به نام خداوند بخشنده مهربان

روش‌های آماری و اقتصادسنجی

در تحلیل و مدل‌سازی داده‌های حمل و نقلی

محمد مهدی بشارتی

besharati@iut.ac.ir

فصل پنجم

توزیع های نمونه ای

Sampling distributions

مقدمه

بخش اول: مفاهیم کلی

بخش دوم: توزیع‌های نمونه‌ای

- **سوال: تفاوت احتمال با آمار در چیست؟**

- **در احتمال،**
  - توزیع پارامتر جامعه «**معلوم**» است
  - هدف: محاسبه احتمال مشاهده یک رویداد خاص

- **در آمار،**
  - مشخصات یا توزیع پارامترهای جامعه (اغلب) «**نامعلوم**» است
  - هدف: ارائه یک برآورد از پارامتر مجهول جامعه براساس داده‌های نمونه

در این فصل و فصول آینده، یاد می‌گیریم که چطور از منطق‌های احتمالاتی برای ارائه برآورد درمورد پارامترهای جامعه استفاده می شود.

یک مثال از احتمال:

- اگر بدانیم تعداد تصادفات در یک بازه زمانی معین از یک توزیع پواسون با پارامتر (λ.)، پیروی می‌کند، آنگاه می‌توان احتمال مشاهده x_0 تصادف در یک بازه زمانی دلخواه را بدست آورد.

یک مثال از آمار:

- داده‌های «سرعت تردد خودروها» در یک قطعه بزرگراهی را با نمونه‌گیری جمع‌آوری کرده باشیم و بخواهیم به کمک آن، «**میانگین واقعی** سرعت تردد خودروها» در آن قطعه (یعنی یک مشخص از جامعه) را برآورد کنیم.

از آنجا که مقادیر «سرعت تردد خودروها» در نمونه، یک **متغیر تصادفی** است؛ بنابراین می‌بایست از **گزاره‌های احتمالاتی** برای استنباط در مورد میانگین سرعت‌ها استفاده کنیم.

فصل پنجم

توزیع های نمونه ای

Sampling distributions

مقدمه

بخش اول: مفاهیم کلی

بخش دوم: توزیع‌های نمونه‌ای

نمونه. هر زیرمجموعه از جامعه، یک نمونه نامیده می‌شود.

مثلا اگر جامعه را بصورت "زمان‌سفر وسایل‌نقلیه در کمان A" در نظر بگیریم. درصورتیکه چند خودرو شناور را از این کمان عبور داده و زمان‌سفر آنها را ثبت نماییم؛ آنگاه یک نمونه چندتایی از زمان‌سفر در این کمان را خوهیم داشت.

پارامتر. مشخصه مجهول جامعه که قرار است در مورد آن استنباطی انجام شود.

مثلا اگر جامعه را بصورت "زمان‌سفر خودرها در کمان A" در نظر بگیریم؛ آنگاه، یکی از پارامترهای جامعه می‌تواند «**میانگین زمان سفر(μ) خودروهای عبوری از کمان موردنظر**» باشد.

مفاهیم کلی

مدرس: محمدمهدی بشارتی

سوال: اغلب، بدست آوردن مقدار دقیق پارامترهای جامعه ممکن نیست. پس برای شناخت جامعه، چکار باید کرد؟

- ✓ بجای مقدار دقیق پارامتر جامعه، از **برآورد** پارامتر جامعه استفاده می‌شود.

برآورد (تخمین)(Estimate). «برآورد»، تخمینی از پارامتر مورد نظر است که به کمک نمونه‌گیری بدست می‌آید.

- ✓ در مثال قبل، میانگین زمان‌سفر نمونه یک برآورد از میانگین زمان‌سفر جامعه می‌باشد.
- ✓ هر کمیتی که براساس نمونه محاسبه می‌شود، برآورد نامیده می‌شود.

تفاوت «نمونه مشاهده شده» و «نمونه تصادفی»

۱- «نمونه مشاهده شده»:

نمونه مشاهده شده فقط شامل اعدادی است که پس از نمونه‌گیری به دست می‌آیند.

۲- «نمونه تصادفی»:

مفهوم نمونه تصادفی به مرحله قبل از نمونه‌گیری مربوط می‌شود؛ و مشخص نیست که چه مقادیری را مشاهده خواهیم کرد.

پس، **نمونه تصادفی مجموعه‌ای از n متغیر تصادفی $(X_1, X_2, \dots, X_n)$ است که مستقل از هم بوده و توزیع هر یک از آن‌ها با توزیع X در جامعه آماری، یکسان است.**

یعنی هر متغیر نمونه (X_i) دقیقاً همان ویژگی‌های متغیر تصادفی X در جامعه اصلی را دارد.

تفاوت «نمونه مشاهده شده» و «نمونه تصادفی»

در حالت کلی،

برداری از متغیرهای تصادفی ($X_1, X_2, \ldots, X_n$) را داریم که مقادیر آن‌ها با متغیر تصادفی ($x_1, x_2, \ldots, x_n$) نشان داده می‌شود. اکنون، توزیع مشترک یا احتمال مشترک این نمونه تصادفی برابر است با؛

$$P(X_1=x_1, X_2=x_2, \ldots, X_n=x_n)$$

این احتمال بیانگر این است که چقدر احتمال دارد در نمونه‌گیری تصادفی، مقادیر $x_1, x_2, \ldots, x_n$ مشاهده شود.

از آنجا که این متغیرها مستقل از هم هستند، این احتمال برابر است با؛

$$P(X_1=x_1, X_2=x_2, \ldots, X_n=x_n) = P(X_1=x_1)\, P(X_2=x_2) \ldots P(X_n=x_n) = \prod_{i=1}^{n} P(X_i=x_i)$$

آماره یا برآوردگر (تابعی از متغیرهای تصادفی)

از آنجا که $X_1, X_2, \ldots, X_n$ متغیرهای تصادفی هستند، هر تابعی از این متغیرها نیز یک متغیر تصادفی است.

مثلا میانگین نمونه یک متغیر تصادفی است.

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

به هر یک از این توابع، آماره (یا برآوردگر) می‌گویند.

مثلا میانگین نمونه تصادفی یک آماره است.

برآورد (تخمین): مقدار به دست آمده از تابع برآوردگر که بعد از نمونه‌گیری به دست می‌آید.

برآوردگر(آماره)(Estimator). هر تابعی که به کمک آن، برآورد انجام می‌گردد؛ برآوردگر نامیده می‌شود.

- ✓ برای مثال، میانگین نمونه یک برآوردگر برای میانگین جامعه است.

در حقیقت توابعی که برخی از آن‌ها در فصل اول مرور شدند؛ برآوردگر نامیده شده و برای برآورد کردن پارامتر مجهول جامعه (مثلا میانگین جامعه) استفاده می‌شود.

برآورد (تخمین): مقدار به دست آمده از تابع برآوردگر که بعد از نمونه‌گیری به دست می‌آید.

بنابراین،

- ✓ **«پارامتر»** یک ویژگی جامعه است،
- ✓ **«آماره (برآوردگر)»** یک ویژگی نمونه می‌باشد. تابعی که بر روی نمونه تعریف می‌شود.
- ✓ **«برآورد (تخمین)»** مقدار به دست آمده از تابع برآوردگر که بعد از نمونه‌گیری به دست می‌آید.

		مشخصه	میانگین	واریانس	انحراف معیار	ضریب همبستگی	اندازه	نسبت
آمار توصیفی	نمونه	آماره (برآوردگر)	$\bar{X}$	$S^۲$	S	r	n	$\hat{p}$
آمار استنباطی	جامعه	پارامتر	μ	$\sigma^۲$	σ	ρ	N	p

به‌طور کلی، **برآورد** به دو صورت انجام می‌شود:

الف) برآورد نقطه‌ای(Point estimation). برآورد نقطه‌ای به کمک آماره‌ی بدست آمده از نمونه محاسبه می‌شود.

