آزمون کای دو

(R)

تدوین: مرکز تحلیل آماری خوارزمی

***

www.kharazmi-statistics.ir

مرکز آماری خوارزمی

مقدمه:

این آزمون یکی از رایج ترین آزمون های ناپارامتریک است و زمانی از آن استفاده می شود که داده ها بصورت فراوانی باشند و آنها را بتوان بصورت دو یا چند طبقه تقسیم بندی کرد(مقیاس اسمی). آزمون مجذور کای بیش از سایر آزمون های ناپارامتریک در پژوهش های علوم انسانی بکار می رود.

مطالب ارائه شده در این متن عبارتند از:

معرفی آماره ی کای دو

توزیع کای دو

آزمون انطباق کای دو

آزمون های داده های مستقل کای دو

معرفی آزمون کای دو

آزمون کای دو یا خی دو، بر مبنای فراوانی (تعداد) مشاهده شده و فراوانی مورد انتظار به بررسی یک متغیر در جامعه می پردازد. و از آنجا که هدف بررسی یک متغیر است، از این آزمون برای بررسی فرضیه های توصیفی استفاده می شود.

پس در واقع هدف این آزمون پاسخگویی به این سوال است که آیا فراوانی یک صفت در جامعه طبق انتظار است یا خیر و به عبارتی دیگر، آیا بین فراوانی های مشاهده شده و فراوانی های مورد انتظار اختلاف معنی دار وجود دارد یا خیر.

فرض صفر بیان می کند که بین این فراوانی ها اختلاف معنی دار وجود ندارد. بنابراین: بین فراوانی های مشاهده شده و فراوانی های فرضی یا مورد انتظار تفاوت معنی دار وجود ندارد.

آماره آزمون یا شاخص آماری برای این فرضیه دارای توزیع کای دو با درجه آزادی (K-1) است و به صورت زیر محاسبه می شود.

علامت اختصاری درجه آزادی هم df است. $x^2 = \sum \frac{(O_i - E_i)^2}{E_i}$

که در آن، O_i فراوانی مشاهده شده و E_i فراوانی مورد انتظار است.

همچنین اگر کای دو از کای دوی جدول بزرگتر باشد فرض صفر رد می شود ولی اگر کای دو مشاهده شده از کایدو جدول یا به اصطلاح کای دوی بحرانی کوچکتر باشد بنابراین فرض صفر هم رد نمی شود.

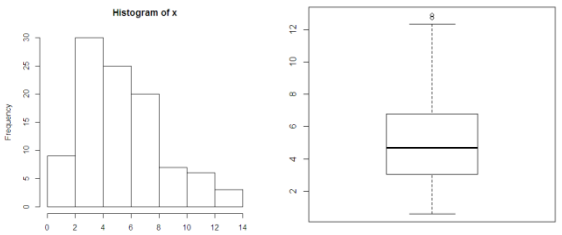
توزیع کای دو

توزیع χ^2 توزیع مجموع مربع متغیرهای تصادفی نرمال است . باید i.i.d نرمال (۱و۰) اعداد تصادفی باشد. پس χ^2 توزیع کای دو با n درجه آزادی دارد. $\chi^2 = \sum_{i=1}^{n} Z^2{}_i$

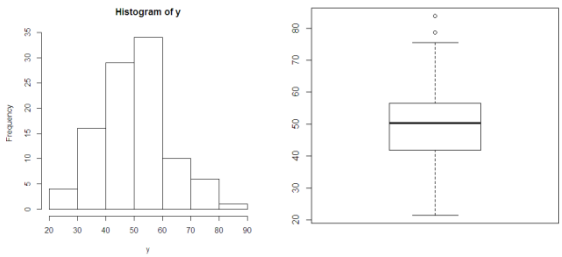
شکل توزیع به درجه آزادی بستگی دارد. در ادامه ۱۰۰ نمونه تصادفی با توزیع کای دو با درجه های آزادی ۵ و ۵۰ تولید کرده و نمودارهای آن های را نشان می دهیم تا تفاوت این دو حالت مشخص تر گردد.

برای تولید اعداد تصادفی با توزیع کای دو عبارت زیر را در نرم افزار بنویسید.

```
> x<-rchisq(100,5)
> y<-rchisq(100,50)
```



نمودار QQ نرمال و جعبه ای متغیر x



نمودار QQنرمال و جعبه ای متغیر y

نمودارهای جعبه ای و QQ نرمال هر یک از متغیرهای رسم شده است. برای آموزش چگونگی رسم نمودار در نرم افزار R به فایل آموزشی "رسم نمودار در نرم افزار R" در صفحه آموزش نرم افزار سایت مرکز تحلیل آماری خوارزمی مراجعه نمایید. همان طور که در نمودار های رسم شده مشاهده می شود برای متغیری که درجه ی آزادی کوچکتری دارد نمودارها نامتقارن هستند. و هر چقدر که عدد بزرگتر باشد توزیع به نظر نرمال می شود.

