1. Доверительные интервалы и однородность нескольких выборок

1.1. Построение доверительных интервалов

Определение 1. Интервал, накрывающий истинное значение параметра распределения с заданной вероятностью P называется $P \cdot 100$**%-доверительным интервалом.**

В случае нормально распределенной выборки $x_1, \ldots, x_n$ при известной дисперсии σ можно построить доверительный интервал для генерального среднего μ на основании теоремы Фишера,

$$\bar{x} \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \iff \frac{\bar{x} - \mu}{\sigma}\sqrt{n} \sim \mathcal{N}(0, 1)\,.$$

В случае произвольного распределения это выражение можно использовать при достаточно большом объеме выборки вследствие ЦПТ. Зададим уровень значимости α. Для стандартно нормально распределенной случайной величины ξ справедливо

$$P\{|\xi| < z_{1-\frac{\alpha}{2}}\} = 1 - \alpha,$$

где через $z_{1-\frac{\alpha}{2}}$ обозначена $(1 - \frac{\alpha}{2})$- квантиль нормального распределения $\mathcal{N}(0, 1)$. Подставим $\xi = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ в выражение

$$P\{-z_{1-\frac{\alpha}{2}} < \xi \le z_{1-\frac{\alpha}{2}}\} = 1 - \alpha,$$
$$P\{-z_{1-\frac{\alpha}{2}} < \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \le z_{1-\frac{\alpha}{2}}\} = 1 - \alpha$$

и построим неравенство для μ

$$P\left\{\bar{x} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu \le \bar{x} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right\} = 1 - \alpha\,.$$

В случае неизвестной дисперсии построение доверительного интервала осуществляется аналогичным образом, – параметр σ^2 заменяется на несмещенную оценку S^2, а так как $\frac{\bar{x}-\mu}{S/\sqrt{n}} \sim \mathbf{T}(n-1)$, то используется квантиль распределения Стьюдента $T_P^{(df)}$ с числом степеней свободы $df = n-1$,

$$P\left\{\bar{x} - T_{1-\frac{\alpha}{2}}^{(n-1)}\frac{S}{\sqrt{n}} < \mu \leq \bar{x} + T_{1-\frac{\alpha}{2}}^{(n-1)}\frac{S}{\sqrt{n}}\right\} = 1-\alpha\,.$$

1.2. Теорема Фишера-Кочрена

Теорема 1. *(Fisher-Cochran) Пусть $Y = (y_1, \ldots, y_n)$ с независимыми компонетами, $y_i \sim \mathcal{N}(0,1)$, квадратичные формы $Q_1, \ldots, Q_k$ рангов $n_1, \ldots, n_k$, такие что*

$$Y^T Y = Q_1 + \ldots + Q_k\,.$$

Тогда $n = n_1 + \ldots + n_k$ равносильно $Q_i \sim \chi^{\mathbf{2}}(\mathbf{n_i})$ и $Q_1, \ldots, Q_k$ независимы.

Доказательство.

Пусть $Q_i = Y^T A_i Y$, где матрица A_i ранга n_i. Тогда существует матрица B_i размерности n_i на n, при помощи которой квадратичная форма приводится к диагональному виду, то есть $Q_i = (B_i Y)^T \Delta_i (B_i Y)$, где Δ_i диагональная матрица с элементами ± 1. Соберем k матриц B_i в одну матрицу B размерности n на n.

$$[B_1^T(n,n_1)|\ldots|B_k^T(n,n_k)] \cdot \begin{bmatrix} \Delta_1 & \ldots & \mathbf{0} \\ \vdots & \ldots & \vdots \\ \mathbf{0} & \ldots & \Delta_k \end{bmatrix} \cdot \begin{bmatrix} B_1(n_1,n) \\ \vdots \\ B_k(n_k,n) \end{bmatrix} = \sum_i B_i^T \Delta_i B_i$$