- ✓ در برآورد نقطه‌ای، ***تنها یک عدد*** بعنوان پارامتر جامعه معرفی می‌شود.
- ✓ برای مثال اگر در یک نمونه تصادفی از جامعه، میانگین نمونه برابر با ۱۲ مشاهده شود و براین اساس بگوییم که $\mu = 12$ است؛ در اینجا به کمک آماره میانگین نمونه، یک برآورد نقطه‌ای از پارامتر میانگین جامعه بدست آورده‌ایم.

سوال: نقص اساسی برآورد نقطه‌ای چیست؟

- ✓ به علت خطای نمونه‌گیری (تصادفی بودن نمونه‌گیری) نمی‌توان مطمئن شد که آماره مشاهده شده در نمونه، با دقت قابل قبولی همان پارامتر جامعه است.

به‌طور کلی، **برآورد** به دو صورت انجام می‌شود:

الف) برآورد نقطه‌ای(Point estimation). برآورد نقطه‌ای به کمک آماره‌ی بدست آمده از نمونه محاسبه می‌شود.

- ✓ در برآورد نقطه‌ای، ***تنها یک عدد*** بعنوان پارامتر جامعه معرفی می‌شود.
- ✓ برای مثال اگر در یک نمونه تصادفی از جامعه، میانگین نمونه برابر با ۱۲ مشاهده شود و براین اساس بگوییم که $\mu = 12$ است؛ در اینجا به کمک آماره میانگین نمونه، یک برآورد نقطه‌ای از پارامتر میانگین جامعه بدست آورده‌ایم.

ب) برآورد فاصله‌ای(Interval estimation).

در برآورد فاصله‌ای، ***یک بازه از مقادیر***، بعنوان مقادیر محتمل برای پارامتر جامعه ارائه می شود.

- ✓ «فاصله‌اطمینان(*Confidence Interval (CI)*)» یکی از ابزارهای اصلی برای انجام برآورد فاصله‌ای

مفاهیم کلی

فاصله‌اطمینان. یک دامنه از مقادیری است که با یک حد بالا و یک حد پایین و با یک سطح اطمینان مشخص (مثلا ۹۵٪) می‌توان گفت که پارامتر جامعه در آن بازه قرار می‌گیرد.

توضیح بیشتر در فصل بعد...

مفاهیم کلی

مدرس: محمدمهدی بشارتی

- ✓ پارامتر یک مقدار **ثابت** است؛
- ✓ آماره (برآوردگر)، یک **متغیر تصادفی** بوده (چون تابعی از متغیرهای تصادفی است) و مقدار آن از یک نمونه به نمونه دیگر متفاوت خواهد بود.
- ✓ **یک پارامتر ممکن است چند برآوردگر داشته باشد.**
- ✓ یعنی برای برآورد یک پارامتر می‌توان برآوردگرهای متعددی را معرفی کرد.
- ✓ مثلا مقدار **میانگین** یک متغیر تصادفی را می‌توان با **میانه**، **میانگین** یا **مد** آن متغیر تصادفی برآورد کرد.

اینجا ۲ سوال مطرح می‌شود؛

۱- از میان برآوردگرهای یک پارامتر، کدامیک بهتر (مناسب‌تر) است (یک برآوردگر خوب چه خواصی دارد)؟

۲- چه روش‌هایی برای به دست آوردن برآوردگرهایی که دارای خواص مناسبی باشند، وجود دارد؟

خواص یک برآوردگر مناسب

۱- از میان برآوردگرهای یک پارامتر، کدامیک بهتر است (یک برآوردگر خوب چه خواصی دارد)؟

یک برآوردگر خوب، چهار ویژگی زیر را داراست:

الف) نااریبی(*Unbiasedness*).

ب) کارایی(*Efficiency*).

ج) سازگاری(*Consistency*).

د) بسندگی(*Sufficiency*).

خواص یک برآوردگر مناسب

الف) نااریبی(*Unbiasedness*). برآوردگر $\hat{\theta}$ را برای θ یک برآوردگر نااریب می‌نامند؛ هرگاه امیدریاضی این برآوردگر، با مقدار پارامتر جامعه برابر باشد. $E(\hat{\theta}) = \theta$ یا $E(\hat{\theta}) - \theta = 0$

- ✓ برای مثال، میانگین نمونه $(\bar{X})$ یک برآوردگر نااریب برای میانگین(μ) جامعه است. $E(\bar{X}) = \mu$

نااریبی یک برآوردگر یعنی آن برآوردگر «به طور متوسط» به پارامتر موردنظر گرایش دارد.

از طرف دیگر به برآوردگری که با افزایش حجم نمونه، مقدار اریب آن کاهش یافته و به صفر میل می‌کند، «برآوردگر نااریب حدی (مجانبی)» می‌گویند.

- ✓ برای مثال، واریانس نمونه (S^2) یک برآوردگر نااریب مجانبی برای واریانس جامعه(σ^2) است. $E(S^2) = \sigma^2 - \frac{\sigma^2}{n}$
- ✓ یعنی S^2 به طور متوسط مقدار σ^2 را به اندازه $\frac{\sigma^2}{n}$ کمتر از «مقدار واقعی» برآورد می‌کند.

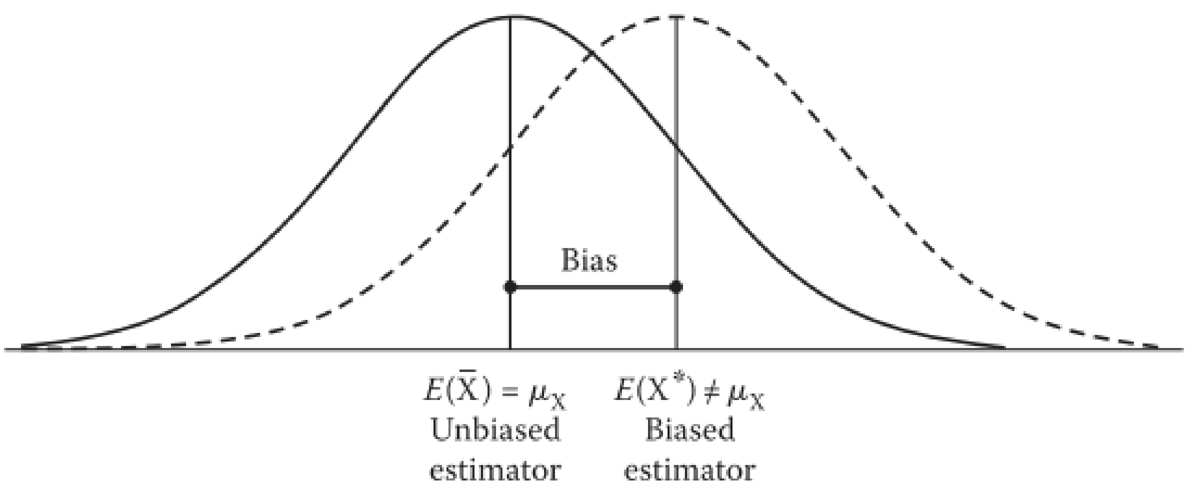
خواص یک برآوردگر مناسب

هرگونه انحراف سیستماتیک برآوردگر از پارامتر جامعه را **«انحراف/اریب/تورش(bias)»** می‌نامند.

