Basit Tesadüfi Sondajda Tahminlerin İsabet Derecesi ve Regresyon Metodu

Dr. M. Kemal YOĞTRTÇUGİL

İstanbul İktisat Fakültesi

1. Giriş :

Birtakım olaylar vardır ki bu olayları veya topluluğu teşkil eden birimler birbirine benzer ve bir birim gerek diğerlerini gerek ait olduğu yığını temsil eder. Hakkında bilgi edinilmek istenen böyle bir toplulukta mevcut birimlerden bir tanesinin müşahede ve tetkiki değerleme için kâfi gelecektir. Buna karşılık bazı olaylar vardır ki bu topluluğa mensup birimler müşterek bazı vasiflara sahip olmakla beraber başka bakımlardan çok fark ederler. Meselâ bir toprağın randımanı yıldan yıla değişir. Bu değişmeye tesir eden faktörler arasında iklim ve toprağın cinsine genel sebepler, uygulanan ziraat usulleri ile tohumun miktar ve kalitesi gibi faktörlere de tesadüfi sebepler denebilir.

İşte bir olay genel sebepler yanında tesadüfi sebeplere de tâbi bulunuyorsa, bu ikincilerin tesiriyle birimler arasında bazı farklar husule gelmekte ve böyle birimlerden ibaret topluluğa Kollektif Olay denmektedir.

Kollektif olayları araştırmaya mahsus olan ve çok sayıda birimi müşahede etmek, saymak veya ölçmek, sonuçları sınıflayıp tahlil etmek suretiyle çalışan ilmi usule ise İSTATİSTİK adı verilmektedir.

Bu tarifte adı geçen «çok sayıda birimin müşahedesi» tâbirinden kasıt nedir ? Bazı istatistikçiler bu çok sayıda birimden bütün birimlerin anlaşılması gerektiğini, diğer bir ifade ile hakkında bilgi edinilmek istenen

topluluğu teşkil eden bütün birimlerin müşahedesi lâzım geldiğini ileri sürerler. Bu görüşü hemen olduğu gibi kabul etmek biraz güçtür. Şöyle ki, bu usulün tatbiki umumiyetle zor, zaman alıcı, çoğu kere pahalı ve bazı hallerde ise imkânsızdır. Ayrıca bütün birimler kavransa bile hatalardan kaçınmak mümkün değildir ve tam sıhhatli olmamakla beraber gerçek durumu ifade edebilen değerlere bazen birimlerin bir kısmını müşahede suretiyle de varılabilmektedir.

Bu gibi sebeplerden ötürü hakkında bilgi edinmek istediğ miz topluluğu teşkil eden birimler arasından bir kısmını seçer ve yalnız bunları müşahedeye tâbi tutarız. Bu şekilde yapılan releveye KISMİ RELEVE, bütün birimlerin meydana getirdiği topluluğa ANA KÜTLE, bu ana kütleden çekilen birimlerden müteşekkil topluluğa NUMUNE denmekte ve bu numune değerlerinden hareketle ana kütle gerçek değerlerinin tahminine girişilebilmektedir. Ancak ana kütle gerçek değerleri ile numunenin verdiği değerler arasında tahminlerin birtakım hatalar ihtiva etmesi bakımından farklar olmakta ve neticeler bu hataları küçültebildiğimiz ölçüde değer kazanmaktadır.

Numuneye dahil edeceğimiz birimleri iki tarzda seçebiliriz.

1. Yapacağımız tahminin mümkün olduğu kadar isabetli olmasını istediğimizde temsili olmadığına kanaat getirdiğimiz birimleri müşahede dışı bırakmak. Böyle bir seçime «İradi Seçim» denir.

2. Numuneye girecek birimlerin seçimini tesadüfi şartlar altında yapmak. Tesadüfi şartlar altında yapılan böyle bir seçimde şu özellikler vardır:

a. Belli bir ana kütleden çekilmesi mümkün bütün numuneleri nazari olarak tarif edebiliriz.

b. Ana kütleden çekilmesi mümkün bütün numunelerin seçilme ihtimali nazari bakımdan belli olup numuneyi bu ihtimallere uygun olarak seçebiliriz.