چند نکته ضروری در آزمون های خی دو:

۱. داده ها بصورت فراوانی و طبقه ای باشند و در مقیاس اسمی باشند.

۲. فراوانی هر مشاهده باید از سایر مشاهدات مستقل باشند.

۳. هر مشاهده باید فقط در یک طبقه قرار داده شود.

۴. فراوانی مورد انتظار ۸۰٪ خانه ها باید بیشتر از ۵ باشد و در صورتی که درجه آزادی یک باشد باید فراوانی مورد انتظار همه خانه ها از ۵ بیشتر باشد.

۵. همچنین، در مواقعی که حجم نمونه کم باشد و فراوانی مورد انتظار در بیشتر ۲۰٪ خانه ها کمتر از ۵ باشد، باید گروه ها را با هم ادغام کنیم که فراوانی مورد انتظار هر خانه بیشتر شود و در مواقعی که درجه آزادی یک باشد حتی یک خانه هم کمتر از ۵ هم نباید باشد.

آزمون انطباق کای دو

آزمون میزان انطباق برای آن است که ببینیم آیا داده ها از تعدادی جمعیت خاص برمی آیند؟ آزمون انطباق کای دو این امکان را فراهم می کند تا آزمایش کنیم آیا داده ها با مدل سازگاری دارند یا نه؟

برای این آزمون مثالی را در نظر بگیرید. فرض کنید تاسی داریم که آن را ۱۸۰ بار پرتاب کرده و هر بار یکی از اعداد تاس رو آمده باشد. (در صورتی که تاس سالم باشد انتظار داریم هر یک از اعداد ۳۰ بار ظاهر شود) تعداد دفعاتی که هر یک از اعداد رو آمده اند در جدول زیر است.

تعداد	عدد
۲۲	۱
۲۷	۲
۳۶	۳
۳۴	۴
۲۹	۵
۳۲	۶

در نرم افزار R باید احتمال ها و میزان مشاهدات هر یک از اعداد روی تاس را برای رسیدن به نتایج دلخواه تعریف کنیم. برای این منظور عبارت های زیر را در نرم افزار تایپ کنید.

۱) > obs<-c(22,27,36,34,29,32)

۲) > probs<-c(1,1,1,1,1,1)/6

۳) >chisq.test(obs,p=probs)

۱) تعریف تعداد مشاهدات هر یک از اعداد روی تاس در ۱۸۰ بار پرتاب

۲) تعریف احتمال مشاهده ی هر یک از اعداد تاس در حالت کاملا سالم بودن تاس

۳)تست انطباق کای دو

```
> obs<-c(22,27,36,34,29,32)
> probs<-c(1,1,1,1,1,1)/6
> chisq.test(obs,p=probs)

        Chi-squared test for given probabilities

data:  obs
X-squared = 4.3333, df = 5, p-value = 0.5025
```

همانطور که در نتیجه ی ارائه شده در روبرو مشاهده می کنید مقدار آماره ی کای دو ۴.۳۳۳۳ و درجه ی آزادی ۵ است. مقدار p-value محاسبه شده در این مثال ۰.۵۰۲۵ است که از مقدار ۰.۰۵ بزرگتر است در نتیجه دلیل بر رد فرض صفر وجود ندارد. در اینجا فرض صفر این است که آیا تاس، تاس خوبی است یا نه؟(آیا احتمال آمدن هر عدد برابر ۱/۶ است یا نه؟) نتیجتا دلیلی بر رد کردن فرضیه نسبی بر اینکه تاس خوبی است وجود ندارد و تاس مورد نظر، تاس خوبی است.

مثالی دیگر:

کیسه ای حاوی ۱۰۰ مهره به رنگ های آبی، قرمز، زرد و سبز داریم. از این مهره ها ۲۰ مهره به رنگ آبی، ۴۸ مهره به رنگ قرمز، ۱۲ مهره به رنگ زرد و ۲۰ مهره به رنگ سبز است و تمامی مهره ی هم اندازه و همگن هستند و تنها در رنگ متفاوت می باشد. ۲۵ انتخاب با جایگذاری از این کیسه انجام می دهیم. انتظار است که با احتمال ۲۰٪مهره آبی، ۴۸٪ مهره ی قرمز، ۱۲٪ مهره زرد و با احتمال ۲۰٪ مهره انتخابی سبز زنگ باشد. در واقع ۵ مهره آبی، ۱۲مهره قرمز، ۳ مهره زرد و ۵ مهره سبز. اما مشاهدات به صورت زیر است.

