Lecture notes for the week between 23. oraz 30. April

Different estimators for the variance

Let us assume that we are given several independent observations from the same distribution $\mathrm{N}(\mu,\sigma^2)$. It is known[1] that the variable $\dfrac{nS_\mu^2}{\sigma^2}$ has $\chi^2(n)$ distribution and[2] $\dfrac{nS^2}{\sigma^2} \sim \chi^2(n-1)$.
Let us also mention that $\chi^2(k) \equiv \mathrm{Gamma}(1/2, k/2)$. Hence for the distribution $\chi^2(k)$: $M_{\chi^2(k)}(t) = (1-2t)^{-k/2}$. This helps us to formulate the following theorem which is easy to prove

Theorem 1. *If the random variable $X \sim \chi^2(k)$, then $\mathrm{E}(X) = k$ and $\mathrm{V}(X) = 2k$.*

Let's consider three estimators for variance σ^2, namely $S_{n-1}^2 = \dfrac{1}{n-1}\sum_{k=1}^{n}\left(X_k - \bar{X}\right)^2$, $S_n^2 = \dfrac{1}{n}\sum_{k=1}^{n}\left(X_k - \bar{X}\right)^2$ and $S_{n+1}^2 = \dfrac{1}{n+1}\sum_{k=1}^{n}\left(X_k - \bar{X}\right)^2$. From the initial remark, the variables $\dfrac{(n-1)S_{n-1}^2}{\sigma^2}$, $\dfrac{nS_n^2}{\sigma^2}$, $\dfrac{(n+1)S_{n+1}^2}{\sigma^2}$ have $\chi^2(n-1)$ distribution each.

Expected value $\mathrm{E}\left(\dfrac{nS_n^2}{\sigma^2}\right) = n-1$. Hence, $\mathrm{E}(S_n^2) = \dfrac{n-1}{n}\,\sigma^2$.

Therefore an approximate for the expected value of S^2 is different from the value of the parameter σ^2. We call such an estimator as **a biased estimator**. If $n \to \infty$, then $\mathrm{E}(S^2) \to \sigma^2$. We say that, S^2 is an estimator of **asymptotically unbiased** parameter σ^2.

Because $S_{n-1}^2 = \dfrac{n}{n-1}\,S_n^2$, therefore $\mathrm{E}(S_{n-1}^2) = \sigma^2$. We say that S_{n-1}^2 is **an unbiased estimator** of the parameter σ^2. Similarly, $S_{n+1}^2 = \dfrac{n}{n+1}\,S_n^2$ and $\mathrm{E}(S_{n+1}^2) = \dfrac{n-1}{n+1}\,\sigma^2$ (S_{n+1}^2 is a biased estimator for σ^2). Therefore the best estimator (taking the expected value into account) is S_{n-1}^2, the worst one is S_{n+1}^2.

Let's compare the variances of the considered estimators. We know that $\dfrac{n\,S_n^2}{\sigma^2} \sim \chi^2(n-1)$. From this it follows that $\mathrm{V}(S_n^2) = \dfrac{2(n-1)}{n^2}\,\sigma^4$, $\mathrm{V}(S_{n-1}^2) = \dfrac{2(n-1)}{(n-1)^2}\,\sigma^4$ and $\mathrm{V}(S_{n+1}^2) = \dfrac{2(n-1)}{(n+1)^2}\,\sigma^4$. The smaller the variance, the more the variable is said to be "stable". The best estimator (based on the variance) is S_{n+1}^2, the worst S_{n-1}^2.

Maximum likelihood estimators (MLE estimators)

We assume that the data are independent observations of a random variable X with the same distribution. The density function of the variable X is $f(x;\theta)$, where θ is the parameter/parameters of the distribution. You can also think that the data are independent variables $X_1,\ldots,X_n$ with the same distribution. Therefore, the event probability can be written as the likelihood function L relative to the variable θ

$$L(x;\theta) = P\left(X_1 = x_1,\ldots,X_n = x_n\right) = \prod_{k=1}^{n} f(x_k;\theta). \tag{1}$$

We assume that the observed values are typical (most likely). Therefore, we want the likelihood function to reach its maximum at some point $\hat{\theta}$. The calculated value of $\hat{\theta}$ is called the *estimator*

[1] Note 6, formula (3).