2. Tahmin Hataları:

Aynı ana kütleden tesadüfi seçimin şartları yerine getirilmek suretiyle alınan çeşitli numunelere istinaden ana kütlenin herhangi bir karakteristiği tahmin edilebilir. Ana kütle mevcudu N, numune mevcudu n ise kabul ettiğimiz seçim usulü gereğince çekilmesi mümkün birbirinden farklı numune sayısı $C_N^n = \dfrac{N!}{n!\,(N-n)!} = m$ dir. Bu numunelerin çekilme ihtimalleri $(p_1, p_2, \ldots p_m)$ tahmin edilmek istenen ana kütle karakteristi-

ğinin hakiki değeri ($\bar{X}$), çekilen numuneye istinaden yapılan tahmini ($\hat{\bar{X}}$) ile gösterelim. ($\hat{\bar{X}}$) nin değeri çekilen numunelere göre ($\bar{x}_1, \bar{x}_2, \ldots\ldots \bar{x}_m$) şeklinde değişecektir. ($\hat{\bar{X}}$) tahmininin matematik ümidi $E(\bar{x}) = \Sigma p_i \bar{x}_i$ dir. Bu değer $E(\bar{x})$, gerçek değer ($\bar{X}$) ya eşit olduğu takdirde yapılan tahminin sistematik bir hata ihtiva etmediği söylenir. Bununla beraber münferit bir tahmin ($\bar{x}_i$), ($\bar{X}$) den farklı olabilir. İşte böyle bir hataya tesadüfi hata veya sondaj hatası diyoruz. Ancak sondaj usullerine başvurularak yapılan tahminlerin bir çoğu sistematik hata [$H = E(\bar{x}) - \bar{X}$] ihtiva ederler. Ana kütle ortalaması tahminin tipik hatası ($s_{\bar{x}}$) ise bu hatanın tahmin edilmek istenen değer hakkında varacağımız hükme tesiri şöyledir : ($\bar{x} - 1.96\ s_{\bar{x}}$) ve ($\bar{x} + 1.96\ s_{\bar{x}}$) hudutları arasında tahminlerin % 95 i bulunur. Diğer bir ifade ile, ($\bar{x} - 1.96\ s_{\bar{x}}$) den küçük veya ($\bar{x} + 1.96\ s_{\bar{x}}$) den büyük bir tahmine rastlamak ihtimali % 5 dir. Aşağıdaki tabloda (Tablo 1.) negatif bir sistematik hata halinde ($s_{\bar{x}}$) cinsinden ifade olunan sistematik hata ($H/s_{\bar{x}}$) nin muhtelif değerleri için ($\bar{x} - 1.96\ s_{\bar{x}}$) den küçük ve ($\bar{x} + 1.96\ s_{\bar{x}}$) den büyük bir tahmin yapma ihtimalleri ve bu ihtimaller toplamı gösterilmiştir. Sistematik hata pozitif olduğu takdirde ihtimaller yer değiştirmelidir.

Tablo 1.

$H/s_{\bar{x}}$ nin muhtelif değerleri için ihtimaller toplamı

$H/s_{\bar{x}}$	$\bar{x} - 1.96\ s_{\bar{x}}$ den küçük bir değere rastlama ihtima.	$x + 1.96\ s_{\bar{x}}$ den büyük bir değere rastlama ihtima.	İhtimaller toplamı
— 0.02	0.0262	0.0238	0.0500
— 0.04	0.0274	0.0228	0.0502
— 0.06	0.0287	0.0217	0.0504
— 0.08	0.0301	0.0207	0.0508
— 0.10	0.0314	0.0197	0.0511
— 0.20	0.0392	0.0154	0.0546
— 0.40	0.0594	0.0091	0.0685
— 0.60	0.0869	0.0052	0.0921
— 0.80	0.1230	0.0029	0.1259
— 1.00	0.1685	0.0015	0.1700
— 1.50	0.3228	0.0003	0.3231

