Youtube - Statquest with Josh Starmer

Odds Ratios and Log (odd ratios), Clearly Explained!!!

$$\text{odds} = \frac{\text{bir şeyin olma sayısı}}{\text{bir şeyin olmama sayısı}}$$

Odds ratio dediğimiz şey ise, farklı odds'ların birbirlerine olan oranı.

$$\text{odds ratio} = \frac{\frac{2}{4}}{\frac{3}{1}} = 0.17 \qquad \text{odds ratio} = \frac{\frac{3}{1}}{\frac{2}{4}} = 6$$

Eğer pay paydadan küçükse odds ratio 0 ile 1 arasına düşerken, eğer pay paydadan büyükse 1 ile sonsuz arasına düşüyor bu değer.

Oddsta olduğu gibi bu odds ratio'ların logaritmasını aldığımızda değerler orijin (0)'den simetrik uzaklıkta oluyor.

$$\log(1.7) = -1.79 \qquad \log(6) = 1.79$$

Odds ratio'lar ile ne yapabiliriz?

		Has Cancer: Yes	Has Cancer: No
Has the mutated gene	Yes	23	117
	No	6	210

> 356 kişi var.

356 kişiden 29'u kanser, 327'si değil. Aynı zamanda bu kişilerin 140'ın mutasyona uğramış gene sahip olurken, 216'sı mutasyona uğramış gene sahip değil.

Mutasyona uğramış gen ve kanser arasında bir ilişki olup olmadığını belirlemek için odds ratio'yu kullanabiliriz. Mutasyona uğramış gene sahip birinin kanser olma olasılığı daha mı yüksek?

→ Bir kişinin mutasyona uğramış gene sahip olduğu göz önüne alındığında, kanser olma ihtimali (odds): $\frac{23}{117}$ → bunu pay kısmına yazacağız.

$$\frac{23}{117} = \frac{23/140}{117/140} = \frac{\text{kanser olma olasılığı}}{\text{kanser olmama olasılığı}}$$

→ Bir kişinin mutasyona uğramış gene sahip olmadığı göz önüne alındığında, kanser olma ihtimali (odds): $\frac{6}{210}$ → bunu payda kısmına yazıyoruz.

$$\frac{6}{210} = \frac{6/216}{210/216} = \frac{\text{kanser olma olasılığı}}{\text{kanser olmama olasılığı}}$$

$$\text{odds ratio} = \frac{\frac{23}{117}}{\frac{6}{210}} = \frac{0.2}{0.03} = 6.88$$

* Odds ratio bize mutasyona sahip birinin kanser olma olasılığının 6.88 kat daha fazla olduğunu söylüyor.

log (odds ratio):
log (6.88) = 1.93

Odds ratio ve log (odds ratio) R^2 gibidir; iki şey arasındaki ilişkiyi gösterirler (bu durumda, mutasyona uğramış gen ile kanser arasındaki bir ilişki). Ve tıpkı R^2 gibi değerler, etki büyüklüğünde karışık gelir. Büyük değerler; mutasyona uğramış gen, kanser riski için iyi bir tahmin edici anlamına gelirken, küçük değerler, mutasyona uğramış genin kanser riski için iyi bir tahmin edici olmadığı anlamına gelir.

* Ancak, tıpkı R^2 gibi, bu ilişkinin istatistiksel olarak anlamlı olup olmadığını bilmemiz gerekiyor.

Odds ratio'nun (veya log (odds ratio)'nun) istatistiksel olarak anlamlı olup olmadığını belirleyebilmemiz için 3 yol var.

1) Fisher's Exact Test
2) Chi-Square Test
3) The Wald Test

Hangi yöntem daha iyidir bilmiyoruz. Bazıları p-value'yu hesaplamak için Fisher's Exact Test ya da Chi-Square Testi, confidence interval'ı hesaplamak için The Wald Testi kullanırken bazısı hem p-value'yu hem confidence interval'ı hesaplamak için The Wald testi kullanıyor.

1) Fisher's Exact Test

İlk adım; Statquest'ten "Enrichment Analysis using Fisher's Exact Test and the Hypergeometric Distribution" videosunu izlemek

Has Cancer

Has the Mutated Gene	Yes	No
Yes	23	117
No	6	210

İnsanları şeker olarak düşünelim. Kanser olan hastalar 23+6 = 29 kırmızı şeker ile kanser olmayan hastalar 117+210 = 327 mavi şeker ile temsil edilsin.

İlgili Statquest'te olduğu gibi bir avuç 23 kırmızı şeker ve 117 mavi şeker almak için p-value hesaplıyoruz ve p-value'yu 0,00001 buluyoruz.

2) Chi-square test

Chi-square test kullanarak p-value'yu nasıl hesapladığımıza bakalım. Ki-kare testi, gözlemlenen değerleri, mutasyona uğramış gen ile kanser arasında hiçbir ilişki olmadığını varsayan expected (beklenen) değerlerle karşılaştırır.