[2] Note 6, Theorem 4.

of the highest likelihood for the parameter θ.
ATTENTION: We often look to maximize the function $\ln L(x;\theta)$, only for computational reasons; logarithm (natural) as an increasing function gives the same answer.

Example:
Ex1: Consider n independent observations from the $\mathrm{Exp}(\lambda)$ distribution. The probability of the event $P(X_1 = x_1, \ldots . X_n = x_n)$ is equal to – due to independence – $L(\lambda) = \lambda^n \exp\left(-\sum_{k=1}^{n} x_i\right)$. We want to compute the value of λ that maximizes the function $L(\lambda)$.[a]

The functions $L(\lambda)$ and $\ln L(\lambda)$ attains the maximum value at the same point $\hat{\lambda}$. By computing the derivative $\dfrac{\partial \ln L}{\partial \lambda}$ and equating it to zero - we get the equation

$$\frac{\partial \ln L}{\partial \lambda} = \frac{n}{\lambda} + \left(\sum_{k=1}^{n} x_i\right) = \frac{n}{\lambda} - n \cdot \bar{x} = 0,$$

where it follows that $\hat{\lambda} = {}^1\!/\bar{x}$. The second derivative of the function $\ln L(\lambda)$ is equal to $\dfrac{\partial^2 \ln L}{\partial \lambda^2} = -\dfrac{n}{\lambda^2}$ which is ≤ 0 at every point λ, i.e. also at the previously determined value $\hat{\lambda}$, which proves that we found the maximum likelihood function $\equiv$ MLE estimator $\hat{\lambda}$ for the parameter λ.

[a] Intuition: What we observe is the most likely outcome.

Example:
Ex2: We consider n independent observations from the $\mathrm{B}(n,p)$ distribution. The likelihood function now has the form

$$\begin{aligned} L(p) &= P(X_1 = x_1, \ldots, X_k = x_k) = \textstyle\prod_{i=1}^{k} P(X_i = x_i) = \\ &= p^{\sum x_i} (1-p)^{nk - \sum x_i} \textstyle\prod_{i=1}^{k} \binom{n}{x_i} = p^{k\bar{x}} (1-p)^{nk - k\bar{x}} \textstyle\prod_{i=1}^{k} \binom{n}{x_i}. \end{aligned} \tag{2}$$

Therefore the logarithm of the likelihood function has the form

$$\ln L(p) = \sum_{i=1}^{k} \ln\binom{n}{x_i} + k\bar{x} \ln p + k(n - \bar{x}) \ln(1-p),$$

and its derivative is

$$\frac{\partial \ln L}{\partial p} = \frac{k\bar{x}}{p} - \frac{k(n-\bar{x})}{1-p} = 0.$$

Solving the above equation, we get the expression $\hat{p} = \dfrac{\bar{x}}{n}$ for the MLE estimator. The second derivative of the likelihood function is

$$\frac{\partial^2 \ln L}{\partial p^2} = -\frac{k\bar{x}}{p^2} - \frac{k(n-\bar{x})}{(1-p)^2} < 0,$$

for $0 < \hat{p} < n$, we have found the maximum of $L(p)$.

> If $\hat{p} = 0$, then $x_1 = \ldots = x_k = 0$. The likelihood function (2) has the form $L(p) = (1-p)^{nk}$ and reaches the maximum for $p = 0$. Similarly, if $\hat{p} = n$, then $x_1 = \ldots = x_k = n$. The likelihood function (2), in this case, has the form $L(p) = p^{nk}$ and reaches the maximum for $p = 1$.

ANalysis Of VAriance - ANOVA

Suppose that the data are independent observations of the random variable $X \sim N(\mu, \sigma^2)$. We group the observations by a certain quality feature, we distinguish I groups. We have J observations for each group. The symbol x_{ij} represents the j-th observation in the i-th group $(\ i = 1, \ldots, I;\ \ j = 1, \ldots, J\ )$.

