International Journal of Eastern Mediterranean Agricultural Research, 5(1):01.-19

Review Article

Bilimsel Araştırmalarda İstatistiksel Anlamlılığın Raporlanmasında Güncel Yaklaşımlar: Hatalar ve Doğrular

Ömer AKBULUT[1*]

[1,] Giresun Üniversitesi Fen Bilimleri Enstitüsü Biyosüreç Mühendisliği Ana Bilim Dalı. Giresun, Türkiye.

[1] https://orcid.org/0000-0002-8860-3513

Sorumlu yazar: omer.akbulut@giresun.edu.tr

Geliş Tarihi: 05.01.2022, Kabul Tarihi: 22.11.2022

To Cite: Akbulut, O. (2022). Bilimsel Araştırmalarda İstatistiksel Anlamlılığın Raporlanmasında Güncel Yaklaşımlar: Hatalar ve Doğrular. International Journal of Eastern Mediterranean Agricultural Research, 5(1): 01-19

Özet

Bilimsel araştırma sürecinin önemli bir bölümünü istatistiksel düşünce ve veri analizi oluşturur. Veri analizi kapsamında en yaygın çıkarımsal istatistik yaklaşımı "Yokluk Hipotezi Anlamlılık Testi (YHAT)" sürecidir. Bu sürecin en son ürünü p-değeridir. Bu yönüyle p çıkarımsal istatistiğin en çok kullanılan ve en ünlü ölçüsüdür. Araştırıcılar büyük çoğunlukla bulgularını p-değeri üzerinden yorumlamaktadır. Ancak gerek p-değerinin rapor edilmesinde gerekse yorumlamasında önemli hatalar yapılmaktadır. Bu makalede önce tanımlayıcı istatistiklerin doğru rapor edilmesi ilkeleri ele alınmıştır. Sonra p-değerinin sunum hataları örneklerle gösterilerek doğru yazım örnekleri verilmiştir. Bu değerin, üç ondalıkla ve olasılığın tam değeri olarak verilmesi önerilmiştir. Bilgisayar çıktılarından alınan p=0.000 değerinin, p<0.001 şeklinde verilmesine özen gösterilmelidir. Araştırıcıların daha fazla hata yaptıkları durum, makalelerini doğrudan p-değeri üzerinden yorumlamaları ve yorumu hatalı yapmalarıdır. Çalışmanın sonunda, p-değeri bakımından yapılan genel hatalar ve literatürde bu konuyu ele alan makalelerin sonuçları derlenerek verilmiştir.

Anahtar Kelimeler: α hatası, istatistiksel anlamlılık, p-değeri, yokluk hipotezi anlamlılık testi

International Journal of Eastern Mediterranean Agricultural Research, 5(1):01.-19
Review Article

Current Approaches in Reporting Statistical Significance in Scientific Research: Errors and Truths

Abstract

Statistical thinking and data analysis constitute an important part of the scientific research process. The most common inferential statistical approach within the scope of data analysis is the "null hypotesis significance test, NHST" process. The final product of this process is the p-value. In this respect, p is the most widely used and famous measure of inferential statistics. Researchers mostly interpret their findings on the p-value. However, important errors are made both in reporting and interpreting the p-value. In this article, firstly, the principles of accurate reporting of descriptive statistics are discussed. Then, the presentation errors of the p-value were shown with examples and correct spelling examples were given. It has been suggested that the P-value be given as three decimals and the exact value of the probability. Care should be taken to give the p=0.000 value taken from the computer printouts as p<0.001. The situation where researchers make more mistakes is that they interpret the articles on the p-value alone and make the interpretation wrong. At the end of the article, the mistakes made in terms of p-value and the results of the articles dealing with this subject in the literature are compiled.

Keywords: α error, statistical significance, p-value null hypotesis significance test

1. Giriş

Birçok bilim dalında olduğu gibi, tarım ve mühendislik alanlarındaki bilimsel araştırmaların önemli bir kısmı, istatistiksel tahmin ve karar teorisine dayalı olarak yürütülmektedir. Araştırmaya esas olan problemin hipoteze dönüştürülmesi adımından, araştırma raporunun yazılması aşamasına kadar tüm aşamalar istatistiksel süreçlere dayalı olarak yürütülür. Hipotezleri kurma, araştırmayı tasarlama, veri toplama, verileri analiz etme, yorumlama ve sonuca ulaşma adımlarının her birinde istatistik yöntemlerin önemli bir işlevi vardır. Veri analiz sürecinin önemli yöntemlerinden biri "Yokluk (sıfır) Hipotezi Anlamlılık Testi (YHAT)" olarak bilinir. Bu kavram İngilizce literatürde "Null Hypotesis Significance Test (NHST)", olarak ifade edilmektedir. Bu yöntem, hipotezlerin kurulması, hata olasılık düzeylerinin (α, β) belirlenmesi, uygun istatistik testin uygulanması, anlamlılık değerine dayalı bulguların yorumlanması adımlarından oluşur. Hipotez testi süreci, adım adım ve sıralı olmak üzere bir şablona bağlı olarak yürütülmektedir (Mark ve ark., 2016). Bu süreç yokluk ritüeli (null ritüel) olarak adlandırmaktadır (Gigerenzer, 2004 ve 2018; Işık, 2014; Mark ve ark., 2016).

YHAT veya kısaca hipotez testi, sıfır hipotezinin (H_0) sorgulanması ve yanlışlanması, araştırma (karşıt) hipotezin (H_1) doğrulanması işlemidir. Örnekleme ait veriler kullanılarak yapılan bir hipotez testi aşağıda verilen dört karardan biri ile sonuçlanır. (Kartal, 2006; Gürkan, 2007; Dahiru, 2008; Meurs, 2016; Cengiz ve Terzi, 2018, Yıldız ve ark., 2020; Şenyay, 2021; Göksöz, 2021) Bu dört sonuçtan ikisinde karar doğru iken diğer iki durum Tip I hata (α hatası) ve Tip II hata (β hatası) ile sonuçlanabilir. Bu durumlar,

1. Durum: H_0 gerçekte doğrudur. Test sonucunda H_0 kabul edilmiştir (Karar doğru 1-α).
2. Durum: H_0 gerçekte doğrudur. Fakat test sonucunda H_0 reddedilmiştir (Tip I hata, α).
3. Durum: H_0 gerçekte yanlıştır. Fakat test sonucunda H_0 kabul edilmiştir (Tip II hata, β).
4. Durum: H_0 gerçekte yanlıştır. Test sonucunda H_0 reddedilmiştir (Karar doğru 1-β).

Bu sonuçlar birçok literatürde (Dahiru, 2008; Işık, 2014; Meurs, 2016; Cengiz ve Terzi, 2018; Yıldız ve ark., 2020; Şenyay, 2021; Göksöz, 2021) Tablo 1'deki gibi özetlenmektedir.

Tablo 1. YHAT Olası Sonuçlarının Matris Gösterimi

		Gerçek Durum	
		H_0 Doğru	**H_1 Doğru**
Test Sonucu	**H_0 Kabul**	1-α Doğru karar	β Hatası II. Tip hata
	H_1 Kabul	α Hatası I. Tip hata	1-β Doğru karar

Bu kararların dayanak noktası önsel (a priori) istatistiksel anlamlılık düzeyi alfa "α" olasılığı ve örneklem verilerin istatistiksel analizi ile üretilen "p" olasılığıdır. Anlamlılık düzeyine, önemlilik düzeyi veya kısaca önemlilik te denir. YHAT sürecinde, $p<\alpha$ ise sonuç istatistiksel olarak "anlamlı" olarak nitelendirilir.

Çıkarımsal istatistiğin uygulandığı hemen hemen tüm bilimsel çalışma alanlarında, YHAT tekniği kullanılmaktadır. Araştırma bulgularının raporlanmasında YHAT sürecinin önemli bir ağırlığı mevcuttur. Bu sürecin son ürünü p-değeri'dir. Araştırıcılar çıkarımsal bulgularını p-değerinin sonucuna göre yorumlamakta ve doğru bir yaklaşım olmamakla birlikte $p<0.05$ olasılığına ulaşmayı hedeflemektedir.

