

Eşdizimlilik (Collocation)

Prof.Dr. Banu Diri

Eşdizimlilik(Collocation) Nedir ?

- İki veya daha fazla kelimenin bir araya gelerek farklı bir şeyi ifade etmesidir (*ağır abi*).
- Kelimeler birlikte kullanıldıklarında daha farklı anlamlar içerebilirler (*disk drive, hot dog, mother in law*).
- Kelimeler üzerinde çalışırken onları bulundukları bağlamdan bağımsız olarak düşünmek imkansızdır. Kelimeler bağlam içerisinde çıplak anlamlarından farklı anlamlar alabilir.
- Metin içerisinde neyin kaç kez görüldüğündense neyle beraber görüldüğü önemlidir (*Türkiye Büyük Millet Meclisi, Türk Hava Yolları, vs.*).

Eşdizimlilik iki ilkeye sahiptir.

- Açık seçim ilkesi
- Deyim ilkesi

Açık seçim ilkesi : Birbirleriyle bağlantılı kelimelerin seçiminde herhangi bir zorlanma yoktur (Mavi gökyüzü).

Deyim ilkesi : Kelimelerin ayrı ayrı anlamlarından farklı bir anlam çıkarılır (Tefe koymak).

- İngilizceden örnek

- noun phrases *strong tea* not *powerful tea*
- phrasal verbs *to make up* and *the rich and powerful*

- Geçerli bir eşdizimlilik mi (collacation) ?

- *a stiff breeze* (sert esen rüzgar) but not *a stiff wind*
(*a strong breeze* or *a strong wind* is okay)
- *broad daylight* (güpegündüz) (but not *bright daylight* or *narrow darkness*)

Eşdizimlilik(Collocations) kriterleri

Eşdizimlilik sınırlı sayıda kelime ile karakterize edilir.

- Eşdizimlilikte 3 farklı kriter vardır.
 - non-compositionality (bir araya getirilemez)
 - non-substitutability (yeri değiştirilemez)
 - non-modifiability (değiştirilemez)
- Eş dizimlik hiç bir zaman bir dilden diğer dile kelime kelime tercüme edilemez.
- Eşdizimlilik için kelimeler arka arkaya gelmek zorunda değildir (*knock . . . door*).

Non-Compositionality

- Kelimelerin herbirinin anlamından birleştirilmiş ifadenin anlamı tahmin edilebiliyorsa bu ifade **compositional**'dır.
 - new companies
- Kelimelerin herbirinin anlamından birleştirilmiş ifadenin anlamı tahmin edilemiyorsa bu ifade **non-compositional**'dır.
 - hot dog
- Kelimelerin herbirinin anlamından birleştirilmiş ifadenin anlamı yakın olarak tahmin edilebilir.
 - *strong tea, powerful drug, not powerful tea*
- non-compositional için en uç örnekler **deyimler**dir.
 - *"it rains cats and dogs", "etekleri zil çalmak"*

Non-Substitutability

- Collocation'nın bir elemanı olarak yakın anlamlı (near-synonyms) bir kelimeyi kullanamayabiliriz.
 - Beyaz şarabın rengini iyi tanımlasa bile *white wine* yerine *yellow wine* kullanılamaz
- Collocation'ların çoğu gramatik olarak bir dönüşüm veya ek bir kelime ile yeniden düzenlenemezler (**Non-modifiability**).
 - **white wine**, but not **whiter wine**
 - **mother in law**, but not **mother in laws**

Collocation'da alt sınıflar

- Light verbs
 - *make, take* ve *do* gibi fiillerin kullanımı
 - *make lunch, take easy*
- Fiil Edat yapıları
 - *to go down*
- Özel isimler (proper nouns)
 - *Mustafa Kemal Atatürk*
- Teknik terimler, teknik alandaki nesne ve kavramlar
 - *Hidrolik yağ filtresi (Hydraulic oil filter)*

Collocation'ları bulmak için genel yaklaşım

Bir text içerisinde yer alan collocation'lar nasıl bulunur ?

- En basit method: *Frekans*'a dayalı collocation seçimi
- Eşdizimliliği oluşturan kelimeler arasındaki uzaklığın ortalama ve varyansına dayalı seçim (**mean and variance**)
- **Hipotez testi (Hypothesis testing)**
- **Karşılıklı bilgi (Mutual information)**

Frekans yaklaşımı (Frequency)

- Meydana gelme sıklığına göre collocation'ın bulunması.
- Size window'a ihtiyaç vardır.
- Döndürülen sonuçlar içerisinde Function word'ler (stop words) olabilir. Bunların filtrelenmesi gerekir.
- Bu filtreden geçen yapılar collocation'a adaydır.

$C(w^1 w^2)$	w^1	w^2
80871	of	the
58841	in	the
26430	to	the
21842	on	the
21839	for	the
18568	and	the
16121	that	the
15630	at	the
15494	to	be
13899	in	a
13689	of	a
13361	by	the
13183	with	the
12622	from	the
11428	New	York
10007	he	said
9775	as	a
9231	is	a
8753	has	been
8573	for	a

Örnek Corpus'daki en sık kullanılan bigram'lar (biword) çıkarılır

New York hariç, listedeki bigram'ların hepsi function word'dür

Tag Pattern	Example	A : adjective (sifat)
A N	<i>linear function</i>	N : noun (isim)
N N	<i>regression coefficients</i>	P : preposition (edat)
A A N	<i>Gaussian random variable</i>	
A N N	<i>cumulative distribution function</i>	
N A N	<i>mean squared error</i>	
N N N	<i>class probability function</i>	
N P N	<i>degrees of freedom</i>	

Part of speech tag patterns for collocation filtering
(Justesen and Katz).

$C(w^1 w^2)$	w^1	w^2	tag pattern
11487	New	York	A N
7261	United	States	A N
5412	Los	Angeles	A N
3301	last	year	A N
3191	Saudi	Arabia	A N
2699	last	week	A N
2514	vice	president	A N
2378	Persian	Gulf	A N
2161	San	Francisco	N N
2106	President	Bush	N N
2001	Middle	East	A N
1942	Saddam	Hussein	N N
1867	Soviet	Union	A N
1850	White	House	A N
1633	United	Nations	A N
1337	York	City	N N
1328	oil	prices	N N
1210	next	year	A N
1074	chief	executive	A N
1073	real	estate	A N

Eğer collocation'ı
oluşturan kelimeler
arası sabit ise Frekans
tabanlı yöntem iyi sonuç
verir.

