فاصله‌ی اطمینان دقیق و تقریبی برای پارامتر نسبت جامعه

رحیم محمودوند[۱]

چکیده

برای محاسبه ی برآورد فاصله ای نسبت جامعه در روش های آماری غالباً از فرمول های تقریبی استفاده می کنند که قطعاً در برخی موارد ممکن است شرایط استفاده از تقریب بطور مطلوبی برقرار نباشد. در این مقاله سعی داریم تا یک روش دقیق برآورد فاصله ای نسبت جامعه را بیان کنیم. در پایان با استفاده از شبیه سازی به مقایسه ی این روش با روش تقریبی والد می پردازیم.

واژه های کلیدی: فاصله اطمینان، روش عمومی، شبیه سازی، نسبت جامعه.

۱–مقدمه

مشاهده های آماری ممکن است در برخی موارد دو حالت ممکن داشته باشند یا اینکه بر اساس یک ویژگی به دو دسته تقسیم شوند. برای مثال ممکن است افراد را از لحاظ عینکی بودن یا نبودن به دو گروه عینکی ها و غیر عینکی ها تقسیم کنیم یا پاسخ افراد به یک سوال ممکن است به صورت بله یا خیر باشد. چنین داده هایی را که دو حالت ممکن دارند، داده های دودویی می نامند. برای راحتی معمولاً یکی از دو حالت را با ۰ و دیگری را با ۱ نشان می دهند. تخصیص ۰ یا ۱ به هر کدام از دو حالت اختیاری است. در چنین داده هایی محقق مایل است درباره ی وجود یا عدم وجود یکی از دو حالت در واحد های جامعه نتیجه گیری کند. پس چنانچه از اعداد ۰ و ۱ به جای حالت ها استفاده کنیم، هدف مقایسه ی تعداد صفرها و یک هاست.

فرض کنید یک جامعه ی N عضوی مورد بررسی قرار گرفته است و نتایج به صورت صفر و یک ثبت شده‌اند. در صورتی که تعداد یک ها در این نمونه ی N تایی برابر X باشد، پارامتر نسبت جامعه را با p نشان داده و $p=X/N$. بنابراین اگر $X_1,\ldots,X_N$ جامعه ی تحت بررسی باشد که در آن هر X_i دو مقدار صفر و یک را می گیرد، داریم $p=\sum_{i=1}^{N}X_i/N=\mu$. بنابراین نسبت جامعه در واقع برابر میانگین جامعه ای است که مقادیر آن صفر و یک است.

در صورتی که در یک نمونه ی n تایی، تعداد یک ها برابر x باشد، یک برآورد شهودی برای پارامتر نسبت جامعه، نسبت در نمونه است که با $\hat{p}=x/n$ نشان داده می شود. همچنین به طور مشابه اگر $x_1,\ldots,x_n$ مقادیر نمونه ای به دست آمده از این جامعه باشد، داریم $\hat{p}=\sum_{i=1}^{n}x_i/n=\bar{x}$. پس یک برآورد برای نسبت جامعه برابر میانگین نمونه ی به دست آمده از یک جامعه با مقادیر صفر و یک است.

تحقیق های زیادی درباره ی فاصله ی اطمینان برای نسبت جامعه وجود دارد که در اینجا به چند مورد اشاره

[1] مدرس دانشگاه پیام نور مرکز تویسرکان

شده است. ویلسون[1][٦] بر مبنای نسبت نمونه ای و با در نظر گرفتن تقریب نرمال برای توزیع آن یک فاصله ی اطمینان تقریبی برای نسبت جامعه به دست آورده است. والد و ولفویتز[2][٥] نیز با استفاده از تقریب نرمال برای توزیع دو جمله ای، تقریب دیگری برای فاصله ی اطمینان نسبت جامعه ارایه نمودند. اگرستی و کول[3][٣] نیز با اعمال یک اصلاح در فرمول والد و ولفویتز فاصله ی اطمینان تقریبی دیگری ارایه کرده اند.

با وجود اینکه استفاده از این روش های تقریبی در بسیاری موارد نتایج بسیار مناسبی ارایه می کنند لیکن در برخی موارد نیز نتایج چندان رضایت بخش نیست و باید به دنبال راه حل های بهتری بود. در این مقاله یک روش دقیق برای فاصله ی اطمینان نسبت جامعه ارایه شده است.

