معیارهای مرکزیت داده‌ها

مد (mode) — داده با بیشترین تکرار

میانه (median) — داده وسطی. (برای تعداد فرد، میانگین دو داده وسط)

میانگین (mean) — $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

معیارهای پراکندگی داده‌ها

دامنه تغییرات — $R = max - min$

میانگین قدر مطلق انحراف از میانگین — $MAD = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \bar{x}\right|$

میانگین مربع انحراف از میانگین — $MSD = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{m}(x_i - \bar{x})^2 \frac{f_i}{n}$

جذر میانگین مربع انحراف از میانگین — $RMSD = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

واریانس — $\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

انحراف معیار — $\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

احتمال

اصول احتمال کولموگروف: $\sum P_i = 1 \qquad P \in [0,1]$

جمع احتمال: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

احتمال متمم: $P(\bar{A}) = 1 - P(A)$

استقلال:

- وقوع یکی از پیشامدها روی دیگری تاثیر ندارد
- $P(A \cap B) = P(A)P(B)$
- $P(A|B) = P(A) \quad or \quad P(B|A) = P(B)$
- $P(A|B) = P(A|\bar{B}) \quad or \quad P(B|A) = P(B|\bar{A})$

احتمال شرطی:

- $P(A|B) = \frac{P(A\cap B)}{P(B)}$
- $P(A \cap B) = P(B) \times P(A|B) = P(A) \times P(B|A)$

قانون بیز: $P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{i=1}^{n} P(A_i)P(B|A_i)}$

متغیر تصادفی گسسته

تابع جرمی احتمال (pmf) $\quad f(x) = P(X = x)$

$$f(x) \geq 0 \quad \forall x \in X, \qquad \sum_x f(x) = 1$$

تابع توزیع جمعی (cdf) $\quad F(x) = P(X \leq x) = \sum_{k \leq x} f(k)$

- میانگین/امید ریاضی $\quad \mathbb{E}[X] = \sum_x x f(x)$
- واریانس $\quad Var(X) = \sigma^2 = \sum_x (x-\mu)^2 f(x) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$

متغیر تصادفی پیوسته

تابع چگالی احتمال (pdf) $\quad P(a \leq X \leq b) = \int_a^b f(x)dx$

$$f(x) \geq 0, \quad \forall x, \qquad \int_{-\infty}^{\infty} f(x)dx = 1$$

تابع توزیع جمعی (cdf) $\quad F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$

- میانگین/امید ریاضی $\quad \mathbb{E}[X] = \int_{-\infty}^{\infty} Xf(X)dx$
- واریانس $\quad Var(X) = \sigma^2 = \int_{-\infty}^{\infty}(x-\mu)^2 f(x)dx = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$

دو متغیر تصادفی

- $\mathbb{E}[aX + bY + c] = a\mathbb{E}[X] + b\mathbb{E}[Y] + c$
- $Var(aX + bY + c) = a^2Var(X) + b^2Var(Y) + 2abCov(X,Y)$
- $Cov(X,Y) = \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$

تابع احتمال شرطی

$$p(x,y) = \frac{p(x,y)}{p(y)} \qquad \mathbb{E}[X|Y=y] = \sum_x xp(x,y)$$

استقلال

$$Pr(X = x, Y = y) = Pr(X = x).Pr(Y = y)$$

کوواریانس معیاری برای ارتباط دو متغیر تصادفی

همراستا $X,Y : Cov(X,Y) > 0$ $\qquad$ غیرهمراستا $X,Y : Cov(X,Y) < 0$

$$\begin{tikzcd}
 & \text{Cov}(X,Y) = 0 \arrow[dr, leftrightarrow] & \\
\text{مستقل } X,Y \arrow[ur] \arrow[rr] & & \mathbb{E}[X]\mathbb{E}[Y] = \mathbb{E}[XY]
\end{tikzcd}$$

جمع و میانگین متغیر تصادفی‌های مستقل

جمع

$Y = a_1X_1 + a_2X_2 + ... + a_nX_n$

$$\mathbb{E}[Y] = \sum_{i=1}^{n} a_i\mathbb{E}[X_i] \qquad Var(Y) = \sum_{i=1}^{n} a_i^2 Var(X_i)$$

میانگین

$\overline{X} = \frac{X_1+X_2+...+X_n}{n}$

$$\mathbb{E}[\overline{X}] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] \qquad Var(\overline{X}) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i)$$