Tarım, biyoloji, sağlık, mühendislik bilimleri alanlarında yayınlanan araştırma makalelerinde YHAT çıkarımsal istatistik tekniği yoğun bir şekilde kullanılmaktadır. Çıkarımsal istatistiğin en yaygın ölçüsü anlamlılık düzeyi olarak adlandırılan p-değeri olmakla

birlikte, bir olasılık olan p-değerinin gerek rapor edilmesinde gerekse yorumlanmasında önemli hatalar yapılmaktadır. Bu nedenle, özellikle p-değerinin tanım, anlam ve yorumlanması hakkında literatürde farklılıklar bulunmaktadır. Bu makalenin amaçları; 1) YHAT sürecinin son ürünü p-değerinin sunumunda ve yorumunda yapılan hataları derlemek, 2) Araştırma sonuçlarının rapor edilmesinde, söz konusu kapsamda yapılabilecek olası hataların giderilmesinde araştırıcılara, hakemlere ve ilgili alan editörlerine güncel bilgiler sunarak, makalelerin en az hata ile yayınlanmasına katkıda bulunmaktır.

Ayrıca bilimsel yazımlarda tanımlayıcı istatistikler bazen hatalı olarak yazılmaktadır. Bazı yayınlarda örneklem büyüklüğü bilgisi eksik verilmektedir. Bu bakımdan, makalenin ilave bir amacı ise tanımlayıcı istatistiklerin doğru sunumunun yapılarak, makalenin ilk bölümünde de tanımlayıcı istatistiklerin doğru sunum şekilleri ele alınmıştır.

2. Tanımlayıcı İstatistiklerin Sunumu

Veri analizinin en çok kullanılan tanımlayıcı istatistik ölçüleri; ortalamalar, oranlar ve katsayılardır. Veriler sayı doğrusu üzerinde çok noktada (sürekli değişkenlerde sonsuz) değer alırken ortalama, oran, katsayı gibi ölçülerin her biri örnekleme bağlı olarak sayı doğrusu üzerinde bir noktada değer alabilir. Bu nedenle bu ölçülere yer veya konum ölçüsü denir. Yer ölçüleri, parametreleri için nokta tahminleridir. Bir diğer grup tanımlayıcı istatistik ölçüleri, standart sapma (standart deviation, S) ve standart hata (standart error, Se) ölçüleridir. Bu ölçülerin kaynağı verilerdeki değişkenlik, istatistiksel anlamıyla "varyasyon"dur. Varyasyonun ölçüsü ise varyanstır. Varyans değişkene (özelliğe) ait verilerin merkezden uzaklıklarını ölçen karakteristik bir ölçüdür. Varyansın pozitif karekökü ise standart sapmadır. Gerek varyans, gerekse standart sapma örneklem büyüklüğünden bağımsızdır. Ayrıca varyans, istatistik analiz ve tahmin teorisinde sağlam bir ölçüdür. Bu nedenlerle hipotez testi ve tahmin sürecinin temel kaynağı ve dayanağı varyasyondur.

Örneklemden hesaplanan ölçüler, parametreleri için tahminler olup, tahminler istatistik olarak ta adlandırılmaktadır. Ortalama oran ve katsayı gibi istatistik ölçüleri standart sapma veya standart hataları ile birlikte verilmelidir. Örneğin ortalama ölçüsü standart sapması ($\bar{X} \pm S$) veya standart hatası ($\bar{X} \pm Se$) ile birlikte sunulur. Ortalamanın yanında "standart sapma mı yoksa standart hata mı verilmelidir?" sorusu araştırıcıların sıkça karşılaştıkları bir durumdur. Standart sapma varyansın (pozitif) karekökü olup, nokta tahmini yapılan özelliğin değişkenliğinin ölçüsüdür. Standart hata ise, nokta tahmininin ne kadar sapmalı tahmin edildiğini gösterir. Standart hata örneklem büyüklüğünün karekökü ile ters orantılı değişir. Yani Se= S/√n olup, örneklem sayısı büyüdükçe küçülür. Eğer araştırıcı ortalama oran gibi nokta

tahminlerini ne kadar sapmalı tahmin ettiğini vurgulamak istiyor ise standart hata (Se) ölçüsünü kullanır. Diğer durumda, yani araştırıcı özelliğin değişkenliğini vurgulamak istiyor ise nokta tahmininin yanına standart sapma ölçüsünü yazmalıdır. Gözlemsel olarak normal dağılımlı verilerde, verilerin %95'i $\bar{X} \pm 2S$ %99.7'si $\bar{X} \pm 3s$ aralığında yer alır. Bu şekilde standart sapma sunulduğunda verilerin değişim aralığı da büyük ölçüde tanımlanmış olur.

Standart sapmanın ve standart hatanın kullanım tercihi, bilim alanlarına göre farklılık gösterir. Örneğin mühendislik alanında daha çok yer ölçüsü ve onun hatası yani ($\bar{X} \pm Se$) kullanılırken sağlık bilimlerinde yer ölçüsü dağılımı ile yorumlanır. Bu nedenle sağlık bilimlerinde daha yaygın istatistikler standart sapması ile birlikte sunulur ($\bar{X} \pm S,\ p\ \pm S_p\ ; b \pm S_b\ gibi$) ve yorumlanır.

İstatistiksel bulgular sunulurken dikkat edilmesi gereken bazı kurallar vardır. Bunlardan bazıları, ortalama, oran gibi istatistikler ile birlikte S veya Se mutlaka verilmelidir. İstatistikler ve standart sapma veya hataları istatistikle aynı duyarlıkta yazılmalıdır. Standart sapma veya standart hata istatistikten daha az hassas yazılamaz. Örneğin ortalama ve standart sapma iki ondalıkla 20.74 ± 3.26, tek ondalıkla 20.7 ± 3.3 şeklinde aynı ondalık düzeyinde yazılır. Daha az tercih edilmekle birlikte standart sapma veya standart hata ortalama için verilen ondalıktan bir basamak daha fazla (Örneğin: 20.7 ± 3.26) yazılabilir.

Yüzdelik (%) değerlerinin ondalık sayısının yazımında örneklem büyüklüğü esas alınır. Oran değeri yazılırken eğer n<100 ise yüzdelik tam sayı, n, 100-1000 arasında ise yüzdelik tek ondalıkla, n>1000 ise yüzdelik iki ondalıkla yazılmalıdır. Örneğin % olarak oran değerleri n=75 ise %45, n=458 ise %35.2, n=1405 ise %68.13 şeklinde yazılmalıdır (Anon, 2021).

Hipotez testi sürecinin önemli bir unsuru ise örneklem büyüklüğüdür. Bilimsel çalışmalarda tanımlayıcı ve çıkarımsal istatistikler sunulurken örneklem büyüklüğü mutlaka belirtilmelidir. Çünkü örneklem büyüklüğü (n), araştırmanın tanımlayıcı istatistiklerinin tahminine ve çıkarımsal istatistiklerin sonucuna çok etkili bir bilgidir. Örneklem büyüdükçe örneklemden elde edilen ölçü (ortalama, oran, katsayı vd.) populasyondaki değerine (parametre) yakınlaşır. Bu durum istatistiğe, parametresinin kararlı bir tahmini olma özelliği kazandırır. Bilimsel makalelerde sıkça karşılaşılan bir kavram hatası ise parametre kelimesinin özellik (değişken) ile eş anlamlı kullanılmasıdır. Parametre ortalama, oran vb. bir ölçünün populasyondaki değeri olup, özellik ile eş anlamlı kullanılmamalıdır. Örneğin "çalışmada sütte pH parametresi incelenmiştir" ifadesi hatalıdır. Ayrıca çıkarımsal istatistiğin p-değeri gibi testin gücü ve güven aralığı ölçüleri de örneklem büyüklüğünden büyük ölçüde etkilenmektedir.

