Elementary Education Online, 10(3), 1112-1123, 2011.
İlköğretim Online, 10(3), 1112-1123, 2011. [Online]: http://ilkogretim-online.org.tr

On Hypothesis Testing, Confidence Interval, Effect Size and Noncentral Probability Distributions

Hüseyin Hüsnü YILDIRIM[1], Selda YILDIRIM[2]

ABSTRACT. The research provides evidence that null hypothesis significance testing (NHST) is frequently misunderstood and, consequently, misused by researchers. Sources of these misunderstandings can be stated as: 1) Interpreting NHST as a probability on accuracy of null hypothesis, 2) regarding NHST as an indicator of effect size, and 3) overlooking the fact that NHST is just an analysis on means. This study investigates these misuses and discusses confidence interval, effect size and noncentral probability distributions through their possible contributions on resolving the misuses.
Key Words: Hypothesis testing, confidence interval, effect size, noncentral probability distributions

SUMMARY

Purpose and significance: Null hypothesis significance testing (NHST) has been an essential part of studies in behavioral and educational research areas. However, almost by the time it was introduced by Ronald Fisher in 1925, NHST has been surrounded by severe critics. Some of these critics were so crucial due to their claim that NHST had a deteriorating effect on foundations of science. For example, some critics argued that due to the dominancy of NHST in research methodologies, information from various studies could not accumulate to provide insights on psychological or educational theories. These critics have continued, leading a considerable amount of literature on limitations of NHST. The article by Berkson (1938) is one of the earliest challenges to the NHST. Articles by Cohen (1990, 1994) and Kirk (1996) may be regarded as classics bringing the shortcomings of NHST to the attention of research community. However, despite the generous effort of these researchers to point out serious flaws in use of NHST, latest research shows that dominancy of NHST still continues (Alhija & Levy, 2009). And to make the things worse, there is also evidence that NHST is frequently misunderstood and, as a result, misused by researchers. So, articles pointing out the limitations of NHST and providing suggestions on ways to supplement NHST still seem necessary. Thus, the purpose of this article is twofold. First, it summarizes the major criticisms of NHST. Second, it provides information on ways to supplement NHST.

Discussion: Three major criticisms are discussed in the article. The first criticism is that NHST does not tell anything about the probability that null hypothesis is true given the set of data obtained from the sample. However it is frequently used in that sense. The second criticism of NHST is on its triviality. In fact, a null hypothesis is always false, and any attempt failing to falsify the null hypothesis is due to shortage of sample size. However, it is not uncommon interpreting the failure in rejecting the null hypothesis as a sign that treatment under investigation is not useful in the real world. The last criticism is that the results from NHST are usually overgeneralized. NHST uses cumulative statistics such as sample mean. Thus, individual differences or multiple aspects of treatments used in the study may easily fade away within these generalized statistics. Too much focus on NHST may create an illusion that inference on reality is a simple dichotomous endeavor of rejecting or not rejecting the null. Three ways to supplement NHST are provided in the article. These are, using confidence intervals, using measures of effect magnitudes and, in order to take the combined advantage of both, calculating confidence intervals for the effect size measures. The later one requires the use of noncentral probability distributions.

Conclusion: Statistical techniques provide ways to make inferences about different aspect of reality through the data collected in research studies. Investigating these techniques in a way just to increase researchers' procedural fluency is not enough. Researchers should also gain conceptual understanding on statistical techniques. Critical evaluation of the research methodologies may supply important contributions to this purpose.

[1] Abant İzzet Baysal Üniversitesi, Eğitim Fakültesi Oda No:326, Gölköy /Bolu, Tel: 374 254 10 00 (1719), yildirim.huseyin@ibu.edu.tr
[2] Abant İzzet Baysal Üniversitesi, Eğitim Fakültesi Oda No:326, Gölköy /Bolu, Tel: 374 254 10 00 (1719), cet_s@ibu.edu.tr

Hipotez Testi, Güven Aralığı, Etki Büyüklüğü ve Merkezi Olmayan Olasılık Dağılımları Üzerine

Hüseyin Hüsnü YILDIRIM[3], Selda YILDIRIM[4]

ÖZ. Araştırmalar, yokluk hipotezi anlamlılık testinin (YHAT), çoğunlukla yanlış anlaşıldığını ve yanlış kullanıldığını göstermektedir. Bu yanlışlıklar üç ana başlıkta toplanabilir: 1) YHAT'ın, yokluk hipotezinin doğru olma ihtimali olarak ele alınması, 2) YHAT'ın, etki büyüklüğü hakkında bilgi verdiğinin sanılması ve 3) YHAT'ın sadece ortalamalar üzerinden bir hesaplama olduğunun gözden kaçması. Bu çalışmada, söz konusu yanlışlıklar üzerinde durularak güven aralığı, etki büyüklüğü ve merkezi olmayan olasılık dağılımlarının kullanılmasının, bu yanlışlıkların giderilmesine olan katkıları ele alınmıştır.

Anahtar Sözcükler: Hipotez testi, güven aralığı, etki büyüklüğü, merkezi olmayan olasılık dağılımları

GİRİŞ

Bilimi diğer düşünme disiplinlerinden ayıran temel özelliklerden biri, bilimin olgusallığıdır (Yıldırım, 1996). Diğer bir deyişle, bilim, gözlemlenebilir olgusal verilere dayanarak çıkarımlarda bulunmaya, genellemelere ulaşmaya dayalı bir düşünme disiplinidir. Bu çıkarımların nasıl yapılacağını, genellemelere nasıl ulaşılabileceğini belirleyen metodolojinin belkemiğini ise istatistik oluşturur. Olasılık teorisine dayanan istatistik bu sebeple, bilimin mantığı olarak da betimlenir (Jaynes, 2003).

Yokluk hipotezi anlamlılık testi (YHAT), çıkarımsal (inferential) istatistikte sıklıkla kullanılan bir yöntemdir. Bu yöntemde, yokluk hipotezinin geçerli olduğu varsayıldığında, araştırmada elde edilen verinin (örneğin, deney ve kontrol grupları arasındaki ortalama farkının), görülme olasılığı hesaplanır (istatistik yazılımlarının çıktılarında, çoğunlukla *p* veya *significance* değeri olarak gösterilen değer). Elde edilen bu olasılık değeri "çok küçük" ise (örneğin, çoğunlukla kabul edildiği üzere, α=0,05'ten küçükse), eldeki verinin, yokluk hipotezinin geçerli olmayabileceğini gösteren bir kanıt olduğu kabul edilir ve buna dayanarak, yokluk hipotezi reddedilir (Fisher, 1925). Bu yöntem, ilerleyen bölümlerde daha detaylı bir şekilde ele alınmaktadır.

YHAT'ın temelleri Ronald Fisher (1925) tarafından atılır. Üç yıl sonra da, Jerzy Neyman ve Egon Pearson (1928) Tip 1 ve Tip 2 hata kavramlarını tanımlarlar. Böylece neredeyse bugünkü şeklini alan YHAT, ortaya atıldığı günden itibaren genel bir kabul görür ve yaygın bir şekilde kullanılmaya başlanır. Cohen (1990) bunu YHAT'ın cezbedici bir özelliğine bağlar; Cohen'e göre, YHAT'ın belirlenimci (determinist) bir algoritmaya sahip olması, aslında karmaşık bir süreç olan gözlem verilerine dayalı karar vermeyi, ret-kabul cevaplı mekanik bir probleme indirgemiştir. Bu sayede YHAT, kolay kullanır ve kolay öğretilir bir yöntem olarak hızla yayılmıştır.

