Лекция 10

Однофакторный дисперсионный анализ

Однофакторный дисперсионный анализ

(One-way ANOVA)

Есть k независимых выборок:

$$\left.\begin{array}{l} Y_{11}, \ldots, Y_{1n_1} \\ Y_{21}, \ldots, Y_{2n_2} \\ \quad \ldots \\ Y_{k1}, \ldots, Y_{kn_k} \end{array}\right\} \text{все эти величины независимы}$$

Почему Y, а не X?

Дисперсия во всех наблюдениях одинакова.
Математическое ожидание может отличаться от выборки к выборке.

$$Y_{ij} \sim N(\mu_i, \sigma^2)$$

ещё пишут так: $Y_{ij} = \mu_i + \epsilon_{ij},\ \epsilon_{ij} \sim N(0, \sigma^2)$

Общее число наблюдений:

$$n = n_1 + n_2 + \ldots + n_k$$

Гипотезы:

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_k$$

H_A: хотя бы в одной выборке м.о. отличается

Зачем это нужно? Мы же умеем сравнивать средние попарно.

Однофакторный дисперсионный анализ

Есть k независимых выборок:

$$Y_{11}, \ldots, Y_{1n_1}$$
$$Y_{21}, \ldots, Y_{2n_2}$$
$$\ldots$$
$$Y_{k1}, \ldots, Y_{kn_k}$$

Общее число наблюдений:

$$n = n_1 + n_2 + \ldots + n_k$$

Разные средние:

$$\bar{Y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$$ - выборочные (групповые) средние

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}$$ - общее среднее

За основную гипотезу: все $\bar{Y}_i$ близки к $\bar{Y}$.

Против основной гипотезы: разброс между средними.

Разложение разброса

Разные средние:

$$\bar{Y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$$ - выборочные (групповые) средние

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}$$ - общее среднее

Остатки (residuals):

$$e_{ij} = Y_{ij} - \bar{Y}_i$$

Потом пригодится:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i} e_{ij} = 0$$

$$\sum_{j=1}^{n_i} e_{ij} = 0, \quad j = 1,\dots,k$$

Разложение признака на объяснённую и необъяснённую составляющие:

$$Y_{ij} = \mu_i + \epsilon_{ij} = \bar{Y}_i + e_{ij}$$

Зачем эти отстатки нужны?

Разложение разброса

Разные средние:

$$\bar{Y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$$ - выборочные (групповые) средние

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}$$ - общее среднее

Туманная дребедень:

$$\mathrm{TSS} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}\right)^2$$ - общая сумма квадратов (Total Sum of Squares)

$$\mathrm{ESS} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\bar{Y}_i - \bar{Y}\right)^2 = \sum_{i=1}^{k} n_i\left(\bar{Y}_i - \bar{Y}\right)^2$$ - объяснённая сумма квадратов (Explained Sum of Squares)

$$\mathrm{RSS} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_i\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} e_{ij}^2$$ - остаточная сумма квадратов (Residual Sum of Squares)

Теорема (Пифагора):

$$\mathrm{TSS} = \mathrm{ESS} + \mathrm{RSS}$$

Примечание. Обозначения сильно варьируются. Вместо ESS пишут MSS (Model Sum of Squares) или SSW (Sum of Squares Within groups). Да как угодно всё это называется и обозначается...

$$\mathrm{TSS} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}\right)^2 \qquad \mathrm{ESS} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(\bar{Y}_i - \bar{Y}\right)^2 \qquad \mathrm{RSS} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}_i\right)^2$$

$$\mathrm{TSS} = \mathrm{ESS} + \mathrm{RSS}$$

Доказательство. ◄Представим отклонения от общего среднего в таком виде:

$$Y_{ij} - \bar{Y} = \left(Y_{ij} - \bar{Y}_i\right) + \left(\bar{Y}_i - \bar{Y}\right)$$

$$\mathrm{TSS} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}\right)^2 \qquad \mathrm{ESS} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(\bar{Y}_i - \bar{Y}\right)^2 \qquad \mathrm{RSS} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}_i\right)^2$$

$$\mathrm{TSS} = \mathrm{ESS} + \mathrm{RSS}$$

Доказательство. ◄Представим отклонения от общего среднего в таком виде:

$$Y_{ij} - \bar{Y} = \left(Y_{ij} - \bar{Y}_i\right) + \left(\bar{Y}_i - \bar{Y}\right)$$

Возведём в квадрат:

$$\left(Y_{ij} - \bar{Y}\right)^2 = \left(Y_{ij} - \bar{Y}_i\right)^2 + \left(\bar{Y}_i - \bar{Y}\right)^2 + 2\left(Y_{ij} - \bar{Y}_i\right)\left(\bar{Y}_i - \bar{Y}\right)$$

$$\mathrm{TSS} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}\right)^2 \qquad \mathrm{ESS} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\bar{Y}_i - \bar{Y}\right)^2 \qquad \mathrm{RSS} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_i\right)^2$$

$$\boxed{\mathrm{TSS} = \mathrm{ESS} + \mathrm{RSS}}$$

Доказательство. ◄Представим отклонения от общего среднего в таком виде:

$$Y_{ij} - \bar{Y} = \left(Y_{ij} - \bar{Y}_i\right) + \left(\bar{Y}_i - \bar{Y}\right)$$

Возведём в квадрат:

$$\left(Y_{ij} - \bar{Y}\right)^2 = \left(Y_{ij} - \bar{Y}_i\right)^2 + \left(\bar{Y}_i - \bar{Y}\right)^2 + 2\left(Y_{ij} - \bar{Y}_i\right)\left(\bar{Y}_i - \bar{Y}\right)$$

