DOI: 10.48305/jims.v41.i732.0725

مجله دانشکده پزشکی اصفهان — سال چهل و یکم/شماره‌ی ۷۳۲/هفته‌ی چهارم مهر ۱۴۰۲

تاریخ دریافت: ۱۴۰۲/۳/۷ — تاریخ پذیرش: ۱۴۰۲/۶/۱۴ — تاریخ چاپ: ۱۴۰۲/۷/۲۸

مروری بر مفاهیم معنی‌داری آماری و بالینی با رویکرد آزمون فرضیه (p-value)

مهدی سپیدارکیش[۱]، زهرا محمدی پیروز[۲]

مقاله مروری

چکیده

کاربرد و تفسیر معنی‌داری آماری برای اثبات اثربخشی یک مداخله و یا وجود رابطه بین دو متغیر، یک اصل اساسی و ضروری در مطالعات است. بطور سنتی تجزیه و تحلیل داده‌های یک مطالعه با استفاده از آزمون فرضیه و گزارش p-value انجام می‌شود. در دو دهه‌ی اخیر متخصصان آمار و متدولوژی، محاسبه‌ی p-value و به ویژه استفاده از آستانه‌ی پنج درصد برای تأیید معنی‌داری آماری را نادرست می‌دانند. محدودیت‌های p-value مانند وابستگی مقدار آن به حجم نمونه و منعکس نکردن اهمیت بالینی به کرات اشاره شده است. متخصصان آمار و متدولوژی، گزارش تنهای p-value را کافی نمی‌دانند و گزارش شاخص اندازه‌ی اثر و حدود اطمینان بطور ملموسی تأکید شده است. با این حال مقالات متعددی به ویژه در مطالعات غیربالینی، به این امر توجه نکرده و حتی تفسیر صحیحی از p-value انجام نمی‌دهند. هدف نویسندگان این مقاله، ارائه‌ی یک دستورالعمل یکپارچه به پژوهشگران و متخصصین بالینی، به جهت گزارش صحیح معنی‌داری آماری و بالینی یافته‌ها بر اساس اهداف و طراحی مطالعات در علوم پزشکی با رویکرد آزمون فرضیه (گزارش p-value) بود.

واژگان کلیدی: معناداری آماری؛ معناداری بالینی؛ آزمون فرضیه؛ p-value

ارجاع: سپیدارکیش مهدی، محمدی پیروز زهرا. **مروری بر مفاهیم معنی‌داری آماری و بالینی با رویکرد آزمون فرضیه (p-value).** مجله دانشکده پزشکی اصفهان ۱۴۰۲؛ ۴۱ (۷۳۲): ۷۳۵-۷۲۵

معنی‌داری آماری با رویکرد آزمون فرضیه

پیش‌نیاز درک آسان مفهوم معنی‌داری آماری با رویکرد آزمون فرضیه (گزارش p-value) در این مقاله، دانستن کامل و عمیق مفهوم برآورد و حدود اطمینان است. بنابراین برای درک این مهم، الزامی است که در ابتدا مقاله‌ی چاپ شده در شماره ۷۳۱ / هفته سوم مهر۱۴۰۲ سال چهل و یکم با عنوان (مروری بر مفاهیم معنی‌داری آماری و بالینی: استفاده و تفسیر حدود اطمینان) را مطالعه بفرمایید. در مقاله‌ی مذکور توضیح داده شد، با بررسی حدود اطمینان (۹۵ درصد مطالعات) می‌توان معنی‌داری آماری و حتی بالینی را تأیید و یا رد نمود. اگر حدود اطمینان (۹۵ درصد مطالعات) برای شاخص اختلاف میانگین عدد صفر (در حجم نمونه‌ی مناسب و بدون وجود خطای منظم) را در بر بگیرد، نتایج غیرمعنی‌دار آماری هستند و در صورتی که عدد صفر را در بر نگیرد، نتایج معنی‌دار آماری هستند (۱، ۲). در رویکرد آزمون فرضیه که پایه‌ریزی آن با پیشتازان آمار Neyman و Pearson از سال‌های ۱۹۰۰ شروع شد، ابتدا در مورد پارامتر (در اینجا اختلاف میانگین جامعه/ $\mu_1 - \mu_2$) فرضیه می‌سازیم و در نهایت بر اساس اطلاعات نمونه (اختلاف میانگین نمونه)، فرضیه را رد یا قبول می‌کنیم (۳، ٤). معمولاً آزمون فرضیه در چندین مرحله انجام می‌شود، با این رویکرد می‌توان به راحتی مراحل را پیگیری کرد و فرضیه را قبول و یا رد نمود.

مراحل آزمون فرضیه

برای روشن شدن مفهوم معنی‌داری آماری با رویکرد آزمون فرضیه (گزارش p-value)، مراحل آزمون فرضیه را بر مبنای مثال ۱ بیان می‌کنیم.

مثال ۱: در یک مطالعه‌ی کارآزمایی بالینی تصادفی شده، پژوهشگر به جهت بررسی اثربخشی مکمل کورکومین بر وزن بیماران مبتلا به سرطان معده، آن‌ها را بطور تصادفی به دو گروه مکمل کورکومین (۱۰۰ نفر) و دارونما (۱۰۰ نفر) تقسیم کرد. میانگین و انحراف معیار وزن بعد از مداخله (۲٤ هفته)، در دو گروه مکمل

۱- استادیار، گروه آمار زیستی و اپیدمیولوژی، دانشکده‌ی بهداشت، دانشگاه علوم پزشکی بابل، بابل، ایران

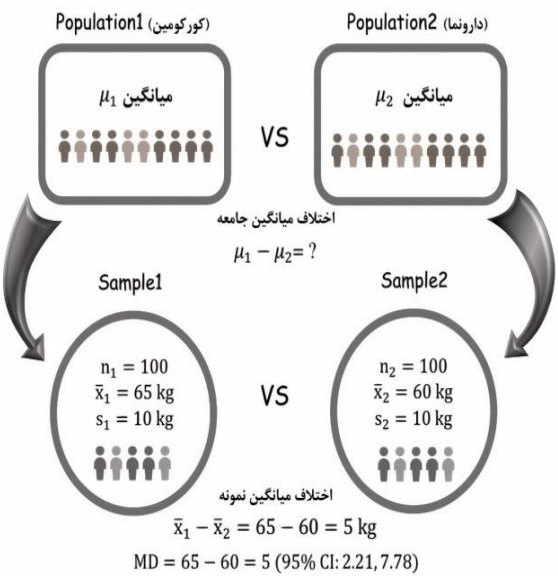
۲- دانشجوی کارشناسی ارشد، گروه آمار زیستی و اپیدمیولوژی، دانشکده‌ی بهداشت، دانشگاه علوم پزشکی بابل، بابل، ایران

نویسنده‌ی مسؤول: مهدی سپیدارکیش: استادیار، گروه آمار زیستی و اپیدمیولوژی، دانشکده‌ی بهداشت، دانشگاه علوم پزشکی بابل، بابل، ایران

Email: mahdi.sepidarkish@gmail.com

کورکومین و دارونما به ترتیب ۱۰ ± ۶۵ و ۱۰ ± ۶۰ کیلوگرم بدست آمد. اختلاف میانگین وزن بین دو گروه، ۵ کیلوگرم محاسبه شد. سؤال اینجاست که آیا با مشاهده‌ی اختلاف میانگین ۵ کیلوگرم بین دو گروه، می‌توان اثربخشی مکمل کورکومین را تأیید نمود؟ باید به این نکته توجه کرد که این اختلاف میانگین از مقایسه‌ی دو نمونه‌ی ۱۰۰ نفری بدست آمده است و این اختلاف ممکن است به دلیل خطای تصادفی یا خطای نمونه‌گیری باشد. تصور کنید اگر پژوهشگر مکمل کورکومین را به یک بیمار و دارونما را به یک بیمار دیگر می‌داد و وزن این دو نفر به نفع فرد مصرف‌کننده‌ی مکمل کورکومین، ۵ کیلوگرم افزایش می‌یافت، آیا همچنان می‌توانیم مکمل کورکومین را در افزایش وزن اثربخش بدانیم؟ بطور شهودی می‌توان دریافت که نتیجه‌ی حاصل ممکن است تصادفی باشد. در دو نمونه‌ی ۱۰۰ نفری نیز، همین حالت وجود دارد. اگر میانگین وزن افراد گروه کورکومین ۵ کیلوگرم افزایش یافته است، نمی‌توان به راحتی مکمل را اثربخش دانست و آن را برای همه‌ی بیماران مبتلا به سرطان معده توصیه کرد. همانطور که در شکل شماره‌ی ۱ نشان داده شده است، دو نمونه‌ی مورد مداخله، از دو جامعه‌ی بالا دست انتخاب شده‌اند. جامعه‌ی شماره‌ی یک شامل بیماران مبتلا به سرطان معده که مکمل کورکومین را مصرف می‌کنند و جامعه‌ی شماره‌ی دو شامل بیماران مبتلا به سرطان معده که مکمل دارونما مصرف می‌کنند. میانگین وزن جامعه‌ی شماره‌ی یک را با μ_1 و میانگین وزن جامعه‌ی شماره‌ی دو را با μ_2 نشان می‌دهیم. اختلاف میانگین این دو جامعه با $\mu_1 - \mu_2$ نشان داده شده است. در واقع صرفاً با دانستن اختلاف میانگین دو جامعه (پارامتر مجهول) می‌توان اثربخشی مکمل کورکومین را تأیید و یا رد کرد. فرض کنید تعداد کل بیماران مبتلا به سرطان معده در سرتاسر جهان یک میلیون نفر باشند و آن‌ها را به دو گروه ۵۰۰ هزار نفری تقسیم کنیم. یک گروه مکمل کورکومین و گروه دیگر دارونما مصرف کنند. میانگین را در هر گروه محاسبه و از هم کم کنیم. این شاخص اختلاف میانگین جامعه بوده و تنها با دانستن آن می‌توان اثربخشی دارو را تأیید و یا رد کرد (۵–۷).