Ne var ki, YHAT bir yandan hızla yayılmaya devam ederken, diğer yandan YHAT'a karşı eleştiriler de ortaya çıkmaya başlar. Eleştiriler, YHAT'a dayanarak verilen kararların geçerliğini sorgulamakta, sadece YHAT dikkate alındığında, karar verme sürecini etkileyebilecek diğer bazı önemli değişkenlerin gözden kaçtığına dikkat çekmektedir. Joseph Berkson'ın (1938) makalesi YHAT'ın bu bağlamdaki ilk sistemli eleştirilerinden biri olarak görülür.

Yukarıda da belirtildiği gibi, YHAT'a yöneltilen eleştiriler, analizlerde YHAT'ın kullanılmasına değil, sadece YHAT'ın kullanılmasına karşı çıkmaktadır. İstatistiği sadece YHAT'tan ibaret görmenin, bir araç olması gereken istatistik analizlerini, istatistiksel anlamlılık peşinde koşulan bir amaca dönüştürdüğü uyarısında bulunan eleştiriler, bunun bilimsel genellemelere ulaşmanın önünde büyük bir engel olduğunu

[3] Abant İzzet Baysal Üniversitesi, Eğitim Fakültesi Oda No:326, Gölköy /Bolu, Tel: 374 254 10 00 (1719), yildirim.huseyin@ibu.edu.tr
[4] Abant İzzet Baysal Üniversitesi, Eğitim Fakültesi Oda No:326, Gölköy /Bolu, Tel: 374 254 10 00 (1719), cet_s@ibu.edu.tr

dile getirir (Bakan, 1966; Cohen, 1994; Schmidt, 1996; Shrout, 1997). O kadar ki, Hunter (1997), YHAT'ın, özellikle psikoloji alanındaki çalışmaların başına gelmiş büyük bir felaket olduğunu iddia eder. Hunter'e (1997) göre, sadece YHAT kullanıldığında, önemli bulgular gözden kaçabilmekte, önemsiz bulgular gereksiz yere rapor edilebilmekte ve farklı araştırmalardan elde edilen bilgileri bir araya getirerek daha genel sonuçlara ulaşmak zorlaşmaktadır.

İstatistik analizlerin sadece YHAT kullanılarak yapılmasını önlemeye çalışan bilim insanları, American Educational Research Association (AERA) veya American Psychological Association (APA) gibi kuruluşların çıkardığı yayım kılavuzlarında, bu konuyla ilgili uyarı ve sınırlamalar olması için de yoğun gayret göstermektedirler (Wilkinson & Task Force on Statistical Inference, 1999; Finch, Thomason & Cumming, 2002).

APA'nın 1952'de yayımlanan ilk yayım kılavuzu, 2010'da yayımlanan 6. yayım kılavuzuyla karşılaştırıldığında, bu gayretlerin meyvelerini vermiş olduğu görülmektedir. İlk kılavuzda sadece YHAT önerilirken, son kılavuzda YHAT'ın sadece bir başlangıç olarak ele alınabileceği vurgulanmaktadır. YHAT'ı desteklemek üzere, istatistiklere ait etki büyüklüğü ve güven aralığının da rapor edilmesi ve mümkün olduğunda, etki büyüklüğü için de güven aralığı hesaplanması önerilmektedir (APA, 2010 : 30).

Sonuç olarak, veri analizlerinde istatistik kullanımıyla ilgili bu tartışmaları takip etmenin, iki boyutta önem kazandığı görülmektedir. İlk olarak, analizlerde sıklıkla kullanılan istatistik yöntemlerinin neler olduğu, bunların hangi yönleriyle eleştirildiği ve görülen eksiklerinin giderilmesi için neler önerildiği üzerine yapılan tartışmaların takibi, özellikle sosyal bilimlerde, istatistik yöntemlerinin bilimsel düşünceye bir rehber olarak kullanılabilecek şekilde anlaşılmasını sağlayacaktır. Aksi halde, Cohen'in de (1994) belirttiği gibi, bu metotlarının birer şablon olarak algılanması ve bilimsel düşünceye rehber olacak yerde, onu belirlemesi tehlikesi ortaya çıkacaktır. İkinci olarak ise, bu tartışmalar sayesinde dergilerin yayım kriterleri daha iyi anlaşılabilecek ve istatistik analizlerin bu doğrultuda yapılmasıyla, çalışmaların yayımlanma şansı da artacaktır. Diğer yandan, Türkiye'de, Sosyal ve Beşeri Bilimler alanında ulusal veri tabanı tarandığında, YHAT'a eleştirel olarak yaklaşan bir çalışmanın olmadığı da görülmektedir.

Tüm bunlar göz önüne alındığında, istatistik analizlerinde sıklıkla yapılan hatalara dikkat çeken, bu alandaki tartışmaları sunarak hataların giderilmesi yönündeki önerileri özetleyen bu çalışmanın, istatistik analizlerinin daha doğru bir şekilde yürütülmesine katkıda bulunacağı düşünülmektedir. Çalışmada, söz konusu tartışmalar aşağıdaki başlıklar altında ele alınmıştır.

1) YHAT ile ilgili eleştiriler nelerdir?

2) YHAT ile ilgili eleştirilerde belirtilen eksiklerinin giderilmesine yönelik öneriler nelerdir?

3) Merkezi olmayan olasılık dağılımlarının, YHAT'ın eksiklerinin giderilmesindeki rolü nedir?

TARTIŞMA

Aşağıda, YHAT'la ilgili alan yazınında yer alan eleştiriler ve bu eleştirilerde belirtilen eksiklerin giderilmesine yönelik öneriler verilmiştir. Ancak herhangi bir yanlış anlaşılmaya yol açmamak için, bu eleştirilerin doğrudan doğruya YHAT'a değil, YHAT'ın aslında cevaplayamayacağı bazı soruların cevabı olarak kullanılmasına, diğer bir deyişle sınırlarının çok ötesinde kullanılmasına karşı çıktığı tekrar vurgulanmalıdır.

YHAT'la İlgili Eleştiriler

Bu bölümde, YHAT'a yöneltilen eleştiriler 3 başlık altında sunulmuştur. Tartışmaları somutlaştırmak üzere, bu ve sonraki bölümlerde, *t*-dağılımı üzerinde durulmuştur. Ancak, ele alınan tartışmalar kay-kare ve F-testlerine de genellenebilmektedir (Kirk, 2003; Smithson, 2000). Eleştirilere geçmeden önce, ilerleyen bölümlerde de sıklıkla kullanılacak bazı kavramlar üzerinde durmak, yazının anlaşılırlığını artıracaktır.

