4.2.2. Bağıntı ve Anlamlılık Sınamalarının Yanlış Kullanımı: Gazeteciler Nasıl Kandırmış Olur?

Geçersizliği Yenmek

Önce olanaksızlık nedir onu ele alalım. Matematik kuramlarının bazıları neredeyse çocukluktan gelen güdülerimize uygundur. Bunlar arasında sayılar, geometri, aritmetik gelir. Doğal içgüdülerimize uygundur bu konular. Ancak olasılıkçılık bunlardan biraz farklıdır. Bu nedenle de matematik tarihinde oldukça geç oluşmuş kuramlardan biridir. Yazı-tura atarken, turanın gelmesi olasılığı yarı yarıyadır dersek, Büyük Sayılar Yasası'na gönderme yapmış oluruz. Ne kadar çok sık yazı-tura atarsak, turanın yüzde 50 gelme olasılığı o kadar artar. Bu yaklaşıma olasılığın tekrarlamacı bakışı diyebiliriz. Ancak bazı durumlar vardır ki sık tekrarlama mümkün olmaz. Örneğin cep telefonlarınızdaki hava durumuna baktığınızdaki yağmur olasılığı yüzdesi kafanızı karıştırmalıdır. Yani bugünden yarının yüzde 20 yağışlı olacağını söylediğimizde, yarını çok sayıda tekrar etmek mümkün olmaz (Ellenberg, 2015: 110-111).

Buna karşın matematikçiler ayaklarına dar gelen ayakkabıları giyebilmekte ustadırlar. Biraz zorlamayla, bu tekrarlamacı bakışı yağmur olasılığına uygulamak mümkün. Sonuç olarak "bugüne" ilişkin çok sayıda verimiz varsa, bugüne bakarak yarına ilişkin bir olasılık vermek mümkün hale gelir. Buraya kadar anlaşılabilir görülebilir bu durum. Ama şu soruya nasıl yanıt verirsiniz: "Önümüzdeki 1,000 yılda insan ırkının yok olma olasılığı kaçtır?" Bu açıkça tekrarlamasını çok sayıda oluşturamayacağınız bir durumdur. Elinizde çok sayıda yok oluşa ilişkin veri de bulunmamaktadır. Bu sorulara pek çok soru eklemek mümkündür: Daha fazla zeytinyağı tüketmenin kanser hastalığını önleme olasılığı nedir? Orhan Pamuk'un kitaplarını Orhan Pamuk'un yazmış olma olasılığı nedir? Bu tür sorulara yanıt verirken yazı tura atışlarında kullandığımız dilin kullanılması oldukça zordur. Buna karşın, bu tür sorularada "pek mümkün değil gibi" veya "öyle gibi gözüküyor" gibi yanıtlar verebiliriz. Bu yanıtlardan sonra bir adım daha atınca şu soruyu sormaktan kaçınmak zordur: "Ne kadar mümkün değil?" veya "Ne kadar öyle?"

Tabii sormak başka yanıtlamak başka. Ayaklarını dar ayakkabıya sokabilmeyi başaran metamatikciler bu tür sorulara da yanıt verme başarısını geliştirecek yaklaşımlar oluşturmuşlardır. Örneğin talih oyunlarında, piyangolarda, haftada iki kez birbirinin aynısı olan sayıların kazanması mümkün gözükmemektedir. Bu doğru bir yanıt olur, yeter ki sayıların çekilmesinin bütünüyle rastgele olduğu denencesi geçerli olsun. Ama rastgeleliliğin ortadan kaldırıldığını düşünüyor olabilirsiniz. Örneğin piyangoyu çekenlerin hile yaptığını düşünebilirsiniz. Böyle bir durumda temel denenceyi, sayıların rastgele çekildiğini kabul etmemiş oluruz. Öylesi bir durumda aynı sayı çiftlerinin bir hafta içinde 2 kez aynı çıkmaması durumu "mümkün olmayabilir". Mümkün olmama durumu mutlak değil, görelidir. Bir haftada aynı sayıların 2 kere çıkmamasının mümkün olmama durumundan söz edince, ister açık ister örtük olarak mümkün olmama durumunun bir dizi ön kabule göre olduğunu söylemekteyiz (Ellenberg, 2015:112).

