"مجله علوم زراعی ایران"
جلد دوم، شماره ۱، ۱۳۷۹

نکات اساسی در تجزیه و تحلیل داده‌های آماری

Main Points in Statisical Data Analysis

عباس گرامی [۱]

چکیده

هر تجزیه و تحلیل آماری مبتنی بر یک مدل آماری است. مدل آماری خود شامل اجزایی است که مهم‌ترین آن‌ها رابطه ریاضی بین متغییر وابسته با عوامل و فاکتورهایی است که محقق علاقمند به بررسی اثرات آن‌ها را روی میانگین این متغیر می‌باشد. علاوه بر این فاکتورها، عوامل مزاحم دیگری هستند که اگر در رابطه ریاضی مدل محلوظ نشوند، از صحت، اعتبار و دقت تجزیه و تحلیل داده‌های آماری می‌کاهد. مورد مهم دیگر، مفروضات پیرامون عوامل تصادفی موجود در رابطه ریاضی مدل است که نقشی کمتر از مورد اول ندارد. این مفروضات عمدتاً به گشتاورهای اول و دوم عوامل تصادفی و توزیع احتمال آن‌ها مربوط می‌شود. صحت نتیجه‌گیری از تجزیه داده‌های آماری منوط به برقراری نسبی این مفروضات و رابطه صحیح است. در بسیاری از تجزیه و تحلیل داده‌های آماری این نکات اساسی به فراموشی سپرده می‌شوند و بدون توجه به صحت یا عدم صحت برقراری آن‌ها، داده‌ها تجزیه و تحلیل می‌شوند. در این مقاله به تشریح بعضی انواع مدل‌های آماری از دو جنبه فوق پرداخته می‌شود.

واژه‌های کلیدی: گشتاورها، همگن بودن واریانس‌ها، تبدیل، نرمال بودن، نزدیکترین همسایه

مقدمه

در تجزیه و تحلیل داده‌های حاصل از اجرای اغلب طرح‌های تحقیقاتی معمولاً تجزیه واریانس اولین و مهم‌ترین مرحله را شامل می‌شود. هر تجزیه واریانس مبتنی بر یک مدل آماری است که خود در حالت کلی شامل سه قسمت زیر است:

- رابطه ریاضی
- مفروضات
- شرایط و محدودیت‌ها

۱ـ۱ـ رابطه ریاضی

این قسمت از مدل متغیر وابسته را با عوامل و فاکتورها به صورت زیر مرتبط می‌کند:

$$y=f(\mu, \alpha, \beta)+e$$

که در آن y مقدار صفت مورد اندازه‌گیری μ میانگین کل y در جامعه، بردار α اثر عوامل ثابت و بردار β اثر عوامل تصادفی را شامل شده و e اثر سایر عوامل (اشتباه) است. در صورت حضور عوامل ثابت، به تنهایی، مدل را با اثرات ثابت، تنها حضور عوامل تصادفی مدل را تصادفی و در حضور عوامل ثابت و تصادفی در کنار هم آن را مختلط گویند. تابع $f=(\mu, \alpha, \beta)$ در ساده‌ترین و معمول‌ترین شکل خود از نوع خطی است و برای بردار مشاهدات Y به صورت زیر و در قالب ماتریس قابل بیان است:

$$Y=\mu 1+X_1\alpha+X_2\beta+e$$

که در آن X_1 و X_2 در صورت عدم حضور متغیرهای کمکی (Covariates) ماتریس هایی هستند که عناصر آن را

تاریخ دریافت: ۱۳۷۸/۷/۱۸
تاریخ پذیرش: ۱۳۷۸/۱۰/۱۲

۱ـ استادیار پژوهش

صفر و یک تشکیل می‌دهند و به ماتریس‌های طرح (Design Matrices) مشهورند چه، مقادیر تشکیل دهنده آن‌ها بستگی به نوع طرح دارد و 1 نیز یک بردار واحد است. در صورت ضرورت می‌توان در رابطه فوق متغیرهای کمکی را نیز ملحوظ کرد.

۱ـ۲ مفروضات

این قسمت شامل مفروضاتی پیرامون عوامل تصادفی β و e است که عمدتاً به گشتاورهای اوّل، دوّم و گشتاورهای مشترک آن‌ها مربوط است. متداول‌ترین این مفروضات عبارتند از (Searle 1987):

$$E(\beta)=0 \quad Var(\beta)=\Sigma_\beta \quad , \quad E(e)=0 \quad Var(e)=\sigma_e^2 I$$

$$Cov(\alpha, \beta)=0$$

که Σ_β به ماتریس واریانس کواریانس β مشهور است واریانس اجزاء β و کوواریانس بین دو به دوی اجزاء آن را شامل می‌شود. این مفروضات برای برآورد کردن پارامترهای ثابت (α) مدل و پیش بینی عوامل تصادفی مدل (β) از روش حداقل مربعات (Least Squares) الزامی است و برای انجام آزمون‌های فرض پیرامون اثرات ثابت مدل معین بودن توزیع احتمالی عوامل تصادفی β و e ضرورت دارد.