3. Çıkarımsal İstatistik Ölçüler

Yukarıda ifade edildiği gibi veri analizi, diğer ifade ile çıkarımsal istatistiğin en yaygın yöntemlerinden biri YHAT'dir. Bu bağlamda çıkarımsal istatistiğin en yaygın kullanılan ölçüleri, önsel (a priori) olasılık "Alfa (α)" ve YHAT sürecinin son ürünü olan "istatistiksel anlamlılık (p)" olasılık değerleridir. İstatistik alanında bilişim teknolojisinin yaygın kullanımı sonucu bu kavramların yazım ve sunumunda güncellemeler söz konusudur. Ayrıca bu kavramların yorumunda önemli hatalar yapılmaktadır. Çalışmanın bu bölümünde, öncelikle bu kavramların tanımları yapılmış ve devamında da söz konusu kavramlarının rapor edilmesinde yapılan hatalı veya eksik yazılımlar değerlendirilmiştir. Ek olarak, istatistiksel anlamlılık p-değerinin yorumunda yapılan hatalara yer verilmiştir.

3.1 Alfa (α) Nedir?

Araştırma sürecinde hipotezlerin kurulması aşamasında ön görülen Tip I hata düzeyi "α" olasılığıdır. Yanlış pozitif olarak da adlandırılan bu hata olasılığı için "kesme değeri" adı verilen 0.05 (genellikle) değeri kullanılır. Bu değer 1925 yılında R. Fisher tarafından istatistiksel anlamlılığın "eşik değeri (kesme değeri)" olarak önerilmiştir (Lehman, 1993; Lu ve Belitskaya-Levy, 2015). Neyman ve Pearson tarafından 1930'lu yıllarda (1930-1933) "hipotez testi" tekniği istatistik literatürüne kazandırılmıştır (Lehman, 1993). Fisher ve Neyman-Pearson'un bu teorileri daha sonra birleştirilerek YHAT (NHST) olarak (Mark ve ark., 2016) yaklaşık 100 yıllık bir süredir veri analizinde kullanılmaktadır. Bu hata oranı veya olasılığı, testin ön görülen anlamlılık düzeyini (significance level) ifade eder. Diğer bir kavram ise, önsel hata adı verilen "α" değeri, veri analizinden önce, hatta araştırmanın örneklem büyüklüğü belirleme aşamasında belirlenir. Bu hatanın değeri, genellikle 0.05 alınmakla birlikte bazı durumlarda 0.01 nadiren ise 0.10 olarak alınır. Mühendislik ve sağlık alanlarında daha duyarlı ve hassas çalışmalar için gerekli örneklem büyüklüğünü belirlemede bu olasılık 0.01 olması önerilir. Araştırmanın metot bölümünde kısa bir cümle ile "istatistiksel anlamlılık düzeyi $p<0.05$ alındı" şeklinde ifade edilir. Bu ifade ile α için kesme değerinin 0.05 olarak alındığı anlaşılır.

Kesme veya eşik değeri $\alpha=0.05$, Fisher'in önerdiği zamandan günümüze kadar değiştirilmeyen bir sınır değerdir. Ancak Piana (2019) Wallach'a atfen 0.05 eşik değerinin yanlışı doğru hipotezden ayırma özelliğini kaybetmiş olabileceğini ve 0.05'in

düşürülebileceğini, hatta bu eşiğin kullanımının terk edilmesinin yararlı olabileceğini vurgulamaktadır. Ayrıca Consonni ve Bertazzi (2017), yeni bir yaklaşım olarak, makalelerin istatistiksel analiz bölümünde "istatistiksel anlamlılık düzeyi $p<0.05$ alındı" cümlesine artık yer verilmemesini önermektedirler. Bu değer makale tasarımında öngörülen güçte (espected power) örneklem büyüklüğünü belirlemede kullanılan bilgi olarak rapor edilmesi açıklayıcı olabilir. Ancak makalenin metot bölümünde bu bilginin istatistiksel anlamlılık düzeyi olarak verilmesinin bir katkısı yoktur.

3.2 İstatistiksel Anlamlılık p Nedir?

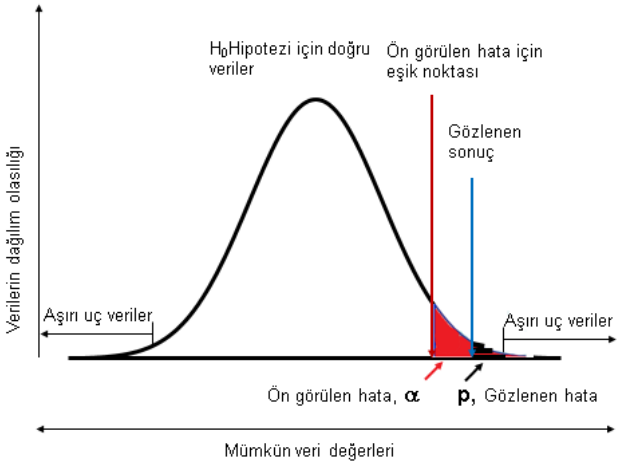
İstatistiksel anlamlılık p-değeri, örneklem verilere YHAT veya diğer istatistiksel testlerin uygulamasının bir ürünüdür. Bilimsel makalelerde istatistiksel anlamlılık değeri kısaca "p" veya "P" ile sembolize edilmektedir (Consonni anad Bertazzi, 2017). Enformel olarak p-değeri belirli bir istatistiksel model altında verilerin istatistiksel bir özetinin (örneğin karşılaştırılan iki ortalama arasındaki fark gibi) gözlemlenen verilere eşit veya daha uçta olma olasılığıdır (Wasserstein ve Lazar 2016). Burada daha çok tercih edilen model sıfır hipotezidir. Bu nedenle çoğu istatistiksel çıkarım, makalenin başında anlatılan YHAT'ne dayanmakla birlikte diğer ilgili modellere de dayanabilir (Consonni anad Bertazzi, 2017).

Olasılık yazımıyla p-değeri $P(V \mid H_0)$ şeklinde ifade edilen şartlı olasılık olarak tanımlanabilir. Burada V araştırma verilerinden elde edilen değer olmak üzere p, H_0'ın geçerli olduğu varsayıldığında V'nin elde edilme olasılığıdır (Cohen, 1994; Yıldırım ve Yıldırım, 2011; Işık, 2014; Vidgen ve Yasseri, 2016; Startz, 2019). Örneğin bir çalışmada $p=0.02$ bulunmuş ise aynı şartlarda, aynı büyüklükte örneklemler kullanıldığında örneklemlerin p kadar yüzdesi H_0 hipotezi değerinde veya daha uçta olabilir (Erkuş, 2017; Erkan, 2018). Gigerenzer (2004) bu olasılığı, "anlamlılığın tam seviyesi" olarak adlandırmıştır.

Araştırıcılar YHAT sürecinde, p-değerinin α değerinden düşük olması yönünde bir beklentiye sahiptir. Çünkü $\alpha=0.05$ alındığında $p<\alpha$ olması durumunda istatistiksel anlamlılık araştırıcıyı Işık (2014)'ın da belirttiği gibi "aynı araştırmanın, benzer özelliklerdeki örneklemlerde tekrarlanması halinde, elde edilen ilişki ya da farka ilişkin istatistik test değerlerinin %5'inden daha azı, gözlenen bulgulardan daha uçlarda (düşük ya da yüksek) bir sonuç verecektir" sonucuna götürür. Bu sonuç elde edildiğinde araştırıcı H_0'ni ret ederek önermesi (iddiası) H_1 için destek sağlamış olur.

Özet olarak ifade etmek gerekirse α, veri analizi öncesinde karar verilen olasılık değeri, p ise veri analizinin ürettiği anlamlılığın tam seviyesi olan olasılık değeridir. YHAT sürecinde

$p<\alpha$ ise sonuç istatistiksel olarak anlamlı olarak nitelendirilir. Eşik değeri α, ve p'nin konumu şekilsel olarak aşağıdaki gibi gösterilmektedir (Şekil 1).