Görülüyor ki sistematik hata arttıkça tesadüfi tahmin hatalarını $(\bar{x} - 1.96\, s_{\bar{x}})$ ve $(\bar{x} + 1.96\, s_{\bar{x}})$ hudutları arasında tutabilme ihtimali azalmaktadır. Ancak sistematik hatanın mutlak değeri tahminlerin tipik hatasının % 10 nu aşmadıkça tesiri pek mühim telâkki edilemez. (Tablodaki 0.0511 eğer sistematik hata olmasaydı 0.0500 olacaktı).

Kısaca şunu belirtmek isteriz ki, sondaj teorisinde işlenen sistematik hatalar meydana geliş sebepleri bulunduğu takdirde ortadan kaldırılabilmekte, tesadüfi hatalar ise büyük sayılar kanununun işlemesi ile her iki istikamette tezahür ettiklerinden birbirlerini götürmektedirler. Bununla beraber ana kütleden çekilmesi mümkün bütün numuneleri değil, yalnız birini göz önünde bulundurduğumuzdan tesadüfi seçimin şartları yerine getirilmiş olsa bile sondaj hatası önlenememekte, ancak bunu kontrol etmek, muayyen hudutlar arasında tutmak ve ehemmiyetini azaltmaya çalışmak mümkün olmaktadır.

3. Basit Tesadüfi Sondaj ve Tahminlerin isabet derecesi :

Bir ana kütlede mevcut birimlerin herbirine eşit seçilme şansı tanıyarak çekilen bir numuneye göre yapılan tahmine «Basit Tesadüfi Sondaj» denilmektedir. Bu usulde ana kütle ortalaması $(\bar{X})$ numune ortalaması $(\bar{x})$ nin sistematik hata ihtiva etmiyen bir tahmini olarak alınmakta ve $(\hat{\bar{X}} = \bar{x})$ yazılmaktadır.

Tahminlerin isabetine tesir eden başlıca iki unsur numune mevcudu ve ana kütle mütehavvilliğidir. Numune mevcudu arttıkça veya numunenin içinden çekildiği ana kütlenin mütehavvilliği azaldıkça tahminlerin değişim sahası veya kısaca tipik hatası azalır ve dolayısıyle isabet derecesi artar. Ancak numune birim sayısının daima büyük sayılar kanununun az çok işlemesine müsaade edecek çoklukta olması arzu edilirse de, araştırıcı bazen elinde olmıyan sebeplerden numune büyüklüğünü küçük tutmak mecburiyetinde kalabilir. Bilhassa Tıp ve Biyoloji ile ilgili incelemelerde karşılaşılan bu durumda numunenin tesadüfi sayılıp sayılamıyacağı da ayrıca tetkik edilmelidir. Bunun yanında araştırıcı numuneye girecek birim sayısını ayarlıyabilecek durumda olsa bile tipik hatayı küçültmek gayesiyle çoğu kere numune mevcudunu istediği çoklukta tesbit edememe durumuyla karşı karşıya kalır ve numune büyüklüğü kısaca ne derece kesin neticelere ihtiyaç olduğu, teferruat ve tetkik konusu olayın değişkenlik derecesi, mâli imkân gibi çeşitli faktörlere göre belirir.

Basit tesadüfi sondajda ana kütle birimleri arasındaki mütehavvillik dereceleri düşünülmediğinden tahminlerin isabet derecesi düşük kalmakta, tahminin değerini yükseltmek için birimler arasındaki mütehavvilli

nazara alan zümrelere göre sondaj usulüne başvurulmakta ve teşkil edilen zümrelerden uygun görülecek miktarlarda birimi ihtiva eden numune incelenmektedir.

