Review. Sample correlation

Statistics

Anton Afanasev

Higher School of Economics

DSBA 221

January 13, 2024

Seminar Overview

1. Quiz
2. Review
3. Bivariate normal distribution
4. Estimation of correlation coefficient
   - Sample correlation coefficient
   - Fisher transformation
   - Confidence interval for correlation coefficient
   - Spearman's rank correlation coefficient

Quiz

Find a match:

1. Cumulative distribution function
2. Quantile function
3. Pooled variance
4. Exponential distribution
5. Degrees of freedom
6. Central Limit Theorem
7. Independence
8. Correlation

A. Waiting time
B. Separability
C. Approximation
D. Antiderivative
E. Weighted average
F. Goodness-of-fit
G. Constraints
H. Inverse

A random sample of 400 married couples was selected from a large population of married couples.

- Heights of married men are approximately normally distributed with mean 70 inches and standard deviation 3 inches.
- Heights of married women are approximately normally distributed with mean 65 inches and standard deviation 2.5 inches.
- There were 20 couples in which wife was taller than her husband, and there were 380 couples in which wife was shorter than her husband.

Problem 1

Objectives

1. Find a 95% confidence interval for the proportion of married couples in the population for which the wife is taller than her husband.
2. Suppose that a married man is selected at random and a married woman is selected at random. Find the approximate probability that the woman will be taller than the man.
3. Based on your answers to 1 and 2, are the heights of wives and their husbands independent? Explain your reasoning.

Problem 2

Suppose 2000 points are selected independently at a random from the unit square $S = \{(x, y) : 0 \leq x \leq 1,\ 0 \leq y \leq 1\}$. Let W be the number of points that fall into the set $A = \left\{(x, y) : \ x^2 + y^2 < 1\right\}$.

1. How is W distributed?
2. Find the mean, variance and standard deviation of W.
3. Estimate probability that W is greater than 1600.

Problem 3

Distribution of X is uniform $\mathcal{U}(-a, a)$. Sample of size $n = 2$ is available. Consider $\widehat{a} = c \cdot (|X_1| + |X_2|)$ as a class of estimators for the parameter a. Find c such that

1. Estimator $\widehat{a}$ is unbiased.
2. Estimator $\widehat{a}$ is the most efficient in the class. (In terms of mean square error.)

Problem 4

Consider random variables X and Y with joint density function

$$f(x,y) = \begin{cases} \frac{1}{2} + cx, & x + y \leq 1,\ x \geq 0,\ y \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

1. Find c.
2. Find $f_X(x)$. Evaluate $\mathsf{E}(X)$.
3. Write down an expression for $f_{Y|X}(x,y)$. Find $\mathsf{E}(Y \mid X = x)$.

Problem 5

Internal angles $\theta_1, \theta_2, \theta_3, \theta_4$ of a certain quadrilateral, located on the ground, were measured by the aerial system. It is assumed that those observations x_1, x_2, x_3, x_4 were taken with minor and independent errors, which have zero mean and identical variance σ^2.

1. Find the LSE of $\theta_1, \theta_2, \theta_3, \theta_4$.
2. Find an unbiased estimate of σ^2 in the case, described in part 1.
3. Let's assume now that the considered quadrilateral is a parallelogram with $\theta_1 = \theta_3$ and $\theta_2 = \theta_4$. How values of internal angles LSE would change? Find an unbiased estimate of σ^2 in this particular case.

Problem 6

Suppose that student's grade for a statistics exam, X, has continuous uniform distribution at the interval $[0, 100]$. But less then 25 points means "failed", and more than 80 points is "excellent", hence the final grade Y is calculated as follows:

$$Y = \begin{cases} 0, & X < 25 \\ X, & 25 \leq X < 80 \\ 100, & X \geq 80 \end{cases}$$

1. Find c.d.f. of Y. Sketch the plot.
2. Find p.d.f. of Y. Sketch the plot.
3. Find mean and variance of X and Y.
4. Find $\mathsf{E}(Y \mid Y > 0)$.
5. Find $\mathsf{Corr}(X, Y)$.