Şekil 1. Normal Dağılımlı Verilerde α ve p Değerlerinin Şekilsel Gösterimi

Bu açık tanımlamaların yanında, bazı yayınlarda ise p-değerinin yaygınlığı ve önemi vurgulanmaktadır. Örneğin Cohen'in (1994) makalesinin başlığı "Dünya $p<0.05$'in etrafında dönüyor, The earth is round $p<0.05$" şeklindedir. Bu makale aynı zamanda ironik bir eleştiri de içermektedir. Nuzzo (2014) p-değerinin en yaygın istatistik ölçü olduğunu bildirmektedir. Bokai ve ark. (2019) ise p-değerini bilimsel yayınların en ünlü terminolojisi olduğunu ifade etmişlerdir. İstatistik tabanlı araştırmaların çoğu hipotezlerin kurulması ile başlayan YHAT sürecinin sıralı birçok işlemlerden sonra p istatistiğinin elde edilmesi ve yorumlanması ile sonuçlandırılır. Bu süreçte p-değeri için Leek ve Peng (2015) "p-değeri istatistiksel analizde buzdağının görülen ucudur" ifadesini kullanmaktadır.

Bu yükseltici ifadelerin yanında YHAT ve p için literatürde son 20 yılda özellikle son 10 yılda çok tartışmalı makaleler yayınlanmıştır (Cohen, 2011; Nuzzo, 2014; Lu ve Belitskaya-Levy, 2015; Aschwanden, 2016; O'Leary, 2021). Bu bağlamda Nuzzo (2014) makalesini "istatistiksel geçerliliğin 'altın standardı' olan p değerleri, birçok bilim insanının sandığı kadar güvenilir değildir" ifadesine odaklamıştır. Cohen (2011) tarafından yayımlanan makalede, p-değeri ve onun tıp literatüründe yanlış kullanımını kaleme almıştır. Lu ve Belitskaya-Levy (2015) tarafından "p hakkındaki tartışma" başlıklı makalelerinde konuyu irdelemişlerdir. Araştırıcılar istatistiksel anlamlılık p-değerinin araştırmalarında hipotezleri test etmek için kabul edilebilir bir çıkarımsal istatistik olduğunu belirterek, bir çalışmanın bilimsel değerini yargılamak için yalnızca tek bir p-değerine güvenmek, p-değerinin kötüye kullanılması olarak yorumlamışlardır. Yazarlar p-değerlerinin sınırlılığını, değişkenliğini doğru anlaşılmasını dikkate alarak, araştırma bulgularını tek bir p-değerine dayalı değil, ilişkili başka ölçülerinde

verilmesi gerekli olduğunu vurgulamışlardır. Murtaugh (2014) p-değerinin içsel tanımından çok, yorumunun hatalı yapıldığını vurgulamıştır. Yazar p-değeri yerine önerilen diğer bazı çıkarımsal ölçülerde de (güven aralığı gibi) benzer yorum hatalarının yapılabileceğine dikkat çekmiştir. Türkçe literatürde konuyu irdeleyen az sayıda olsa da bazı yayınlar mevcuttur. Bu bağlamda Yıldırım ve Yıldırım (2011) birçok yayında, hatta temel başucu kitaplarında bile, p'nin hatalı tanımlandığını ve yorumlandığını bildirmektedir. Başka bir yayında Kılıç (2011) eleştirel makalesini "Neyin peşindeyiz? Kutsal p-değerinin mi yoksa klinik önemliliğin mi?" başlığı altında bir makale kaleme almıştır. Makalede araştırıcıların klinik anlamlılığı veya farksızlığı göz ardı ederek bulgularını daha çok $p<0.05$ üzerine yoğunlaştıkları vurgulanmaktadır. Kanık (2014) ve Akoğlu (2015) p-değerinin yorumlanmasında yapılan hataları eleştirel yaklaşımla düşünce yazısı bağlamında ele almışlardır. Kanık (2014) "P değeri dedikleri" başlıklı makalesinde p-değerinin yorumunda yapılan hatalara ve yazarların, hakemlerin ve editörlerin makalelerde p-değerini anlamlı ($p<0.05$) bulma yönünde zorlayıcı girişimlerde bulunduğuna dikkat çekmiştir. Akoğlu (2015) ise ironi yaparak örneklendirdiği makalesinde yeterli büyüklükte bir örneklem ile çok küçük farklılıkların bile istatistiksel olarak önemli "anlamlı" olacağının vurgulayarak p-değerinin "anlam" kelimesi ile ifade edilmesinin de hatalı olduğunu bildirmiştir. Her iki yazar bilimsel makalelerde p-değerinin değil, onun yerine etki büyüklüğü, güven aralığı gibi diğer çıkarımsal ölçülerin yorumlanması gerektiğini vurgulamışlardır.

Aschwanden, (2016) "İstatistikçiler üzerinde anlaşabilecekleri bir şey buldu: P değerlerini kötüye kullanmayı durdurmanın zamanı geldi" başlıklı makalesinde Amerikan İstatistik Derneği'nin (ASA) istatistiksel anlamlılık p değerleri üzerinde bir fikir birliği oluşturmak için 26 uzmanının katıldığı toplantıda uzmanların p-değeri konusunda yorum ve görüşlerini derlemiştir. Aschwanden (2016), p-değerinin tanımı ve yorumunun uzman görüşlerine göre tartışmalı olduğunu vurgulamıştır. Aschwanden (2016) makalesini Michigan Üniversitesi biyoistatistik uzmanı Roderick Little'nin p-değerine araştırıcıların bakışını ifade eden ironik bir şiir ile sonlandırmıştır. Söz konusu şiirin bir mısraının özeti literatüre "$p<0.05$ publish, else perish; $p<0.05$ ise yayınla değilse yok ol" şeklinde geçmiştir.

O'Leary (2021), hipotez testine eleştirel bir yaklaşımla yayınladığı makalesinde, YHAT sürecini bir p avcılığı olarak niteleyerek 0.05 eşik alınarak, $p=0.051$ anlamsız; $p=0.049$ anlamlı kabul edilmesini eleştirerek bu bağlamda universal bir eşiğin olamayacağını ifade etmiştir. Yazar istatistik anlamlılığı p-değerine indirgenmesinin doğru olmadığını vurgulayarak p-değerinin yanında etki büyüklüğü ve güven sınırları ile desteklenmesinin gerekliliğini

vurgulamıştır. Bütün bu tartışmaların sonucu olsa gerek Temel ve Uygulamalı Sosyal Psikoloji Dergisi (Basic and Applied Social Psychology) editörleri makalelerde hipotez testi ve p-değerinin kullanımını yasaklama yoluna gitmişlerdir (Tramifow ve Marks, 2015; Wasserstein ve Lazar, 2016).

Anlamlılık düzeyi p-değerinin bilimsel araştırmalardaki konumu ve yorumu konusunda eleştirel nitelikte yukarıda atıf yapılan makalelerin dışında daha çok sayıda yayın mevcuttur. Bu eleştirel düşünceleri şimdilik öteleyerek p-değerinin sunumunda yapılan hatalar ele alınacaktır.

3.3 P-değeri Nasıl Rapor Edilmelidir?

Veri analiz sonucunda ulaşılan p-değeri geleneksel yazımda $p<0.05$, $p<0.01$ veya $p<0.001$ şeklinde yazılmakta idi. Günümüzde çoğu yazılımlar istatistiksel anlamlılık p-değerini gerçek değeri ile sunmaktadır. Genel bir kural ve APA standardı olarak (APA, 2010) p-değeri, gerçek değeri ile en az iki veya üç ondalıkla yazılmalıdır. Bu nedenle p-değeri makalede rapor edilirken olasılığın gerçek değeri ile (örneğin $p=0.026$ gibi) yazılmalıdır. Bu değerin $p<0.05$ olarak yazılması da doğru olmakla birlikte $p=0.026$ yazılımı tercih edilmelidir. Çünkü $p<0.05$ yazılımı durumunda bir miktar bilgi kaybı söz konusudur (Greenland ve ark., 2016).