- Bunu yapmak için kanserli toplam kişi sayısını, toplam kişi sayısına bölüyoruz.

$$\frac{23+6}{23+6+117+210} = \frac{29}{356} = 0.08$$

Dolayısıyla kanser olma olasılığı 0.08

$$p(\text{has cancer}) = \frac{29}{356} = 0.08$$

- Yani gen, mutasyona uğramış 23+117=140 gen ile ilişkili değilse,

kanser olma olasılığı X mutasyona uğramış gene sahip insan sayısı

$$0.08 \times 140 = 11.2$$

- Böylece, mutasyona uğramış gen ve kansere sahip beklenen (expected) kişi sayısı: 11.2
- Kalan mutasyona uğramış gene sahip 140−11.2=128.8 kişinin kanser olmaması bekleniyor.

- Aynı şekilde, gen, mutasyona uğramış gen olmadan (6+210)=216 kişi ile ilişkili değilse, o zaman,

kanser olma olasılığı X mutasyona uğramış gene sahip olmayan kişi sayısı

$$0.08 \times 216 = 17.3$$

- Böylece, mutasyona uğramış gene sahip olmayan kanserli kişi sayısı 17,3'tür.
- Kalan mutasyona uğramış gene sahip olmayan 216−17.3=198.7 kişinin kanser olmaması bekleniyor.

Observed Values

Has the Mutated gene	Has Cancer: Yes	Has Cancer: No
Yes	23	117
No	6	210

Expected Values

Has the Mutated gene	Has Cancer: Yes	Has Cancer: No
Yes	11.2	128.8
No	17.3	198.7

Gözlemlenen değerleri ve beklenen değerleri karşılaştırmak için Ki-kare testi yapıyoruz ve p-value'yu continuity correction ile 0.00001, continuity correction olmadan 0.000004 buluyoruz.

3) The Wald Test

Bu test, lojistik regresyonda odds ratio'ların önemini belirlemek ve güven aralıklarını hesaplamak için yaygın olarak kullanılır.

Wald Testi, log (odds ratios)'ların tıpkı log (odds) gibi normal olarak dağıldığı gerçeğinden yararlanır. Bu, mutasyona uğramış gen ile kanser arasında bir ilişki yoksa ne bekleyeceğimizi söyleyen, rastgele oluşturulmuş 10.000 log (odds ratio)'dan oluşan bir histogramdır.

Bu histogramı (normal dağılım gibi görünen) şu şekilde oluşturuyoruz:

1) Random olarak 300 ile 400 arasında bir sayı çekiyoruz. Diyelim ki 325 çektik → toplam sample size

2) Her sample için 0 ile 1 arasında random bir sayı çekiyoruz. Örneğin; 0.01, 0.43, 0.95...

3) 0.08'den (kanserli insanların oranı) küçük random sayılar kanserli örneklerdir. Örneğin 0.05 çektik. Bu 325 örneğin 17'sinin kanserli olduğu anlamına geliyor.

4) 0 ile 1 arasında başka bir random sayı çekiyoruz.

5) 0.39'dan (mutasyona uğramış gene sahip kişi sayısı) küçükse çektiğimiz sayı mutasyona uğramış gene sahiptir.

Bu, bize mutasyona uğramış gen ile kanser arasındaki ilişkiye bağlı olmayan random değerler matrisi verir.

Random Values

Has the mutated gene	Has Cancer: Yes	Has Cancer: No
Yes	14	141
No	17	122

Son olarak log (odds ratio)'yu hesaplıyoruz. Bunu 10.000 kere yapıp histogramı çizdiriyoruz. Histograma normal eğri oldukça iyi oturuyor. Histogram ve eğri 0 noktasında merkezleniyor. → Eğer odds'lar arasında kraker fark yoksa (log(1)=0) log (odds ratio)=0 oluyor.

10.000 log (odds ratio)'nun standart sapması 0.43. Fakat standart sapmayı gözlemlenen değerlerden hesaplamak daha yaygın.

Bunu tüm değerleri paydayal alıp toplama ve karekök ilemi yaparak buldurduk.

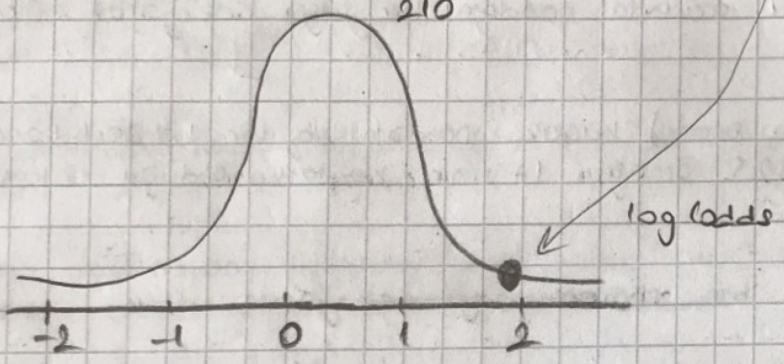
gözlemlenen değerler ← $\sqrt{\frac{1}{23}+\frac{1}{117}+\frac{1}{6}+\frac{1}{210}} = 0.47$

Örnekinde 0,43 bulmuştuk. Oldukça benzer 0,47'yle.

