


DT/NT : DATA SCIENCE
LESSON : STATISTICS -2
SUBJECT: INDEPENDENT T TEST
DEPENDENT T TEST
ANOVA
CATEGORICAL TEST

BATCH: 223



TECHPRO
EDUCATION

techproeducation.com [+1 \(585\) 304 29 59](https://wa.me/915853042959)

[f](#) [X](#) [in](#) [v](#) [i](#) [@](#)

STATISTICS - 2

Data Science Program
Statistics Session -9

Session - 9 Content

Content

- Independent Samples T Test
- Dependent T test
- One Way ANOVA
- Categorical Data Analysis



RECAP

**Herkes önceki dersten hatırladığı
1 cümle yazabilir mi?**



Recap – Previous Lesson



Review Recap



Significance Test Steps

HYPOTHESIS TESTS:

INDEPENDENT SAMPLES T TEST

Bağımsız t testi
(Unpaired t testi)

Independent Samples T test

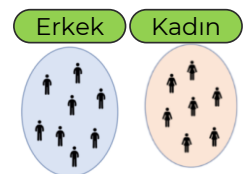
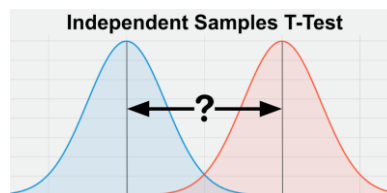
Bağımsız t testi

- Aynı continuous, dependent değişken **üzerinde iki independent grup arasındaki ortalamaları karşılaştırır**
- İki farklı grup üzerinden tek değişkenini analizi için Independent t test kullanılabilir

Dependent Değişken

2 Independent Grup

Maaş	Cinsiyet
135.000	E
145.000	E
138.000	K
150.000	K
128.000	E



Independent Samples T test

Assumptions

- 2 grup için quantitative-nicel bir değişken
- Rasgele örneklemden bağımsız rasgele değişkenler
- Her grup için Normal dağılım

Hypothesis

- Null Hipotez:
 - $H_0: \mu_1 = \mu_2$
(İki grubun ortalamaları arasında fark yok)
- Alternative Hipotez:
 - $H_a: \mu_1 \neq \mu_2$
(Significant difference between the means of the two groups)



Independent Samples T test

Test Statistics

- **Equal Variances not assumed** (2 grubun varyanslarının eşit olmadığı varsayılırsa)

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1} \left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1} \left(\frac{s_2^2}{n_2}\right)^2}$$

- **Equal Variances assumed** (2 grubun varyansları eşit varsayılırsa)

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}}$$

$$df = n_1 + n_2 - 2$$



Independent Samples T test - Example

Örnek

- Bir kimyasal prosesin ortalama verimini nasıl etkilediklerini belirlemek için iki katalizör analiz edilecektir.
- Halen kullandığı 1.katalizör yerine verimi düşürmemek kaydıyla daha ekonomik 2.katalizör araştırılmaktadır.
- 2.katalizör için bir pilot uygulama alanında test edilip tablodaki değerler elde edilmiştir.
- 0,05 için, eşit varyans olduğu varsayılarak, ortalama verimler arasında anlamlı bir fark var mı?

Test No	1.Katalizör	2.Katalizör
1	91,50	89,19
2	94,18	90,95
3	92,18	90,46
4	95,39	93,21
5	91,79	97,19
6	89,07	97,04
7	94,72	91,07
8	89,21	92,75
Xort	92,225	92,733
S	2,39	2,98



Independent Samples T test - Example

Step 1

**Assumptions
(Varsayımlar)**

**2 grup için Quantitative
değişken**

**Independent random
örnekler**

2 grup da Normal dağılım

Step 2

**Hypotheses
(Hipotezler)**

Null Hypothesis ($H_0: \mu_1 = \mu_2$)

Alternate Hypothesis ($H_a: \mu_1 \neq \mu_2$)



Independent Samples T test - Example

$$S_{Pooled} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Sample Sizes

- $n_1 = 8$
- $n_2 = 8$

Sample Std.Spm

- $s_1 = 2.39$
- $s_2 = 2.98$

$$s_p = \sqrt{\frac{(8 - 1)2.39^2 + (8 - 1)2.98^2}{8 + 8 - 2}} = 2.70$$

Step 3

t test

Test Statistics
(Test İstatistikleri)

$$\bar{x}_1 = 92.255, \bar{x}_2 = 92.733, s_p = 2.70$$

$$t_0 = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$t = \frac{\bar{x} - \bar{y}}{s_{Pooled} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{92.255 - 92.733}{2.70 \sqrt{\frac{1}{8} + \frac{1}{8}}} = -0.35$$



Independent Samples T test - Example

Step 4

P - Value
(P- Değeri)

```
In [1]: import scipy.stats as stats
In [2]: 2*stats.t.cdf(-0.35, 14)
Out[2]: 0.7315482686624126
```

P-value =

$$2 * t_{0.35, 14} = .7315$$

Step 5

Conclusion
(Sonuç)

P - Value = 0, 7315

 $\alpha = 0,05$ P-Value > $\alpha = 0,05$

P-değeri
önceden
belirlenen α
değerinden
büyük olduğu
için
Null hipotezi
fail to reject olur

2.Katalizörün birinciden farklı olduğuna dair elimizde
yeterli güçlü bir kanıt yoktur.