سوال: منظور از انحراف سیستماتیک چیست؟

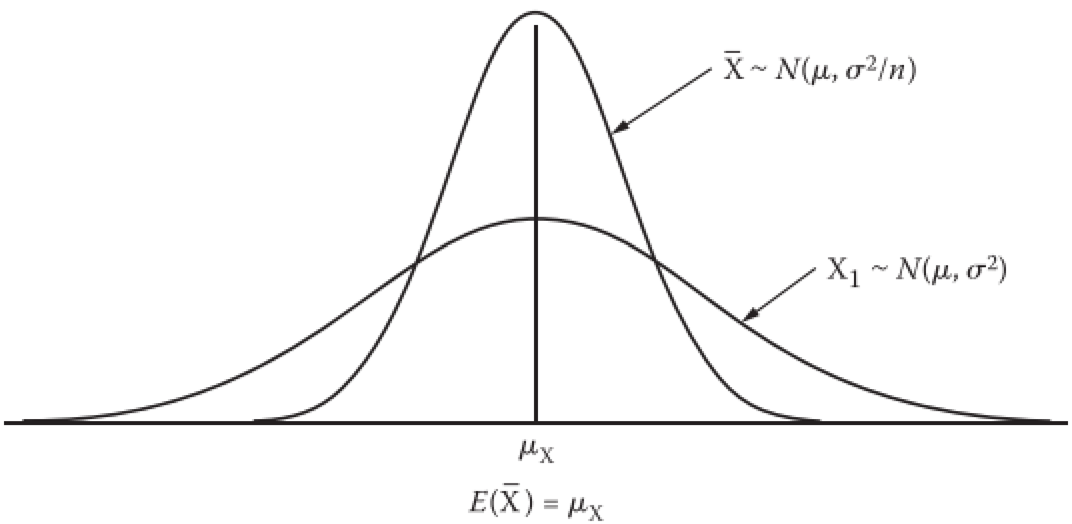
مثلا اگر یک آماره، همواره مقادیری بیشتر از پارامتر جامعه داشته‌باشد، می گوییم انحراف سیستماتیک دارد.

برای مثال، هنگامی که گفته می‌شود میانگین نمونه، برآوردگر ناریبی از میانگین جامعه است؛ بدین معناست که میانگین نمونه، گاهی اوقات مقادیر کمتر از میانگین واقعی جامعه را به‌خود می‌گیرد و برخی اوقات نیز مقادیر بدست‌آمده برای میانگین نمونه، از میانگین جامعه بیشتراست.



ب) کارایی (حداقل واریانس) (*Efficiency*).

کارایی یک ویژگی نسبی است. بدین معنا که از میان چندین برآوردگر نااریب $\hat{\theta}_1$ و $\hat{\theta}_2$ و ...، هرکدام که توزیع نمونه‌گیری آن دارای واریانس کمتری باشد؛ کارایی بیشتری خواهد داشت.



- ✓ می‌توان ثابت کرد برای یک جامعه با توزیع نرمال، میانگین نمونه ($\bar{X}$) یک برآوردگر نااریب با کمترین واریانس برای میانگین جامعه(μ) است (برابر با $\frac{\sigma^2}{n}$).
- ✓ هنگامی که دو یا چند برآوردگر نااریب موجود باشد، برای انتخاب برآوردگر برتر، از شاخص «کارایی» استفاده می‌شود.

خواص یک برآوردگر مناسب

میانگین مجذور خطا ((Mean Squared Error (MSE))

یکی دیگر از معیارهایی که برای ارزیابی و مقایسه برآوردگرها به کار می‌رود، میانگین مجذور خطا است که برابر است با «**میانگین توان دومِ انحراف $\hat{\theta}$ از θ**»

$$MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = var(\hat{\theta}) + \left[E(\hat{\theta}) - \theta\right]^2$$

جمله دوم برابر با مجذور اریب است.

بنابراین، میانگین مجذور خطا، ترکیبی از کارایی (واریانس) و اریب را منعکس می‌کند.

برای میانگین توان دوم خطاها از یک مجموعه با n داده؛

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

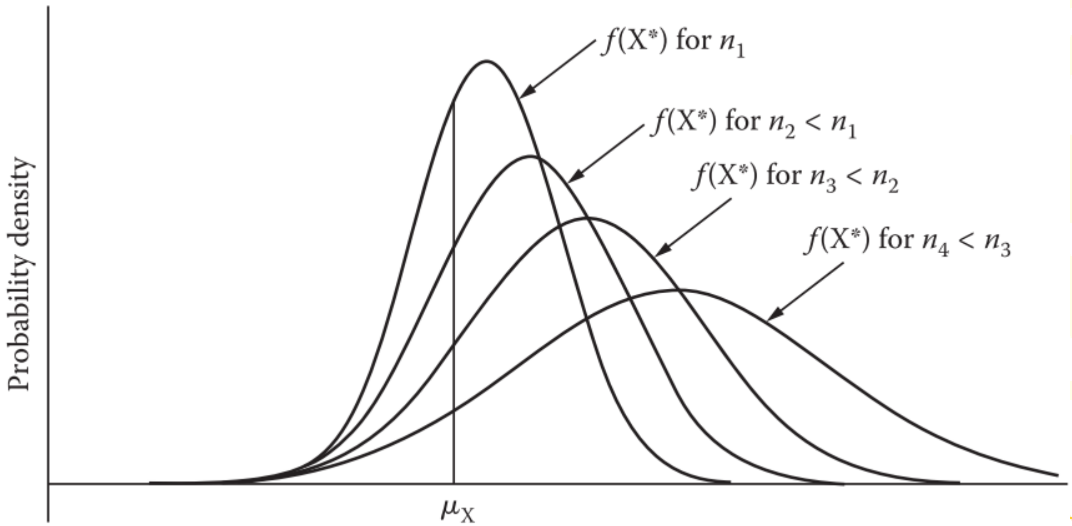
خواص یک برآوردگر مناسب

ج) سازگاری(*Consistency*). اگر برای برآوردگر $\hat{\theta}$ ، با افزایش اندازه نمونه، احتمال نزدیک شدن مقدار برآوردگر به مقدار واقعی پارامتر جامعه(θ) افزایش یابد؛ آنگاه می‌گوییم این برآوردگر دارای ویژگی «سازگاری» است.

یعنی وقتی حجم نمونه افزایش یابد، مقدار $\hat{\theta}$ کاملا به θ نزدیک خواهد شد.

اگر با افزایش اندازه نمونه واریانس $\hat{\theta}$ به سمت صفر میل کند (یعنی انحراف $\hat{\theta}$ از θ به صفر میل کند)، $\hat{\theta}$ یک برآوردگر سازگار است.

- ✓ یک برآوردگر ممکن است ناااریب نباشد، اما سازگار باشد. مثلا، واریانس نمونه (S^2) یک برآوردگر اریب اما سازگار است.



تصویر: رفتار برآوردگر فرضی X^* برای میانگین جامعه با افزایش اندازه نمونه

د) بسندگی یا کفایت (*Sufficiency*).

اگر یک برآورد تمام اطلاعات موجود درون داده‌ها درمورد پارامتر موردنظر را درون خود گنجانده باشد، آنگاه می‌گویند آن برآوردگر، «بسندگی» دارد.

یعنی $\hat{\theta}$ از تمام اطلاعات نمونه که مربوط به برآورد θ است استفاده کند.

یعنی تمام n داده موجود در نمونه را استفاده کند و نهایتاً یک مقدار به عنوان برآورد θ ارائه کند.

مثال؛ داشتن یک *نمونه n تایی برای محاسبه برآوردگر t* (مجموع نمونه) و یا داشتن مقدار *برآورد t* برای نمونه موردنظر،

- **آیا برآورد t دقیقا همان مقداری را ارائه می‌کند که با داشتن مقادیر نمونه می‌توانستیم حساب کنیم؟**
- **اگر بله؛ پس t یک برآوردگر کافی است (بسندگی دارد).**

د) بسندگی یا کفایت (*Sufficiency*).

به عبارت دیگر، توزیع مشترک X_i ها (اطلاعات نمونه) به شرط $\hat{\theta}$ (یعنی $f\left(x_1,\dots x_n\middle|\hat{\theta}\right)$) مستقل از θ باشد (یعنی در تابع $f\left(x_1,\dots x_n\middle|\hat{\theta}\right)$ پارامتر θ وجود نداشته باشد).

اگر این احتمال مستقل از θ نباشد و به θ بستگی داشته باشد، آنگاه مشاهده یک نمونه خاص بستگی به θ دارد.

اگر این احتمال مستقل از θ باشد، آنگاه نمونه خاصی مانند $(x_1,\dots x_n)$ که منجر به برآورد θ می‌شود، برای هر مقدار از θ احتمال یکسانی با سایر نمونه‌ها دارد.

$\bar{X}$ یک برآوردگر بسنده برای μ (میانگین جامعه نرمال با واریانس معلوم σ^2) است.