Wald Testinin tek yaptığı, gözlemlenen log (odds ratio)'nun 0'dan kaç standart sapma olduğunu görmektir. Wald Test genel olarak "estimated" standart sapmayı kullandığı için histogramımızın yerine standart sapma 0,47 olacak şekilde değiştirebiliriz normal eğrisiyle birlikte.

log (odds ratio) bizim daha önce hesapladığımızla aynı.

$$\log(\text{odds ratio}) = \log\frac{\frac{23}{117}}{\frac{6}{210}} = \log(6.88) = 1.93$$



standart sapma = 0.47

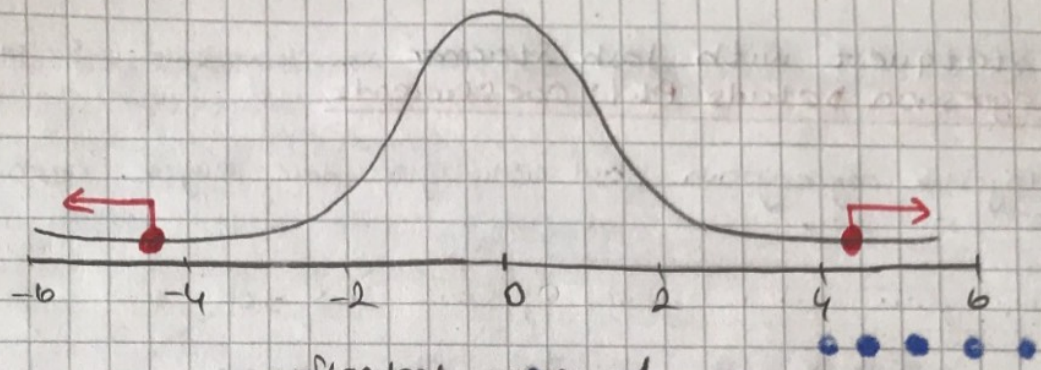
log(odds ratio), 0'dan kaç standart sapma uzakta bulabilmek için log (odds ratio)'yu standart sapmaya bölüyoruz.

$$\frac{1.93}{0.47} = 4.11$$

Dolayısıyla log(odds ratio) dağılımın ortalamasından (0) 4,11 standart sapma uzaklıktadır. Normal dağılımlarla ilgili genel kural, ortalamadan 2 standart sapmanın ötesindeki herhangi bir şeyin p-value'sunun 0,05'ten küçük olacağıdır. Bu yüzden log (odds ratio)'nun istatistiksel olarak anlamlı olduğunu biliyoruz.

Kesin bir 2-taraflı p-value elde etmek için 1.93'ten büyük noktalar ve -1.93'ten küçük noktalar için eğrinin altındaki alanları toplayabiliriz. Ancak bu, geleneksel olarak <u>standart bir normal eğri</u> kullanılarak yapılır (yani ortalama = 0 ve standart sapma = 1 olan normal bir eğri.

Ve bu, 4.11'den büyük olan noktalar ve -4.11'den küçük olan noktalar için eğrinin altındaki alanların toplanması anlamına gelir; burada 4.11, log (odds ratio)'nun ortalamadan uzakta olduğu standart sapmaların sayısıdır.



Standart sapma = 1

Sonuçta, mutasyona uğramış gen ile kanserle ilişkinin olmadığı p-value 0.00005'tir.

* log (odds ratio) ile kullanabileceğimiz bu 3 testle ilgili olarak!

10.000 random log (odds ratio) ürettiğim zaman, 3 testi bunlar üzerinde çalıştırdım. Eğer testler beklendiği gibi çalıştıysa p-value'nun < 0.05 olmalı. Bizim bulduklarımız:

1) Fisher's Exact Test: 4% of the p-values were < 0.05

With continuity correction ... 3%

2) Chi-Square Test:

Without continuity correction ... 5%

3) The Wald Test: 5% of the p-values were < 0.05

Tüm testler anlamlı sonuçlar üretti. Kendi alanımızda genelde hangi testin kullanıldığını bulmak gerekiyor.

Özet olarak;

odds ratio ve log(odds ratio) bize iki şey arasında güçlü veya zayıf bir ilişki olup olmadığını söyler, örneğin mutasyona uğramış bir gene sahip olup olmamanın kanser olma ihtimalini (odds) artırması gibi.

Ve, çalışılan alana bağlı olarak insanlar Fisher's Exact Test, Chi-Square veya Wald Test kullanıyorlar ilişkinin anlamlılığına dair p-value'yu belirlemek için.