Ancak p-değerinin çok küçük olması durumunda istatistiksel yazılımlar p-değerini $p=0.000$ olarak sunmaktadır. P değeri sıfırdan büyük, birden küçük bir olasılık olduğuna göre, p hiçbir zaman 0.000 değerinde olmayacaktır. Ancak istatistiksel paket programlar sonuçları genellikle üç ondalıkla yuvarlayarak sundukları için anlamlılık sonucu 0.000 olarak sunulmaktadır. Yani 0.00002 gibi çok küçük bir p-değeri çıktıda 0.000 olarak ifade edilmektedir. Bu 0.0002 değerinin $0.2x10^{-4}$ olarak yazılımı da yazım ve anlatımda kolaylık sağlamayacağı için $p=0.00002$ olasılığının doğru yazılımı $p<0.001$ şeklinde olmalıdır. Bununla birlikte son yıllarda etki faktörü yüksek popüler bazı dergilerde yayınlanan makalelerde p-değerinin çok küçük olduğu durumda p-değeri üslü olarak örneğin $p=2.8x10^{-5}$ şeklinde gerçek değeri ile rapor edilmektedir (Lamount ve ark., 2021; Fang ve ark., 2021; Sun ve ark., 2021).

Anlamlılık düzeyi p-değerinin rapor edildiği makalelerde yaygın bir hata da tablolarda gerçek değeri ile ($p=0.014$ veya $p=0.003$ gibi) veya p-değerinin çok küçük olması durumunda $p<0.001$ olarak ifade edilen değerin metin içinde $p<0.05$ şeklinde kısıtlanmasıdır. Yani tabloda p-değerinin gerçek değeri ($p=0.014$ gibi) yazılmalı ve bu değer bulgular, tartışma hatta özet bölümlerine aynı şekilde raporlanmalıdır.

Bir diğer yaygın hatalı durum, p-değerinin gerçekteki değerinin ne olduğu rapor edilmeden, tablo ve metin içinde $p<0.05$ veya $p>0.05$ eşik değerinin kullanılmasıdır. Eğer sonuç

çok sayıda grubun ortak bir bulgusu ise (örneğin Duncan, veya SNK çoklu karşılaştırma testleri gibi) ancak bu durumda $p<0.05$ veya $p<0.01$ şeklinde rapor edilebilir. Diğer durumlarda yukarıda vurgulandığı gibi p-değerinin gerçek değeri rapor edilmelidir. Bu bilgiler ışığında bazı istatistiksel anlamlılık (olasılık) değerlerinin doğru yazılımı aşağıdaki Tablo 2' de verilmiştir.

Tablo 2. İstatistiksel Anlamlılık p-değerinin Hatalı ve Doğru Yazım Örnekleri

Değişken	**Analiz Çıktısı Olasılık Değeri**	**Doğru Yazım**	**Hatalı Yazım**	**Hatalı Yazım Nedeni**
X1	0.0000025	$p<0.001$ veya $p=2.5x10^{-6}$	p=0.000	Olasılık sıfır olmaz
X2	0.000	$p<0.001$	p=0.000	Olasılık sıfır olmaz
X3	0.003	p=0.003	$p<0.01$	Bilgi kaybı, eski yazım şekli
X4	0.042	p=0.042	$p<0.05$	Bilgi kaybı, eski yazım şekli
X5	0.087	p=0.087	$p>0.05$	Bilgi kaybı, eski yazım şekli

Ayrıca tablolarda p-değerinin hesaplanmasında kullanılan örneklem büyüklüğü, kullanılan dağılım veya test, dağılımın test istatistiği (z, t, *F*, χ^2) tabloda veya başka bir şekilde rapor edilmemiş ise p-değeri metin içine yazılırken, test istatistiği ve serbestlik derecesi (sd) ile birlikte yazılmalıdır. Yani eğer analiz t dağılımı ile yapılmış ise bu istatistiksel bulgu "$t_{(sd)}$ = … ; p=…" formatında yazılmalıdır. Örneğin t dağılımı kullanılarak 25 örneklem büyüklüğü (sd=25-1=24) ile test istatistiği t = 2.104 ve p=0.023 elde edilmiş ise bu bulgu makale metninde [$t_{(24)}$=2.104; p=0.023] şeklinde yazılmalıdır. Ki-Kare testi için benzer şekilde [$\chi^2_{(sd)}$=… ; p=…] yazılmalıdır. *F* dağılımı pay *v1*, ve payda *v2* şeklinde iki serbestlik derecesine sahip olacağı için [$F_{(v1,v2)}$ =...; p=...] formunda yazılması gerekir. Yani 4 bağımsız grup ve her grupta 5'er gözlem kullanılarak elde edilen veriler *F* dağılımı ile analiz edilmiş olsun. Burada *F* test istatistiği 4.254 ve p=0.017 elde edilmiş ise bu bulgu [$F_{(3,16)}$ =4.254; p=0.017] şeklinde yazılır.

4. P-Değerinin Yorumlanması

Anlamlılık değeri p, büyüklüğü ve büyüklüğü ile birlikte taşıdığı anlam esas alınarak iki farklı boyutta yorumlanmaktadır.

4.1 P-değerinin Sayısal Büyüklüğüne Göre Yorumlanması

P-değerinin büyüklüğüne göre yorumlanmasında literatürde kabul görmüş standart bir uygulama mevcuttur. Araştırıcılar bu standarda göre bulgularını yorumlarken genelde hata yapmazlar. Kabul görmüş APA stiline göre p-değerinin büyüklüğü dikkate alınarak anlamlılık düzeyi "üç yıldız işaret" (asteriks) sistemi kullanılarak aşağıda Tablo 3'te verildiği gibi yorumlanır (Leahey, 2005; APA, 2010; Kul, 2014; Akbulut ve ark., 2015).

Tablo 3. P-değerinin Büyüklüğüne Göre Yorumlanması

p-değeri	Yorumu	Asteriks (İşaret)
$0.01 \leq p < 0.05$	İstatistiksel anlamlı.	*
$0.001 \leq p < 0.01$	Yüksek düzeyde istatistiksel anlamlı.	**
$p < 0.001$	Çok yüksek düzeyde istatistiksel anlamlı.	***
$0.05 \leq p < 0.10$	Sınırda anlamlılık veya anlamlılık eğilimi var.	+
$p > 0.10$	İstatistiksel anlamlılık yoktur.	NS: not significant

Anlamlılık düzeyi p-değerinin aldığı bu büyüklük değerlerinin işaretli (asteriks kullanarak sunum) sunumunda daha az hata yapılmaktadır. İstatistik analizin yapıldığı makalelerde "*" asteriksi istatistiksel anlamlılığın düzeyini göstermek için kullanılır. Yıldız asteriksi bazen testin türünü, (örneğin Mann Whitney U*, Friedman** gibi) göstermek amacıyla veya açıklama işareti olarak kullanılmaktadır. Özetle "*" işareti istatistiksel anlamlılık düzeyi için kullanılır. Başka bir durumu açıklamak amacıyla kullanılmamalıdır.

