Veri Bilimi İçin İstatistik

DOÇ. DR. İHSAN HAKAN SELVİ

ÖĞR. GÖR. DR. DENİZ DEMİRCİOĞLU DİREN

13.HAFTA

12.Haftanın Konuları

- P değeri
- Hipotez Testi Adımları
- Parametrik Hipotez Testleri
  - Tek Örneklem Z Testi
    - Tek Anakütle Ortalamasının Z Testi
    - Tek Anakütle Oranının Z Testi

p değeri

p değeri, kurduğumuz hipotezi değerlendirebilmek için kullanacağımız bir değerdir.

$p<0,05$ olduğunda H_0 hipotezi reddedilir

$p>0,05$ ise H_0 hipotezi kabul edilmektedir.

Hipotez Testi Adımları

Adım 1: Hipotezlerin kurulması ve yönlerinin belirlenmesi

Adım 2: Anlamlılık düzeyi () ve tablo değerlerinin belirlenmesi

Adım 3: Test istatistiğinin belirlenmesi ve test istatistiğinin hesaplanması

Adım 4: Hesaplanan test istatistiği ile alfa değerine karşılık gelen tablo değerinin karşılaştırılması

Test istatistiği > Tablo değeri olduğunda H0 reddedilir

Adım 5: Yorumlama

Parametrik Hipotez Testleri

İstatistiksel testlerin türüne karar vermek için dağılıma bakmamız gerekmektedir.

Normal dağılıma uygun verilerin analizi parametrik testlerle yapılmaktadır.

Normal dağılıma uymayan verilerin analizi ise parametrik olmayan testler ile yapılmaktadır.

Parametrik testler

- Z testi
- t testi
- F testi

Z testi

Örneklemin alındığı anakütle varyansının bilindiği ve anakütlenin normal dağılıma uygun olması durumunda standart normal dağılım kullanılarak Z testi uygulanır.

Eşit aralıklı ve oransal veri olduğunda kullanılır.

Ana kütleden elde edilen örnek büyüklüğü 30 veya daha fazla ise (n>=30) için Z testi kullanılır.

Tek örneklem Z testi

İki örneklem Z testi

Tek Örneklem Z testi

Bu test, tek örnek için ana kütle ortalaması ve oranını incelemektedir.

Popülasyon ortalamasının varsayılmış değerden farklı olup olmadığını doğrulamak için kullanılmaktadır.

Tek Örneklem Z testi

Tek örneklem Z testinin uygulaması için kriterler;

- Örneklemin alındığı popülasyon normal olarak dağılması
- Örneklem boyutunun 30’dan büyük olması
- Tek bir örneklem alınması
- Popülasyon ortalamasının test edilmesi
- Popülasyon standart sapmasının bilinmesi

Tek Örneklem Z testi

Tek örneklem Z testi için ortalamalara ait hipotezler ve matematiksel ifadeleri aşağıdaki gibidir:

H_0: $\mu = \mu_0$

H_1: (Tek yönlü sağ kuyruk testi için)

H_1: (Tek yönlü sol kuyruk testi için)

Test istatistiği

$$z = \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$\bar{X}$: Örnek Ortalaması
μ : Ana kütle Ortalaması
σ : Popülasyon Standart Sapması
n : Örnek Büyüklüğü

Örnek Soru

Bir okul, öğrencilerin üniversite sınav puanlarının ortalama puandan yüksek olduğunu iddia etmektedir. Okulda 50 öğrencinin sınav puanlarının ortalaması hesaplanmış olup 110 olarak bulunmuştur. Buna göre sınavın ortalama değeri 100 ve standart sapması 15 olduğuna göre müdürün iddiasının %5 anlamlılık düzeyinde doğru olup olmadığını belirtin.

Cevap: Formül Çözüm

H_0: $\mu = 100$

H_1: $\mu > 100$

$\alpha = 0.05$

$z = \frac{110-100}{15/\sqrt{50}} = 4,71$

$\alpha = 0,05$

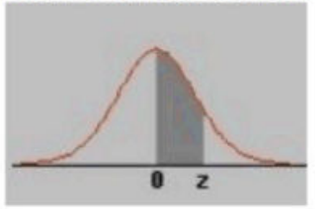
0,5-0,05= 0,45

Z test = 1,65

$Z_{hesap} > Z_{test}$ ise, H_0 reddedilir

Standard Normal (Z) Table

Area between 0 and z



	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545

Cevap: Python Çözüm

Öncelikle çalışmanın hipotezleri oluşturulur.

H_0: $\mu = 100$

H_1: $\mu > 100$

Çalışmada yer alan 50 öğrenci verisi bulunmadığından dolayı ortalaması 110 ve standart sapması 15 olan bir örnek bir veri oluşturulmuş.