Просуммируем по всем наблюдениям:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_i\right)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\bar{Y}_i - \bar{Y}\right)^2 + 2\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_i\right)\left(\bar{Y}_i - \bar{Y}\right)$$

$$\mathrm{TSS} = \mathrm{RSS} + \mathrm{ESS} + 2\sum_{i=1}^{k}\sum_{j=1}^{n_i} e_{ij}\left(\bar{Y}_i - \bar{Y}\right)$$

$$\mathrm{TSS} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}\right)^2 \qquad \mathrm{ESS} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\bar{Y}_i - \bar{Y}\right)^2 \qquad \mathrm{RSS} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_i\right)^2$$

$$\boxed{\mathrm{TSS} = \mathrm{ESS} + \mathrm{RSS}}$$

Доказательство. ◄Представим отклонения от общего среднего в таком виде:

$$Y_{ij} - \bar{Y} = \left(Y_{ij} - \bar{Y}_i\right) + \left(\bar{Y}_i - \bar{Y}\right)$$

Возведём в квадрат:

$$\left(Y_{ij} - \bar{Y}\right)^2 = \left(Y_{ij} - \bar{Y}_i\right)^2 + \left(\bar{Y}_i - \bar{Y}\right)^2 + 2\left(Y_{ij} - \bar{Y}_i\right)\left(\bar{Y}_i - \bar{Y}\right)$$

Просуммируем по всем наблюдениям:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_i\right)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\bar{Y}_i - \bar{Y}\right)^2 + 2\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_i\right)\left(\bar{Y}_i - \bar{Y}\right)$$

$$\mathrm{TSS} = \mathrm{RSS} + \mathrm{ESS} + 2\sum_{i=1}^{k}\sum_{j=1}^{n_i} e_{ij}\left(\bar{Y}_i - \bar{Y}\right)$$

Избавимся от суммы справа:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i} e_{ij}\left(\bar{Y}_i - \bar{Y}\right) = \sum_{i=1}^{k}\left[\left(\bar{Y}_i - \bar{Y}\right)\underbrace{\sum_{j=1}^{n_i} e_{ij}}_{0}\right] = 0$$

Получили, что $\mathrm{TSS} = \mathrm{ESS} + \mathrm{RSS}$ ►

F-статистика

Проверяем гипотезы:

$$H_0: \mu_1 = \mu_2 = ... = \mu_k$$

H_A: хотя бы в одной выборке м.о. отличается

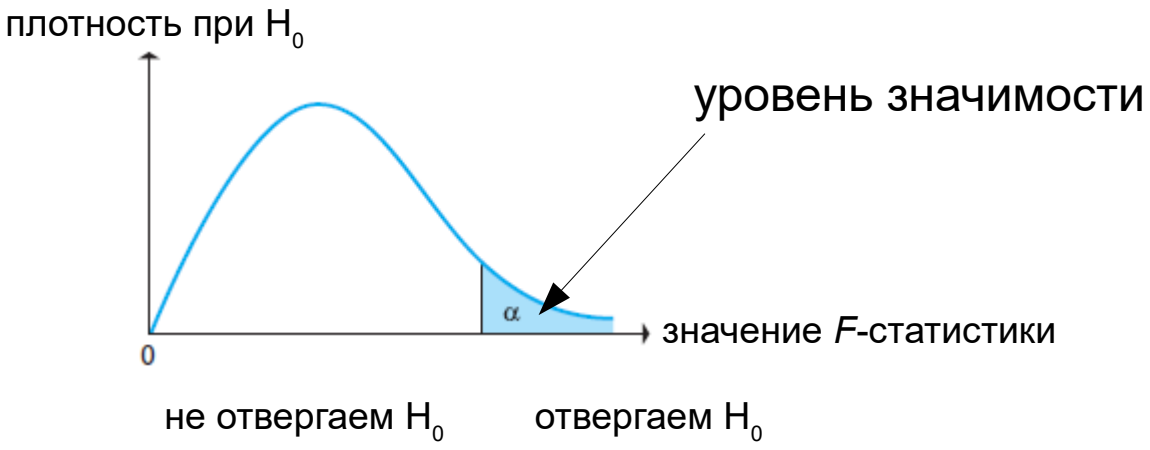
Статистика:

$$F = \frac{\mathrm{ESS} / (k-1)}{\mathrm{RSS} / (n-k)} \overset{H_0}{\sim} F_{k-1,n-k}$$

доказательство - отдельно

Решающее правило:

отвергнуть H_0, если $F > F_{k-1,n-k,\alpha}$



Пример

Отдел маркетинга в компании, владеющей сетью супермаркетов, пытается определить оптимальное расположение для продаваемых пончиков. В течение четырёх недель пончики ставились на верхнюю полку, в течение шести недель — на полку на уровне глаз, и в течение пяти недель — на нижнюю полку. В таблице приведены количества проданных упаковок с пончиками за все 15 недель эксперимента:

Нижняя полка	На уровне глаз	Верхняя полка
228	226	175
173	237	211
190	225	214
195	261	159
	243	233
	224	

Даёт ли проведённый эксперимент основания считать, что объём продаж пончиков зависит от их расположения на полках? Выполните проверку на уровне значимости 5%.