Следовательно,

$$Y^T Y = \sum_i Q_i = \sum_i (B_i Y)^T \Delta_i (B_i Y) =$$
$$= Y^T(\sum_i B_i^T \Delta_i B_i) Y = Y^T B^T \Delta B Y, \;\Rightarrow\; B^T \Delta B = \mathbf{I}\,.$$

Ранг матрицы B должен быть равен n, так как с одной стороны, он не может быть больше n, а с другой ранг произведения матриц не превышает минимального ранга сомножителей,

$$n = rank((B^T\Delta)B) \leq \min(rank(B^T\Delta), rank(B).$$

Тогда $\Delta = (B^T)^{-1}B^{-1}$ положительно определенная матрица, следовательно, $\Delta = \mathbf{I}$ и матрица B является ортогональной. Тогда компоненты вектора $X = BY$ независимы и нормальны, кроме того

$Y^TY = X^TX$.

$$Q_1 = x_1^2 + \ldots x_{n_1}^2,$$
$$Q_2 = x_{n_1+1}^2 + \ldots x_{n_1+n_1}^2,$$
$$Q_3 = x_{n_1+n_2+1}^2 + \ldots x_{n_1+n_2+n_3}^2,$$
$$\ldots$$

Осюда все Q_i независимы и имеют распределение $\chi^{\mathbf{2}}(\mathbf{n_i})$ соответственно. Таким образом достаточность $n = n_1 + \ldots + n_k$ установлена. Необходимость очевидна.

1.3. Однофакторный дисперсионный анализ

Пусть имеется выборка наблюдений, которая разбивается на r групп. Каждая группа содержит n_i величин $\mathcal{N}(\mu_i, \sigma)$, где σ не зависит от i, $n_1 + \ldots + n_r = n$. Требуется проверить о равенстве средних $H_0 : \mu_1 = \ldots = \mu_r$.

- Выборка представляет собой вектор размерности n вида

$$x = (x_{11}, \ldots, x_{1n_1}, \ldots, x_{r1}, \ldots, x_{rn_r})^T; \tag{1}$$

- x_{ij} — j-я величина ($j = 1, \ldots, n_i$) в i-й группе ($i = 1, \ldots, r$);
- $\bar{x}_i = \frac{1}{n_i}\sum\limits_{j=1}^{n_i} x_{ij}$ — среднее в i-й группе;
- $\bar{x} = \frac{1}{n}\sum\limits_{i=1}^{r}\sum\limits_{j=1}^{n_i} x_{ij}$ — общее среднее.

Разбиение суммы квадратов отклонений

Покажем, что общий источник вариации Q в виде суммы квадратов отклонений от общего среднего можно разложить на сумму источников вариации, обусловленных различием между группами Q_1 и внутри групп

Q_2.

$$Q=\sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij}-\bar{x})^2=\sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i+\bar{x}_i-\bar{x})^2=$$

$$=\underbrace{\sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2}_{Q_2}+\underbrace{\sum_{i=1}^{r}\sum_{j=1}^{n_i}(\bar{x}_i-\bar{x})^2}_{Q_1},\text{ так как}$$

$$\sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)(\bar{x}_i-\bar{x})=\sum_{i=1}^{r}(\bar{x}_i-\bar{x})\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)=0\,.$$

Предложение 1. *Ранги квадратичных форм Q, Q_1 и Q_2 равны $n-1$, $r-1$ и $n-r$ соответственно.*