Cevap

```
import math
import numpy as np
from numpy.random import randn
from statsmodels.stats.weightstats import ztest

ort_iq=110
ss_iq= 15/math.sqrt(50)
veri=ss_iq*randn(50)+ort_iq
print('ort=%.2f ss=%.2f'%(np.mean(veri), np.std(veri)))
```

```
ort=110.30 ss=1.77
```

Cevap

Aşağıdaki kod yardımı ile, örnek veri için değerlerin dağılımı yandaki gibi gözükmektedir.

```
veri
```

```
array([110.35419869, 108.9662533 , 109.64706145,
107.82839561,
       108.57500393, 112.57917635, 112.37379761,
108.55840863,
       107.93172374, 111.27201103, 109.40576082,
109.09705537,
       110.84115638, 109.88947358, 106.61013292,
110.45165836,
       112.59922868, 109.94595398, 111.9784853 ,
111.97259794,
       110.69707258, 108.15188769, 107.64538646,
111.29988367,
       112.87595484, 109.13152105, 112.67180908,
111.49877509,
       112.05576116, 109.89326779, 110.38643443,
111.13649811,
       109.95547799, 112.57323447, 110.25704531,
108.32556246,
       112.29063244, 109.66336107, 107.40785517,
106.85931338,
       110.31544413, 110.68062612, 114.93680088,
111.05047682,
       113.48367667, 109.91819503, 110.32115872,
109.33592914,
       110.37547734, 109.2382674 ])
```

Cevap

Z-testinin yapılması ‘ztest (veri, value = 100, alternative = ‘large’)’ modülü kullanılmaktadır.

‘value’ değeri karşılaştırma yapılacak değerdir.

Ayrıca alternatif hipotezde belirtilen testin yönü büyüktür olduğu için alternative = ‘larger’ şeklinde belirtilir.

Cevap

```
ztest_Score, p_value = ztest (veri,value=100,alternative='larger')
ztest_Score, p_value
```

```
(33.66644551796583, 8.958391720870862e-249)
```

```
alfa = 0.05
if(p_value < alfa):
   print('Boş hipotezi reddet')
else:
   print('Boş hipotez reddedilmez')
```

```
Boş hipotezi reddet
```

Cevap

Boş hipotez olan örnek ortalamasının 100 olması %5 önem seviyesinde ret edilir ve alternatif hipotez kabul edilir.

Yani müdürün iddia ettiği öğrencilerin sınav puanlarının yüksek olduğu iddiası doğrudur.

Örnek

Bir kurye şirketinin ortalama teslimat süresi 45 dakika ve standart sapması 5 dakikadır. Şirket yeni araçlar alarak teslimat süresini düşürmek istemektedir. Araç alımlarından sonra 40 sipariş incelenip ortalama teslimat süresi 44 dakika bulunmuştur. Yeni araçlarla teslimat sürelerinin azalacağı iddiasını %1 anlamlılık düzeyinde inceleyiniz.

Cevap

Hipotez testi için öncelikle boş ve alternatif hipotez μ kuryenin teslimat süresi olmak üzere tanımlanmıştır.

Boş hipotez: $H_0 : \mu = 45$ (Teslimat süresinde değişiklik olmamıştır)

Alternatif hipotez: $H_1 : \mu < 45$ (Teslimat süresi düşmüştür) (Tek yönlü test)

Uygulamada anlamlılık düzeyi α = 0.01 için örnek sayısı: 1, örnek boyutu: n = 40 ve örnek ile popülasyon ortalaması, μ = 45 st edilmek istenmektedir.

Burada popülasyon standart sapması, σ = 5 bilindiğinden tek örneklem Z testi kullanılır.

Cevap

Öncelikle test için örneklem verileri bulunmadığından örnek bir veri elde edilmiştir.

```
import math
import numpy as np
from numpy.random import randn
from statsmodels.stats.weightstats  import ztest

ort = 44
sh = 5/math.sqrt(40)
data = sh*randn(40)+ort
data
```

```
array([44.9215998 , 42.91345698, 42.92511334, 45.11288996, 43.66917866,
       43.73500242, 44.28155736, 43.18477801, 43.95724439, 44.24535755,
       43.79764803, 43.77456535, 44.46751588, 45.21859472, 43.11562744,
       44.54753134, 44.48078641, 44.42612826, 44.25282044, 44.28788988,
       43.42599457, 43.33709172, 44.14701152, 43.86085987, 45.59480083,
       43.72957274, 43.96008418, 42.85929839, 43.39738828, 44.74378395,
       44.8422737 , 45.54249649, 44.9422983 , 44.79644079, 44.34069234,
       43.25825001, 46.43096297, 43.84155789, 44.49163431, 43.61732253])
```

Cevap

Örnek verileri için Z testi uygulamasında alternatif hipotez küçüktür olduğundan dolayı modülde ‘alternative=>smaller>’ ve test değeri ise yüz olduğundan ‘value=45’ olarak girilmiştir.

```
ztest_değeri, p_değeri = ztest (data,value=45,alternative='smaller')
ztest_değeri, p_değeri
```

```
(-6.614871599161315, 1.8593719128919458e-11)
```

```
alfa = 0.01
if(p_değeri < alfa):
   print('Boş hipotezi reddet')
else:
   print('Boş hipotez reddedilmez')
```

Boş hipotezi reddet

Cevap

Elde edilen test sonucuna göre boş hipotez ret edilir ve alternatif hipotez kabul edilir.