Bilimsel soruların büyük çoğunluğu evet/hayır sorularına indirgenebilir. Bir şey oluyor mu, olmuyor mu? Yeni ilaç sağaltmayı hedeflediği hastayı iyileştiriyor mu, iyileştirmiyor mu? Psikolojik bir değişken sizi daha mutlu/daha çekici/ yapıyor mu yapmıyor mu? Bu sorulara olumsuz yanıt verdiğinizde, yani olmuyor, yapmıyor, iyileştirmiyor dediğiniz durumlardaki senaryo geçersizlik denencesi olarak tanımlanır. Geçersizlik, araştırdığınız konuyla ilgili yapılan müdahalenin bir etkisi olmadığı denencesidir. Yeni bir ilaç geliştiren bir araştırmacıysanız, geçersizlik denencesi uykularınızı kaçıran bir durumdur. Uykularınızın kaçmaması için öncelikle ilacınızın hiç bir etkisi olmadığı durumunun elenmesi gerekir. Yani geçersizlik denencesinin reddedilmesi gerekir. Bunu nasıl yapabiliriz? Matematikteki standard çerçevesi geçersizlik denencesi anlamlılık sınamasıdır ve R. A. Fisher isimli istatistikçi tarafından 20. YY'ın başlarında geliştirilmiştir. Bu çerçeve yeni ilacımıza ilişkin olarak şöyle gelişir. Birinci olarak bir deney yapmalısınız. Yüz kadar denekle başlayıp bunları rastgele ikiye ayırıp, bir yarısına harika ilacınızı uygularsınız. Diğer yarısınaysa placebo olarak adlandırılan ilaç olmayan hap, yani toz verirsiniz. Beklentiniz, ilacınızı alan hastaların, sahte hapı alanlara göre daha az ölme şansına sahip olduğudur.

Bundan sonra yapılacaklar daha basittir: Eğer ilacınızı alan hastalar grubunda daha az ölüm gözlemliyorsanız, zaferinizi ilan edip yetkili mercilere ilaç başvurusu yapabilirsiniz. Ama bu yanlıştır. Verilerin sizin kuramınızla tutarlı olması yeterli değildir, verilerin kuramınızın olumluzlanması (geçersizlik denencesinin reddi) noktasında da tutarsızlık sergilemelidir. Bir kişi telekinetik becerilere sahip olduğunu, ve bu yeteneğiyle güneşi doğurabileceğini, kanıtın da sabaha karşı kalkınca gündoğumunun gözlenmesi olduğunu ileri sürebilir. Ama bu kanıt geçerli bir kanıt değildir, çünkü telekinetik becerilere sahip olmadığı geçersizlik denencesine, göre güneş sabaha karşı aynı saatte doğacaktır.

Tıptaki deneylerin sonuçlarının yorumlanması da benzeri bir dikkat gerektirir. Yukarıdaki harika ilacımıza geri dönersek, diyelim ki geçersizlik denencesi ülkesinde, ilacı alan 50 hastada ve ilaç yerine sahte tozu alan diğer 50 hastada ölüm şansı yüzde 10'dur. Ancak bu sahte tozu alanlarda 5 ve ilacı alanlarda 5 kişi ölecektir anlamına gelmez. Gerçekten de ilacı alan hastalarda tam 5 kişinin ölmesi şansı yüzde 18,5'tur ve yazı-tura atılmasında atışların arttırılması durumunda ancak turaların yarı yarıya geleceği durumuna benzer biçimde, olması pek mümkün gözükmemektedir. Deneyin gerçekleştirilme süresinde ilaç verilen hastalarla sahte toz verilen hastalardaki ölümlerin sayısının aynı sayı olması mümkün gözükmemektedir.

Yukarıdaki örneğin istatistiki olarak şans durumu şöyledir:

- Hem ilaç alanlarda hem de sahte toz alanlarda eşit ölümlerin olması şansı yüzde 13.3'tür.
- Sahte toz alan hastaların, ilaç alan hastalardan daha az ölmesi şansı yüzde 43.3'tür.
- İlacı alan hastaların, sahte toz alan hastalardan daha az ölmesi şansı yüzde 43.3'tür.

Yukarıdaki oranlara bakarak ilacı alan hastalarda daha iyi sonuçların ortaya çıkması bize hiç bir şey anlatmaz, çünkü geçersizlik denecesi olan ilacınızın hiç bir etkisi olmaması durumunda bile ilacı alan grupta ölümlerin azalması istatistiki olarak mümkün gözükmektedir (Ellenberg, 2015:114-118). Durum ne zaman değişir? Eğer ilacı alan hasta grubundaki durum çok daha iyi olursa (daha az sayıda ölüm olması). Diyelim ki sahte toz verilen hasta grubunda 5 ölüm ortaya çıktı ve bizim ilacımızı alan hasta grubunda hiç ölüm olmadı. Eğer geçersizlik denencesi doğruysa, her iki grubun yüzde 90 yaşama şansı vardır. Fakat bu durumda bile, ilacı aran gruptaki 50 kişinin tamamının yaşaması pek mümkün gözükmemektedir. İlacı alan gruptaki birinci kişinin yaşama şansı yüzde 90'dır. Sadece birinci hastanın değil ikinci hastanın da yaşama şansı yüzde 90'nin yüzde 90'ıdır: yüzde 81. Bunlar yanında üçüncü hastanın da yaşama şansını hesaplamak istersek, yüzde 81'in yüzde 90'ıdır: yüzde 72.9 kısacası. Yaşama şansı, her yeni hastayla birlikte aşınmaktadır ve sürecin sonunda tüm grubun yani 50 kişinin yaşaması şansı oldukça küçüktür: 50 kere 0.9 çarpılırsa = 0.000515. Geçersizlik denencesine göre, ancak 200 kerede 1 bu kadar iyi bir sonucun alınması söz konusu olabilir. Eğer bir kişi ben güneşi zihnimle doğurabiliyorum derse ve güneş de doğarsa, bu kişinin güçlerinden etkilenmenize gerek yoktur. Ama eğer bir kişi güneşin doğmasını engelleyebiliyorum diyorsa ve güneş doğmazsa, o kişi geçersizlik hipotezine göre önemli bir sonuç elde etmiştir.