Для доказательства введем преобразование $y=Ax$ вектора x из (1) при помощи ортогональной матрицы A, $A^{-1}=A^T$, у которой последняя строка имеет вид

$$\frac{1}{\sqrt{n}}+\ldots+\frac{1}{\sqrt{n}}.$$

Тогда $x=A^{-1}y=A^Ty$, $\sum_{ij}x_{ij}^2=x^Tx=(A^Ty)^TA^Ty=y^TAA^Ty=y^Ty=\sum\limits_{k=1}^{n}y_k^2$, а квадраточная форма

$$Q=\sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij}-\bar{x})^2=\sum_{ij}x_{ij}^2-n\bar{x}^2=$$

$$=\sum_{ij}x_{ij}^2-\left(\frac{x_1}{\sqrt{n}}+\ldots+\frac{x_n}{\sqrt{n}}\right)^2=\sum_{k=1}^{n}y_k^2-y_n^2=\sum_{k=1}^{n-1}y_k^2$$

имеет ранг $n-1$. Источник вариации, обусловленный различием между группами, можно представить в виде

$$Q_1=\sum_{i=1}^{r}\sum_{j=1}^{n_i}(\bar{x}_i-\bar{x})^2=\sum_{i=1}^{r}L_i^2,\quad\text{где}\quad L_i=\sqrt{n_i}(\bar{x}_i-\bar{x}),$$

$$\sum_{i=1}^{r}L_i\sqrt{n_i}=0,\quad\text{так как}\quad n\bar{x}=\sum_{i=1}^{r}n_i\bar{x}_i\,.$$

Следовательно, $rank(Q_1)\le r-1$. Источник вариации, обусловленный различием внутри групп, можно представить в виде

$$Q_2=\sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2=\sum_{i,j}L_{ij}^2,\text{ где }L_{ij}=x_{ij}-\bar{x}_i,$$

$$\sum_{j=1}^{n_i}L_{ij}=0\quad\forall i=1,\ldots,r.\Longrightarrow\ rank(Q_2)\le n-r.$$

Воспользуемся свойством рангов квадратичных форм

$$rank(Q_1+Q_2)\le rank(Q_1)+rank(Q_2)\ \text{ Следовательно,}$$

$$n-1\le(r-1)+(n-r)=n-1,$$

и ранги Q_1 и $Q-2$ в точности равны $r-1$ и $n-r$. Воспользуемся теоремой Кохрена: *пусть* Q_i, $i=1,\ldots,k$, *— неотрицательные квадратичные формы ранга* r_i,

$$\sum_{i=1}^{n} x_i^2 = Q_1 + \ldots + Q_k.$$

Если $\sum\limits_{i=1}^{k} r_i = n$, *то существует ортогональное преобразование* $x = Cy$, *переводящее все* Q_i *в суммы квадратов такого вида:*

$$Q_1 = \sum_{i=1}^{r_1} y_i^2,\ Q_2 = \sum_{i=r_1+1}^{r_1+r_2} y_i^2, \ldots,\ Q_k = \sum_{i=n-r_k+1}^{n} y_i^2,$$

где никакая пара форм не содержит общей переменной y_i. *Из* $Q = Q_1 + Q_2$ *при помощи ортогонального преобразования получаем*

$$\sum_{i=1}^{n-1} y_i^2 = \sum_{i=1}^{r-1} y_i^2 + \sum_{i=r}^{n-1} y_i^2.$$

Из независимости $x_{11},\ldots,x_{rn_r}$ *следует независимость* $y_1,\ldots,y_{n-1}$, *отсюда,* Q_1 *и* Q_2 *независимы.*

Статистика критерия Фишера

Предположим, что нулевая гипотеза $H_0 : \mu_1 = \ldots = \mu_r = \mu$ верна. Тогда $x_{ij} = \mu + \xi_{ij}$, где $\xi_{ij} \sim \mathcal{N}(0,1)$ независимы.

$$Q = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij} - \bar{x})^2 = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(\xi_{ij} - \bar{\xi})^2,$$

$$Q_1 = \sum_{i=1}^{r} n_i(\bar{x}_i - \bar{x})^2 = \sum_{i=1}^{r} n_i(\bar{\xi}_i - \bar{\xi})^2,$$

$$Q_2 = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2 = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(\xi_{ij} - \bar{\xi}_i)^2.$$