روش‌های به دست‌آوردن یک برآوردگر مناسب

مدرس: محمدمهدی بشارتی

۲- چه روش‌هایی برای به دست آوردن برآوردگرهایی که دارای خواص مناسبی باشند، وجود دارد؟

برای برآورد پارامترهای موردنظر، روش‌هایی وجود دارد که هریک دارای ویژگی‌هایی هستند.

دو مورد از مهمترین روش‌ها؛

الف) روش حداقل مربعات معمولی

ب) روش حداکثر درستنمایی

در واقع این روش‌ها به ما کمک می‌کنند که مناسب‌ترین برآوردگرها (و بالتبع، بهترین برآوردها) را برای پارامترهای مجهول پیدا کنیم.

روش حداقل مربعات معمولی ((Ordinary Least Squares (OLS)

به دلیل خواص مطلوبی که دارد (البته هنگامی که فروض این روش برقرار باشند) به عنوان پرکاربردترین و غالبترین روش شناخته شده است.

این روش به کارل فردریک گوس، ریاضیدان نامی آلمان نسبت داده می‌شود.

از طریق **حداقل کردن مجموع مربعات انحرافات (جملات اخلال، یا باقیمانده‌ها)** تلاش می‌کند که بهترین خط رگرسیونی برای داده‌ها را برازش نماید.

روش حداکثر درستنمایی (Maximum Likelihood Estimation (MLE))

فرض کنید X دارای توزیع معینی باشد که فقط یک پارامتر (θ) دارد.

تابع توزیع X : $P(x,\theta)$

وقتی نمونه‌گیری را انجام دهیم، مقادیر x_1، x_2، ... و x_n به دست می‌آید.

احتمال مشاهده شدنِ این نمونه چقدر است؟

$$\mathrm{P}(\mathrm{X}_1 = x_1, \dots, \mathrm{X}_n = x_n) = \mathrm{P}(\mathrm{X}_1 = x_1) \dots \mathrm{P}(\mathrm{X}_n = x_n) = \mathrm{P}(x_1, \theta) \dots \mathrm{P}(x_n, \theta)$$

مقدار این احتمال به مقدار θ بستگی دارد (مقادیر متغیرهای تصادفی $\mathrm{X}_1, \dots, \mathrm{X}_n$ مشخص شده‌اند).

بنابراین، مقدار این احتمال، تابعی از θ است. آن را با $L(\theta)$ نمایش می‌دهیم.

$L(\theta)$: تابع درستنمایی: احتمال مشاهده شدنِ نمونه موردنظر (که آن را به صورت تابعی از پارامتر (های) توزیع نمایش می‌دهیم).

تابع درستنمایی برابر است با احتمال آنکه به ازای مقادیری معین برای پارامترها، نتایج مشاهده شده حاصل شود.

روش حداکثر درستنمایی ((Maximum Likelihood Estimation (MLE)

تابع درستنمایی همان **تابع توزیع مشترک** است.

$$L(\theta) = f(x_1, \theta) \ldots \mathrm{f}(x_n, \theta) = \prod_{i=1}^{n} \mathrm{f}(x_i, \theta)$$

روش حداکثر درستنمایی ((Maximum Likelihood Estimation (MLE))

سوال: از بین تمام مقادیری که می‌توانستند نمایان شوند، چرا مقادیر x_1، x_2، ... و x_n مشاهده شدند؟

پاسخ: چون این مشاهدات، بیشترین احتمال یا شانس را برای نمایان شدن داشته‌اند.

پس $L(\theta)$ باید بیانگر حداکثر احتمال باشد.

بنابراین، مقدار θ که برحسب مشاهدات نمونه تعیین می‌شود، باید به گونه‌ای باشد که احتمال مذکور ($L(\theta)$) را حداکثر نماید.

پس برآوردگر حداکثر درستنمایی از طریق حداکثر شدنِ تابع درستنمایی نسبت به پارامترهای مجهول (ضرایب θ) به دست می‌آید.

برای حداکثر شدن $L(\theta)$ شرط زیر را به کار می‌بریم:

$$\frac{\partial L}{\partial \theta} = 0 \quad \rightarrow \quad \hat{\theta}_{ML} = g(x_1, \ldots x_n)$$

$\hat{\theta}_{ML}$: برآوردگر حداکثر درستنمایی نامیده می‌شود.

نکته: معمولاً بهتر است ابتدا از $L(\theta)$ لگاریتم گرفته و سپس نسبت به θ مشتق بگیریم.

روش حداکثر درستنمایی (Maximum Likelihood Estimation (MLE))

سوال: فرض کنید X دارای توزیع برنولی است. پارامتر p این توزیع را با روش حداکثر درستنمایی برآورد کنید؟

$$P(x) = p^x(1-p)^{1-x}$$

پاسخ:

اگر یک نمونه n تایی از این جامعه انتخاب شود، داریم؛

$$L(p) = P(x_1, p) \ldots P(x_n, p) = p^{x_1}(1-p)^{1-x_1} \ldots p^{x_n}(1-p)^{1-x_n} = p^{\sum x_i}(1-p)^{n-\sum x_i}$$

$$\ln L(p) = \sum x_i \ln p + \left(n - \sum x_i\right) \ln(1-p)$$

$$\frac{\partial \ln L}{\partial p} = \sum x_i \frac{1}{p} + \left(n - \sum x_i\right)\frac{-1}{1-p} = 0 \quad \rightarrow \quad \hat{p}_{ML} = \frac{\sum x_i}{n} = \bar{x}$$

برآوردگر حداکثر درستنمایی پارامتر p :

بنابراین، هنگامی که پارامتر p برابر با $\hat{p} = \frac{\sum x_i}{n}$ یا همان میانگین $(\bar{x})$ باشد، تابع $L(p)$ حداکثر می‌شود.

روش حداکثر درستنمایی (Maximum Likelihood Estimation (MLE))

سوال: فرض کنید X دارای توزیع نرمال است. پارامترهای μ و σ^2 این توزیع را با روش حداکثر درستنمایی برآورد کنید؟

$$f(x;\mu,\sigma^2)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

پاسخ:

تابع درستنمایی؛

$$L(\mu,\sigma^2)=f(x_1,\mu,\sigma^2)\ldots f(x_n,\mu,\sigma^2)=\frac{1}{(\sqrt{2\pi})^n\sigma^n}e^{-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}}$$

$$\ln L(\mu,\sigma^2)=-n\ln\sqrt{2\pi}-\frac{n}{2}\ln\sigma^2-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}$$

روش حداکثر درستنمایی ((Maximum Likelihood Estimation (MLE)

سوال: فرض کنید X دارای توزیع نرمال است. پارامترهای μ و σ^2 این توزیع را با روش حداکثر درستنمایی برآورد کنید؟

$$f(x;\mu,\sigma^2)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

پاسخ:

مشتق‌گیری نسبت به پارامترها؛

$$\frac{\partial \ln L}{\partial \mu}=0 \qquad , \qquad \frac{\partial \ln L}{\partial \sigma^2}=0$$

با حل معادلات بالا، برآوردگرهای پارامترهای این متغیر تصادفی را به روش حداکثر درستنمایی به دست آورده‌ایم؛

$$\hat{\mu}_{ML}=\frac{\sum x_i}{n}=\bar{x} \qquad , \qquad \hat{\sigma}^2{}_{ML}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}$$

روش حداکثر درستنمایی (Maximum Likelihood Estimation (MLE))

دو نکته:

- تابع درستنمایی بعد از مشخص شدن مقادیر x_i **فقط تابعی از ضرایب θ** است.
- تعداد پارامترها به هر میزان که باشد، تابع درستنمایی، تابعی از همه پارامترهاست و لازم است نسبت به همه پارامترها مشتق جزئی گرفته و برآوردگر هرکدام از آن پارامترها را به دست آورد.
- ممکن است تابع درستنمایی دارای یک «فرم بسته» نباشد. در این شرایط نمی‌توان معادلات مشتق جزئی را حل کرد.