آزمون فرضیه در چندین مرحله انجام می‌شود. در مرحله‌ی اول فرضیه را بر اساس پارامتر (در اینجا اختلاف میانگین وزن در دو جامعه)، بیان می‌کنیم. از منظر بالینی، می‌توان فرض نمود، که مداخله (مکمل کورکومین) اثر دارد و یا اینکه فرض می‌کنیم مداخله (مکمل کورکومین) اثر ندارد. به زبان آماری عدم اثربخشی مداخله (مکمل کورکومین) را با فرضیه‌ی صفر (H0) نشان می‌دهند و اثربخشی مداخله (مکمل کورکومین) را با فرضیه‌ی یک (H1) نمایش می‌دهند. فرضیه‌ی صفر را به اصطلاح فرضیه‌ی خنثی (Null hypothesis) و فرضیه‌ی یک را فرضیه‌ی پژوهشگر یا فرضیه‌ی جانشین (Alternative hypothesis) می‌نامند (۸، ۹).



شکل ۱. اختلاف میانگین معنادار دو گروه تحت درمان با کورکومین و دارونما با واریانس‌های برابر

به زبان آماری می‌توان فرضیات را مطابق شکل شماره‌ی ۲، به سه صورت نوشت. در حالت اول میانگین جامعه‌ی یک از میانگین جامعه‌ی دو بیشتر است ($\mu_1 > \mu_2$)، در حالت دوم میانگین جامعه‌ی یک از میانگین جامعه‌ی دو کمتر است ($\mu_1 < \mu_2$) و یا در حالت سوم میانگین جامعه‌ی یک با میانگین جامعه‌ی دو، متفاوت ($\mu_1 \neq \mu_2$) است. به دو فرضیه‌ی اول، اصطلاحاً فرضیات یک دامنه و به فرضیه‌ی آخر، فرضیه‌ی دو دامنه اطلاق می‌شود. بر اساس مثال ۱، فرضیات را می‌توان به سه حالت زیر بیان کرد. در حالت اول میانگین وزن کسانی که مکمل کورکومین مصرف می‌کنند بیشتر از کسانی است که دارونما مصرف می‌کنند، در حالت دوم، میانگین وزن کسانی که مکمل کورکومین مصرف می‌کنند، کمتر از کسانی است که دارونما مصرف می‌کنند و یا در حالت سوم، میانگین وزن کسانی که مکمل کورکومین مصرف می‌کنند با میانگین وزن کسانی که دارونما مصرف می‌کنند، متفاوت است. البته فرضیات بر اساس مرور متون، اهداف اصلی و اختصاصی مطالعه و جهت اثربخشی مداخله، نوشته می‌شوند. در مثال ۱ منطقی است که بعد از مداخله، میانگین وزن بیمارانی که مکمل کورکومین مصرف می‌کنند، بیشتر از بیمارانی باشد که دارونما مصرف می‌کنند و یا اینکه میانگین وزن بیمارانی که مکمل کورکومین مصرف می‌کنند با میانگین وزن بیمارانی که دارونما مصرف می‌کنند، متفاوت است. در ابتدا، مثال ۱ را با فرضیه‌ی یک دامنه ($\mu_1 > \mu_2$)، بررسی می‌کنیم. یعنی فرض می‌کنیم میانگین وزن کسانی که مکمل کورکومین مصرف می‌کنند بیشتر از کسانی است که دارونما مصرف می‌کنند (۵–۷).

$$H_0: \mu_1 > \mu_2 \quad VS \quad H_1: \mu_1 = \mu_2$$
$$H_0: \mu_2 > \mu_1 \quad VS \quad H_1: \mu_1 = \mu_2$$
$$H_0: \mu_1 \neq \mu_2 \quad VS \quad H_1: \mu_1 = \mu_2$$

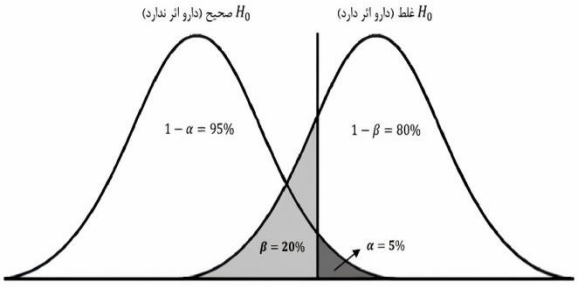
شکل ۲. فرضیات آماری

در مرحله‌ی دوم خطاهای آماری را تعیین می‌کنیم. به جهت درک ساده‌تر، خطاهای آماری، در شکل شماره‌ی ۳، نشان داده شده است.

نتایج مطالعه	واقعیت: H_0 صحیح (دارو اثر ندارد)	واقعیت: H_0 غلط (دارو اثر دارد)
H_0 رد (دارو اثر دارد)	تصمیم صحیح	خطای نوع اول (α)
H_0 قبول (دارو اثر ندارد)	خطای نوع دوم (β)	تصمیم صحیح