در بخش دوم روش های تقریبی ذکر شده در مقدمه به طور مختصر بیان می شوند. در بخش سوم روش دقیق ارایه می شود و در بخش پایانی بر مبنای یک مثال به مقایسه ی این روش ها پرداخته می شود.

۲- فاصله ی اطمینان تقریبی برای نسبت جامعه

ابتدا توجه کنید که اگر X_iها متغیرهای تصادفی مستقل دو حالتی با مقادیر صفر و یک باشند $\sum_{i=1}^{n} X_i \sim B(n,p)$ است. از طرفی می توان تحت برخی شرایط ($np \geq 5$, $n(1-p) \geq 5$) توزیع دو جمله ای را با استفاده از توزیع نرمال تقریب زد. در صورت برقراری شرایط داده شده داریم:

$$Q = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \sim N(0,1).$$

یعنی Q یک کمیت محوری است. بر اساس این کمیت ویلسون[٦] فاصله ی اطمینان تقریبی زیر رابه دست آورد (جزئیات مربوط به نحوه ی به دست آوردن این فاصله در هاگ و تانیس[٤] داده شده است).

$$\frac{\hat{p} + z^2/(2n) - z\sqrt{\hat{p}(1-\hat{p})/n + z^2/(4n^2)}}{1+z^2/n} \leq p \leq \frac{\hat{p} + z^2/(2n) + z\sqrt{\hat{p}(1-\hat{p})/n + z^2/(4n^2)}}{1+z^2/n}, \quad (1)$$

که در این رابطه z برابر چندک توزیع نرمال استاندارد است. با صرف نظر کردن از عبارات z^2/n، $z^2/(4n^2)$ و $z^2/(2n)$ در رابطه ی (۱) به رابطه ی زیر می رسیم

$$\hat{p} - z\sqrt{\hat{p}(1-\hat{p})/n} \leq p \leq \hat{p} + z\sqrt{\hat{p}(1-\hat{p})/n}. \quad (2)$$

بر این مبنا والد و ولفویتز[٥] فاصله ی اطمینان تقریبی زیر را، که شاید بتوان گفت پر استفاده ترین فرمول برای برآورد فاصله ای نسبت جامعه است، معرفی نمودند.

$$\max\left\{0, \hat{p} - z_{1-\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}\right\} \leq p \leq \min\left\{1, \hat{p} + z_{1-\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}\right\}. \quad (3)$$

چنان که گفته شد، فرمول های تقریبی (۱) و (۳) با فرض برقراری شرایط $np \geq 5$, $n(1-p) \geq 5$ مناسب هستند. اما این تقریب برای مثال زمانی که $n = 300$ و $p = 0/01$ خیلی خوب نیست. پس روش اخیر در بعضی شرایط حتی برای n های بزرگ نیز خوب نخواهد بود. در این مقاله سعی داریم تا خواننده را با یک روش دقیق که در هر شرایطی قابل استفاده است آشنا کنیم. برای این منظور از روش عمومی برای بدست آوردن برآورد فاصله ای نسبت استفاده می کنیم. برای بررسی بیشتر در این زمینه به مشکانی[٢] و یا پارسیان[١] مراجعه کنید.

۳- ارتباط توزیع گسسته دو جمله ای و توزیع پیوسته فیشر

در این بخش به اختصار رابطه ی بین این توزیع ها را بیان می کنیم.

لم ۱- اگر $V \sim Beta(r, n-r+1)$ و $U \sim B(n,p)$ آنگاه $P(U \leq r-1) = P(V > p)$.

اثبات: پارسیان[١] را ببینید.

لم ۲- اگر $W \sim F_{(m,n)}$ آنگاه

[1] Wilson
[2] Wald and Walfowitz
[3] Agresti and Coull

$$V=\frac{\frac{m}{n}W}{\frac{m}{n}W+1}\sim Beta\left(\frac{m}{2},\frac{n}{2}\right).$$

اثبات: پارسیان[۱] را ببینید.

با استفاده از لم ۱ داریم:

$$P\left[V\le p\right]=\sum_{t=r}^{n}\binom{n}{t}p^{t}(1-p)^{n-t}, \qquad (٤)$$

با به کار بردن لم ۲ در رابطه ی (٤) نتیجه می شود که:

$$\begin{aligned} P\left[V\le p\right]&=P\left[\frac{\frac{r}{n-r+1}W}{\frac{r}{n-r+1}W+1}\le p\right]\\ &=P\left[W\le\frac{r}{n-r+1}.\frac{p}{1-p}\right] \qquad (٥)\\ &=\sum_{t=r}^{n}\binom{n}{t}p^{t}(1-p)^{n-t}, \end{aligned}$$

که در این رابطه W دارای توزیع فیشر $F_{(2r,2(n-r+1))}$ است.