Problem 6

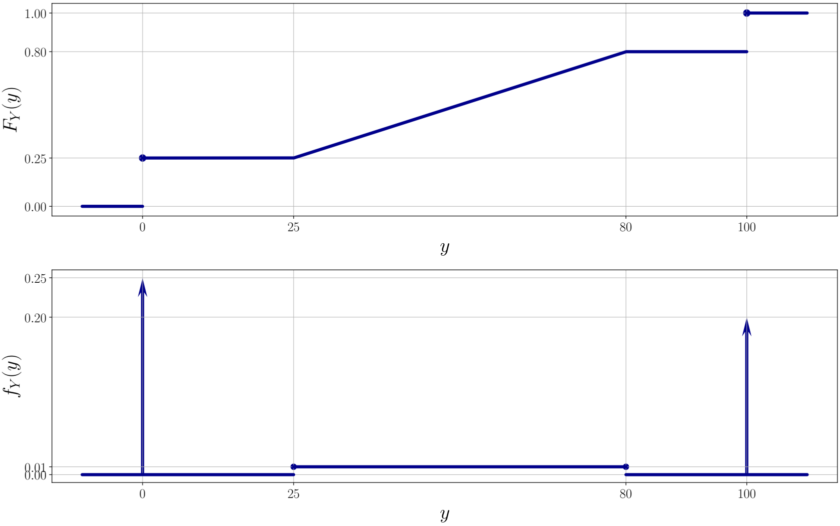


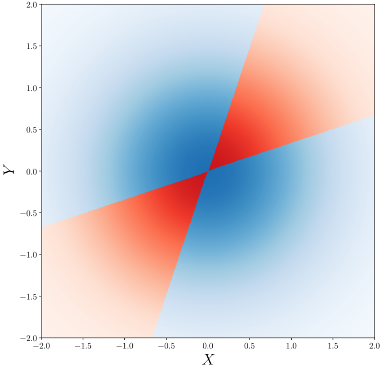
Figure: C.d.f. and generalized p.d.f. of the random variable Y.

Problem 7

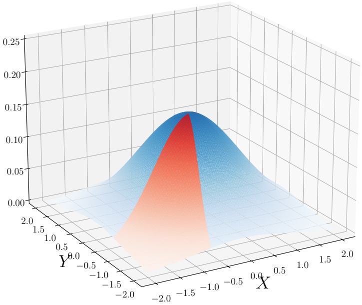
Let X and Y be two independent standard normal random variables. Find

1. $\mathsf{P}(|X+Y| > |X-Y|)$.
2. $\mathsf{P}(|X+Y| > 2|X-Y|)$.

Problem 7



(a) top view



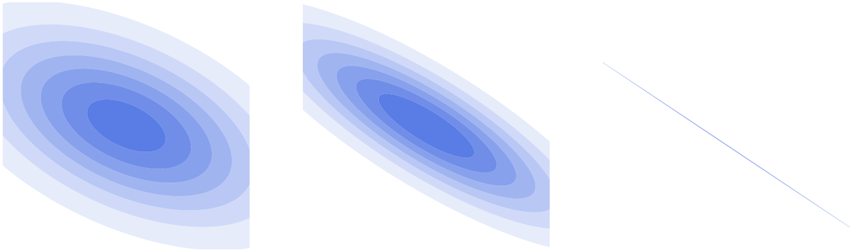
(b) side view

Figure: Region $(3X - Y)(3Y - X) > 0$ of variable $(X \;\; Y)^\top \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Problem 8

Two random variables are given: $X \sim \mathcal{N}(0, 9)$ and $Y \sim \mathcal{N}(0, 4)$. $\mathsf{Corr}(X, Y) = -1$. Evaluate $\mathsf{P}(2X + Y > 3)$.

Problem 8



(a) $\rho = -\frac{1}{2}$ (b) $\rho = -\frac{5}{6}$ (c) $\rho \to -1$

Figure: P.d.f. of a bivariate normal distribution with $\sigma_X = 3$ and $\sigma_Y = 2$.

Sample from bivariate normal distribution

- Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a random sample from population $\mathbf{X} = \begin{pmatrix} X & Y \end{pmatrix}^\top$ with bivariate normal distribution $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}$$

  is a mean vector, and

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}$$

  is a covariance matrix.