При ортогональном преобразовании получаем, что

$$Q = \sum_{k=1}^{n-1}\eta_k^2, \quad Q_1 = \sum_{k=1}^{r-1}\eta_k^2, \quad Q_2 = \sum_{k=r}^{n-1}\eta_k^2,$$

где $\eta_k \sim \mathcal{N}(0,\sigma)$ независимы. Таким образом, при справедливости нулевой гипотезы $\frac{Q_1}{\sigma^2}$ и $\frac{Q_2}{\sigma^2}$ имеют распределение хи-квадрат с числом степеней свободы соответственно $r-1$ и $n-r$. Отсюда получаем статистику

$$F = \frac{\frac{1}{r-1}Q_1}{\frac{1}{n-r}Q_2} \ \sim\ F(r-1, n-r)\,,$$

которая используется на практике для проверки гипотезы однородности в случае нескольких групп.

Пример кода и его результата в R: $summary(aov(yield \sim block, npk))$

	Df	$SumSq$	$MeanSq$	$Fvalue$	$Pr(>F)$
$block$	5	343.3	68.66	2.318	0.0861
$Residuals$	18	533.1	29.61		

Различие между средними по блокам можно считать значимым при уровне значимости $\alpha = 0.1$.

1.4. Линейная модель с фиксированными эффектами

Статистике Фишера для проверки равенства средних эквивалентна статистика критерия равенства нулю фиксированныхэффектов в линейной модели. Однако линейная модель имеет свои преимущества, поскольку ее проще обобщить на случаи большего числа факторов. Предполагаем, что у нас есть r нормально распределенных популяций $\mathcal{N}(\mu_1, \sigma), \ldots, \mathcal{N}(\mu_r, \sigma)$, где через μ_i обозначены генеральные средние внутри каждой популяции. Переменная x_{ij} означает j-е наблюдение в i-й подпопуляции. Это предположение можно записать в виде:

$$x_{ij} = \mu_i + e_{ij}, \quad i = 1, \ldots, r, \ j = 1, \ldots, n_i, \tag{2}$$

где "ошибки" e_{ij} независимы и распределены нормально по $\mathcal{N}(0, \sigma)$. Соотношения (2) представляют собой одну из форм модели дисперсионного анализа.

Во многих случаях желательно выразить i-е среднее μ_i в виде суммы генерального среднего μ и дифференциальных или главных эффектов α_i, определяемых для каждой подпопуляции. Перепишем модель однофакторного дисперсионного анализа в виде

$$x_{ij} = \mu + \alpha_i + e_{ij}, \quad i = 1, \ldots, r, \ j = 1, \ldots, n_i. \tag{3}$$

Наилучшими оценками параметров модели (3) в смысле метода наименьших квадратов являются

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n_i} x_{ij}, \quad \hat{\alpha}_i = \bar{x}_{i\cdot} - \bar{x} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij} - \bar{x}.$$

Ошибкой в этой модели является выражение

$$Q_2 = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij} - \hat{\mu} - \hat{\alpha}_i)^2 = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_{i\cdot})^2.$$

Ее число степеней свободы равно $\nu_2 = n - r$. Принятие гипотезы $H_0 : \alpha_1 = \ldots = \alpha_r = 0$ означает справедливость эквивалентной гипотезы $H_0 : \mu_1 = \ldots = \mu_r = \mu$, что все средние по подпопуляциям равны генеральному среднему. Для проверки этой гипотезы используется критерий Фишера. Приведем обоснование этого критерия в рамках общей линейной модели.