Пример

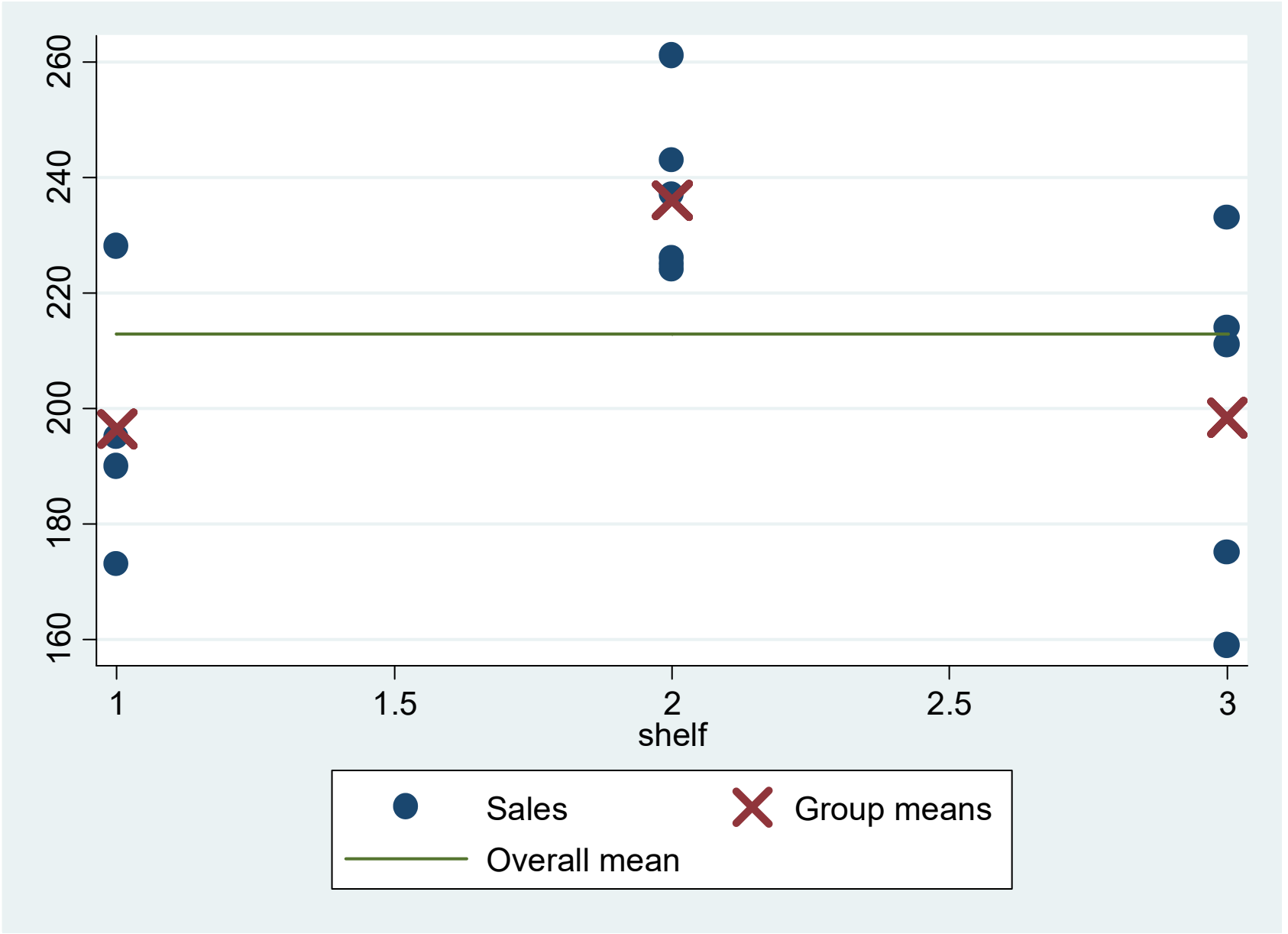
Рассчитываем средние:

	Y_{1j}	Y_{2j}	Y_{3j}	
	Нижняя полка	На уровне глаз	Верхняя полка	
	228	226	175	
	173	237	211	
	190	225	214	
	195	261	159	
		243	233	
		224		
Среднее:	196.5	236.0	198.4	← $\bar{Y}_i$
Общее среднее:		212.933		$\bar{Y}$

$$n_1 = 4, \qquad n_2 = 6, \qquad n_3 = 5,$$

$$n = 4 + 6 + 5 = 15.$$

То же на картинке



Когда не слишком мало наблюдений, лучше использовать усатые ящики

Рассчитываем статистику и делаем вывод

	Нижняя полка		На уровне глаз		Верхняя полка	
	Продажи	Остатки	Продажи	Остатки	Продажи	Остатки
	228	31.5	226	-10	175	-23.4
	173	-23.5	237	1	211	12.6
	190	-6.5	225	-11	214	15.6
	195	-1.5	261	25	159	-39.4
			243	7	233	34.6
			224	-12		
Среднее:	196.5		236.0		198.4	
Общее среднее:			212.933			

$$\text{ESS} = 4\times(196.5-212.933)^2 + 6\times(236-212.933)^2 + 5\times(198.4-212.933)^2 = 5328.73$$

$$\text{RSS} = 31.5^2 + (-23.5)^2 + \ldots + (-39.4)^2 + 34.6^2 = 6328.2$$

$$\text{TSS} = (228-212.933)^2 + (173-212.933)^2 + \ldots + (233-212.933)^2 = 11656.93$$

Рассчитываем статистику и делаем вывод

	Нижняя полка		На уровне глаз		Верхняя полка	
	Продажи	Остатки	Продажи	Остатки	Продажи	Остатки
	228	31.5	226	-10	175	-23.4
	173	-23.5	237	1	211	12.6
	190	-6.5	225	-11	214	15.6
	195	-1.5	261	25	159	-39.4
			243	7	233	34.6
			224	-12		
Среднее:	196.5		236.0		198.4	
Общее среднее:			212.933			

$$\text{ESS} = 4\times(196.5-212.933)^2 + 6\times(236-212.933)^2 + 5\times(198.4-212.933)^2 = 5328.73$$