P-değeri hangi büyüklükte ve düzeyde olursa olsun, büyük ölçüde örneklem büyüklüğüne bağlı olduğu dikkatten kaçırılmamalıdır. Aynı büyüklükteki bir fark veya katsayı için, örneklem büyüklüğü arttıkça test istatistiği büyür ve "p" olasılığı küçülür. Örneğin populasyon ortalaması 165 ve örneklem ortalaması 171.5, standart sapması 26.5 olan çalışmada aynı örneklem ortalaması ve standart sapma için sağ kuyrukta tek ortalamanın hipotez testi için farklı örneklem büyüklükleri için test istatistikleri ve p değerleri aşağıdaki gibi (Tablo 4) şekillenir. Burada n=25 için $t = \frac{171.5-165}{\frac{26.5}{\sqrt{25}}} = \frac{6.5}{5.30} = 1.226$ ve p= 0.112 elde edilir. Bu olasılık düzeyi yorumlandığında bulgu anlamsız olarak nitelendirilir. Aynı olay için örneklem 10 kat artırılarak 250 olduğunda p-değeri 0.00007 olarak hesaplanır ve 25 örneklem ile 6.5 birimlik fark istatistiksel olarak anlamsız bulunurken aynı büyüklükteki fark 250 birimlik örneklemde p=0.00007 değeriyle çok çok anlamlı konuma gelir. Örneklem büyüklüğüne bağlı p değerlerinin değişimi Tablo 4'den izlenebilir. Tablo 4'de verildiği gibi, istatistiksel

anlamlılığın yorumlanmasında p değerlerinin örneklem sayısı ile yakından ilişkili olduğu dikkatten kaçırılmamalıdır.

Tablo 4. Aynı Dağılımlı Örneklemlerde p-Değerlerinin Örneklem Büyüklüğüne Göre Değişimi

Uygulama	n	S	Se	Fark	t	p
1	25	26.5	5.30	6.5	1.226	0.116
2	50	26.5	3.75	6.5	1.734	0.045
3	100	26.5	2.65	6.5	2.453	0.008
4	250	26.5	1.676	6.5	3.878	0.00007
5	1000	26.5	0.84	6.5	7.757	1.07254E-14
	Kitle Ortalaması =165		Örneklem Ortalaması =171.5			Fark=6.5

4.2 P-değerinin Anlam Yorumlanmasında Yapılan Hatalar

P-değerinin tanımında yaşanan tartışmalar p-değerinin anlam yorumlanmasında hatalara neden olmuştur. Bu bağlamda yazılan makalelerin bazıları yukarıda 3.2 numaralı "İstatistiksel Anlamlılık p Nedir?" başlığı altında özetlenmiştir. Makalenin bu kısmında literatür ışığında p-değerinin yorumlanmasında yapılan hatalar ele alınacaktır.

İstatistiksel anlamlılık p-değerinin tanımı incelendiğinde tanımın Tip I hatanın ve gözlenen bulgunun ortaya çıkma olasılığına vurgu yaptığı görülür. Halbuki araştırmacıların esas amacı, ilgilenilen değişkenler arasında olası bağıntıların varlığını ve bu bağıntıların gücünü anlamak ve açıklamaktır (Işık 2014). Yani aslında amaç, kısa ifade ile "etki büyüklüğü" değerini anlamaktır. Ancak istatistiksel anlamlılık p bu konuda bilgi vermediği gibi, araştırıcıları fark veya ilişkilerin önemsenecek düzeyde olup olmadığı sorusundan da uzaklaştırır (Işık 2014).

İstatistiksel anlamlılığın hatalı yorumlarından bir diğeri şu şekilde özetlenebilir. Araştırıcılar p-değerini α değerinden küçük bulmaları durumunda, doğru bir öneri ortaya koydukları ve iddialarını doğruladıkları düşüncesine sahip olmalarıdır. Tersi durumda ise araştırıcı yaptığı işin herhangi bir değere sahip olmadığı ve boşuna emek harcadığı duygusunu yaşamaktadır. Halbuki her iki düşüncede eksiktir ve hatalı yorumdur.

İlk durumda ($p<\alpha$), yani H_1 hipotezinin kabul edilmesi bu bulgunun kesin doğru olduğu anlamına gelmez. Örneğin istatistiksel anlamlılık değeri p=0.04 olması durumunda sıfır hipotezi H_0, reddedilir. Ancak bu kitleden aynı şartlarda ve büyüklükte 100 örneklemden 4'ünde sıfır hipotezinde ifade edilen durum geçerli olabilecektir. Yani araştırma hipotezi H_1

doğrulandığında araştırıcı p düzeyinde H_0 ile ifade edilen sonuçlara ulaşabilecektir. Şüphesiz bilimsel doğrulamalar kesin sonuçlar değildir. Yeni doğru üretilinceye kadar kabul gören geçici doğrulardır. Araştırma hipotezinin doğrulanması bu hipotezinde yanlışlanabilir olması ilkesini değiştirmez yani her teori yanlışlanabilme özelliğine sahiptir. Bilimsel gelişmelerin sürekliliği de bu ilkeye dayanır.

Diğer durumda ($p>\alpha$), genellikle $p>0.05$ gibi bir sonuca ulaşmak araştırıcıda faydasız bir iş yapma, yanlış bir araştırma tasarlama duygusu oluşturmaktadır. Erkuş (2017), bu durumu "yokluk hipotezinin reddedilememesinin dayanılmaz ağırlığı" tanımı ile açıklamıştır. Araştırıcılar $p>0.05$ bulgusuna ulaştıklarında 'yanlış araştırma tasarladım' düşüncesine kapılmaktadır. Bu düşünce aşağıdaki nedenlerden dolayı doğru değildir. Çünkü:

- $p>\alpha$ durumu gerçekten sonucun anlamsız olduğunu göstermez. Belki ulaşılan sonucun hatasının yüksek olabileceğini gösterir. Bununla birlikte p-değerinin anlamsız bulunması ($p>\alpha$), çalışmada bazı olası çevre etkilerinin kontrol edilememesinden veya örneklem hacminin yeterli büyüklükte olmamasından veya etki büyüklüğünün çok küçük tasarlanmış olmasından kaynaklanabilir. Çalışma daha kontrollü ve gerekli ve yeterli örneklem ile yürütüldüğünde daha gerçekçi sonuçlar bulunulabilecektir.
- Etkinin, farkın veya ilişkinin anlamsız bulunması daha sonra aynı konuda yapılacak çalışmalarda araştırmacılara bilgi ve deneyim sunar. Ayrıca araştırıcı yaptığı çalışma ile en azından ilgili konuyu tartışılabilir konuma getirmiştir.
- Araştırma sonucunun istatistiksel anlamsız ($p>0.05$) durumunda, araştırıcı araştırmanın yanlış tasarladığı düşüncesi ile yayınlamaz veya dergi editörleri bu tip makaleleri yeni bir şey bulunmadı gerekçesi ile ret ederler. Böylece aynı konuda hep $p<0.05$ bulgusuna ulaşan çalışmalar literatüre girer. Bu durum hem "tekrarlanabilirlik krizine; replicability crisis" hem de daha sonra yapılacak Meta Analiz çalışmalarında hatalı sonuca ulaşılmasına neden olur.

P-değerinin yorumlanmasında yapılan bir diğer hata, küçük p-değerinin, daha büyük bir etkiye işaret ettiği yönündeki algı veya yorumlamadır. P, etki büyüklüğü konusunda bilgi vermez. Birçok araştırıcı p-değerinin çok küçük olması durumunda daha büyük bir etki büyüklüğünü anlamlı buldukları şeklinde yorumlamaktadırlar. P-değerinin aldığı sayısal değer örneklem büyüklüğü ile doğrudan ilişkilidir. Örneklem büyüklüğü arttıkça p-değeri küçülür küçük farklılıklar veya ilişkiler anlamlı çıkar. Yani çok büyük örneklemler ile çok küçük farklılıklar dahi istatistiksel olarak anlamlı bulunabilir. P-değeri etkinin veya farkın olup olmadığını gösterir, ancak bunların büyüklüğü veya derinliği hakkında bilgi vermez. Etkinin

derinliği için "etki büyüklüğü" ölçüsünü belirlemek ve p-değerinin doğru yorumlanması için p-değerinin yanında raporlamak gereklidir.