Özet olarak geçersizlik denencesini eleme sürecinin protokolü şöyledir:

- Bir deneyi gerçekleştirin.
- Geçersizlik denencesinin doğru olduğunu varsayarak p değeri gözlemlenen sonuçlar kadar aşırı olasılık düzeyini oluştursun.
- P kodu p-değeri olarak tanımlanır. Eğer bu değer çok küçükse sevinin çünkü deneyin sonucunun istatistiki olarak anlamlı olduğunu söyleyebilirsiniz. Eğer büyükse, geçersizlik denencesinin elenemediğini itiraf edin.

"Çok küçük" ne kadar küçük olmalıdır? Anlamlı olanı anlamlı olmayandan ayırdetmek için keskin bir çizgi çekmek kolay değildir. Ancak Fisher'ın kendisiyle başlayan bir gelenek vardır ve oldukça geniş bir biliminsanları çevresinde destek görmektedir. Bu gelenek p değerinin p=0,05 (yani 20 de 1) olması eşiğidir.

Geçersizlik denencesi anlamlılık testi, belirsizliğin olduğu koşullarda sezgisel akıl yürütmeyi sağladığı için çok popülerdir. Fisher'ın katkısı anlamlılık sınamasını bir sisteme bağlayarak bir deneyin verilerinin anlamlı olup olmadığını nesnel bir olgu olarak ortaya koymasıdır. Fisher'ın geliştirdiği biçim, yaklaşık 100 yıldır bilimsel araştırmaların sonuçlarını değerlendirmeye yarayan standard metod olarak alınmaktadır. Örneğin Psikolojiye ilişkin herhangi bir üniversite derst kitabı, anlamlılık sınamasını "psikolojik araştırmaların omurgası" olarak tanımlamaktadır. Deneylerin başarılı mı başarısız mı olduğunu karar verilen standard olmuştur. Tıptan, iktisada kadar her alanda yayımlanan bir araştırmanın bir yerinde (eğer sayısal yöntemler kullanılmışsa) Fisher'ın standardı bulunmaktadır.

Sınamanın sınanması

Buna karşın, anlamlılık sınamasını "devasa bir hata" olarak gören başka bir grup matematikçi grubu da bulunmaktadır (Ellenberg, 2015:117-120). Psikolog David Bakan 1966 yılında psikolojinin krizinin aslında istatistik kuramının krizi olduğunu ileri sürdü. Bu tarihten 50 yılın üzerinde bir zaman geçmesine karşılık eleştiriler artarak devam etmektedir.

Anlamlılıkta yanlış olan nedir? Başlangıç olarak sözcüğün İngilizcesinde (significance test olarak geçen prosedür İngilizce önemlilik olarak anlaşılır) kendisinde kaymalar vardır 117. İngilizce'de significant sözcüğü önemli veya anlamlı olarak anlaşılır. Ancak bilim insanlarının kullandığı anlamlılık sınaması önemi ölçmez. Bir ilacın etkisini ölçerken, geçersizlik denencesi ilacın hiç bir etkisi olmadığır. Bu nedenle geçersizlik denencesinin reddi sadece ve sadece ilacın etkisinin "sıfır" olmadığını gösterir. Buna karşın bu etki son derece düşük olabilir ki matematikçi olmayan bir kişiye göre bu neredeyse ilacın hiç bir etkisinin olmadığı sonucunu gösterir.

İngiltere İlaç Güvenliği Komitesi (CSM) 18 Ekim 1995 yılında 200 bin doktora gönderdiği mektupta üçüncü kuşak, ağızdan alınan gebelik önleme ilaçlarından bazılarının "yeni kanıtlara" göre damarlarda pıhtı oluşturma şansını diğer ilaçlara göre 2 kat arttırdığını açıkladı. Bu mektup aynı zamanda ağızdan alınan gebelik önleyici ilaçların kadınların büyük çoğunluğu için güvenli olduğunu ve tıbbi bir öneri alınmadıkça ilaçların kullanımına devam edilmesi gerektiğini de vurguluyordu. Ancak bu mektubun başına yansıması "Doğum kontrol hapları öldürüyor!" olunca insanlar telaşa kapıldı. 19 Ekim 1995 tarihli The Associated Press (AP) bülteninin giriş paragrafı "Perşembe günü devlet yeni tip doğum kontrol haplarını kullanan 1,5 milyon İngiliz kadınını bu hapların damarlarda pıhtı oluşturabilmesi olasılığına karşı uyardı," olmuştu. Bültenin gövdesinde de devletin ilaçların lisansını iptal edip geri çekmeyi değerlendirdiğini, ancak bazı kadınların başka tür hapları tolere edemeyeceği gerekçesiyle bu kararından vazgeçtiği bilgisini içeriyordu. Haberler çıktıktan sonra hastaların yüzde 12'si ilaçları almaktan vazgeçmişti. Bazılarıysa başka ilaçlara geçmiş ve dolayısıyla daha çok gebeliğin olmasına yol açılmış olabilirdi. 1995 yılına kadar gebelik oranları sürekli düşmesine karşın, bu haberin çıkmasından sonra İngiltere'de doğum oranları bir kaç puan yükseldi. Bir önceki yıla göre 26 bin fazla gebelik ve 13 bin fazla kürtaj oldu.