$$\text{RSS} = 31.5^2 + (-23.5)^2 + \ldots + (-39.4)^2 + 34.6^2 = 6328.2$$

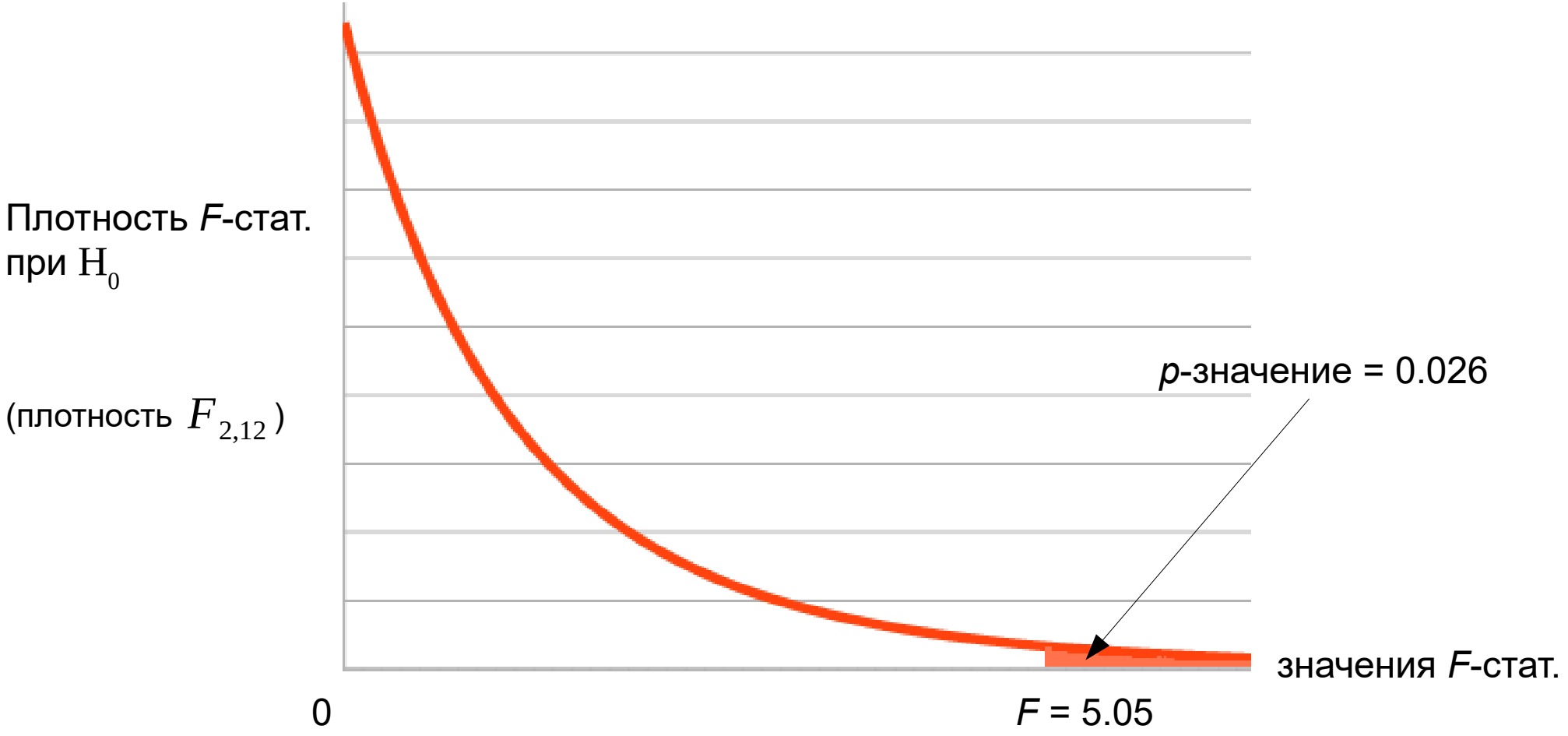
$$\text{TSS} = (228-212.933)^2 + (173-212.933)^2 + \ldots + (233-212.933)^2 = 11656.93$$

Статистика: $F = \dfrac{\text{ESS} / (k-1)}{\text{RSS} / (n-k)} = \dfrac{5328.73 / (3-1)}{6328.2 / (15-3)} = 5.05.$

Критическое значение: $F_{k-1,n-k,\alpha} = F_{2,12,0.05} = 3.89.$

Вывод: $F = 5.05 > 3.89$ => объём продаж в среднем зависит от расположения.

P-значение



Гипотеза о равенстве средних (продажи не связаны с расположением) отвергается на уровне 2.6% и выше.

Оценка дисперсии

$$\hat{\sigma}^2 = \frac{\mathrm{RSS}}{n-k} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} e_{ij}^2}{n-k}.$$

Докажем несмещённость. ◄ Несмещённая оценка по отдельной выборке:

$$\hat{\sigma}_i^2 = \frac{1}{n_i-1}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_i\right)^2 = \frac{1}{n_i-1}\sum_{j=1}^{n_i} e_{ij}^2$$

$$\sigma^2 = \mathrm{E}\left(\hat{\sigma}_i^2\right) = \frac{1}{n_i-1}\mathrm{E}\left(\sum_{j=1}^{n_i} e_{ij}^2\right) \quad \Rightarrow \quad \mathrm{E}\left(\sum_{j=1}^{n_i} e_{ij}^2\right) = \sigma^2\left(n_i - 1\right)$$

Оценка дисперсии

$$\hat{\sigma}^2 = \frac{\mathrm{RSS}}{n-k} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} e_{ij}^2}{n-k}.$$

Докажем несмещённость. ◄ Несмещённая оценка по отдельной выборке:

$$\hat{\sigma}_i^2 = \frac{1}{n_i-1}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_i\right)^2 = \frac{1}{n_i-1}\sum_{j=1}^{n_i} e_{ij}^2$$

$$\sigma^2 = \mathrm{E}\left(\hat{\sigma}_i^2\right) = \frac{1}{n_i-1}\mathrm{E}\left(\sum_{j=1}^{n_i} e_{ij}^2\right) \quad \Rightarrow \quad \mathrm{E}\left(\sum_{j=1}^{n_i} e_{ij}^2\right) = \sigma^2\left(n_i - 1\right)$$