Bilindiği gibi bir popülasyondan, yeterince büyük (*n*) bir örneklem, rasgele (seçkisiz olarak) seçildiğinde, örneklemin ortalaması $\overline{x}$, çoğunlukla popülasyonun ortalaması μ_0'a yakın bir değer çıkar. Ancak, aynı popülasyondan birçok farklı örneklem seçildiğinde, örneklem ortalamaları, denemeden denemeye farklılık gösterebilir. İstatistikte bu farklılıkların kaynağı örnekleme salınımı (sampling fluctuation) olarak adlandırılır. Örnekleme salınımı sonucunda, popülasyon ortalamasından görece çok daha farklı örneklem ortalamaları da elde edilebilir, ancak böyle durumlarla nadiren karşılaşılır.

Rasgele seçilen örneklemlerin ortalaması yerine, bu örneklemlerin popülasyon ortalamasından farkları alındığında ise, yukarıdaki paragraftakilere denk olarak kısaca şu söylenebilir: denemelerin çoğunluğunda $(\overline{x} - \mu_0)$ farkı sıfıra yakın çıkar. Bazı denemelerde sıfırdan oldukça farklı değerler de elde edilebilir, ancak böyle durumlarla nadiren karşılaşılır.

Yukarıda sözü edilen denemeleri, yani bir popülasyondan birçok örneklem seçme işini pratikte yapmak çok zordur. Olasılık dağılımı denilen teorik modeller, bu güçlüğün aşılmasında önemli bir rol üstlenir. Örneğin, teorik bir model olan *t*-dağılımı sayesinde, birçok gerçek deneme yapmak zorunda kalmadan, denemelerin yapılmış olması durumunda, $(\overline{x} - \mu_0)$ değerlerinin yaklaşık kaçta kaçının, belirli bir değerden küçük veya büyük veya belirli değerler arasında olacağı hesaplanabilir.

Ancak, *t*-dağılımı, doğrudan $(\overline{x} - \mu_0)$ farkının değil, aşağıda verilen birinci eşitlikte görüleceği gibi, bu farkın, standart hata denilen bir değere bölünmesiyle elde edilen ve *t* istatistiği olarak adlandırılan değerlerin dağılımını vermek üzere geliştirilmiştir. Böylece, çoğunlukla -4 ile 4 arasında çıkan standart bir istatistik elde edilmiştir. Bu sayede, ortalamalar ister yüz üzerinden, ister on üzerinden, isterse başka bir değer üzerinden ölçeklenmiş olsun, *t* değerleri standart değerler çıkacağı için, aynı *t*-dağılımı kullanılabilmektedir. Eşitlikteki *s* değişkeni, rasgele seçilen *n* büyüklüğündeki örneklemin standart sapmasıdır.

$$t = \frac{(\overline{x} - \mu_0)}{\frac{s}{\sqrt{n}}} \qquad (1)$$

Özetle, bir popülasyondan *n* büyüklüğünde farklı örneklemler rasgele seçildiğinde, bu örneklemlerin her biri için ayrı ayrı hesaplanacak *t* istatistikleri, örnekleme salınımından ötürü, belirli bir farklılık gösterecektir. Bu değerler çoğunlukla sıfıra yakın olacaktır. Sıfırdan uzak bir *t* değerinin elde edilmesi olasılığı ise, sıfırdan ne kadar uzak olduğuna bağlı olarak düşecektir. Bilindiği üzere bu olasılık değerleri, *t*-dağılımı sayesinde, teorik olarak kolayca hesaplanabilmektedir. İstatistik kitaplarında bulunabilecek bir *t*-dağılım tablosu incelendiğinde, örneğin, bir popülasyondan rasgele seçilecek, 100'er kişilik binlerce farklı örneklemden elde edilecek *t*-değerlerinin yaklaşık %40'ının -0,5 ile +0,5 arasında olacağı görülecektir. Diyelim ki 2,36 dan büyük bir *t* değeri ise, denemelerin ancak %1'inde görülecektir. Statistical Package for the Social Sciences (SPSS) vb. istatistik programlarının verdiği *p* olasılık değerleri, bu oranların teorik değeridir.

Bu *p* değerlerinin YHAT'taki yerini, bir örnek üzerinde detaylı bir şekilde ele almak yararlı olacaktır. Genel matematik dersinin belirli bir yöntemle işlendiği bir üniversitedeki tüm öğrencilerin (popülasyon), matematik sınav notları ortalaması μ_0 olsun. Bu üniversitede, söz konusu matematik dersinde kullanılmak üzere yeni bir öğretim yöntemi geliştirildiğini, ancak bu yöntemi tüm üniversitede uygulamaya geçmeden önce, yeni yöntemin eskisine göre daha etkili olduğuyla ilgili kanıt arandığını varsayalım. Bu sebeple, matematik dersini ilk kez alacaklardan *n* öğrencinin rasgele seçildiğini; sonrasında bu öğrencilere yeni yöntemin uygulandığını ve nihayetinde öğrencilerin matematik sınav notları ortalamasının da $\overline{x}$ olarak bulunduğunu kabul edelim. Ayrıca, $\overline{x}$'ın, μ_0'dan büyük çıktığını da kabul edelim.

Çıkarımsal istatistik, $\overline{x}$ ile μ_0 arasındaki bu farkın, aranan kanıt olup olamayacağı sorusuyla ilgilenir. Bu soruya hemen evet cevabının verilememesinin sebebi, örnekleme salınımıdır. Çünkü, rasgele seçilen bu *n* öğrenci, yeni yerine, yine eski yöntemle öğrenim görmüş olsaydı, $\overline{x}$ ile μ_0 arasında, örnekleme salınımından kaynaklanan bir fark yine de olacaktı.

Sonuçta, yeni yöntemle öğretim yapıldığında elde edilen $(\bar{x} - \mu_0)$ farkının, örnekleme salınımından kaynaklanan bir fark olma olasılığının çok düşük olduğu gösterilmediği müddetçe, bu farka dayanarak, yeni öğretim yönteminin daha etkili olduğunu iddia etmek mümkün olmayacaktır. İşte YHAT, $(\bar{x} - \mu_0)$ farkının örnekleme salınımından kaynaklanmış olma olasılığının hesabına dayanan bir karar verme yöntemidir. Hesaplanan bu olasılık değeri, *p* değeri olarak bilinir. Bu *p* değeri, α ile gösterilen ve çoğunlukla 0,05 olarak alınan "anlamlılık düzeyinden" küçükse, bu olasılık değeri "çok küçük bir değer" olarak tanımlanır.

Böylece, *t* dağılımı kullanılarak yapılacak hesaplamalar sonucunda 0,05'ten (α = 0,05 alındıysa) küçük bir *p* değeri bulunması, $(\bar{x} - \mu_0)$ farkının örnekleme salınımından kaynaklanmış olma olasılığının "çok düşük" olduğunu gösterir. Bu da, yeni yöntemin öğrenci ortalamasını yükselttiğini gösteren bir kanıt olarak kabul edilir. Ancak aşağıdaki ilk eleştiride bahsedildiği üzere, bu *p* değeri, sıklıkla, yanlış yorumlanmaktadır.

$P(V|H_0)$ ve $P(H_0|V)$ farklı olasılıklardır.