Önce $C(w^1 w^2)$ a filtre
uygulandıktan sonra,
geride kalan en yüksek
kullanım sıklığına sahip
ifadeler

w	C (strong,w)	w	C(powerful,w)
support	50	force	13
safety	22	computers	10
sales	21	position	8
opposition	19	man	8
showing	18	computer	8
sense	18	man	7
message	15	symbol	6
defense	14	military	6
gains	13	machines	6
evidence	13	country	6
criticism	13	weapons	5
possibility	11	post	5
feelings	11	people	5
demand	11	nation	5
challenges	11	forces	5
challenge	11	chip	5
case	11	Germany	5
supporter	10	senators	4
signal	9	neighbor	4
man	8	magnet	4

Strong challenge, powerful computer

Not powerful challenge, strong computer

Collocational Window

Çoğu collocation farklı değişken uzunluklarda bulunabilir.

Bu tip collocation'ların bulunmasında *Frekans Tabanlı* yaklaşımlar kullanılmaz.

she knocked on his door	distance=3
they knocked at the door	distance=3
100 women knocked on Donaldson's door	distance=5
a man knocked on the metal front door	distance=5

Sentence: *she knocked on his door*

Bigrams:

<i>she knocked</i>	<i>she on</i>	<i>she his</i>	
	<i>knocked on</i>	<i>knocked his</i>	<i>knocked door</i>
		<i>on his</i>	<i>on door</i>
			<i>his door</i>

3 kelimelik collocation window kullanılarak bigram'lar çıkarılır.

Genelde 3, 4 kelimelik window'lar kullanılır.

Mean and Variance

Knocked and *door* arasındaki ilişkiyi keşfetmenin bir yolu, corpus içerisinde yer alan iki kelime arasındaki ofsetin (işaretli uzaklık) *mean (ortalama)* ve *variance (varyans)* hesaplamaktır.

Ortalama(mean= μ), iki kelime arasındaki ofsetin ortalamasıdır.

she **knocked** on his **door**

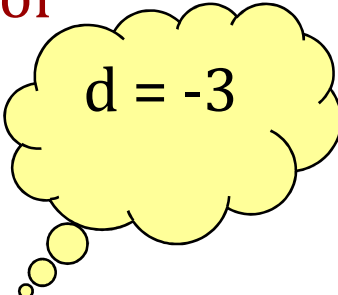
they **knocked** at the **door**

100 women **knocked** on Donaldson's **door**

a man **knocked** on the metal front **door**

Mean ?

$$\mu = \frac{1}{4} (3+3+5+5) = 4.0$$


$$d = -3$$

Bazen distance negatif bir sayı olabilir. *The door that she knocked on*

Mean and Variance

- Varyans : Değerlerin ortalamanın çevresindeki dağılımını ölçmek için kullanılan bir niceliktir. Ortalamanın örneklem değerlerinden çıkarılmasıyla bulunan sapmaların karelerinin ortalaması alınarak hesaplanır.