Bu durum insanların canlarının kurtarılmasının bir bedeli olarak düşünülebilirdi. Ancak damarlarda tıkanma riskinin artması sonucu olabilecek ölüm sayısının sadece 1 olduğunu ileri süren bilim insanları oldu. Kısacası, üçüncü kuşak doğum kontrol haplarının damarda tıkanma yaratma riski Fisher anlamlılık sınamasına göre önemliydi ama kamu sağlığı anlamında bir önemi yoktu. Haberin çerçevelenmesinin yarattığı etki olayı büyütmüştü. Çünkü CSM bir risk oranı açıklamıştı: bu ilaçlar kullanan kadınlarla damar tıkanması riski 2 katına çıkmıştır. Bu kulağa çok kötü geliyor ama şunu bilmezseniz: damar tıkanması oranı çok düşüktür. Birinci ve ikinci kuşak doğum kontrol haplarında 7,000 kişiden bir kişide tıkanma riski vardı. Üçüncü kuşak ilaçlarda bu risk 2'ye katlanında 7,000 kişide 2'ye çıkmıştı. Bu oldukça düşük bir risktir, çünkü düşük bir riski 2 ile çarparsanız gene küçük bir risk ortaya çıkar (Ellenberg: 119).

ABD'de sosyogların yaptığı bir araştırma dadılar/bakıcılar tarafından bakılan çocukların ölüm oranının anaokulunada bakılanlara göre 7 kat fazla olduğunu söylüyordu (Ellenberg: 120-122). Buna dayanarak bakıcıyı işten atmadan önce evde bakılan çocuk ölümlerinin yıllık oranının 100 bin bebekte sadece 1,6 olduğunu düşünmek gerekir. Ama gerçekten de anaokulu/kreş gibi yerlerdeki kazalardan olan yıllık çocuk ölümlerinin 100 bin bebekte 0,27 olduğu, yani evde bakılan çocuklara göre kat ve kat düşük olduğu bilinmektedir. Her iki sayıda aşağı yukarı sıfırdır. 2010 yılındaki 1,100 kazada ölen bebeğin büyük çoğunluğunun çarşafa dolanarak boğulma olduğunu, ayrıca 2,063'ünün bebeklik ölüm sendromu hastalığı olduğunu gözönüne alıp, bakıcıların yol açtığı kazalarda 8-9 bebeğin öldüğünü bilmek gerekir. ABD'li sosyologların araştırmasına göre her şey eşitse, kreş/anaokulu bakımının bakıcı/dadı/anne bakımına göre yeğelenmesi sonucu önerilir. Ancak hayatta hiç bir veri eşit değildir. Evinizin yakınındaki kreşin yıllık bebek ölümleri oranı, bakıcı/dadı ölüm oranlarına göre 2 kat yüksek olabilir!

Anlamlılık sınaması bilimsel bir araçtır ve tüm diğer araçlar gibi belli bir oranda kesinliği vardır. Örneğin çalışılan nüfusun boyutunu genişleterek sınamayı daha hassas hale getirebilirsiniz. Böylece daha küçük etkileri de görmeniz mümkün olur. Yöntemin gücü buradadır, ama bu durum aynı zamanda zayıflığıdır. İşin doğrusu, geçersizlik denencesi, muhtemelen her zaman yanlıştır. Bir hastaya verilen herhangi bir ilacın tam olarak 0 etkide bulunması durumunun gerçekleşmesini beklemek mümkün değildir. Müdahele aracınız, bu durumda ilacınız başka etkilerde giderek başka hastalıklara yol açıyor olabilir. Eğer yeteri kadar araştırma yaparsanız başka hangi etkilere yol açtığını ortaya çıkartabilirsiniz. Ama bu etkileri bulmak demek onların bir sonuç yaratağı anlamına gelmez. Çünkü etkiler çok küçüktür.

Ellenberg'e göre eğer geriye gidebilseydik ve Fisher'ın sınamasının "istatistiki açıdan farkedilen", "istatistiki olarak saptanabilen" sınama olarak adlandırsaydık çok daha iyi olurdu. Çünkü bu kavramlar yöntemin bize sadece bir etkinin olduğunu söylediğini, ancak bu etkinin önemi ve boyutu hakkında bir fikir vermediğini daha doğru anlatırdı (121). İstatistiki olarak bir olguyu saptamak için gerekli boyuttan küçük gösteren araştırmalara, bu duruma düşük güçlü araştırma denir. Bu gezegenlere dürbünle bakmaya benzer. Her baktığınızda hiç gezegen olmadığı sonucuna varırsınız. Ancak telekopla bakıldığında gezegenler görünmeye başlar. Öte yandan yüksek güçlü araştırma, daha önce verilen doğum kontrol hapı örneğinde olduğu gibi, gerçekte çok önemli olmayan bir etkiyi olduğundan büyük gösterebilir. Düşük güçlü araştırmaysa küçük etkiyi yanlış olarak gözardı etmeyle sonuçlanabilir.

