VYBRANÉ APLIKACE MATEMATICKÉ STATISTIKY. SYNOPSIS OF COURSE AT OU, WINTER SEMESTERS 2015/2016, 2016/2017, SUMMER SEMESTER 2018/2019

PASHA ZUSMANOVICH

Each class lasted 1.5 hours. *Stuff typed in italic was not covered during summer semester 2018/2019.*

Class 1. (February 14, 2019)

Subject of statistics.

Interesting applications of statistics: detection election frauds ([KSP], [KYHT]*), patterns in citations* [SR].

Types of data: numerical and categorical. Statistical data always contains errors and is incomplete.

Bar charts and histograms.

Average, standard deviation, their meaning. Assuming $\bar{x} = (x_1, \dots, x_n)$,

$$m(\bar{x}) = \frac{x_1 + \dots + x_n}{n}$$

$$\sigma(\bar{x}) = \sqrt{\frac{\big(x_1 - m(\bar{x})\big)^2 + \dots + \big(x_n - m(\bar{x})\big)^2}{n}}.$$

Median, quantiles (generalization of median).

A glimpse into R: installation, usage as calculator, assignments, 1-dimensional arrays, functions. `help()`, `example()`, `mean()`, `sd()`, `median()`, `quantile()`, `plot()`, *barplot()*. Drawing histograms in different ways.

A toy example: plot *and linear regression* of air pollution against temperature for a 24 hour period in Ostrava.

Class 2. (February 21, 2019)

Mode.

Discrete vs. continuous distributions.

Density function of the normal distribution:

$$f_{m,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-m}{\sigma})^2}.$$

Its importance, its properties: symmetry, maximum (by calculating derivative). Central Limit Theorem.

Density function of the logistic distribution:

$$f_{m,s}(x) = \frac{e^{-\frac{x-m}{s}}}{s(1 + e^{-\frac{x-m}{s}})^2}.$$

Its standard deviation: $\sigma = \frac{\pi s}{\sqrt{3}}$.

Density function of the Laplace distribution:

$$f_{m,b}(x) = \frac{1}{2b} e^{-\frac{|x-m|}{b}}.$$

Its standard deviation: $\sigma = b\sqrt{2}$.

Date: last modified May 9, 2019.

These distributions have, roughly, the same properties as a normal distribution: the same "bell-shaped" form, attain one maximum "in the middle" (average), and are symmetric, but logistic distribution is "heavier on tails", and Laplace distribution has a sharp peak in the middle.

Fitting in R real data to normal, Laplace and logictic distributions.

Class 3. (February 28, 2019)

Demonstration in R of the central limit theorem. (An alternative demonstration, using a different R code, can be found in [Cr1, §7.3.2]).

Skewness and kurtosis, their meaning (according to [Cr2, pp. 84–87]) and Wikipedia ([$\mathrm{Sk_w}$], [$\mathrm{K_w}$]).

$$skewness(\overline{x}) = \frac{\text{3rd moment}(\overline{x})}{\sigma(\overline{x})^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - m(\overline{x}))^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - m(\overline{x}))^2\right)^{\frac{3}{2}}};$$

$$kurtosis(\overline{x}) = \frac{\text{4th moment}(\overline{x})}{\sigma(\overline{x})^4} - 3 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - m(\overline{x}))^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - m(\overline{x}))^2\right)^{2}} - 3\,.$$

Kurtosis of a normal distribution is equal to 0.

"Paradoxes" in statistics (according to [T, §6.5]).

Weighted mean. `weighted.mean()` in R. The usual mean of a discrete statistical distribution $(x_1, \dots, x_n)$ can be interpreted as a weighted mean, if we assume that all x_i's are pairwise distinct, and each appears with a frequency (probability) p_i. Then

$$m(\overline{x}) = p_1 x_1 + \dots + p_n x_n = \frac{p_1 x_1 + \dots + p_n x_n}{p_1 + \dots + p_n}$$

(as $p_1 + \dots + p_n = 1$).

Computation of weighted population density: if the whole area is divided to n regions with population $x_1, \dots, x_n$ and areas $s_1, \dots, s_n$, then the weighted population density is the weighted mean of densities per region, with weights given by population:

$$\frac{x_1 \frac{x_1}{s_1} + \dots + x_n \frac{xn}{s_n}}{x_1 + \dots + x_n}.$$

Simpson's paradox (according to [$\mathrm{Si_w}$]). UC Berkeley suitcase.

	men	men admitted	women	women admitted
Department 1	m_1	$\lambda_1 m_1$	w_1	$\mu_1 w_1$
Department 2	m_2	$\lambda_2 m_2$	w_2	$\mu_2 w_2$

It could be that $\lambda_1 < \mu_1$ and $\lambda_2 < \mu_2$, but

$$\frac{\lambda_1 m_1 + \lambda_2 m_2}{m_1 + m_2} > \frac{\mu_1 w_1 + \mu_2 w_2}{w_1 + w_2}.$$

Another example of Simpson's paradox often occurs in US election system, see, e.g. [W].

Class 4. (March 7, 2019)

Discussion of homeworks 1-2.

Confidence intervals for normal distribution. Standard error. (According to [D, pp. 63-64] and [$\mathrm{C_w}$]).

Using confidence intervals to determine sample size.

$$\text{Margin error} = \frac{\sigma}{\sqrt{n}} N_{1-\frac{1-\alpha}{2}},$$

where α is confidence interval (say, $\alpha = 0.95$), and N are quantiles for the normal distribution.

Q-Q plot of one data against another.

Tests for normality: normal scores, Q-Q plots (according to [CC, pp. 220–223]).

Class 5. (March 21, 2019)

Discussion of homework 3.