- We want to estimate correlation coefficient $\rho = \mathsf{Corr}(X, Y)$. Values of parameters μ_X, μ_Y, σ_X and σ_Y are unknown.

Sample covariance

- Correlation coefficient is defined as:

$$\rho = \frac{\mathsf{Cov}(X,Y)}{\sigma(X)\cdot\sigma(Y)} = \frac{\mathsf{E}\left(X-\mathsf{E}(X)\right)\left(Y-\mathsf{E}(Y)\right)}{\sqrt{\mathsf{E}\left(X-\mathsf{E}(X)\right)^2}\cdot\sqrt{\mathsf{E}\left(Y-\mathsf{E}(Y)\right)^2}}.$$

- Naturally, a point estimator of ρ should look like:

$$\widehat{\rho} = \frac{\widehat{\mathsf{Cov}}(X,Y)}{\widehat{\sigma}(X)\cdot\widehat{\sigma}(Y)} = \frac{S_{XY}}{S_X\cdot S_Y},$$

constituted of unbiased point estimates of variances and covariance

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2, \qquad S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2,$$

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right).$$

Corrected sums

- We will have a specific notation for corrected sums of cross-products of X and Y:

$$SS_{XY} = \sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right) = \sum_{i=1}^{n} X_i Y_i - n\overline{X}\,\overline{Y}.$$

  Substituting X for Y and vice versa, we get corrected sums of squares:

$$SS_{XX} = \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}^2,$$

$$SS_{YY} = \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2 = \sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2.$$

- The notation could define something more exotic:

$$SS_{X\sin X} = \sum_{i=1}^{n} X_i \sin X_i - n\overline{X}\,\overline{\sin X}.$$

Sample correlation coefficient

- Sample variances and covariance are corrected sums, divided by a number of degrees of freedom:

$$S_X^2 = \frac{SS_{XX}}{n-1}, \qquad S_Y^2 = \frac{SS_{YY}}{n-1}, \qquad S_{XY} = \frac{SS_{XY}}{n-1}.$$

(*meaning*: how much change is allocated onto one degree of freedom)

- Reducing degrees of freedom, the equation for a sample correlation coefficient:

$$\widehat{\rho} = \frac{SS_{XY}}{\sqrt{SS_{XX} \cdot SS_{YY}}}.$$

- The formula is true for any population distribution.

Variance-stabilizing transformations

- The distribution of $\widehat{\rho}$ is a complex hypergeometric function, and moreover, its variance depends on ρ.
- Let's transform $\widehat{\rho}$ into another variable, which will be well-approximated with table function and will have constant variance. Such transformations are called variance-stabilizing.

Example (Poisson distribution)

- Let $P \sim \text{Poisson}(\lambda)$ with $\mathsf{E}(P) = \lambda$ and $\mathsf{V}(P) = \lambda$.
- In order to allow analysis of variances techniques, we want all sources to have identical variance.
- Using Anscombe transform: $Q = 2\sqrt{P + \frac{3}{8}}$,

  new variable will have $\mathsf{E}(Q) \approx 2\sqrt{\lambda + \frac{3}{8}} - \frac{1}{4\sqrt{\lambda}}$ and $\mathsf{V}(Q) \approx 1$ for larger parameters λ.

Fisher z-transform

- Fisher z-transform of a sample correlation coefficient $\widehat{\rho}$:

$$\widehat{Z} = \operatorname{artanh}\left(\widehat{\rho}\right),$$

where "artanh" is an inverse hyperbolic tangent function

$$\operatorname{artanh}(x) \equiv \frac{1}{2}\ln\left(\frac{1+x}{1-x}\right).$$

- Transformed variable is very close to normal distribution with constant variance:

$$\widehat{Z} \overset{\text{approx}}{\sim} \mathcal{N}\left(\operatorname{artanh}(\rho), \frac{1}{n-3}\right).$$

- Approximation becomes better for larger n, though it's good for any $n > 3$.