Yani teslimat sürelerinde %1 anlamlılık düzeyinde azalma meydana gelmiştir.

Tek Anakütle Oranının Z testi

Anakütle oran testi, anakütlede bulunan özelliğinin oranı (p) ve anakütleden alınan örnekte aynı özelliğin rastlanma oranı (p') olmak üzere bunlar hakkında karar vermek için kullanılır.

Oran Z-testi sadece iki kategori varken, gözlemlenen oran ile beklenen oranın karşılaştırılmasında kullanılır.

Oran testleri nominal verilerle kullanılır ve yüzdeleri veya oranları karşılaştırmak için kullanışlıdır.

Tek Anakütle Oranının Z testi

Test istatistiği;

$$Z = \frac{p' - p}{\sqrt{(p(1-p))/n}}$$

p' : Örnek özellik oranı

p: Anakütle özellik oranı

n: Örnek büyüklüğü

Örnek Soru

Bir firma müşterilerinin %40’ının hizmetlerden memnun olduğunu iddia etmektedir. Firma şikayetler arttığı için müşteri memnuniyeti oranını değerlendirmek amacıyla bir anket yapmaya karar verir. Yapılan anket sonucunda 100 müşteriden 30’unun memnun olduğu tespit edilmiştir. Son durumda müşterilerin memnuniyet düzeyinde değişiklik olup olmadığını %5 anlamlılık düzeyinde test ediniz.

Cevap

Öncelikle boş ve alternatif hipotez yazılmalıdır. Ortalama müşteri memnuniyeti oranı p olsun;

H_0: p = 0.4

H_1: p ≠ 0.4

Eşitlik işareti çift kuyruklu bir test olduğunu göstermektedir. Anlamlılık düzeyi α = 0.05'tir. Örnek boyutu n = 100 için popülasyon oranındaki değişiklik test edilmektedir. Bundan dolayı; oranlar için tek örnekli Z testi uygulanmıştır. İlgili test istatistiğini ve p değerini bulalım.

Cevap

$$Z = \frac{p' - p}{\sqrt{(p(1-p))/n}}$$

Burada, p' = 0.3, p = 0.4 n=100

Cevap

Scipy.stats kütüphanesinde "stats.norm.cdf()" modülü ile test istatistiği değeri hesaplanmıştır.

```
import scipy.stats as stats
z=(0.3-0.4)/((0.4)*(1-0.4)/100)**0.5
z
```

```
-2.041241452319316
```

```
p=stats.norm.cdf(z)
p
```

```
0.020613416668581 79
```

Test istatistiği olan p-değeri=0,02<0,05 olduğundan boş hipotez ret edilir. Yani, %5 önem düzeyinde, müşterilerin firmanın hizmetlerinden memnuniyet yüzdesi değişmiştir.

Örnek Soru

Bir çiftlikte yarısı erkek yarısı dişi ineklerden oluşan bir popülasyon vardır (p=0.5=%50). Bazı ineklerde (n=200) bilinmeyen bir sebepten dolayı kanser gelişmiştir. Kanser gelişen ineklerin 120 tanesi erkek, 80 tanesi dişidir.

- Erkek ineklerin gözlemlenen oranı (p_0)=120/200=0.6
- Dişi ineklerin gözlemlenen oranı (q)=80/200=0.4
- Erkek ineklerin beklenen oranı (p_e)=0.5
- Gözlem sayısı (n)=200

Örnek Soru

İneklerde meydana gelen kanser vakası ile ilgili erkek kanserli inekler için 3 farklı durum incelenmek istenmiştir. Burada dikkat edilmesi gereken 2. ve 3. durum için hipotezlerin ters kurulmasıdır. Uygulamada 2. durum olmaması gerekirken örnek olması adına incelenmiştir.

1. Durum: erkeklerin gözlemlenen oranı beklenen orana eşittir.

2. Durum: erkeklerin gözlemlenen oranı beklenen orandan küçüktür.

3. Durum: erkeklerin gözlemlenen oranı beklenen orandan büyüktür.

Cevap

Öncelikle incelenecek durumlar için boş ve alternatif hipotezler belirlenmiştir.

