DSE210 HW5 - Ryan Inghilterra

Ryan Inghilterra
03/15/2018
DSE 210

HW 5 - Worksheet 10 - PCA and SVD

① set of vectors $\begin{pmatrix}3\\4\\0\end{pmatrix}, \begin{pmatrix}4\\-3\\0\end{pmatrix}, \begin{pmatrix}0\\0\\1\end{pmatrix}$ an orthonormal basis of $\mathbb{R}^3$?

to be orthonormal first off all must be unit vectors

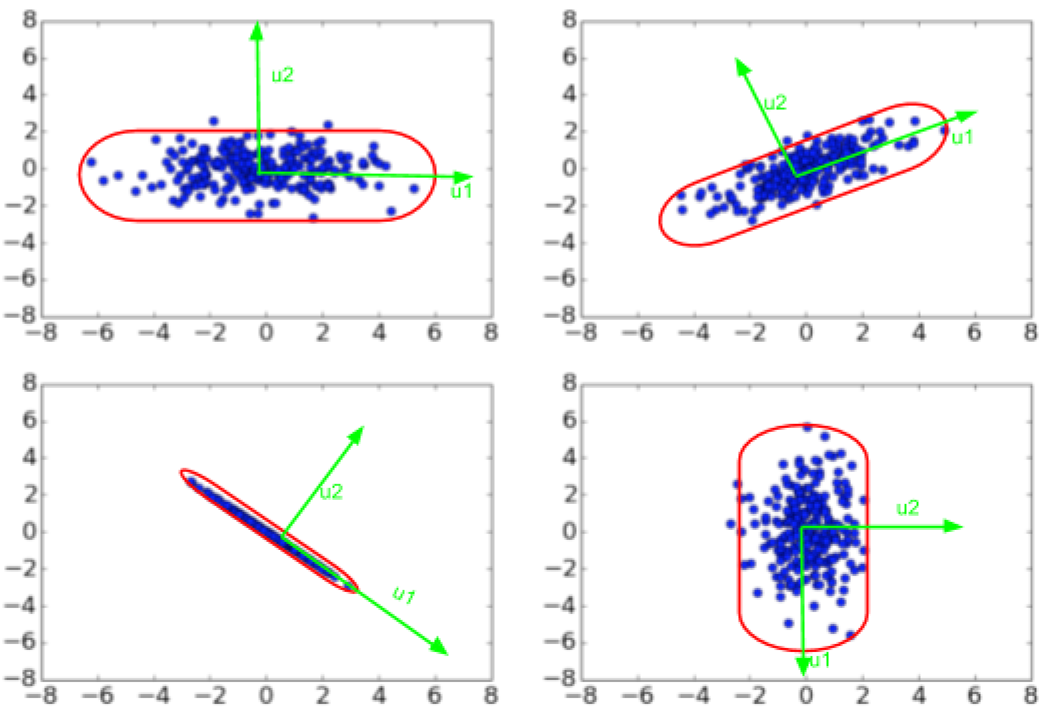
so $\|u\|=1$, we can see that for $V_1 = \begin{pmatrix}3\\4\\0\end{pmatrix}$, $\|V_1\| = \sqrt{3^2+4^2+0^2} = \sqrt{25} = 5$

$\|V_1\| = 5$, so not a unit vector

therefore $\boxed{NO}$ set of vectors not an orthonormal basis of $\mathbb{R}^3$

② SEE NEXT PAGE

2. The following four figures show different 2-dimensional data sets. In each case, make a rough sketch of an ellipsoidal contour of the covariance matrix and indicate the directions of the first and second eigenvectors (mark which is which).



(3) Let $u_1, u_2 \in \mathbb{R}^d$ be two vectors, $\|u_1\| = \|u_2\| = 1$ and $u_1 \cdot u_2 = 0$

Define $U = \begin{pmatrix} \uparrow & \uparrow \\ u_1 & u_2 \\ \downarrow & \downarrow \end{pmatrix}$

$d \begin{pmatrix} \uparrow & \uparrow \\ u_1 & u_2 \\ \downarrow & \downarrow \end{pmatrix}$ $\times$ 2

(a) what are dimensions of following?