Складываем квадраты остатков по всем выборкам:

$$\mathrm{E}\left(\mathrm{RSS}\right) = \mathrm{E}\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i} e_{ij}^2\right) = \sum_{i=1}^{k}\mathrm{E}\left(\sum_{j=1}^{n_i} e_{ij}^2\right) = \sum_{i=1}^{k}\sigma^2\left(n_i - 1\right) = \sigma^2\left(n_1+\ldots+n_k-k\right)=\sigma^2\left(n-k\right).$$

Так что

$$\mathrm{E}\left(\sigma^2\right) = \frac{\mathrm{E}\left(\mathrm{RSS}\right)}{n-k} = \frac{\sigma^2\left(n-k\right)}{n-k} = \sigma^2.\blacktriangleright$$

Коэффициент детерминации

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

Отражает долю разброса признака Y, которую можно объяснить принадлежностью к разным выборкам (долю *объясняемого* разброса).

Интерпретация крайних значений:

$R^2 = 0$ - средние во всех выборках совпадают, никаких различий

$(\text{ESS}=0, \text{TSS}=\text{RSS}>0)$

$R^2 = 1$ - внутри выборок нет вариации признака,

$(\text{RSS}=0, \text{TSS}=\text{ESS}>0)$ разброс объясняется исключительно различием между выборками

Скорректированный (нормированный, поправленный) **коэффициент детерминации**:

$$R^2_{adj} = 1 - \frac{\text{RSS}/(n-k)}{\text{TSS}/(n-1)}$$

он может быть <0 !

зачем он?

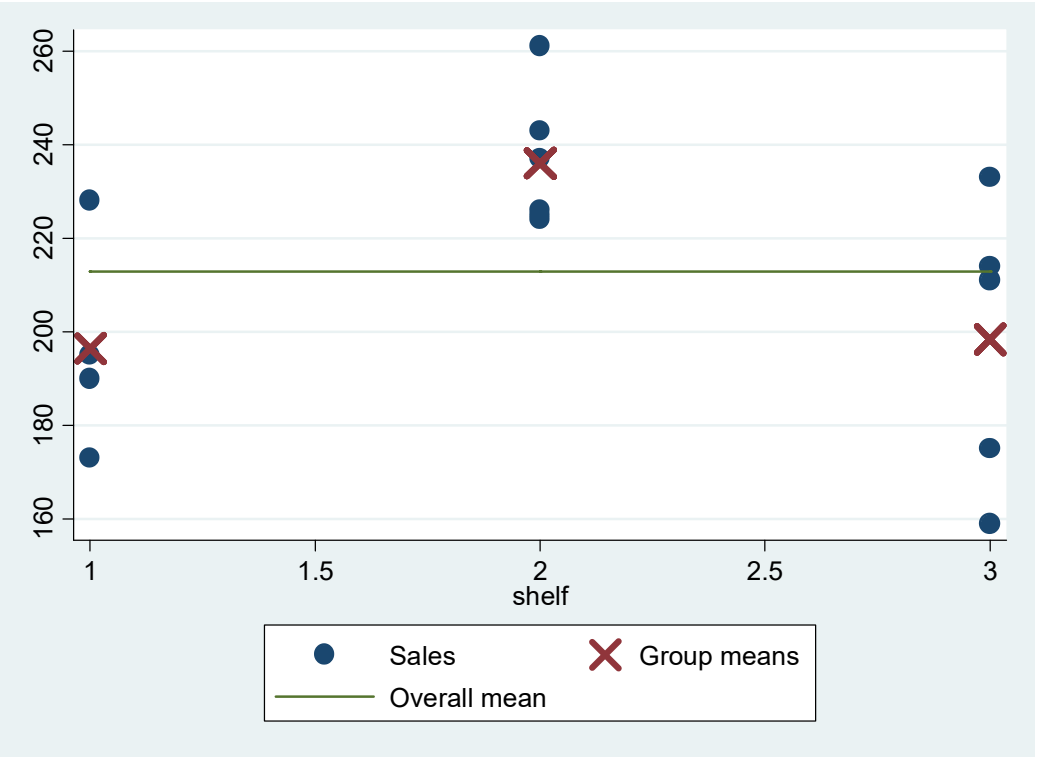
Возвращение к пончикам

Мы уже успели рассчитать суммы квадратов:

$$\text{ESS} = 5328.73$$

$$\text{RSS} = 6328.2$$

$$\text{TSS} = 11656.93$$



Коэффициенты детерминации:

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = \frac{5328.73}{11656.93} = 0.46$$

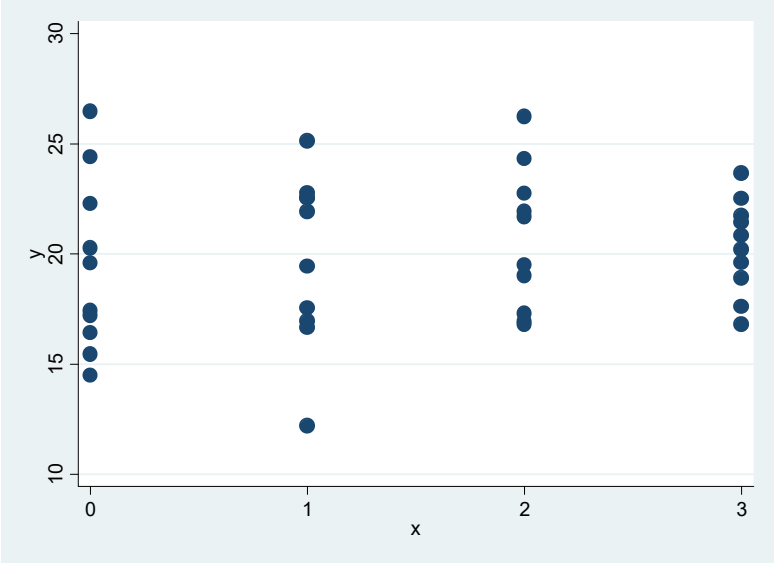
$$R^2_{adj} = 1 - \frac{6328.2/(15-3)}{11656.93/(15-1)} = 0.37$$