При справедливости нулевой гипотезы $H_0 : \alpha_1 = \ldots = \alpha_r = 0$ имеет место усеченная модель вида

$$x_{ij} = \mu + e_{ij}, \; i = 1, \ldots, r; \; j = 1, \ldots, n_i. \tag{4}$$

Наилучшей оценкой параметра модели (4) в смысле метода наименьших квадратов является

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n_i} x_{ij}.$$

Ошибкой $\widetilde{Q_2}$ в этой модели является выражение

$$\widetilde{Q_2} = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij} - \hat{\mu})^2 = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij} - \bar{x})^2 = Q.$$

Ее число степеней свободы равно $\widetilde{\nu_2} = n - 1$. Статистика критерия правдоподобия имеет вид:

$$F = \frac{(\widetilde{Q_2} - Q_2)/(\widetilde{\nu_2} - \nu_2)}{Q_2/\nu_2} = \frac{(Q - Q_2)/(n - 1 - n + r)}{Q_2/(n - r)} = \frac{Q_1/(r - 1)}{Q_2/(n - r)}. \tag{5}$$

Естественно, что ошибка $\widetilde{Q_2}$ в усеченной модели (4) больше, чем ошибка в расширенной модели (3). Если разность $\widetilde{Q_2} - Q_2$ мала, соответственно мало значение статистики F, то усеченная модель не хуже основной модели. Если разность $\widetilde{Q_2} - Q_2$ велика, то соответствие усеченной модели реальным данным гораздо хуже, поэтому влияние фактора, разделяющего наблюдения по подпопуляциям, значимо. Степень значимости определяется при помощи доверительного уровня вероятности

$$\alpha_* = P\{F > F_*\},$$

где F_* наблюдаемое значение статистики F. При $\alpha_* < \alpha = 0.05$ гипотеза $H_0 : \alpha_1 = \ldots = \alpha_r = 0$ отвергается.

1.5. Матричная модель дисперсионного анализа

Модель однофакторного дисперсионного анализа (3)

$$x_{ij} = \mu + \alpha_i + e_{ij}, \quad i = 1, \ldots, r, \; j = 1, \ldots, n_i, \quad \text{где } \sum_{i=1}^{r} \alpha_i = 0,$$

можно записать в матричном виде:

$$Y = X\beta + \epsilon,$$

где $Y = (x_{11}, \ldots, x_{1n_1}, x_{21}, \ldots, x_{2n_2}, \ldots, x_{r1}, \ldots, x_{rn_r})'$ –вектор наблюдений, $\sum\limits_{i=1}^{r} n_i = n$, $\beta = (\mu, \alpha_1, \ldots, \alpha_{r-1})'$ – вектор параметров и матрица плана размерности n на r имеет вид:

$$X = \left[\begin{array}{ccccc} 1 & 1 & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & 1 & 0 & \ldots & 0 \\ \hline 1 & 0 & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & 0 & 1 & \ldots & 0 \\ \hline \ldots & \ldots & \ldots & \ldots & \ldots \\ \hline 1 & -1 & -1 & \ldots & -1 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & -1 & -1 & \ldots & -1 \end{array}\right], \qquad \beta = \left[\begin{array}{c} \mu \\ \alpha_1 \\ \alpha_2 \\ \ldots \\ \alpha_{r-1} \end{array}\right]$$

Оценки параметров по методу наименьших квадратов имеют вид:

$$\hat{\beta} = (X'X)^{-}X'Y\,. \tag{6}$$

Для проверки гипотезы $H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_r = 0$ используют модель с ограничением на параметры при помощи матрицы H размерности r (число параметров) на $s = r - 1$

$$H'\beta = \theta_0, \quad \text{где} \quad H' = \left[\begin{array}{ccccc} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & 1 \end{array}\right] \quad \text{и} \quad \theta_0 = \left[\begin{array}{c} 0 \\ 0 \\ \ldots \\ 0 \end{array}\right]$$