P-değerinin çok küçük olması farklılığın veya ilişkinin biyolojik, teknik veya klinik anlamlı olduğunu göstermez. Biyolojik teknik veya klinik farklılığın anlamlı olup olmadığı etki büyüklüğü ile anlaşılır. Buradaki yanılgı istatistiksel anlamlılığın etki büyüklüğü gibi algılanmasıdır. Bu nedenle istatistiksel anlamlılık ile birlikte etki büyüklüğü rapor edilmeli ve yorumlanmalıdır. Ayrıca önsel olasılık α, rapor edilmeden yani 0.05 eşik değeri kullanılmadan p-değerinin büyüklüğü ve tespit edilen farklılık veya ilişki katsayılar ile birlikte yorumlanabilir. Yani p-değerinin tek başına "anlamlı bulundu" veya "anlamsız bulundu" şeklindeki kesin ve tekil yargılardan kaçınılmalıdır.

P-değeri tekil olarak yorumlandığında p-değerinin küçüklüğü nispetinde H_0 hipotezine karşı kanıtlar güçlenir. P-değeri yeterince küçükse H_0 hipotezi ret edilerek yerine H_1 hipotezi ikame edilebilir.

P-değerinin anlamlı bulunması olayın tüm tekrarlarında aynı sonucun bulunmasını garanti etmez ve tekrarlarda p-değerini öngörmek için kullanılamaz. Bir araştırma için elde edilen p-değeri o örneklem için geçerlidir. Ayrıca düşük güce sahip testler ile hesaplanan p- değerlerinin değişkenliği yüksektir. İstatistiksel anlamlılığın değişkenliğine neden olan bir diğer etken örneklemin küçük olmasıdır (Kabasakal 2019).

P-değerinin sunumu ve özellikle yorumlanması konusunda yapılan hataları Goddman (2008), 12 madde ile özetlerken, Greenland ve ark. (2016) ise bilimsel makalelerde anlamlılık değeri p için hatalı yorumları 18 başlık altında, güven aralığı ve testin gücü istatistikleri için yanlış yorum hatalarını 7 alt başlık da olmak üzere toplam 25 alt başlık altında irdelemişlerdir.

P-değerinin yanlış yorumlanmasında özet ve daha ileri bilgilere ulaşmak isteyen araştırıcılara (Goddman, 2008; Greenland ve ark., 2016; Wasserstein ve Lazar, 2016; Mark ve ark., 2016; Gao, 2020; Balkin ve Lenz, 2021) tarafından yayınlanan makalelerin incelenmesi önerilir.

5. **Sonuç**

Bilimsel çalışmalarda Yokluk Hipotezi Anlamlılık Testinin (YHAT) kullanımı çok yaygındır. YHAT sürecinin son ürünü anlamlılık seviyesi olarak adlandırılan p-değeridir. P-değeri istatistik biliminin en ünlü kavramıdır. Veri analizine dayalı çalışmalarda verilere ait tanımlayıcı istatistiklerin konum ve değişim ölçüleri ile birlikte doğru sunumu önemlidir.

Tanımlayıcı istatistikler örneklem büyüklüğü n ve S veya Se ölçüleri ile birlikte verilmeli, Konum ve değişim ölçüleri aynı ondalık düzeyinde sunulmalıdır.

YHAT dayalı araştırmalarda önsel hata düzeyi, Fisher (1925) tarafından yapılan önermesinden günümüze, α=0.05 olarak kullanılmaktadır. YHAT sonucunda p<α durumunda sonuç anlamlı olarak yorumlanmaktadır. Araştırmaya özgü p-değerinin sunumunda ve yorumunda önemli hatalar yapılmaktadır. Geleneksel olarak anlamlı bulunan p-değerinin raporlanması p<0.05 veya p<0.01 iken güncel yazımda p-değerinin gerçek değeri ile ve en az üç ondalıkla sunulması (Örneğin p=0.036 gibi) önerilmektedir. P-değerinin gerçek p=0.00 veya p; 0.000 şeklinde verilmesi son derece hatalı olup bu yazımlardan kaçınılmalıdır.

P-değerinin büyüklük olarak rapor edilmesinin yanında p-değerinin yorumlanmasında önemli hatalar yapılmaktadır. P-değerinin enformel tanımı için istatistik uzmanları farklı tanımlar yapmakta ve p-değerinin yorumu oldukça tartışmalıdır. Bununla birlikte p-değerinin yorumlanması aşağıdaki ilkeler doğrultusunda yapılmalıdır.

P-değerine mutlak doğru olarak bakmak hatalı bir yorumdur. P-değeri yokluk hipotezi ile ifade edilen durumun hangi olasılıkla gerçekleşebileceğini ifade eder. Keza p<0.05 araştırmanın doğru tasarlandığının ölçüsü olmadığı gibi, p>0.05 durumu araştırmanın hatalı tasarlandığını göstermez. Araştırmalarda p<0.05 durumu değerli olduğu kadar p>0.05 durumu da bilgi vericidir. Yani p<0.05 ise araştırma yayınlanabilir, p>0.05 ise bulgular yayınlanamaz duygusundan kaçınılmalıdır. Ayrıca p<0.05 durumu üzerinde aşırı yoğunlaşarak, p-değerine taşıdığı anlamdan daha fazla anlam yüklenmemelidir.

Araştırıcılar araştırma sonuçlarının p<0.05, yani anlamlı bulunması, yönünde eğilim göstermektedir. Bilimsel periyodiklerde genellikle p<0.05 bulgusuna ulaşan araştırmalar yayınlanmakta, diğer durum kabul görmemektedir. Bu durum literatürde bulguların tekrarlanabilirlik özelliğini kısıtlamakta, özellikle meta analiz çalışmalarında, yanlılığa neden olmaktadır.

P-değerinin küçük bulunması etkinin büyük olduğu anlamına gelmez, p-değeri etkinin olup olmadığı konusunda bilgi verir. Araştırma sonunda, fark, ilişki, katsayı vb. ölçüler için bulunan istatistiksel anlamlılık sonuçların teknik, klinik veya ekonomik anlamlı olduğu anlamına gelmez. Teknik, ekonomik veya klinik anlamlılık etki büyüklüğü ölçüsü ile belirlenir.

Bütün bu tartışmalara rağmen p-değeri istatistiksel çıkarımım önemli bir özeti olarak hala kullanılmaktadır. Daha güçlü alternatif çıkarım ölçüsü veya ölçüleri geliştirilinceye kadar bu ölçü kullanılacaktır. Ancak p-değerinin gerek yazımında gerekse yorumunda araştırıcıların yukarıda açıklanan hususları dikkate almaları önerilir.

Genel bir sonuç olarak araştırma raporlarının çıkarımsal istatistiği olarak p olasılığı tek başına verilmemelidir. Ayrıca yorumları p-değerinin anlamlılığı üzerine odaklamaktan

kaçınılmalıdır. Çünkü p düşünüldüğü ve ünlü olduğu kadar güvenilir bir ölçü değildir. Makalelerde p ile birlikte etki büyüklüğü (d), güven aralığı (CI), ve testin gücü (1-β, power) gibi istatistikler de rapor edilmelidir. P-değerinin yanında rapor edilmesi gereken bu çıkarımsal istatistik ölçüleri ve bu ölçülerin güçlü ve zayıf yönleri bir başka çalışma ile Türkçe literatüre kazandırılmalıdır.

Kaynaklar

Akbulut, Ö., Yıldız, N., & Orhan, H. (2015). İstatistik Analizlerde Temel Formüller ve Tablolar. Aktif Yayınevi s:8-9.

Akoğlu, H. (2015). https://acilci.net/kategori/eğitim-yöntem/akademi/istatistik ve metodoloji /P değeri ve Güven aralıkları: anlatılmaz yaşanır. Erişim 15.12.2021.

Anonim. (2021). Guidelines for Reporting Statistics.https://slideplayer.com/slide/6366206/ Erişim 4 Eylül 2021.

APA. (2010). American Psychological Association. (2010). Publication Manual of the American Psychological Association (6th Ed.). Washington, D.C.