Yukarıda açıklanan *p,* bir koşullu olasılıktır. Çünkü bu *p* değeri, yeni öğretim yöntemiyle eskisinin özdeş olduğu veya buna denk olarak, yeni yöntem kullanıldığında da eski yöntem kullanıldığındaki popülasyon ortalamasının elde edileceği, diğer bir deyişle, yöntemler arasında fark olmadığı varsayımı (koşulu) altında hesaplanır. Bilindiği üzere bu varsayım, yokluk veya sıfır hipotezi olarak adlandırılır ve H_0 ile gösterilir. Sonuçta, yeni yöntem uygulandığında elde edilen veri *V* ile gösterilirse, *p* değeri, $P(V|H_0)$ koşullu olasılık değeridir. Diğer bir deyişle, H_0'ın geçerli olduğu kabul edildiğinde, *V*'nin elde edilme olasılığıdır.

Bu olasılık, $P(H_0|V)$'den yani, *V* verisi elde edildiğine göre, H_0'ın doğru olma olasılığından farklıdır. Oysa *p*'yi bu ihtimalmiş gibi yanlış yorumlayan azımsanamayacak sayıda makele ve hatta kitap bulunmaktadır (Gliner, Leech & Morgan, 2002; Kirk, 2003; Alhija & Levy, 2009). Bu hata, *p*'nin yokluk hipotezinin doğru olma ihtimali olarak algılanması hatasıdır. Oysa tek bir bilimsel çalışmada elde edilen veriye dayanarak, yokluk hipotezinin doğru (veya yanlış) olma ihtimali hesaplanamaz (Hedges, 1987).

$P(V|H_0)$ ve $P(H_0|V)$'nin farklılığı üzerine, ders kitaplarında da sıklıkla yer alanlara benzer şöyle bir örnek verilebilir. Diyelim ki, 5'i gözlüklü 9 erkek ve 2'si gözlüklü 8 kız öğrenciden oluşan bir sınıftan, rasgele seçilen bir öğrencinin, kız olduğu (*K*) bilindiğine göre, gözlüklü olma (*G*) ihtimali, $P(G|K)$, 2/8'dir. Oysa $P(K|G)$, yani gözlüklü olduğu bilindiğine göre kız olma ihtimali 2/7'dir.

Bu iki olasılığın karıştırılmasında, Pollard ve Richardson (1987) ile Cohen'in (1994) dikkat çektiği mantıksal bir hata söz konusudur. Bu araştırmacılar, Aristoteles'in klasik mantığında geçen 'modus tolens'in YHAT'ta da geçerli olduğunun sanıldığı uyarısında bulunmaktadırlar. Bilindiği gibi, Modus tolens, 'P varsa Q vardır' ve 'Q yoktur' öncüllerinden hareket ederek, bu durumda 'P yoktur' sonucuna varılabileceğini söyleyen bir usa vurma tekniğidir.

Pollard ve Richardson (1987) ile Cohen'in (1994), bu mantığın geçerli olması için istatistiksel anlamlılık testlerinin mutlak ifadeler üzerine kurulmuş olması gerektiğinin altını çizer. Öyle olsaydı; 'H_0 yokluk hipotezi doğruysa *V* gibi bir veri elde edilemez'; '*V* verisi elde edilmiştir' öncüllerinden, demek ki, 'H_0 hipotezi doğru olamaz' sonucuna varılmasında bir mahsur olmayacaktı. Ancak YHAT, olasılığa dayalı bir tekniktir. Bu durumda, 'H_0 yokluk hipotezi doğruysa *V* gibi bir verinin elde edilmesi olasılığı çok düşüktür'; '*V* verisi elde edilmiştir' öncüllerinden, demek ki, 'H_0 hipotezinin doğru olma olasılığı çok düşüktür' çıkarımında bulunulamaz. Yukarıda belirtilen yanlışlık, bu çıkarımın yapılmasından kaynaklanır.

Olasılığa dayalı ifadelerde 'modus tolens'in geçerli olamayacağı, Cohen'in (1994) verdiği bir örnek uyarlanarak gösterilebilir: 'Bir Türkiye Cumhuriyeti vatandaşının, Cumhurbaşkanı olma ihtimali çok düşüktür'; 'A kişisi Cumhurbaşkanıdır'. Demek ki, 'A kişisinin Türkiye Cumhuriyeti vatandaşı olma ihtimali çok düşüktür'. İşte, YHAT'ta *p* değerini doğrudan yokluk hipotezinin doğru olma ihtimali olarak yorumlayanlar, bu örnekte apaçık görünen hatayı yapmaktadırlar.

Özetle, YHAT çerçevesinde, *p* değerinin α değerinden küçük olması, H_0 yokluk hipotezinin geçerli olmayabileceğini gösteren veya daha teknik bir şekilde söylersek, yokluk hipotezinin reddedilmesine yeterli, dolaylı bir kanıt olarak görülür. Ancak, *p* değeri, yokluk hipotezinin doğruluğunu veya yanlışlığını gösteren bir olasılık değeri değildir. Kaldı ki, bir sonraki eleştiride tartışıldığı üzere, kavramsal olarak ele alındığında yokluk hipotezinin doğru olması zaten mümkün değildir.

Yokluk hipotezi her zaman yanlıştır.

Yukarıdaki örneği ele alarak devam edersek, yokluk hipotezi, yeni veya eski öğretim yönteminin kullanılmasının, popülasyonun matematik sınavı ortalamasını değiştirmeyeceğini iddia eder. Oysa, madem uygulanan yöntemler farklıdır, farklı yöntemler uygulandığında elde edilecek ortalamalar arasında da bir farklılık olacaktır. Bu fark belki yüzde birler, belki binde birler basamağında, çok küçük bir fark da olabilir. Ama bir fark mutlaka olacaktır. Dolayısıyla, yokluk hipotezi, en azından mutlak bir anlamda ele alındığında, her zaman yanlıştır.

Teknik açıdan bakıldığında da benzer bir durum söz konusudur. Yukarıda da tartışıldığı üzere, YHAT'ta yokluk hipotezinin yanlış olma ihtimali hesaplanamaz. Söz konusu olan, *p*'nin, çalışmadan önce belirlenen bir sabit olan, α 'dan küçük olmasını, yokluk hipotezinin yanlış olabileceğine dolaylı bir kanıt saymaktır. *p* ise, yine yukarıda örneklendiği gibi, *t* değeri arttıkça küçülür.

Eşitlik 1 incelendiğinde, *t* değerinin sadece $(\overline{x} - \mu_0)$ farkına değil, örneklem standart sapması (*s*) ve örneklem büyüklüğüne de (*n*) bağlı olduğu görülecektir. Örneklem büyüklüğü arttıkça, *t* değeri de büyüyecek ve dolayısıyla böyle bir *t* değerinin görülme ihtimali *p* de küçülecektir. Sonuçta, $(\overline{x} - \mu_0)$ farkı sıfır olmadığı müddetçe, yeterince büyük bir örneklemle, α 'dan küçük bir *p* değerine mutlaka ulaşılacaktır. Sonuç olarak, YHAT'ta, yokluk hipotezi, yeterince büyük bir örneklemle, her zaman yanlışlanabilecektir. Eleştirmenler, haklı olarak, bu nokta gözden kaçtığında, araştırmacıların, yeterince büyük bir örneklemle zaten yanlışlanacak bir hipotezin yanlışlanabilirliğini araştırmak gibi boş bir uğraşa girmiş olacağını dile getirmektedir (Murphy, 1990; Kirk, 1996).