۴– فاصله اطمینان دقیق برای نسبت

همانطور که در مقدمه گفتیم برای معرفی روش جدید می خواهیم از روش عمومی استفاده کنیم. فرض کنید یک نمونه تصادفی از چگالی $f_X(x;\theta)$ در اختیار است و می خواهیم بر اساس این نمونه یک برآورد فاصله ای با ضریب اطمینان $(1-\alpha)$ برای پارامتر مجهول θ بدست آوریم. در روش عمومی ابتدا یک آماره مناسب مانند T، که می تواند آماره بسنده یا برآوردگر درستنمایی ماکزیمم باشد، انتخاب می کنیم. فرض کنید تابع چگالی احتمال آماره ی T با $f_T(t;\theta)$ نشان داده می شود. برای بدست آوردن فاصله اطمینان مورد نظر از دو رابطه زیر استفاده می کنیم:

$$\sum_{t=-\infty}^{h_1(\theta)}f_T(t;\theta)=\alpha_1 \quad , \qquad \sum_{t=h_2(\theta)}^{\infty}f_T(t;\theta)=\alpha_2,$$

که در آن $\alpha_1+\alpha_2=\alpha$. (به پارسیان[۱] مراجعه شود)

در اینجا فرض بر این است که یک نمونه ی تصادفی $X_1,\dots,X_n$ از جامعه ای برنولی در اختیار است و می خواهیم نسبت در این جامعه را برآورد کنیم. بر مبنای روش عمومی $T=\sum_{i=1}^{n}X_i$ یک آماره بسنده ی کامل برای پارامتر p است و بعلاوه $T\sim B(n,P)$ است. بنابراین فاصله اطمینان مورد نظر از معادلات زیر محاسبه می شوند.

$$\sum_{t=0}^{h_1(P)}\binom{n}{t}P^{t}(1-P)^{n-t}=\frac{\alpha}{2},$$

$$\sum_{t=h_2(P)}^{n}\binom{n}{t}P^{t}(1-P)^{n-t}=\frac{\alpha}{2},$$

اکنون اگر (P_L,P_U) فاصله اطمینان مورد نظر باشد و t_0 مقدار مشاهده شده آماره T باشد، آنگاه فاصله ی مورد نظر از حل معادلات زیر بدست می آیند:

$$\sum_{t=0}^{t_0}\binom{n}{t}P_U^{t}(1-P_U)^{n-t}=\frac{\alpha}{2},$$

$$\sum_{t=t_0}^{n}\binom{n}{t}P_L^{t}(1-P_L)^{n-t}=\frac{\alpha}{2},$$

برای حل این معادلات می توان از جدول توزیع دوجمله ای استفاده نمود. حل معادلات فوق به دلیل گسسته بودن توزیع دو جمله ای کار مشکلی است. جواب های این معادلات با استفاده از معادله (۵) به صورت زیر است:

$$\begin{aligned} P_L&=\frac{m_L F_{m_L,n_L,\alpha/2}/n_L}{1+m_L F_{m_L,n_L,\alpha/2}/n_L},\\ m_L&=2t_0,\ n_L=2(n-t_0+1),\\ P_U&=\frac{m_U F_{m_U,n_U,1-\alpha/2}/n_U}{1+m_U F_{m_U,n_U,1-\alpha/2}/n_U},\\ m_U&=2(t_0+1),\ n_U=2(n-t_0). \end{aligned}$$

۵– مقایسه ی فاصله ی اطمینان دقیق و فاصله‌ی اطمینان والد

برای این منظور از شبیه سازی استفاده شده است. با استفاده از نرم افزار S-plus برای هر دو زوج (n,p) و برای مقادیر۰/۹، ۰/۸، ۰/۷، ۰/٦، ۰/٥، ۰/٤، ۰/۳، ۰/۲، ۰/۱ =p و مقادیر۱۰۰، ۵۰، ۳۰، ۱۰ =n به تعداد ۱۰۰۰۰ بار از توزیع دوجمله ای نمونه ی تصادفی ایجاد شده است. برای هر تکرار فاصله ی اطمینان ۹۵ درصدی از هر دو روش دقیق و روش والد محاسبه شده است. برای مقایسه ی دو روش از احتمال پوشش و متوسط طول فاصله اطمینان های محاسبه شده استفاده شده است. نتایج محاسبه در جدول ۱

داده شده اند. در این جدول مقادیر داخل پرانتز متوسط طول بازه ی اطمینان ها و اعداد خارج پرانتز مقدار احتمال پوشش هر حالت است.