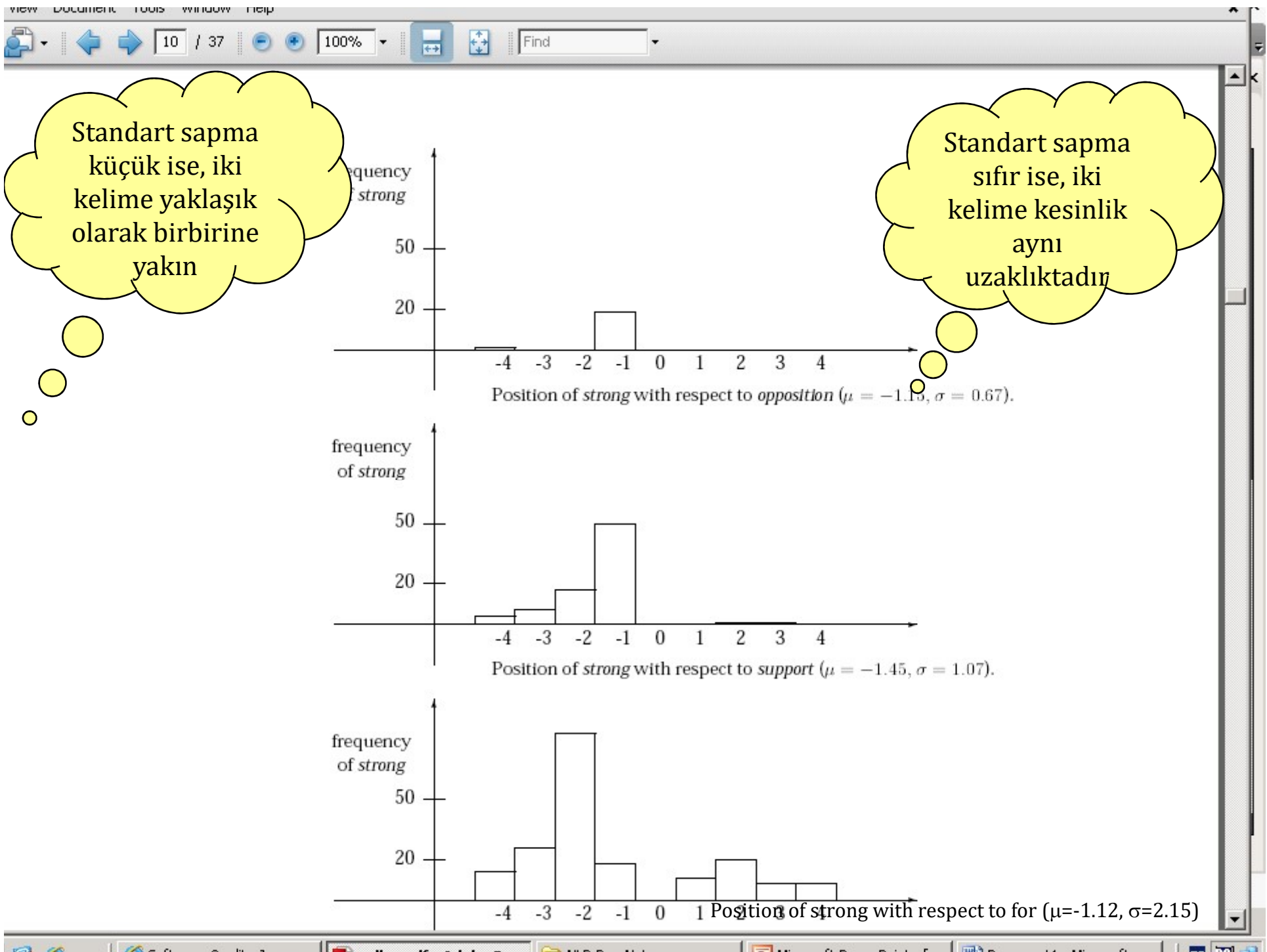
- $\sigma^2 = s$

$$\sigma^2 = \frac{\sum_{i=1}^n (d_i - \mu)^2}{n - 1}$$

- n iki kelimenin birlikte kullanılma sayısı, d_i i. birlikte görülmenin uzaklık değeri, ve μ ortalama

-
- Ortalama ve varyans, iki kelime arasındaki mesafenin dağılımını karakterize eder.
 - Yüksek varyansın anlamı, birlikteliklerin şans eseri gerçekleştiğidir.
 - Düşük varyansın anlamı, birlikteliklerin aynı uzaklıklara sahip olduğudur.

$$\sigma = \sqrt{\frac{1}{3} \left((3 - 4.0)^2 + (3 - 4.0)^2 + (5 - 4.0)^2 + (5 - 4.0)^2 \right)} \approx 1.15$$



Ortalama ve varyansa dayalı Collocation'ların bulunması

s	$\bar{d}$	Count	Word 1	Word 2
0.43	0.97	11657	New	York
0.48	1.83	24	previous	games
0.15	2.98	46	minus	points
0.49	3.87	131	hundreds	dollars
4.03	0.44	36	editorial	Atlanta
4.03	0.00	78	ring	New
3.96	0.19	119	point	hundredth
3.96	0.29	106	subscribers	by
1.07	1.45	80	strong	support
1.13	2.57	7	powerful	organizations
1.01	2.00	112	Richard	Nixon
1.05	0.00	10	Garrison	said

▪ σ küçük, μ 1'e yakın ise NY frekans tabanlı yöntem ile bulunur.

▪ σ küçük, μ , 1'den büyük ise üzerinde durulması gereken ilginç bir durumdur.

The pair *previous / games* (distance 2) corresponds to phrases like *in the previous 10 games* or *in the previous 15 games*; *minus / points* corresponds to phrases like *minus 2 percentage points*, *minus 3 percentage points* etc; *hundreds / dollars* corresponds to *hundreds of billions of dollars* and *hundreds of millions of dollars*.

▪ Eğer σ çok büyük ise bu kelime çiftleri ile ilgilenilmez.

▪ *strong {business} support, powerful {lobbying} organizations, Richard*

{M.} Nixon, and Garrison said / said Garrison (remember that we tokenize *Richard M. Nixon* as four

Şansın bertaraf edilmesi...

- İki kelime şans eseri birlikte olabilir.
 - Frekansı yüksek ve varyansı düşük ise
- **Hipotez Testini (Hypothesis Testing)** kullanarak bu birlikteliğin gerçek mi yoksa şans eseri mi olduğu ölçümlenebilir.



Hipotez Testleri

Dr. İrfan Yolcubal
Kocaeli Üniversitesi
Jeoloji Müh.



Hipotez

- Örneklemeye dayalı bir popülasyon parametesinin değeri hakkında ileri sunulan iddia
Örnekler:
 1. İstatistik Vize sınavının ortalaması 50'nin altındadır.
 2. Televizyon izleyicilerin %70 i günlük haber programlarını izlemektedir.
 3. Firestone ve Lassa tarafından üretilen lastiklerinin ömrü aynıdır.

Hipotez Testleri

- Bir popülasyon hakkında ileri sunulan hipotezinin kabul edilip edilmeyeceğini belirlemek için örnekleme dayalı sistematik izlenen bir seri işlemler.

5 aşamadan oluşur.

1. Null ve alternatif hipotezin belirlenmesi

Null hipotezi: Bir popülasyon parametresi hakkında ileri sürülen varsayım. Genellikle bu varsayımda popülasyon parametresinin belli bir değeri olduğu varsayılır.

- H_0 = null hipotezi yada sıfır hipotez

Alternatif hipotez: Örnekleme ait veriler null hipotezonin yanlış olduğuna ait deliller sunduğu durumlarda kabul edilen hipotezdir

- H_A = alternatif hipotez

Hipotez Testinin Aşamaları

2. Önem veya Risk Derecesinin

Belirlenmesi(α): Aslında doğru olan Null hipotezinin rededilme olasılığı:

Risk derecesinin seçimi tercihe dayalı

- Genelde 0.05 yani % 5 ve % 1 risk dereceleri araştırmalarda kullanılmakta



Hata Tipleri

- **I. tip hata:** Null hipotezi doğru iken reddedilir.
- I. Tip hata yapma olasılığı α olarak bilinmektedir.
- **II. tip hata:** Null hipotezi yanlış iken rededilmez.
- II. Tip hata yapma olasılığı β olarak bilinmektedir.
- Daima bu hatalardan birini yapma ihtimali vardır. Bu ihtimalleri risk derecesini belirleyerek azaltmak isteriz.



