

Parametrik ve Parametrik Olmayan Testler

Parametrik testlerde değişkenlerin ^{*}Normal Dağılması,
^{*}Varyansların homojen olması ve bazı durumlarda gözlem
sayısının 30 ve üzeri olması gibi varsayımlar vardır.

→ Bu varsayımları Test ettikten sonra eğer varsayımlar
gerçekleşirse t-testi; Z-testi ya da ANOVA
testleri gibi testler gerçekleştirilebilir.

Parametrik olmayan Test'ler ise bu varsayımların geçerli
olmadığı durumlarda kullanılır.

Parametrik Testler:

- Verilerin normal dağılıma uyması
- Varyansların homojen olması (eş varyanslılık)
- Deneklerin evrenden rastgele seçilmiş olması
- Deneklerin birbirinden bağımsız olması
- Denek sayılarının yeterli olması **Gerekir.**

Parametrik olmayan Testler:

- Evrene yönelik tahminde bulunmazlar
- Verinin normal dağılımı gerekmez. (Bu yüzden kategorik verilerde kullanılır)

★

Kategorik verilerde (Nominal, ordinal) parametrik olmayan yöntemler;

Sürekli verilerde (aralıklı, oransal) parametrik yöntemler kullanılır.

Nümerik veri tek grup

Nümerik veri 2 bağımsız grup

Nümerik veri 2 bağımlı grup

Nümerik veri 2'den fazla bağımsız grup

Kategorik veri 2 veya daha fazla bağımsız grup

Kategorik veri 2 bağımlı grup

Tek örnekleme t-test

Bağımsız örnekleme t-test

Bağımlı örnekleme t-test

One way Anova

Ki-kare Test

McPemar Test

Nonparametric

Veriden kastımız >> İncelediğimiz değişken yarı ölçmek istenilen değişkendir.

Örneğin = Sigara kullanma durumu (Kategorik)

Kilo (Nümerik)

.../.../...

	<u>Parametrik</u>	<u>Non-parametrik</u>
* Correlation test	Pearson	Spearman

* t-test	} Parametrik Yöntemler
z-test	
Anova (F-test)	

* Ki-Kare	} Nonparametrik Yöntemler
Mann-Whitney	
Wilcoxon	
McNemar	

Anova Dedik? Nedir bu Anova.

Varyans Analizi (ANOVA):

Ne zaman kullanılır?

1) Ortalamalar arasında fark olup olmadığına bakmak istediğimizde

2) Continuous Data ve 2 ya da daha fazla grubumuz olduğu zaman.

→ Bize grupların birbirinden farklı olup olmadığını bilgisini verir ancak hangi grupların birbirinden farklı olduğunu bilgisini vermez

→ Hangi grupların birbirinden farklı olduğuna bakmak için Post-hoc testleri yapılır

ANOVA, bağımsız değişkenlerin kendi aralarında nasıl etkileşime girdiklerini ve bu etkileşimlerin bağımlı değişken üzerindeki etkilerini analiz etmek için kullanılır.

| Anova'nın t testinden farkı

- Örneğin 3 grup olsun t-testini 1-2, 1-3, 2-3 grupları için ayrı ayrı yapmamız gerekli.
- Her testin kendine özgü Tür 1 hata olasılığı var (0.05 olsun).
- Yani her t testinde Tür 1 hatası yapma olasılığı (0.95).
- 3 t testi olduğuna göre hata yapma olasılığı $0.95 \times 0.95 \times 0.95 = 0.857$.
- Yani Tür 1 hatası yapma olasılığımız 0.05'ten 0.143'e yükseldi.

!!! Bu kabul edilemez (Grup sayısı 3 yerine 5 olsaydı)

* Hata = 0.14 oldu

t - testi 2 örneğin ortalamalarını eşit olup olmadığını hipotezini test eder

* Anova ise 3 ya da daha fazla ortalamanın eşit olup olmadığını test eder.

→ Anova F istatistikini verir. F verilerindeki sistematik varyans miktarını sistematik olmayan varyansla karşılaştırır.

$$F = \frac{\text{Between groups variance}}{\text{within - group variance}}$$

Levene testi anlamlı olduğu durumda, varyansları eşitlemek için veri dönüştürümü yapılır

Varyansların homojenliğine bakarken "Levene Testi" kullanılır.