Correlation:

$$cov(\overline{x}, \overline{y}) = \sum_{i=1}^{n} \Big(x_i - m(\overline{x})\Big)\Big(y_i - m(\overline{y})\Big)$$

$$cor(\overline{x}, \overline{y}) = \frac{cov(\overline{x}, \overline{y})}{\sqrt{cov(\overline{x}, \overline{x}) cov(\overline{y}, \overline{y})}}.$$

Properties of correlation:

$$cor(\overline{x}, \overline{x}) = 1$$
$$cor(\overline{x}, \overline{y}) = cor(\overline{y}, \overline{x})$$
$$-1 \le cor(\overline{x}, \overline{y}) \le 1.$$

The latter one follows from the Cauchy–Schwarz inequality:

$$\Big(\sum_{i=1}^{n} a_i b_i\Big)^2 \le \Big(\sum_{i=1}^{n} a_i^2\Big)\Big(\sum_{i=1}^{n} b_i^2\Big).$$

Correlation between linearly dependent datasets is equal to 1 or -1.

Example of two datasets with correlation zero: let $\overline{x}$ be any vector of even length whose alternating sum is zero, for example, $x_1 = x_2, x_3 = x_4, \dots, x_{2n-1} = x_{2n}$, and $\overline{y}$ is oscillating, say, $y_i = 1$ for i odd and $y_i = 0$ for i even. Then $m(\overline{y}) = \frac{1}{2}$, and

$$\begin{aligned} cov(\overline{x}, \overline{y}) &= \sum_{i=1,3,\dots,2n-1} \Big(x_i - m(\overline{x})\Big)(1 - \frac{1}{2}) + \sum_{i=2,4,\dots,2n} \Big(x_i - m(\overline{x})\Big)(0 - \frac{1}{2}) \\ &= \frac{1}{2}(x_1 - x_2 + x_3 - x_4 + \dots + x_{2n-1} - x_{2n}) = 0, \end{aligned}$$

and hence $cor(\overline{x}, \overline{y}) = 0$.

Use and misuse of correlation. "Correlation is not causation".

Correlation matrix, its properties (symmetric, 1's on the main diagonal).

Class 6.

Discussion of homework 6 (R code demonstrating Central Limit Theorem for any distribution).

Iterative correlation matrices (according to [Ch]).

The case of 2×2 matrices:

$$\begin{pmatrix} 1 & a \\ a & 1 \end{pmatrix} \mapsto \begin{pmatrix} 1 & t(a) \\ t(a) & 1 \end{pmatrix}$$

where

$$t(a) = cor((1, a), (a, 1)) = \frac{(1 - \frac{1+a}{2})(a - \frac{1+a}{2}) + (a - \frac{1+a}{2})(1 - \frac{1+a}{2})}{\sqrt{((1 - \frac{1+a}{2})^2 + (a - \frac{1+a}{2})^2)((a - \frac{1+a}{2})^2 + (1 - \frac{1+a}{2})^2)}} = -1$$

unless $a \ne 1$.

Class 7.

Descriptive and inferential statistics.

Statistical models. Linear regression. Explanatory and response variables.

$$Y = \alpha + \beta X + \varepsilon.$$

Y - response vaiable, X - predictor, ε - error term.

$$\varepsilon \sim N(0, \sigma^2).$$

Simple (one predictor) and multiple (several predictors) regressions (according to [AR, Chapter 7]).

Least squares. Derivation of coefficents for simple linear regression via least squares:

$$\beta = cor(\overline{x}, \overline{y}) \frac{\sigma(\overline{y})}{\sigma(\overline{x})}$$
$$\alpha = m(\overline{y}) - \beta m(\overline{x}).$$

Residuals. Standard deviation of residuals and test of residuals for normality as criteria for "goodness" of a linear model.

Class 8.

Generalized additive models (according to [Cr2, pp. 146–148]).

Class 9.

Clustering: Hirearchical, K-means, gravitational algorithms.

Examples: genetic analysis; transportation, traffic.

Class 10.

Null and alternative hypotheses. Hypotheses testing, p-values (according to Pruim, pp. 71 onwards).

References

[AR] J. Albert and M. Rizzo, *R by Example*, Springer, 2012.

[Ch] C.-H. Chen, *Generalized association plots: information visualization via iterative generated correlation matrices*, Statistica Sinica 12 (2002), 7–29.

[CC] Y. Cohen and J.Y. Cohen, *Statistics and Data with R*, Wiley, 2008.

[Cr1] M.J. Crawley, *The R Book*, 2nd ed., Wiley, 2013.

[Cr2] ———, *Statistics: An Introduction Using R*, 2nd ed., Wiley, 2014.

[D] P. Dalgaard, *Introductory Statistics with R*, 2nd ed., Springer, 2008.

[KYHT] P. Klimek, Y. Yegorov, R. Hanel, and S. Thurnera, *Statistical detection of systematic election irregularities*, Proc. Nat. Acad. Sci. USA **109** (2012), 16469–16473.

[KSP] D. Kobak, S. Shpilkin, and M.S. Pshenichnikov, *Statistical anomalies in 2011-2012 Russian elections revealed by 2D correlation analysis*, arXiv:1205.0741v2.

[SR] M.V. Simkin and V.P. Roychowdhury, *An introduction to the theory of citing*, Significance **3** (2006), №4, 179–181; arXiv:math/0701086.

[T] T. Tao, *Compactness and Contradiction*, AMS, 2013.

[W] S. Wang, *The great gerrymander of 2012*, The New York Times, February 2, 2013.

Wikipedia articles

[$\mathrm{C_w}$] *Confidence interval.*
[$\mathrm{K_w}$] *Kurtosis.*
[$\mathrm{Si_w}$] *Simpson's paradox.*
[$\mathrm{Sk_w}$] *Skewness.*

Email address: pasha.zusmanovich@gmail.com