Group	Observations				Mean
1	x_{11}	x_{12}	$\ldots$	x_{1J}	$x_{1\bullet}$
2	x_{11}	x_{12}	$\ldots$	x_{1J}	$x_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
I	x_{I1}	x_{I2}	$\ldots$	x_{IJ}	$x_{I\bullet}$

The last column of the above table contains the average of the groups (rows), i.e., $x_{k\bullet} = \dfrac{1}{J} \sum\limits_{j=1}^{J} x_{kj}$.

The symbol $\bar{x}$ denotes the mean of all observations: $\bar{x} = \dfrac{1}{IJ} \sum\limits_{i=1}^{I} \sum\limits_{j=1}^{J} x_{ij}$. Examples of data grouping: 3 tire groups (winter, summer, universal) and we note the degree of wear after a specified mileage; the effectiveness of a certain drug in groups: the initial stage of the disease, the disease in its full manifestation, severe condition; we compare similar drugs from 3 manufacturers, etc.

We assume - according to the initial assumption - that each of the random variables $X_{i\bullet}$ has $N\left(\mu, {}^{\sigma^2}\!/_{n}\right)$ distribution. In fact, we want to reject this assumption, to say from the data that one of the groups (or some, or many) is different from the others. The next calculations should indicate which group stands out "on the plus side", but for now we postpone it. What interests us is the answer of the form: Tires of type A are better than the others; some medicine is best suited for some stage of the disease; producer C has the best product.

An interesting fact is that we can say something about the mean of each group $(x_{i\bullet})$ taking into consideration the variance of all observations and variance within the groups (rows). Let's start with the variance of all observations. We have used the formula many times.

$$\begin{array}{ccccc|c}
\sum\limits_{k=1}^{n} (X_k - \mu)^2 & = & \sum\limits_{k=1}^{n} \left(X_k - \bar{X}\right)^2 & + & n \cdot \left(\bar{X} - \mu\right)^2 & \text{formula} \\
\dfrac{nS_\mu^2}{\sigma^2} & = & \dfrac{nS^2}{\sigma^2} & + & \left(\dfrac{\bar{X} - \mu}{\sigma\sqrt{n}}\right)^2 & \text{shortcuts} \\
\chi^2(n) & = & \chi^2(n-1) & + & \chi^2(1) & \text{distribution}
\end{array} \tag{3}$$

The basic fact is: for observation x_{ij} the random variable $\dfrac{nS^2}{\sigma^2} = \sum\limits_{ij} \left(X_{ij} - \bar{X}\right)^2$ has $\chi^2(IJ-1)$ distribution, because we have I groups with J observation each and one degree of freedom "disappears", and it follows from the formula (3).

We present the diversity of observations as $\mathrm{SS_{Tot}} = \sum\limits_{i,j} (x_{ij} - \bar{x})^2$. Up to constant we have the variance of all observations. [3] Therefore: $\dfrac{\mathrm{SS_{Tot}}}{\sigma^2} \sim \chi^2(IJ - 1)$.

[3] SS $\equiv$ sum of squares.

Group diversity (**between groups variation**) can be expressed by the mean of the groups $x_{i\bullet}$. Because the variables X_{ij} are independent, the variables $X_{i\bullet}$ are also independent. Therefore we have I independent random variables $X_{1\bullet}, \ldots, X_{I\bullet}$. The mean $\bar{X}$ of all observations is also the average taken from the mean of the individual groups. [4] Treating the group as a "generalized observation" we find that the expression $\mathrm{SSA} = J \cdot \sum_{i=1}^{I} (x_{i\bullet} - \bar{x})^2$ – up to a constant – has $\chi^2(I-1)$ distribution. The second component of variation remains to be considered, the size $\mathrm{SSE} = \sum_{i,j} (x_{ij} - x_{i\bullet})^2$, is called **within groups variation**.