Independent Samples Z test – Example 2

Örnek - σ 'nın bilinmesi durumuna örnek

- Özel bir ürün geliştiricisinin ürettiği bir boya için, 2 boya formülasyonu test ediliyor. 1. ürün standart bir üründür ve 2. ürün ise yeni bir üründür.
- Önceki tecrübelerle göre 1. ürünün kullanım süresi için $\sigma=8$ dakikadır. Yeni üründe bu süre değişmemelidir.
- 10 kişiye ilk ürün verilmiş, 10 kişiye de 2. ürün verilmiş ve bu randomly yapıldı
- Grupların ortalaması $x_1=121$ ve $x_2=112$ dakikadır.
- $\alpha = 0,05$ için buradan nasıl bir sonuç çıkar

$\sigma = 8$
Sample Sizes
 ○ $n_1 = 10$
 ○ $n_2 = 10$
Sample Means
 ○ $\bar{x}_1 = 121$
 ○ $\bar{x}_2 = 112$
 $\alpha = 0.05$



Independent Samples Z test - Example

Step 1

**Assumptions
(Varsayımlar)**

2 grup için Quantitative değişken ✓

Independent random örnekler ✓

Popülasyonun σ biliniyor veya 30 gözlem olması ✓

Step 2

**Hypotheses
(Hipotezler)**

Null Hypothesis ($H_0: \mu_1 - \mu_2 = 0$)

Alternate Hypothesis ($H_a: \mu_1 > \mu_2$)



Independent Samples Z test - Example

Step 3

Z test

**Test Statistics
(Test İstatistikleri)**

$\bar{x}_1 = 121$ and $\bar{x}_2 = 112$

$$z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{121 - 112}{\sqrt{\frac{8^2}{10} + \frac{8^2}{10}}} = 2.52$$

Step 4

**P - Value
(P- Değeri)**

```
In [1]: import scipy.stats as stats
In [2]: 1-stats.norm.cdf(2.52)
Out[2]: 0.005867741715332553
```

P-value =

$P(z > 2.52 \mid H_0 \text{ true}) = .0059$

		Second decimal place of z									
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09	
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641	
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247	
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859	
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483	
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121	
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776	
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451	
0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148	
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867	
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611	
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379	
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170	
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985	
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823	
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0722	.0708	.0694	.0681	
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559	
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455	
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367	
1.8	.0359	.0352	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294	
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233	
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183	
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143	
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110	
2.3	.0108	.0105	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084	
2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064	
2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048	
2.6	.0047	.0046	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036	



Independent Samples T test - Example

Step 5

**Conclusion
(Sonuç)**

P - Value = 0,0059

$\alpha = 0,05$

P-Value < $\alpha = 0,05$

P-değeri önceden belirlenen α değerinden küçük olduğu için Null hipotezi **reject olur**

2.Ürüne uygulanan proses olumlu bir katkı sağlamıştır



Large Sample, σ bilinmiyor – Example 3

Örnek - σ 'nın bilinmediği duruma örnek

- Egzersiz yapmanın kan basıncı üzerindeki etkileri inceleniyor.
- $n_1=500$ hasta nın yüksek kan basıncına sahip olduğu görülüp bunlara bir egzersiz programı uygulanıyor
- $n_2= 400$ olan diğer bir yüksek kan basıncı olan hastaya da egzersiz öneriliyor.
- 1 yıl sonra alınan ortalamalar şöyle oluyor.
- $X_{1ort}: 10,67$ $x_{2ort}: 7,83$
- $S_1 : 3,895$ $s_2: 4,224$

$$\bar{x} = \frac{\sum_{i=1}^{n_1} x_i}{n_1} = 10.67 \quad \bar{y} = \frac{\sum_{i=1}^{n_2} y_i}{n_2} = 7.83$$

$$s_1 = \sqrt{\frac{\sum_{i=1}^{n_1} (x_i - \bar{x})^2}{n_1 - 1}} = 3.895$$

$$s_2 = \sqrt{\frac{\sum_{i=1}^{n_2} (y_i - \bar{y})^2}{n_2 - 1}} = 4.224$$



Large Sample, σ bilinmiyor – Example

Step 1

Assumptions
(Varsayımlar)

Örnekler random seçilmiş ✓

Independent random
örnekler ✓

Popülasyonun gözlem
sayısı 30'dan fazla ✓

Step 2

Hypotheses
(Hipotezler)

Null Hypothesis ($H_0: \mu_1 - \mu_2 = 0$)

Alternate Hypothesis ($H_a: \mu_1 > \mu_2$)



Large Sample, σ bilinmiyor – Example

Step 3

Z test

Test Statistics
(Test İstatistikeleri)