Bir anlamlılık sınamasının en kaygan felsefi noktası en başta olur. Yani "varsayalım ki geçersizlik denencesi doğrudur." Aslında pek çok durumda ispatlamaya çalıştığımız geçersizlik denencesinin doğru olmamasıdır. Mantıksal olarak hedeflediğimiz şeyin tersini ispat etmeye çalışmamız garip bir durum gibi gelir. Doğru olduğuna inandığımız şeyin yanlışlanması Aristotales'e kadar gider ve çelişkiyle ispat veya "reductio ad absurdum" denir. Eğer bir denence yanlışlık içeriyorsa, denencenin kendisi de yanlış olmalıdır. Dolayısıyla planlama şöyle olur:

- Denence H'nin doğru olduğunu düşünelim
- H denencesinin sonucu belli bir olgu olan F olamaz
- Ama eğer durum F ise
- Bu durum H'nin yanlış olduğunu gösterir

Bir arkadaşınızın Ankara'da 2016 yılında kurşunlanarak ölen kadınların sayısının 212 olduğunu söylediğini düşünün. Bu bir denencedir. Ama bunu doğrulamak mümkün olmayabilir. İnternet tarayıcılarına sorduk ve bir sonuç alamadık. Öte yandan eğer denencenin doğru olduğunu kabul edersek, Ankara'da cinayette ölenlerin sayısının 212'den az olmaması gerekir. Çünkü arkadaşın ileri sürdüğü gibi 200 kadın silahla vurularak öldürüldüyse, Ankara'da tüm nesneleri kullanarak öldürülenlerin sayısının 212'den az olması mümkün değildir. Buna tek tek haberlerden bakıp aslında tüm nesnelerle öldürülen kadın sayısının 88 olduğunu öğrenirsek, arkadaşımızın denencesi yanlışlanır. Burada zihnimizde yarattığımız dünyada yanlış denencenin bir an için öyle kabul ettik ve sonra onun gerçek verilerin baskısı altında parçalanmasına tanıklık ettik 131. Bu şekilde bakarsak, reductio (zıttıyla ispatlamak) önemsiz görülebilir ve bir anlamda ne kadar güçlü bir araç olduğunu bilmeden ne kadar çok durumda kullandığımızı unutabiliriz.

Anlamlılık sınama testini zıddıyla ispatlamanın biraz sislisi olarak düşünmek mümkündür:

- Geçersizlik denencesi H doğrudur.
- Buradan giderek belli bir sonuç O'nun olması pek mümkün değildir (yani, Fisher'ın 0,05 eşiğinden düşüktür)
- Ama O'nun olduğu gözlemlenmiştir.
- Demek ki H'nin olması pek mümkün değildir.

Yukarıdaki prosedüre zıttıyla ispatlama değil de zıttıyla pek mümkün olmadığını göstermek diyebiliriz. Ellenberg'e göre zıttıyla pek mümkün olmadığını göstermeyi anladıysanız, kötü haber şudur: Aristotales'in zıddıyla ispat yöntemine göre genelde kulağa pek mantıklı gelmez (133-135). Mayo Hastanesi Tıbbi İstatistik Bölümü Başkanı Joseph Berkson, yöntemin sallantılı bir yöntem olduğunu kanıtlamak için ünlü bir örnekle tuzakları gösterdi. 50 denek seçtiğiniz ve denenceniz bunların insan olduğudur. Oysa bir kişinin albino (beyaz kaş, kirpik) olduğunu gözlemlediniz. Albino olanların oranı olağanüstü düşüktür ve 20 bin kişinin sadece birinde görülür. H denencesinin doğru olması söz konusuysa, ve 50 kişilik denek grubunda albino olan biri varsa, bunun olma şansı 0,0025'ten küçüktür. Dolayısıyla p-değeri H denencesine göre O gözlemlenmesi olasılığının 0,0025'ten küçük olduğunu göstermektedir. Kaçınılmaz olarak yüksek düzeyde istatistiki bir güvenle H'nin doğru olmadığı sonucuna varırız: Bu gruptaki denekler insan değildir! Dolayısıyla pek mümkün olmadığı ile kesinlikle mümkün olmadığı arasındaki ayrımı yapmalıyız. Bu sınamalar pek mümkün olmaması durumunu gösterir ancak hayatta pek mümkün olmayan şeyler bir anda olabilir. Yani bir hafta içinde 2 günde aynı piyango sayılarının kazanması durumu 300 milyarda birdir (her biri 2 veya tek rakamlı 5 grup sayının olduğu bir piyango çekilişinde). Ama olabilir ve tarihsel olarak da bir kere olmuştur. ABD Kuzey Karolayna Eyalet Piyangosu'nda… . Bu bölümü okumuş olduğunuza göre, Türkiye'de Milli Piyango çekilişinde benzer bir durumda komplo twitleri atanlar arasında olmamanız gerektiğini biliyorsunuz.