رنگ مهره	تعداد مشاهده انتظاری	تعداد مشاهدات واقعی
آبی	۵	**۴**
قرمز	۱۲	**۷**
زرد	۳	**۵**
سبز	۵	**۹**

میخواهیم به وسیله آزمون کای دو منطبق بودن این مهره ها را با انتظار و مدل انتظاری بسنجیم؟

فرض صفر: مهره ها با توجه به انتظار موجود مدل بندی شده اند.

عبارت زیر را در نرم افزار تایپ نمایید.

```
>a<-c(4,7,5,9)
>probs<-c(20,48,12,20)/100
>chisq.test(a,p=probs)
```

```
> a<-c(4,7,5,9)
> probs<-c(20,48,12,20)/100
> chisq.test(a,p=probs)

        Chi-squared test for given probabilities

data:  a
X-squared = 6.8167, df = 3, p-value = 0.07798

Warning message:
In chisq.test(a, p = probs) : Chi-squared approximation may be incorrect
```

با توجه به عدد بدست آمده برای p-value و بزرگتر بود این عدد نسبت به مقدار آلفا ۰.۰۵ در نتیجه دلیلی بر رد فرض وجود ندارد و فرض پذیرفته می شود.

در انتهای نتیجه نمایش داده شده پیام هشداری توسط نرم افزار بیان شده که ممکن است تقریب کای دو درست نباشد. این اخطار به واسطه ی آن است که باید احتمالات را شناسایی کرد. فرضیات آزمون کای دو به مستقل بودن نیاز دارند و باید این موضوع در مراحل آزمون در نظر گرفته شود و بررسی شود. ما در اینجا احتمالات را مستقل در نظر گرفته ایم.

آزمون های داده های مستقل کای دو

این آزمون زمانی کاربرد دارد که بخواهیم وابسته بودن و یا استقلال دو ردیف در یک جدول را با هم آزمون کنیم. یعنی فرض صفر آن است که این دو ردیف مستقل هستند و فرض مقابل آن است که آنها مستقل نیستند. به طور مثال در این بخش فرض کنید داده هایی از شدت سقوط هواپیما در اختیار است که برای مواردی که مسافران کمربند ایمنی داشتند و یا نه جدول زیر دسته بندی شده است.

میزان خسارت

		اصلا	حداقل	کمتر از	بیشتر از
کمربند ایمنی	بله	۱۲۸۱۳	۶۴۷	۳۵۹	۴۲
	خیر	۶۵۹۶۳	۴۰۰۰	۲۶۴۲	۳۰۳

آیا دو ردیف داده های مستقل هستند یا بستن کمربندها تفاوتی ایجاد کرده است؟ آماره کای دو یک خصیصه را ایجاد می کند. اما تعداد داده های مورد نظر(انتظار) کدامند؟

برای انجام تست مورد نظر عبارت زیر را در نرم افزار تایپ نمایید.

```
>yes<-c(12813,647,359,42)
>no<-c(65963,4000,2642,303)
>chisq.test(data.frame(yes,no))
```

```
> yes<-c(12813,647,359,42)
> no<-c(65963,4000,2642,303)
> chisq.test(data.frame(yes,no))

        Pearson's Chi-squared test

data:  data.frame(yes, no)
X-squared = 59.224, df = 3, p-value = 8.61e-13
```

با توجه به مقدار محاسبه شده برای مقدار آماره p-value و کوچکتر بودن این مقدار از مقدار آلفا ۰.۰۵ در نتیجه فرض صفر مورد نظر در سطح ۹۵٪ رد می شود. در نتیجه دو ردیف داده از هم مستقل نیستند.

دقت داشته باشید که تمامی این مراحل برای شرایطی که توزیع دو متغیر متفاوت باشد نیز برقرار است. در این شرایط فرض صفر اینگونه بیان میشود که: آیا ردیف هایی از توزیع های های مختلف هستند توزیعشان مناسب است؟

منبع:

- http://khi-2.loxblog.com/

- استفاده از R در آمار مقدماتی، مولف:JOHN VERZAN، ترجمه: صغری طاهرخانی، نجم الدین گنج خانلو، علی فصیحی