- U - $\boxed{d \times 2}$
- U^T - $\boxed{2 \times d}$
- UU^T - $\boxed{d \times d}$
- $u_1 u_1^T$ - $\boxed{d \times d}$

(b) what are the differences between following four mappings?

1. $x \mapsto (u_1 \cdot x, u_2 \cdot x)$
2. $x \mapsto (u_1 \cdot x) u_1 + (u_2 \cdot x) u_2$
3. $x \to U^T x$
4. $x \to UU^T x$

mapping 1 and 3 are equivalent, both are projections of x, from $x \in \mathbb{R}^d$ to $\mathbb{R}^2$ $\in \mathbb{R}^2$ (d to K dimensions) $K=2$

mapping 2 and 4 are equivalent both are reconstructions $x \in \mathbb{R}^d$ projected to $\mathbb{R}^2$ reconstructed back to $\mathbb{R}^d$ $K=2$

Ryan Inghilterra
DSE 210
03/15/2018

HW - Worksheet 11 - Sampling

(1) 9 red marbles, 1 blue marble. 900 rand. draws w/ replacement.

$n = 900$ $Pr(red) = 0.9 = p$

$N(n \cdot p, n \cdot p(1-p))$ since observational number

$N(900 \cdot (0.9), 900 \cdot (0.9) \cdot (0.1))$

$= \boxed{N(810, 81)}$

(2) in world say 1% people are left handed. sample 200 people at random. Give 99% confidence interval for # of them left handed

$N = 200$ Find STD, $var = np(1-p)$ $Pr(L) = 0.01 = p$

$STD = \sqrt{200(0.01)(0.99)}$ mean $M = n \cdot p$

$= \sqrt{1.98} \approx 1.41$ $M = 200 \cdot 0.01$

mutiply std by 3 since 99% $M = 2$

99% confidence interval left handed $= 2 \pm \sqrt{1.98}(3)$

$\boxed{= 2 \pm 4.22}$

(3) 20 wedges numbered 1-20. Half red, half black, 100 darts thrown

$n = 100$ $Pr(R) = 0.5$ $Pr(B) = 0.5$ $E(X_i) = 1/20$

(a) X_i # darts fall in wedge i. What is $E(X_i)$ and $Var(X_i)$

$E(X_i) = (\frac{1}{20})(100) = \boxed{5 = E(X_i)}$

$\quad\quad\quad\quad p \quad\quad n$

$var(X_i) = n \cdot p(1-p) = (100)(.05)(.95) = \boxed{4.75 = Var(X_i)}$

(b) upper bound on X_i, 95% confidence. (normal approx)

$M \pm 2 \cdot std \rightarrow 5 \pm 2 \cdot \sqrt{4.75} \rightarrow \boxed{5 \pm 2.18}$

Z_r = # wedges red, Z_b = # wedges black, $Z = |Z_r - Z_b|$

$Y_i = \begin{cases} 1 & \text{if dart on red} \\ -1 & \text{if dart on black} \end{cases}$, $Z_r - Z_b = Y_1 + Y_2 \ldots Y_{100}$ want 99% confidence interval

(c) $\boxed{E(Y_i) = 0}$ $\boxed{var(Y_i) = 1}$

d) central limit theorem $Z_r - Z_b$ approx. normal

$E(Y) = 0$ $var(Y) = n$, $std(Y) = \sqrt{n} = \sqrt{100} = 10$

$\boxed{M = 0, \; std = 10}$, $\boxed{N(0, 100)}$

(e) 99% conf. interval $\rightarrow$

$M \pm 3 \cdot std = 0 \pm 3(10)$ $\boxed{0 \pm 30}$

④ colorblind appear 1% people. How large sample for prob of it containing at least 1 colorblind person to be at least 95%.