Hipotez Testlerinde Hatalar

Karar	Null Hipotezi doğru	Null Hipotezi Yanlış
Null hipotezi Kabul etme	Doğru Karar	I. tip hata
Null hipotezi redetme	II. tip hata	OK



Hipotez Testinin Aşamaları

3. İstatistiksel test metodunun belirlenmesi: Null hipotezin rededilip edilmeyeceğinin belirlenmesinde kullanılan ve popülasyon örneklemeinden elde edilen değer

örnek: t , F , ve χ^2 kare istatistik testleri

4. Null hipotezinin hangi koşullarda kabul ve hangi koşullarda rededileceğinin belirlenmesi

5. Karar verilmesi: Null hipotezinin alınan risk derecesi doğrultusunda reddi yada kabülü.

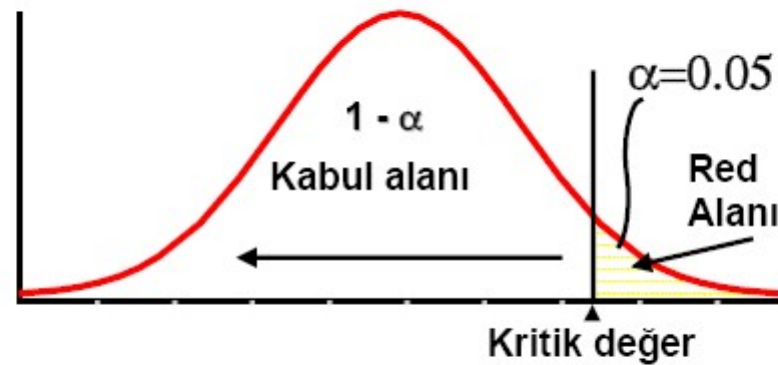


1: Null ve Alternatif hipotezleri ileri sürmek

- Farzedelim öğrencilerin ders geçmek için 60 almaları gerekmekte.
- Rastgele 40 öğrenci secelim ve onların ortalamalarının 64 olduğunu varsayalım
- Araştırma sorusu: Popülasyonun gerçek ortalaması 60 in üzerinde midir?
 - $H_0: \mu \leq 60$
 - $H_A: \mu > 60$

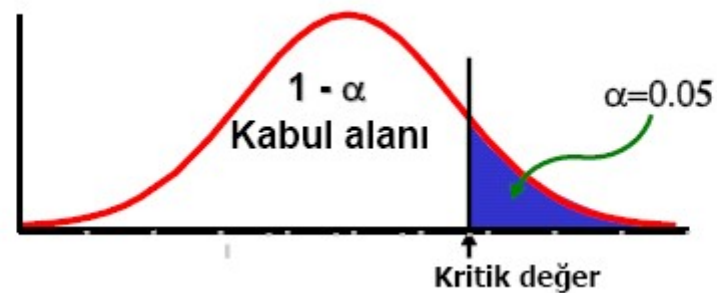
2: Önem Derecesini(α) belirlemek

- Önem derecesi, null hipotezi gerçekten doğru iken, null hipotezini redetme olasılığıdır
- Örnek: $\alpha = 0.05$ seçelim



3: Hipotez testinin 1 veya 2 yönlü olup olmadığının belirlenmesi

- Eğer alternatif hipotez ortalamanın belli bir değere eşit yada ondan büyük olduğunu ifade ediyor ise hipotez tek yönlüdür.
- Örnek: $H_A: \mu \geq 60$



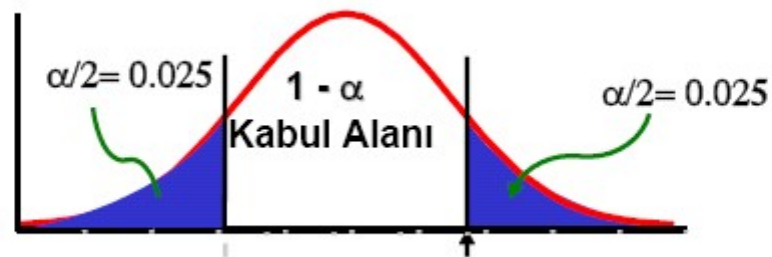
3: (Devam)

- Eğer alternatif hipotez ortalamasının belli bir değere eşit yada ondan küçük olduğunu ifade ediyorsa, hipotez tek yönlüdür. Örnek: $H_A: \mu \leq 60$



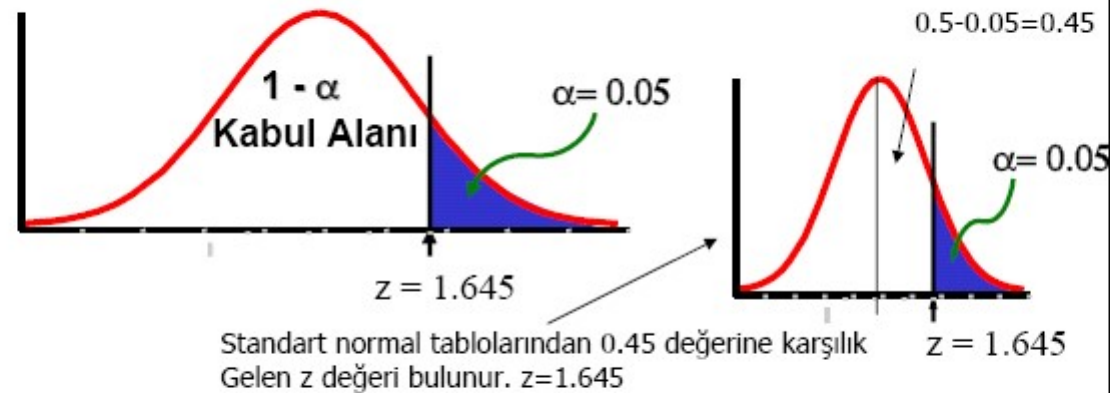
3: (Devam)

- Eğer Alternatif hipotez ortalamasının belli bir değere eşit olmadığını ifade ediyor ise bu hipotez çift yönlüdür. $H_A: \mu \neq 60$



4: Kritik değeri veya değeleri belirlemek

- Bilinmek istenilen – Null hipotezi doğru varsayarsak dağılımın $1-\alpha$ yüzdesine karşılık gelen kritik değer.
- Eğer popülasyonun standart sapması (σ) biliniyor ise yada σ bilinmiyor fakat $n \geq 30$ ise standart normal tabloları kullanılarak risk derecesine karşılık gelen z kritik değeri belirlenir.



