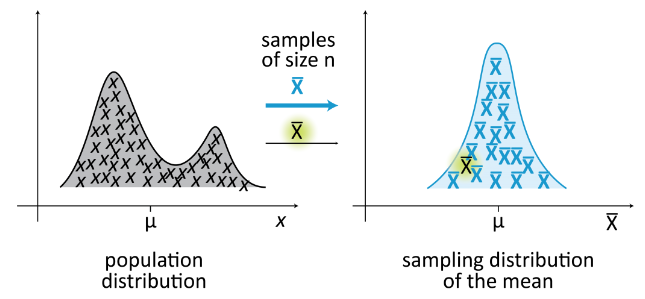
**Central Limit Theorem**

**Sampling Distribution of the Sample Mean**

Bir tıp araştırmacısı olduğunuzu ve 2015 ile 2020 yılları arasında ABD'de doğan tüm bebeklerin ortalama ağırlığı hakkında araştırma yaptığınızı düşünün. Şimdi, örneklem büyüklüğü 100 (n=100) olan 50 rastgele örnek aldığınızı varsayalım. Yani 50 farklı grubunuz (örnek) var ve birlikte çalıştığınız toplam bebek sayısı 50\*100=5000. Her örneğin (grubun) ortalama değerini hesaplar ve bu 50 ortalama değeri çizerseniz, örnek ortalamalarının örnekleme dağılımının grafiğini elde edersiniz. Tekrarlanan örneklemede, rastgele örneklerin popülasyon ortalamasına göre ortalama almasını bekliyoruz.

Numune araçlarının örnekleme dağılımı, birçok insanın takılıp kaldığı yerdir. Normalde, ortalamanın örnekleme dağılımı mevcut değildir. Teoriktir. Çünkü sonsuz sayıda örnek alırsak elde edeceğimiz araçların dağılımıdır (olası her örnek). Bu, 50 örnek değil, sonsuz sayıda örnek kullanmamız gerektiği anlamına gelir. Nüfusun 10 milyon bebek içerdiğini ve örneklem büyüklüğünün n=100 olduğunu hayal edin. Bir seferde 100 alınan 10 milyon bebek için kombinasyonlar neredeyse sonsuz bir sayıdır. Teorik numune sayısı için tüm olası kombinasyonları kapsamanız gerekir. Ancak, ortalama kavramının örneklem dağılımını daha kolay hayal edebilmeniz için yukarıda belirtilen örneği sınırlı sayıda örnekle (50) sağladık.



Bu şemada popülasyon dağılımının normal bir dağılımdan uzak olduğunu görebilirsiniz. Popülasyondan alınan bir örnek, örnek ortalama x-bar'a yol açar. "Tüm numune araçları", "numune araçlarının numune dağılımını" oluşturur. Ortalamanın örneklem dağılımı ise normal dağılım formuna sahiptir ve popülasyon dağılımından daha dardır. Bu, bir sonraki konuda tartışacağımız Merkezi Limit Teoremi ile açıklanmaktadır.

**Central Limit Theorem**

Merkezi Sınır Teoremi, örneklem boyutu büyüdükçe, örneklem araçlarının örnekleme dağılımının, popülasyon dağılımının şekli ne olursa olsun, normal bir dağılıma yaklaştığını belirtir. Bu gerçek özellikle örneklem büyüklüğü (n) 30'un üzerindeyse geçerlidir. Çünkü istatistikçiler birçok popülasyon dağılımı için en az 30'luk bir örneklem büyüklüğünün "yeterince büyük" olduğunu bulmuşlardır.

**Sampling Error**

Yalnızca bir örnek almak ve ortalama gibi bir örnek istatistiğinin sonuçlarını elde etmek, popülasyon parametresinin bir nokta tahminini oluşturur. Farklı bir örnek seçilirse, bu tek tahmin neredeyse kesinlikle aynı olmayacaktır. Çünkü örnek istatistikleri örnekten örneğe farklılık gösterir.

