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?"

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.

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.

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.

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

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.

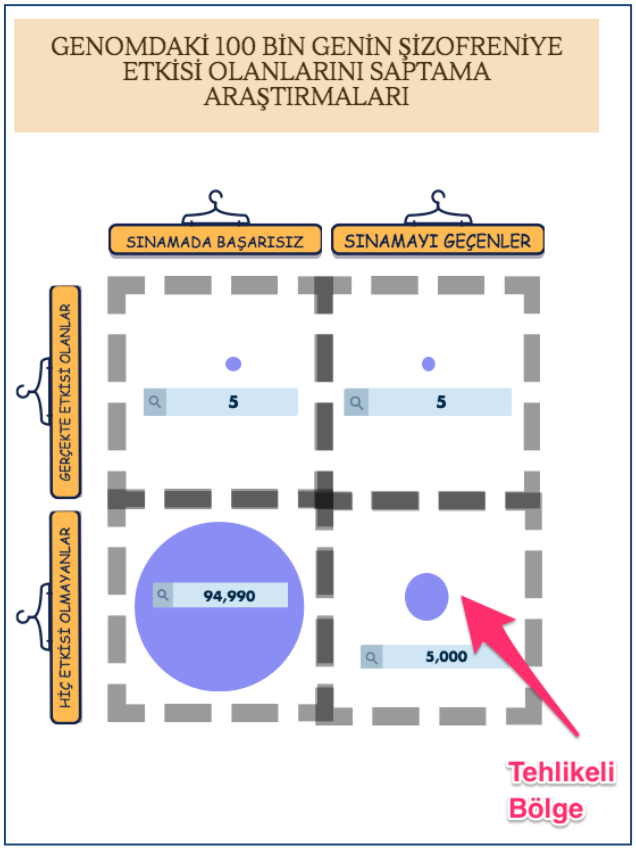
Bu muhaliflerin başında John Ionnidis ve 2005 yılında yayımladığı "Neden Basılan Araştırmaların Büyük Çoğunluğu Yanlıştır?" makalesi geliyor. Biomedikal alanda araştırmalar yapan ve Yunan kökenli metamatik yıldızı olan Ionnidis'in bu makalesi bilim alanında tıbbi araştırmacılarda bir özeleştiri ve savunma 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 : Genom ve şizofreni ilişkisindeki durum



4.2.3. Çaprazlama ve Çapraz Tablolar

Sınamaların ve mutlaka bağıntı kurmaya zorlamanın tehlikelerini gördükten sonra, betimleyici düzeyde olguları sergilemek ama daha fazla iddiada bulunmamak mümkündür. İki değişken arasındaki ilişkiyi çaprazlama (İngilizce cross tabulation) tekniği ile tabloya dökerseniz, yaptığınız bu olur. Bunu gerçekleştirmek için istatistik yazılımının anlayze/descriptives/crosstabs yolunun izlenmesi gerekir (Şekil 4.8 ve Şekil 4.9). Karşınıza çıkan menüden, çaprazlayacağınız değişkenleri ilgili kutucukların içine tıklayarak aldıktan sonra "Ok" tuşuna basınca çapraz tablo ortaya çıkar. Örnek çaprak tablomuzda gelir düzeyiyle, cep

telefonunu kullanarak uzaktan evdeki bazı aygıtların çalıştırılmasına yönelik isteklilik arasındadır (Tablo 4.4). Öncelikle, "İstemem" sütunundaki vurgulanmış hücreleri ele alalım. En üst vurgulu hücrede görünen 825 sayısı, yoklamaya katılanların tamamının bu hizmeti istemeyen alt-orta gelir grubundan bireyler olduğunu anlatmaktadır. "İstemem" sütunundaki Bu durum, "Yoklamaya katılan alt-orta gelir grubundaki 825 bireyin bu hizmeti istemediği" anlamına gelir. Bunun yüzdesini gösteren aynı hizadaki ikinci vurgulu hücreye baktığımız zaman, "Yoklamaya katılan alt-orta gelir grubundakilerin yüzde 44.3'ünün hizmeti istemediğini" anlarız. Aynı hizadaki üçüncü vurgulu hücre, yoklama katılıp da "İstemem" diyenler içinde, alt-orta gelir grubundan olanların oranını gösterir. Bu bilgiyi kullanarak "Yoklamaya katılarak 'İstemem" yanıtı verenler içinde alt-orta gelir grubunun oranı yüzde 29.80'dir" diyebiliriz.

Tablo : Gelir Gruplarına Göre Cep Telefonunu Kullanarak Uzaktan Evdeki Araçların Kontolü Hizmetine Yönelik İstekliliğin İşlenmemiş Verileri

Cep Telefonunu Kullanarak Uzaktan Kombi Gibi Ev Araçlarının Kontrolü					
Gelir Grupları		İsteklilik			
		İstemem	Fark	İsterim	
Alt gelir	Birim sayısı	1385	128	826	2339
	Gelir grubu	59.20%	5.50%	35.30%	100.00%
	İsteklilik içinde	50.00%	39.00%	33.00%	41.80%
	Toplamda	24.70%	2.30%	14.70%	41.80%
Alt orta gelir	Birim sayısı	825	116	920	1861
	Gelir grubu	44.30%	6,20%	49,40%	100.00%
	İsteklilik içinde	29.80%	35,40%	36,80%	33.20%
	Toplamda	14.70%	2,10%	16,40%	33.20%
Orta gelir	Birim Sayısı	325	47	435	807
	Gelir grubu	40.30%	5.80%	53.90%	100.00%
	İsteklilik içinde	11.70%	14.30%	17.40%	14.40%
	Toplamda	5.80%	0.80%	7.80%	14.40%
Üst orta gelir	Birim sayısı	177	27	236	440
	Gelir grubu	40.20%	6.10%	53.60%	100.00%
	İsteklilik içinde	6.40%	8.20%	9.40%	7.90%
	Toplamda	3.20%	0.50%	4.20%	7.90%
Üst gelir	Birim sayısı	60	10	85	155
	Gelir grubu	38.70%	6.50%	54.80%	100.00%
	İsteklilik içinde	2.20%	3.00%	3.40%	2.80%
	Toplamda	1.10%	0.20%	1.50%	2.80%
	Birim Sayısı	2772	328	2502	5602
	İsteklilik içinde	100.00%	100.00%	100.00%	100.00%
	Toplamda	49,50%	5.90%	44,70%	100,00%

İkinci olarak "Farketmez" sütunundaki vurgulanmış hücrelere odaklanalım. Birinci hücrede görünen 328 sayısı, yoklamaya katılanların tümü arasında farketmez diyenlerin sayısını göstermektedir. Bu sayı tüm katılanlara oranla yüzde olarak ikinci vurgulu hücrede yer almaktadır (yüzde 5,9). Tüm katılımcıların toplam sayısıysa en sağdaki sütunun en altındaki vurgulu hücrede bulunmaktadır. Bu sütunda bulunan en yukarıdaki vurgulu hücrede yazan 1861 sayısı aslında alt-orta gelir grubundakilerin toplam sayısını vermektedir. Bu grubun toplam katılımcılara oranı aynı sütundaki üçüncü ve dördüncü hücrede yazmaktadır (yüzde 33.2). Verilen tablo örnek çalışma tablosudur, metnin içinde bu tablo ayıklanarak kullanılır. Tablo 4.5, istemem diyenler arasında en yüksek oranın, alt gelir grubunda olduğunu ortaya koymaktadır (%24.7) . Tablo yüzde olarak verildiği için toplam katılımcı sayısı, tablonun altındaki N (birim sayısı) verilmiştir.

Cep Telefonunu Kullanarak Uzaktan, Kombi Gibi Ev Aygıtlarını Kulanmaya İsteklilik (%)			
Gelir Grupları	İsteklilik		
	İstemem	Fark etmez	İsterim
Alt	24,70	2,30	14,70
Alt-Orta	14,70	2,10	16,40
Orta	5,80	0,80	7,80
Orta-Üst	3,20	0,50	4,20
En-üst	1,10	0,20	1,50
Genel Toplam	49,50	5,90	44,70