Diğer yandan, örneklemin yeterince büyük olmaması ise, $(\overline{x} - \mu_0)$ gerçekte önemli bir farka işaret ediyor olsa bile, α 'dan küçük bir *p* değeri elde edilmesini engelleyecektir. Bu durum, istatistik testinin gücünün düşük olması olarak bilinir. Görüldüğü üzere, örneklem büyüklüğü, göz önünde bulundurulması gereken önemli bir faktördür. Yokluk hipotezi, ortada önemli bir fark yokken sırf örneklem çok büyük diye reddedilmiş olabileceği gibi, fark önemli olduğu halde, örneklem yeterince büyük olmadığı için reddedilememiş de olabilir.

Örneklem büyüklüğünün bu etkisine dikkat çekmek için, "istatistiksel anlamlılık" ifadesi geliştirilmiştir. Örneğin, yeni öğretim tekniğinin eskisinden daha etkili olup olmadığının araştırıldığı yukarıdaki örnekte, H_0 reddedilebilirse bu, yeni teknik ile eskisi arasında "istatistiksel olarak anlamlı" bir fark olduğu şeklinde yorumlanır. Bu ifadenin, araştırmacının eleştirilerde verilen tehditlerden haberdar olduğunu gösterdiği kabul edilir. Ancak yayımlanmış çalışmalar incelendiğinde, istatistiksel anlamlılık ifadesinin, ya kullanılmadığı ya da, sıklıkla basmakalıp bir ifade olarak kullanılabildiği görülmektedir (Kirk, 1996).

İstatistiksel anlamlılık aşırı genellenmektedir.

Her şeyden evvel, *p* değerinin, ancak ve ancak, örneklemlerin rasgele seçildiği varsayımı altında tanımlanmış olduğu gözden kaçmamalıdır. Dolayısıyla, bilimsel bir çalışmada, üzerinde çalışılan örneklem, rasgele seçilmemişse, rasgeleliğe dayalı *p* değerinin kullanılması da sorgulanabilir hale gelir. Bilindiği gibi, sosyal bilimlerde, örneklemlerin rasgele seçilmesi çoğunlukla mümkün olmaz. Dolayısıyla, bu durumdan kaynaklanabilecek hataları telafi etmek için, istatistiksel anlamlılığı destekleyecek ek kanıtlar da aranmalıdır. Elde edilen sonuçların, farklı mekan ve zamanda yinelenip yinelenmediğinin araştırılması (replication) böyle kanıtların başında gelir (Shaver, 1993). Ne var ki, istatistiksel

anlamlılığın sınırlılıkları anlaşılmadığında, sadece YHAT sonuçlarını rapor etmenin yeterli olduğu düşünülmektedir. Bu sebeple ek kanıtlar elde edilmesine yeterince önem verilmemektedir.

Diğer yandan, örneklemin rasgele seçilmemesinden kaynaklanan bir sorun olmadığı varsayılsa da YHAT'ın sınırlılıkları bitmez. Bunlardan biri, *p*'nin ortalamaya dayalı bir istatistik olmasıdır. Ne yazık ki bu da sıklıkla gözden kaçmaktadır. (Bakan, 1966; Cohen, 1994). Örneğin, yeni öğretim yönteminin eskisiyle arasında istatistiksel olarak anlamlı bir fark olup olmadığını araştırmak, ortalamalar üzerinden yapılan hesaplamalara dayanır; öğretim yöntemlerinin bireysel farklılıklarla ilişkisi hesaba katılmaz.

Ayrıca, *p*'ye dayanarak yapılabilecek çıkarımların içinde, yeni öğretim yönteminin konulara göre etkisi de yoktur. Örneğin, yeni öğretim yöntemi, sadece belirli bir matematik konusunda çok etkili olduğu için ortalamayı yükseltmiş de olabilir.

Sıklıkla düşülen yanılgılardan bir diğeri ise, *p*'nin etkinin büyüklüğü olarak yorumlanmasıdır (Kirk, 2003). Yukarıdaki örneğe dönecek olursak, *p* ne kadar küçükse, yeni öğretim yöntemi de eskisine kıyasla o kadar etkilidir şeklinde bir yorum bu tür bir hatadır. Çünkü bu durumda, çok küçük bir *p* değerinin, örneklemin çok büyük olmasından da kaynaklanmış olabileceği ihtimali gözden kaçmış olacaktır.

Tüm bu eleştiriler göz önüne alındığında, istatistiksel anlamlılık testinin ancak sınırlı bir bilgi verdiğinin gözden kaçtığı görülmektedir. Bu da, YHAT üzerinde gereğinden fazla durulmasına, araştırmacının farklı bilgi kaynaklarına yönelmek için kullanması gereken enerjisinin heba olmasına ve sonuçta gerçeklikle ilgili önemli olabilecek ipuçlarının elde edilmemesine sebep olmaktadır (Schmidt, 1996). Örneğin, Kirk (2003) verdiği bir örnekte, aynı konu üzerinde yapılan yineleme (replication) çalışmalarında bulunan farklar, istatistiksel olarak anlamlı olmamasına rağmen, bu çalışmalardan elde edilen sonuçlar bir arada ele alındığında, önemli bulgulara ulaşılabileceğini göstermektedir. Oysa YHAT tek kriter olarak ele alındığında, bu çalışmalar dikkate alınmayacak ve çok önemli bilgilere ulaşılamayacaktır.

Öneriler

Bu bölümde, YHAT'ın yukarıda belirtilen sınırlılıklarını mümkün olduğunca tamamlamaya yönelik 2 temel öneri üzerinde durulmaktadır: Popülasyon parametreleri için güven aralığı hesaplanması ve etki büyüklüğünün göz önünde bulundurulması. Bunlara ek olarak, APA (2010) yayım kılavuzunda da geçen, etki büyüklüğü için de güven aralığı hesaplanması ele alınmıştır. Ancak bu hesaplama için merkezi olmayan olasılık dağılımı gerektiğinden, iki konu ortak bir başlık altında incelenmiştir.

Güven aralığı.

Bir popülasyondan, örneğin, 100'er kişilik bir çok örneklem seçilerek *t* değerleri hesaplanırsa, bu *t* değerlerinin yaklaşık %95'inin -1,96 ile 1,96 arasında olacağını, *t*-dağılımı sayesinde biliriz. Bu olguyu olasılık olarak şöyle de ifade edebiliriz: Bir popülasyondan rasgele seçilecek 100 kişilik bir örneklemden elde edilecek *t* değerinin, -1,96 ile 1,96 arasında olma ihtimali 0,95'tir. Matematiksel olarak ise bu ifade, $P(-1{,}96 < t < 1{,}96) = 0{,}95$ şeklinde yazılır.

Bu matematiksel ifadedeki *t* değeri yerine, Eşitlik 1'deki karşılığı yazılarak, basit bir kaç cebirsel işlemle popülasyon ortalaması yalnız bırakıldığında, aşağıdaki eşitlik elde edilir.

$$P(\bar{x} - 1{,}96.\frac{s}{\sqrt{n}} < \mu_0 < \bar{x} + 1{,}96.\frac{s}{\sqrt{n}}) = 0{,}95 \qquad (2)$$

Rasgele bir örneklem seçildiğinde, bu örneklemin ortalaması, standart sapması ve büyüklüğü bilineceği için, Eşitlik 2'de bilinmeyen sadece popülasyon parametresi kalacaktır.