Можно показать, что $Z = H'\hat{\beta} \;\sim\; \mathcal{N}_s(H'\beta, \sigma^2 D)$, где $D = H'(X'X)^{-}H$ и $R_0^2 = (Y - X\hat{\beta})'(Y - X\hat{\beta}) \;\sim\; \sigma^2\chi^2(n-r)$ распределены независимо. Кроме того $(Z - \theta_0)'D^{-1}(Z - \theta_0) = R_1^2 - R_0^2 \;\sim\; \sigma^2\chi^2(s)$, где $R_1^2 = (Y - X\beta^*)'(Y - X\beta^*)$, β^* оценка параметров усеченной модели. Тогда при справедливости нулевой гипотезы

$$F = \frac{n-r}{s}\frac{R_1^2 - R_0^2}{R_0^2} \;\sim\; F(s, n-r)\,.$$

1.6. Дифференцирование по вектору параметров

Этот раздел предназначен для тех, кто забыл, как осуществляется дифференцирование по вектору параметров и как получается система нормальных уравнений, откуда возникают оценки (6).

$$A\beta = \begin{bmatrix} a_{11} & \dots & a_{1m} \\ \dots & \dots & \dots \\ a_{n1} & \dots & a_{nm} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \dots \\ \beta_m \end{bmatrix} = \begin{bmatrix} \sum\limits_{i=1}^{m} a_{1i}\beta_i \\ \dots \\ \sum\limits_{i=1}^{m} a_{ni}\beta_i \end{bmatrix}$$

$$\frac{\partial A\beta}{\partial \beta} = \begin{bmatrix} \frac{\partial A\beta}{\partial \beta_1} & \dots & \frac{\partial A\beta}{\partial \beta_m} \end{bmatrix} = \begin{bmatrix} a_{11} & \dots & a_{1m} \\ \dots & \dots & \dots \\ a_{n1} & \dots & a_{nm} \end{bmatrix} = A$$

Для получения оценок МНК дифференцируем квадратичную форму.

$$\mathcal{L}_1 = (A\beta)'(A\beta) = \left(\sum_{i=1}^{m} a_{1i}\beta_i\right)^2 + \ldots + \left(\sum_{i=1}^{m} a_{ni}\beta_i\right)^2 .$$

Система нормальных уравнений имеет вид

$$\begin{cases} \frac{\partial \mathcal{L}_1}{\partial \beta_1} = 2\left(a_{11}\left(\sum\limits_{i=1}^{m} a_{1i}\beta_i\right) + \ldots + a_{n1}\left(\sum\limits_{i=1}^{m} a_{ni}\beta_i\right)\right) = 0, \\ \dots \\ \frac{\partial \mathcal{L}_1}{\partial \beta_m} = 2\left(a_{1m}\left(\sum\limits_{i=1}^{m} a_{1i}\beta_i\right) + \ldots + a_{nm}\left(\sum\limits_{i=1}^{m} a_{ni}\beta_i\right)\right) = 0, \end{cases}$$

$$\frac{\partial (A\beta)'(A\beta)}{\partial \beta} = 0 \Longleftrightarrow \quad 2\frac{\partial (A'\beta)}{\partial \beta} A\beta = 2A'A\beta = 0\,.$$

1.7. Наведение контрастов

Для проверки $H_0 : \mu_i = \mu_k$ значимости отклонений внутригрупповых средних используем отношение Стьюдента

$$t = \frac{\xi_0}{\sqrt{\eta}} = \frac{\xi_0}{\sqrt{\frac{1}{M}\zeta}},$$

где $\xi_0 \sim \mathcal{N}(0,1)$, $\zeta \sim \chi^2(M)$. Для выражения ξ_0 рассмотрим разность $\bar{x}_i - \bar{x}_k$ с характеристиками:

$$\mathbf{E}(\bar{x}_i - \bar{x}_k) = \mu_i - \mu_k,$$
$$\mathbf{D}(\bar{x}_i - \bar{x}_k) = \frac{\sigma^2}{n_i} + \frac{\sigma^2}{n_k},$$
$$\xi_0 = \frac{(\bar{x}_i - \bar{x}_k) - (\mu_i - \mu_k)}{\sigma\sqrt{\frac{1}{n_i} + \frac{1}{n_k}}}.$$