شکل ۳. انواع خطاها

در واقعیت یا مداخله (در اینجا مکمل کورکومین) اثر دارد و یا ندارد (به عناوین ستون‌ها نگاه کنید). بر اساس نتایج مطالعه نتیجه می‌گیریم یا مداخله اثر دارد یا ندارد (به عناوین ردیف‌ها نگاه کنید). در دو حالت، درست نتیجه می‌گیریم، جایی که در واقعیت مداخله اثر دارد و نتایج مطالعه نیز اثربخشی مداخله را نشان می‌دهد و همچنین در جایی که در واقعیت مداخله اثر ندارد و نتایج مطالعه اثربخشی مداخله را نشان نمی‌دهد. به زبان آماری اگر فرضیه‌ی صفر (H0) (عدم اثربخشی مداخله) در واقعیت غلط باشد و ما فرضیه‌ی صفر (H0) را رد کنیم، به درستی نتیجه گرفتیم. اگر فرضیه‌ی صفر (H0) در واقعیت صحیح باشد ما آن را رد نکنیم، باز هم به درستی نتیجه‌گیری کرده‌ایم. در دو حالت ما اشتباه نتیجه‌گیری می‌کنیم. اگر مداخله در واقعیت اثر نداشته باشد و به غلط نتیجه بگیریم، مداخله اثر دارد، مرتکب خطای نوع یک یا آلفا (Type one error) α شده‌ایم. از نظر بالینی، مفهوم خطای نوع یک به این‌صورت است که دارو در واقعیت اثر نداشته باشد و ما به اشتباه دارو را اثربخش بدانیم. مثال خطای نوع یک مانند این است که فردی مبتلا به ایدز باشد و ما به قصد درمان به او پشمک (بله درست فهمیدید پشمک!) بدهیم. در حالت دوم، اگر مداخله در واقعیت اثر داشته باشد و ما به غلط نتیجه‌گیری کنیم دارو اثر ندارد، مرتکب خطای نوع دوم یا بتا (Type two error) β شده‌ایم. از نظر بالینی معنای خطای نوع دوم به اینصورت است که دارو در واقعیت اثر دارد و ما به اشتباه دارو را بی‌اثر بدانیم و بیمار را از دارو محروم کنیم. مثال خطای نوع دوم مانند این است که فرد مبتلا به دیابت را از داروی اثربخش متفورمین (متفورمین در واقعیت بر کاهش قند خون اثر دارد) محروم کنیم. رابطه‌ی خطای نوع یک و دو مانند دو پدال قایق پدالی است، یعنی با کاهش یکی، دیگری افزایش می‌یابد و یا برعکس (شکل ٤). از این دو خطا، خطای نوع یک موذی‌تر و خطرناکتر است. با رجوع به متن، متوجه می‌شویم که اگر فرد مبتلا به ایدز به اشتباه پشمک بخورد (و داروی دیگر مصرف نکند)، بسیار خطرناک‌تر از حالتی است که فرد مبتلا به دیابت از متفورمین محروم شود، چون در این حالت به سراغ داروی دیگر می‌رود. معمولاً خطای نوع یک را در سطح ۵ درصد (البته بسته به شرایط ۱ تا ۱۰ درصد انتخاب می‌شود) و خطای نوع دوم را در سطح ۲۰ درصد در نظر می‌گیرند. در مثال ۱، احتمال اینکه مکمل کورکومین در واقعیت اثر نداشته باشد ولی به اشتباه آن را اثربخش بدانیم، خطای نوع یک دانسته و مقدار آن را ۵ درصد در نظر می‌گیریم. از طرف دیگر احتمال اینکه مکمل کورکومین در واقعیت اثر داشته باشد ولی به اشتباه آن را بی‌اثر بدانیم، خطای نوع دوم دانسته و مقدار آن را ۲۰ درصد در نظر می‌گیریم. از آنجایی که دامنه‌ی حرکتی احتمال بین صفر تا یک است، با در نظر گرفتن خطای نوع دوم در سطح ۲۰ درصد، احتمال متمم آن را ۸۰ درصد در نظر می‌گیریم. این احتمال را توان مطالعه تعریف می‌کنند. به مفهوم دیگر توان مطالعه را «احتمال دیدن اثربخشی مداخله، زمانی‌که در واقعیت مداخله اثربخش باشد» تعریف می‌کنند. در مثال یک، اگر در واقعیت مکمل کورکومین اثربخش باشد، ما بتوانیم در ۸۰ درصد مواقع آن را بدرستی تشخیص دهیم. از نظر آماری توان مطالعه را «احتمال رد فرضیه صفر (H0) غلط» تعریف می‌کنند (۱۰-۱۲). توضیح مبسوط مفهوم توان (با ذکر مثال) در انتهای مقاله آمده است.

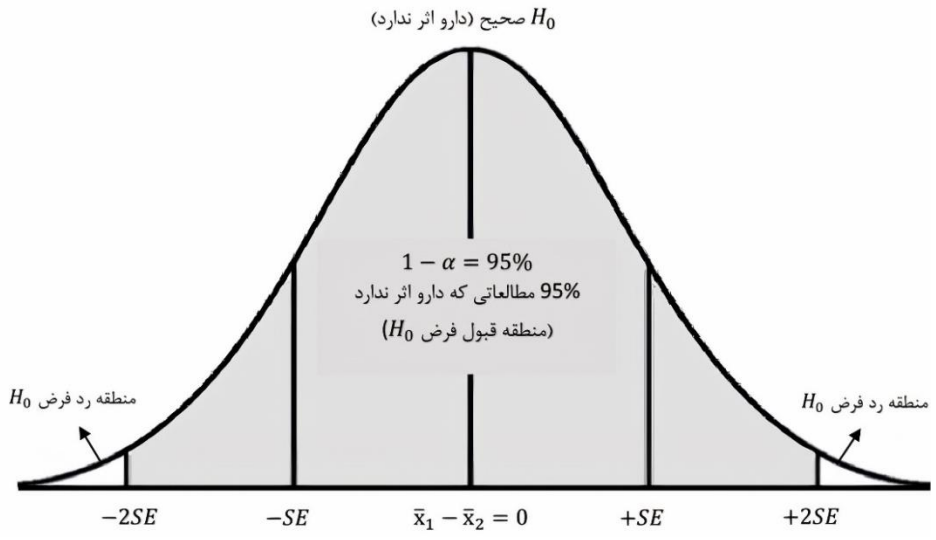


شکل ٤. رابطه خطای نوع اول و خطای نوع دوم

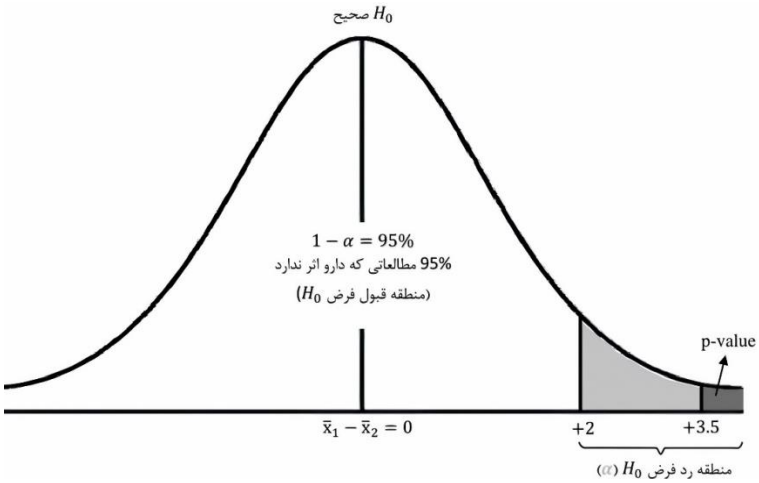
در مرحله‌ی سوم از اطلاعات نمونه برای قبول و یا رد فرضیه‌ی صفر (H0) استفاده می‌کنیم. به مفهوم دیگر از اختلاف میانگین نمونه

برای قبول و یا رد فرضیه‌ی صفر (H0) استفاده می‌کنیم. اما در نظر بگیرید که در مثال ۱، آیا می‌توان بر اساس اختلاف میانگین ۵ کیلوگرم به نفع گروه مکمل کورکومین، آن را در واقعیت اثربخش دانست؟ بر اساس اختلاف میانگین یک مطالعه (نمونه) نمی‌توان به راحتی فرضیه‌ی صفر (H0) را رد کرد. لذا باید شاخص اختلاف میانگین نمونه را با خطای نمونه‌گیری تلفیق کرده و به مفهوم دیگر با در نظر گرفتن خطای تصادفی، فرضیه‌ی صفر (H0) را رد کرد. در آمار به شاخص ترکیبی اندازه‌ی اثر (Effect size) (اختلاف میانگین نمونه) و خطای تصادفی (خطای معیار در اینجا جانشین خطای تصادفی است)، آماره‌ی آزمون (Statistic) می‌گویند (۱۳). بنابراین برای رد و یا قبول کردن فرضیه‌ی صفر (H0)، باید آماره‌ی آزمون را محاسبه کرد. در فرمول $T = \frac{\bar{X}_1 - \bar{X}_2}{S.E}$، حرف T نشانگر آماره‌ی آزمون است. صورت کسر آماره‌ی آزمون، اختلاف میانگین مطالعه (تنها مطالعه‌ی واقعی یا نمونه) و مخرج کسر، خطای معیار است. در واقع هر چه اختلاف میانگین مطالعه (تنها مطالعه‌ی واقعی یا نمونه) بزرگ و یا حجم نمونه زیاد (خطای معیار کوچک شود) شود، آماره‌ی آزمون بزرگ می‌شود و با بزرگ شدن آماره‌ی آزمون، احتمال رد فرضیه‌ی صفر (H0)، افزایش می‌یابد (۱۴، ۱۵).