Hakemli dergilerin görmediği

Aynı zamanda şunu da anlamamız gerekiyor: Bilimsel olarak p-değeri 0.05'ten küçük olduğu için hakemli dergilerde yayımlanmış olan her araştırmanın sonucu doğru olmayabilir. 0.05'i 20 de bir olarak okumak da mümkündür, 1'i 20'ye bölerseniz 0.05 sayısını bulursunuz. P-değerinin tanımını, eğer geçersizlik denencesi bir araştırma/deney için doğruysa, bu araştırmanın istatistiki olarak anlamlı bir sonuç üretmesi şansı 20'de 1'dir. Eğer geçersizlik denencesi hep doğruysa 20 araştırmadan sadece 1'i yayımlanabilir. Yapılan binlerce araştırma arasında sadece p değeri 0.05'ten küçük olanların yayımlanması, benzer bir araştırmanın p değeri açısından sınıfta kalmış olması böylece gözardı edilmiş olur. Uzmanlar bu durumu araştırmadaki şu veya bu nedene bağlayarak, istatistiki olarak anlamlı olanların geçerli olduğu algısını sürdürmeye devam ederler. Ancak geçtiğimiz 10 yıl içinde muhalif bilim insanları herkesi rahatsız edecek şu mesajı vermeye devam ediyorlar: araştırma sonuçlarının içinde kabul etmek istediğimizden çok daha fazla bilimsel açıdan önemli olmayan okumalar yapıyoruz.

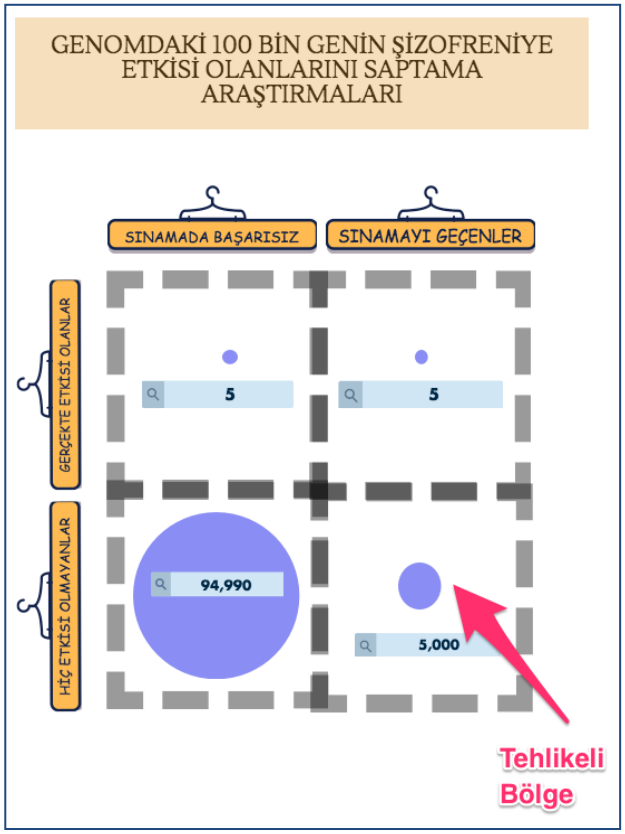
Bu muhaliflerin başında John Ionnidis ve 2005 yılında yayımladığı "Neden Basılan Araştırmaların Büyük Çoğunluğu Yanlıştır?" makalesi geliyor. Biomedikal alanda araştırmalar yapan ve Yunan kökenli metamatik yıldızı olan Ionnidis'in bu makalesi bilim alanında tıbbi araştırmacılarda bir özeleştiri ve savunma dalgasına yol açtı. Ionnidis, makalesinde tıbbi araştırmaların en önemli özelliğinin "geçersizlik" alanları olduğunu ve gerçekte etkisi olmayan etkilerin bulunduğunu ileri sürdü ve bunun ispatlanabileceğini yazdı (Ellenberg:146). İspatlama iddiası her ne kadar tartışılır olsa da önemli noktalara dikkat çekiyordu. Örneğin hastalıkların genetik yapılarla ilişkilendirilmesi araştırmalarına bakıldığında genomdaki çok sayıda genin insanlarda kanser, depresyon veya şişmanlık gibi doğrudan etkiler yaratmasının doğrudan tanınabilir sonuçları olmayacağını ileri sürdü. Ionnidis özellikle şizofreniyle genler arasındaki ilişkiyi kuran araştırmaları mercek altına alıyordu. Rahatsızlığın akrabalarında rahatsızlık bulunan kişilere transfer olduğu hakkında bildiklerimiz, bu tür bir etkinin neredeyse kesin olduğunu gösteriyor. Ancak genomda nerede? Araştırmacılar "büyük veri" olarak tanımlanan bu alanda yüzbin gene bakıp hangilerinin şizofreniyle ilişkili olduğunu görmeye çalışıyorlar. Ionnidis, bunlardan 10'unun tıbbi olarak ilişkilendirilebilecek olduğunu ileri sürüyor. Geriye kanan 99,990'ı? Bunların şizofreniyle hiç bir ilişkisi yok. Ama istatistiki anlamlılık açısından 20'de birinde veya beşbininde p-değeri sınaması eşiği yakalanabilecek. Başka biçimde söylenirse, "Şizofreni genini" bulduğunu söyleyebilecek ve basım hakkı kazanacak araştırmaların saptadıklarına kıyasla 500 kat fazla sahte şizofreni geni saptanmış olacak.