посчитано просто для примера

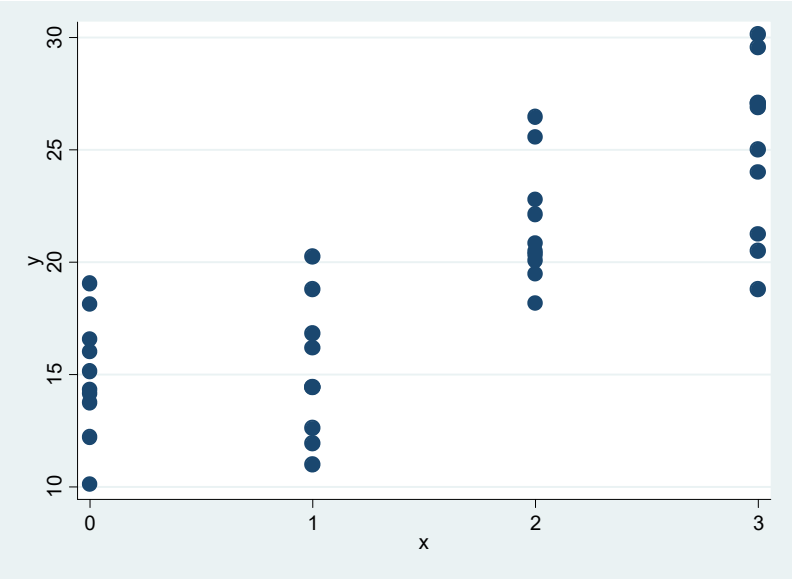
46% разброса наблюдаемых объёмов продаж можно объяснить за счёт смены расположения пончиков.

Картинки и циферки

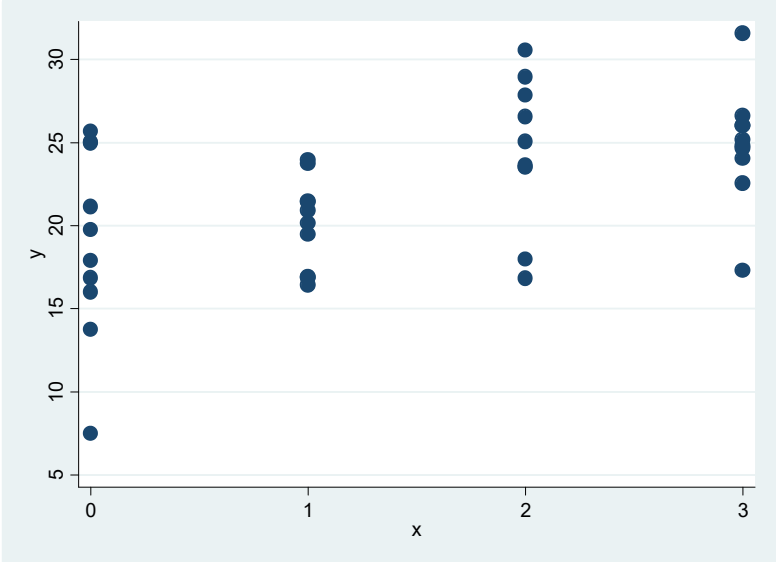
$R^2 \approx 0$



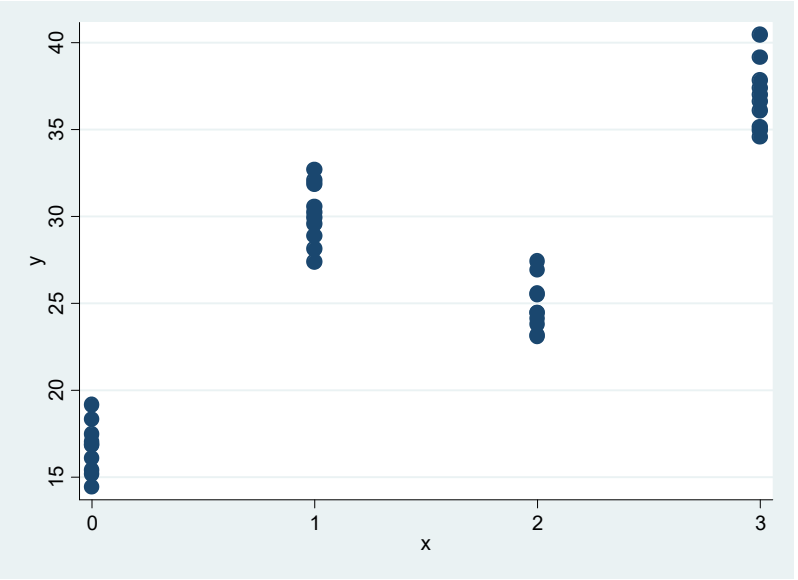
$R^2 \approx 0.7$



$R^2 \approx 0.3$



$R^2 \approx 0.95$



Снова *F*-статистика

Проверяем гипотезы:

$$H_0: \mu_1 = \mu_2 = ... = \mu_k$$

H_A: хотя бы в одной выборке м.о. отличается

Статистика:

$$F = \frac{\text{ESS} / (k-1)}{\text{RSS} / (n-k)} \overset{H_0}{\sim} F_{k-1, n-k}$$