Theorem 2.

$$\mathrm{SS}_{\mathrm{Tot}} = \mathrm{SSA} + \mathrm{SSE}. \tag{4}$$

COMMENTS:

- The claim of the theorem is: the total variance is divided into the sum of the variances between groups and the variances within groups.
- If most of the variances are within groups, then we are inclined to consider that the average of the groups are the same (or close to each other).
- Inversely: if the variance between the groups prevails over the variance within the groups then we can think that the average of the groups differ.
- In summary: based on variance (or rather its division into two components), we can draw conclusions about the mean within groups.

Proof.

$$\begin{aligned}\mathrm{SS}_{\mathrm{Tot}} &= \sum_{i,j} (x_{ij} - \bar{x})^2 = \sum_{i,j} (x_{ij} - x_{i.\bullet} + x_{i\bullet} - \bar{x})^2 = \\ &= J \cdot \sum_{i} (x_{i\bullet} - \bar{x})^2 + \sum_{i,j} (x_{ij} - x_{i.\bullet})^2 + 2 \cdot \sum_{i,j} (x_{ij} - x_{i.\bullet}) \cdot (x_{i\bullet} - \bar{x}).\end{aligned}$$

The third component in the last equality can be transformed into the following form

$$\begin{aligned}&\sum_{i,j} (x_{ij} - x_{i.\bullet}) \cdot (x_{i\bullet} - \bar{x}) = \sum_{i} (x_{i\bullet} - \bar{x}) \sum_{j} \cdot (x_{ij} - x_{i.\bullet}) = \\ &= (x_{i\bullet} - \bar{x}) \cdot (n \cdot x_{i\bullet} - n \cdot x_{i\bullet}) = 0.\end{aligned}$$

From this, we have

$$\mathrm{SS}_{\mathrm{Tot}} = \sum_{i,j} (x_{ij} - \bar{x})^2 = J \cdot (x_{i\bullet j} - \bar{x})^2 \sum_{i,j} + (x_{ij} - x_{i.\bullet})^2 = \mathrm{SSA} \ + \ \mathrm{SSE}. \tag{5}$$

□

Two-way ANOVA

Suppose that the data are independent observations of the random variable $X \sim N(\mu, \sigma^2)$. The observations are grouped according to the quality feature (factor) A and quality feature B, and we distinguish I and J groups, respectively. We have one observation for each combination of the groups.[5] The variable x_{ij} denotes i,j-th observation for which the feature A has i-th value, while the feature B has j-th value ($i = 1, \ldots, I$; $j = 1, \ldots, J$).

[4] All groups have the same number of observations

[5] We are talking about 2-factor ANOVA analysis without repetition.

Group	1	2	...	J	Mean
1	x_{11}	x_{12}	...	x_{1J}	$x_{1\bullet}$
2	x_{11}	x_{12}	...	x_{1J}	$x_{2\bullet}$
⋮	⋮	⋮		⋮	⋮
I	x_{I1}	x_{I2}	...	x_{IJ}	$x_{I\bullet}$
Mean	$x_{\bullet 1}$	$x_{\bullet 2}$	...	$x_{\bullet J}$	

The symbols $x_{i\bullet}, x_{\bullet j}$ represent – respectively – the mean value of the i-th group of feature A and the mean value of j-th group of feature B. The symbol $\bar{x}$ represents the mean of all observations. In addition, assume

$$\begin{aligned} \mathrm{SSTot} &= \sum_{ij}(x_{ij}-\bar{x})^2, & \mathrm{SSA} &= J\cdot\sum_{i}(x_{i\bullet}-\bar{x})^2 \\ \mathrm{SSB} &= I\cdot\sum_{j}(x_{\bullet j}-\bar{x})^2, & \mathrm{SSE} &= \sum_{ij}(x_{ij}-x_{i\bullet}-x_{\bullet j}+\bar{x})^2. \end{aligned} \tag{6}$$

Theorem 3.