<u>اعمال تابع</u>

$$\mathbb{E}[g(X)] = \sum_{x} g(x)p(x) \quad \mathbb{E}[g(X,Y)] = \sum_{x} g(x,y)p(x,y)$$

بیشینه و کمینه متغیرهای تصادفی مستقل

برای متغیر تصادفی X_i، cdf را با $F_{X_i}(x)$ و pdf را با $f_{X_i}(x)$ نشان می‌دهیم

بیشینه ($V = max\{X_1, X_2, ..., X_n\}$)

$$\begin{aligned} F_V(v) &= P(V \le v) = P(X_1 \le v, X_2 \le v, ..., X_n \le v) \\ &= P(X_1 \le v)P(X_2 \le v)...P(X_n \le v) = \prod_{i=1}^{n} F_{X_i}(v) \end{aligned}$$

$$f_V(v) = F_V'(v) = \frac{d}{dv}F_V(v)$$

کمینه ($U = min\{X_1, X_2, ..., X_n\}$)

$$\begin{aligned} F_U(u) &= P(U \le u) = 1 - P(U > u) = 1 - P(X_1 > u, X_2 > u, ..., X_n > u) \\ &= 1 - P(X_1 > u)P(X_2 > u)...P(X_n > u) \\ &= 1 - \prod_{i=1}^{n}[1 - F_{X_i}(u)] \end{aligned}$$

$$f_U(u) = \frac{d}{du}F_U(u)$$

توزیع $t - student$

فقط برای نمونه‌های با اندازه کوچک که σ نامعلوم است استفاده می‌شود.
همچنین جامعه باید نرمال باشد.

توزیع‌های پیوسته

توزیع یکنواخت (Uniform) : $X \sim U(a,b)$

میانگین: $\mu = \mathbb{E}(X) = \frac{a+b}{2}$

واریانس: $\sigma^2 = Var(X) = \frac{(b-a)^2}{12}$

<u>pdf</u>

$$f_X(x) = \begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$

<u>cdf</u>

$$F_X(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \le x \le b \\ 1 & x > b \end{cases}$$

توزیع نمایی (Exponential) : $X \sim Exp(\lambda)$

میانگین: $\mu = E(X) = \frac{1}{\lambda}$

واریانس: $\sigma^2 = Var(X) = \frac{1}{\lambda^2}$

<u>pdf</u>

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \ge 0 \\ 0 & x < 0 \end{cases}$$

<u>cdf</u>

$$F_X(x) = \begin{cases} 1 - e^{-\lambda x} & x \ge 0 \\ 0 & x < 0 \end{cases}$$

توزیع نرمال

توزیع نرمال $X \sim N(\mu, \sigma^2)$

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\left(\frac{x-\mu}{\sigma}\right)^2/2}$$

نرمال استاندارد $Z \sim N(0,1)$ و $z = \frac{X-\mu}{\sigma}$

$$p(z) = \frac{1}{\sqrt{2\pi}}e^{-z^2/2}$$

- تقریبا 68% داده‌ها در فاصله σ میانگین اند.
- تقریبا 95% داده‌ها در فاصله 2σ میانگین اند.
- تقریبا 99.7% داده‌ها در فاصله 3σ میانگین اند.

قضیه حد مرکزی (CLT)

اگر X_1, X_2,...,X_n، n نمونه تصادفی از یک جامعه/توزیع با میانگین μ و واریانس σ^2 باشند، برای n به اندازه بزرگ ($n \ge 20$) آنگاه:

$$\overline{X} = \frac{X_1 + X_2 + ... + X_n}{n} \sim N(\mu, \frac{\sigma^2}{n})$$

این قضیه در مورد جمع نمونه‌ها نیز برقرار است:

$$T = X_1 + X_2 + ... + X_n = n\overline{X} \qquad T \sim N(n\mu, n\sigma^2)$$

متغیر تصادفی برنولی و دوجمله‌ای

برنولی

فضای نمونه دو حالته.

$$Pr(Success) = p, \quad Pr(Failure) = 1 - p$$

توزیع برنولی:

$X \sim Bernoulli(p)$

pmf: $P(X = x) = p^x(1-p)^{1-x}, \quad x = 0, 1$

میانگین: $\mu = \mathbb{E}[X] = p$

واریانس: $\sigma^2 = Var(X) = p(1-p)$

دوجمله‌ای

متغیر تصادفی دوجمله‌ای X، تعداد موفقیت‌ها در n تلاش مستقل که همه موفقیت‌ها، احتمال برابر p دارند.