5: Test istatistiğini belirlemek ve kritik değerle karşılaştırmak

Popülasyonun standart sapması biliniyor ise z = kritik değer sağdaki formül vasıtasıyla hesaplanır.

$$z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

$\bar{X}$ = örnekleminin ortalaması

μ = populasyon ortalaması

σ = populasyonun standart sapması

- Popülasyonun standart sapması bilinmiyorsa ve $n \geq 30$, örnekleminin standart sapması (s) popülasyonun standart sapması yerine kullanılabilir.
- Populasyon normal dağılım sergilemekte
- Hipotez testinde kullanacak değer:

$$z = \frac{\bar{X} - \mu}{s / \sqrt{n}}$$



Hipotez Test Aşamalarını Özetlersek

1. Null ve Alternatif Hipotezleri Belirlemek: H_0 , H_a
2. Önem yada Risk Derecesini Belirlemek: α
3. Hipotezin tek mi çift mi yönlü olduğunu belirlemek
4. Kritik değerleri belirlemek
5. Test istatistik değerlerini hesaplamak ve kritik değerle karşılaştırmak



Örnek 1

$$H_0: \mu = 50$$

$$H_1: \mu \neq 50$$

Örnek ortalaması 49, örneklemedeki veri sayısı da 36dır. Popülasyonun standart sapması ise 5 dir. Hipotez testinde % 5 risk alırsak

- Hipotez testi tek mi yoksa çift mi yönlüdür
- Null hipotezi hakkındaki kararınız nedir
- Bu kararı almakta nekadar kendinize güveniyorsunuz yani p değeri nedir.

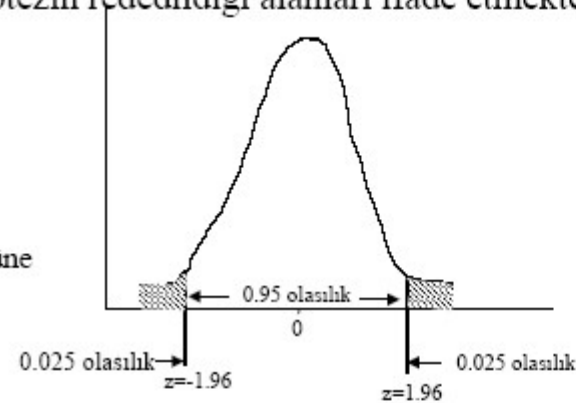
Örnek 1. Çözüm

a) Hipotez testi iki taraflı bir hipotezdir çünkü alternatif hipotezin yönü yoktur yada belli değildir. Popülasyon ortalaması 50 den farklı olabilir ifadesi büyükte olabilir ve küçükte olabilir gibi 2 ihtimal içermektedir. Bu nedenle hipoteze 2 taraflı hipotez denilmektedir.

b) %5 riskle taraflı alanlar hipotezin rededildiği alanları ifade etmektedir.

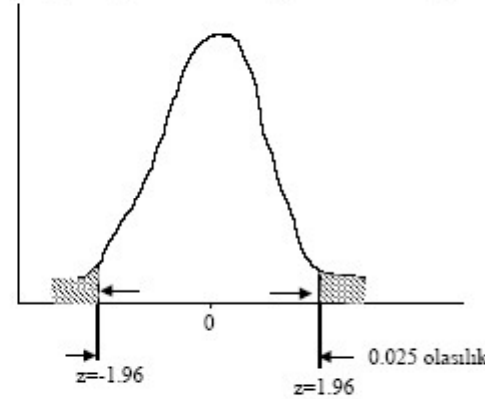
$$z = \frac{49 - 50}{\frac{5}{\sqrt{36}}} = -1.2$$

Hesaplanan z değeri bu taraflı alanlar dışında kalan bölgeye düştüğüne göre Null hipotezini kabul edebiliriz



Örnek 1. Çözüm (Devam)

c) Null hipotezini kabul etmede ne kadar eminiz ? Bunu belirleye bilmek için hesaplanan z değerinin 0 değerinin üzerinde bulunma olasılığını yani p değerini hesaplamamız gerekecektir.



-1.2 ve altında bir değer olma olasılığı $0,1151$ dir ($0.5-0.3849$). p değerini hesaplayabilmek için z değerinin -1.2 den az ve 1.2 den fazla olma ihtimalini hesaplamamız gerekmektedir çünkü hipotez iki taraflı olup iki farklı red bölgesi içermektedir. Bu nedenle p değeri $2 \times 0,1151$ dir. p değeri risk derecesinden 0.05 büyük olduğundan null hipotezi kabul edilir. p değeri popülasyonun ortalamasının 50 nin üzerinde veya altında olma olasılığının $\%11.51$ olduğunu ifade eder.



Örnek 2: Tek yönlü z testi

- Bir kutu mısır gevreği 368 gramın üzerinde midir?
- Rastgele seçilen 25 kutunun ortalaması $\bar{X} = 372.5$ gr.
- Üretici firma ürün miktarı için standart sapmayı σ 15 gram olarak belirlemiştir.
- Hipotezi 0.05 önem derecesi ile test edelim.



Tek yönlü hipotez test çözümü

Test İstatistiği:

$$H_0: \mu \leq 368$$

$$H_A: \mu > 368$$

$$\alpha = 0.05$$

$$n = 25, \sigma \text{ bilinmemekte}$$

Kritik değerler

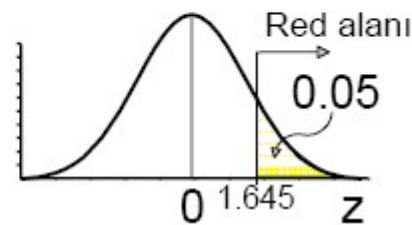
$$z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{3725 - 368}{\frac{15}{\sqrt{25}}} = +1.50$$

Karar:

Null hipotez $\alpha = 0.05$ ile rededilmez

Sonuç:

Ortalamanın 368 gr üzerinde olduğuna ait yeterli delil yoktur.





Çift yönlü z Testi Çözümü

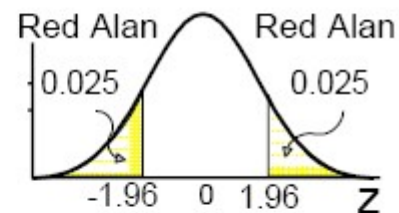
$$H_0: \mu = 368$$

$$H_A: \mu \neq 368$$

$$\alpha = 0.05$$

$$n = 25, \sigma \text{ bilinmekte}$$

Kritik değerler



Test İstatistiği:

$$z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{3725 - 368}{\frac{15}{\sqrt{25}}} = +1.50$$

Karar:

Null hipotezi $\alpha = 0.05$ ile red edilmez

Sonuç:

Ortalama miktarın 368 olduğu hakkında yeterli bir delil yoktur



z-testi and t-testi karşılaştırılması

- **z-test istatistiği**

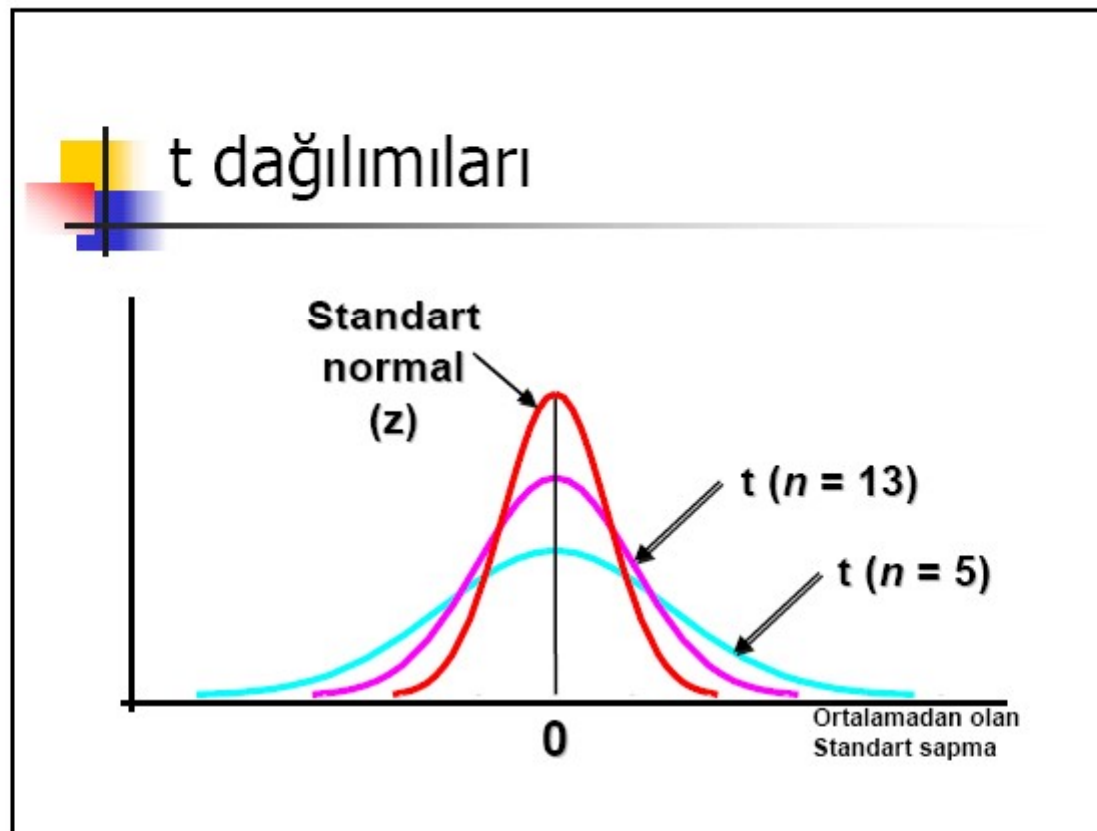
- **Normal dağılıma dayalı**

- Popülasyonun varyansı bilindiğinde yada örneklemedeki veri sayısı büyük olduğunda örneklerin ortalamaları hakkındaki hipotezleri test etmek için kullanılır

- **t-test istatistiği**

- t dağılımına dayalı

- t dağılımının şekli örneklemedeki veri sayısına bağlı olarak değişmektedir
 - Serbestlik derecesine bağlıdır $df : n-1$
 - Örneklemedeki veri sayısı artınca t dağılımı normal dağılıma yaklaşır
 - Popülasyonun varyansı yada standart sapması bilinmediğinde ve örneklemedeki veri sayısı küçük olduğunda ($n < 30$) örneklerin ortalamaları hakkındaki hipotezleri test etmek için kullanılır





t-testleri

Varyans hakkında kesin bir bilgiye sahip olmadığımız için (sadece tahmin), t dağılımını kullanırız

■ t-testinin ortalaması
$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

$\bar{X}$ = örnek ortalaması

μ = test edilen popülasyonun ortalaması

s = örnek standart sapması

n = örneklemedeki veri sayısı



Hipotez testi: σ bilinmemekte

- Yüksek lisans dersindeki öğrencilerin araştırma metodları hakkında iyi bir bilgiye sahip olup olmadıklarını öğrenmek istiyorum
- 6 öğrenci sınıftan rastgele seçilir ve sınava tabii tutulur
- Sınıfın test den en az 70 alabilmesini istiyorum. 6 öğrencinin notları 62, 92, 75, 68, 83, and 95.



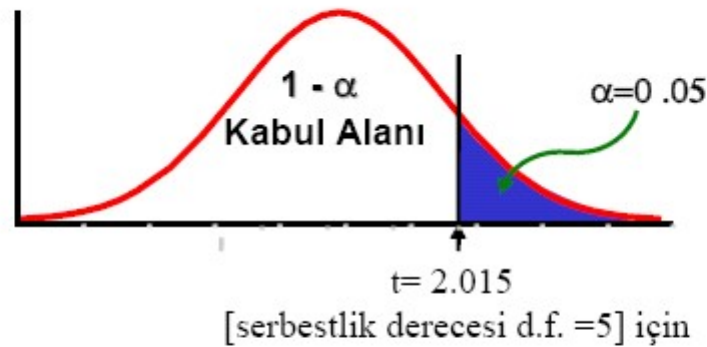
1 & 2: Hipotezleri belirle, önem derecesini α belirle

- Sınıfın ortalama notu 70 ve üstüdür:
 - $H_0: \mu \leq 70$
 - $H_A: \mu > 70$
- $\alpha = 0.05$

3: Tek veya çift yönlü bir hipotez mi?