Confidence interval for correlation coefficient ρ

- Confidence interval for transformed ρ:

$$\mathrm{CI}_{1-\alpha}(\operatorname{artanh}(\rho)) = \operatorname{artanh}\left(\widehat{\rho}\right) \pm z_{\alpha/2} \cdot \frac{1}{\sqrt{n-3}}.$$

- Applying inverse Fisher transform $\left(\widehat{\rho} = \tanh\left(\widehat{Z}\right)\right)$ to those endpoints gives a result for ρ:

$$\mathrm{CI}_{1-\alpha}(\rho) = \left(\tanh\left(\operatorname{artanh}\left(\widehat{\rho}\right) - z_{\alpha/2} \cdot \frac{1}{\sqrt{n-3}}\right);\right.$$
$$\left.\tanh\left(\operatorname{artanh}\left(\widehat{\rho}\right) + z_{\alpha/2} \cdot \frac{1}{\sqrt{n-3}}\right)\right),$$

  where “tanh” is a hyperbolic tangent function

$$\tanh(x) \equiv \frac{e^{2x}-1}{e^{2x}+1}.$$

- For simulations refer to the link:

Confidence interval for ρ

The sample from bivariate normal distribution with random variables X and Y is following:

X	1.59	-2.20	-0.06	-1.45	-1.02	-2.59	-1.14	-3.25
Y	3.24	0.44	-1.14	5.40	2.09	5.33	1.25	8.72

Find 90% confidence interval for a population correlation coefficient ρ.

Spearman's rank correlation coefficient

- Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a random sample from any population.
  Spearman's rank correlation coefficient is

$$\widehat{\rho}_s = \frac{S_{\text{rank}(X)\text{rank}(Y)}}{S_{\text{rank}(X)} \cdot S_{\text{rank}(Y)}},$$

  where rank is an ordered number of each score X_i and Y_i.
- If all n ranks are distinct integers, the following formula is applicable:

$$\widehat{\rho}_s = 1 - \frac{6 \sum\limits_{i=1}^{n} d_i^2}{n\left(n^2 - 1\right)},$$

  where $d_i = \text{rank}(X_i) - \text{rank}(Y_i)$.

Spearman's rank correlation coefficient

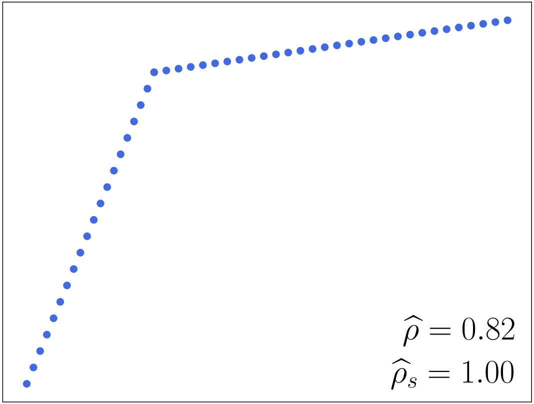


Figure: Spearman's rank correlation may follow nonlinear monotonicity.

Spearman's rank correlation coefficient

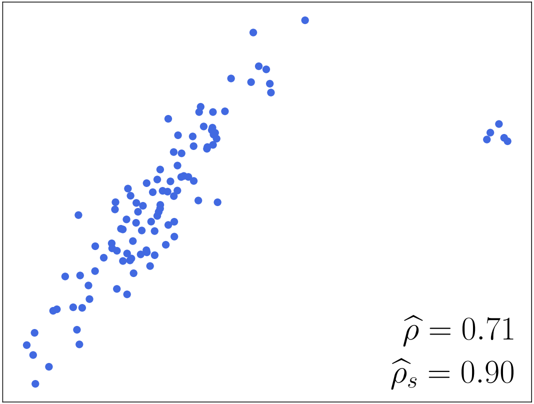


Figure: Spearman's rank correlation is more robust to outliers.

Problem 10

Consider observations in the table below:

x	0	2	6	-3	4	1	-2	5	-1
y	8	2	0	6	1	5	7	3	4

1. Find Spearman's rank correlation coefficient r_s.
2. Find sample correlation coefficient r and compare it with r_s.

That's all Folks