۱ـ۳ـ شرایط و محدودیت‌ها

اصولاً فرض می‌شود که سایر شرایط تأثیر گذار روی y برای کلیه عوامل مدل یکسان است. صحت نتیجه‌گیری از داده‌های آماری مبتنی بر مصداق نسبی موارد سه گانه فوق است. از آنجایی که اکثر تحقیقات کشاورزی، به جز در شرایط آزمایشگاهی و گلخانه‌ای در مزارع آزمایشی صورت می‌گیرد قالب کلی طرح‌ها بلوکی و در موارد نادر از نوع پیچیده‌تری خواهد بود. بر این اساس در این مقاله اجزاء مدل در طرح بلوک مورد بحث و بررسی قرار می‌گیرد و بسیاری از رهنمودهای آن قابل استفاده در طرح‌های پیچیده‌تر نیز خواهد بود. از آنجایی که نگارنده ظرف چند سال گذشته به طور مستقیم و غیر مستقیم در جریان بسیاری از طرح‌های کشاورزی قرار گرفته است که تجزیه و تحلیل آن‌ها بدون توجّه به صحت نسبی اجزاء مدل مربوطه انجام شده است ارائه این مقاله را خالی از فایده نمی‌داند. البته در سال‌های اخیر گرایش بیشتری در توجّه به این مسایل ملاحظه می‌شود که خود جای بسی خوشحالی است.

۲ـ مدل طرح بلوک‌های تصادفی

رابطه ریاضی مدل آماری طرح بلوک‌های تصادفی در حالت کلی به صورت زیر بیان می‌شود:

$$y_{ij}=\mu+\alpha_i+\beta_j+e_{ij}$$

که در آن y_{ij} مشاهده صفت یا متغیر مورد نظر برای تیمار iاُم در بلوک jاُم، μ میانگین کل، α_i اثر تیمار iاُم (معمولاً ثابت) و β اثر بلوک jاُم (ثابت یا تصادفی) و e_{ij} اشتباه آزمایشی مربوط به تیمار iاُم در بلوک jاُم می‌باشد که در آن t و ... و ۲ و ۱ =i و b و... و ۲ و ۱ =j می‌باشد. اگر فرض کنیم تعداد واحدهای آزمایشی تمام بلوک‌ها با هم مساوی و برابر k باشد طرح را proper گویند. در صورتی که k=t باشد و تعداد تکرار هر تیمار در داخل هر بلوک فقط و فقط یک باشد طرح را بلوک‌های کامل نامند. در صورتی که k<t باشد و هر تیمار در یک بلوک یا اصلاً تکرار نداشته یا حداکثر یک تکرار داشته باشد طرح را بلوک‌های ناقص از نوع دوگانه (Binary) گویند، طرح‌های لاتیس از این گونه‌اند. در صورتی که k>t باشد طرح را بلوک‌های کامل تعمیم یافته (Generalized Blocks) گویند. طرح‌های از نوع اوّل و دوّم متداول‌ترین طرح‌های مورد استفاده در تحقیقات کشاورزی مزرعه‌ای هستند.

مفروضات مدل به شرح زیر می‌باشد: Mead *et al.*, (1993), Mead (1994):

الف: نرمال بودن اشتباهات با عبارت $e_{ij} \sim N(0,\sigma_e^2)$ بیان می‌شود که به طور ضمنی یکسان بودن واریانس مشاهدات از تیماری به تیمار دیگر را نیز نشان می‌دهد.

ب: استقلال اشتباهات آزمایش از یکدیگر به مفهوم صفر بودن کواریانس بین اشتباه آزمایشی تیمارها و تکرارهای متفاوت.

