yaş ortalaması 50 ise, yaş potansiyel olarak sonuçları etkileyebilir; genç insanların daha fazla satın alma olasılığı daha yüksektir ve bu da deneyi uygulama lehine taraflı hale getirir.

The gold standard of experiments

Kontrollü deneylerde önyargıyı ortadan kaldırmaya yardımcı olan ilk yöntem rastgeleleştirmedir (Randomization).

- Katılımcılar Uygulama/Kontrol grubuna belirli özelliklerine göre değil rastgele atanır.
- Rastgeleleştirme grupların karşılaştırılabilir olmasını sağlamaya yardımcı olur.
- Buna randomize kontrollü çalışma adı verilir.

İkinci yöntem körleme(Blinding) kullanmaktır.

- Katılımcılar hangi grupta olduklarını bilmezler. Bu tedavinin etkisinin tedavi olam fikriden değil, tedavinin kendisinden kaynaklanmasını sağlar.
- Bu tedaviye benzeyen fakat hiçbir etkisi olmayan bir plasebo kullanımını içerebilir.
- Bu bir ilacın etkinliğini test eden klinik deneylerde yaygındır.

Üçüncü yöntem Çift kör randomize kontrollü bir çalışma (double-blind randomized controlled trial) kullanmaktır.

- Tedaviyi uygulayan veya deneyi yürüten kişi gerçek tedaviyi mi yoksa plaseboyu mu uyguladığını da bilmez.
- Bu, sonuçların analizinin yanı sıra yanıtta da önyargıya karşı koruma sağlar.

- Bu farklı araçların hepsi aynı ilkeye dayanır: Deneye önyargı girmesi için ne kadar az fırsat olursa, tedavinin yanıtı etkileyip etkilemediği sonucuna o kadar güvenilir bir şekilde varılabilir.

Randomized Controlled Trials vs. A/B testing

Amaç, bir ilacın farklı dozajları gibi birden fazla tedavi arasındaki farkı test etmekse, randomize kontrollü çalışmalar birden fazla tedavi grubuna sahip olabilir. Bunlar akademide, özellikle de bilimsel ve klinik araştırmalarda popülerdir.

Randomize kontrollü denemeler, genellikle pazarlama ve mühendislik gibi sektörlerde kullanıldığında A/B testi olarak da adlandırılır. Aradaki fark, A/B testinin katılımcıları yalnızca uygulama ve kontrol olmak üzere iki gruba ayırmasıdır.

Correlation

Değişkenler arası ilişkiler daha önceki konularda anlatıldı. Şimdi bu ilişkiyi ölçmenin bir yolu olan Korelasyon konusundan bahsedilecektir.

Relationships between two variables

İki değişken arasındaki ilişkiyi görselleştirebilmek için dağılım grafiği kullanılır.

Pearson correlation coefficient

İşte bu noktada, genellikle korelasyon katsayısı olarak adlandırılan Pearson korelasyon katsayısı işe yarar.

Karl Pearson tarafından geliştirilmiş ve 1896 yılında yayınlanmıştır.

İki değişken arasındaki ilişkinin gücünü ölçer ve eksi bir ile bir arasında bir değer üretir. Bu sayı, değişkenler arasındaki ilişkinin gücüne karşılık gelir ve pozitif veya negatif işaret, ilişkinin yönüne karşılık gelir.

Linear relationships

Pearson korelasyon katsayısı yalnızca doğrusal ilişkiler için kullanılabilir, yani değişkenler arasındaki değişiklikler orantılıdır.

Korelasyon nedenselliğe eşit değildir. Bu, su maliyetinin artmasının ortalama yaşam süresini artıracağı anlamına mı geliyor? Bir ilişkinin var olmasının, su maliyetlerindeki değişikliklerin yaşam beklentisinde bir değişikliğe yol açacağı anlamına gelmediğini ayırt etmek önemlidir.

Confounding variables

Veriler arasındaki ilişkilere bakarken, değerleri başka nelerin etkiliyor olabileceğini sormak önemlidir. Bir şişe suyun maliyeti, daha güçlü ekonomilere sahip yerlerde genellikle daha yüksektir ve bu yerler yüksek kaliteli sağlık hizmetlerine daha iyi erişim sunabilir. Dolayısıyla, belki de yaşam beklentisi bir şişe suyun maliyetinden etkilenmiyor, aslında ekonominin gücünden etkileniyordur. Bu, analiz ettiğimiz verileri etkileyen ancak değişkenler arasındaki ilişkiyi değerlendirirken hesaba katılmayan bir şey olan karıştırıcı değişken olarak bilinir.