4: Kritik değerleri belirle

- Tek yönlü
- Kritik değerleri t tablosundan belirlenir





5: Test istatistiklerini hesapla & değerlendir

- t değeri kritik değerinden küçüktür. Null
- $$\bar{x} = \frac{475}{6} = 79.17$$

- Hipotezi kabul edilir $s = 13.17$

$$t = \frac{79.17 - 70}{\frac{13.17}{\sqrt{6}}} = 1.71$$



Örnek: 2 yönlü t testi

- Kuzey Kıbrıstaki seçim noktalarının her birinde az yada çok 368 seçmen oy kullanmış mıdır?
- 36 rastgele seçim noktasındaki ortalama seçmen sayısı 372.5 ve standart sapma 12 seçmendir.
- Hipotezi 0.05 önem derecesi ile test edelim

Çift yönlü t Testi: Çözüm

Test İstatistiği:

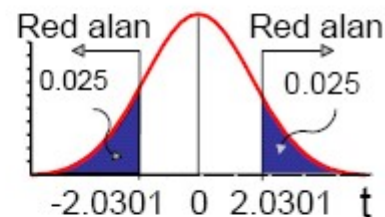
$$H_0: \mu = 368$$

$$H_1: \mu \neq 368$$

$$\alpha = .05$$

$$df = 36 - 1 = 35$$

Kritik değerler



$$t = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} = \frac{372.5 - 368}{\frac{12}{\sqrt{36}}} = +2.25$$

Karar:

Null hipotezi $\alpha = 0.05$ ile red edilir

Sonuç:

Popülasyonun ortalamasının 368 olmadığına ait delil vardır



Örnek: Tek yönlü t testi

- Kuzey Kıbrıstaki seçim noktalarının her birinde 368 den fazla seçmen oy kullanmış mıdır?
- 36 rastgele seçim noktasındaki ortalama seçmen sayısı 372.5 ve standart sapma 12 seçmendir.
- Hipotezi 0.05 önem derecesi ile test edelim

Tek yönlü t testi - Çözüm

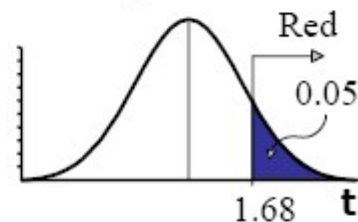
$$H_0: \mu \leq 368$$

$$H_1: \mu > 368$$

$$\alpha = .05$$

$$df = 36 - 1 = 35$$

Kritik değerler



Test İstatistiği:

$$t = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} = \frac{372.5 - 368}{\frac{12}{\sqrt{36}}} = +2.25$$

Karar:

Null hipotezi $\alpha = 0.05$ red edilir

Sonuç:

368 den fazla seçmenin ortalama oy sandıklarında oy kullandığına ait delil vardır

t-Test: Örnek

- Corpus içerisinde, *new* kelimesi 15,828 kez, *companies* kelimesi de 4,675 kez geçmiş olsun, ve corpusta toplam 14,307,668 kelime olsun.

$$P(\text{new}) = 15828 / 14307668$$

$$P(\text{companies}) = 4675 / 14307668$$

Null hipotez bu iki kelimenin bağımsız olarak meydana geldiği olsun.

$$H_0 : P(\text{new companies}) = P(\text{new})P(\text{companies})$$

$$= \frac{15828}{14307668} \times \frac{4675}{14307668} \approx 3.615 \times 10^{-7}$$

Eğer bu null hipotez doğru ise rasgele Bigram'lar üretelim. *New company* gelirse 1, diğer durumlarda 0 olsun (Bernoulli trial – sadece iki durum söz konusu)

t-Test: Örnek

$$P = 3.615 \times 10^{-7}$$

$$\mu = 3.615 \times 10^{-7}$$

$$\sigma^2 = p(1-p) \cong p$$

- 14,307,668 adet bigram içerisinde *new companies* kelimesi ile 8 kez karşılaşılсын

$$\bar{X} = \frac{8}{14307668} \approx 5.591 \times 10^{-7}$$

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \approx \frac{5.591 \times 10^{-7} - 3.615 \times 10^{-7}}{\sqrt{\frac{5.591 \times 10^{-7}}{14307668}}} \approx 0.999932$$

$\alpha = 0.005$ için kritik değer 2,576 olsun, (df=sonsuz)

$t < 2,576$ null hipotez kabul edilir.
New company collocation değildir.

t	C(w ¹)	C(w ²)	C(w ¹ w ²)	w ¹	w ²
4.4721	42	20	20	Ayatollah	Ruhollah
4.4721	41	27	20	Bette	Midler
4.4720	30	117	20	Agatha	Christie
4.4720	77	59	20	videocassette	recorder
4.4720	24	320	20	unsalted	butter
2.3714	14907	9017	20	first	made
2.2446	13484	10570	20	over	many
1.3685	14734	13478	20	into	them
1.2176	14093	14776	20	like	people
0.8036	15019	15629	20	time	last

Hipotez red ediliyor. İlk 5 bigram collocation için adaydır.

Hipotez kabul ediliyor. Son 5 bigram collocation'a aday değildir.

H_0 : bu ikililer birbirlerinden bağımsızdır.

$\alpha = 0.005$ için değer 2,576 ise

Hypothesis testing of differences-İki ortalama Farkın Testi (Church and Hanks, 1989)

Bazı durumlarda iki popülasyonun ortalamalarının karşılaştırılması gerekebilir. Amaç, 2 örnek ortalamasının aynı ortalamalı 2 popülasyondan gelip gelmediğini test etmektir.

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

H_0 =farkların ortalaması sıfırdır. $\mu = 0$

$$\bar{x} - \mu = \bar{x} = \frac{1}{n} \sum (x_{1i} - x_{2i}) = \bar{x}_1 - \bar{x}_2$$

$$\bar{x}_1 = s_1^2 = P(v^1 w)$$

$$\bar{x}_2 = s_2^2 = P(v^2 w)$$

$$s^2 = p - p^2 \approx p$$

$$t \approx \frac{C(v^1 w) - C(v^2 w)}{\sqrt{C(v^1 w) + C(v^2 w)}}$$

Örnek: 2 popülasyonun ortalamalarının karşılaştırılması

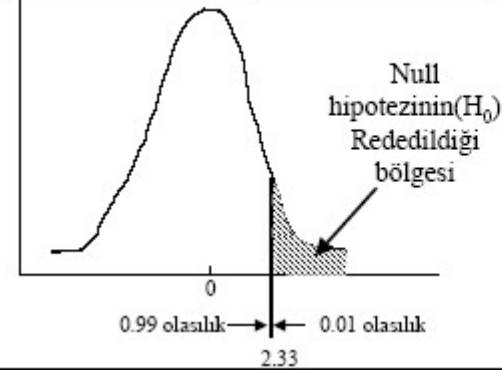
2 farklı hastanenin acil servisine gelen hastalara müdahale süresi aşağıda sunulmaktadır. Bu araştırmaya göre %1 riskle numune hastanın acil servisi, sigorta hastanesinin acil servisinden daha mı hızlı hastalara ilk müdahaleyi yapmaktadır?