فصل پنجم

توزیع های نمونه ای

Sampling distributions

مقدمه

بخش اول: مفاهیم کلی

بخش دوم: توزیع‌های نمونه‌ای

توزیع‌های نمونه‌ای

مدرس: محمدمهدی بشارتی

- ✓ مقادیر به‌دست‌آمده برای یک برآوردگر در نمونه‌گیری‌های مختلف با یکدیگر متفاوت است.
- ✓ بنابراین، هر برآوردگر، خود یک متغیر تصادفی است.
- ✓ توزیع احتمال مقـادیری که یک برآوردگر می‌تواند داشته باشد را توزیع نمونه‌ای آن برآوردگر می‌نامند.
- ✓ **توزیـع نمونه‌ای (یا توزیع نمونه‌گیری):** نمودار فراوانی مقادیر مشاهده‌شده برای یک آماره.
- ✓ **توزیـع نمونه‌ای (یا توزیع نمونه‌گیری): توزیع احتمال مقادیر موردانتظار برای یک آماره.**

- ✓ ***میانگین نمونه*** یکی از برآوردگرهای پرکاربرد است و به توزیع نمونه‌ای آن، «توزیع نمونه‌ای میانگین» گفته می‌شود.
- ✓ واضح است که مقدار میانگین نمونه، از یک نمونه به نمونه دیگر، متفاوت است.
- ✓ اگر توزیع مقادیر میـانگین نمونه‌هـای مختلف استخراج‌شده از یک جـامعه را در یک نمودار ترسیم کنیم؛ توزیع نمونه‌ای میانگین آن جامعه را به دست خواهیم آورد.

مثال؛

یک جامعه کوچک و متناهی شامل ۵ داده سرعت به‌دست‌آمده از ۵ خودرو را در نظر بگیرید. مقادیر میانگین (μ) و واریانس (σ^2) این جامعه

((**چون** داده‌های مربوط به تمام اعضای جامعه را برداشت کرده‌ایم، توانستیم میانگین و واریانس جامعه را محاسبه نماییم))

شماره خودرو	۱	۲	۳	۴	۵
سرعت (X)	۶۰	۵۷	۶۶	۶۱	۵۴
احتمال (X)	۰/۲	۰/۲	۰/۲	۰/۲	۰/۲

$$\mu=\frac{۶۰+۵۷+۶۶+۶۱+۵۴}{۵}=۵۹/۶$$

$$\sigma^{۲}=\frac{\sum_{i=۱}^{۵}(x_i-\mu)^{۲}}{N}=۱۶/۲۴$$

مثال؛

فرض کنید به‌جای برداشت داده‌های کل جامعه، تنها می‌توانیم نمونه‌های دوتایی (بدون جایگذاری) از جامعه برداشت کنیم.
بر این اساس، به ۱۰ شکل مختلف می‌توان نمونه‌هایی ۲ تایی از این جامعه برداشت کرد.

نمونه‌های برداشت شده به همراه میانگین متناظر با هر نمونه:

شماره نمونه (i)	خودروهای نمونه‌گیری شده	سرعت خودروها (X_{i1}، X_{i2})	میانگین نمونه ($\bar{X}_i$)
۱	۱،۲	۵۷،۶۰	۵۸/۵
۲	۱،۳	۶۶،۶۰	۶۳
۳	۱،۴	۶۱،۶۰	۶۰/۵
۴	۱،۵	۵۴،۶۰	۵۷
۵	۲،۳	۶۶،۵۷	۶۱/۵
۶	۲،۴	۶۱،۵۷	۵۹
۷	۲،۵	۵۴،۵۷	۵۵/۵
۸	۳،۴	۶۱،۶۶	۶۳/۵
۹	۳،۵	۵۴،۶۶	۶۰
۱۰	۴،۵	۵۴،۶۱	۵۷/۵

مثال؛

با داشتن داده‌های مربوط به «میانگین نمونه‌ها» و نیز احتمال‌های متناظر با هرکدام (که در این مثال برابر با ۰.۱ است)؛ می‌توان توزیع نمونه‌ای میانگین ($\bar{X}$) را به‌صورت جدول زیر ارائه کرد.

$\bar{x}$	۵۸/۵	۶۳	۶۰/۵	۵۷	۶۱/۵	۵۹	۵۵/۵	۶۳/۵	۶۰	۵۷/۵
احتمال ($\bar{x}$)	۰/۱	۰/۱	۰/۱	۰/۱	۰/۱	۰/۱	۰/۱	۰/۱	۰/۱	۰/۱

شماره نمونه (i)	خودروهای نمونه‌گیری شده	سرعت خودروها (x_{i1}, x_{i2})	میانگین نمونه ($\bar{x}_i$)
۱	۱،۲	۵۷،۶۰	۵۸/۵
۲	۱،۳	۶۶،۶۰	۶۳
۳	۱،۴	۶۱،۶۰	۶۰/۵
۴	۱،۵	۵۴،۶۰	۵۷
۵	۲،۳	۶۶،۵۷	۶۱/۵
۶	۲،۴	۶۱،۵۷	۵۹
۷	۲،۵	۵۴،۵۷	۵۵/۵
۸	۳،۴	۶۱،۶۶	۶۳/۵
۹	۳،۵	۵۴،۶۶	۶۰
۱۰	۴،۵	۵۴،۶۱	۵۷/۵

$$\mu_{\bar{X}} = \sum_{i=۱}^{۱۰} \bar{x}_i \left(\frac{۱}{۱۰}\right) = ۵۹/۶$$

$$\sigma^۲_{\bar{X}} = \frac{\sum_{i=۱}^{۱۰} (x_i - \bar{x}_i)^۲}{N} = ۶/۰۹$$

با مقایسه مقادیر میانگین و واریانس جامعه، با میانگین و واریانس توزیع نمونه‌ای $\bar{X}$ می‌توان نتیجه گرفت که:

۱. مقدار $\bar{X}$ از نمونه‌ای به نمونه دیگر، متغیر است (یک متغیر تصادفی است) و لزوماً با میانگین جامعه (μ) برابر نیست.

۲. امید ریاضی توزیع نمونه‌ای $\bar{X}$ (که به‌صورت $\mu_{\bar{X}}$ نمایش داده‌شده)، با میانگین جامعه برابر است.

۳. واریانس توزیع نمونه‌ای $\bar{X}$ (که به‌صورت $\sigma^{۲}_{\bar{X}}$ نمایش داده‌شده)، کمتر از واریانس جامعه است.

توزیع‌های نمونه‌ای

مدرس: محمدمهدی بشارتی

$$\mu_{\bar{X}}=\mu$$

برای یک جامعه نامتناهی

$$\sigma_{\bar{X}}^{2}=\frac{\sigma^{2}}{n}$$

در حالتی که جامعه متناهی باشد

$$\sigma_{\bar{X}}^{2}=\frac{\sigma^{2}}{n}\left(\frac{N-n}{N-1}\right)$$

N اندازه جامعه و n اندازه نمونه می‌باشد.

ضریب $\left(\frac{N-n}{N-1}\right)$ تحت عنوان ضریب تصحیح جامعه متناهی شناخته می‌شود.

در شرایطی که اندازه جامعه (N) بسیار بزرگ‌تر از اندازه نمونه باشد؛ (مثلاً هنگامی‌که $\left(\frac{n}{N}\right)$ کمتر از ۵٪ یا ۱۰٪ باشد) می‌توان ضریب تصحیح جامعه متناهی را برابر با ۱ در نظر گرفت.

در این مثال، اندازه جامعه برابر با ۵ و اندازه نمونه‌ها برابر با ۲ می‌باشد؛ بنابراین،

$$\left(\frac{N-n}{N-1}\right)=\left(\frac{5-2}{5-1}\right)=0.75$$

با توجه به روابط قبلی، با داشتن میانگین و واریانس جامعه، می‌توان میانگین و واریانس توزیع نمونه‌ای میانگین را به دست آورد:

$$\mu_{\bar{X}}=\mu=59.6$$

$$\sigma_{\bar{X}}^{2}=\frac{\sigma^{2}}{n}\times\left(\frac{N-n}{N-1}\right)=\frac{16.24}{2}\times(0.75)=6.09$$

دو ویژگی مهم توزیع نمونه‌ای میانگین:

۱– فارغ از این‌که توزیع جامعه به چه شکل است؛ درصورتی‌که یک نمونه تصادفی جمع‌آوری کرده و میانگین نمونه را حساب نماییم و این کار را تکرار کنیم؛ آنگاه، مقادیر میانگین نمونه‌ها حول میانگین جامعه متمرکز خواهنــد شد.