Eğer araştırma yüksek güçlü bir araştırma değilse, bütün bu araştırmalarda gerçek etkili olan genlerin istatistiki olarak anlamsız bulunması da mümkün. Eğer araştırmalar düşük güç araştırmalarıysa, gerçekten fark yaratacak genlerin sadece yarısı anlamlılık sınaması eşiğini geçecek. Bunun anlamı, p-değeri ile şizofreniye etkisi olduğu sertifikasını alacak genlerden sadece 5'i gerçekten bu etkiye sahip olacak. Oysa 5 bin gen sadece şans eseri bu eşiği geçmiş olacak. Sonuç olarak şunun söylenmesi gerekiyor ki, anlamlılık sınaması aslında problem değil. Bu sınama ondan bekleneni yapıyor. Şizofreniye etkisi olmayan genler pek az sınamayı geçiyor, buna karşılık gerçekten bizi ilgilendiren genlerin yarısı bu testi geçebiliyor. Ancak etkili olmayan genlerin sayısının büyüklüğü o kadar yüksekki, yanlış ama testi geçen genlerin sayısı, gerçekten doğru olup da testi geçenler kadar. Asıl tehlikeli bölgeyse şizofreniye hiç bir etkisi olmayan sağ alt kutucuktaki genlerin p-değeri testini geçmesi.

Daha kötüsü de var. Düşük güçlü bir araştırma oldukça yüksek bir etki bulabilir. Ama biliyoruz ki, bu etki varsa bile gerçekte küçük bir etki. Başka bir deyişle bir genin etkisini hassas olarak ölçen bir araştırmada istatistiki olarak anlamlı değil olarak çıkarken, p değerinin 0.05'ten küçük olması sınamasını yanlış olumlu veya genin şizofreniye etkisini çok daha büyük gösteren hakiki olumlu çıkabilir.

Etki boyutlarını ılımlı olarak gösteren düşük güç, küçük araştırmaların yürütüldüğü alanlarda özellikle tehlikeli duruma gelir. Psikoloji Bilimi isimli bir alanın önde gelen hakemli dergisinde, yumurtlama dönemindeki evli kadınların Obama'ya karşı Cumhuriyetçi adayı desteklemelerinin anlamlı olarak mümkün göründüğünü gösteren bir makale yayımlanmıştı. Bu araştırmada yumurtlamanın en tepe noktasında Cumhuriyetçi aday için yüzde 40.4 destek mümkün görülürken, aynı kadınlar yumurtlama dönemleri dışında Cumhuriyetçi adayı yüzde 23,4 oranında oy vereceklerini söylüyorlardı.

Şekil 4-10: Genom ve şizofreni ilişkisindeki durum



Araştırma örneklemi sadece 228 kişi kapsayan küçük bir örneklemdi ama fark çok büyüktü. Öylesine büyüktü ki, p-değeri sınamasını 0.03 ile geçiyordu.

Problem tam da budur — farkın çok yüksek olması. Yani Cumhuriyetçi adayı destekleyen kadınların yarısının, ayın bir yarısında Demokrat Barak Obama için destek vermeleri mümkün müdür (Ellenberg:149-150)? Eğer gerçekten de yumurtalama dönemleri siyaseten sağa kaymaya yol açıyorsa, bunun etkisinin çok daha küçük olması beklenir. Ama paradoksal olarak örneklemin görece küçük boyutu, etkinin gerçekçi değerlendirilmesini sağlayacak denenceyi geçersiz kılacaktır. Diğer bir ifadeyle emin olabiliriz ki bu araştırmadaki bildirilen büyük etki, araştırmada sadece kanal gürültüsüdür. Ama gürültü araştırmacıları gerçek etkinin tersine doğru yönlendirmektedir. Sonuç olarak öylesine bir sonuçla karşı karşıyayız ki istatistiki anlamlılığı yüksek ancak güvenilirliği düşük bir veriyle karşı karşıyayız. Bilim insanları, etkileyici ve çok ses getiren deneysel sonuçların bu deneylerin tekrarlanması durumunda hayal kırıcı sonuçlar ortaya çıkarmasını, "şampiyonun laneti" olarak tanımlarlar. Bu yinelenebilirlik krizi gerçekten bilimsel araştırmalarda önemlidir. 2012 yılında Kaliforniya'daki Amgen firması kanser biyolojisindeki en ünlü 53 deneysel sonucu yineleyecek araştırmalar yaptı. Araştırmalardan sadece 6 tanesini yinelendiğinde aynı sonucu verdi.

