

## Statquest ~~with~~ Josh Starmer - Youtube

### p-values: What they are and how to interpret them

\* A ve B diye iki ilacın olduğunu düşünelim. A ilacının B ilacından farklı olup olmadığını öğrenmek istiyoruz.

- Bir kişiye A ilacını, başka bir kişiye B ilacını verdik. A ilacını alan kişi iyileşti, B ilacını alan kişi iyileşmedi.

→ A ilacının B ilacından daha iyi olduğu sonucuna ulaşabilir miyiz?  
Hayır.

B ilacı farklı sebeplerden dolayı başarısız olmuş olabilir. Belki bu ilacı alan hastalar, B ilacıyla kötü etkileşimi olan başka bir ilaç daha almıştır. Belki bu adamın B ilacına karşı nadir bir alerjisi vardır. Belki bu adam B ilacını düzenli bir şekilde almamıştır ve bir dozunu kaçırmıştır.

Veya belki A ilacı aslında işe yaramıyor placebo etkisinden dolayı, sadece işe yaramıştır A ilacını alan hastada.

Test yapılırken bunlar gibi birçok sonuç eseri bir rejler olmuş olabilir. Bu yüzden her ilaç birden fazla kişiyle tekrar etmeliyiz.

- Bu sefer A ilacını 2 kişiye, B ilacını başka 2 kişiye verdik. A ilacını alan 2 kişi iyileşti, B ilacını alan 2 kişiden biri iyileşti diğer iyileşmedi.

→ A ilacı B ilacından daha mı iyi?

→ Her iki ilaç aynı mı?

Bu sonuçlara cevap veremeyiz. Çünkü belki B ilacını alıp iyileşmeyen kişinin iyileşmemesinin sebebi farklı şeyler olabilir.

Belki B ilacını alıp iyileşen kişi aslında yanlış etiketlendiği için A ilacını aldığı için iyileşti.

Bu yüzden deneyi çok daha fazla kişiyle tekrar ediyoruz.

- A ve B ilacını çok daha fazla kişiyle tekrar ettiğimizde şu sonuçlara ulaşıyoruz:

A	
Cured	Not Cured
1,043	3

99.7% Cured

B	
Cured	Not Cured
2	1,432

0.1% Cured

Bu sonuçlara göre A'nın B'den daha iyi bir ilaç olduğu açık. Diğer bir deyişle, bu sonuçların rastgele gerçekleştiğini A ilacıyla B ilacı arasında bir fark olmadığını varsaymak gerçekçi olmayacaktır.



A ilacını alıp iyileşen kişilerin placebo etkisiyle iyileşmesi ve B ilacını alıp iyileşmeyen kişilerin alerjilerinden dolayı iyileşmemesi olması mümkün, fakat bu sonuçların rastgele olduğunu düşünmemiz için A ilacıyla iyileşen çok fazla insan ve B ilacıyla iyileşen çok az insan var.

- Peki bu sonuçları nasıl değerlendirebiliriz?

A	
Cured	Not Cured
73	125

37% Cured

B	
Cured	Not Cured
59	131

31% Cured

A ilacı, B ilacından daha fazla insana etki etmiş. Çalışma mükemmel olmadığında ve her zaman sonuç eseri bir seçimin olma ihtimali söz konusu olduğunda A ilacının daha iyi bir ilaç olduğundan nasıl emin olabiliriz?  
p-values burada devreye giriyor.

### p-values

p-values; bu örnekte A ilacının B ilacından farklı olduğundan ne kadar emin olamamız gerektiğini ölçen 0 ile 1 arasındaki sayılardır.

p-value 0'a ne kadar yakınsa, A ve B ilacının birbirinden farklı olduğuna dair güvenimiz o kadar artar. Soru şu: A ilacının B ilacından farklı olduğuna emin olamamız için p-value'nun ne kadar küçük olması gerekir? Başka bir deyişle, iyi bir karar vermek için threshold olarak ne kullanmamız gerekir?

Pratikte genelde kullanılan threshold 0,05. Bunun anlamı, A ilacı ve B ilacı arasında hiçbir fark yoksa ve aynı deneyi birkaç kez yaparsak, bu deneylerin sadece %5'i yanlış sonuçla sonuçlanacaktır.

↳ Bunu örneklendirecek olursak:

- İki gruba aynı ilacı (A) verdiğimizizi düşünelim.

A	
Cured	Not Cured
73	125

$$p = 0.9$$

A	
Cured	Not Cured
71	127

Sonuçlardaki farklılık, bir kişide nadir görülen bir alerji veya başka bir kişide güçlü bir placebo etkisi gibi garip şeylere atfedilebilir.

Bu örnekte p-value 0,9, 0,05'ten çok büyük. Bu nedenle iki grup arasında bir fark göremediğimizi söyleyebiliriz.