$\bar{x}_1 = 10.67$, $s_1 = 3.895$ and $\bar{x}_2 = 7.83$, $s_2 = 4.224$

$$z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{10.67 - 7.83}{\sqrt{\frac{3.895^2}{500} + \frac{4.224^2}{400}}} = 10.37$$

Step 4

P - Value
(P- Değeri)

```
In [1]: import scipy.stats as stats
In [2]: 1-stats.norm.cdf(10.37)
Out[2]: 0.0
```

P-value =

$P(z > 10.37 \mid H_0 \text{ true}) = .0000$



Large Sample, σ bilinmiyor – Example

Step 5

Conclusion
(Sonuç)

P - Value = 0,0000

$\alpha = 0,05$

P-Value < $\alpha = 0,05$

P-değeri önceden belirlenen α değerinden küçük olduğu için Null hipotezi **reject** olur

Yapılan egzersizlerin kan basıncını düşürme üzerinde önemli bir etkisi vardır



YOUTUBE ONERİ VIDEO

<https://www.youtube.com/watch?v=NkGvw18zlGQ>

Two-sample t test for difference of means

Kaito grows tomatoes in two separate fields. When the tomatoes are curious as to whether the sizes of his tomato plants differ between random samples of plants from each field and measures the heights, summary of the results:

$\alpha = 0.05$ $H_0: \mu_A = \mu_B$
Assume

	Field A	Field B
Mean	1.3 m	1.6 m
Standard deviation	0.5 m	0.3 m
Number of plants	22	14

$t = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$

HYPOTHESIS TESTS: DEPENDENT T TEST

Bağımlı t testi
(Paired t test)

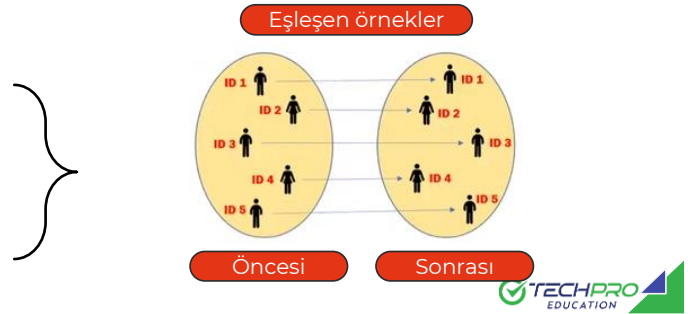
Dependent t test

Bağımlı t testi

- paired sample t test de denir
- Aynı continuous, dependent değişken üzerinde ilgili iki grup arasındaki ortalamaları karşılaştırır

Örnek: 2 aylık sigara bırakma tedavisi alan kişilerin önceki ve sonraki günlük sigara tüketimleri örneği

ID	Öncesi	Sonrası	Fark
1	12	10	2
2	18	7	11
3	23	22	1
4	10	12	-2
5	8	4	4



Dependent t test

Assumptions

- Bağımlı değişken sürekli ve aynı denek örneğinde iki kez ölçülür.
- Bağımsız değişken iki kategorik, "ilgili grup" veya "eşleşen çiftlerden" (paired) oluşan 2 kategorik gruptur
- Değişkenlerin skorları arasındaki fark normal dağılmıştır.

Hypothesis

- Null Hipotez:

$$H_0: \mu_1 - \mu_2 = 0 \text{ veya } \mu_D = 0$$

(Eşleşmiş paired popülasyonların ortalamaları arası fark 0'dır)

- Alternative Hipotez:

$$H_a: \mu_D \neq 0$$

(eşleştirilmiş popülasyon ortalamaları arasındaki fark 0 değildir)

Dependent t test

Test Statistics

$$t = \frac{\bar{x}_{diff}}{s_{\bar{x}}} \quad s_{\bar{x}} = \frac{s_{diff}}{\sqrt{n}}$$

\bar{x}_{diff} = Sample mean of the differences

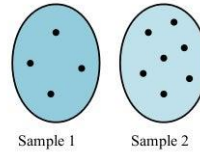
n = Sample size (i.e., number of observations)

s_{diff} = Sample standard deviation of the differences

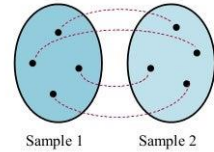
$s_{\bar{x}}$ = Estimated standard error of the mean (s/\sqrt{n})

Independent and Dependent Samples

Independent Samples



Dependent Samples



Dependent t test - Example

Örnek

- Bir yayında verilen çelik kirişlerin shear strength değerlerin tahmini verilmiştir.
- Buradaki 9 kirişe 2 metot(Karlsruhe-Lehigh) uygulanmıştır.
- $\alpha=0,05$ için ortalamalar açısından bu 2 metot arasında fark olup olmadığını inceleyelim.
- (Girder: Kiriş)

Girder	Karlsruhe Method	Lehigh Method	Difference d_j
S1/1	1.186	1.061	0.125
S2/1	1.151	0.992	0.159
S3/1	1.322	1.063	0.259
S4/1	1.339	1.062	0.277
S5/1	1.2	1.065	0.135
S2/1	1.402	1.178	0.224
S2/2	1.365	1.037	0.328
S2/3	1.537	1.086	0.451
S2/4	1.559	1.052	0.507