$M \pm 2 \cdot std$ since 95% conf. interval $M = (0.1 \cdot N)$ $var = (0.1)n(0.99)$

think in terms of prob nobody colorblind less than 5%

$.99^N < .05$ $N > 298.073$

$N \ln(.99) < \ln(.05)$ (sample size need to be 299 or greater)

⑦ ~~tolerate charac~~ errors of 1% or so of estimate. Use sample size 100, 250, 10,000? Aux. source of info suggests population % range 20% to 40%

want estimate to be $M \pm .01$ $\sigma = \sqrt{\frac{p \cdot (1-p)}{n}}$ so $p = .40$ since will be worse then $p = 0.2$

for $N = 100$ $\sigma = \sqrt{\frac{(0.4)(0.6)}{100}} = .05 \leftarrow$ too high!

For $N = 250$ $\sigma = \sqrt{\frac{(0.4)(0.6)}{250}} = .03 \leftarrow$ too high! $M = p = 0.4$

For $N = 10,000$ $\sigma = \sqrt{\frac{(0.4)(0.6)}{10,000}} = .0048 \leftarrow$ good ✓

since we want 95% conf. interval that is

$M \pm 0.01$ so where $2 \cdot \sigma \leq 0.01$

only for $N = 10,000$, is $2\sigma < 0.01$

$2(.0048) = .0096 < .01$ ✓

so should use sample size of $\boxed{10000}$

⑧ city there are 100,000 people age 18 to 24. random sample 500 people 194 enrolled in college. Estimate % of all people 18-24 in city in college give 95.5% confidence interval.

$n = 500$ $194/500 = \hat{p}$ (fraction)

want $M + 2\sigma$ 0.388 $M = \hat{p} = 0.388$ $\sigma = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

$= \sqrt{\frac{(.388)(1-.388)}{500}}$

so 95.5 conf. interval is

0.388 ± 0.02

$\sigma = 0.02$

Ryan Inghilterra
DSE 210
03/15/2018

HW 5 Worksheet 11 - Sampling Continued

(9) $n = 1000$ what sample size should they use?

SAME sample size of 1000

what matters is the sample size, not overall population size.

(10) $n = 100$ $\hat{\mu} = 307$ $\hat{\sigma} = 30$

$$\sigma = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{30}{\sqrt{1000}} = 0.949 = \sigma$$

$$\boxed{\mu = 307 \quad \sigma = 0.95}$$

we dont know, so use formula w/ sample std $\hat{\sigma}$

Ryan Inghilterra
DSE 210
03/15/2018

(HW5)
Worksheet 12 - Hypothesis Testing

① 1990 US, 2.1 million deaths from all causes, compared to 1.7 million in 1960
25% increase, show public health got worse over 1960 - 1990?
$\boxed{\text{NO}}$, not necessarily, all causes can include causes other than health, also not taking into account population increase.

② smoking study
ⓐ controlled or observational? observational
ⓑ don't want to confound different biases between age groups
ⓒ that conclusion is not necessarily true and can be misleading. People could have stopped smoking b/c of health issues. And those who were healthy had no reason to stop smoking.

③ oral contraceptive study
ⓐ controlled or observational? observational
ⓑ want to control for confounding factors. Age, education levels, and marital status can have a big influence on the results of this expirement.
ⓒ study only showed there exists a relationship between the pill and cervical cancer. Did not prove it is a causal relationship

⑥ 10,000 tossings, coin comes up heads 5,400 times. Conclude coin is biased
ⓐ null hypothesis: coin is unbiased
ⓑ compute Z-statistic and p-value
observed = 5400 expected = 5000

std $= \sqrt{n \cdot p(1-p)}$ $p = 0.5$
std $= \sqrt{(10{,}000)(.5)(.5)}$ $n = 10{,}000$
std $= \sqrt{2500} = 50$

$$Z\text{-stat} = \frac{\text{observed} - \text{expected}}{\text{std}}$$

$$= \frac{5400 - 5000}{50} = \frac{400}{50} = 8 = Z\text{-stat}$$

p-value is < 0.00001

ⓒ strong evidence against the null, reject at 0.05 significance level
conclude coin is biased