$$\mathrm{SSTot} = \mathrm{SSA} + \mathrm{SSB} + \mathrm{SSE}. \tag{7}$$

Proof. Note that $\mathrm{SSTot} = \sum_{ij}(x_{ij}-\bar{x})^2 = \sum_{ij}\left(\underbrace{x_{ij}-x_{i\bullet}-x_{\bullet j}+\bar{x}}_{(a)}+\overbrace{x_{i\bullet}-\bar{x}}^{(b)}+\underbrace{x_{\bullet j}-\bar{x}}_{(c)}\right)^2.$

Note that the sum of the squares of expressions denoted as $(a),(b),(c)$ give the components SSA, SSB, SSE to the right of the equation (7). Therefore it remains to be shown that the sum of the products of $(a)\cdot(b),(a)\cdot(c),(b)\cdot(c)$ result in 0.

$(b)\cdot(c) = \sum_{ij}(x_{i\bullet}-\bar{x})\cdot(x_{\bullet j}-\bar{x}) = \sum_{i}(x_{i\bullet}-\bar{x})\cdot\sum_{j}(x_{\bullet j}-\bar{x}) = 0$, bo $\sum_{i}(x_{i\bullet}-\bar{x}) = I\cdot\bar{x} - I\cdot\bar{x} = 0$.

$(a)\cdot(b) = \sum_{ij}(x_{i\bullet}-\bar{x})\cdot(x_{ij}-x_{i\bullet}-x_{\bullet j}+\bar{x}) = \sum_{i}(x_{i\bullet}-\bar{x})\cdot\sum_{j}(x_{ij}-x_{i\bullet}-x_{\bullet j}+\bar{x}) = (\star)$

For a fixed i, let's consider the internal sum over j: $\sum_{j}(x_{ij}-x_{i\bullet}) = 0$. Hence

$$(\star) = \sum_{i}(x_{i\bullet}-\bar{x})\cdot\sum_{j}(\bar{x}-x_{\bullet j}) = 0.$$

The proof for the products of the form $(a)\cdot(c)$ is practically the same, just switch the indices i, j. □

Random number generator from $\mathbf{N}(0,1)$ distribution

Example:

Let's assume that we have a random number generator from the $U[0,1]$ distribution and we draw two values u_1, u_2. Therefore the two-dimensional random variable (U_1,U_2) has a distribution with the density function $f_{U_1,U_2}(u_1,u_2) = 1$ for $((u_1,u_2)\in[0,1]\times[0,1]$. Let us consider the new variables $Y_1 = -2\ln U_1$, $Y_2 = 2\pi\,U_2$. Obviously, $Y_1\in[0,\infty)$ and $Y_2\in[0,2\,\pi)$. Interpreting Y_1,Y_2 as the polar coordinates of a point on the plane, we can say that we draw the square of the radius and the point's argument. Let's determine the density $f_{Y_1,Y_2}(y_1,y_2)$ of the variable (Y_1,Y_2).

$$\left\{\begin{array}{l} U_1 = \exp\left(-\dfrac{Y_1}{2}\right) \\ U_2 = \dfrac{Y_2}{2\,\pi} \end{array}\right., \text{ where } \operatorname{abs}(J) = \operatorname{abs}\left(\begin{vmatrix} -\dfrac{1}{2}\exp\left(-\dfrac{Y_1}{2}\right) & 0 \\ 0 & \dfrac{1}{2\,\pi} \end{vmatrix}\right) = \frac{1}{4\cdot\pi}\exp\left(-\frac{Y_1}{2}\right). \tag{8}$$

In the above formula, we want to calculate the absolute value from the Jacobian determinant. Unfortunately, both operations (absolute value and determinant) are often marked with the same $|\ |$. As a result: $\operatorname{abs}(\det(A)) \equiv ||A||$ – which in turn could suggest that we're talking about the **norm** of the matrix A.
For the density $f_{Y_1,Y_2}(y_1,y_2)$ we have the formula

$$f_{Y_1,Y_2}(y_1,y_2) = \frac{1}{4\cdot\pi}\exp\left(-\frac{Y_1}{2}\right). \tag{9}$$

From the polar coordinates (Y_1,Y_2) let's now go to the Cartesian coordinates (X_1,X_2), i.e.