Так как $Q_1 = \sum_i n_i(\bar{x}_i - \bar{x})^2$ не зависит от Q_2, при ортогональном преобразовании переменных x_{ij} в y_k, $k = 1, \ldots, n$, разности вида

$$\bar{x}_i - \bar{x}_k = (\bar{x}_i - \bar{x}) - (\bar{x}_k - \bar{x})$$

выражаются через линейные комбинации переменных $y_1, \ldots, y_r$ и не зависят от Q_2, выражаемой через переменные $y_{r+1}, \ldots, y_n$. Следовательно, в отношении Стьюдента можно использовать выражение $\eta = \frac{s_2^2}{\sigma^2}$, где

$$s_2^2 = \frac{Q_2}{(n-r)}, \quad \frac{Q_2}{\sigma^2} \sim \chi^{\mathbf{2}}(n-r),$$
$$t = \frac{(\bar{x}_i - \bar{x}_k) - (\mu_i - \mu_k)}{Q_2\sqrt{\frac{1}{n_i} + \frac{1}{n_k}}}\sqrt{n-r} \ \sim \ T(n-r)\,.$$

Поправки Бонферрони применяются для того, чтобы частота ложноположительных результатов с поправкой на эффект множественных сравнений не превышала заданное значение.

Пусть имеются нулевые гипотезы $\mathcal{H}_1, \ldots, \mathcal{H}_m$ с соответствующими p-значениями: $p_1, \ldots, p_m$. Например, $m = C_r^2$, если проверяются гипотезы о

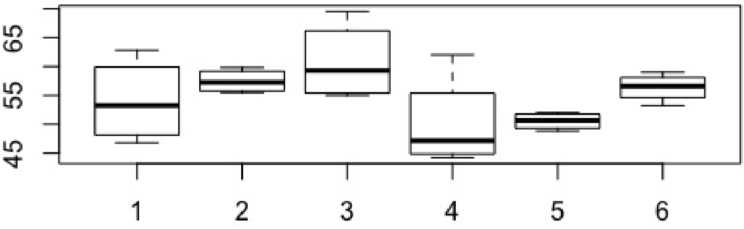

Рис. 1. Бокс-плот тестовых данных.

равенстве всех попарных средних. Определим как вероятность совершить хотя бы одну ошибку в выводе из m выводов. Если осуществляется n независимых испытаний с вероятностью ошибки α в каждом, то $\tilde{\alpha} = 1-(1-\alpha)^n$. В случае зависимых испытаний $\tilde{\alpha} \leqslant 1-(1-\alpha)^n$. Заметим, что по неравенству Буля

$$\tilde{\alpha} = 1 - (1 - n\alpha + C_n^2\alpha^2 - \ldots) \leqslant n \cdot \alpha.$$

Таким образом, если мы хотим, чтобы вероятность наличия хотя бы одного неверного вывода из m была равна 0.05, то достаточно установить вероятность неверного отклонения нулевой гипотезы равной $0.05/m$ для каждого вывода. Такой метод называется поправкой Бонферрони. Если гипотезы имеют различные доверительные уровни вероятностей, то достаточно отвергнуть гипотезы, имеющие $\alpha < \tilde{\alpha}/m$.

1.8. Множественные сравнения

Определение 2. Сравнением параметров $\beta_1, \ldots, \beta_p$ называется линейная функция $\sum\limits_{i=1}^{p} c_i\beta_i$, где $\sum\limits_{i=1}^{p} c_i = 0$.