(7) die is rolled 100 times. total # of spots = 368 instead of expected 350.
Can this be explained as a chance variation or die loaded?

null hypothesis: die is fair   mean of single dice roll = 3.5

$std(\text{dice roll}) = \sqrt{\frac{1}{6}\left((1-3.5)^2 + (2-3.5)^2 + \dots + (6-3.5)^2\right)} = \sqrt{35/12} = 1.71$

divide std 1.71 by $\sqrt{\text{sample size } n}$, $\frac{1.71}{\sqrt{100}} = 0.171$

mean of sample dice roll = 3.68

$Z\text{-stat} = \frac{(3.68 - 3.5)}{0.171} = \frac{0.18}{0.171} = 1.05$ p-value is 0.146859

result is not significant at $p < 0.05$
therefore we cannot reject the null, we accept the null.
Conclusion: die are not loaded

(8) other things being equal, which is better, for null hypothesis:
a higher p-value or lower p-value?
a higher p-value, higher p-value more likely to accept the null. since p-value is the probability of observing the expirement results under the null.

Ryan Inghilterra
DSE 210
03/15/2018

(HW 5) - Worksheet 12 - Hypothesis Testing Cont.

(9) drug abuse survey, each year 1985 and 1992, 700 random people sampled

(a) age 18 to 25, percentage of marijuana users dropped from 21.9% to 11.0%. Real or chance variation? $n = 700$

observed difference $= .219 - .11 = .109$

$\rightarrow \sigma_{1985} = \sqrt{\frac{(.219)(.781)}{700}} = .01563$ $\quad \sigma_{1992} = \sqrt{\frac{(.11)(.89)}{700}} = .01182$

null hypothesis: $X_{1985} = X_{1992}$ (drop is from chance in variation)

calculate $\sigma = \sqrt{\sigma_{1985}^2 + \sigma_{1992}^2}$ $\quad \sigma = \sqrt{(.01563)^2 + (.01182)^2} = .0196$

so $z\text{-stat} = \frac{.109}{.0196} = 5.56$, reject the null,

Conclusion: Drop in marijuana use from 1985 to 1992 is real!

(b) cigarette 36.9 to 31.9, difference real or chance in variation?

observed difference $= .369 - .319 = .05$

null hypothesis: $X_{1985} = X_{1992}$ (drop is from chance variation)

$\sigma_{1985} = \sqrt{\frac{(.369)(.631)}{700}} = .01823$ $\quad \sigma_{1992} = \sqrt{\frac{(.319)(.681)}{700}} = .01761$

calculate $\sigma = \sqrt{(.01823)^2 + (.01761)^2} = .02534$

so $z\text{-stat} = \frac{.05}{.02534} = 1.97 \rightarrow$ p-value is .024419

we are testing at 95% confidence interval, result is significant at $p < 0.05$ level

therefore reject null

Conclusion: Drop in cigarette use from 1985 to 1992 is real!

⑩ 1000 freshman sample. Average # hours = 12.2 worked
std = 10.5 for public school

- other survey at private university Average # hours = 9.2 std = 9.9
difference between two averages due to chance?

null hypothesis: means of the two distributions (public vs private universities) are the same. Differences due to chance

$\sigma = \sqrt{\sigma_1^2 + \sigma_2^2} \rightarrow$ $\sigma = \sqrt{(10.5)^2 + (9.9)^2} = 14.43$

z stat = $\frac{12.2 - 9.2}{14.43} = \frac{3}{14.43} = 0.207$ p-value ≈ 0.418 05
↑ high p-value!

result is not significant at $p < 0.05$

therefore we cannot reject the null, we accept the null.

Conclusion: the differences between the two averages are due to chance