Korelasyon Uyarıları (Correlation caveats)

Korelasyon, ilişkileri ölçmek için yararlı bir yol olsa da, bazı uyarıları dikkate almak gereklidir.

Non-linear relationships

- Yukarıdaki grafiği gözönünde bulunduralım. X ve Y arasında açıkça bir ilişki var, ancak korelasyonu hesapladığında 0.18 elde edilir. Bunun nedeni, iki değişken arasındaki ilişkinin doğrusal bir ilişki değil, ikinci dereceden bir ilişki olmasıdır. Korelasyon katsayısı sadece doğrusal ilişkilerin gücünü ölçer.
- Tüm özet istatistiklerde olduğu gibi korelasyon da körü körüne kullanılmamalı ve mümkün olduğunca veriler görselleştirilmelidir.

Other Transformations

Log dönüşümüne ek olarak, bir ilişkiyi daha doğrusal hale getirmek için bir değişkenin karekökünü veya tersini almak gibi kullanılabilecek birçok başka dönüşüm vardır.

- Log Transformation (log(x))
- Square Root Transformation (sqrt(x))
- Reciprocal Transformation (1 / x)
- Bunların Kombinasyonları
  log(x) ve log(y)
  sqrt(x) ve 1 / y

Dönüşümün seçimi veriye ve ne kadar çarpık olduğuna bağlı olacaktır.

Transformation Neden Kullanılır?

Bazı istatistiksel yöntemler, korelasyon katsayısının hesaplanması gibi değişkenlerin doğrusal bir ilişkiye sahip olmasına dayanır.

Doğrusal regresyon, değişkenlerin doğrusal bir şekilde ilişkili olmasını gerektiren başka bir istatistiksel tekniktir.

Korelasyon nedensellik anlamına gelmez

Korelasyon nedensellik anlamına gelmez. Örneğin, burada ABD'de her yıl kişi başına düşen margarin tüketimi ile Maine eyaletindeki boşanma oranını gösteren bir dağılım grafiği yer almaktadır. Bu iki değişken arasındaki korelasyon 0.99'dur, yani neredeyse mükemmeldir. Ancak bu, daha fazla margarin tüketmenin daha fazla boşanmaya neden olacağı anlamına gelmez. Bu tür bir korelasyon genellikle sahte korelasyon olarak adlandırılır.

Confounding (Karıştırma)

Karıştırma adı verilen bir olgu sahte korelasyonlara yol açabilir. Diyelim ki kahve içmenin akciğer kanserine neden olup olmadığını bilmek istiyoruz. Verilere baktığımızda, kahve içme ve akciğer kanseri arasında korelasyon olduğunu görürüz; bu da bizi daha fazla kahve içmenin akciğer kanserine yol açacağını düşünmeye sevk edebilir.

Ancak, üçüncü ve gizli bir değişken daha vardır ki o da sigaradır.

Sigara içmenin kahve tüketimi ile ilişkili olduğu bilinmektedir.

Sigaranın akciğer kanserine neden olduğu da bilinmektedir.

Gerçekte, kahvenin akciğer kanserine neden olmadığı ve sadece onunla ilişkili olduğu, ancak üçüncü değişken olan sigara nedeniyle nedensel göründüğü ortaya çıkmıştır. Bu üçüncü değişkene karıştırıcı ya da gizlenen değişken denir. Bu da kahve ve akciğer kanseri arasındaki ilişkinin sahte bir korelasyon olduğu anlamına gelmektedir.

Bunun bir başka örneği de tatiller ve perakende satışlar arasındaki ilişkidir. Her ne kadar insanlar bayramlarda kutlama amacıyla daha fazla alışveriş yapıyor olsa da, satışlardaki artışın ne kadarının bayramlardan, ne kadarının ise bayramlarda yapılan özel fırsat ve promosyonlardan kaynaklandığını söylemek zordur. Burada, özel fırsatlar tatil ve satışlar arasındaki ilişkiyi karıştırmaktadır.

Interpreting Hypothesis Test Results

Chicago ve Bangkok'da Yaşam Süreleri

Chicago ve Bangkok'ta ortalama yaşam süreleri arasında bir fark olup olmadığının test edilmek istendiğini varsayalım.

- H0 = Chicago ve Bangkok'ta ortalama yaşam süreleri arasında bir fark yoktur.
- H1 = Chicago'da yaşayanlar Bangkok'ta yaşayanlardan daha uzun bir yaşam süresine sahiptir.