جدول۱: احتمال پوشش و متوسط طول فاصله ی اطمینان در دو روش دقیق و والد

		n=۱۰	n=۳۰	n=۵۰	n=۱۰۰
p=۰/۱	والد	۰/۶۵(۰/۲۹)	۰/۸۰(۰/۲۰)	۰/۸۶(۰/۱۶)	۰/۹۲(۰/۱۲)
	دقیق	۰/۶۵(۰/۳۱)	۰/۹۵(۰/۲۳)	۰/۹۶(۰/۱۷)	۰/۹۶(۰/۱۲)
p=۰/۳	والد	۰/۸۴(۰/۵۳)	۰/۹۵(۰/۳۲)	۰/۹۳(۰/۲۵)	۰/۹۴(۰/۱۸)
	دقیق	۰/۹۷(۰/۵۵)	۰/۹۷(۰/۳۳)	۰/۹۷(۰/۲۶)	۰/۹۶(۰/۱۸)
p=۰/۵	والد	۰/۸۹(۰/۵۹)	۰/۹۵(۰/۳۵)	۰/۹۲(۰/۲۷)	۰/۹۵(۰/۲۰)
	دقیق	۰/۹۸(۰/۵۹)	۰/۹۶(۰/۳۶)	۰/۹۷(۰/۲۸)	۰/۹۷(۰/۲۰)
p=۰/۷	والد	۰/۸۴(۰/۵۳)	۰/۹۴(۰/۳۲)	۰/۹۳(۰/۲۵)	۰/۹۵(۰/۱۸)
	دقیق	۰/۹۷(۰/۵۵)	۰/۹۸(۰/۳۴)	۰/۹۷(۰/۲۶)	۰/۹۶(۰/۱۸)
p=۰/۹	والد	۰/۶۴(۰/۲۹)	۰/۸۰(۰/۲۰)	۰/۸۷(۰/۱۶)	۰/۹۳(۰/۱۲)
	دقیق	۰/۶۷(۰/۳۰)	۰/۹۶(۰/۲۳)	۰/۹۷(۰/۱۷)	۰/۹۶(۰/۱۲)

از این مثال می توان چند نتیجه مطلوب گرفت.

الف- در همه ی موارد محاسبه شده احتمال پوشش برای روش دقیق بزرگتر از روش والد است.

ب- طول بازه اطمینان در روش دقیق در بیشتر موارد بزرگتر از طول بازه اطمینان ها در روش تقریبی است.

۶- نتیجه گیری و آینده ی تحقیق

در این مقاله برای به دست آوردن فاصله ی اطمینان دقیق برای نسبت جامعه یک روش دقیق معرفی شد. این روش بر مبنای روش عمومی برای به دست آوردن فاصله ی اطمینان ها به دست آمد. با توجه به شبیه سازی نتیجه شد که علی رغم برقرار بودن شرایط استفاده از تقریب نرمال این روش منجر به جواب های بهتری نسبت به روش والد می شود.

برای توسیع و تعمیم نتایج حاصل از تحقیق می توان روش های تقریبی دیگری مانند روش ویلسون را نیز با روش دقیق مقایسه نمود.

منابع

[۱]. پارسیان، احمد. (۱۳۷۸). مبانی آمار ریاضی. انتشارات دانشگاه صنعتی اصفهان. چاپ اول.

[۲]. مشکانی، علی. (۱۳۷۷). مقدمه ای بر نظریه آمار. انتشارات دانشگاه فردوسی. چاپ اول.

[3]. Agresti, A. and Coull, B. A. (1998). Approximate is better than exact for interval estimation of binomial proportions. The American Statistician, 52, 119-126.

[4]. Hogg, R. V. and Tanis, E. A. (2001). Probability and statistical inference. (6th ed.). Upper Saddle River, NJ: Prentice Hall.

[5]. Wald, A., and Wolfowitz, J. (1939). Confidence limits for continious distribution functions. The Annals of Mathematical Statistics, 10, 105-118.

[6]. Wilson, E. B. (1927). Probable infere-nce, the law of succession, and statistical inference. Journal of the American Statistical Association, 22, 209-212