1. Durum: H_0: $p_0 = p_e$ ve H_1: $p_0 \neq p_e$ (Çift yönlü test)

2. Durum: H_0: $p_0 = p_e$ ve H_1: $p_0 < p_e$ (Tek yönlü test)

3. Durum: H_0: $p_0 = p_e$ ve H_1: $p_0 > p_e$ (Tek yönlü test)

Cevap

Python’da oran testi yapmak için statsmodels kütüphanesi içindeki "proportions_ztest" modülü kullanılmaktadır. Uygulamada öncelikle gerekli olan kütüphaneler yüklenmiştir ve "proportions_ztest" modülü içe aktarılmıştır.

```
import pandas as pd
from statsmodels.stats.proportion import proportions_ztest
```

Cevap

Uygulamada öncelikle 1. Durum olan erkeklerin gözlemlenen oranının beklenen orana eşit olduğu iddiası test edilmiştir. "alternative" parametresi ile alternatif hipotezin yönü belirtilmektedir. Çift taraflı test yapmak istenildiğinden dolayı modülde "alternative=two-sided" değeri girilmiştir.

```
count=120 #kanserli erkek inek
nobs=200 #toplam kanserli inek
value=0.5 #test edilmek istenilen oran
ztest_Score, p_value=proportions_ztest(count, nobs, value, alternative="two-sided")
ztest_Score, p_value #1. Durum
```

```
(2.886751345948128, 0.0038924171227786367)
```

Bu sorguda yer alan modülde "count" parametresi ile başarılı deneme sayısını, "nobs" parametresi ile gözlem sayısı, "value" parametresi ile test etmek istenilen oranı belirtmektedir. Bu oran burada erkek ve dişi ineklerin eşit olma durumu olan 0.5'tir.

Cevap

```
alpha = 0.05
if(p_value < alpha):
    print("Boş Hipotezi Reddet")
else:
    print("Boş Hipotez Ret Edilemez")
```

```
Boş Hipotezi Reddet
```

1. Durum için boş hipotez ret edilerek alternatif hipotez kabul edilir. Yani erkeklerin gözlemlenen oranı beklenen orana yani 0.5'e %5 anlamlılık düzeyinde eşit değildir.

Cevap

2. Durum için erkeklerin gözlemlenen oranının beklenen orandan (0.5) küçük olduğu iddiası test edilmiştir. Tek taraflı test yapmak istenildiğinden dolayı alternatif hipotez küçüktür olduğundan modülde "alternative=smaller" değeri girilmiştir.

```
ztest_Score, p_value=proportions_ztest(count, nobs, value, alternative="smaller")
ztest_Score, p_value #2. Durum
```

```
(2.886751345948128, 0.9980537914386107)
```

NOT: Bu durum sorunun başında belirtildiği gibi aslında 3.hipotezle çelişmektedir. Böyle bir hipotez kurulmamalıdır. Ancak ‘smaller’ alternatif hipotez kodunu gösterebilmek için konulmuştur.

Cevap

```
alpha = 0.05
if(p_value < alpha):
    print("Boş Hipotezi Reddet")
else:
    print("Boş Hipotez Ret Edilemez")
```

```
Boş Hipotez Ret Edilemez
```

2. Durum için boş hipotez ret edilemez. Yani erkeklerin gözlemlenen oranı beklenen orana yani 0.5'e %5 anlamlılık düzeyinde eşittir.

Cevap

3. Durum için erkeklerin gözlemlenen oranının beklenen orandan (0.5) büyük olduğu iddiası test edilmiştir. Tek taraflı test yapmak istenildiğinden dolayı alternatif hipotez büyüktür olduğundan modülde "alternative=larger" değeri girilmiştir.

```
ztest_Score, p_value=proportions_ztest(count, nobs, value, alternative="larger")
ztest_Score, p_value #3. Durum
```

```
(2.886751345948128, 0.0019462085613893183)
```

Cevap

```
alpha = 0.05
if(p_value < alpha):
    print("Boş Hipotezi Reddet")
else:
    print("Boş Hipotez Ret Edilemez")
```

```
Boş Hipotezi Reddet
```

3. Durum için boş hipotez ret edilerek alternatif hipotez kabul edilir. Yani erkeklerin gözlemlenen oranı beklenen orandan yani 0.5'ten %5 anlamlılık düzeyinde büyüktür.

Yardımcı Kaynaklar

Sel, A., «Python Uygulamalı İstatistiksel Veri Bilimi ve Analizi», Akademisyen Kitabevi

Gayathri Rajagopalan - A Python Data Analyst’s Toolkit_ Learn Python And Python-based Libraries With Applications In Data Analysis And Statistics-Apress (2021).pdf