Ortalama yaşam beklentisinin örnekleme dağılımı

- Tüm nüfus verilerini topanamaz, bu nedenle bir yaklaşım, her şehirden orijinal veriler üzerinde değiştirme ile örnekleme yapmak ve her örnek için ortalama yaşam beklentisini hesaplamaktır.
- Bu 10000 kez tekrarlanarak ve sonuçları görselleştirilerek, Bangkok ve Chicago'daki ortalama yaşam beklentisi için normal dağılımlar görülebilir ve Chicago daha büyük bir beklenen değere sahiptir!

p-value

- Hipotez testlerinde sonuç çıkarırken p-değeri adı verilen bir ölçüt kullanılır.
- Bu, sıfır hipotezinin doğru olduğunu varsayarak en az gözlemlediğimiz kadar uç bir sonuç elde etme olasılığıdır.
- Diyelim ki, 79,3'lük bir popülasyon ortalaması göz önüne alındığında, Chicago yaşam beklentisi için örnek ortalamasının 82'ye eşit veya daha fazla olma olasılığını bilmek istiyoruz. Örneklem ortalamaları dağılımını görselleştirebilir ve 82'den itibaren toplam alana bakarak p-değerinin 0,037 olduğunu, yani ortalama yaşam beklentisinin 82 veya daha fazla olduğunu gözlemleme şansımızın yüzde 3,7 olduğunu belirleyebiliriz.

Significance level (α)

- Yanlış bir sonuca varma riskini azaltmak için, sıfır hipotezini yanlışlıkla reddetmek için bir olasılık eşiği belirlenir. Bu olasılık eşiği alfa veya anlamlılık düzeyi olarak bilinir.
- Önyargıyı en aza indirmek için veri toplanmadan önce karar verilir, çünkü bir araştırmacı verileri gördükten sonra kendi çıkarlarına hizmet eden bir sonuç çıkarmak için farklı bir eşik seçebilir.
- Bunun için tipik bir değer 0,05'tir, yani Chicago sakinlerinin Bangkok sakinlerinden daha uzun yaşadığı sonucuna yanlış bir şekilde varmak için yüzde beş şans vardır.
- Veri toplandıktan sonra, p-değerinin alfa değerinden küçük ya da eşit olup olmadığına bakılır. Eğer p-değeri bu kriteri karşılıyorsa, sıfır hipotezini reddetme konusunda güvenilebilir. Bu gerçekleşirse sonuçlar istatistiksel olarak anlamlı olarak tanımlanır.

Type I/II Error

- Yanlış olduğu halde sıfır hipotezi yanlışlıkla kabul edilebilir. Bu, ikinci tip hata olarak bilinir.
- Doğru olduğunda boş hipotezi yanlış bir şekilde reddedilebilir. Bu birinci tip hata olarak kabul edilir.

Sampling and point estimates (Örnekleme ve nokta tahminleri)¶

Estimating the population of France

Fransa'da milyonlarca insan olduğu için bu gerçekten pahalı bir işlemdir. Modern veri toplama teknolojisine rağmen çoğu ülke maliyet nedeniyle sadece beş veya on yılda bir nüfus sayımı yapmaktadır.

Hanehalklarının örneklenmesi

1786'da Pierre-Simon Laplace, nüfusun daha az çabayla tahmin edilebileceğini fark etti. Orada yaşayan her haneye sormak yerine, az sayıda haneye sordu ve istatistikleri kullanarak tüm nüfustaki insan sayısını tahmin etti. Tüm nüfusun bir alt kümesiyle çalışma tekniğine **örnekleme (sampling)** denir.

Population vs. sample

- Population, ilgilenilen verilerin tam kümesidir.
- Sample, üzerinde çalışılan verilerin alt kümesidir.

Population parameters & point estimates

- Bir popülasyon parametresi, popülasyon veri kümesi üzerinde yapılan bir hesaplamadır.
- Bir nokta tahmini veya örnek istatistiği, örnek veri kümesine dayalı bir hesaplamadır.

Convenience sampling (Kolay Örnekleme)

Önceki alıştırmalarda hesaplanan nokta tahminleri, bunların dayandığı nüfus parametrelerine oldukça yakındı, ancak durum her zaman böyle midir?