Aynı grup: p-value büyük  $\rightarrow$  gruplar demek ki aynı  
p-value küçük  $\rightarrow$  gruplar demek ki farklı

- Aynı deneyi birden fazla kez tekrar ederseniz çoğunlukla benzer şekilde büyük p-value alındık.

(A)

Cured	Not Cured
71	127

$p=1$

(A)

Cured	Not Cured
72	126

(A)

Cured	Not Cured
75	123

$p=0.7$

(A)

Cured	Not Cured
70	128

etc...

- Bununla birlikte, orada bir ilaca alerjisi olan kişilerin tümü soldaki gruba dâhil olur ve plasebo etkisinde olan hastaların hepsi sağdaki gruba dâhil olur.

(A)

Cured	Not Cured
60	138

$p=0.01$

30% Cured

(A)

Cured	Not Cured
84	114

42% Cured

Bununla birlikte, sonuçlar oldukça farklı olduğundan deneyin bu özel çalışması için p-value 0.01'dir.

Bu bağlamda; aynı ilacı almalarına rağmen iki grubun farklı olduğunu söyleyebiliriz.

### False Positive

Fark yokken küçük bir p-value elde etmeye False Positive denir.

p-value için 0.05 eşiği, (farklılıkların yalnızca rastgele şeylerden kaynaklandığı) deneylerin %5'inin 0.05'ten küçük bir p-value ürettiği anlamına gelir.

Bazı bir diziyle, eğer A ilacıyla B ilacı arasında bir farklılık yoksa deneylerin %5'inde 0.05'ten daha küçük bir p-value olacağı, yani bir False Positive (%5'inde daha büyük bir p-value)

(A)

Cured	Not Cured
73	125

(B)

Cured	Not Cured
59	131

100 deneyin 5'i yanlış sonuç  
%5'inde p-value büyük  
Aynı ilacı aldı.  
Aynı ilacı aldı.  
Aynı ilacı aldı.  
Farklı ilacı aldı.



**Note!** İlaçlar farklıdır dediğimizde haklı olmamız son derece önemlidir, 0.00001 gibi daha küçük bir epik kullanabiliriz 0.00001 gibi bir epik kullanmak, her 100.000 deneyde yalnızca 1 false positive alacağımız anlamına gelir. (farklı diye tahmin etmek ama bu yanlış → sadece 1 örnek için yanlış tahmine izin vermişiz)

Fakat o kadar da önemli olmayan (dondurma arabasının yanında gelip gelmeyeceğine karar vermezse çalışıyorsa) o zaman 0.2 gibi daha büyük bir epik kullanabiliriz (10'da 2'sinde yanlış tahminde bulunurum, false positive → zannında gelecek diyorum ama gelmiyor)

Bununla birlikte en yaygın epik 0.05'tir, çünkü false positive'lerin sayısını %5'in altına düşürmeye çalışmak genellikle değerinden daha pahalıya mal olur.

• Bu bağlamda aynı p-value'yu hesapladığımızda;

Cured	Not Cured
73	125

$$p = 0.24$$

Cured	Not Cured
59	131

p-value büyük → aynı

p-value < 0.05 ise o zaman A ilacının B ilacından farklı olduğuna karar verilir.

Fakat p-value burada 0.24. Dolayısıyla A ilacının B ilacından farklı olduğundan emin değiliz.

## Hypothesis Testing

İstatistiksel dilde, bu ilaçların aynı olup olmadığını belirlemeye çalışır. Akrine Hypothesis Testi denir.

Null Hypothesis: İlaçlar aynıdır.

p-value, null hypothesis'ni reddedip reddetmemek konusunda bize yardımcı olur. (Alternatif hipotezimiz: bunların farklı olması)

(Kendi yorumum: son örnekte p-value büyük çıktığı için null hypothesis reddedemiyorduk)

• Küçük p-value A ilacının B ilacından farklı olup olmadığını konusunda bize yardımcı olur ama ne kadar farklı olduklarını bize söylemez.

Cured	Not Cured
73	125

$$p = 0.24$$

37% Cured

Cured	Not Cured
59	131

29% Cured

Null hypothesis  $\rightarrow$  Aynı  
alternative "  $\rightarrow$  Farklı  $\rightarrow$  buna göre p-value dağılımına.  $p < 0.05$  ise  $\checkmark$   
25.10.2021

p value  $< 0.05$  ise null hypothesis reddedilir.

Örneğin; bu deney bize %8'lik bir fark olduğunu söylemesine rağmen 0.24 gibi büyük bir p-value veriyor.

Aksiine;

A	
Cured	Not Cured
5005	9868

34 % Cured

$$p = 0.04$$

Cured	Not Cured
4800	9000

35 % Cured

Bu örnekte A ilacıyla B ilacı arasında sadece %1 fark olmasına rağmen çok daha küçük bir p-value veriyor.

Yani p-value'nun büyüklüğüyle A ve B'nin farklılık oranları arasında bir ilişki yok.