Например, имеются три градации категориального признака, из которых две подвергнуты некоторому условию, а одна нет (контрольная группа здоровых и две группы больных, которых лечили разными препаратами). Для проверки этого утверждения можно рассмотреть разность

$$\psi = \frac{\beta_1 + \beta_2}{2} - \beta_3\,.$$

Обозначим через $\hat{\psi} = \sum_{i=1}^{p} c_i \hat{\beta}_i$ несмещенную оценку сравнения. В матричном виде

$$\psi_{q,1} = C_{q,p}\beta_{p,1}, \quad rank(C) = q,$$
$$\hat{\psi} = C\hat{\beta} = C(X^T X)^{-} X^T Y = AY.$$

Ковариационная матрица оценок находится по формуле

$$\Gamma_\psi = \sigma^2 A A^T, B = AA^T,$$

несмещенной оценкой σ^2 является средний квадрат ошибок

$$s^2 = \frac{R_0^2}{n-r}$$

Теорема 2. *(Метод множественных сравнений по Шеффе) Если вектор наблюдений $Y \sim \mathcal{N}(X\beta, \sigma^2 I)$, $rank(X) = r$, то случайная величина $\hat{\psi}$ не зависит от $R_0^2/\sigma^2 \sim \chi^{\mathbf{2}}(n-r)$ и имеет нормальное распределение $\mathcal{N}(\psi, \Gamma_\psi)$. Поэтому*

$$\frac{(\hat{\psi} - \psi)^T B^{-1} (\hat{\psi} - \psi)}{qs^2} \sim F(q, n-r).$$

Пусть элементы выборки $x_1, \ldots, x_n \sim \mathcal{N}(\mu, \sigma)$ упорядочены в виде $x^{(1)} < \ldots < x^{(n)}$, разность $R = x^{(n)} - x^{(1)}$ размах выборки. Пусть s^2 является независимой среднеквадратичной оценкой σ^2 с ν степенями свободы. Таким образом $\nu s^2/\sigma^2 = \chi^2_\nu$ не зависит от R. Случайную величину $R/s = q_{n,\nu}$ называют стьюдентизированным размахом.

Метод Тьюки можно применять для получения совместных доверительных утверждений о сравнениях множеста параметров $\theta_1, \ldots, \theta_k$ в терминах несмещенных оценок. Ограничением метода Тьюки является требование одинаковых дисперсий для оценок θ_i. Поэтому если нужна классификация по одному признаку, то объемы должны быть равными.

Теорема 3. *(Метод множественных сравнений Тьюки) Пусть* $\hat{\theta}_1, \ldots, \hat{\theta}_k$ *независимы,* $\mathcal{N}(\theta_i, a^2\sigma^2)$, $\frac{\nu s^2}{\sigma^2} = \chi^2_\nu$ *не зависит от* $\{\theta_i\}$, $\hat{\psi} = \sum\limits_{i=1}^{k} c_i \hat{\theta}_i$, $T = aq_{\alpha,k,\nu}$, *где* $q_{\alpha,k,\nu}$ *верхний* α *предел стьюдентизированного размаха. Тогда вероятность того, что все* $\frac{k(k-1)}{2}$ *разностей* $\{\theta_i - \theta_j\}$ *одновременно удовлетворяют неравенствам*

$$\hat{\theta}_i - \hat{\theta}_j - Ts \le \theta_i - \theta_j \le \hat{\theta}_i - \hat{\theta}_j + Ts$$

равна $1 - \alpha$.

Кроме того, существует обобщение для $\hat{\theta}_1, \ldots, \hat{\theta}_k$, не обязательно независимых, но имеющих одинаковые ковариации и дисперсии.

В критерии Пиллая используется верхний α предел стьюдентизированного максимума модулей $M = max|x_i|/s$. Увеличенным размахом R' называется $max(R, M)$.

В тех случаях, когда главный интерес представляют все разности, причем никакой из них не отдается предпочтения, метод Тьюки дает более узкие интервалы, но применим только в случае одинаковых дисперсий. Преимущество метода Шеффе в том, что он менее чувствителен к нарушению о предположении нормальности и равенства дисперсий.