همچنین، واریـانس $\bar{X}$ مقداری کمتر از واریـانس جـامعه خواهـد داشت (بـا ضریب کاهش $\frac{\sigma^2}{n}$ برای جامعه نامتناهی و $\frac{\sigma^2}{n}(\frac{N-n}{N-1})$ برای جامعه متناهی).

۲– هرچه اندازه نمونه‌ها (n) بزرگ‌تر باشد، تغییرات مشاهده‌شده در توزیع نمونه‌ای کمتر خواهد بود.

به‌عبارت‌دیگر، مقدار میانگین نمونه‌ای که بر اساس یک نمونه با اندازه بزرگ‌تر به‌دست‌آمده باشد، نسبت به میانگینی که از یک نمونه با اندازه کوچک‌تر به‌دست‌آمده باشد، به میانگین واقعی جامعه نزدیک‌تر است.

این‌که انتظار داریم با افزایش اندازه نمونه، برآورد به‌دست‌آمده از پارامتر موردنظر، به مقدار واقعی آن نزدیک‌تر شود؛ یک ویژگی خوب برای نمونه‌گیری تصادفی به‌حساب می‌آید (**ویژگی سازگاری؟**).

دو نکته مهم در مورد رابطه زیر:

$$\sigma_{\bar{X}}^{2}=\frac{\sigma^{2}}{n}$$

1. این رابطه همواره صحیح نیست. گاهی اوقات ممکن است یک نمونه تصادفی ۱۰ تایی، نسبت به یک نمونه ۱۰۰ تایی، برآورد نزدیک‌تری برای میانگین جامعه ارائه کند. اما هیچ راهی برای فهمیدن این‌که کدام برآورد به پارامتر جامعه نزدیک‌تر است وجود ندارد (زیرا اگر ما مقدار پارامتر جامعه را می‌دانستیم به سراغ نمونه‌گیری نمی‌رفتیم).

2. درحالی‌که با افزایش اندازه نمونه، واریانس توزیع نمونه‌ای کاهش می‌یابد (و به این‌ترتیب، احتمال این‌که برآورد به‌دست‌آمده، به مقدار واقعی میانگین نزدیک‌تر باشد، افزایش می‌یابد)؛ اما این افزایش، نرخ کاهشی دارد. به عبارت ساده‌تر، سود حاشیه‌ایِ اضافه نمودن ۱۰ مشاهده جدید به ۱۰ مشاهده قبلی، بیشتر از زمانی است که بخواهیم ۱۰ مشاهده را به ۱۰۰ مشاهده قبلی اضافه کنیم.

در مسائل عملی، ما تنها یک نمونه تصادفی با اندازه n (و نه چندین نمونه nتایی) از جامعه برداشت می‌کنیم.

بدین ترتیب، میانگین نمونه برداشت‌شده، به‌عنوان برآوردی از میانگین مجهول جامعه مورداستفاده قرار خواهد گرفت.

سؤال: میانگین به‌دست‌آمده از نمونه، چقدر به میانگین واقعی جامعه نزدیک است؟

برای کمّی کردن میزان نزدیکی میانگین نمونه به میانگین مجهول جامعه، می‌توان میزان این نزدیکی را در قالب یک گزاره احتمالاتی بیان کرد. برای مثال، احتمال این‌که مقدار میانگین یک نمونه ($\bar{X}$) با اندازه n در یک بازه $\pm E$ از میانگین واقعی جامعه قرار داشته باشد، چقدر است؟ (E همان خطای قابل‌قبول می‌باشد).

درصورتی‌که توزیع نمونه‌ای میانگین ($\bar{X}$) را بدانیم (مانند آنچه در **مثال قبل** نشان داده شد)، آنگاه می‌توان به این سؤال پاسخ داد.

ازآنجاکه اندازه جامعه (N) معمولاً نامتناهی یا بسیار بزرگ می‌باشد، ساختنِ توزیع نمونه‌ای غیرعملی است.

نظریه‌ها و قواعدی در مورد توزیع نمونه‌ای چند آماره معروف (مثلاً میانگین نمونه $\bar{X}$) وجود دارد که به ما کمک می‌کند بدون ساختن توزیع احتمالاتی آن آماره‌ها، احتمال‌های متناظر با مقادیر مختلف آن‌ها را به دست آوریم.

توزیع نمونه‌ای میانگین

به‌طورکلی ۳ حالت زیر وجود دارد:

الف) اگر $\bar{X}$ میانگین یک نمونه تصادفی n تایی از **یک جامعه نرمال با میانگین μ و واریانس σ^2** باشد؛ آنگاه، صرف‌نظر از اندازه نمونه، توزیع نمونه‌ای میانگین‌ها ($\bar{X}$) نیز از توزیع نرمال به‌صورت زیر پیروی خواهد کرد.

$$\bar{X} \sim N(\mu , \sigma^2 / n)$$

یا به‌عبارت‌دیگر، توزیع نمونه‌ای $\bar{X}$ استانداردشده، از توزیع نرمال استاندارد پیروی می‌کند

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0 , 1)$$

در بسیاری از مسائل عملی، **واریانس جامعه مجهول است**. برای حل این مشکل، انحراف معیار نمونه را جایگزین انحراف معیار جامعه در رابطه بالا می‌کنند.

مطالعات پیشین نشان داده که این جایگذاری موجب می‌شود که توزیع حاصله، نسبت به توزیع نرمال استاندارد، پراکندگی بیشتری داشته باشد. در این حالت، توزیع نمونه‌ای آماره زیر، از توزیع t با $n-1$ درجه آزادی پیروی می‌کند:

$$T = \frac{(\bar{X} - \mu)}{(S/\sqrt{n})} \sim t_{n-1}$$

توزیع نمونه‌ای میانگین

به‌طورکلی ۳ حالت زیر وجود دارد:

ب) **هنگامی‌که اندازه نمونه (n) به‌قدر کافی بزرگ باشد**. در این حالت از یک قضیه مشهور در علم آمار به‌نام قضیه حد مرکزی استفاده می‌شود. این قضیه در مورد توزیع نمونه‌ای $\bar{X}$ فارغ از توزیع حاکم بر جامعه بیان می‌کند که:

اگر $\bar{X}$ میانگین یک نمونه تصادفی با اندازه n باشد که از جامعه‌ای با میانگین (μ) و واریانس (σ^2) استخراج‌شده باشد؛ آنگاه، با افزایش اندازه نمونه (n)، توزیع $\bar{X}$ به یک توزیع نرمال با میانگین $\mu_{\bar{X}}=\mu$ و واریانس $\sigma_{\bar{X}}^2=\sigma^2/n$ میل می‌کند.

$$\bar{X} \sim N(\mu , \sigma^2/n)$$

به‌عبارت‌دیگر، توزیع نمونه‌ای $\bar{X}$ استانداردشده ($\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$)، از توزیع نرمال استاندارد پیروی می‌کند.

$$Z=\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0 , 1)$$

توزیع نمونه‌ای میانگین

به‌طورکلی ۳ حالت زیر وجود دارد:

بدین ترتیب، با استفاده از قضیه حد مرکزی، می‌توان گفت درصورتی‌که اندازه نمونه n به اندازه کافی بزرگ باشد، آنگاه می-توان توزیع نمونه‌ای $\bar{X}$ را با یک توزیع نرمال تقریب زد. به‌طورکلی می‌توان گفت، این گزاره (در حالتی که توزیع جامعه چولگی شدیدی نداشته باشد) برای $n \geq ۳۰$ قابل‌قبول است.