همانطور که در مقاله‌ی سری آمار و متدولوژی ۱ با عنوان «مروری بر مفاهیم معنی‌داری آماری و بالینی با رویکرد برآورد (فاصله‌ی اطمینان)» توضیح داده شده است، خطای معیار متوسط اختلاف هر مطالعه (اختلاف میانگین هر مطالعه) از میانگین مرکزی (میانگین اختلاف میانگین‌های نمونه‌ای) است. با توجه به شکل ۵، با اضافه و کم کردن دو خطای معیار به اختلاف میانگین مطالعه‌ی واقعی (در عمل یک مطالعه انجام می‌شود)، می‌توان بیان نمود که ۹۵ درصد مطالعات، اختلاف میانگین جامعه را در برگرفته‌اند. در عمل، پژوهشگر فقط یک مطالعه انجام می‌دهد و در واقع ۱۰۰ مطالعه، از روی همان یک مطالعه‌ی واقعی شبیه‌سازی می‌شوند. خطای معیار بر اساس متوسط اختلاف هر اختلاف میانگین، از اختلاف میانگین مرکزی (میانگین اختلاف میانگین‌ها) بدست می‌آید. بنابراین، خطای معیار از تکرار بر روی تنها مطالعه‌ی انجام شده در واقعیت، ساخته می‌شود. در صورتی که حجم نمونه‌ی تنها مطالعه‌ی انجام شده زیاد باشد، خطای معیار کوچک و حدود اطمینان باریک (Narrow) می‌شود (۱۶–۱۸). تصور کنید که مطالعه‌ی کارآزمایی مثال ۱، بر روی یک میلیون بیمار مبتلا به سرطان معده انجام شود، در هر گروه ۵۰۰ هزار بیمار حضور داشته باشد و به ترتیب به مکمل کورکومین و دارونما تخصیص داده شوند. از آن‌جا که برای محاسبه‌ی خطای معیار و ساختن حدود اطمینان برای اختلاف میانگین جامعه، ۱۰۰ مطالعه از روی مطالعه‌ی واقعی (کارآزمایی ۱ میلیونی) ساخته می‌شود، خطای معیار در کم‌ترین مقدار ممکن خواهد بود. با صفر شدن آن، اختلاف میانگین مطالعه‌ی واقعی (کارآزمایی ۱ میلیونی)، با اختلاف میانگین جامعه یکی می‌شود. در حالتی که خطای معیار صفر شده باشد، آماره‌ی آزمون صرفاً بر اساس اختلاف میانگین مطالعه‌ی واقعی ساخته می‌شود، بنابراین می‌توان بر اساس اختلاف میانگین مطالعه‌ی واقعی فرضیه‌ی صفر را رد کرد. بر اساس مثال ۱ که اختلاف میانگین مطالعه ۵ کیلوگرم بدست آمد، اگر خطای معیار یک باشد، آماره‌ی آزمون برابر ۵ می‌باشد و بنابراین براحتی می‌توان با اختلاف میانگین نمونه‌ی فرضیه‌ی صفر را رد کرد.



شکل ۵. منطقه رد فرضیه صفر



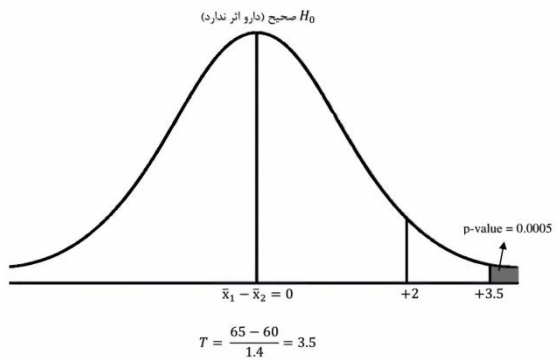
شکل ٦. توزیع اختلاف میانگین نمونه‌ای زمانی‌که دارو اثر ندارد

در مرحله‌ی چهارم، باید آماره‌ی آزمون را با معیاری استاندارد (در اینجا عدم اثربخشی دارو در واقعیت/ اختلاف میانگین دو جامعه صفر باشد / $\mu_1 - \mu_2 = 0$)، مقایسه کرد. از آن‌جا که توزیع اختلاف میانگین جامعه را زمانی‌که صفر باشد (عدم اثربخشی دارو در واقعیت) در اختیار نداریم، از جانشین آن یعنی ۹۵ درصد مطالعاتی که در آن‌ها دارو اثر ندارد (حدود اطمینان شامل صفر است)، استفاده می‌کنیم. در شکل ٦، توزیع اختلاف میانگین نمونه‌ای، زمانی که دارو اثر ندارد (اختلاف میانگین نمونه‌ای صفر باشد)، نشان داده شده است. در توزیع نمونه‌ای، زمانی‌که دارو اثر ندارد (اختلاف میانگین نمونه‌ای صفر باشد)، به جهت تفهیم ساده‌تر، دو منطقه‌ی قبول و رد فرضیه‌ی صفر (H0) را مشخص کرده‌ایم. بنابراین آماره‌ی آزمون، با توزیع نمونه‌ای (زمانی‌که دارو اثر ندارد/ اختلاف میانگین نمونه‌ای صفر باشد)، مقایسه می‌شود. همانطور که در شکل ٦ نشان داده شده است، اگر آماره‌ی آزمون خارج از منطقه قبول توزیع نمونه‌ای قرار بگیرد، می‌توان فرضیه‌ی صفر (H0) را رد کرد و اگر در داخل منطقه قبول توزیع نمونه‌ای قرار بگیرد، فرضیه‌ی صفر (H0) را قبول می‌کنیم. در ادامه‌ی مثال یک، آماره‌ی آزمون در فرمول $T = \frac{65-60}{1.4} = 3.5$، محاسبه شده است. اختلاف میانگین مطالعه، ۵ و خطای معیار ١/٤، محاسبه شده است و از تقسیم اختلاف میانگین مطالعه (۵) به خطای معیار (١/٤)، آماره‌ی آزمون ۳/۵ بدست آمد. همانطور که نشان داده شده است، آماره‌ی آزمون (۳/۵)، در منطقه‌ی رد فرض صفر (بزرگتر از عدد ۲)، قرار گرفته است، بنابراین فرضیه‌ی صفر (H0)، در سطح معنی‌داری ۵ درصد، رد می‌شود. از نظر بالینی می‌توان بیان نمود که اطلاعات منتج از مطالعه، با توزیع اختلاف میانگین نمونه‌ای (زمانی‌که دارو اثر ندارد/ اختلاف میانگین نمونه‌ای صفر باشد)، همخوانی نداشته و بنابراین دارو علاوه بر نمونه، در واقعیت (جامعه) نیز اثربخشی دارد (۵، ۱۹).

مفهوم p-value

بعد از آن‌که مراحل آزمون فرضیه انجام شد و فرضیه‌ی صفر (H0) را رد کردیم، این سؤال مطرح می‌شود که آیا ممکن است، هنوز اشتباه کرده باشیم و در واقعیت مداخله (دارو/ در اینجا مکمل کورکومین) اثر ندارد؟ این احتمال را که ممکن است مداخله در واقعیت اثر نداشته باشد و ما به غلط نتیجه بگیریم، مداخله اثر دارد خطای نوع یک یا آلفا تعریف کردیم و مقدار آن را در ابتدای مطالعه و قبل از جمع‌آوری داده‌ها ثابت و ۵ درصد در نظر گرفتیم. از نظر بالینی، معنای خطای یک به این مفهوم است که دارو در واقعیت اثر نداشته باشد و ما به اشتباه دارو را اثربخش بدانیم. شاخص p-value، از جنس خطای نوع یک یا آلفا است، اما خود خطای نوع یک نیست. شاخص p-value، مقدار احتمال خطای نوع یک، بعد از جمع‌آوری داده‌ها و دیدن آماره‌ی آزمون است. همانطور که در شکل شماره‌ی ۷ نشان داده شده است، p-value (که به رنگ خاکستری تیره نشان داده شده است)، احتمال رخداد خطای نوع یک بعد از محاسبه‌ی آماره است. به این مفهوم بعد از اینکه فرضیه‌ی صفر (H0) را رد کردیم (دارو را اثربخش دانستیم)، آیا ممکن است هنوز اشتباه کرده باشیم و در واقعیت دارو اثر نداشته باشد. بر اساس مثال شماره‌ی ۱، بعد از آن‌که مکمل کورکومین را اثربخش دانستیم، آیا ممکن است به اشتباه نتیجه‌گیری کرده‌ایم که مکمل کورکومین اثر دارد در حالی‌که در واقعیت اثر ندارد؟ مقدار این اشتباه (مقدار رخداد خطای نوع یک بعد از آنالیز داده‌ها و محاسبه‌ی آماره) را باید دقیقاً محاسبه کرده و تا سه رقم اعشار گزارش داد. همانطور که در شکل ۷ نشان داده شده است، مقدار دقیق این احتمال ۰/۰۰۰۵ است. هر چقدر مقدار p-value کوچک‌تر باشد، می‌توان نتیجه گرفت که اشتباه نکرده‌ایم و دارو در واقعیت اثر دارد. سؤال بعدی این است که چه مقدار رخداد p-value مجاز است؟ از آن‌جایی که کوچک بودن p-value ذهنی بوده و مقدار

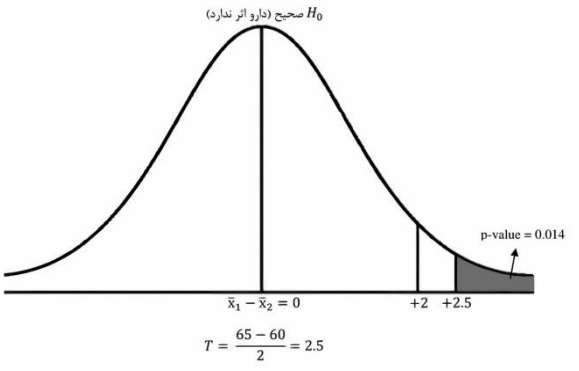
آن را باید با یک معیار استاندارد مقایسه کرد و همچنین خطای نوع یک قراردادی و قبل از انجام مطالعه، ۵ درصد تعیین شده است، ما مقدار p-value را با آن مقایسه می‌کنیم. به مفهوم دیگر حداکثر مجاز خطای نوع یک، ۵ درصد بوده و ما مقدار رخ داده‌ی آن را در عمل محاسبه و گزارش می‌دهیم. بر اساس شیوه‌ی نگارش علمی انجمن روانشناسان آمریکا (American Psychological Association) APA، مقدار p-value باید با p کوچک نوشته شده و باید تا سه رقم اعشار گزارش شود. احتمال کوچک‌تر از ۰/۰۰۱ را بصورت <۰/۰۰۱ می‌نویسند (۲۰-۲۲).