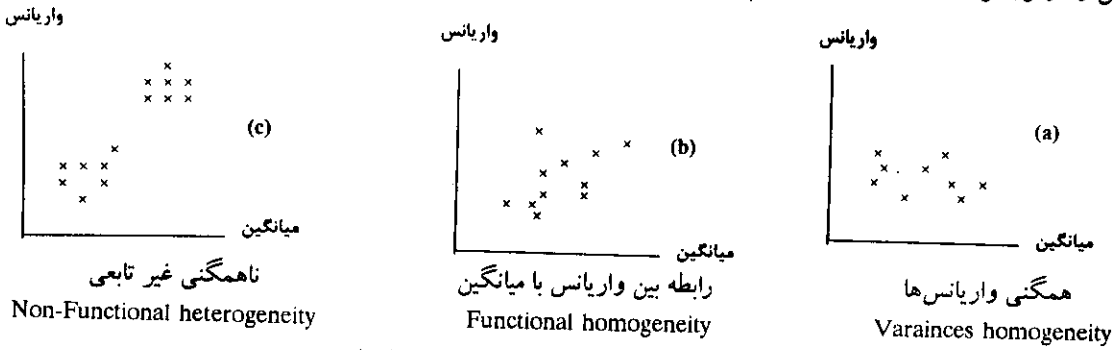
پ: جمع‌پذیر بودن اثرات تیمارها و محیط که در رابطه فوق نشان داده شده است.

غالباً در عمل دلایل خوبی بر عدم صحت مفروضات فوق وجود دارد، انحراف از این مفروضات بر سطح معنی دار شدن اختلاف میانگین صفات مورد نظر در تیمارها و حساسیت آزمون‌های F و t اثر می‌گذارد که ذیلاً شرح داده می‌شود:

۲ـ۱ ـ نرمال نبودن داده‌ها

در صورتی که فرض نرمال بودن داده‌ها مصداق نسبی نداشته باشد تا مرحله تشکیل جدول تجزیه واریانس و محاسبه

Fهای مورد نظر اشکال خاصی وجود ندارد اما آزمون معنی دار شدن F و قضاوت‌های آماری مبتنی بر آن، منوط به مصداق نسبی نرمال بودن است، گرچه از نظر آماری F نسبت به انحراف از نرمال بودن چندان حساس نیست اما برای مواقعی که F در مرز معنی دار شدن است سطح اشتباه واقعی ممکن است بزرگتر یا کوچکتر از سطح اشتباه ظاهری (بر اساس مقایسه F محاسبه شده با جدول) باشد. به طور مثال ممکن است α ظاهری از ۵ درصد کوچکتر باشد که بر اساس α ظاهری فرض H_0 رد می‌شود در حالی که بر اساس α واقعی H_0 رد نخواهد شد. معمول‌ترین حالت نرمال نبودن داده‌ها در تحقیقات کشاورزی شامل صفات شمارشی و داده‌های درصدی و نمره‌ای است که برای این داده‌ها فرم توزیع اولاً متقارن نبوده و ثانیاً میانگین واریانس از نظر ریاضی با یکدیگر مرتبط هستند. برای آزمون فرض نرمال بودن داده‌ها روش‌های مختلفی وجود دارد، از جمله این روش‌ها می‌توان نیکویی برازش (Goodness of Fit) را نام برد و مناسب برای زمانی است که تعداد واحدهای آزمایشی یا کرت‌ها در کل آزمایش (n) از ۱۰۰ بزرگتر باشد. هم چنین از روش Kolmogorov-Smirnov که روشی غیر پارامتری بوده می‌توان برای حالت $50<n<100$ استفاده کرد. روش دیگر عبارت از Q-Q plot است (Johnson and Wichern, 1988). بسته به مورد یکی از روش‌های مذکور برای آزمون نرمال بودن داده‌ها روی برآورد اشتباهات آزمایش یعنی $e_{ij} = \bar{y}_{ij} - \bar{y}_{io} - \bar{y}_{oj} + \bar{y}_{..}$ ها مورد استفاده قرار می‌گیرد که در آن $\bar{y}_{io}$ میانگین تیمار اُم و $\bar{y}_{oj}$ میانگین بلوک اُم و $\bar{y}_{..}$ کل است. در صورتی که فرض نرمال بودن داده از طریق یکی از آزمون‌های مذکور رد نشد تجزیه واریانس و آزمون فرض‌ها مطابق معمول انجام می‌شود اما در صورتی که این فرض رد شود دو حالت متمایز می‌توان در نظر گرفت: حالت اوّل این که واریانس تیمارهای متفاوت با یکدیگر برابر نیستند، حالت دوّم اینکه مشاهدات دارای توزیع نرمال نیستند. عدم همگن بودن واریانس‌ها را در حالت کلی می‌توان ناشی از دو عامل دانست: اولین عامل وجود رابطه بین میانگین و واریانس تیمارها است و دومین عامل آن عبارت از این است که بعضی تیمارها هم به لحاظ میانگین و هم به لحاظ واریانس متفاوت از سایرین باشند. برای کشف این عوامل می‌توان از ترسیم دیاگرام پراکنش مربوط به میانگین و واریانس تیمارها در یک صفحه محورهای مختصات دو بعدی استفاده نمود. سه حالت متمایز از این دیاگرام را می‌توان در قالب حالت‌های (a)، (b) و (c) از شکل ۱ ملاحظه کرد. عدم همگن بودن واریانس‌ها از جمله عواملی است که آزمون F نسبت به آن خیلی حساس است و عدم صحت نسبی آن به طور بارزی سطح معنی دار شدن آزمون‌ها را تحت تأثیر قرار می‌دهد. در صورتی که عدم همگن بودن واریانس‌ها به دلیل وجود رابطه $\sigma^2=\Phi(\mu)$ بین میانگین و واریانس باشد ضرورت دارد تا تبدیلی مناسب به دست آوریم به طوری که اگر بین میانگین و واریانس $Z=f(Y)$ ارتباطی موجود نباشد می‌توان نشان داد که در تبدیل $Z=f(Y)$ به طور تقریب $Var(Z)\approx[f'(\mu)]^2\Phi(\mu)$ است و این باید مقداری ثابت مستقل از μ باشد. برای این منظور باید $f(y)\alpha\int\frac{dy}{\sqrt{\Phi(y)}}$ یعنی $f(y)$ مضربی از تابع متناسب با عبارت انتگرال باشد. به عنوان مثال تئوری مذکور پیشنهاد می‌کند که در صورتی که صفت مورد نظر در طرح آزمایش دارای توزیع پواسن است تبدیل $z=\sqrt{Y}$ همگن بودن واریانس‌ها را تأمین می‌کند. برای توزیع دو جمله‌ای، تبدیل $Z=\arcsin\sqrt{y/n}$ این وضعیت را ایجاد می‌کند.