Dependent t test – (Paired t Test)

Step 1

Assumptions
(Varsayımlar)

Dependent değişken
continous ✓

Gözlemler birbirinden
bağımsız ✓

Her grup normal dağılıma
uygun ✓

Step 2

Hypotheses
(Hipotezler)

Null Hypothesis ($H_0: \mu_D = 0$)

Alternate Hypothesis ($H_a: \mu_D \neq 0$)



Dependent t test – (Paired t Test)

Step 3

Test Statistics
(Test İstatistikeleri)

$\bar{d} = 0.2769$, $s_d = 0.1350$, $n = 9$

$$t_0 = \frac{\bar{d}}{s_d / \sqrt{n}}$$

$$t_0 = \frac{\bar{d}}{s_d / \sqrt{n}} = \frac{0.2769}{0.1350 / \sqrt{9}} = 6.15$$

t test

Step 4

P - Value
(P- Değeri)

```
In [1]: import scipy.stats as stats
```

```
In [2]: 2*(1-stats.t.cdf(6.15, 8))
```

```
Out[2]: 0.00027399606897193785
```

P-value = 0.0003



Dependent T test – (Paired T Test)

Step 5

Conclusion
(Sonuç)

P – Value = 0, 0003

$\alpha = 0,025$

P-Value < $\alpha = 0,025$

P-değeri önceden belirlenen α değerinden küçük olduğu için Null hipotezi **reject** olur

Karslsruhe metodu ortalamalar üzerinden Lehigh e göre daha yüksek bir dayanım tahmini yaptığı görülmüştür.



HYPOTHESIS TESTS:

One-way ANOVA

Tek yönlü ANOVA

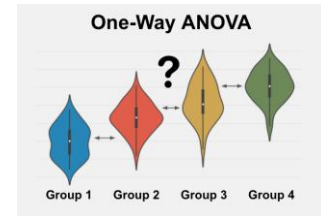
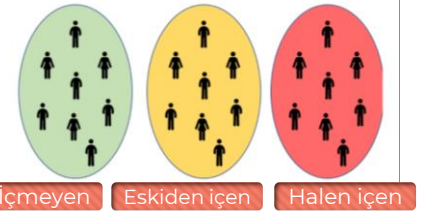
One-way ANOVA test

Tek Yönlü ANOVA testi

- Analysis of Variance kısaltması olarak ANOVA, birkaç grubun ortalamalarını karşılaştırmak için inferential bir yöntem
- Tek Yönlü ANOVA, üç veya daha fazla grubun ortalamaları arasında karşılaştırılabilir.

Periyot	İçme durumu
5,1	0
7,8	2
7,1	1
8,6	2
4,9	0
7,7	1

Bağımsız Gruplar



One-way ANOVA test

Assumptions

- Continuous olan Dependent variable
- Categorical olan Independent variable
- Independent gözlemler
- Her grup için dependent variable'ın normal dağılımı
- Gruplar arası yaklaşık eşit varyanslar

Hypothesis

Null Hipotez:

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

(Tüm k adet popülasyon ortalamaları eşit)

Alternative Hipotez:

$$H_a: \mu_i \text{ diğerlerinden farklı}$$

(k adet popülasyondan en az birinin ortalaması diğerlerine eşit değil)



One-way ANOVA test

Test Statistics – ANOVA Table

	Sum of Squares	df	Mean Square	F
Group (Between)	SSR	k-1	MSR = SSR/(k-1)	MSR/MSE
Error (Within)	SSE	n-k	MSE = SSE/(n-k)	
Total	SST = SSR+SSE	n-1		

regression sum of squares (points to SSR)
 model degrees of freedom (points to k-1)
 regression mean square (points to MSR)
 F statistic (points to MSR/MSE)
 error sum of squares (points to SSE)
 total sum of squares (points to SST)
 error degrees of freedom (points to n-k)
 total degrees of freedom (points to n-1)
 mean square error (points to MSE)

$$F = \frac{MSR}{MSE}$$

TECHPRO EDUCATION

One-way ANOVA test

Test Statistics – ANOVA Table

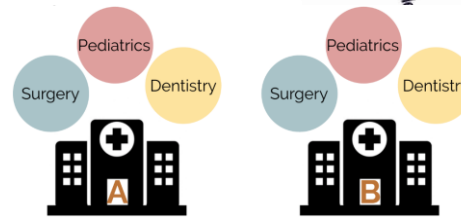
ANOVA Tablosu				
Değişkenlik Kaynağı	Kareler Toplamı	Serbestlik Derecesi	Kareler Ortalaması	Test İstatistiği
Gruplar arası	$SS_B = \sum_{i=1}^k \frac{T_i^2}{n_i} - \frac{T^2}{N}$	$v_1 = k - 1$	$MS_B = S_B^2 = \frac{KT_B}{v_1}$	$F_h = \frac{S_B^2}{S_W^2}$
Gruplar içi	$SS_W = \sum_{i=1}^k \sum_{j=1}^{n_i} X_{ij}^2 - \sum_{i=1}^k \frac{T_i^2}{n_i}$	$v_2 = N - k$	$MS_W = S_W^2 = \frac{KT_W}{v_2}$	-----
Toplam	$SS_T = KT_B + KT_W$	$v = N - 1$	-----	-----