شکل ۷. محاسبه برای p-value مثال یک

آیا مقادیر کوچک‌تر p-value، به مفهوم معنی‌داری بالینی است؟

همانطور که در شکل ۷ نشان داده شده است، مقدار کوچک p-value زمانی بدست می‌آید که آماره‌ی آزمون بزرگ شود و همانطور که توضیح داده شد آماره‌ی آزمون در دو صورت بزرگ می‌شود: ۱) اختلاف میانگین نمونه (صورت کسر آماره‌ی آزمون) بزرگ باشد، ۲) خطای معیار (مخرج کسر آماره‌ی آزمون) کوچک شود و یا به مفهوم دیگر حجم نمونه بزرگ شود. همانطور که در شکل ۷ و ۸ نشان داده شده است، دو مطالعه‌ی جداگانه با حجم نمونه‌ی متفاوت نشان داده شده است (شکل شماره‌ی ۷: حجم نمونه ۲۰۰ نفر (۱۰۰ نفر در هر گروه)؛ شکل شماره‌ی ۸: حجم نمونه ۱۰۰ نفر (۵۰ نفر در هر گروه)). در هر دو مطالعه اثربخشی مکمل کورکومین (صورت آماره‌ی آزمون/ اختلاف میانگین نمونه) ثابت نگه داشته شده. در هر دو مطالعه اختلاف میانگین ۵ کیلوگرم است. در مطالعه‌ی با حجم نمونه‌ی ۲۰۰ نفر (هر گروه ۱۰۰ نفر)، خطای معیار ۱/۴۱ و در مطالعه‌ی با حجم نمونه‌ی ۱۰۰ نفر (هر گروه ۵۰ نفر)، خطای معیار ۲ بدست آمد. همچنین مقدار p-value در دو مطالعه به ترتیب ۰/۰۰۰۵ و ۰/۰۱۴۱ محاسبه شد. بنابراین هر چند که اثربخشی بالینی دارو (صورت آماره‌ی آزمون/ اختلاف میانگین نمونه) در هر دو مطالعه یکسان است اما به دلیل حجم نمونه‌ی بیشتر در مطالعه‌ی اول (شکل شماره‌ی ۷)، آماره‌ی آزمون بزرگتر شده و مقدار p-value کوچک‌تری دارد. بنابراین مقادیر پایین‌تر p-value به مفهوم معنای‌داری بالینی بیشتر نیست. چون مشخص نیست کوچک‌تر بودن p-value به دلیل افزایش حجم نمونه است یا به دلیل اثربخشی بالینی مداخله، ما مقدار آن را صرفاً برای معنی‌داری آماری گزارش و تفسیر می‌کنیم (۲۳).



شکل ۸. محاسبه p-value برای مثال یک با حجم نمونه ۱۰۰ (دو گروه ۵۰ نفری)

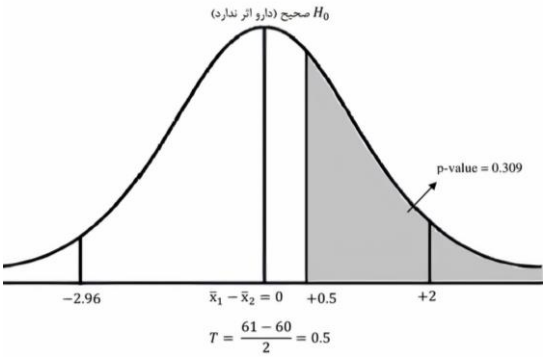
مقایسه‌ی شاخص حدود اطمینان و p-value

همانطور که در مقاله‌ی سری آمار و متدولوژی ۱ با عنوان «مروری بر مفاهیم معنی‌داری آماری و بالینی با رویکرد برآورد (فاصله‌ی اطمینان)» توضیح داده شد، شاخص حدود اطمینان علاوه بر معنی‌داری آماری، معنی‌داری بالینی را نیز نشان می‌داد در حالی‌که شاخص p-value صرفاً معنی‌داری آماری را مشخص می‌کند. توصیه‌ی اکید متخصصین آمار و متدولوژی، گزارش شاخص اندازه‌ی اثر همراه با حدود اطمینان است. شاخص حدود اطمینان برای پژوهشگران در مطالعات آتی نیز مهم است. بطور مثال به جهت محاسبه‌ی حجم نمونه در مطالعات آتی، پژوهشگران نیاز به شاخص اندازه‌ی اثر و حدود اطمینان دارند. همچنین در مطالعات مروری ساختارمند، پژوهشگران بطور معمول بر اساس شاخص‌های اندازه‌ی اثر مطالعات اولیه متاآنالیز انجام می‌دهند و انجام متاآنالیز بر اساس p-value، پیچیدگی‌های خاص خودش را داشته و محدودیت‌های متعددی دارد و مرسوم نیست (۲۴، ۲۵).

آیا امکان دارد که حدود اطمینان از نظر آماری معنی‌دار باشد در صورتی‌که p-value از نظر آماری معنی‌دار نباشد؟

این امر امکان‌پذیر نیست و در صورتی که حدود اطمینان معنی‌دار آماری باشد (در مثال یک عدد صفر را در بر بگیرد)، قطعاً p-value کوچک‌تر از ۵ درصد شده است. به مفهوم دیگر این دو شاخص،

معنی‌داری آماری را با دو رویکرد متفاوت بررسی می‌کنند. همانطور که در شکل ۹، نشان داده شده است، p-value در صورتی بزرگتر از ۵ درصد می‌شود (معنی‌دار آماری نباشد)، که آماره‌ی آزمون در منطقه‌ی قبول فرضیه‌ی صفر (H0) قرار بگیرد. در شکل ۹، میانگین و انحراف معیار وزن در گروه بیماران مبتلا به سرطان معده، که مکمل کورکومین دریافت کرده‌اند، معادل ۱۰ ± ۶۱ و در گروه بیماران مبتلا به سرطان معده، که دارونما دریافت کرده‌اند، معادل ۱۰ ± ۶۰ محاسبه شده است. اختلاف میانگین معادل ۱ و خطای معیار ۲ بدست آمده است. آماره‌ی آزمون با تقسیم عدد ۱ به خطای معیار ۲، معادل ۰/۵ محاسبه شد. آماره‌ی آزمون (۰/۵)، در منطقه‌ی قبول فرضیه‌ی صفر (H0) قرار گرفت و مقدار p-value معادل ۰/۳۰۹ محاسبه گردید. از آنجایی‌که آماره‌ی آزمون در منطقه‌ی رد فرضیه‌ی صفر (H0) قرار نگرفته است، قطعاً اطلاعات نمونه (آماره‌ی آزمون) با مجموعه مطالعاتی (حدود اطمینان ۹۵ درصد) که دارو اثر ندارد، شباهت دارد. در این حالت اگر دقت بفرمایید اختلاف میانگین و حدود اطمینان گزارش شده معادل ۱ و (۲/۹۶- تا ٤/۹٦) بدست آمد. بنابراین حدود اطمینان و p-value کاملاً به یک مفهوم یکسان اشاره می‌کنند و معنی‌داری آماری را بطور یکسان گزارش می‌کنند (۲٤، ۲٦).



شکل ۹. نتایج یکسان حدود اطمینان و p-value

توان مطالعه (Power of study)

همانطور که تعریف شد خطای نوع دوم (یا بتا)، قبول فرض صفر غلط است. به زبان بالینی زمانی که در واقعیت دارو اثر داشته باشد (یا رابطه‌ی بین متغیر مستقل و وابسته وجود داشته باشد)، ما بر اساس نتایج مطالعه‌ی خود (نمونه)، به اشتباه نتیجه بگیریم دارو اثر ندارد (یا رابطه‌ی بین متغیر مستقل و وابسته وجود ندارد). از منظر آماری، توان مطالعه، رد فرضیه‌ی صفر (H0) غلط است. به زبان بالینی یعنی اگر در واقعیت دارو اثر داشته باشد (یا رابطه‌ی بین متغیر مستقل و وابسته وجود داشته باشد)، ما بتوانیم بر اساس نتایج مطالعه‌ی خود این اثر (یا رابطه) را ببینیم. توان مطالعه رابطه‌ی مستقیمی با حجم نمونه دارد. به مفهوم دیگر مطالعه‌ی ما چقدر توان دارد (چقدر حجم نمونه دارد)، تا بتواند اثر دارو (در صورتی که در واقعیت دارو اثر دارد) را کشف کند. البته توان مطالعه نه تنها بر اساس حجم نمونه بلکه با توجه به اثربخشی (اندازه‌ی اثر دارو/ یا اندازه‌ی اثر رابطه‌ی بین متغیر مستقل و وابسته) نیز تعیین می‌شود. اثربخشی (اندازه‌ی اثر) بر اساس مطالعات قبلی و تجربیات متخصص بالینی تعیین می‌شود. بر اساس مثال ۱، متخصص تغذیه در سرطان، اثربخشی مکمل کورکومین را چه مقدار تصور می‌کند و یا بر اساس مطالعات قبلی چه مقدار برای مکمل کورکومین اثر دیده شده است. به زبان ساده‌تر مصرف مکمل کورکومین در مقایسه با دارونما بطور متوسط چند کیلوگرم وزن بیماران مبتلا به سرطان معده را افزایش می‌دهد. بنابراین متخصص آمار و یا متدلوژی بر اساس اثربخشی مفروض (مثلاً ۵ کیلوگرم) حجم نمونه را با توان ثابت در نظر می‌گیرد. بطور معمول خطای نوع دوم (یا بتا)، ۲۰ درصد مفرض می‌شود و بر این اساس توان مطالعه ۸۰ درصد تلقی می‌شود. پس اگر در واقعیت مکمل کورکومین در مقایسه با دارونما بتواند بطور متوسط ۵ کیلوگرم وزن بیماران مبتلا به سرطان معده را افزایش دهد، مطالعه‌ی ما (یافته‌های حاصل از نمونه) می‌تواند این اثربخشی را در ۸۰ درصد از مواقع تشخیص دهد و در ۲۰ درصد از مواقع (خطای نوع دوم)، نمی‌توانیم اثربخشی را ببینیم و به اشتباه فکر می‌کنیم دارو اثر ندارد، در حالی که در واقعیت دارد. هر چقدر اثربخشی مفروض (بطور واقعی و بر اساس یافته‌های مطالعات قبلی) بیشتر تلقی شود، ما حجم نمونه‌ی کمتری را نیاز داریم و با حجم نمونه‌ی کمتر در ۸۰ درصد از مواقع می‌توان اثربخشی را پیدا کرد.

برای اینکه این مفهوم را بهتر توضیح دهیم به شبیه‌سازی زیر توجه فرمایید.

فرض کنید که در یک اتاق، یک شیء پنهان شده است و ما سعی داریم که در یک زمان مشخص این شیء را پیدا کنیم. در این مثال شیء پنهان شده را اثربخشی دارو (اندازه‌ی اثر) و مدت زمان جستجو را توان مطالعه فرض کنید. دو حالت مفروض است: ۱) بعد از مدت زمان جستجوی مشخص، شیء را پیدا می‌کنیم و یا ۲) بعد از مدت زمان جستجوی مشخص شیء را پیدا نمی‌کنیم. در حالت اول که شیء را پیدا کرده‌ایم یعنی مدت زمان جستجو (توان مطالعه) برای پیدا کردن شیء پنهان شده (اثربخشی دارو/ وجود رابطه بین متغیر مستقل و وابسته) کافی بوده است. اما در حالت دوم که شیء را پیدا نکرده‌ایم، دو امکان وجود دارد: ۱) یا شیء وجود نداشته (در واقعیت دارو اثر نداشته و یا رابطه‌ای بین متغیر مستقل و وابسته وجود ندارد) و مدت زمان جستجو (توان مطالعه/ حجم نمونه) کافی بوده و یا ۲) شیء وجود داشته (در واقعیت دارو اثر داشته/ رابطه‌ای بین متغیر مستقل و وابسته وجود دارد) و ما مدت زمان کافی (توان مطالعه/

حجم نمونه) برای پیدا کردن شیء صرف نکرده‌ایم. برای عینیت بخشیدن به مثال بالا تصور کنید که شیء مخفی شده فرش کف اتاق (اندازه‌ی اثر بزرگ) باشد. چقدر زمان جستجو (توان مطالعه/ حجم نمونه) نیاز است تا فرش را پیدا کنیم. قطعاً در یک زمان کوتاه (حجم نمونه‌ی کم) ما فرش را پیدا می‌کنیم. حال فرض کنیم شیء مخفی شده در اتاق یک سنجاق ته گرد (اندازه‌ی اثر کوچک) باشد. چقدر زمان (توان مطالعه/ حجم نمونه) نیاز است تا سنجاق ته گرد را پیدا کنیم. مطمئناً باید زمان بسیار زیادی (حجم نمونه‌ی بالا/ توان بالا) برای پیدا کردن آن صرف کنیم. حال اگر ما با یک زمان جستجوی کوتاه (حجم نمونه‌ی پایین/ توان پایین) مدعی شویم که سنجاق ته گرد در اتاق وجود ندارد، مطمئناً ما اشتباه کرده‌ایم، سنجاق ته گرد (اثربخشی دارو/رابطه‌ی بین متغیر مستقل و وابسته) وجود دارد و ما زمان کافی (حجم نمونه‌ی لازم/ توان کافی) برای پیدا کردن آن تخصیص نداده‌ایم. حال فرض کنید که ما ۷ شبانه روز در اتاق بگردیم. در این حالت می‌توانیم مطمئن باشیم که زمان کافی (توان کافی) گذاشته‌ایم و به احتمال بسیار زیاد (بالای ۸۰ درصد) سنجاق ته گرد وجود ندارد.

معمولاً در مطالعاتی که نتایج غیرمعنی‌دار آماری (عدم دیدن اثر دارو/ عدم دیدن رابطه بین متغیر مستقل و وابسته) است، داوران درخواست محاسبه‌ی مجدد توان مطالعه را می‌دهند. یعنی می‌خواهند مطمئن شوند که آیا در صورت وجود اثربخشی دارو/ وجود رابطه بین متغیر مستقل و وابسته، توان مطالعه همچنان ۸۰ درصد باقی مانده است (مدت زمان کافی برای جستجو صرف شده) و یا اینکه توان مطالعه کافی نبوده (مدت زمان کافی برای جستجو تخصیص داده نشد) و به اشتباه بیان شده دارو اثر ندارد و یا رابطه‌ی بین متغیر مستقل و وابسته وجود ندارد (۲۷، ۲۸).