Hastane	Ortalama süre	Örnek standart sapması	Örnek sayısı
numune	5.5 dak	0.4 dak	50
sigorta	5.3 dak	0.3 dak	100

Null ve alternatif hipotez:

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 > \mu_2$$



Örnek: Devam

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{5.5 - 5.3}{\sqrt{\frac{0.4^2}{50} + \frac{0.3^2}{100}}} = 3.12$$

$$z = 3.12 > 2.33$$

null hipotezi red edilir, alternatif hipotez %1 riskle kabul edilir.

p değeri bu büyüklükte yada onun üzerinde bir değer bulma olasılığıdır. 3.12 ve üzerinde bir z değeri alma olasılığı 0.499(Tabloda 3.12 değeri olmadığından en yakın 3.09 değerine karşılık gelen olasılık esas alınmıştır.

Buna göre 3.12 ve üzeri bir değer olma olasılığı: 0.5-0.499=0.001

Bu değer 0.01 risk derecesinden küçük olduğundan null hipotezinin doğru olmama ihtimali çok yüksektir.

Örnek : *strong* ve *powerful* kelimeleri ile birlikte görülen kelimeleri bulmak isteyebiliriz.

t	C(w)	C(strong w)	C(powerful w)	Word
3.16	933	0	10	computers
2.82	2337	0	8	computer
2.44	289	0	6	symbol
2.44	588	0	5	Germany
2.23	3745	0	5	nation
7.07	3685	50	0	support
6.32	3616	58	7	enough
4.69	986	22	0	safety
4.58	3741	21	0	sales
4.02	1093	19	1	opposition

$$H_0: \mu_1 = \mu_2$$

$\alpha = 0.005$ için değer 2,576 ise

Pearson'ın ki-kare (chi-square) testi

- İki değişkenin birbirine bağımlı olup olmadığı veya bir değişkenin başka bir değişkenle ilişkili olup olmadığını test etmek için kullanılır.
- Popülasyon içerisindeki dağılım bilindiği halde bazen de bilinmeyebilir veya örneklem dağılımının popülasyon dağılımına uyup uymadığı kontrol edilmek istenebilir.
- 2 x 2'lik bir matris kullanılır. Matrisin hücrelerinde *gözlemlenen (observed)* değerler vardır. Bu matris yardımıyla *beklendik (expected)* değerler hesaplanır.
- Sonrasında ki-kare değeri hesaplanır.

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

χ^2 Test: örnek

	$w_1 = new$	$w_1 \neq new$	
$w_2 = companies$	8 (new companies)	4667 (e.g., old companies)	4675
$w_2 \neq companies$	15820 (e.g., new machines)	14287181 (e.g., old machines)	14303001
	15828	14291848	14307676

$E_{ij} = ((\text{satır_toplamı}) \times (\text{sütun_toplamı})) / \text{toplam } N$

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$$\chi^2 \cong 1.55$$

5.17	4669.8
15822.8	14287178.2

H_0 : bu ikililer birbirlerinden bağımsızdır.

χ^2 tablosundan $df(\text{degree of freedom}) = n - 1$ ($n = 2$ $df = 1$) için $\alpha = 0.05$ değeri 3.8 olup, $1.55 < 3.8$ null hypothesis kabul edilir. *new companies collocation değildir.*

χ^2 nin farklı kullanım alanları

	cow	$\wedge$ cow
vache	59	6
$\wedge$ vache	8	570934

İki farklı corpus'tan yararlanarak çeviri yaparken uygun kelimenin bulunması.

H_0 = cow, vache birbirinden bağımsızdır.

$\chi^2 = 456400$ bulunur ve H_0 red edilir

Mutual Information

- Mutual Information, bir kelimenin diğer kelimeler hakkında bize ne söylediğini kabaca anlatır.
- Bazı problemleri mevcuttur.
 - İki olay arasındaki benzerliğin ölçümünde her zaman iyi değildir.
 - Bağımlılığın ölçümünde kötüdür.
 - Sparse data'da kötüdür.

$$\begin{aligned} I(x', y') &= \log_2 \frac{P(x' y')}{P(x') P(y')} \\ &= \log_2 \frac{P(x' | y')}{P(x')} \\ &= \log_2 \frac{P(y' | x')}{P(y')} \end{aligned}$$

$I(w_1, w_2)$	$C(w_1)$	$C(w_2)$	$C(w_1, w_2)$	w_1	w_2
18.38	42	20	20	Ayatollah	Ruhollah
17.98	41	27	20	Bette	Midler
16.31	30	117	20	Agatha	Christie
15.94	77	59	20	videocassette	recorder
15.19	24	320	20	unsalted	butter
1.09	14907	9017	20	first	made
1.01	13484	10570	20	over	many
0.53	14734	13478	20	into	them
0.46	14093	14776	20	like	people
0.29	15019	15629	20	time	last

$$I(\text{Ayatollah}, \text{Ruhollah}) = \log_2 \frac{\frac{20}{14307668} \times \frac{14307668}{42}}{\frac{14307668}{20}} \approx 18.38$$

	Chambre	$\hat{\text{chambre}}$	MI	χ^2
House	31,950	12,004	4.1	553610
$\hat{\text{house}}$	4,793	848,330		
	Communes	$\hat{\text{communes}}$		
House	4,974	38,980		
$\hat{\text{house}}$	441	852,682	4.2	88405

Kanada parlamentosundaki anayasa hem İngilizce hem de Fransızca olarak hazırlanmış.

$$\log \frac{P(\text{house} | \text{chambre})}{P(\text{house})} = \log \frac{\frac{31950}{31950 + 4793}}{P(\text{house})} \approx \log \frac{0.87}{P(\text{house})}$$

$$< \log \frac{0.92}{P(\text{house})} \approx \log \frac{\frac{4974}{4974 + 441}}{P(\text{house})} = \log \frac{P(\text{house} | \text{communes})}{P(\text{house})}$$

Collocation'nın Kullanım Alanları ...

- Corpus Analizlerinde
- Information Retrieval
- Cross-language Information Retrieval