TECHPRO EDUCATION

One-way ANOVA test

ANOVA Senaryo Örnekleri

- Eğitim metotlarının öğrenci üzerindeki etkisinin incelenmesi
- 3 farklı tedavi yönteminin hastayı iyileştirme sürecindeki değişim
- 4 farklı Videoe oluşturma metotlarının aynı kalitede olup olmadığı
- 3 yeni ilacın plasebo ya göre ne kadar farklı olduğu



One-way ANOVA test - Example

Örnek

- Statsmodal ile sunulan diğer bir Cushing data seti (Mass paketi içinde) vardır. (csv dosya) (Cushing diye bir hastalık)
- Bu hastalıkla ilgili 4 farklı grup var. Adenoma (a), bilateral hyperplasia (b), carcinoma (c), unknown (u).
- Bu 4 çeşit hastalıkla ilgili olarak bir ilacın kortizon seviyesine etki derecesi açısından bir fark olup olmadığını görmek için değerlendirme yapılacaktır.

Toplam gözlem sayısı, $n = 27$

Her gruptaki gözlem sayısı, $n_1=6, n_2=10, n_3=5$ ve $n_4=6$

Grupların üstteki sırayla ortalamaları: 3, 8.2, 19.7 ve 14

Degree of freedoms: $df_1 = 4-1 = 3$, $df_2 = 27 - 4 = 23$

$SS_B = 893.5$ ve $SS_W = 2123.6$

One-way ANOVA test - Example

Step 1

Assumptions
(Varsayımlar)

Dependent değişkenler
continous ✓

Independent değişkenler
catagorical ✓

Her grubun dağılımı
normal dağılıma yakın ✓

Varyanslar homojen ✓

Step 2

Hypotheses
(Hipotezler)

Null Hypothesis ($H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$)

Alternate Hypothesis (H_a : En az 1 μ_i farklı)



One-way ANOVA test - Example

Step 3

Test Statistics
(Test İstatistikeleri)

F test

ANOVA Table

	Sum of Squares	df	Mean Square	F
Group (Between)	SSR = 893.5	$k-1 = 3$	$MSR = 893.5 / 3 = 297.8$	$MSR/MSE = 297.8 / 92.3 = 3.226$
Error (Within)	SSE = 2123.6	$n-k = 23$	$MSE = 2123.6 / 23 = 92.3$	
Total	SST = 3017.1	$n-1 = 27$		



One-way ANOVA test - Example

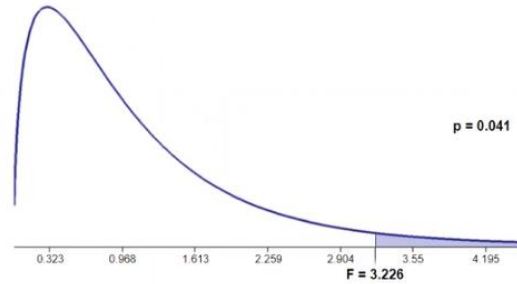
Step 4

**P - Value
(P- Değeri)**

```
In [1]: import scipy.stats as stats
In [2]: 1 - stats.f.cdf(3.226, dfn=3, dfd=23)
Out[2]: 0.041207862659964456
```

F statistic = 3.226

P-value = 0.041



One-way ANOVA test - Example

Step 5

**Conclusion
(Sonuç)**

P - Value = 0, 041

$\alpha = 0,05$

P-Value < $\alpha = 0,05$

P-değeri önceden belirlenen α değerinden küçük olduğu için Null hipotezi **reject olur**

Gruplar arasındaki farkın istatistik olarak anlamlı olduğu söylenir.



YOUTUBE ONERI VIDEO

<https://www.youtube.com/watch?v=WUoVftXvjiQ>

► A One-Way ANOVA
Example

An Example of One-Way
Analysis of Variance (ANOVA)

CATEGORICAL DATA ANALYSIS

Kategorik Data Analizi

Categorical Data Hypothesis Test

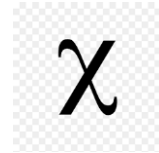
Kategorik Dataların Hipotez Testleri

- Chi-Square (Ki-kare) test kullanılabilir.
- Categorical data var ise population proportion testler kullanılmalıdır
- Sayısal olmayan değişkenler arasındaki herhangi bir ilişkinin var olup olmadığını ileri sürerek (H_0 hipotezi) bu hipotezi red edilip edilip edilmeyeceğinin incelenmesinde uygulanacak test kare testidir.

Kİ-KARE TESTİ



NEDİR?