Ana kütle mütehavvilliğini göz önüne alan zümrelere göre sondaj usulüne girmeden basit tesadüfi sondajda tahminlerin isabet derecesini arttırmak için bir diğer yol da, bu olaya ait iki vasfın şıkları arasındaki münasebetten faydalanarak en küçük kareler metodu yardımıyla tayin edilen bir REGRESYON DOĞRUSU'nu tahminde kullanmaktır. Bu tahminde alınan numunenin (X) vasfı yanında bu vasıfla münasebette olan başka bir (Y) vasfının bilinen özelliklerinden tamamlayıcı bilgi olarak istifade edilir. Bu yolda zaman zaman kullanılan farkla tahmin ve nisbet usulleri ise esas metod olan Regresyon usulünün iki özel halini teşkil ederler. Şöyle ki, farkla tahmin usulünde regresyon doğrusu yine temeldir fakat regresyon katsayısı (b) numune yerine önceden tayin edilen bir usule göre hesap edilmekte, nisbet usulünde ise alınan iki vasıf arasındaki münasebetin doğrusal olduğu ve temsil ettikleri doğrunun orijinden geçtiği kabul edilmektedir. Bu usulde ana kütle ortalamasının tahmini $\left(\hat{\bar{X}} = \frac{\bar{x}}{\bar{y}} \cdot \bar{Y}\right)$ dir. Ancak böyle bir tahminin daima bir sistematik hata ihtiva ettiği görülmekte, buna mukabil tahminlerin standart hatası doğrudan doğruya tahmin usulünden daha küçük çıkmaktadır.

Pratik ihtiyaçlar bakımından aynı olay hakkında yapılması mümkün bütün tahminler ortalamasının gerçek değere uygun olup olmadığı hususundan çok münferit tahminlerin bu değere ne kadar yakın olduğu konusu önemli olduğundan, bu kusuruna rağmen basit tesadüfi sondajda regresyon usulü tercih edilmektedir.

4. Basit Tesadüfi Sondajda Regresyon metodu :

Ana kütleden çekilen n birimlik numune için (X) ve (Y) vasıfları arasında doğrusal bir münasebetin mevcut olduğu tesbit edilmişse, ana kütle birimlerinin kartezyen koordinat sisteminde temsil edeceği $P(X_i : Y_i)$ noktaları, denklemi $\bar{X} = \bar{x} + B(\bar{Y} - \bar{y})$ olan ve orijinden geçmiyen doğru etrafında toplanır.

$$B = \frac{\Sigma (X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma (Y_i - \bar{Y})^2}$$

Varyanslar toplamını minimum yapan ve $(\bar{X} ; \bar{Y})$ noktasından geçen doğrunun eğimi olan (B) ye (X) in (Y) ye nazaran Regresyon katsayısı denir.

$$\overline{X} = \overline{x} + B(\overline{Y} - \overline{y})$$

ile hesaplanan $(\overline{X})$, (X) vasfı için hakiki ortalamanın tahmini değeridir ve sonuç daha doğru şekilde

$$\hat{\overline{X}} = x + B(\overline{Y} - \overline{y})$$

yazılır. Ancak formülde geçen (B) parametresi ana kütle değerlerine dayanılarak hesap edildiğinden bizim için meçhul bir değerdir ve tahmini olarak numune değerlerine göre hesaplanan (b) alınırsa

$$b = \frac{\Sigma(x_i - \overline{x})(y_i - \overline{y})}{\Sigma(y_i - \overline{y})^2} \quad (1)$$

ve ana kütle ortalamasının tahmini

$$\hat{\overline{X}} = \overline{x} + b(\overline{Y} - \overline{y})$$

olacaktır. Regresyon doğrusu ile yapılan bu tahminlerin variyansı ise

$$V(\hat{\overline{X}}) = \frac{N - n}{n \cdot N} \cdot S_x^2 (1 - R^2)$$

dir. Ancak gerek (S_x^2) gerek (R^2) ana kütle değerleri olduklarından bunlar yerine tahminleri olan numune değerlerini koymak suretiyle

$$v(\hat{\overline{X}}) = \frac{N - n}{n \cdot N} \cdot s_x^2 \cdot (1 - r^2)$$

elde edilir.