نحوه‌ی کاربرد و تفسیر p-value

مثال ۲: در یک مطالعه‌ی کوهورت آینده‌نگر که بر روی ۱۹۹ زن باردار انجام شد، هدف، بررسی رابطه‌ی عفونت کووید-۱۹ (زنان باردار مبتلا به کووید-۱۹ (٦٦ نفر) و زنان باردار غیرمبتلا به کووید-۱۹ (۱۳۳ نفر)) و پیامدهای نامطلوب بارداری (سزارین، دیابت بارداری، پراکلامپسی، پارگی زودرس کیسه‌ی آب، تأخیر رشد داخل رحمی، مرده‌زایی، پلی‌هیدرامینوس، الیگوهیدرامینوس، زایمان زودرس، وزن کم هنگام تولد، بستری نوزاد در بخش مراقبت‌های ویژه‌ی نوزادان) بود. تمامی پیامدهای سنجیده شده کیفی و دو حالتی بودند. شاخص اندازه‌ی اثر گزارش شده در این مطالعه به دلیل پیگیری ۹ ماهه‌ی مادران باردار، نسبت خطر بود. نسبت خطر گزارش شده برای رخداد سزارین معادل ۱/٥٤ (حدود اطمینان ۹۵ درصد: ۱/۰٤ تا ۲/۲۷) بدست آمد. مقدار p-value بدست آمده معادل ۰/۰۳۰ بدست آمد. مقدار p-value،کوچک‌تر از خطای نوع یک (۵ درصد مفروض شده) بوده و بنابراین یافته‌ها معنی‌دار آماری‌ست. مقدار p-value نشان می‌دهد که ما به درستی فرضیه‌ی صفر را رد کرده‌ایم و آماره‌ی آزمون، خارج از منطقه‌ی قبول فرضیه‌ی صفر قرار دارد و احتمال اینکه ما اشتباه کرده باشیم و فرضیه‌ی صفر را به غلط رد کنیم (تعریف p-value) فقط ۰/۰۳۰ است. همچنین با بررسی حدود اطمینان متوجه می‌شویم که در حدود اطمینان عدد یک قرار ندارد و به مفهوم دیگر نسبت خطر بین دو جامعه (که در محدوده‌ی حدود اطمینان قرار دارد)، عدد یک نبوده و ارتباط معناداری بین عفونت کووید-۱۹ و رخداد سزارین وجود دارد. همانطور که گفته شد، نتایج آزمون فرضیه (گزارش p-value) و برآورد (گزارش حدود اطمینان) کاملاً یکسان و معنی‌دار آماری‌ست. بر اساس مقدار p-value نمی‌توان اهمیت بالینی عفونت کووید-۱۹ بر رخداد سزارین را تفسیر نمود، اما با بررسی برآورد نقطه‌ای نسبت خطر (۱/٥٤) و حدود اطمینان ۹۵ درصد: ۱/۰٤ تا ۲/۲۷، می‌توان به اهمیت بالینی عفونت پی برد. متوسط اثر بالینی عفونت بر رخداد کووید-۱۹، ٥٤ درصد است که عدد به نسبت بالایی بوده و اهمیت این عفونت بر رخداد سزارین را نشان می‌دهد (۲۹).

مثال ۳: در یک مطالعه‌ی کارآزمایی بالینی تصادفی شده که بر روی ۷۲ بیمار مبتلا به سندرم تخمدان پلی‌کیستیک انجام شد، پژوهشگران اثر مکمل کورکومین را بر شاخص‌های قندی و لیپیدی بررسی کردند. در این مطالعه پژوهشگران، روزانه کپسول کورکومین (دو بار در روز؛ مقدار ۵۰۰ میلی‌گرم) به گروه مداخله و کپسول دارونما (دو بار در روز؛ مقدار ۵۰۰ میلی‌گرم) به بیماران تجویز کردند. نمونه‌های سرم قبل و ۱۲ هفته بعد از مداخله از بیماران گرفته شد. میانگین تغییرات قند خون در گروه مداخله و دارونما به ترتیب معادل ۵/۰۹- (انحراف معیار: ۷/۲۹) و ۰/۹۸- (انحراف معیار: ۹/۱۱) بدست آمد. اختلاف میانگین تغییرات و مقدار p-value به ترتیب معادل ٤/۱۱- (حدود اطمینان ۹۵ درصد: ۸/۳۵- تا ۰/۳۵-) و ۰/۰٤۸ بدست آمد. مقدار آماره‌ی آزمون و خطای معیار به ترتیب معادل ۲/۰٤- و ۲/۰۱ بدست آمد. شاخص p-value نشان می‌دهد که ما به درستی فرضیه‌ی صفر را رد کرده‌ایم و آماره‌ی آزمون، خارج از منطقه‌ی قبول فرضیه‌ی صفر قرار دارد و احتمال اینکه ما اشتباه کرده باشیم و فرضیه‌ی صفر را به غلط رد کنیم (تعریف p-value) فقط ۰/۰٤۸ است. همچنین با بررسی حدود اطمینان متوجه می‌شویم که در حدود اطمینان عدد صفر قرار ندارد و به مفهوم دیگر اختلاف میانگین تغییرات بین دو جامعه (که در محدوده‌ی حدود اطمینان قرار دارد)، عدد صفر نبوده و مکمل کورکومین در واقعیت بر قند خون زنان مبتلا

به سندرم تخمدان پلی‌کیستیک اثر دارد. مقدار p-value صرفاً معنی‌داری آماری را نشان می‌دهد و کوچکی آن نشان‌دهنده‌ی اثر بالینی مکمل کورکومین نیست. با بررسی شاخص اختلاف میانگین تغییرات و حدود اطمینان متوجه می‌شویم که هر چند یافته‌ها معنی‌داری آماری دارد اما از نظر بالینی اهمیت خاصی ندارد. کاهش متوسط قند خون ٤/١١ واحد از نظر بالینی کاملاً بی‌اهمیت است. شاید اگر متوسط کاهش قند خون بین ٢٠ تا ٣٠ واحد بود، می‌توانستیم اثر بالینی مکمل کورکومین بر قند خون را با اهمیت بدانیم. اگر در این مطالعه اثربخشی بالینی مکمل کورکومین را صرفاً بر اساس مقدار p-value تعیین می‌کردیم، نتیجه‌گیری نادرستی انجام می‌شد و شواهد غلطی در اختیار متخصصین بالینی قرار می‌گرفت (٣٠).

مثال ٤: در یک مطالعه‌ی بالینی شبه‌تجربی (بدون تخصیص تصادفی) که بر روی ١٥٠ بیمار مبتلا به فیبروم رحمی انجام شد، پژوهشگران با دو روش میومکتومی باز و لاپاروسکوپی، فیبروم‌ها را خارج نمودند و در انتها با شستشوی پریتوئن تعداد سلول‌های اسپایندل (Spindle cells) را بین دو گروه شمرده و مقایسه کردند. تعداد سلول‌های اسپایندل در گروه لاپاروسکوپی و میومکتومی باز به ترتیب معادل ٢ (٢/٦ درصد) و ٥ (٦/٩ درصد)، گزارش شد. مقدار p-value گزارش شده معادل ٠/٢٠٤ گزارش شد. پژوهشگران نتیجه‌گیری کردند که تفاوت معنی‌داری بین دو گروه گزارش نشده است. همچنین نسبت شانس گزارش شده، معادل ٣/٦٦ و حدود اطمینان ٩٥ درصد: ٠/٦٢ تا ٢١/٤٣ بدست آمد. هر چند که مقادیر p-value و حدود اطمینان ٩٥ درصد، معنی‌داری آماری را تأیید نکردند اما حدود اطمینان ٩٥ درصد اطلاعات بسیار بیشتری را در اختیار پژوهشگران قرار می‌دهد. حدود اطمینان گزارش شده بشدت پهن (Wide) بود. حدود اطمینان پهن نشان‌دهنده‌ی خطای تصادفی بالا و به مفهوم دیگر توان پایین مطالعه است. هر چند که تعداد نمونه‌ها در دو گروه (٧٨ نفر در گروه لاپاروسکوپی و ٧٢ نفر در گروه میومکتومی باز) مناسب بود اما رخداد پیامد (فراوانی تعداد سلول‌های اسپایندل) نادر بود که این امر خطای معیار را بشدت افزایش می‌دهد و در نهایت حدود اطمینان پهن شد و توان مطالعه کاهش یافت. پژوهشگران در یافته‌های مطالعه تأکید نمودند که توانی معادل ٢٣ درصد داشته‌اند. همانطور که قبلاً توضیح داده شد، توان مطالعه در ابتدای طراحی و قبل از جمع‌آوری داده‌ها، حداقل معادل ٨٠ درصد مفروض می‌شود. یعنی اگر در واقعیت، فراوانی سلول‌های اسپایندل بین دو گروه متفاوت باشد ما بتوانیم در ٨٠ درصد مواقع این اختلاف را ببینیم. اما در این مطالعه توان صرفاً ٢٣ درصد بوده و می‌توان حدس زد که ممکن است در واقعیت فراوانی سلول‌های اسپایندل بین دو گروه متفاوت بوده و پژوهشگران این اختلاف را نتوانسته‌اند پیدا کنند. در این موارد توصیه بر تکرار مطالعه با حجم نمونه‌های بالاتر است و همچنین می‌توان با انجام مطالعات ثانویه و متاآنالیز شاخص‌های اندازه‌ی اثر چندین مطالعه در این زمینه را تجمیع کرده و توان را افزایش داد (٣١).