Öğretim yöntemiyle ilgili örneğe dönerek, rasgele seçilen 100 öğrenciye yeni öğretim yöntemi uygulandığında, bu öğrencilerin matematik notları ortalamasının 70 ve standart sapmasının da 20 bulunduğunu varsayalım. Bu durumda, Eşitlik 2'den, yaklaşık olarak, $P(66 < \mu_0 < 74) = 0{,}95$ çıkacaktır. İşte bu (66 ; 74) reel sayı aralığına %95'lik güven aralığı denir.

Başka güven aralıkları da benzer şekilde hesaplanır. Örneğin %90'lık güven aralığı hesaplamak için Eşitlik 2'nin solunda geçen 1,96 değerinin , yaklaşık 1,65 ve eşitliğin sağındaki değerin de 0,90 olarak değiştirilmesi yeterli olacaktır.

Görüldüğü üzere, güven aralığı, YHAT gibi popülasyondan örnekleme değil, örneklemden popülasyona doğru hareket eden bir çıkarım yöntemidir. Örneğin, %95'lik güven aralığı, öyle reel sayılar kümesidir ki, ortalaması bu aralıkta kalan bir popülasyondan, eldeki örneklemin seçilme ihtimali en az %5'tir (%100 – %95 = %5 olduğu için). Bunun sonucunda, ortalaması bu aralıktaki değerlerin dışında olan bir popülasyondan, eldeki örneklemin seçilme ihtimali de %5'ten az olacağı için, %95'lik güven aralığı aynı zamanda, α =0,05 anlamlılık düzeyinde reddedilemeyecek yokluk hipotezi değerleri kümesi olarak da tanımlanabilir. Böylelikle, YHAT ile yapılabilecek çıkarımların, güven aralığıyla da yapılabileceği görülür.

Diğer yandan, güven aralığı YHAT'a kıyasla şu avantajları da sunmaktadır. 1) Çalışma öncesinde bir yokluk hipotezi kurulmasını gerektirmez. 2) Güven aralığı, muhtemel değerlerden oluştuğu için, araştırmalarda elde edilen verinin mutlak olarak algılanmasını önler. Bir araştırmada elde edilen verinin, elde edilebileceklerden biri olduğunun anlaşılmasını kolaylaştırır. 3) Güven aralığının genişliği, çalışmanın tutarlığı, diğer bir deyişle, çalışmanın yinelenmesi halinde benzer sonuçlar elde edilme ihtimali hakkında fikir verir. 4) Meta-analiz çalışmalarında, farklı çalışmalardan elde edilen bilgilerin birleştirilmesini kolaylaştırır (Kirk, 2003; Cumming & Finch, 2001).

Etki büyüklüğü.

Etki büyüklüğü, biraz basitleştirerek tanımlarsak, yeni denenen bir yöntemin, eskisine kıyasla ne kadar fark yarattığıdır. Bunun dışında, korelasyona dayalı etki büyüklüğü tanımları da vardır, ancak bu çalışmanın bütünlüğünü bozmamak açısından ele alınmamıştır (Ek bilgi için bakınız: Cohen, 1988).

Bu tanımdan yola çıkıldığında, yukarıda, yeni öğretim yöntemi örneğinde geçen, $(\bar{x} - \mu_0)$ farkının aslında etki büyüklüğünün bir göstergesi olduğu açıktır. Bu fark, yine yukarıda söz edildiği gibi, kullanılan ölçeklerden bağımsız, standart bir değer elde etmek üzere standartlaştırılmalıdır. Bu standartlaştırma, $(\bar{x} - \mu_0)$ farkını, örneklemin standart sapması *s*'ye bölerek yapılmaktadır. Cohen (1988) tarafından geliştirilen bu $(\bar{x} - \mu_0)/s$ oranı, *d* gösterilir. Formülden görüldüğü gibi, etki büyüklüğü, yeni yöntemin, örneklemin standart sapmasının kaç katı bir fark yarattığıdır. Cohen (1988) genel bir öneri olmak üzere, *d* değerinin 0,2'den küçük olması durumunda, yeni tekniğin etkisinin zayıf, 0,8'den büyük olması durumunda ise kuvvetli olarak tanımlanabileceğini söylemektedir. Ancak, 0,2'lik bir *d* değerinin bile kuvvetli bir etki olarak ele alınabileceği özel durumların olabileceği de unutulmamalıdır (Rosnow & Rosenthal, 1989).

Eşitlik 1'de verilen *t* istatistiğiyle karşılaştırıldığında, *d*'nin tek farkı, örneklem büyüklüğünü (*n*) içermemesidir. Dolayısıyla, *d* istatistiği, kullanılan yöntemin etkisini, örneklem büyüklüğünden etkilenmeden gösteren bir indistir. Bu özelliğinden ötürü, çalışmalarda *d* istatistiğinin de rapor edilmesi, eleştiriler bölümünde verilen istatistiksel anlamlılığın, örneklem büyüklüğünden de kaynaklanmış olabileceği şüphesinin giderilmesine olanak sağlar. Ayrıca, istatistik analizlerde kullanılan program *d* istatistiğini vermiyorsa bile, *d*'nin, *t* değerini $\sqrt{n}$'e bölerek kolayca bulunabileceği dikkatlerden kaçmamalıdır.

d etki büyüklüğüyle ilgili gözden kaçmaması gereken bir diğer önemli nokta ise, bunun da bir istatistik olduğudur. Yani, bir başka örneklem seçilip çalışma yinelendiğinde farklı bir *d* değeri elde edilecektir. Dolayısıyla, *d* değeri için de güven aralığı hesaplanmalıdır. Böylelikle, güven aralığı başlığı altında verilen avantajlar etki büyüklüğü için de sağlanmış olacaktır. Ne var ki, *d* istatistiği için güven aralığı hesaplanması, ancak merkezi olmayan *t*-dağılımı sayesinde mümkündür.

Merkezi olmayan t-dağılımları ve etki büyüklüğü için güven aralığı.

Bilindiği üzere, örneklemlerden elde edilen "istatistikler", popülâsyondaki karşılıklarının, yani "parametrelerin" tahmin edicisidir. Örneğin, rasgele seçilen bir örnekleme, yeni yöntemle öğretim

yapıldıktan sonra elde edilen, örneklemdeki öğrencilerin matematik sınav notları ortalaması $\bar{x}$, bu yeni yöntem tüm popülasyona uygulandığında elde edilmesi muhtemel, popülasyondaki bütün öğrencilerin matematik notları ortalaması μ'nün bir tahmin edicisidir. Benzer şekilde, örneklem standart sapması *s*, popülasyon standart sapması σ'nın bir tahmin edicisidir. Böylelikle, Eşitlik 1'de verilen *t* istatistiğinin de, aşağıda verilen Δ'nın bir tahmin edicisi olduğu kolaylıkla görülür.