در صورت وجود چولگی شدید در توزیع جامعه، نمونه‌ای با اندازه به‌مراتب بزرگ‌تر از ۳۰ موردنیاز است تا بتوان گفت قضیه حد مرکزی برقرار است.

توزیع نمونه‌ای میانگین

به‌طورکلی ۳ حالت زیر وجود دارد:

ج) هنگامی‌که توزیع جامعه غیر نرمال بوده و اندازه نمونه نیز کوچک است.

برای این حالت، یک قضیه کلی که بتوان به کمک آن، توزیع نمونه‌ای $\bar{X}$ را توضیح داد، وجود ندارد.

زیرا در این حالت، توزیع نمونه‌ای $\bar{X}$ به توزیع جامعه بستگی دارد. در این شرایط، نمی‌توان به‌راحتی در مورد کیفیت برآورد انجام‌شده، اظهارنظر کرد.

برای حل این مشکل، از **روش‌های ناپارامتری** استفاده می‌شود.

توزیع نمونه‌ای میانگین

در ادامه و در قالب یک مثال، به‌صورت اجمالی خواهیم دید که چگونه می‌توان تنها با انجام یک مرتبه نمونه‌گیری و استفاده از مفهوم توزیع نمونه‌ای، در مورد مقدار تقریبی میانگین واقعی جامعه اظهارنظر کرد.

فرض کنید توزیع نمونه‌ای میانگین یک جامعه، به صورت یک توزیع نرمال با میانگین مجهول μ و واریانس معلوم σ^2 باشد. (حالت الف از سه حالت تشریح شده در قسمت قبل)

اگر تنها یک‌بار نمونه‌گیری کنیم و میانگین نمونه برابر $\bar{x}$ شود؛

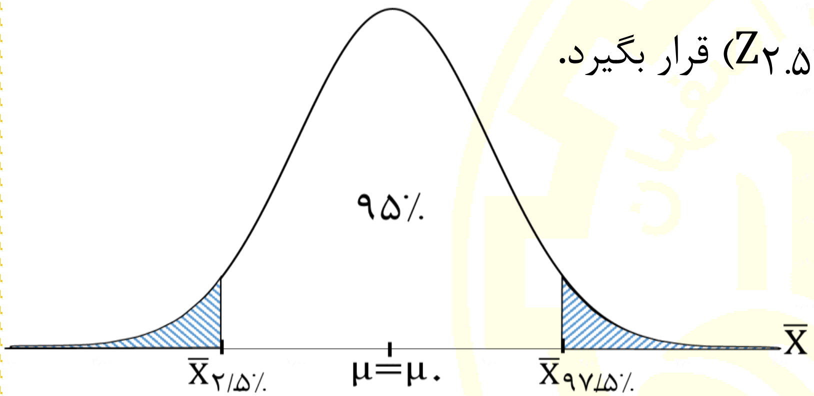
آنگاه با داشتن خطای استاندارد میانگین (به‌صورت $\sigma/\sqrt{n}$)، به دو شکل می‌توان در مورد میانگین مجهول جامعه (μ) استنباط انجام داد؛

1. **حالت اول. اظهارنظر در مورد میزان نزدیکیِ یک مقدارِ مفروض به مقدار واقعی میانگین جامعه.**
2. **حالت دوم. ارائه بازه‌ای از مقادیر که با یک درصد اطمینان مشخص، شامل میانگین جامعه است.**

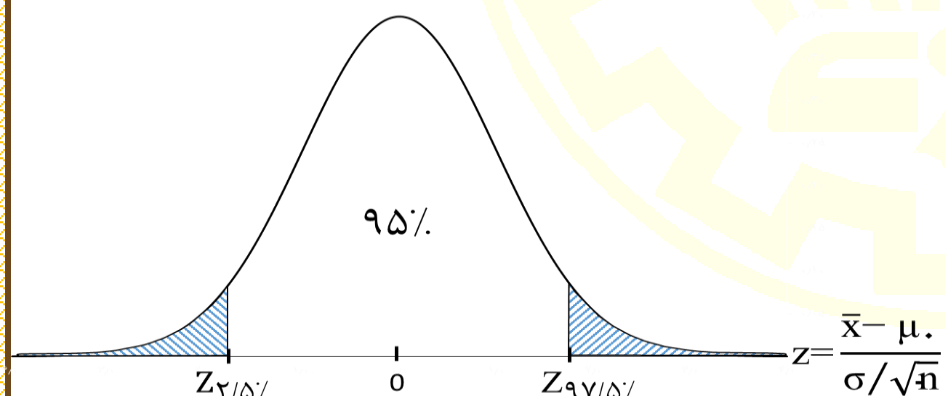
توزیع نمونه‌ای میانگین

حالت اول. اظهارنظر در مورد میزان نزدیکیِ یک مقدارِ مفروض به مقدار واقعی میانگین جامعه.

می‌توانیم بگوییم اگر مقدار مفروض برای میانگین جامعه (μ_0)، صحیح باشد؛ آنگاه، اگر یک نمونه تصادفی با میانگین $\bar{x}$ از جامعه برداشت شده باشد؛ در ۹۵٪ اوقات انتظار داریم مقدار آماره $z=\dfrac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$ در بازه ($Z_{۲.۵\%}$, $Z_{۹۷.۵\%}$) قرار بگیرد.



به‌عبارت‌دیگر، اگر نمونه‌ای تصادفی از جامعه‌ای با میانگین μ انتخاب شده باشد؛ احتمال این‌که میانگین نمونه مذکور در قسمت ۲.۵٪ ابتدا یا انتهای توزیع نمونه‌ای قرار بگیرد، ۰.۰۵ است.



پس اگر مقدار آماره موردنظر، در این بازه قرار نگرفت؛ بدین معناست که؛

۱) یا یک رویداد نادر اتفاق افتاده (در این حالت، با احتمال ۰.۰۵) و

۲) یا این‌که مقدار فرض شده برای میانگین جامعه (یعنی μ_0)، با مقدار واقعی میانگین جامعه (μ) تفاوت دارد.

توزیع نمونه‌ای میانگین

حالت دوم. ارائه بازه‌ای از مقادیر که با یک درصد اطمینان مشخص، شامل میانگین جامعه است.

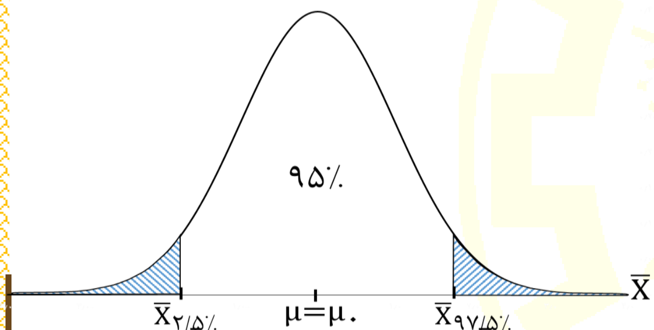
می‌دانیم که میانگین نمونه‌ها یک متغیر تصادفی است.

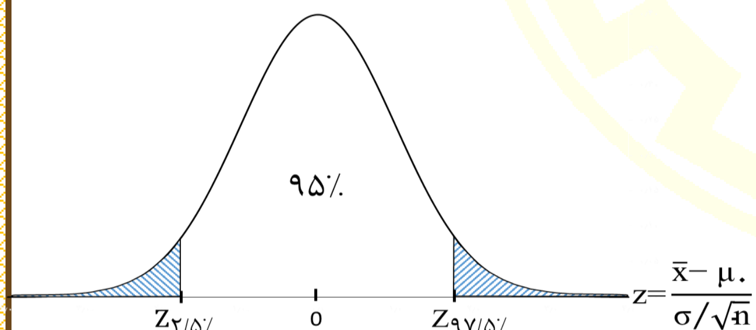
اگر توزیع نمونه‌ای میانگین نمونه‌ها را داشته باشیم، می‌توانیم بازه‌ای از مقادیر را ارائه کنیم که با یک درصد اطمینان مشخص، شامل میانگین جامعه است (Confidence Interval).