$$\begin{cases} X_1 = \sqrt{Y_1}\,\cos Y_2 \\ X_2 = \sqrt{Y_1}\,\sin Y_2 \end{cases}, \text{ i and } J = \begin{vmatrix} \dfrac{\cos Y_2}{2\sqrt{Y_1}} & -\sqrt{Y_1}\sin Y_2 \\ \dfrac{\sin Y_2}{2\sqrt{Y_1}} & \sqrt{Y_1}\cos Y_2 \end{vmatrix} = \frac{1}{2}. \tag{10}$$

We finish the transformation with two remarks:

1. The Jacobian computed above should be reversed. We customarily compute Jacobian of the "old" variables relative to the "new" ones. Here: it is more convenient to determine the inverse Jacobian of the "new" variables compared to the "old" variables.
2. We also use the relationship: $Y_1 = X_1^2 + X_2^2$ [a].

The final result ≡ density $f_{X_1,X_2}(x_1,x_2)$ is:

$$f_{X_1,X_2}(x_1,x_2) = \frac{1}{2\pi}\exp\left(-\frac{x_1^2+x_2^2}{2}\right) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x_1^2}{2}\right)\times\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x_2^2}{2}\right), \tag{11}$$

which means that the variables X_1, X_2 are independent and have $\mathrm{N}(0,1)$ distribution each.

[a] equation (10)

[Popular|favourite] formulas and distributions

1. Suppose the random variables X, Y are independent and subject to $X \sim \chi^2(n)$, $Y \sim \chi^2(k)$ distributions. Then the random variable $Z = X + Y$ has $Z \sim \chi^2(n+k)$ distribution.
2. Suppose the variable X is distributed by $\mathrm{N}(\mu,\sigma^2)$. Let additionally $Y = \dfrac{X-\mu}{\sigma}$. Fact: $X \sim \mathrm{N}(\mu,\sigma^2) \iff Y \sim \mathrm{N}(0,1)$.
3. $\mathrm{Gamma}\,(1/2, n/2) \equiv \chi^2(n)$.
4. Let's assume that the variables $X_1,\ldots,X_n$ are independent and are subject to the $\mathrm{N}(\mu,\sigma^2)$ distribution each. Then the variables $Z = \sum_{k=1}^{n}\left(\dfrac{X_k-\mu}{\sigma}\right)^2$ have $\chi^2(n)$ distribution.
5. The independent variables X, Y have the $X \sim \chi^2(k)$, $Y \sim \chi^2(l)$ distribution respectively. We say that the variable $F(k,l) = \dfrac{X}{Y}\cdot\dfrac{l}{k}$ has a F-Fisher distribution with (k,l) degrees of freedom.
6. The independent variables X, Y have the distributions $X \sim \mathrm{N}(0,1)$, $Y \sim \chi^2(k)$ respectively. We say that the variable $t(k) = \dfrac{X}{\sqrt{Y/k}}$ has the t-Student distribution with k degrees of freedom.

7. Intuition: the quotient of two independent and the normalized χ^2 distributions is the F-Fisher distribution and the quotient of the standard normal distribution and the square root of normalized χ^2 distribution is the t -Student distribution.
8. Let's assume that the variables $X_1, \ldots, X_n$ are independent and are subject to the $\mathrm{N}(\mu, \sigma^2)$ distribution each. Additionally assume that $S_\mu^2 = \frac{1}{n}\sum_{k=1}^{n} (X_k - \mu)^2$. Then $\frac{nS_\mu^2}{\sigma^2} \sim \chi^2(n)$.
9. Let's assume that the variables $X_1, \ldots, X_n$ are independent and are subject to the $\mathrm{N}(\mu, \sigma^2)$ distribution each. Additionally assume that $S^2 = \frac{1}{n}\sum_{k=1}^{n} \left(X_k - \bar{X}\right)^2$. Then $\frac{nS^2}{\sigma^2} \sim \chi^2(n-1)$.
10. …one thousand and one formulas (like in oriental fairy tales).

↤

Regards,
Pratik Ghosal & Witold Karczewski