5. Misal :

On birimli bir ana kütlede birimlerin gelirleri itibariyle bölünmesi aşağıda gösterilmiştir.

Tablo 2.

Birim	Gelir	Birim	Gelir
A	100	F	250
B	150	G	250
C	180	H	300
D	200	İ	400
E	220	J	450
Toplam	2500		
Ortalama	250		

Ortalama 250 liralık gelire sahip böyle bir ana kütlenin aslında bu karakteristiğinin bilinmediğini ve tesadüfen çekilecek 4 birim ihtiva eden bir numune yardımıyle tahmin edilmek istendiğini kabul edelim. Bu ana kütleden tesadüfi seçimin şartları yerine getirilmek suretiyle herbiri 4 birim ihtiva eden $C_{10}^{4} = 210$ numune çekebileceğimizi ve ana kütleye dahil her birimin çekilecek numunede bulunma ihtimalinin $(n/N = 4/10 = 0.4)$ olduğunu biliyoruz. Çekilmesi muhtemel 210 numune adayından bir tanesi olan ve tesadüfen çekilen (B E G İ) yardımıyla ana kütle karakteristiğini tahmine çalışalım.

Tablo 3.

Numune	değerleri
Birim	Gelir (lira)
B	150
E	220
G	250
İ	400
Toplam	1020
Ortalama	255

Basit tesadüfi sondaj şartları yerine getirilmişse numune ortalaması ana kütle ortalamasının sistematik hata ihtiva etmeyen bir tahmini olduğundan $(\widehat{\bar{X}} = \bar{x})$ yazar ve ana kütle ortalaması tahmininin (255) lira olduğunu söyleyebiliriz. Eğer çekilmesi muhtemel 210 numunenin hepsini çekebilsek ve ortalamalarını hesap edebilsek, bu ortalamalardan ibaret bölünmenin ortalaması (250) lira olacaktı. Numune değeri (255) ile ana kütle değeri (250) arasındaki (5) liralık fark sondaj hatasıdır. 4 birim ihtiva eden diğer bir numune (A E H J) ise bu takdirde işlenen hata $267.5 - 250 = 17.5$ liradır.

Ortalamadan inhiraf mahiyetinde olan bu hataların ölçüsü tipik hata adını almakta

$$s_{\bar{x}} = s_x \cdot \sqrt{\frac{N-n}{n \cdot N}} \qquad s_x = \sqrt{\frac{\Sigma (x_i - \bar{x})^2}{n-1}}$$

ve numunemiz (B E G İ) için

$$s_x = \sqrt{\frac{(105)^2 + (35)^2 + (5)^2 + (145)^2}{3}} = \sqrt{11.100} = 105.37$$

$$s_{\bar{x}} = 105.37 \sqrt{\frac{6}{40}} \simeq 41 \quad \text{olmaktadır.}$$

Oldukça yüksek çıkan bu tipik hatayı azaltmak için iki yol mevcuttur. Numune mevcudunu arttırmak veya tahmini regresyon doğrusundan istifade ile yapmak. İkinci şekli tercih ediyoruz. Bu takdirde ana kütle birimlerinin diğer bir vasfını bulmamız ve bu iki vasfın şıkları arasındaki münasebetten faydalanmamız gerekecektir.

Farzedelim ki ana kütleye ait birimlerin gelirleri yanında ödedikleri kira miktarları da bilinmektedir.

Tablo 4.