Aschwanden, C. (2016). Statisticians Found One Thing They Can Agree On: It's Time To Stop Misusing P-Values. Five Thirty Eight. 17 Haziran 2016 tarihinde kaynağından arşivlendi. Erişim tarihi: 25 Temmuz 2016.

Balkin, R.S., & Lenz S. A. (2021). Contemporary Issues in Reporting Statistical, Practical, and Clinical Significance in Counseling Research. Journal of Counseling & Development, 99: 227-237

Vidgen, B., & Yasseri, T. (2016). P-Values: Misunderstood and Misused. Frontiers in Physics, 4(6): 1-5.

Bokai, W., Zhirou, Z., Hongyue, W., Xin, M T, & Changyong, F. (2019). The p-value and model specification in statistics, General Psychiatry,32(3): e100081.

Cengiz, M.A. & Terzi, Y. (2018). Hipotez Testleri Ders Notları. Ondokuz Mayıs Üniversitesi Fen-Edebiyat Fakültesi İstatistik Bölümü, Samsun.

Cohen, H.W. (2011). P values: use and misuse in medical literature. American Journal of Hypertension, 24(1):18-23.

Cohen, J., (1994). The earth is round (p<0.05). American Psychologist, 49(12): 997-1003.

Consonni, D., & Bertazzi P.A. (2017). Health significance and statistical uncertainty. The value of P-value. La Mediciana del lavoro, 108(5): 327-331.

Dahiru T. (2008). P – value, a true test of statistical significance? A cautionary note. Annals of Ibadan Postgraduate Medicine, 6(1):21-26.

Erkan, D.Ö. (2018). Güç analizi, etki büyüklüğü, p-değeri ve ötesi Akdeniz Üniversitesi İstatistik Danışma Birimi. Erişim: http://akhadyek.akdeniz.edu.tr/.

Erkuş, A. (2017). Denence Testi ve H_0 Denencesinin Reddedilememesinin Dayanılmaz Ağırlığı, Düşünce yazısı-Opinion paper İlköğretim Online, 16(4), dy: 12-16. [Online]: http://ilkogretim-online.org.tr

Fang C., Dong Xu, D., Su J., Dry J.R., & Bolan Linghu, B. (2021). DeePaN: deep patient graph convolutional network integrating clinico-genomic evidence to stratify lung cancers for immunotherapy npj Digital Medicine 4:14. https://doi.org/10.1038/s41746-021-00381-z

Gao, J. (2020). P-values – a chronic conundrum. BMC Medical Research Methodology. 20:167. https://doi.org/10.1186/s12874-020-01051-6

Gigerenzer, G. (2004). Mindless statistics. The Journal of Socio-Economics, 33:587–606.

Gigerenzer, G. (2018). Statistical rituals: The replication delusion and how we got there. Advances in Methods and Practices in Psychological Science,1(2): 198–218.

Goodman, S. (2008). A Dirty Dozen: Twelve P-Value Misconceptions. Seminars in Hematology, 45(3):135-140.

Göksöz, F. (2021). Hipotez Testleri I. Bölüm https://acikders.ankara.edu.tr/pluginfile. php/117324/mod_resource/content/1/10 Erişim tarihi: 23.08.2021

Greenland, S.J., Rothman, K.J., Carlin, J.B., Poole, C., Goodman S.N., & Altman, D.G., (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. European Journal of Epidemiology, 31:337–350.

Gürkan, A. (2007). Klinik Peridontolji araştırmalarında bağımsız iki grup ortalamasının karşılaştırılmasında örnek genişliği, istatistiksel güç ve anlamlılık. EÜ Diş Hekimliği Fak. Dergisi, 28: 123-134.

Işık, İ. (2014). Yokluk hipotezi anlamlılık testi ve etki büyüklüğü tartışmalarının psikoloji araştırmalarına yansımaları. Eleştirel Psikoloji Bülteni, 5:55-80

Kabasakal, C. (2016). P-değeri hakkında 3 yaygın yanılgı. BioMedya.com Erişim tarihi: 18.10. 2021.

Kanık, E.A. (2014). P değeri dedikleri. http://mestacon.tumblr.com/post/80738519327/p-degeri-dedikleri. Erişim: 15.12.2021.

Kartal, M. (2006). Bilimsel Araştırmalarda Hipotez Testleri, Nobel Dağıtım Ankara.

Kılıç, S. (2011). Neyin peşindeyiz? Kutsal p-değerinin mi (istatistiksel önemlilik) yoksa klinik önemliliğin mi? Journal of Mood Disorders, 1:46-48.

Kul, S. (2014) İstatistik Sonuçlarının Yorumu: P-değeri ve Güven Aralığı Nedir? Türk Toraks Derneği Dergisi, 11-13. DOI:10.5152/pb.2014.003.

Leahey, E. (2005). Alphas and Asterisks: The development of statistical significance testing standards in sociology, Social Forces, 84 (1): 1-24.

Leek, J.T., & Peng, R.D. (2015) Statistics: P values are just the tip of the iceberg. Nature. 520 (612). Doi: 10.1038/520612a.

Lehmann, E. L. (1993). The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two? Journal of the American Statistical Association, 88(424): 1242-1249.

Lu, Y., & Belitskaya-Levy, I. (2015). The debate about p-values. Shanghai Archives Psychiatry, 27 (6): 381-385.

Mark, D.B., Lee, K.L., & Jr Harrell, F. E. (2016). Understanding the role of P values and hypothesis tests in clinical research. JAMA Cardiology,1(9):1048-1054. Doi:10.1001/jamacardio.2016.3312

Meurs, J. (2016). The experimental design of postmortem studies: the effect size and statistical power. Forensic Science Medicine & Pathology,12:343–9 Doi 10.1007/s12024-016-9793-x

Murtaugh, P.A. (2014). In defense of P values Ecology. Ecological Society of America, 95(3): 611–617.

Nuzzo, R. (2014). Scientific method: Statistical errors. Nature, 506:150–152.

O'Leary, T.J. (2021) Rigor, Reproducibility, and the P value (Commentary). The American Journal of Pathology, 191(5): 806-808.

Piana, R. (2019). https://ascopost.com/issues/july-25-2019/is-it-time-to-reevaluate-the-p-value-in-biomedical-research/ Erişim tarihi: 6 Eylül 2021.

Startz, R. (2019). Not p-Values, Said a Little Bit Differently. Econometrics, 7(1):11; doi:10.3390/econometrics7010011

Sun, P., Lu, Q., Li, Z., Qin, N., Jiang, Y., Ma, H., Jin, G., Yu, H., & Dai J. (2021). Assessment of prognostic prediction models for gastric cancer using genomic and transcriptomic profiles. Elsevier Publ. Meta Gene 28 100890.

Szucs, D., & Ioannidis, J. P. A. (2017). When Null Hypothesis Significance Testing Is Unsuitable for Research: A Reassessment. Frontiers in Human Neuroscience, 11:390-410. doi: 10.3389/fnhum.2017.00390

Şenyay, L. (2021). Hipotez Testleri ve Güven Aralıkları. Ders Notu (9. Bölüm) Dokuz Eylül Üniversitesi İktisadi ve İdari Bilimler Fakültesi İzmir. https://pdf4pro.com Erişim tarihi: 24.9.2021.

Tramifow, D., & Marks M. (2015). Editorial. Basic and Applied Social Psychology,37: 1-2.

Wasserstein, R.L, & Lazar, N.A. (2016). Editorial. The ASA's statement on P-values: context, process, and purpose. The American Statistician,70(2): 129-133.

Yıldırım, H.H., & Yıldırım, S. (2011). Hipotez Testi, Güven Aralığı, Etki Büyüklüğü ve Merkezi Olmayan Olasılık Dağılımları Üzerine İlköğretim Online, 10(3): 1112-1123, [Online]: http://ilkogretim-online.org.tr

Yıldız, N., Akbulut, Ö., & Bircan, H. (2020). İstatistiğe Giriş (14. Basım) Kültür ve Eğitim Vakfı Yayınevi, Erzurum.