همگنی واریانس‌ها
Varainces homogeneity

رابطه بین واریانس با میانگین
Functional homogeneity

ناهمگنی غیر تابعی
Non-Functional heterogeneity

شکل ۱ـ حالت‌های مختلف واریانس تیمارها
Figure 1. Different cases of treatments variances

مثال ۱: برای بررسی اثرات چند سم برای از بین بردن حشرات، آزمایشی اجرا شده است که نتایج آن در جدول ۱ آمده است:

جدول ۱ـ متوسط تعداد تخم حشره در هر متر مکعب برای تیمارهای مختلف

Table 1. Mean eggs counts per cubic meter for different insecticides.

تیمارها Treatments	شاهد(بدون حشره‌کش) Control	A	B	C
متوسط تخم موجود حشره در هر متر مکعب Mean egg/m^{-3}	150	85	8	1.2

با فرض همگن بودن واریانس‌ها به دست آورده‌ایم $S_{\bar{y}}=6.4$ بدیهی است در چنین آزمایشی واریانس مربوط به حشره‌کش C نمی‌تواند زیاد بزرگ باشد. چنین مطلبی در مورد حشره‌کش B هم تقریباً درست است. صحت نسبی همگنی واریانس‌ها برای داده‌هایی نظیر این باید قبل از تجزیه بررسی شود.

مثال ۲: فرض کنید آزمایشی در قالب یک طرح بلوک‌های کامل تصادفی شامل ۴ تکرار و ٦ تیمار انجام پذیرفته و نتایج حاصل در جدول ۲ آمده است:

جدول ۲ـ مشاهدات آزمایش طرح بلوک‌های کامل تصادفی با ٦ تیمار

Table 2. Data from a complete randomized block with design 6 treatments.

تیمار Treatment	بلوک ۱ B1	بلوک ۲ B2	بلوک ۳ B3	بلوک ۴ B4
A	538	422	377	315
B	438	442	319	380
C	77	61	157	52
D	115	57	100	45
E	17	31	87	16
F	18	26	77	20

از تشکیل جدول تجزیه واریانس برای داده‌های طرح فوق بدست خواهیم آورد $S^2=2653$ که نتیجه می‌دهد انحراف معیار هر کرت 51.5 بوده و اگر تیمار F را در نظر بگیریم، با استفاده از توزیع نرمال، 25% عملکردها باید کمتر از صفر باشند که این نامعقول است. به منظور انجام تجزیه و تحلیل روی داده هایی که توأم با عدم برقراری این فرض هستند دو حالت متمایز، یکی وجود رابطه بین میانگین و واریانس تیمارها و دیگری عدم وجود رابطه مشهود است.

در بعضی مواقع صفت مورد بررسی ممکن است برای تیمارهای متفاوت نرمال باشد اما دارای واریانس‌های متفاوت باشند و طبیعی است در چنین حالتی لزوماً رابطه تابعی بین میانگین و واریانس وجود ندارد. مثلاً صفتی در نسل F_2 ممکن است دارای واریانس بزرگتری نسبت به سایر نسل‌ها باشد، در اینجا همانند آن چه که در قبل بحث شد انجام تبدیل نه تنها مشکلی را حل نمی‌کند بلکه انجام تبدیل به نرمال بودن مشاهدات نیز لطمه خواهد زد.

مشکل نابرابر بودن واریانس‌ها از حالت $t=2$ جامعه شروع می‌شود که به مسئله Behrens-Fisher معروف است که برای آن راه حل دقیق اما مشکل وجود دارد. در مقابل راه حل‌های تقریبی اما ساده نیز پیشنهاد شده است (Gerami 1986). برای حالت $t>2$ یک راه حل پیشنهادی توسط استیل و توری Steel & Torrie (1980) توضیح داده شده است. در این حالت،

بر اساس مقایسات مورد نظر بین تیمارها، SSE را تجزیه نموده به طوری که هر جزء دارای درجه آزادی برابر بوده و برای یک مقایسه خاص مورد استفاده قرار می‌گیرد. طبیعی است که این راه حل قدری محدود و محاسبات آن تا حدودی وقت گیر است. نگارنده نرم‌افزاری آماری سراغ ندارد که این نوع تجزیه را به طور مستقیم انجام دهد. این روش اشتهار به "تقسیم SS خطای آزمایشی" دارد. برای آزمون فرض یکسان بودن واریانس‌ها در ادبیات مربوطه روش‌های متفاوتی ذکر شده است که همگی روش‌های آزمون تقریبی ارائه می‌دهند و معروف‌ترین آن‌ها آزمون Bartlett است. ویراهاندی Weerahandi (1995) در مقاله‌ای ضمن بیان روش آزمون دقیقی در این خصوص، مسئله Behrens-Fisher را به پیش از t>2 جامعه تعمیم داده است. اخیراً نیز زاهدیان و گرامی (۱۳۷۸) روش‌های مختلف پیشنهادی را در حالت دو جامعه‌ای و چند جامعه‌ای مورد بررسی قرار داده و با روش پیشنهادی ویراهاندی Weerahandi (1995) مورد مقایسه قرار داده‌اند.

مثال ۳: فرض کنید هدف از اجرای یک آزمایش مقایسه میانگین چهار تیمار باشد و واریانس‌های درون تیمارها با یکدیگر تفاوت فاحش نشان دهند به طوری که این تفاوت‌ها ناشی از تفاوت تعداد نمونه‌ها نباشد. یک مثال فرضی از این قبیل آزمایش‌ها در جدول ۳ آورده شده است.

جدول ۳ـ تعداد نمونه، میانگین و انحراف استاندارد مشاهدات حاصل از یک آزمایش با ۴ تیمار

Table 3. Number of observations, sample means and sample standard deviations of an experiment.

Treatment تیمار	n	$\bar{y}$	s
1	6	13.1	1.9
2	8	14.1	1.7
3	5	14.6	0.89
4	7	12.9	0.55

نتایج حاصل از تجزیه آماری داده‌های آزمایش فوق الذکر در قالب یک طرح کاملاً تصادفی در جدول ۴ آمده است:

جدول ۴ـ تجزیه واریانس سن آزمایش مثال ۳

Table 4. Analysis of variance for data of example 3.

S.O.V منابع تغییرات		df	SS	MS	F
Treatment	تیمار	3	11.88	3.96	1.712
Error	اشتباه	22	50.90	2.31	

ملاحظه می‌شود که فرض H_0 در مورد برابری میانگین این چهار تیمار رد نمی‌شود. (p=۰/۱۹۴) حال اگر فرض برابر بودن واریانس‌ها را نپذیریم ویراهاندی Weerahandi (1995) نشان داده است که با ۰/۰۵ > ۰/۰۴۶=p فرض برابری میانگین‌ها رد می‌شود.

یک راه عملی برای پیدا کردن بهترین نوع تبدیل روش توکی است که در حالت کلی به صورت $Z_{ij} = Y^{p}_{ij}$ بیان می‌شود که بسیاری از انواع تبدیلات از آن بدست می‌آیند. مثلاً برای $p=\frac{1}{2}$ تبدیل ریشه دوّم است، برای p=-1 تبدیل معکوس است و برای p خیلی کوچک تبدیل لگاریتمی و غیره. این روش شامل مراحلی است از جمله:

الف: محاسبه مقادیر اشتباهات به صورت

$$e_{ij} = \bar{y}_{ij} - \bar{y}_{io} - \bar{y}_{oj} + \bar{y}..$$

ب: به دست آوردن ضریب رگرسیون e_{ij} با $\mu_{ij} = (\bar{y}_{io} - \bar{y}..)(\bar{y}_{oj} - \bar{y}..)$ که آن را با b نشان می‌دهیم.

پ: محاسبه p از رابطه p=1-b ...$\bar{y}$

ت: در صورتی که b تقریباً برابر با صفر باشد p تقریباً برابر با یک است و تبدیل ضرورتی ندارد.

ث: انجام تبدیل $Z_{ij} = Y^{p}_{ij}$ و انجام محاسبات تجزیه واریانس در صورتی که داده‌های تبدیل یافته خود تبدیل لازم نداشته باشند.

معمول‌ترین تبدیل‌ها در تحقیقات کشاورزی عبارتند از:

الف: تبدیل لگاریتمی: این تبدیل معمولاً وقتی بکار برده می‌شود که اثرات جمع‌پذیر نبوده و یا انحراف معیار هر تیمار مناسب با میانگین آن تیمار باشد. در این صورت از تبدیل $Z_{ij} = \log y_{ij}$ استفاده می‌کنیم. در صورتی که در مشاهدات مقادیر کمتر از ۱۰ موجود باشد تمام مشاهدات را به صورت $Z_{ij} = \log (y_{ij} + 1)$ تبدیل می‌کنیم.

ب: تبدیل ریشه دوم: این تبدیل معمولاً برای اعداد کوچک مانند تعداد گیاه آسیب دیده و تعداد علف هرز در هر کرت و یا داده هایی که که از طریق شمارش بدست آمده و به صورت درصد بیان می‌شدند و این درصد همگی بین 0%-30% یا 70%-100% باشند بکار برده می‌شود و وقتی مقادیر خیلی کوچک باشد مشاهدات را به صورت $Z_{ij} = \sqrt{y_{ij} + 0.5}$ تبدیل می‌کنیم.

پ: تبدیل سینوسی: این تبدیل مناسب برای داده‌های درصدی و شمارشی است. در این حالت مشاهده صفر به وسیله $\frac{1}{4n}$ - 100 جایگزین می‌شود که در آن n مقسوم علیه مربوطه به همه نسبت هاست. موارد زیر برای بکار بردن این تبدیل باید مد نظر قرار می‌گیرند.

ج ۱) برای داده‌های واقع در محدوده 30%-70% تبدیل ضرورت ندارد.

ج ۲) برای داده‌های واقع در محدود 0%-30% و با 70%-100% از تبدیل ریشه دوم به شرح بند ب استفاده می‌کنیم.

ج ۳) در سایر موارد از تبدیل $Z_{ij} = \arcsin\sqrt{y_{ij}}$ استفاده می‌شود.

۲-۲- استقلال اشتباهات

در آزمایشات مزرعه‌ای میزان محصول در کرت‌های مجاور مشابه ترند و در نتیجه دقت مقایسات بین تیمارها در کرت‌های مجاور از سایر مقایسات بیشتر بوده و ممکن است نتایج گمراه کننده باشند مگر چاره جویی لازم صورت پذیرد. چاره جویی در این مورد انتساب کاملاً تصادفی تیمارها به کرت‌های داخل هر بلوک است که اگر به درستی صورت گیرد، مشکل عدم استقلال اشتباهات از بین می‌رود. طرح‌های انتساب سیستماتیک این مفروض را نقض می‌کنند. کوچارن و کاکس (Cocharn and Cox 1992) اظهار می‌دارند. "انتساب تصادفی اقدامی است احتیاطی در مورد مسایلی که ممکن است اتفاق بیافتند یا نیافتند. توصیه موکد این است که زحمت تصادفی منتسب نمودن را به جان بخریم حتی اگر چندان انتظار نداشته باشیم که عدم انجام آن باعث لطمه‌ای شود. بدین وسیله محقق در مقابل حوادث غیر مترقبه تجزیه داده‌ها بیمه می‌شود."