Örneklem büyüklüğü 100 (n=100) olan 10 milyon bebeği ve 50 örneği hatırlayın. Her numunenin numune ortalamaları (100 bebeğin ortalama ağırlığı) numuneden numuneye farklılık gösterecektir. Örneğin birinci grupta 3.5 kg, ikinci grupta 3,6 kg olabilir vs. 3,2 ile 3,8 kg arasında değiştiğini düşünün. Örnek ortalamalarından bazıları popülasyon ortalamasından daha yüksek olacak ve örnek ortalamalarından bazıları popülasyon ortalamasından daha düşük olacaktır. İstatistikte, örnek ortalamalarının bu beklenen varyasyonu örnekleme hatasıyla açıklanır.

Bu durumda örnekleme hatasının "popülasyonun tamamını kullanmak yerine bir popülasyondan örnek alınması sonucunda ortaya çıkan hata" olduğunu söyleyebiliriz.

Örnekleme hatasının miktarı öncelikle popülasyonun kendisindeki varyasyona ve seçilen örneğin büyüklüğüne (n) bağlıdır. Daha büyük örnekler (n daha büyük olduğunda) daha az örnekleme hatasına sahip olacak ancak üzerinde çalışmak daha maliyetli olacaktır.

**Standard Error**

Standart Hata (SE), standart sapmaya benzer. Her ikisi de yayılma ölçütleridir. İki terim aslında önemli bir farkla eşittir. Standart sapmalar popülasyon verilerini kullanırken, standart hata örnek verileri kullanır.

İstatistikte, "ortalamanın standart hatası" terimiyle çok sık karşılaşırız. Standart hata bize örnek istatistiğimizin (örnek ortalaması gibi) gerçek popülasyon ortalamasından ne kadar saptığını söyler. Örnek boyutunuz (n) ne kadar büyük olursa, standart hata o kadar küçük olur.

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

**Confidence Interval**

Uygulamada, bir anakütle parametresini (örneğin anakütle ortalaması) tahmin etmek için temel olarak yalnızca bir örnek kullanılır. Ancak, birkaç örnek arasındaki farklılıkları göz önünde bulundurarak (bir deneyi veya anketi tekrar tekrar yaparsak), istatistikçiler, alt ve üst limitli belirtilen bir aralığın uygun şekilde olma olasılığını gösteren bir "güven aralığı" kavramını geliştirdiler. parametreyi tahmin eder.

Güven aralığı kavramı her zaman güven düzeyi ile birlikte kullanılır. Güven düzeyi yüzde olarak ifade edilir (örneğin, %90, %95 veya %95 güven düzeyi). Bu, bir deneyi veya anketi tekrar tekrar tekrarlamanız gerektiği anlamına gelir, sonuçlarınızın zamanın yüzde 90'ında bir popülasyondan elde ettiğiniz sonuçlarla eşleşecektir. Güven aralığı ise alt ve üst limitler arasındaki boşluktur.

**Confidence Interval Formula**

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

Örneğin, bir grup evcil hayvan sahibiyle yılda kaç kutu kedi maması aldıklarını görmek için anket yaparsınız. Örnek ortalaması 125, örnek boyutu 25, örnek standart sapması 30'dur. Bu aralığı %90 güven seviyesinde hesapladığınızı ve 125±10 = [115, 135] güven aralığı elde ettiğinizi hayal edin. Yılda ortalama 115 ile 135 kutu arasında %90 güvenle aldıklarını söyleyebiliriz. Bu, yalnızca 25 örneğe dayanarak popülasyon ortalamasının 115 ile 135 arasında bir yerde olduğundan %90 emin olabileceğimiz anlamına gelir.

Örneklem büyüklüğü arttıkça daha dar aralık (aynı güven düzeyi ile) hesaplayabiliriz. Bir önceki örneğimizde, örneklem boyutunu 25'ten 100'e çıkarırsak, %90 güven aralığı yaklaşık [120, 130] arasında olacaktır.