$$\Delta = \frac{(\mu - \mu_0)}{\frac{\sigma}{\sqrt{n}}} \tag{3}$$

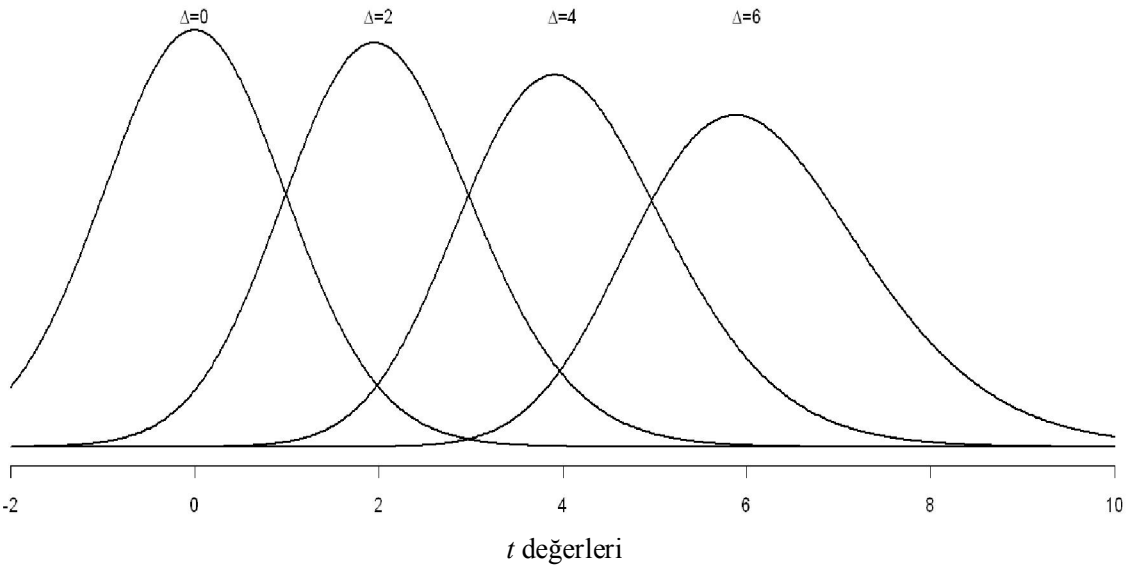
Bu Δ parametresi, *t*-dağılımı için merkezdışılık (noncentrality) parametresi olarak adlandırılır (Cumming & Finch, 2001). İstatistikte, etki büyüklüğü için güven aralığı hesaplamak için, önce Δ merkezdışılık parametresi için güven aralığı hesaplanır. Hesaplanan bu güven aralığının alt ve üst sınırları, *n* örneklem büyüklüğünü göstermek üzere, $\sqrt{n}$'e bölündüğünde de, etki büyüklüğü için güven aralığı bulunmuş olur. Dolayısıyla, öncelikle, *t* istatistiğine dayanarak Δparametresi için güven aralığı hesaplama üzerinde durulmalıdır.

Bilindiği üzere YHAT'ta yokluk hipotezleri, yeni ve eski yöntemlerin popülasyon ortalamasını değiştirmediği, yani $(\mu - \mu_0) = 0$ olduğu şeklinde kurulur. Bu varsayım, Eşitlik 3'ün paydasındaki ifadenin sıfır olduğunu ve dolayısıyla $\Delta = 0$ olduğunu varsaymakla aynı şeydir. İşte, istatistik kitaplarında verilen *t*-dağılımı da, Δ'nın sıfır olduğu bir popülasyondan rasgele seçilen sonsuz sayıda örneklemden elde edilecek *t* değerlerinin dağılımıdır. Diğer bir deyişle, bu dağılım, *t*-dağılımları ailesi içinde, $\Delta = 0$ şartıyla elde edilmiş özel bir dağılımdır. Ancak çoğunlukla bu şarttan söz edilmediği için, bu *t*-dağılımının mümkün olan tek *t*-dağılımı olduğu düşünülür.

Kitaplarda sıklıkla geçen bu *t*-dağılımı, $\Delta = 0$ koşulu altında elde edildiği içindir ki, ancak ve ancak 'yeni ve eski yöntem uygulandığında elde edilecek popülasyon ortalamalarının farklı olmayacağı' gibi hipotezleri test etmek üzere kullanılabilir. Diğer yandan, örneğin, 'yeni öğretim yöntemi uygulandığında popülasyon ortalaması eskisine göre 15 puan yükselecektir' gibi bir iddiayı YHAT ile test etmek mümkün olmaz. Çünkü bu, sembolle yazarsak $(\mu - \mu_0) = 15$ iddiasıdır ve bu durumda $\Delta \neq 0$ olacağı için, $\Delta = 0$ varsayımı altında üretilmiş *t*-dağılımı kullanılamayacaktır. Bu iddiayı test edebilmek için, $(\mu - \mu_0) = 15$ koşulu altında üretilmiş *t*-dağılımı gereklidir.

Şekil 1'de, Δ değerleri farklı dört popülasyonun her biri için, bu popülasyondan seçilen çok sayıdaki 30'ar kişilik örneklemden elde edilen *t* istatistiklerinin dağılımı verilmiştir. Görüldüğü üzere, Δ değerlerinin değişmesi, *t*-dağılımlarının sadece konumunu değil, şeklini de değiştirmektedir. Bu durumda, Δ parametresi için güven aralığı, aşağıda tartışıldığı üzere, cebirsel değil, ancak numerik olarak hesaplanabilir.

Örneklem ortalaması $\bar{x}$'e dayanarak, popülasyon ortalamasının büyük bir ihtimalle içinde kalacağı güven aralığının, alt ve üst sınırlarının nasıl bulunacağı, Eşitlik 2'de verilmişti. Ancak, Δ parametresi için güven aralığını vermek üzere benzer bir eşitlik kurulamaz. Çünkü Δ parametresinin belirli bir aralıkta değişmesi, aynı zamanda bu parametreye bağlı olarak, kullanılacak *t*-dağılımının da değişmesi demektedir. Oysa, popülasyon ortalamasının belirli bir aralıkta değişmesi, daima $\Delta = 0$ koşulu altında olduğu için, sabit bir *t*-dağılımı sayesinde, alt ve üst sınırlar tespit edilirken kullanılacak *t*-değerleri hesaplanabilmektedir. Δ parametresi için güven aralığını hesaplarken böyle bir olanak bulunmadığı için, cebirsel bir çözüm mümkün olmaz.



Şekil 1. Merkezdışılık, Δ, parametreleri farklı olan popülasyonlardan elde edilen *t*-dağılımları. (serbestlik derecesi, *df* = 29)

Bilgisayar teknolojisindeki gelişim bu sorunun, iterasyon yardımıyla aşılmasını sağlamıştır. Bu yöntemde, temel olarak, önce Δ parametresi için güven aralığının alt ve üst sınırlarına belirli değerler atanır. Daha sonra, alt sınır değeri (Δ_{alt}) belirli miktarda artırılarak ve üst sınır değeri ($\Delta_{üst}$) ise belirli miktarda azaltılarak, elde edilen yeni sınırların istenen koşulu sağlayıp sağlamadığı kontrol edilir. Diyelim ki, %95'lik güven aralığı belirlenmeye çalışılıyorsa, bu koşul, $P(\Delta_{alt} < \Delta < \Delta_{üst}) = 0{,}95$ olacaktır. Belirli bir hata payı dahilinde, bu koşula ulaşılana kadar alt ve üst sınır değerleri güncellenerek iterasyon devam eder (Cumming & Finch, 2001).