The Literary Digest election prediction

- 1936'da The Literary Digest adlı bir gazete, bir sonraki ABD başkanlık seçimlerini tahmin etmeye çalışmak için kapsamlı bir anket düzenledi.
- On milyon seçmeni aradılar ve iki milyondan fazla yanıt aldılar.
- Yaklaşık bir buçuk milyon kişi Landon'a oy vereceğini söyledi ve bir milyondan biraz fazla kişi Roosevelt'e oy vereceğini söyledi. Yani Landon'ın oyların yüzde elli yedisini, Roosevelt'in ise oyların yüzde kırk üçünü alacağı tahmin ediliyordu.
- Örneklem büyüklüğü çok büyük olduğundan, bu anketin çok doğru olacağı varsayılıyordu. Ancak seçimde Roosevelt oyların yüzde altmış ikisini alarak ezici bir çoğunlukla kazandı.
- Peki ne ters gitti? 1936'da telefonlar bir lükstü, bu yüzden The Literary Digest tarafından ulaşılan tek kişiler nispeten zengindi. Seçmen örneği, seçmen nüfusunun tamamını temsil etmiyordu ve bu yüzden anket örneklem yanlılığından (sample bias) zarar gördü.
- Veriler en kolay yöntemle, bu durumda insanları telefonla arayarak toplandı. Buna **kolay örnekleme** denir ve sıklıkla örneklem yanlılığına eğilimlidir. Örneklemeden önce, yanlı sonuçlardan kaçınmak için veri toplama sürecinin düşünülmesi gerekir.

Finding the mean age of French people

- Disneyland Paris'te tatildeyken, Fransızların ortalama yaşını merak etmiş olun.
- Bir cevap almak için, yakınınızda duran on kişiye yaşlarını sorun.
- Ortalama yaşları 24.6 olduğunu görün.

- Bunun tüm Fransız vatandaşlarının ortalama yaşı için iyi bir tahmin olacağını düşünebilir misiniz?

Pseudo-random number generation

Bir bilgisayar bu rastgele örneklemeyi nasıl yapar?

Rastgele ne demek?

Bir popülasyondan rastgele veri noktaları seçmek istiyorsak, hangi veri noktalarının önceden sistematik bir şekilde seçileceğini tahmin edememeliyiz.

True random numbers

- Gerçekten rastgele sayılar üretmek için genellikle yazı tura atmak veya zar atmak gibi fiziksel bir işlem kullanmak zorundayız.
- Hotbits hizmeti radyoaktif bozunmadan sayılar üretir ve RANDOM.ORG, yıldırım tarafından üretilen radyo sinyalleri olan atmosferik gürültüden sayılar üretir.
- Ne yazık ki, bu işlemler rastgele sayılar üretmek için oldukça yavaş ve pahalıdır.

Pseudo-random number generation

- Çoğu kullanım durumu için, ucuz ve hızlı olduğu için sözde rastgele sayı üretimi daha iyidir.
- Pseudo-random'da, her değer rastgele görünse de aslında bir önceki rastgele sayıdan hesaplandığı anlamına gelir.
- Hesaplamalara bir yerden başlamanması gerektiğinden, ilk rastgele sayı, tohum değeri(seed value) olarak bilinen değerden hesaplanır.
- Rastgele kelimesi tırnak içindedir ve bu sürecin gerçekten rastgele olmadığını vurgular. Belirli bir tohum değerinden başlanırsa, gelecekteki tüm sayılar aynı olacaktır.

Visualizing random numbers

Her rastgele sayı fonksiyonunun ilk argümanları dağıtım parametrelerini belirtir. Size argümanı kaç tane sayı üretileceğini belirtir, bu durumda beş bin. Beta dağılımını seçtik ve parametreleri a ve b olarak adlandırıldı. Bu rastgele sayılar sürekli bir dağılımdan gelir, bu yüzden bunları görselleştirmenin harika bir yolu bir histogram kullanmaktır. Burada, sayılar beta dağılımından üretildiği için tüm değerler sıfır ile bir arasındadır.

Random numbers seeds

NumPy ile rastgele bir başlangıç noktası ayarlamak için np.random.seed() metodu kullanılır. np.random.seed() başlangıç noktası değeri için istenilen herhangi bir sayı olabilen bir tam sayı alır. np.random.normal() dağılımdan sözde rastgele sayılar üretir. np.random.normal() ikinci kez çağırılırsa iki farklı rastgele sayı elde edlir. Aynı başlangıç noktası değeriyle np.random.seed() çağrılarak başlangıç sıfırlanırsa ve ardından np.random.normal() tekrar çağrılırsa, öncekiyle aynı sayıları elde edilir.

Sampling Methods

Simple random and systematic sampling

Bir popülasyondan örnekleme yapmanın çeşitli yöntemleri vardır.

Basit rastgele örnekleme

Basit rastgele örnekleme, bir çekiliş veya piyango gibi çalışır. Çekiliş biletleri veya piyango topları popülasyonuyla başlanır ve bunlar rastgele birer birer seçilir.