مثال ٥: در یک مطالعه‌ی کارآزمایی بالینی تصادفی شده که بر روی ٩٤ زن ٤٥ تا ٥٥ ساله انجام شد، پژوهشگران اثر کوچینگ سلامت (Health coaching) را بر علائم منوپوز و افسردگی بررسی کردند. مداخلات توسط مامای آموزش دیده در پنج جلسه (جلسات ٣٠ تا ٤٥ دقیقه) انجام شد و اثر آن چهار ماه بعد از مداخله بر روی نمره‌ی پرسش‌نامه‌ی علائم منوپوز و افسردگی ارزیابی شد. اختلاف میانگین علائم منوپوز و افسردگی بین دو گروه مداخله و شاهد به ترتیب معادل ١٢/٥١- (حدود اطمینان ٩٥ درصد: ١٠/٥٩- تا ١٤/٤٢-) و ٥/٧٢- (حدود اطمینان ٩٥ درصد: ٣/٨٣- تا ٧/٦١-) بدست آمد. یافته‌ها نشان داد که مقادیر p-value برای هر دو پیامد معادل ٠/٠٠١> بدست آمد. شاخص p-value نشان می‌دهد که ما به درستی فرضیه‌ی صفر را رد کرده‌ایم و آماره‌ی آزمون، خارج از منطقه‌ی قبول فرضیه‌ی صفر قرار دارد و احتمال اینکه ما اشتباه کرده باشیم و فرضیه صفر را به غلط رد کنیم (تعریف p-value) ٠/٠٠١> است. بنابراین می‌توان مطمئن بود که این مداخله در واقعیت نیز اثر دارد و شانسی نیست. با بررسی مقادیر اختلاف میانگین و حدود اطمینان‌ها می‌توان علاوه بر معنی‌داری آماری به معنی‌داری بالینی مداخله نیز اطمینان داشت. بر اساس دامنه‌ی نمرات هر دو پیامد (نمرات پرسش‌نامه‌ی علائم منوپوز و افسردگی)، اهمیت بالینی این مداخله قابل توجه است (٣٢).

References

1. Nakagawa S, Cuthill IC. Effect size, confidence interval and statistical significance: a practical guide for biologists. Biol Rev Camb Philos Soc 2007; 82(4): 591-605.
2. Schober P, Bossers SM, Schwarte LA. Statistical significance versus clinical importance of observed effect sizes: what do P values and confidence intervals really represent? Anesth Analg 2018; 126(3): 1068-72.
3. Perezgonzalez JD. Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. Front Psychol 2015; 6: 223.
4. Mayo DG, Spanos A. Severe testing as a basic concept in a Neyman-Pearson philosophy of

induction. Br J Philos Sci 2006; 57(2).
5. Davis RB, Mukamal KJ. Hypothesis testing: means. Circulation 2006; 114(10): 1078-82.
6. Tello R, Crewson PE. Hypothesis testing II: means. Radiology 2003; 227(1): 1-4.
7. Allua S, Thompson CB. Hypothesis testing. Air Med J 2009; 28(3): 108-53.
8. Christensen R. Testing fisher, neyman, pearson, and bayes. Am Stat 2005; 59(2): 121-6.
9. Huberty CJ. Historical origins of statistical testing practices: The treatment of Fisher versus Neyman-Pearson views in textbooks. J Exp Educ 1993; 61(4): 317-33.
10. Sedgwick P. Pitfalls of statistical hypothesis testing: type I and type II errors. BMJ 2014; 349: g4287.
11. Lu J, Qiu Y, Deng A. A note on Type S/M errors in hypothesis testing. Br J Math Stat Psychol 2019; 72(1): 1-17.
12. Newman MC. "What exactly are you inferring?" A closer look at hypothesis testing. Environ Toxicol Chem 2008; 27(5): 1013-9.
13. Emmert-Streib F, Dehmer M. Understanding statistical hypothesis testing: The logic of statistical inference. Mach Learn Knowl Extr 2019; 1(3): 945-62.
14. Boulesteix AL, Hable R, Lauer S, Eugster MJ. A statistical framework for hypothesis testing in real data comparison studies. Am Stat 2015; 69(3): 201-12.
15. Gill J. The insignificance of null hypothesis significance testing. Polit Res Q 1999; 52(3): 647-74.
16. Altman DG, Bland JM. Standard deviations and standard errors. BMJ 2005; 331(7521): 903.
17. Lee DK, In J, Lee S. Standard deviation and standard error of the mean. Korean J Anesthesiol 2015; 68(3): 220-3.
18. O'Brien SF, Yi QL. How do I interpret a confidence interval? Transfusion 2016; 56(7): 1680-3.
19. Guyatt G, Jaeschke R, Heddle N, Cook D, Shannon H, Walter S. Basic statistics for clinicians: 1. Hypothesis testing. CMAJ 1995; 152(1): 27-32.
20. Dahiru T. P-value, a true test of statistical significance? A cautionary note. Ann Ib Postgrad Med 2008; 6(1): 21-6.
21. Andrade C. The P value and statistical significance: misunderstandings, explanations, challenges, and alternatives. Indian J Psychol Med 2019; 41(3): 210-5.
22. Ioannidis JPA. The proposal to lower P value thresholds to. 005. JAMA 2018; 319(14): 1429-30.
23. Dick F, Tevaearai H. Significance and Limitations of the p Value. Eur J Vasc Endovasc Surg 2015; 50(6): 815.
24. Lee DK. Alternatives to P value: confidence interval and effect size. Korean J Anesthesiol 2016; 69(6): 555-62.
25. Ranstam J. Why the P-value culture is bad and confidence intervals a better alternative. Osteoarthritis Cartilage 2012; 20(8): 805-8.
26. Altman DG, Bland JM. How to obtain the P value from a confidence interval. BMJ 2011; 343: d2090.
27. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol 2016; 31(4): 337-50.
28. Shreffler J, Huecker MR. Type I and Type II errors and statistical power; 2020.
29. Pirjani R, Hosseini R, Soori T, Rabiei M, Hosseini L, Abiri A, et al. Maternal and neonatal outcomes in COVID-19 infected pregnancies: a prospective cohort study. J Travel Med 2020; 27(7): taaa158.
30. Heshmati J, Moini A, Sepidarkish M, Morvaridzadeh M, Salehi M, Palmowski A, et al. Effects of curcumin supplementation on blood glucose, insulin resistance and androgens in patients with polycystic ovary syndrome: A randomized double-blind placebo-controlled clinical trial. Phytomedicine 2021; 80: 153395.
31. Asgari Z, Hashemi M, Hosseini R, Sepidarkish M, Seifollahi A. Comparison of the number of spindle cells in peritoneal washings between laparoscopic myomectomy with morcellation and open myomectomy without morcellation. J Minim Invasive Gynecol 2021; 28(7): 1391-6.
32. Shokri-Ghadikolaei A, Bakouei F, Agajani Delavar M, Azizi A, Sepidarkish M. Effects of health coaching on menopausal symptoms in postmenopausal and perimenopausal women. Menopause 2022; 29(10): 1189-95.

DOI: *10.48305/jims.v41.i732.0725*

Journal of Isfahan Medical School **Vol. 41, No. 732, 4th Week, October 2023**

Received: 28.05.2023 Accepted: 05.09.2023 Published: 20.10.2023

An Overview of Statistical and Clinical Concepts with the Approach of Hypothesis Testing (P-value)

Mahdi Sepidarkish[1], Zahra Mohammadi-Pirouz[2]

Review Article

Abstract

Application and interpretation of statistical significance of association are the basic and necessary principle in medical research. Traditionally, hypothesis testing and reporting p-values are widely used to quantify the statistical significance of observed results. In the last two decades, the calculation of a p-value in research and especially the use of a threshold to declare the statistical significance of the p-value have been challenged. The limitations of p-value, such as the dependence of its value on the sample size and not reflecting the clinical significance, have been repeatedly mentioned. The statisticians and methodologists recommend do not report p-value alone, and reporting of effect size index with corresponding confidence interval is mandatory. However, many researchers do not pay attention to this and do not even interpret the p-value correctly. The present study intended, to provide an integrated instruction for reporting the statistical and clinical significance in medical sciences with the approach of hypothesis testing (reporting p-value).

Keywords: Confidence interval; Data analysis; Data interpretation; Hypothesis; Statistics

Citation: Sepidarkish M, Mohammadi-Pirouz Z. **An Overview of Statistical and Clinical Concepts with the Approach of Hypothesis Testing (P-value).** J Isfahan Med Sch 2023; 41(732): 725-35.

1- Assistant Professor, Department of Biostatistics and Epidemiology, School of Public Health, Babol University of Medical Sciences, Babol, Iran

2- MSc Student, Department of Biostatistics and Epidemiology, School of Public Health, Babol University of Medical Sciences, Babol, Iran

Corresponding Author: Mahdi Sepidarkish, Assistant Professor, Department of Biostatistics and Epidemiology, School of Public Health, Babol University of Medical Sciences, Babol, Iran; Email: mahdi.sepidarkish@gmail.com