Categorical Data Hypothesis Test

Kategorik Dataların Hipotez Testleri

Karşılaştırılacak 2 Kategori Örneği:

- 1. Sigara içme durumu (İçer – İçmez)
 - Akciğer kanseri (Kanserdir – Değildir)
- 2.
 - İrk grubu – Şeker hastalığı meyli
- Testte sorumuz şu:
 - Bir faktörün (değişkenin) varlığı/yokluğu, diğer faktörün (değişken) varlığını/yokluğunu etkiler mi?

Vaka

Üretim sektöründe faaliyet göstermekte olan bir firmada **ürün kalitesi** ile **çalışanların eğitim durumları** arasında bir ilişki olduğu düşünülmektedir. Bu tezin incelenmesi için ki-kare testi kullanılır.



Categorical Data Hypothesis Test

- Oranlar arasındaki karşılaştırma

Tedavi	İyileşme var	İyileşme yok	Toplam
Yeni ilaç	18	6	24
Plasebo	9	11	20
Toplam	27	17	44

Tabloya göre ilaç ile iyileşme oranı: $18/24 = \%75$

Tabloya göre plasebo ile iyileşme oranı: $9/20 = \%45$



Categorical Data Hypothesis Test

- Oranlar arasındaki karşılaştırma

Tedavi	İyileşme var	İyileşme yok	Toplam
Yeni ilaç	18(a)	6 (b)	24
Plasebo	9 (c)	11 (d)	20
Toplam	27	17	44

$$\chi^2 = \sum \frac{(obs - exp)^2}{exp}$$

Seçilen kişinin yeni ilaç almış gruptan olma ihtimali: $24/44$

Seçilen kişinin yeni iyileşmiş gruptan olma ihtimali: $27/44$

Seçilen kişinin iyileşmiş olup yeni ilaç almış gruptan olma ihtimalinde a hücre için beklenen değer (expected value): $24 \cdot 27 / 44 = 14,73$



Categorical Data Hypothesis Test

- Oranlar arasındaki karşılaştırma

Tedavi	İyileşme var	İyileşme yok	Toplam
Yeni ilaç	18 (14,73)	6 (9,27)	24
Plasebo	9 (12,27)	11 (7,73)	20
Toplam	27	17	44

$$\chi^2 = \sum \frac{(obs - exp)^2}{exp}$$

$$\sum \frac{(obs - exp)^2}{exp} = \frac{(18 - 14.73)^2}{14.73} + \frac{(6 - 9.27)^2}{9.27} + \frac{(9 - 12.27)^2}{12.27} + \frac{(11 - 7.73)^2}{7.73}$$

$$\chi^2 = 4,14$$

- Df = (Tablodaki satır sayısı -1)*(tablodaki Sütun sayısı -1)=1



Categorical Data Hypothesis Test

χ^2 table

Critical values in the distributions of chi-squared for different degrees of freedom

df	.05	.02	.01	.001
1	3.841	5.412	6.635	10.827
2	5.991	7.824	9.210	13.816
3	7.815	9.837	11.345	16.266
4	9.488	11.668	13.277	18.467
5	11.070	13.388	15.086	20.515
6	12.592	15.033	16.812	22.457
7	14.067	16.622	18.475	24.322
8	15.507	18.168	20.090	26.125
9	16.919	19.679	21.666	27.877
10	18.307	21.161	23.209	29.588
11	19.675	22.618	24.725	31.264
12	21.026	24.054	26.217	32.909
13	22.362	25.372	27.688	34.528
14	23.585	26.673	29.141	36.123
15	24.996	28.259	30.578	37.697
16	26.296	29.633	32.000	39.252
17	27.587	30.995	33.409	40.790
18	28.869	32.346	34.805	42.312
19	30.144	33.687	36.191	43.820
20	31.410	35.020	37.566	45.315
21	32.671	36.343	38.932	46.797
22	33.924	37.659	40.289	48.268
23	35.172	38.968	41.638	49.728
24	36.415	40.270	42.980	51.179
25	37.652	41.566	44.314	52.620
26	38.885	42.856	45.642	54.052
27	40.113	44.140	46.963	55.476
28	41.337	45.419	48.278	56.893
29	42.557	46.693	49.588	58.302
30	43.773	47.962	50.892	59.703