**Sampling Error and Confidence Intervals**

**t Dağıtım**

t dağılımı (veya Student's t Dağılımı), normal dağılım eğrisine çok benzeyen, yalnızca biraz daha kısa ve şişman olan bir dağılım ailesidir. t dağılımı, örneklem büyüklüğü küçük olduğunda ve/veya popülasyon standart sapması (σ) bilinmediğinde popülasyon parametrelerini tahmin etmek için kullanılır.

Örneklem boyutu ne kadar büyük olursa, t dağılımı o kadar normal dağılıma yaklaşır. 30'dan büyük örneklem büyüklükleri için dağılım normal dağılıma çok benzer.

**💡t Dağıtımı**

t, z'den daha fazla yayılmış. Normalden "daha düz". Bunun nedeni, s kullanımının daha fazla belirsizliğe yol açarak t'yi z'den daha "eğik" hale getirmesidir.

Yayılma miktarı örnek boyutuna bağlıdır. Örnek boyutu ne kadar büyük olursa, s'nin σ'ya yakın olduğundan o kadar emin olabiliriz ve t, z'ye ne kadar yakınsa, normaldir.

Serbestlik dereceleri (dof veya d.f.): Verilen n adet veri için, x̄'yi hesaplarken bir "serbestlik derecesi" kullanırsınız ve n-1 bağımsız bilgi parçası bırakırsınız.

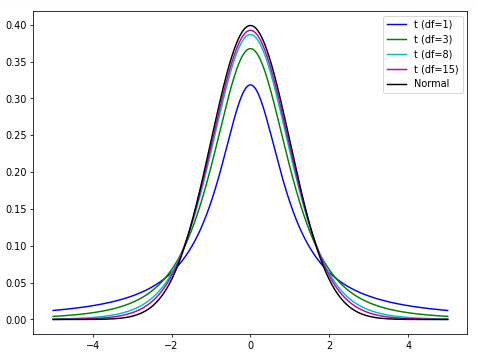
Kaynak: Gonick, L., Smith, W. ve Smith, W. (1993). İstatistikler için çizgi film rehberi. New York: HarperPerennial.

Serbestlik derecesi (n-1) arttıkça, t-dağılımı standart normal dağılıma (z-dağılımı) yaklaşacaktır.

[jupyter notebook#1](file:///C:\Users\Arslan\Desktop\GitHub\daily_work_file\09.07.2022)

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu



**SciPy Kullanan Nüfus Ortalaması için Güven Aralığı**

**Bilinen Büyük Örneklem veya Popülasyon Varyansı (z-skoru kullanılarak)**

Büyük numunelerle (n≥30) çalışırken, numune ortalamasının numune dağılımının yaklaşık olarak normal dağıldığını varsayabiliriz (Merkezi Limit Teoremi sayesinde). Bu durumda, Scipy'deki stats.norm.interval() işlevini kullanarak ortalamanın güven aralığını hesaplayabiliriz.

[jupyter notebook#2](file:///C:\Users\Arslan\Desktop\GitHub\daily_work_file\09.07.2022)

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

scipy.stats kullanarak ortalama etrafında %95 Güven Aralığını hesaplayın:

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

**Popülasyon Varyansı Bilinmiyor (t-puanı kullanılarak)**

Popülasyon standart sapması σ bilinmiyorsa, örnek standart sapması s kullanılır. Bu, örnekten örneğe değiştiği için ekstra belirsizliğe neden olur. Bu nedenle normal dağılım yerine t dağılımı kullanılır. Bu durumda, Scipy'deki stats.t.interval() işlevini kullanarak ortalamanın güven aralığını hesaplayabiliriz.

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

scipy.stats kullanarak ortalama etrafında %95 Güven Aralığını hesaplayın:

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu