***Vaka Takdimi: Doğru Bilinen Yanlışlar Serisi (2): “P” ve Hipotez Doğrulamak!*[[1]](#footnote-1)**

**Prof. Dr. Atila Yüksel**

**Adnan Menderes Üniversitesi**

**Vaka Takdimi**

Meşhur " olasılık (*p) değerine*" (*probablility value*) değinmeden, doğru bilinen yanlışlar serisine devam etmek mümkün değil. Neden mi olasılığı mercek altına alıyorum? Kendisi küçük etkisi büyük de ondan! Hani bilim insanlarına çalışma sonuçlarını genelleme şansı veren ve genelde yanlış anlaşılan, kendisinin de hatalı olma ihtimali unutulan, şu “başka etkenlere göre değişen”, **şartlı** “olasılık” değeri. Aşırı hızdan, başkasını taklit etmekten dolayı çıkan sonuçların doğru olup olmadığı tam da denenmeden, sonuçların kendini tekrar edip etmediğine bakmadan araştırmalardan bir kesinlik vurgulamasıyla genellenebilir yargılar çıkıveriyor. Bu “kesinlik” saplantısı nedense *antaoinizmin* bir tezahürü gibi. İşte olasılığa göre denenmiş, test edilmiş “bilimsel kesin” bulgulardan bazıları... Güzel (yakışıklı) ebeveynlerin (anne-babanın) çirkin ebeveynlere  göre kız çocuğu sahibi olma ihtimali daha yüksektir (Kanazawa 2006). Köri yemek beynin bilişsel kapasitesini yükseltir (Tze-Pin vd. 2006). Video oyunu oynayan cerrahlar ameliyatta daha başarılı performans sergiler (Rosser  vd 2007). Solak çocukların gelecekte alacakları maaş, sağlak çocuklara göre daha yüksektir  (Ruebeck vd. 2006)!!!

Bu sonuçları ortaya koyan olasılık testleri “0.05” ya da “0.01” düzeyinde denenmiş, istatistiksel açıdan “anlamlı” fark bulunmuş, farktan hareketle köri yemeyenlere kıyasla köri yiyenlerin beyinlerinin daha iyi çalıştığı iddia edilmiştir. Peki, bütün testlerde, hesaplamalarda hata ihtimali var da **olasılık testlerinin hatalı olma ihtimali** yok mu? Her hesap gibi, istatistiksel olasılık testinin de hatası varsa ve bunu “0.05” gibi masum bir düzeyde olduğunu varsayarsak, o zaman olasılık testiyle yola çıkan yayınlanmış her yirmi çalışmadan birinin sonucunun belki de yanlış olduğunu da kabul etmemiz gerekmez mi? Meraklanmayın, bende hipotez testinde diğerleri gibi aynı yolu izledim. Örneklemler, anketler oluşturdum, veri topladım, iki gurup, kavram vb arasında analizler sonucunda ilişki ya da fark buldum. Kavramlardan birinin diğerini etkilediği kanaatine vardım. Ama gerçekte ya fark yoksa ya da olasılık değerine bakarak bulduğumu iddia ettiğim etkinin uygulamada bir anlamı yoksa endişesi beni sardığında hipotez testine yönelik diğer bilim alanlarında giderek yükselen eleştirilere kulak vermeye başladım. Ardı ardına sorular geldi…

**Araştırmacının test ettiği gerçek evrensel hipotez nedir?**

Hipotezin doğrulanması/yanlışlanmasında başvurulan bu “olasılık değerini” (0.01, 0.05) belirleyen ne? Asıl olan “Tip1” ve “Tip2” olarak adlandırılan hipotezlerin testi değil midir? Yani “*bir şey buldum, bu doğru mu yoksa tesadüfi mi?”* ve “*bulduğum şeyin uygulamada etkisi gerçekten var mı?*” Bu sorular terk edilmiş gibi.

Gerçekte doğruyken yanlış bir şekilde hipotezi ret etmek ya da gerçekte yokken yanlış bir şekilde fark ya da ilişkiye var diyebilme olasılığı üzerinde neden durmuyoruz? Araştırmaya başlamadan gerçek olmayanı yanlışlıkla doğrulama olasılığımızı, yani Tip1 hata yapma olasılığımızın %5 düzeyinde kalacağını kabul edersek, fark yoktur sonucu çıktığında bu %95 olasılıkla fark yoktur anlamına geliyorken gerçekte var olan farkı saptayamama (Tip2) ihtimalimizin %20’lerde olacağını da düşünmemiz gerekir. Sadece olasılık değeri, sıfır (farksızlık) hipotezin yanlışlanması için yeterli mi diye sormalıyız artık! Çalışma yeterli örneklem sayısına (ki nasıl seçildiği de genelde bir muamma) sahip değilse, sonuçların gerçekten istatistiksel olarak önemli olup olmadığını saptamak mümkün mü? Örneklem sayıları gerçekçi mi? Şu meşhur örneklem belirleme formülüne bir bakmanın ve formüllerin neden her yerde sihirli bir şekilde 386 sayısını çıkarttığını tartışmanın zamanı gelmedi mi? Örneklemde sayının nasıl belirlendiğinde kalıyorum, örneklemin nasıl seçildiğine, yanıtlayıcı psikolojisine girmiyorum bile!

“*İstatistiksel anlamda önemli*” ya da “*olasılık değerine göre*” ifadesinin kullanımı ve bir hipotezin doğrulanması/yanlışlanmasında tek başına kanıt olarak sunulması sosyal bilimlerde artmıştır (Armstrong, 2007; Yüksel ve Tufan, 2017). Sosyal bilimlerdeki yüksek kullanımına karşın, “*olasılık testi*” fen bilimlerinde (ör., fizik, kimya) nadiren kullanılır (Schmidt & Hunter, 1997). Oldukça fazla sayıda bilim insanı tarafından olasılık değerinin hipotezin kanıtı olarak kullanımı “bilim dışı” kabul edilmektedir (Armstrong, 2007; Hubbard ve Armstrong, 1997; Wright ve Armstrong, 2008). Bazıları bu yaklaşımın hipotez testinde ***yasaklanması gerektiğini*** şiddetle tavsiye etmektedir. Bazıları ise hipotezleri test etmede dikkatli olunmasını desteklemektedir (Chow, 1998). Kısacası, “p değeri” giderek ŞİDDETLENEN bir tartışmanın ortasındadır. “Tek başına p değeri, hipotezin doğrulanması/yanlışlanmasında kanıtların rasyonel bir ölçümü olarak kabul edilemez. Ayrıca, tek başına bir p-değerinin, etkinin büyüklüğü hakkında hiçbir şey söylemediği de unutulmamalıdır” (Lecoutre, Poitevineau & Lecoutre, 2005, s. 252). Kısaca, **p değeri** sıfır hipotezini yanlışlamak için yeterli olmayabilir. Gerçekte p değeri “sonucun örnekleme hatasının bir sonucu olup olmadığının ölçüsüdür. Örneğin p <.05 sonucu, bu sonucun tesadüfen gerçekleşme olasılığının % 5'ten az olduğunu ifade eder” (Kline, 2004, s. 63). Hipotez hakkında doğrudur ya da yanlıştır diyemez. Nickerson'ın dediği gibi “p” sıfır hipotezinin testinde “şansa karşı bir ihtimal” vehiminin doğru olma olasılığını temsil eder (Nickerson, 2000, s. 247). Nickerson (2000) şöyle uyarıyor: “pek çok araştırmacının inandığının aksine, sıfır hipotezinin istatistiksel testinden elde edilen p değeri, H0'ın doğru olma olasılığı değildir. Sıfır hipotezini reddetmek için, diyelim ki bulgumuz p <.05 i gösteriyor, bu sıfır hipotezin gerçek olma olasılığının .05 veya daha az olduğu anlamına gelmez. Dahası, p değeri, sıfır hipotezin gerçek olma olasılığını temsil etmediği için, yanlışlandığında sıfır hipotezinin tamamlayıcısı olarak görülen alternatif hipotezin (araştırma hipotezi) doğru olduğu anlamına da gelmez ”(s. 246). Ama nedense araştırma hipotezinin doğru olduğu sıfır hipotezinin yanlışlığına bağlayanların sayısı hiç az değil.

İstatistiksel olasılık, kullanıcılar, hakemler, okuyucular ve editörler tarafından hatalı yorumlanmasından dolayı sorunlu olmasına rağmen bu sorun göz ardı edilmekte, doğru bilinen yanlış geleneği devam etmektedir (Armstrong, 2007, s. 322). Liu (2013), bunun sorumlusunun “p” değeri olmadığını iddia etmesine karşın, yükselen eleştirinin hipotez test etme prosedürünün kötüye kullanımının gittikçe artmasından kaynaklandığını ileri sürmektedir: “Eğer prosedür doğru bir şekilde kullanılmıyorsa, o zaman doğal yol, olasılık testini top yekün terk etmek yerine yanlış kullanımı düzeltmek olmalıdır” önerisinde bulunmaktadır. Liu (2013), “hipotez testindeki p değerinin, bir etkinin ya da farkın gerçek boyutunu göstermediği halde, tesadüfi etkinin makul bir açıklaması olarak şans faktörünü dışlamamıza izin verdiğini” belirtmektedir (Liu, 2013). Yine de, diğerleri doğru bir şekilde yapıldığında bile “olasılık testleri tehlikelidir” (Armstrong, 2007, s.2) denmektedir. Schmidt ve Hunter (1997, s. 38) “olasılık testine güvenmenin mantıksal olarak savunulamaz olduğunu ve bu testin sosyal bilimi geliştirmeyi zorlaştırarak, araştırma girişimini geciktirdiğini” belirtmiştir. Armstrong (2007) ile uyumlu olarak Hunter’a göre (1997, s. 3) “şu anda kullanılan olasılık testi bir felakettir. Araştırmacıların çoğu, olasılık testinin% 5'lik bir hata oranına sahip olduğuna inanmakla birlikte, ampirik çalışmalar, psikolojideki ortalama hata oranının, araştırmacıların düşündüğünden, %60 düzeyinde, 12 kat daha yüksek olduğunu göstermektedir .”

Özetlemek gerekirse bu “p” totemine yönelik iki eleştiri var. Birincisi, “p değeri” nin sıfır hipotezinin doğru bir şekilde reddetmesi/desteklemesinde kullanılması tartışmalıdır çünkü olasılık bulgudaki şans faktörünün ifadesidir. İkinci eleştiri, “acele” davranılarak, “güç analizi (statistical power)” ve “etki seviyesi (effect size)” gibi destekleyici kanıtları dikkate almadan, sadece olasılık değeriyle sıfır hipotezini reddetme / destekleme geleneğiyle ilgilidir. Olasılık testinin hipotez hakkında karar vermede yeterli olmadığı diğer bilim alanlarında yüksek sesle dile getiriliyor. Farkın, tesadüfi olduğu, şans eseri bulunup bulunmadığı ya da gerçekte var olup olmadığını anlamak için daha fazlası gerekirken turizm araştırmacılarındaki sonuca yönelik bu acelecilik neden? Hemşirelik alanında yapılan çalışmalarda olasılık değerinin yetersizliğinin farkına vararak, istatiksel güç ve etki seviyesinin kullanılması gerektiğine yönelik benzer bir saptama yapan Çapık (2014, s. 268) “En yalın tanımı ile istatistiksel güç, analizlerde anlamlı bir fark saptanması durumunda bu farkın gerçek olabilme ihtimalidir. İstatistiksel güç, örneklem sayısından etkilendiği için araştırmacılar örneklem sayısını güç analizi ile hesaplayabilmektedirler. Bir çok akademik derginin çalışmanın gücünü istemesi, etik komitelerin örneklem sayısı hakkında açıklama istemesi ve araştırmacıların daha kesin bulgular sunmak istemesi gibi faktörler güç analizinin daha fazla dikkate alınmasını gerektirmektedir” demektedir. Çapık gibi artan sayıda bilim insanı, yetersiz etki büyüklüğüne ve yetersiz istatistiksel güce sahip ama anlamlı “olasılık” değerinin olduğu çalışmaların tartışmalı olduğunu savunmaktadır (Prajapati vd., 2010). P değeri başka kanıtların desteği olmaksızın anlamsızdır. Çalışmanın yetersiz istatistiksel gücü olduğunu umursamayan bir araştırmacı, anlamlı düzeyde bulduğu p değeri sonucunda sıfır hipotezini yanlış bir şekilde reddedecektir (Cohen, 1988). Sıfır hipotezi hakkında verilen acele ve yersiz kararlar bilimi ileride hayal kırıklığına uğratacaktır. Araştırmada örneklem büyüklüğü ve istatistiksel güç, bir popülasyondaki istatistiksel olarak anlamlı küçük farklılıkları tespit etmek için yeterli düzeyde ayarlanmamışsa, bu hipotez reddetme kararlarıyla ilgili kaygı uyandırmalıdır. Büyük farklılıkların bulunmaması, küçük ve anlamlı farklılıkların mevcut olmadığı anlamına gelmeyecektir. Küçük etkilerin (farklar) saptanması, daha fazla bilginin (örneklemin) toplanmasına ihtiyaç duyacağından, büyük etkilere göre daha zor tespit edilecektir. Kısaca, daha fazla örneklemle gözlem yaparak, parametreler ekleyerek ve daha iyi bir model oluşturarak analizler yapılmalıdır (Westland, 2010, s. 6).

İstatistiksel güç hakkında bilgi olmadan, sıfır hipoteziyle ilgili kararları desteklemek veya reddetmek için sadece p değerine dayanarak sağlam sonuçlar çıkarmak büyük bir eksikliktir. Peki turizm araştırmacıları bu eksikliğin farkında mıdır? Yakın bir zamanda merakımıza yenik düşerek bir meslektaşımla çalışma yaptık (Yüksel ve Tufan, 2017). Uluslararası prestijli dergilerde yayınlanmış, özellikle sıfır hipotezini test ettiğini vurgulayan 155 makaleyi bu hipotez testlerini nasıl yaptıkları, hangi kanıtlarla kararlarını destekledikleri açısından inceledik. İlginç bir şekilde hipotez test ederken 155 makale içinde sadece 2 makalenin etki seviyesi ve istatiksel güç kavramını dikkate aldığını saptadık. İstatiksel güç hakkında bilgi sunmayan 153 yayında belirtilen örneklem sayısı vb parametrelere göre yaptığımız istatiksel güç hesaplaması sonucunda hiçbir çalışmanın istatiksel gücünün (ortalaması 0.37!) Cohen tarafından önerilen 0.80’i aşamadığını fark ettik. Bu şu anlama gelmektedir. İncelenen çalışmalarda, gruplar arasında anlamlı olduğu düşünülen farkların gerçek olabilme ihtimali **%63 ORANINDA TARTIŞMALIDIR**. Gerçekte yanlış olan H1 (araştırma, alternatif) hipotezi ret edilmeyerek Tip2 hata işlenmiştir. Hesaplamalarımızda yüzde 10 hata olma ihtimali de olsa yüzde elliler önemli bir oran! Yani yayınlanmış olsa dahi her iki makaleden biri şüpheli olabilir. Yayınlanan çalışmalar, yeterli düzeyde istatistiksel güç ve etki büyüklüğüne sahip değilse, salt “olasılık değeri” bulgusuna dayanmışsa turizm, seyahat ve konaklama disiplininde yer alan bilgilerin doğruluğu tartışmaya açıktır. O nedenle *“tüm bulguların %90’ının yanlış olduğunu söylemek yanlış değildir”* diyen Kollat, Engell ve Blackwell’e (1972), *“modern bilimin ürettiği bilginin büyük bir bölümü yanlıştır”*  deme cesaretini gösteren Ioannidis’e (2005) katılmamak elde değil.