توزیع دوجمله‌ای:

$X \sim Bin(n, p)$

pmf $\quad P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$

cdf $\quad P(X \le x) = \sum_{i=0}^{x} \binom{n}{i} p^i (1-p)^{n-i}$

میانگین: $\mu = \mathbb{E}[X] = np$

واریانس: $\sigma^2 = Var(X) = np(1-p)$

تخمین نرمال توزیع دوجمله‌ای

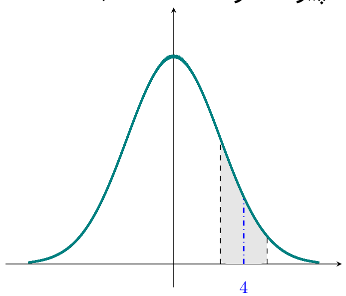
اگر $X \sim Bin(n, p)$ و با تصحیح پیوستگی، داشته باشیم $np \ge 5$ و $n(1-p) \ge 5$، آنگاه :

$$X \sim N(np, np(1-p)), approx.$$

اصلاح پیوستگی

اگر Y پیوسته و X گسسته باشد:

- $P(X > 4) = P(X \ge 5) = P(Y \ge 4.5)$
- $P(X \ge 4) = P(Y \ge 3.5)$
- $P(X < 4) = P(X \le 3) = P(Y \le 3.5)$
- $P(X \le 4) = P(Y \le 4.5)$
- $P(X = 4) = P(3.5 \le Y \le 4.5)$



بازه اطمینان

بازه اطمینان $(1-\alpha)\%$:

- σ^2 معلوم: $\overline{x} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$
- σ^2 نامعلوم: $\overline{x} \pm t_{(n-1)}^{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$

نمونه‌گیری

- با جاگذاری: نمونه مستقل
- بدون جاگذاری: نمونه وابسته. واریانس از با جاگذاری کمتر است. با نمونه کمتری می‌توان μ را تخمین زد
- نمونه‌گیری نسبتا کوچک از جامعه بزرگ: فرقی بین با/بدون جاگذاری نیست

$$\mu_{\bar{X}} = \mu \qquad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

بدون جاگذاری

$$Var(\bar{X}) = \frac{\sigma^2}{n} \frac{m-n}{m-1} \xrightarrow{m \to \infty} Var(\bar{X}) = \frac{\sigma^2}{n}$$

- به $(m-n)/(m-1)$ ضریب کاهش گویند
- اگر $n = 1$، ضریب کاهش برابر ۱
- اگر $m \gg n$ می‌توان نمونه‌ها را مستقل گرفت. (فرقی بین با/بدون جاگذاری نیست)

نمونه‌گیری از جامعه برنولی

$$P = \frac{\sum X_i}{n} \qquad P \sim \mathcal{N}\left(\pi, \frac{\pi(1-\pi)}{n}\right)$$

تفاضل میانگین دو جامعه

با فرض برابری واریانس دو جامعه، تخمین‌گر واریانس جامعه $\mu_1 - \mu_2$

$$s_p^2 = \frac{1}{n_1 + n_2 - 2}\left[\sum_{i=1}^{n_1}(x_{i1} - \bar{x}_1)^2 + \sum_{i=1}^{n_2}(x_{i2} - \bar{x}_2)^2\right]$$

بازه اطمینان برای $\mu_1 - \mu_2$:

$$(\overline{x}_1 - \overline{x}_2) \pm t_{(n_1+n_2-2)}^{\frac{\alpha}{2}} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

جامعه برنولی

برای n بزرگ:

$$\pi_1 - \pi_2 = P_1 - P_2 \pm z_{\frac{\alpha}{2}} \sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}}$$

تخمین‌گر نقطه‌ای

<u>محک</u>

- اریبی

$$B_{\hat{\theta}} = \mathbb{E}[\hat{\theta}] - \theta$$

اریبی صفر، تخمین‌گر خوب

- کارایی

کارایی $\hat{\theta}$ نسبت به $\hat{\hat{\theta}}$:

$$\frac{Var(\hat{\hat{\theta}})}{Var(\hat{\theta})}$$

برای اریب‌ها:

$$\frac{\mathbb{E}[\hat{\hat{\theta}} - \theta]^2}{\mathbb{E}[\hat{\theta} - \theta]^2}$$

تخمین‌گر کارا، بهترین تخمین‌گر

- سازگاری

$$\lim_{n\to\infty} \mathbb{E}[(\hat{\theta}-\theta)^2] = 0 \iff B_{\hat{\theta}}^2 \to 0, \quad \mathrm{Var}(\hat{\theta}) \to 0$$

در $n \to \infty$، سازگاری بهترین معیار و در n کوچک، اریبی.

