

Линейная регрессия. Тест на равенство коэффициента заданному числу

Call:

```
lm(formula = "log(price) ~ log(sqft_living)", data = data)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.94260	-0.21572	0.01598	0.23361	0.83623

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	7.10768	0.63400	11.21	< 2e-16 ***
log(sqft_living)	0.78143	0.08322	9.39	2.55e-15 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.329 on 98 degrees of freedom

Multiple R-squared: 0.4736, Adjusted R-squared: 0.4682

F-statistic: 88.17 on 1 and 98 DF, p-value: 2.555e-15

Как проверить, гипотезу о том, что коэффициент при $\log(\text{sqft_living})$ равен 0.5?

Свойства параметров регрессии

$$y = \beta x + \alpha$$

предполагаем, что истинная зависимость
устроена так

$$y = \hat{\beta}x + \hat{\alpha}$$

Из выборки получаем оценки на параметры

$$E(\hat{\beta}) = \beta$$

$$E(\hat{\alpha}) = \alpha$$

$$se(\hat{\beta}) = \frac{\sigma}{\sqrt{SSX}} = \frac{\sqrt{\frac{SSE}{n-2}}}{\sqrt{SSX}}$$

$$se(\hat{\alpha}) = \sigma \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{s_x^2}}$$

T-test для slope (beta)

$$H_0 : \beta = b$$

$$H_1 : \beta \neq b$$

$$t_{score} = \frac{\hat{\beta} - b}{se(\hat{\beta})} \sim t_{n-2}$$

T-test для slope (beta)

$$H_0 : \beta = 0.5$$

$$H_1 : \beta \neq 0.5$$

$$t_{score} = \frac{\hat{\beta} - 0.5}{se(\hat{\beta})} \sim t_{98}$$

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	7.10768	0.63400	11.21	< 2e-16	***
log(sqft_living)	0.78143	0.08322	9.39	2.55e-15	***

$$t_{score} = \frac{\hat{\beta} - 0.5}{se(\hat{\beta})} = \frac{0.78143 - 0.5}{0.08322} = 3.3818$$

$$2 \cdot p(t_{98} > t_{score}) = 0.001$$

На уровне значимости 0.05 отвергаем H_0

```
Call:  
lm(formula = "log(price) ~ log(sqft_living)", data = data)  
  
Residuals:  
    Min      1Q  Median      3Q     Max  
-0.94260 -0.21572  0.01598  0.23361  0.83623  
  
Coefficients:  
              Estimate Std. Error t value Pr(>|t|)  
(Intercept) 7.10768   0.63400 11.21 < 2e-16 ***  
log(sqft_living) 0.78143   0.08322  9.39 2.55e-15 ***  
---  
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1  
  
Residual standard error: 0.329 on 98 degrees of freedom  
Multiple R-squared:  0.4736,    Adjusted R-squared:  0.4682  
F-statistic: 88.17 on 1 and 98 DF,  p-value: 2.555e-15
```

Существует альтернативная гипотеза, что price квадратично от sqft_living - как ее проверить, не строя новой регрессии?

$$y = \beta x^2$$

$$\log y = \log \beta x^2$$

$$\log y = \log x^2 + \log \beta$$

$$\log y = 2 \cdot \log x + \alpha^*$$

Существует альтернативная гипотеза, что price квадратично зависит от sqft_living - как ее проверить, не строя новой регрессии?

Достаточно проверить гипотезу о том, что коэффициент в log-log регрессии равен 2

T-test для slope (beta)

$$H_0 : \beta = 2$$

$$H_1 : \beta \neq 2$$

$$t_{score} = \frac{\hat{\beta} - 2}{se(\hat{\beta})} \sim t_{98}$$

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	7.10768	0.63400	11.21	< 2e-16	***
log(sqft_living)	0.78143	0.08322	9.39	2.55e-15	***

$$t_{score} = \frac{\hat{\beta} - 2}{se(\hat{\beta})} = \frac{0.78143 - 2}{0.08322} = -14.66$$

$$2 \cdot p(t_{98} > |t_{score}|) \approx 0$$

На уровне значимости 0.05 отвергаем H_0

Call:

```
lm(formula = "log(price) ~ log(sqft_living)", data = data)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.94260	-0.21572	0.01598	0.23361	0.83623

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	7.10768	0.63400	11.21	< 2e-16 ***
log(sqft_living)	0.78143	0.08322	9.39	2.55e-15 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.329 on 98 degrees of freedom

Multiple R-squared: 0.4736, Adjusted R-squared: 0.4682

F-statistic: 88.17 on 1 and 98 DF, p-value: 2.555e-15

Как интерпретировать коэффициент при $\log(\text{sqft})$?