Δ parametresi için güven aralığı bu şekilde bulunduktan sonra, etki büyüklüğü için güven aralığı, Δ parametresi güven aralığının alt ve üst sınırları, etki büyüklüğü ile *t*-istatistiği arasında, yukarıda açıklanan ilişkiden ötürü, $\sqrt{n}$'e bölünerek bulunur.

Etki büyüklüğü ve diğer önemli bazı istatistikler için güven aralığı hesaplamak amacıyla kullanılabilecek SPSS komut dosyaları hazırlanmıştır (Smithson, 2000). Bu sayede yukarıda verilen hesaplamaların yapılması çok kolaylaşmıştır. Örneğin, Smithson'ın (2000) hazırlamış olduğu *NoncT.sav* dosyası açılıp, araştırmada elde edilen *t*-değeri, serbestlik derecesi ve istenilen güven aralığı düzeyi girilerek çalıştırıldığında, *d* etki büyüklüğü için güven aralığı saniyeler içinde elde edilmektedir.

SONUÇ

Bilimsel çalışmalar, gerçeğin ufak da olsa bir parçasını anlamak üzere yürütülür. Bu iş herhangi bir şablon prosedürle yürütülemeyecek kadar karmaşık ve çok boyutlu bir iştir. Dolayısıyla, ne YHAT, ne de güven aralığı hesaplama, etki büyüklüğü hesaplama gibi yöntemler bilimsel gerçeği anlama amacına tek başına hizmet edebilir. Bu sebeple, bilimsel araştırmalarda, mümkün olduğunca farklı bakış açısı sağlamak üzere, bu yöntemlerin birlikte kullanılmasına gayret edilmelidir.

Ayrıca, bilim insanı yetiştirme programlarında, gözlem verilerine dayanarak çıkarımlar yapmayı mümkün kılan istatistik tekniklerin, kavramsal düzeyde de ele alınması, tercihten ziyade bir zorunluluk olarak görülmelidir. Özellikle yüksek lisans ve doktora düzeyindeki derslerde, istatistik teknikler sayesinde hangi çıkarımlarda bulunulabileceği ve belki bundan da önemli olarak, hangi çıkarımlarda bulunulamayacağı tartışmaları mutlaka yapılmalıdır. Aynı zamanda, bu alanlarda yazılan kitapların da, istatistik tekniklerini eleştirilen ve aksayan yanlarıyla birlikte ele alması, nitelikli bilim insanlarının

yetişmesine katkıda bulunacaktır. Aksi halde, özellikle sosyal bilimlerde, bilimsel uğraş bir teknisyenliğe indirgenmiş olacaktır.

Kuhn (2006) bilimi, bilim insanlarının yaptığı iş olarak tanımlar. Bu tanımdan yola çıkıldığında, bilim insanlarının niteliğinin aynı zamanda bilimin niteliğini de belirlediği söylenebilir. Eğitim ve psikoloji alanında yayımlanmış makaleleri ve yazılmış kitapları inceleyen Gliner ve arkadaşları (2002) ile Alhija ve Levy (2009), istatistik tekniklerin kullanımı açısından, güven aralığı hesaplamama, etki büyüklüğü rapor etmeme gibi ciddi eksiklerin olduğuna dikkat çekmektedir. Bu çalışmanın yazarları da, yüksek lisans düzeyinde yürüttükleri seminer derslerinde, Türkiye'de yayımlanmış makalelerde de benzer eksiklere sıkça rastlanabileceğini gözlemlemişlerdir. Sonuç olarak, bu eksiklerin giderilmesine yönelik tedbirlerin, bilimsel nitelik açısından, önemli gereklilikler olarak görülmesi gerektiği bir kez daha vurgulanmalıdır.

Cohen (1990) *t*-testinin, 'student' rumuzlu William Sealy Gosset tarafından geliştirilip yayımlandıktan yaklaşık elli yıl sonra ders kitaplarına girdiğini belirterek, gelişmelerin zaman aldığına dikkat çeker. Ne var ki, gelişmenin kendiliğinden olamayacağının, ancak gelişim yönünde bilinçli bir gayretin sonucunda ortaya çıkabileceğinin altını çizmek gerekir.

KAYNAKÇA

Alhija, F. N. & Levy, A. (2009). Effect size reporting practices in published articles. *Educational and Psychological Measurement*, *69(2),* 245 – 265.

American Psychological Association. (1952). Publication manual of the American Psychological Association. *Psychological Bulletin*, *49*, 389 – 450.

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th Ed.). Washington, DC.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, *66(6),* 423 – 437.

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association*, *33(203),* 526 – 536.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. (2nd Ed.). Hillsdale, NJ: Lawrence Erlbaum.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45(12),* 1304 – 1312.

Cohen, J. (1994). The earth is round *(p* < .05). *American Psychologist*, *49(12),* 997 – 1003.

Cumming, G. & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement*, *61(4),* 532 – 574.

Finch, S., Thomason, N. & Cumming, G. (2002). Past and future American Psychological Association guidelines for statistical practice. *Theory and Psychology, 12(6),* 825 – 853.

Fisher, R. A. (1925). *Statistical methods for research workers*. (12. Baskı, 1954). Edinburgh: Oliver & Boyd.

Gliner, J. A., Leech, N. & Morgan, G. A. (2002). Problems with null hypothesis significance testing: what do the textbooks say? *The Journal of Experimental Education*, *71(1),* 83 – 92.

Hedges, L. (1987). How hard is hard science, how soft is soft science? *American Psychologist*, *42(2),* 443 – 455.

Hunter, J. E. (1997). Needed: a ban on the significance test. *Psychological Science*, *8(1),* 3 – 7.

Jaynes, E. T. (2003). *Probability Theory the Logic of Science*. New York: Cambridge University.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56*, 746 – 759.

Kirk, R. E. (2003). The importance of effect magnitude. In S. F. Davis (Ed.) *Handbook of research methods in experimental psychology* (pp. 83 – 105). Madlen, MA: Blackwell.

Kuhn, T. (2006). *Bilimsel Devrimlerin Yapısı*. (Çev: Nilüfer Kuyaş). İstanbul.

Murphy, K. R. (1990). If the null hypothesis is impossible, why test it? *American Psychologist, 45(3),* 403 – 404.

Neyman, J. & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika*, *20A (1/2),* 175 – 240.

Rosnow, R. L. & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, *44(10),* 1276 – 1284.

Pollard, P. & Richardson, J. T. E. (1987). On the probability of making type 1 errors. *Psychological Bulletin*, *102(1),* 159 – 163.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: implications for training of researchers. *Psychological Methods*, *1(2),* 115 – 129.

Shaver, J. P. (1993). What statistical testing is, and what it is not? *Journal of Experimental Education, 61*, 293 – 316.
Shrout, P. E. (1997). Should significance tests be banned? *Psychological Science*, *8(1),* 1 – 2.
Smithson, M. (2000). *Statistics with Confidence*. London.
Wilkinson, L. & Task Force on Statistical Inference, APA Board of Scientific Affairs. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594 – 604.
Yıldırım, C. (1996). *Bilim Felsefesi* (5. Baskı). İstanbul.