Anova'nın varsayımları → 1) Varyansların homojen olması
2) Değişkenlerin normal dağılması

$$F = \frac{\text{Between Group Variance}}{\text{Within Group Variance}}$$

* Toplam Değişkenlik = Gruplar Arası Değişkenlik + Gruplar İçi Değişkenlik

$$GKT = GAKT + GİKT$$

GKT ⇒ Genel Kareler Toplamı

GAKT ⇒ Örnek ortalamalarının genel ortalamadan gösterdiği sapmalar.

GİKT ⇒ Her bir örnek teki değerlerin kendi örnek ortalamalarından gösterdiği sapmalardır.

Sonuç olarak;

GAKT = Gruplar arası Kareler toplamı

GİKT = Grup içi Kareler toplamı

$$GAKT = n \sum_{i=1}^k (x_i - \bar{x})^2$$

$$GİKT = \sum_{i=1}^k \sum_{j=1}^n (x_{ij} - \bar{x}_i)^2 \quad GKT = \sum_{i=1}^k \sum_{j=1}^n (x_{ij} - \bar{x})^2$$

*
$$F = \frac{GAKT}{GİKT}$$

$$H_0: \mu_1 = \mu_2 = \mu_3$$

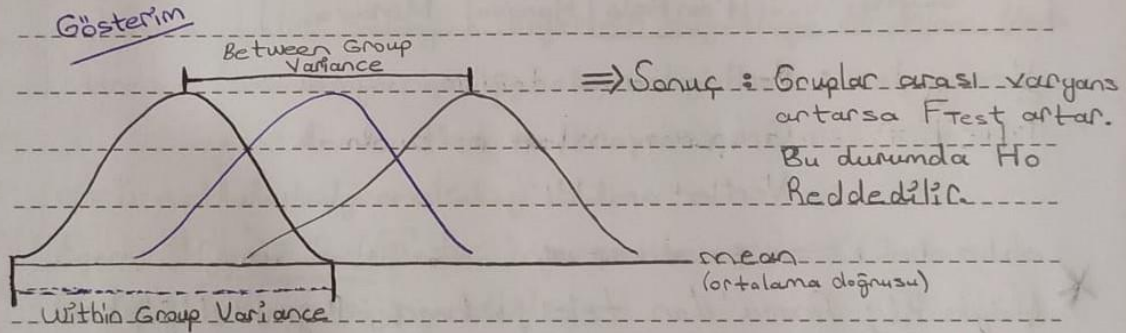
H_1 : En az 2 anakütle ortalaması eşit değildir.

Diye lim ki Test İstatistiği $\Rightarrow F = 16,97$ ve

Kritik Değer $\Rightarrow F_c = 3,24$

Bu durumda %5 önem seviyesinde H_0 hipotezini

reddederek en az 2 örnek ortalamasına birbirinden farklı olduğunu karar verilir.



* $F_{Tablo (kritik)} < F_{Test (hesaplanan)} \Rightarrow H_0 \text{ Red}$

* $F_{Tablo (kritik)} > F_{Test (hesaplanan)} \Rightarrow H_0 \text{ Reddedilemez}$

Anova Testi Örneği:

Bir satış sorumlusu 3 farklı deterjan markasından hangisinin daha çok satıldığını merak ediyor. Satılan deterjan markalarına ortalama olarak birbirinden farklı olup, farklılaşmadığını hesaplayınız

$$H_0: \mu_1 = \mu_2 = \mu_3$$

$$H_1: \mu_i \neq \mu_j \text{ (En az biri farklı)}$$

Post hoc \Rightarrow Hangi gruplar arasında fark var ara bakar.

F-test bize hangi gruplar arasında bütün farkın olduğunu söyleyemez.

Manova

Bağımlı Değişken Sayısı	Bir	1'den fazla	Bir	2 ve daha fazla
			Anova	Anova
Bağımlı Değişken Sayısı	Bir	1'den fazla	Manova	Manova

Varsayımları: Veriler normal dağılıma uymalı

Grup varyansları eşit olmalı

Veriler aralıklı yada oransal olmalı

* Şöyle ki; Anova'dan tek farkımız Bağımlı değişken sayımızın 1'den fazla olmasıdır.

Örnek soru:

Öğrencilerin eğitim bilimleri ve genel kültür KPSS puanları mezun oldukları lisans programına göre anlamlı farklılık gösterir mi?