df	.05	.02	.01	.001
1	3.841	5.412	6.635	10.827

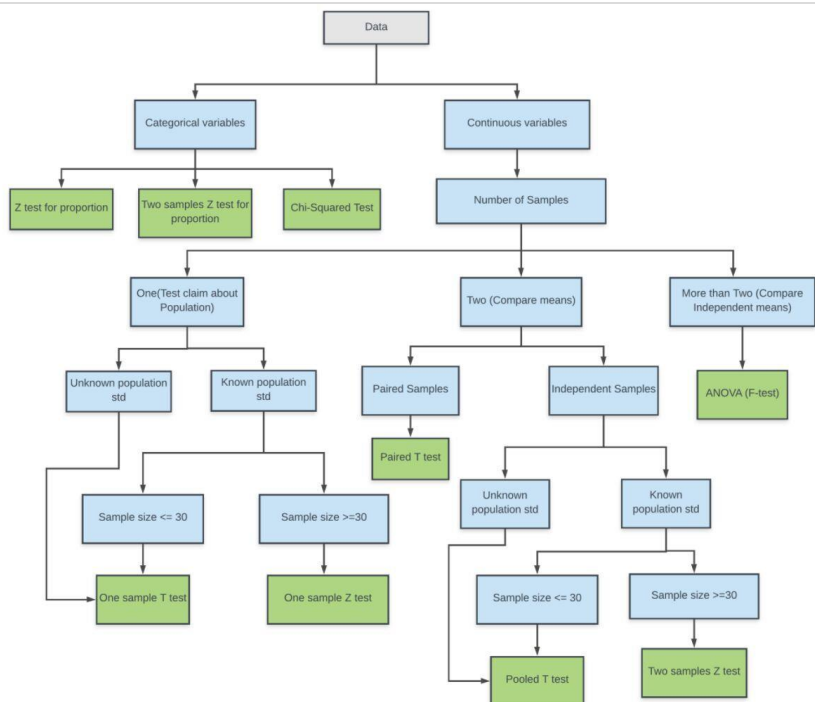
Ki-kare değerimiz olan 4,41 değeri bu 2 değer arasında düşmektedir. Yani 0,05 probability (%95) ile 0,02 probability (%98) arasındadır.

Bu sonuca göre %95 seviyesinde Null hipotez reddedilir, %98 seviyesinde Null hipotez kabul edilir.



WHICH TEST FOR THE PROBLEMS

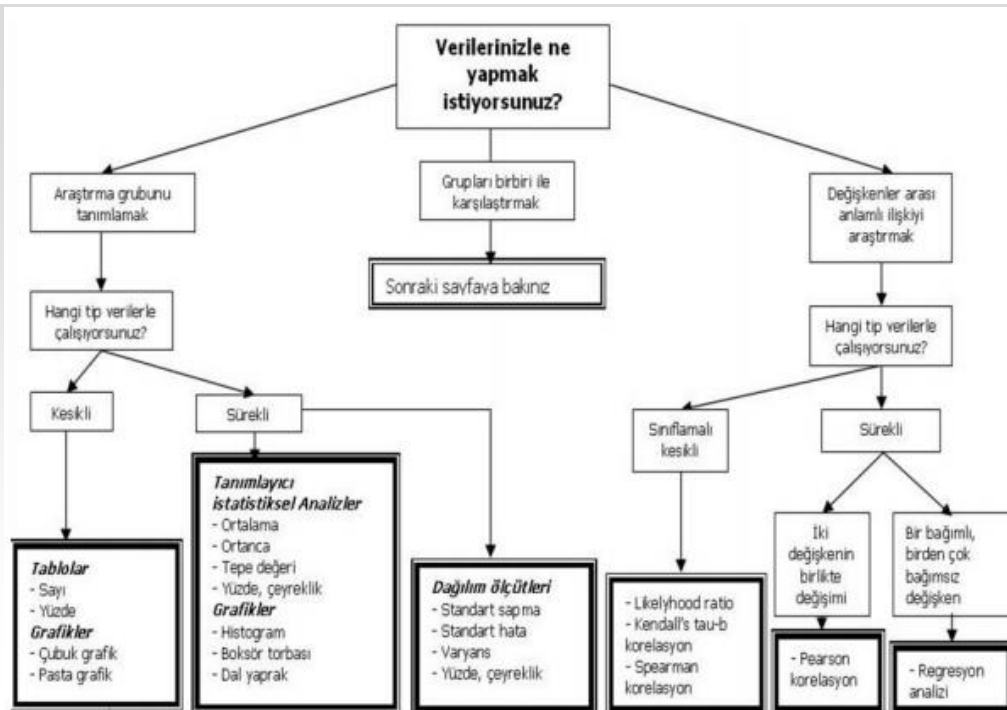
Hangi problem için hangi istatistiksel testi yapabiliriz ?