ANA KÜTLE		NUMUNE	
Birim	Kira (lira)	Birim	Kira (lira)
A	20	B	25
B	25	E	45
C	30	G	45
D	35	İ	50
E	45	Toplam	165
F	40	Ortalama	41.25
G	45		
H	50		
İ	50		
J	60		
Toplam	400		
Ortalama	40		

$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(y_i - \bar{y})^2$
— 105	— 16.25	+ 1706.25	264.1
— 35	+ 3.75	— 131.25	14.1
— 5	+ 3.75	— 18.75	14.1
+ 145	+ 8.75	+ 1268.75	76.6
		+ 2975.—	368.9
		— 150.—	
		+ 2825.—	

Yukarıdaki verilerden hareketle regresyon parametresi

$$b = \frac{2825}{368.9} = 7.66 \quad \text{ve}$$

ana kütle ortalamasının tahmini

$$\hat{\bar{X}} = \bar{x} + b\,(\bar{Y} - \bar{y}) \qquad \text{den}$$

$$\hat{\bar{X}} = 255 - 7.66\,(40 - 41.25) = 245.4 \quad \text{liradır.}$$

Sondaj hatası $250 - 245.4 = 4.6$ lira ve tahminin tipik hatası

$$v\,(\hat{\bar{x}}) = \frac{N - n}{n \,.\, N} \cdot s_x^2\,(1 - r^2) \quad \text{den}$$

$$= \frac{10 - 4}{4 \,.\, 10} \cdot 11100\,[1 - (0.8)^2] \qquad (2)$$

$$= 599.4 \qquad \text{ve}$$

$$s_{\bar{x}_{reg}} = 24.5 \qquad \text{dür.}$$

6. Sonuç :

Sondaj teorisi bir ana kütle ve bu ana kütleden çekilen numune arasında mevcut münasebetleri tetkik eder. Bu münasebet, ana kütle bilinmeyen değerlerinin «parametrelerin», numune değerinden veya statistiklerden tahminine dayanır. Bu tahminin muteber olabilmesi için numunenin içinden çekildiği ana kütleyi temsil etmesi lâzımdır. Temsili numune elde edebilmek için kullanılan usullerden bir tanesi de Basit Tesadüfi Sondaj'dır. Bu usulde, çekilen numune statistiğinin ana kütle parametresi ile tamamen aynı olması beklenemez, sadece gerçek değerin numune değerine yakın bir yerde olduğu söylenebilir. Gaye, bulunan numune değerinin gerçek değerin ne kadar iyi bir tahmini olduğunu tayin etmektir. Bir numune değeri üzerinden tesis edilebilecek bu itimadın derecesi bu numune ile aynı büyüklükte çekilmiş çok sayıdaki numuneler statistiklerinin dağılmasına bağlıdır. İşte bu dağılmanın ölçüsü olan tipik inhiraf ele alınan numune değerinin tipik hatasıdır ve bu kıymeti küçültebildiğimiz derecede tahminlerimizin isabet derecesi artmaktadır.

Ana kütle mütehavvilliğine tesir edemediğimiz gözönüne alınırsa tipik hatayı azaltmanın ilk bakışda numune mevcudunu arttırmak suretiyle sağlanabileceği zannedilmekte, ancak çoğu kere araştırıcı, elinde

olmayan türlü sebeplerden bunu da gerçekleştirememektedir. Bu takdirde ele alınan olayda mevcut iki vasfın şıkları arasındaki münasebetten faydalanmak ve bu suretle bulunan Regresyon Doğrusunu tahminde kullanmak en kolay yol olmaktadır.

(1) Numune mevcudunun çok olması halinde $B = b$ olduğunun ispat için bak. Doç. Dr. Kenan URAL «Sondajda özel bir tahmin metodu ve diğer metodlarla mukayesesi» Doç. TEZİ İstanbul 1967

(2) $$r = \frac{\Sigma (x - \bar{x})\ (y - \bar{y})}{\sqrt{\Sigma (x - \bar{x})^2 \cdot \Sigma (y - \bar{y})^2}} = \frac{2825}{\sqrt{33300.369}} = \frac{2825}{3505} = 0.8$$