Её можно выразить через коэффициент детерминации:

$$F = \frac{R^2 / (k-1)}{(1-R^2) / (n-k)}$$

Образец вывода (Stata)

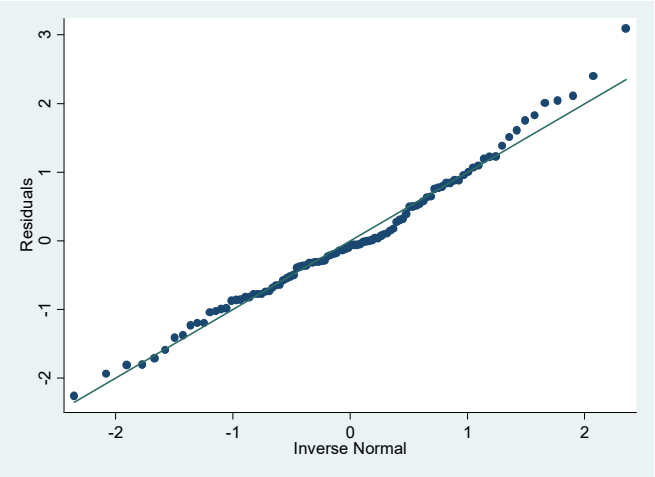
```
. anova y shelf

                          Number of obs =        15     R-squared     =  0.4571
                          Root MSE      =   22.9641     Adj R-squared =  0.3667

                 Source | Partial SS      df       MS           F     Prob > F
------------------------+----------------------------------------------------
                  Model |  5328.73333       2  2664.36667        5.05     0.0256
                        |
                  shelf |  5328.73333       2  2664.36667        5.05     0.0256
                        |
               Residual |     6328.2      12       527.35
------------------------+----------------------------------------------------
                  Total |  11656.9333      14  832.638095

. reg

      Source |       SS       df       MS              Number of obs =      15
-------------+------------------------------           F(  2,    12) =    5.05
       Model |  5328.73333     2  2664.36667            Prob > F      =  0.0256
    Residual |     6328.2     12       527.35           R-squared     =  0.4571
-------------+------------------------------           Adj R-squared =  0.3667
       Total |  11656.9333     14  832.638095           Root MSE      =  22.964

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       shelf |
          2  |        39.5   14.82326     2.66   0.021     7.202881    71.79712
          3  |         1.9   15.40479     0.12   0.904    -31.66415    35.46415
             |
       _cons |       196.5   11.48205    17.11   0.000     171.4828    221.5172
------------------------------------------------------------------------------
```

Сравнение средних по трём выборкам

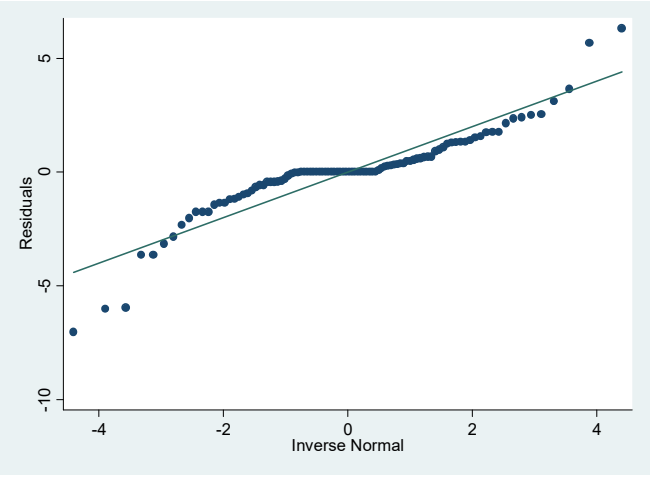
Сравнение с «базовой категорией» (нижней полкой)

Проверка предпосылок

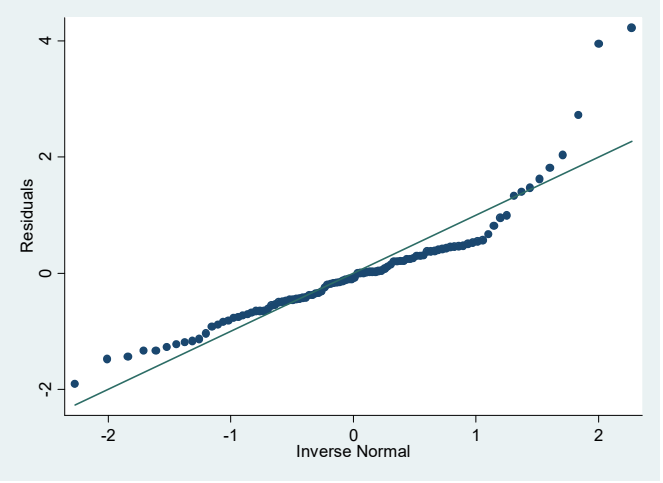
Можно изучить распределение остатков — гистограмма или «квантиль-квантиль».