ناگفته نماند که در انجام تجزیه آماری می‌توان از همبستگی بین واحدهای آماری مجاور استفاده نمود به طوری که دقت استنباطات آماری افزایش یابد. این همبستگی ممکن است ناشی از طبیعت واحدهای آزمایشی، شمای کلی واحدهای آزمایشی، اثرات تجمعی باقی مانده در طول زمان، آلودگی ناشی از واحدهای مجاور و یا سایر عوامل باشد که با بلوک بندی هنوز از بین نمی‌روند. پاپاداکیس (Papadakis 1937) پیشنهاد کرد که به جای انجام تجزیه واریانس از تجزیه کواریانس استفاده نموده و میانگین مشاهدات کرت‌های مجاور به هر کرت به عنوان متغیر کمکی بکار برده شوند. بعداً این پیشنهاد با تغییراتی توسط بارتلت (Bartlett 1938, 1978, 1981) نیز توصیه شد. پیشنهادهایی از این قبیل در ادبیات مربوط به طرح‌های همسایه نزدیک (Nearest Neighbour Designs) اشتهار یافته است (Ipinymoi, 1985، کیانی و گرامی ۱۳۷۷).

۲-۳- جمع پذیری اثرات تیمارها و محیط

برای بیان بهتر این مطلب به مثال زیر توجّه شود:

مثال ۴: فرض کنید آزمایشی شامل ۲ تیمار در طرح بلوک‌های کامل تصادفی با ۲ تکرار انجام شده باشد. مشاهدات مربوط به این آزمایش در جدول ۵ آورده شده است.

جدول ۵ ـ مشاهدات آزمایش طرح بلوک‌های کامل تصادفی با ۲ تیمار و ۲ بلوک

Table 5. Observations of a complete randomized block designs with 2 treatments and 2 blocks

Treatments تیمارها	بلوک ۱ B_1	بلوک ۲ B_2	تفاضل بلوک‌ها (B_1-B_2)
A	180	120	60
B	160	100	60
(A-B)	20	20	-

ملاحظه می‌شود که اختلاف دو بلوک از تیماری به تیمار دیگر و اختلاف دو تیمار از بلوکی به بلوک دیگر یکسان است. این مثال مصداق جمع‌پذیر بودن اثرات تیمارها و یا اثرات تکرار یا محیط است. حال به مثال زیر توجّه شود:

مثال ۵: فرض کنید در آزمایش مثال ۴ به جای داده‌های جدول ۵، داده‌های جدول ٦ حاصل شده باشد:

جدول ٦ـ مشاهدات آزمایش طرح بلوک‌های کامل تصادفی با ۲ تیمار و ۲ بلوک

Table 6. Observations of a complete randomized block designs with 2 treatments and 2 blocks.

Treatments تیمارها	بلوک ۱ B_1	بلوک ۲ B_2	تفاضل بلوک‌ها B_1-B_2	تفاضل بلوک‌ها به بلوک ۱ $(B_1-B_2)/B_1$
A	180	120	60	1/3
B	150	100	50	1/3
A-B	30	20		

ملاحظه می‌شود که این اختلافات دیگر همانند مثال ۴ یکسان نیستند. حال اگر اختلافات را به طور نسبی به دست آوریم ملاحظه می‌شود که با یکدیگر برابر هستند در این مثال اثرات تیمارها و محیط به جای جمع‌پذیر بودن ضرب‌پذیر هستند و می‌توان به جای $y_{ij} =\mu+\alpha_i+\beta_i+e_{ij}$ از $Log\ y_{ij} =\mu+\alpha_i+\beta_j+e_{ij}$ استفاده نمود. اگر داده‌های مثال ۵ را بدون توجّه به جمع‌پذیر نبودن تجزیه کنیم با CV=3.7% که معنی دار نیست، F=25 خواهد بود در حالی که با انجام تبدیل لگاریتمی، با CV=0.21% که در سطح بالایی معنی دار است F=b241 خواهد شد. جمع‌پذیر نبودن اثرات تیمارها و محیط را می‌توان با استفاده از جدول تجزیه واریانس برای داده‌های اصلی دریافت. بدین منظور SSE را به دو قسمت زیر می‌توان تجزیه نمود (Steel and Torrie,1980).

الف: اثر غیر افزایشی $SS(NA)=N^2/D$. این روش به نام آزمون غیر افزایشی توکی معروف است و

$$N=\Sigma_i \Sigma_j y_{ij} d_i f_j, D=(\Sigma d_i^2)(\Sigma_j f_j^2)$$

$$d_i = \bar{y}_{io} - \bar{y}_{..},\ f_j = \bar{y}_{oj} - \bar{y}_{..}$$

ب: باقیمانده

مثال ٦: برای نشان دادن اثرات غیر افزایشی و تأثیر آن‌ها روی سطح معنی دار بودن مقایسات مربوط به تیمارها از آزمایشی با داده‌های فرضی جدول ۷ استفاده شده است:

جدول ۷ـ مشاهدات فرضی یک آزمایش بلوک‌های کامل تصادفی

Table 7. Observations of a complete randomized block designs.