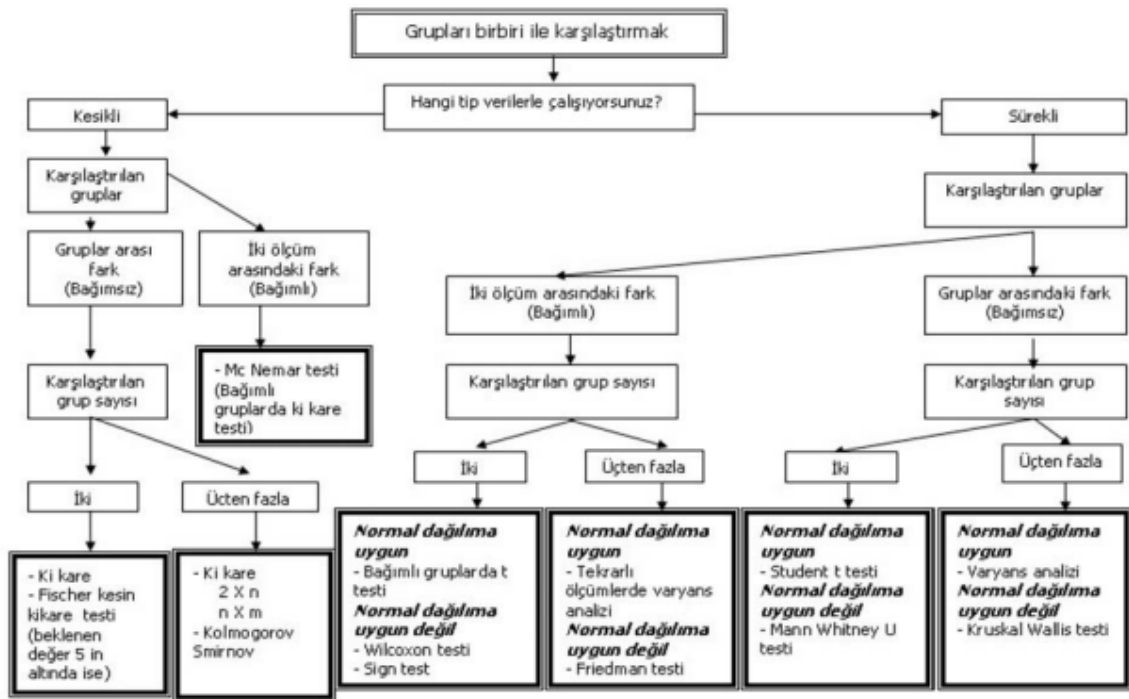




Statistical Tests:

- T-Test:
 - Compare the **Means** between **two** groups
 - **Small** Sample Size
- Z-Test:
 - Compare the **Means** between **two** groups
 - **Large** Sample Size
- ANOVA:
 - Compare the **Means** between **two+** groups
- Chi-Square:
 - Compares **Proportions** between **two** groups





Examples

- Her iki günde de **aynı** 20 istasyonu kontrol ederek, **Çarşamba ve Cumartesi** benzin fiyatları ortalamaları (**mean**) arasında önemli bir fark var mı?
- Önümüzdeki tatil döneminde uçakla seyahat edenlerin yüzdesi (**percentage**), araba ile seyahat edenlerin yüzdesinden (**percentage**) daha fazla olacak mı?
- İstanbuldaki ortalama (**average**) mazot fiyatı ile Karstaki ortalama (average) mazot fiyatı arasındaki anlamlı bir fark var mı?
- Dört** farklı araç tipinin **ortalama** olarak yaptıkları kilometreler arası fark
- İki havayolu şirketinin seyahatinde yaşanan **ortalama** gecikmeler arasında fark var mıdır

Examples

Örnek-2

- Bir trafik memuru bir otoyoldaki güvenlik için sürücülerle alakalı bazı testler yapmak istemektedir. Hangi durum için hangi testi uygulamalıdır ?

1. Yoldaki hız limiti 90 km/h ama memur sürücülerin daha hızlı gittiğini düşünüyor. Bu şüphe doğru mu?

2. Kadın ve erkeklere hızlarını sorup, aralarında bir fark olup olmadığına bakmak istiyor

3. İnsanların Hafta içi ve hafta sonu hızları arasında fark var mıdır ?

4. Sürücülerin 2 yıl önce %60 ının sürüş esnasında telefon kullandığı biliniyor. Memur bunun şimdi daha fazla olduğunu düşünüyor. İspatı ?

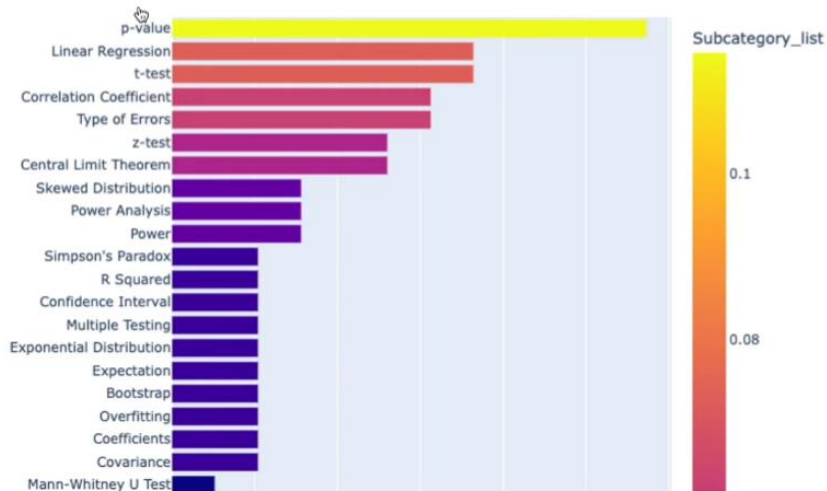
5. Genç-orta yaş ve yaşlı grupları arasında genç sürücülerin daha hızlı gittiğini düşünüyor. İspatı



★ These results are based on 300+ statistics interview questions from 50+ companies.

+ ::

Top Statistics Concepts in Data Science Interviews



Python Coding

› Solution

arsenic_data.ipynb

Python Coding

› chi-square yi anlatan bir notebook

chi-square.ipynb