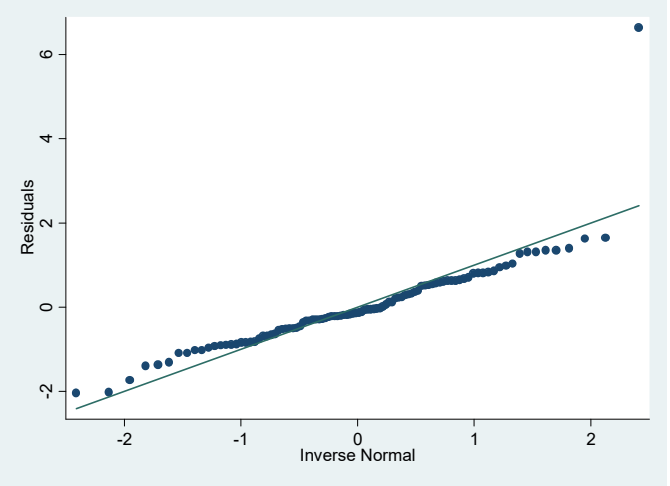
все предпосылки выполнены



нормальное распределение,
разные дисперсии (гетероскедастичность)

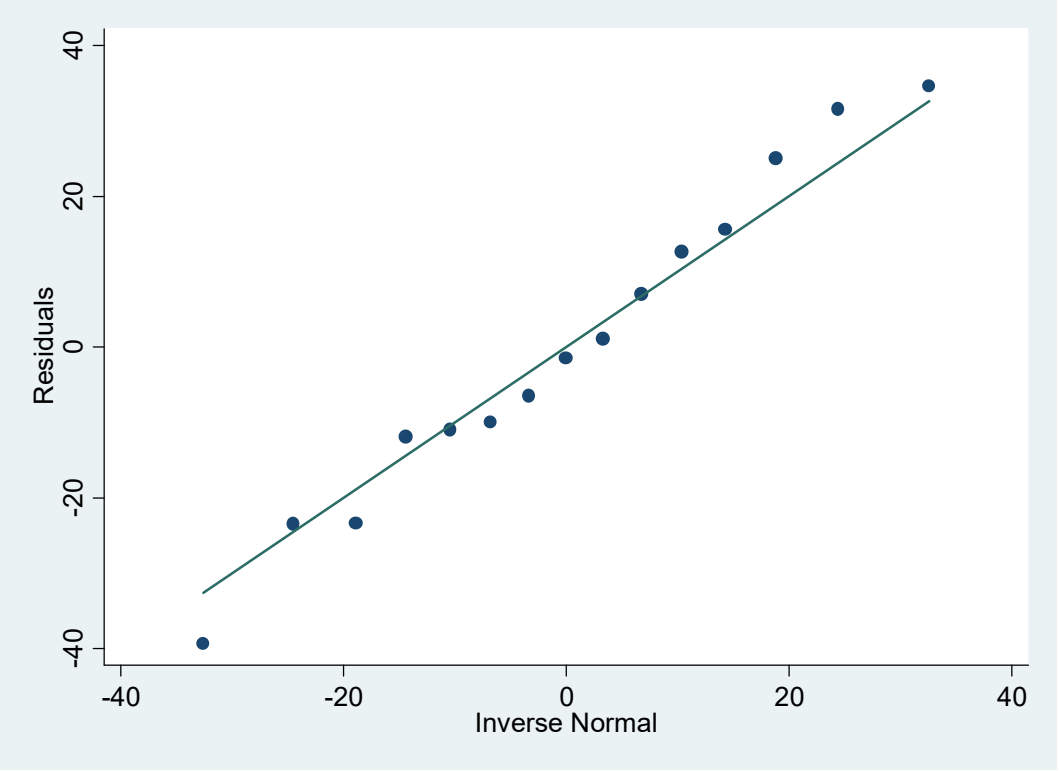


не нормальное распределение,
разные дисперсии



выброс

QQ для задачи с пончиками



Для размышления

Пусть для выборок все предпосылки выполнены.

$$\left.\begin{array}{l} Y_{11}, \ldots, Y_{1n_1} \\ Y_{21}, \ldots, Y_{2n_2} \\ \quad \ldots \\ Y_{k1}, \ldots, Y_{kn_k} \end{array}\right\} \text{независимы, } N(\mu_i, \sigma^2).$$

Будут ли остатки

- независимыми?
- нормальными?
- гомоскедастичными*?

* гомоскедастичность — равенство дисперсий.

Подытожим

Предпосылки: $Y_{11}, \ldots, Y_{1n_1}; \quad Y_{21}, \ldots, Y_{2n_2}; \quad \cdots \quad ; \quad Y_{k1}, \ldots, Y_{kn_k}$ независимы,

$Y_{ij} \sim N(\mu_i, \sigma^2)$.

Гипотезы:

$H_0: \mu_1 = \mu_2 = \ldots = \mu_k$

H_A: хотя бы в одной выборке м.о. отличается

Разложение разброса:

$$\mathrm{TSS} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y})^2; \qquad \mathrm{ESS} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{Y}_i - \bar{Y})^2; \qquad \mathrm{RSS} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2.$$

$$\mathrm{TSS} = \mathrm{ESS} + \mathrm{RSS}.$$

Коэффициенты детерминации:

$$R^2 = \frac{\mathrm{ESS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}, \quad 0 \le R^2 \le 1; \qquad R^2_{adj} = 1 - \frac{\mathrm{RSS}/(n-k)}{\mathrm{TSS}/(n-1)}.$$

F-статистика:

$$F = \frac{\mathrm{ESS}/(k-1)}{\mathrm{RSS}/(n-k)} = \frac{R^2/(k-1)}{(1-R^2)/(n-k)} \overset{\mathrm{H}_0}{\sim} F_{k-1,n-k}.$$

Критическое правило: отвергнуть H_0, если $F > F_{k-1,n-k,\alpha}$.

О чём я не говорил

> Зачем нужен дисперсионный анализ, если мы можем сравнивать выборки попарно.

> Попарные сравнения в дисперсионном анализе.

> Сравнение дисперсий в нескольких выборках.

> Критерий Краскелла-Уоллиса (он полезен, но не уместился).

> Многофакторный дисперсионный анализ.

> Как сравнить распределение качественного признака по нескольких выборкам.

Следующая лекция

Корреляционный анализ, часть I:
выборочные коэффициенты корреляции Пирсона и Спирмена.