Amerikan Psikoloji Derneği'nin (APA) İstatistiksel Çıkarımına İlişkin Görev Gücü, araştırmacıların “bir p değerini bildirirken her zaman bir etki seviyesi tahmini sunmalıdır” uyarısının üzerinden neredeyse 20 yıl geçmesine karşın turizm araştırmalarında daha araştırmayı planlarken ya da uygularken istatistiksel gücü ve etki seviyesini umursamamak dikkat çekicidir (Wilkinson ve APA İstatistiksel Çıkarım Görev Gücü, 1999, s. 599). Ayrıca, “… etki seviyesinin daha önce bildirilen etkiler bağlamında raporlanması ve yorumlanması, iyi bir araştırma için şarttır” (s. 599). Benzer şekilde, APA (2001) Yayın El Kitabının 5. Baskısında etki seviyesinin kritikliğini vurgulayarak, “okuyucunun bulgularınızın önemini tam olarak anlayabilmesi için, bulgular bölümünüzde bazı etki seviyesi endekslerini ya da ilişki gücünü de dahil etmek hemen hemen her zaman gereklidir ”(s. 25) uyarısında bulunmasına karşın üst düzey dergilerdeki yayınlara bakıldığında yayınların üçte ikisi hiç bir zaman istatiksel güç analizi kullanmamıştır (Mone ve ark. 1996, s.110); “yapılan çalışmaların % 5'inden az kısmında güç düzeyi incelemesi yapıldığı”, hipotez test ederken hemen hemen hiçbir yayınlanmış çalışmanın güç seviyelerini hesaplanmadığı görülmektedir” (Cashen ve Geiger, 2004, s.152). Daha vahim bir uyarı Deng’in sonuçlarında görülmektedir. Deng (2015, s.4) incelediği yayınlarda yer alan 2629 olasılık testinin,% 91'inin Cohen'in küçük etkilerin tespit edilmesi için önerdiği etki seviyesi olan 0.80'nden düşük olduğunu, yayınlanmış çalışmaların %12 sinin büyük etkileri bulmak için önerilen güç seviyesinin altında olduğu, yayınların % 47'sinde hesaplanan güç seviyesinin 0.30 larda kaldığı halde bulunan farkların gerçekçi olduğu sonucuna varıldığını rapor etmektedir.

Şimdi eminim “peki ama ne yapmalı?” sorusu kafaları meşgul ediyor.

Kendime tavsiyem “Ölçüm bir teoridir, henüz kesin değildir. Bilimin doğumevi ise gözlemdir. Daha iyiyi bulana kadar p değerine dayalı ampirik çalışmaya ara vermek” oldu.

**Okuma Listesi**

Armstrong,J. S., (2007), “Significance Tests Harm Progress in Forecasting.”International Journal of Forecasting 23, 321-327

Bruno Lecoutre, Jacques Poitevineau, Marie-Paule Lecoutre. (2005). A reason why not to ban Null Hypothesis Significance Tests. Revue MODULAD.

Çapık, C. (2014). İstatı̇stı̇ksel güç analı̇zı̇ ve hemşı̇relı̇k araştırmalarında kullanımı: temel bı̇lgı̇ler. Anadolu Hemşirelik ve Sağlık Bilimleri Dergisi,17:4

Chow, S.L.,(1998). “Pr´ecis of Statistical significance: Rationale, validity, and utility.”Behavioral and Brain Sciences, 21: 169–239.

Cohen, J.(1988). “Statistical Power Analysis for the Behavioral Sciences.” Hillsdale, NJ: Erlbaum

Deng, H.,(2005).“Does It Matter If Non-Powerful Significance Tests Are Used in Dissertation Research?”Practical Assessment, Research & Evaluation 10(16): 1-13

Hubbard, R. and Armstrong, J. S., (1994). “Replication and extensions in marketing: Rarely published but quite contrary.” International Journal of Research in Marketing 11: 233–248.

Hunter, J. E., “Needed: a ban on the significance test.(1997).” Psychological Science 8(1):1-5.

Ioannidis JPA (2005) Why Most Published Research Findings Are False. PLoS Med 2(8): e124.

Ioannidis, J. P. A. (2007). Limitations are not properly acknowledged in the scientific literature. Journal of Clinical Epidemiology, 60, 324–329.

Kline, Rex B., (2004). “Beyond significance testing; reforming data analysis methods in behavioral research.” Am. Psychological Assn Books, 69-91.

Kollat, D. T., R. D. Blackwell, and J. E. Engel.(1972). “The Current Status of Consumer Behaviour Research: Development during the 1968-1972 Period.” In Proceedings of the 3rd Annual Conference of the Association of Consumer Research, edited by M. Venkatesan, pp. 576-84

Liu, X. S., (2013). “Comparing Sample Size Requirements for Significance Tests and Confidence Intervals.” Counseling Outcome Research and Evaluation 4(1): 3-12

Mortel, F. (2008). Faking it: social desirability response bias in selfreport research. Australian Journal of Advanced Nursing, 25(4): 40-48

Prajapati, B.; Dunne, M and Armstrong, R., (2010). “Sample size estimation and statistical power analyse.” Optometry Today no. (July 2010).

Raymond S. Nickerson (2000). Null Hypothesis Significance Testing: A Review of an Old and Continuing Controversy. Psychological Methods. 5, 2: 241-301

[Rosser JC Jr](https://www.ncbi.nlm.nih.gov/pubmed/?term=Rosser%20JC%20Jr%5BAuthor%5D&cauthor=true&cauthor_uid=17309970), [Lynch PJ](https://www.ncbi.nlm.nih.gov/pubmed/?term=Lynch%20PJ%5BAuthor%5D&cauthor=true&cauthor_uid=17309970), [Cuddihy L](https://www.ncbi.nlm.nih.gov/pubmed/?term=Cuddihy%20L%5BAuthor%5D&cauthor=true&cauthor_uid=17309970), [Gentile DA](https://www.ncbi.nlm.nih.gov/pubmed/?term=Gentile%20DA%5BAuthor%5D&cauthor=true&cauthor_uid=17309970), [Klonsky J](https://www.ncbi.nlm.nih.gov/pubmed/?term=Klonsky%20J%5BAuthor%5D&cauthor=true&cauthor_uid=17309970), [Merrell R](https://www.ncbi.nlm.nih.gov/pubmed/?term=Merrell%20R%5BAuthor%5D&cauthor=true&cauthor_uid=17309970). (2007) The impact of video games on training surgeons in the 21st century. [Arch Surg.](https://www.ncbi.nlm.nih.gov/pubmed/17309970) 142(2): 181-6;

Rossiter, J. R. (2011). Measurement for the Social Sciences: The C-OAR-SE Method and Why It Must Replace Psychometrics. New York: Springer

Ruebeck, C. S., Harrington, J. J. E., & Moffitt, R. (2007). Handedness and earnings. Laterality,12, 101–120.

Satoshi Kanazawa (2007) Beautiful parents have more daughters: A further implication of the generalized Trivers–Willard hypothesis (gTWH). Journal of Theoretical Biology 244;133–140

Schmidt, F. L. and Hunter, J.,(1997). “Eight common but false objections to discontinuation of significance testing in the analysis of ersearch data, In What if there were no significance tests?” Harlow, L., Mulaik, S. Steiger, J. (Eds). NJ: Lawrence Erlbaum Associates Publishers, 37-64.

Tze-Pin Ng Peak-Chiang Chiam Theresa Lee Hong-Choon ChuaLeslie Lim Ee-Heok Kua. Curry (2006), consumption and cogitive function in the elderly. American Journal of Epidemiology, e 164, Issue 9, 1 898–906.

Yüksel, A. Ve Tufan, E. (2017). An Analysis of “Publication Bias” in the Travel, Tourism and Hospitality Research.Int. Journal of Applied Behavioural Economics. 6, 19-37.

Westland, J. C., (1999), “Lower bounds on sample size in structural equation modeling.”Electronic Commerce Research and Applications 9: (2010), 476–487.

Wilkinson, L. and Task Force on Statistical Inference, APA Board of Scientific Affairs,“Statistical Methods in Psychology Journals: Guidelines and Explanations.”American Psychologist 54: 594-604.

Wright, M. and J. S. Armstrong, (2008). “The Ombudsman: Verification of Citations: Fawlty Towers of Knowledge? (with commentary and reply).”Interfaces 38(2): 125-139.

1. Yüksel, A. (2018). Aykırı: Araştırma Yanılsamaları. Detay Yayıncılık. [↑](#footnote-ref-1)