Treatment تیمار	بلوک ۱	بلوک ۲	بلوک ۳
	B1	B2	B3
1	19.1	50.1	123.0
2	23.4	166.1	407.4
3	24.5	223.9	398.1
4	23.4	58.9	229.1
5	16.6	64.6	251.2

با تشکیل جدول تجزیه واریانس به طریق معمول F=3.4 غیر معنی دار خواهد بود در حالی که SS(NA)=24327 با یک درجه آزادی و باقیمانده 6280 با هفت درجه آزادی بوده و F=27.1 مربوط به تیمار در سطح احتمال ۱٪ معنی دار است.

علاوه بر مطالب فوق باید توجّه داشت که قبل از هر تجزیه آماری باید داده‌ها را بررسی نمود و داده‌های غیر عادی را شناسایی نمود. در بعضی از موارد یک مشاهده ممکن است نسبت به سایر مشاهدات غیر عادی باشد که به آن داده پرت (Outlier) گفته می‌شود. دورافتاده بودن یک داده ممکن است به چند دلیل باشد:

الف: اشتباه اندازه‌گیری

ب: اشتباه ثبت

پ: استثنایی بودن تیمار

در صورتی که دور افتاده بودن به خاطر مواد الف و ب باشد، باید سعی در تصحیح داده مورد نظر شود، در غیر این صورت می‌توان آن را داده گم شده تلقی نمود و از برآورد آن استفاده کرد. اگر دور افتاده بودن از نوع پ باشد باید داده را مثل سایر مشاهدات مستقیماً مورد استفاده قرار داد.

۳ـ پیشنهادات

به منظور بهره‌گیری از زمان و امکانات و اعتلاء سطح کیفی تحقیقات موارد زیر توصیه و پیشنهاد می‌گردند:

الف: تهیه طرح آماری مناسب با اهداف تحقیق و تبیین مدل مربوطه

ب:رعایت نکات تصادفی نمودن تیمارها

پ: اجرای آزمایش بر اساس طرح

ت: ثبت دقیق مسایل غیر مترقبه پیش آمده در آزمایش و ملحوظ داشتن آن در تجزیه داده‌ها

ث: آزمون صحت نسبی مفروضات مدل و انجام اقدامات مناسب به شرحی که بیان شد، به طوری که مفروضات مدل از صحت نسبی برخوردار شوند.

منابع مورد استفاده

References

زاهدیان، ع. ر. و ع. گرامی ۱۳۷۷. مقایسه میانگین‌های چند جامعه نرمال با واریانس‌های نابرابر. پایان نامه کارشناسی ارشد. دانشگاه تربیت مدرس.

کیانی، م و ع. گرامی، ۱۳۷۷. مدل‌های نزدیک‌ترین همسایه یک بعدی. پایان نامه کارشناسی ارشد. دانشگاه تربیت مدرس.

BARTLETT, M.S. 1938. The approximate recovery of information from field experiments with large blocks. J.Agric Sci., **28**:418-427.

BARTLETT, M.S. 1978. Nearest neighbour models in the analysis of field experiments (with) discussion. J.Roy. Statist. Soc. **40**:147-174.

BARTLETT, M.S. 1981. A further note on the use of neighbouring plot values in the analysis of field experiments. J.Roy. Ststist. Soc. **43**:100-102.

COCHRAN, W.G and COX, G.M. 1992. Experimental design. Second edition Wiley.

GERAMI, A. 1986. Behrens-fisher problem. Msc. dissertation. University of Southampton. U.K.

IPINYMOI, R.A. 1985. Equineighboured experimental designs. Ph.D Thesis. University of Southompton. U.K.

JOHNSON, R.A. and WICHERN, D.W. 1988. Applied multivariate statisttical analysis. Secound edition. Prentice-Hall International Editions.

MEAD, R., CURNOW, R.N. and HASTED, A.M. 1993. Statistical methds in agriculture and experimental Biology. Secound edition. Champan and Hall.

MEAD, R. 1994. The design of experiments. Statistical principles for fractical application. Cambridge univ. Press.

PAPADAKIS, J.S. 1937. Methode statistique pur des experiences sur. Champ. Bull. Inst. Amel. Plantes a Salonique, No. 23.

SEARLE, S.R. 1987. Linear Models for Unbalanced Data. John Wiley.

STEEL, R.G.D. & TORRIE, J.H. 1980. Principles and procedures of statistics. A biometrical approach. McGraw-Hill International Editions Statistics Series. Singapore.

WEERAHANDI, S. 1995. ANOVA under unequal error variances, biometrics **51**:589-599.