Bağımsız Değişken \Rightarrow Mezun Olunan Lisans programı

Bağımlı Değişkenler \Rightarrow Eğitim Bilimleri KPSS puanı

Genel Kültür KPSS puanı

\Rightarrow Bağımlı değişkenlerin hiçbirinde, bağımsız değişkenin gruplarına göre bir ortalama farklılık olmadığı durumda H_0 Kabul edilir.

Korelasyon \Rightarrow Değişkenlerin hareket yönünü ve ilişkisinin gücünü gösterir.

Kovaryans \Rightarrow 2 değişkenin birbirine ne kadar benzerlikte hareket ettiğini gösterir.

Eğer Korelasyon var ise bu şu demektir; Bir değişken artar/azalır iken diğeri de artıyor/azalıyor. İşte bu birlikte değişim "Kovaryans" ile ölçülür. Kovaryans birlikte değişmek anlamına gelir. Kovaryans değerinin büyüklüğü bir anlam ifade etmez; tamamen veriye bağlı olarak değişir.

n = örneklem sayısı olsun.

Eğer; $|r| \geq 2/\sqrt{n}$ ise, gerçekten X_1 ve X_2 arasında bir ilişki olduğunu söyleye biliriz.

! Korelasyon Yeden - Sonuç ilişkisi belirlenmez.

Örneğin, dondurma tüketimi ile suf oranlarının artışı arasında bir korelasyon bulunabilir. Fakat dondurma tüketimi arttığında ötürü suf oranları artar diyemeyiz. Bunun altında yatan farklı bir neden, örneğin sıcaklıkların artması olabilir.

$$\text{Cov formülü} \Rightarrow \text{Cov}(X, Y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$\text{Cov}(X, Y)$ = X bağımsız değişkeni ile Y bağımsız değişkeninin kovaryansı

x_i = X değişkeni \bar{x} = X ortalaması

y_i = Y değişkeni \bar{y} = Y ortalaması

* Değişken Seçimi Sırasında
F-value 'dan nasıl yararlanılır ve Nasıl yorumlanır?

$$F = \frac{\text{Between Group Variability}}{\text{Within Group Variability}} = \frac{\text{Explained Variance}}{\text{Unexplained Variance}}$$

Aynı şey

F test basically represent the Ratio Between the explained and the unexplained variance.

* F value 'su yüksek olan değişkenler daha informative'dir.

F1 skor for Classification:

"Öncelikle şunu netleştirelim. F1 skor ile F test aynı şeyler değil ve tamamen isim benzerlikleri var.

F test regresyon ; F1 skor Sınıflandırmada kullanılır.

Sınıflandırmadan Başlayalım:

	Predicted		
Gerçekleşen (Actual)	TP	FN	FP = Type 1 error
	FP	TN	

FN = Type 2 error

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1 skoru bize Kesinlik (Precision) ve

Duyarlılık (Recall) değerlerinin harmonik

ortalamasını gösterir.

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \text{Modelin } F_1 \text{ skoru}$$

Peki sınıflandırma problemi adeki feature'ların değişken bazlı F_1 skor değeri neyi gösterir;

Sınıflandırmada univariate yaklaşım vardır. Şöyle ki; ben sadece o değişken ile modeli kuraydım. Recall ve Precision ne olurdu diye oradan sadece o değişkenle kurulmuş modelin model F_1 skorunu alacaktım.

ÖR

	F_1
X_1	0.57
X_2	0.32
X_3	0.17

X_1 'i tek başına alıp modeli kurduğumuzda modelin F_1 skoru = 0.57 olmuştur. Aynı şekilde X_2 değişkeni ile tek başına model kurulduğunda modelin F_1 skoru = 0.32 olmuştur.

F skor for Regression:

Model overall anlamlılığı için $\Rightarrow H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0$

H_1 : En az Biri 0'dan farklı.

F istatistiğini Overall signifikant olarak kullanıyoruz.

Modelin Overall signifikant olması en az 1 bağımsız değişkenin istatistiksel olarak anlamlı olmasını gerektirir.

$$F = \frac{\text{Açıklanan Varyans}}{\text{Açıklanamayan Varyans}} = \frac{MSR}{MSE} = \frac{SSR / k}{SSE / (n - k - 1)}$$