<u>انتخاب</u>

- MME عملی که قرار است روی جامعه انجام شود، روی نمونه انجام بده.
- MLE

$$Pr(X_i = x_i; \theta) = \prod_{i=1}^{n} F(X_i = x_i; \theta) := L(\theta)$$

$L(\theta)$ را بیشینه کن

رگرسیون +

<u>سری‌های زمانی:</u>
بررسی روند تغییر یک پارامتر در طول بازه زمانی.
روش‌های تصحیح داده‌ها:

- متغیر کمکی dummy با مقادیر گسسته
- تصحیح فصلی با میانگین‌گیری. مناسب داده‌های periodic

<u>قضیه گاوس-مارکوف:</u> در کلاس تخمین‌گرهای خطی نااریب برای α, β، کمترین مجموع مربعات، کمترین واریانس را دارد.

آزمون فرض

- H_0 : باور عمومی. چیزی که می‌خواهیم در برابر آن بجنگیم
- H_1 : مشاهده ما. ادعای ما. چیزی که تلاش می‌کنیم اثبات کنیم.

در آزمون فرض، صرفا با مقایسه دو فرضیه، درستی یا نادرستی H_0 را بررسی می‌کنیم.
<u>انواع خطا</u>

- خطای نوع اول α : رد H_0 به شرط درستی H_0 یا به عبارتی: $Pr(H_0\text{✗}|H_0)$
- خطای نوع دوم β : پذیرفتن H_0 به شرط نادرستی H_0 یا به عبارتی: $Pr(H_0\checkmark|H_1)$

H_1 در خطای نوع اول دخیل نیست.
خطای نوع اول، علاوه بر H_0 به اندازه نمونه نیز بستگی دارد.

خطای اول و دوم، رفتار الاکلنگی ($trade-off$) دارند.

α خطای مهمتری از β است. مثال قاضی

با افزایش اندازه نمونه، α کاهش می‌یابد.

<u>تابع قدرت:</u> مقدار $1-\beta$ را قدرت تست و برای فرضیات مرکب، تابع قدرت گوییم.
دوست داریم H_1 را بپذیریم. پس دوست داریم α کم و قدرت تست بالا داشته باشیم.

<u>p-value:</u>
کمترین احتمال خطای نوع اول که آماره آزمون موجب رد فرض صفر شود، به شرط آنکه فرض صفر صحیح باشد: $p-value = Pr(X > x|H_0)$

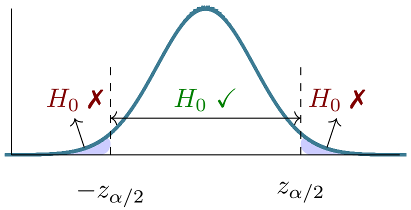
$\alpha > p-value \Rightarrow$ reject H_0

H_0 قابل پذیرش $\iff$ بازه اطمینان شامل H_0 باشد

بازه اطمینان، مجموعه‌ای از فرض‌های قابل قبول است.

Rejection Rules:
Consider test statistic z, and significance value α.

- Lower-tail test: Reject H_o if $z \leq z_\alpha$
- Upper-tail test: Reject H_o if $z \geq z_\alpha$
- Two-tail test: Reject H_o if $|z| \geq z_{\frac{\alpha}{2}}$



ANOVA

برای مقایسه r میانگین. مساله رگرسیون خطی با متغیرهای تماما dummy

تعداد جامعه‌ها	r	
تعداد نمونه‌های هر جامعه	n	
انحراف معیار نمونه i-ام	s_i^2	
میانگین نمونه i-ام	$\bar{x}_i$	$\bar{x}_i = \frac{1}{n}\sum x_i$
میانگین $\bar{X}_i$ها	$\bar{\bar{x}}$	$\bar{\bar{x}} = \frac{1}{r}\sum \bar{x}_i$
واریانس $\bar{X}_i$ها	$s^2_{\bar{X}_i}$	$s^2_{\bar{X}_i} = \frac{1}{r-1}\sum(\bar{x}_i - \bar{\bar{x}})^2$
میانگین واریانس‌ها	s_p^2	$s_p^2 = \frac{1}{r}\sum^r s_i^2$

$$F = \frac{ns^2_{\bar{x}}}{s_p^2} = \frac{Explained\ Var}{Unexplained\ Var}$$

فرضیات:

- برای هر جامعه، متغیر مورد نظر، توزیع نرمال دارد.
- واریانس متغیر مورد نظر، برای همه جوامع برابر σ^2 است.
- مشاهدات مستقلند

آزمون فرض:

$H_0: \mu_1 = \mu_2 = ... = \mu_r$
$H_1: \mu_i \neq \mu_j$ for $i \neq j$

بازه اطمینان همزمان

$$\sum C_i\mu_i = \sum C_i\bar{X}_i \pm \sqrt{F_{0.05}}s_p\sqrt{\frac{(r-1)(\sum C_i^2)}{n}}, \qquad \sum C_i = 0$$

The ANOVA Table:

F-ratio	Mean Square	Sum of Squares	df	Variation of Source
$\frac{\text{MSSr}}{\text{MSSu}}$	$\frac{\text{SSr}}{r-1}$ = MSSr	SSr	$r-1$	Explained
	$\frac{\text{SSu}}{r(n-1)}$ = MSSu	SSu	$r(n-1)$	Unexplained
		SST	$nr-1$	Total

Test Statistic:
$v_1 = df(\text{SSr}) = r-1$
$v_2 = df(\text{SSu}) = r(n-1)$

$$F_{obs} = \frac{\text{MSSr}}{\text{MSSu}} \sim F_{v_1,v_2}$$

اگر $F_{obs} \geq F_{\alpha,v_1,v_2}$ آنگاه H_0 رد می‌شود.

$$\boxed{\text{reject } H_0 \iff F > 1}$$

اگر F نزدیک ۱ باشد، H_0 را می‌پذیریم.

رگرسیون خطی ساده

کمترین مجموع مربعات:

$$\underset{\alpha,\beta}{\text{argmin}}\sum_{i=1}^{n}(y_i - (bx_i + a))^2 \Rightarrow a = \overline{Y} \qquad b = \frac{\sum Y_i x_i}{\sum x_i^2}$$

مدلسازی:

$$Y_i = \alpha + \beta x_i + \varepsilon_i \sim \hat{Y}_i = a + bx_i \quad d.f = n-2$$

$\mathbb{E}[Y_i]$	$\alpha + \beta x_i$	$\mathbb{E}[\varepsilon_i]$	0	$\mathbb{E}[a]$	α	$\mathbb{E}[b]$	β
Var(Y_i)	σ^2	Var(ε_i)	σ^2	Var(a)	$\frac{\sigma^2}{n}$	Var(b)	$\frac{\sigma^2}{\sum x_i^2}$

Y_iها متغیرتصادفی مستقلند

بازه اطمینان:

$$\beta = b \pm t_{0.025}^{(n-2)}\frac{s}{\sqrt{\sum x_i^2}} \qquad \alpha = a \pm t_{0.025}^{(n-2)}\frac{s}{\sqrt{n}}$$

بازه اطمینان پیش‌بینی:

$$Y_0 = \hat{Y}_0 \pm t_{0.025}^{(n-2)} s\sqrt{\frac{1}{n} + \frac{x_0^2}{\sum_{i=1}^{n} x_i^2} + 1}$$

رگرسیون چندگانه

مدلسازی:

$$Y_i = \alpha + \beta x_i + \gamma z_i + \varepsilon_i \sim \hat{Y}_i = a + bx_i + cz_i \quad d.f = n-3$$

کمترین مجموع مربعات:

$$\underset{\alpha,\beta,\gamma}{\text{argmin}}\sum_{i=1}^{n}(y_i - (a + bx_i + cz_i))^2 \Rightarrow a = \overline{Y} \qquad \begin{cases}\sum Y_i x_i = b\sum x_i^2 + c\sum x_i z_i \\ \sum Y_i z_i = b\sum x_i z_i + c\sum z_i^2\end{cases}$$

همخطی چندگانه (multicollinearity): یک بعد را از دست می‌دهیم! وقتی یکی از پارامترها ضریب دیگری باشد یا ارتباط خطی داشته باشند اتفاق می‌افتد.

- رگرسیون استاندارد: رگرسیون با متغیرهای عددی
- ANOVA : رگرسیون با متغیرهای dummy
- ANOCOVA : رگرسیون استاندارد + ANOVA