توزیع نمونه‌ای میانگین

حالت دوم. ارائه بازه‌ای از مقادیر که با یک درصد اطمینان مشخص، شامل میانگین جامعه است.

مثلا، اگر توزیع مقادیر میانگین نمونه‌ها، به صورت توزیع نرمال با میانگین μ و واریانس $\sigma^{۲}/n$ باشد؛ آنگاه، میانگین به‌دست‌آمده از یک‌بار نمونه‌گیری تصادفی، با احتمال ۹۵٪ در بازه ($\bar{X}_{\%97.5}$، $\bar{X}_{\%2.5}$) از توزیع نمونه‌ای میانگین، قرار خواهد گرفت. یا به‌عبارت‌دیگر، اگر یک نمونه تصادفی از جامعه برداشت شود (و میانگین آن برابر با $\bar{x}$ به دست آید)، در ۹۵٪ اوقات انتظار داریم مقدار آماره $\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ در بازه ($Z_{\%97.5}$ ، $Z_{\%2.5}$) قرار بگیرد.





تحلیل‌گر از این‌که مقدار میانگین به‌دست‌آمده از نمونه تصادفی، کدام‌یک از مقادیر $\bar{X}_1$، $\bar{X}_2$، $\bar{X}_{\%2.5}$، $\bar{X}_{\%99}$ و یا ... می‌باشد؛ اطلاعی ندارد.
بنابراین، با در نظر گرفتن کمترین و بیشترین مقدار محتمل، بازه‌ای به‌صورت زیر ارائه کرده و با ۹۵٪ اطمینان می‌گوییم این بازه شامل میانگین واقعی جامعه است؛

$$Z_{\%2.5} < \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} < Z_{\%97.5}$$

$$\left[\bar{x}-|Z_{\%2.5}|\times\sigma/\sqrt{n} < \mu < \bar{x}+|Z_{\%97.5}|\times\sigma/\sqrt{n}\right]$$

توزیع نمونه‌ای واریانس یک نمونه

هنگامی‌که داده‌ها به‌عنوان یک نمونه تصادفی از جامعه هستند و واریانس جامعه (σ^2) مجهول است؛ واریانس نمونه (S^2) می‌تواند برای ارائه برآوردی از واریانس جامعه (σ^2)، مورداستفاده قرار گیرد.

اما واریانس نمونه نیز مانند میانگین آن، برای نمونه‌های مختلف می‌تواند مقادیر متفاوتی داشته باشد.

بنابراین، واریانس نمونه‌ها (S^2) نیز یک متغیر تصادفی است و به توزیع مقادیر آن، **توزیع نمونه‌ای واریانس** گفته می‌شود.

اما توزیع نمونه‌ای S^2 در آمار کاربرد اندکی دارد. به‌جای آن، نسبت زیر، یک متغیر تصادفی پرکاربرد به‌حساب می‌آید.

اگر S^2، واریانس به‌دست‌آمده از یک نمونه تصادفی n تاییِ انتخاب‌شده از یک جامعه نرمال با واریانس σ^2 باشد؛ آنگاه، آماره زیر از **توزیع کای‌دو با $n-1$ درجه آزادی** پیروی خواهد کرد؛

$$\chi^2=\frac{(n-1)s^2}{\sigma^2}=\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sigma^2}\sim\chi^2_{n-1}$$

- توجه: این رابطه در صورتی برقرار است که جامعه، توزیع نرمال داشته باشند.
- اگر جامعه توزیع نرمال نداشته باشد، از روش‌های غیرپارامتری استفاده می‌شود.
- از رابطه اخیر، به‌عنوان پایه‌ای برای انجام استنباط در مورد واریانس مجهول جامعه استفاده می‌شود.

توزیع نمونه‌ای نسبت در یک نمونه

اگر داده‌ها از نوع طبقه‌ای (گسسته) باشند، می‌توان نسبت‌هایی از جامعه را برآورد کرد.

برای مثال، یک جامعه شامل تمام مسافرینی که از مترو تهران استفاده می‌کنند را در نظر گرفته و متغیر تصادفی X را به‌صورت «جنسیت مسافری که از این شیوه حمل‌ونقلی استفاده می‌کند»، تعریف می‌کنیم:

$$X_i=\begin{cases} 0 & \text{اگر مسافر مذکر باشد} \\ ۱ & \text{اگر مسافر مؤنث باشد} \end{cases}$$

فرض کنید می‌خواهیم بدانیم چه نسبتی از مسافرینی که از مترو استفاده می‌کنند، مؤنث هستند؟ بنابراین، پارامتر p برابر خواهد بود با نسبت زن‌ها از کل افرادی که از مترو استفاده می‌کنند.

همانند میانگین و واریانس نمونه، مقدار یک نسبت در یک نمونه ($\hat{p}$) نیز از نمونه‌ای به نمونه دیگر متفاوت می‌باشد.

این تغییرپذیری سبب می‌شود مفهوم توزیع نمونه‌ای، برای برآوردگر $\hat{p}$ نیز قابل‌استفاده باشد.

توزیع نمونه‌ای نسبت در یک نمونه

فرض کنید یک نمونه تصادفی با اندازه n به صورت ($X_۱$، $X_۲$،...،X_n) برای برآورد کردن مقدار p جمع‌آوری شده است.

تعداد موفقیت‌ها در n آزمایش برنولی را می‌توان با توزیع دوجمله‌ای مدل کرد.

بنابراین، تعداد زن‌های مشاهده‌شده در نمونه تصادفی n تایی از مسافران، از یک توزیع دوجمله‌ای به‌صورت زیر پیروی می‌کند:

$$\sum_{i=۱}^{n} X_i \sim B(n,p)$$

به‌علاوه، نسبت نمونه $\hat{p}$ از تقسیم کردن تعداد مسافرین زن درون نمونه، بر اندازه نمونه (n) به شکل زیر به دست می‌آید

$$\hat{p}= \frac{\sum_{i=۱}^{n} X_i}{n}$$

همان‌طور که مشاهده می‌شود، شکل رابطه نسبت موفقیت در یک نمونه را می‌توان مشابه با میانگین نمونه در نظر گرفت.

می‌توان ثابت کرد که میانگین و واریانس توزیع نمونه‌ای $\hat{p}$ برابر است با

$$E(\hat{p})=\mu_{\hat{p}}=p$$

$$\sigma_{\hat{p}}^{۲}=Var(\hat{p})=\frac{p(1-p)}{n}$$

توزیع نمونه‌ای نسبت در یک نمونه

به‌طورکلی، در ارتباط با توزیع نمونه‌ای $\hat{p}$، یکی از سه حالت زیر ممکن است رخ دهد:

الف) اگر نمونه‌ها از **جامعه‌ای متناهی** برداشت شده باشند؛ توزیع نمونه‌ای $\hat{p}$، از **توزیع فوق هندسی** پیروی می‌کند.

ب) اگر نمونه‌ها از **جامعه‌ای نامتناهی** برداشت شده باشند؛ توزیع نمونه‌ای $\hat{p}$، از **توزیع دوجمله‌ای** پیروی می‌کند.

ج) به‌عنوان یک قاعده کلی، اگر دو شرط $n\hat{p}\geq ۵$ و $n(۱-\hat{p})\geq ۵$ برقرار باشد؛ آنگاه می‌توان توزیع نمونه‌ای آماره $\hat{p}$ را با یک **توزیع نرمال با میانگین $\mu_{\hat{p}}$ و واریانس $\sigma^{۲}_{\hat{p}}$** تقریب زد.

$$\hat{p} \sim N\left(p\,,\, \frac{p(1-p)}{n}\right)$$

بنابراین، قضیه حد مرکزی برای توزیع نمونه‌گیری $\hat{p}$ برقرار خواهد بود.

به‌علاوه، درصورتی‌که آماره Z را به‌صورت زیر تعریف کنیم، آنگاه، این آماره از توزیع نرمال استاندارد پیروی خواهد کرد ($Z\sim N(0\,,۱)$).

$$Z=\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0\,,۱)$$