Araştırmacılar ve etik

Bilim dünyasında bu krizi katlayacak bazı uygulamalar vardır. Bunlardan biri bilimsel makale yayımlanmasındaki hatalarımızdır. Diyelim ki 20 genetik işaretle belirli bir hastalık arasında bağ kuran bir araştırma yaptınız ve p değeri 0.05'ten küçük çıktı. Artık matematik uzmanı haline geldiğine göre bileceksiniz ki, 20 araştırmadan birisi eğer işaretlerin hiç bir etkisi olmasa bile doğru olacaktır. Eğer aynı araştırmayı, aynı genetik işaretleyicileri 20 bağımsız grup aynı anda gerçekleştirseydi bu gruplardan 19 tanesi istatistiki sınama açısından sınıfta kalacaktı, ancak bir tanesi anlamlılık testinden sınıftan geçecekti. O zaman, yeşil bamyanın sivilceye yol açtığını araştıran 20 gruptan biri başarılı olacak, 19'u istatistiki olarak anlamsız bulunacaktı. Peki bu ilişkisizlik makalesini bugünün bilim dünyasında kim yazabilir 152? Hiç kimse. Çünkü hakemli makale yayımcıları, başarısız olmuş ve istatistiki olarak anlamsız olan bir araştırmaya yayın hakkı vermemektedir.

Bir başka tehlike de insan doğasından kaynaklanır. Baskı altında olan, büyük sonuçlar beklenen araştırma ekiplerinin tüm unsurlarının dürüstlük abidesi olduğunu düşünemezsiniz. Birazcık rakamlarla oynayarak, p-değeri testini istediğiniz yöne çekebilirsiniz. Uri Simonsohn, bunu "p korsanlığı" (p hacking) olarak adlandırıyor (Ellenberg: 152-153). İstatistikçiler sayılara işkence yaptıkça, sayılar istediğimiz sonucu itiraf ederler derler. Araştırmacı, p-korsanlığı yapmazsa bile başka yollar vardır. Bilim insanları sözcük oyunlarıyla araştırmanın anlamlılık sınamasından geçemediğini söylerken başarılı bir araştırma izlenimi verebilirler. "Neredeyse istatistiki olarak anlamlı," "anlamlılığa doğru yaklaşan," "anlamlılıkla uçuca" gibi sözler kullanarak araştırmalarını kamuoyuna sunabilirler 154. Bunun nedeni araştırmacıların kandırma isteği değil, istatistiki açıdan anlamlılık sınamasını geçmeyen araştırmaların yayımlanmaması geleneğidir 155. Oysa bunlar da yayımlanabilmelidir. Sonuçta 0.05 sayısı Fisher tarafından kullanılmak üzere seçilen bir tür sözleşmedir.

Anlamlılık sınaması dışında basit ve popüler bir yol p-değerlerine ek olarak güven aralıklarının da kullanılmasıdı. Ellenberg'e göre güven aralığı sadece geçersizlik denencesinin değil başka alternatiflere de bakılması gerektiğini gösterir (158-159). Diyelim ki çevrimiçi satış yapmak için bir Web sitesi oluşturmak istediğiniz ve elinizdeki 2 arayüzden hangisinin daha çok satış sağlayacağını anlamak istiyorsunuz. Buna A/B sınaması denir. O zaman siteye gelenlerin yarısını A arayüzüne, diğer yarısını B arayüzüne yönlendirebilirsiniz. Satışların B arayüzünde yüzde 10 daha fazla olduğunu buldunuz. Hemen karar vermeyin, çünkü bu sadece rastgele bir yükselme olabilir. Bunu anlamak için p-değeri sınaması yaptınız ve bu sonucun ortaya şans olarak çıkma durumunun 0.05'ten küçük olan 0.03 olduğunu buldunuz dolayısıyla geçersizlik denenceniz olan "Satışların yüzde 10 artmasında sitenin B sitesi olmasının bir etkisi yoktur," denencesi reddedilmiş olur. Burada durmamakta yarar vardır.

Güven aralığı size gerçekte gözlemlenen (yüzde 10 artış) sonuçla, bununla güvenli olarak tutarlı olan denence yüzdelerinden vazgeçmemenizi sağlayabilir. Yani, acaba B sitesinin arayüzü sadece yüzde 5 veya yüzde 25 artış getirebilir mi? Bu örneğimizde güven aralığı +3% ile +17% olabilirdi. Şöyle düşünebiliriz aslında, denencemizde yüzde 10 artış yerine +3% ile +17% arasında artışı koymuş oluruz. Artı işareti, etkinin yönünün olumlu olduğunu gösterir. Geçersizlik denencesi olan 0.XX düzeyinin güven aralığımızda söz geçmemesinin nedeni, sonuçların istatistiki olarak bu bölümde söz ettiğimiz gerçek anlamda etkili olabilme durumudur. Ama eğer -5% ile +5% arasındaki bir güven aralığından söz etseydik, burada sıfır noktası da işin içine girmiş olacaktı. Güven aralığıyla birlikte kullanılan p-değeri tek bir sayıda geçersizlik denencesinin reddedilmemesi durumunu tersine istatistiki olarak anlamlılık sonucuna ulaşabilir.