Конкретно, как изменится цена на квартиру при увеличении sqft_living в 3 раза?

$$\log y_1 = \beta \log x_1 + \alpha$$

$$\log y_2 = \beta \log x_2 + \alpha$$

$$\log \frac{y_2}{y_1} = \beta \log \frac{x_2}{x_1}$$

Допустим, x_2 больше x_1 в 3 раза

$$\log \frac{y_2}{y_1} = \beta \log 3 = 1.1\beta$$

$$\log \frac{y_2}{y_1} = 1.1 \cdot 0.7814 = 0.8595$$

$$\frac{y_2}{y_1} = e^{0.8595} = 2.36$$

На каждое увеличение переменной `sqft_living` в 3 раза, значение цены изменяется примерно в 2.4 раза

Call:

```
lm(formula = "log(price) ~ log(sqft_living)", data = data)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.94260	-0.21572	0.01598	0.23361	0.83623

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	7.10768	0.63400	11.21	< 2e-16 ***
log(sqft_living)	0.78143	0.08322	9.39	2.55e-15 ***

Signif. codes:	0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1			

Residual standard error: 0.329 on 98 degrees of freedom

Multiple R-squared: 0.4736, Adjusted R-squared: 0.4682

F-statistic: 88.17 on 1 and 98 DF, p-value: 2.555e-15

Как изменится цена на квартиру при увеличении sqft_living в 3 раза? Постройте доверительный интервал

**Как изменится цена на квартиру при увеличении
 $\log(\text{sqft_living})$ в 3 раза? Постройте доверительный интервал**

Сначала строим доверительный интервал для β_1

**Как изменится цена на квартиру при увеличении
 $\log(\text{sqft_living})$ в 3 раза? Постройте доверительный интервал**

Сначала строим доверительный интервал для β

$$\beta \in [\hat{\beta} \pm t_{n-2}(0.05/2) \cdot SE(\hat{\beta})]$$

$$\beta \in [\hat{\beta} \pm t_{n-2}(0.05/2) \cdot SE(\hat{\beta})]$$

$$\beta \in [0.6163, 0.9466]$$

**Как изменится цена на квартиру при увеличении
 $\log(\text{sqft_living})$ в 3 раза? Постройте доверительный интервал**

Сначала строим доверительный интервал для beta

$$\beta \in [\hat{\beta} \pm t_{n-2}(0.05/2) \cdot SE(\hat{\beta})]$$

$$\beta \in [\hat{\beta} \pm t_{n-2}(0.05/2) \cdot SE(\hat{\beta})]$$

$$\beta \in [0.6163, 0.9466]$$

И подставляем в ранее полученную формулу

$$\log \frac{y_2}{y_1} = \beta \log 3 = 1.1\beta$$

$$\log \frac{y_2}{y_1} \in [0.6779, 1.0413]$$

$$\frac{y_2}{y_1} \in [1.9698, 2.8328]$$

Вопрос - лог-трансформации чего достаточно, чтобы доверительный интервал все равно надо было обратно трансформировать?

Вопрос - лог-трансформации чего достаточно, чтобы предыдущие вычисления были так же валидны?

Достаточно лог-трансформации зависимой переменной y.

Упражнение - переделайте вычисления так, для регрессии, где x не логарифмировано. Теперь нас интересует изменение y, которое соответствует увеличению x на 1:

Call:

```
lm(formula = "log(price) ~ sqft_living", data = data)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.88611	-0.21032	0.01592	0.23574	0.88344

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.227e+01	9.030e-02	135.888	< 2e-16 ***
sqft_living	3.600e-04	3.867e-05	9.311	3.78e-15 ***

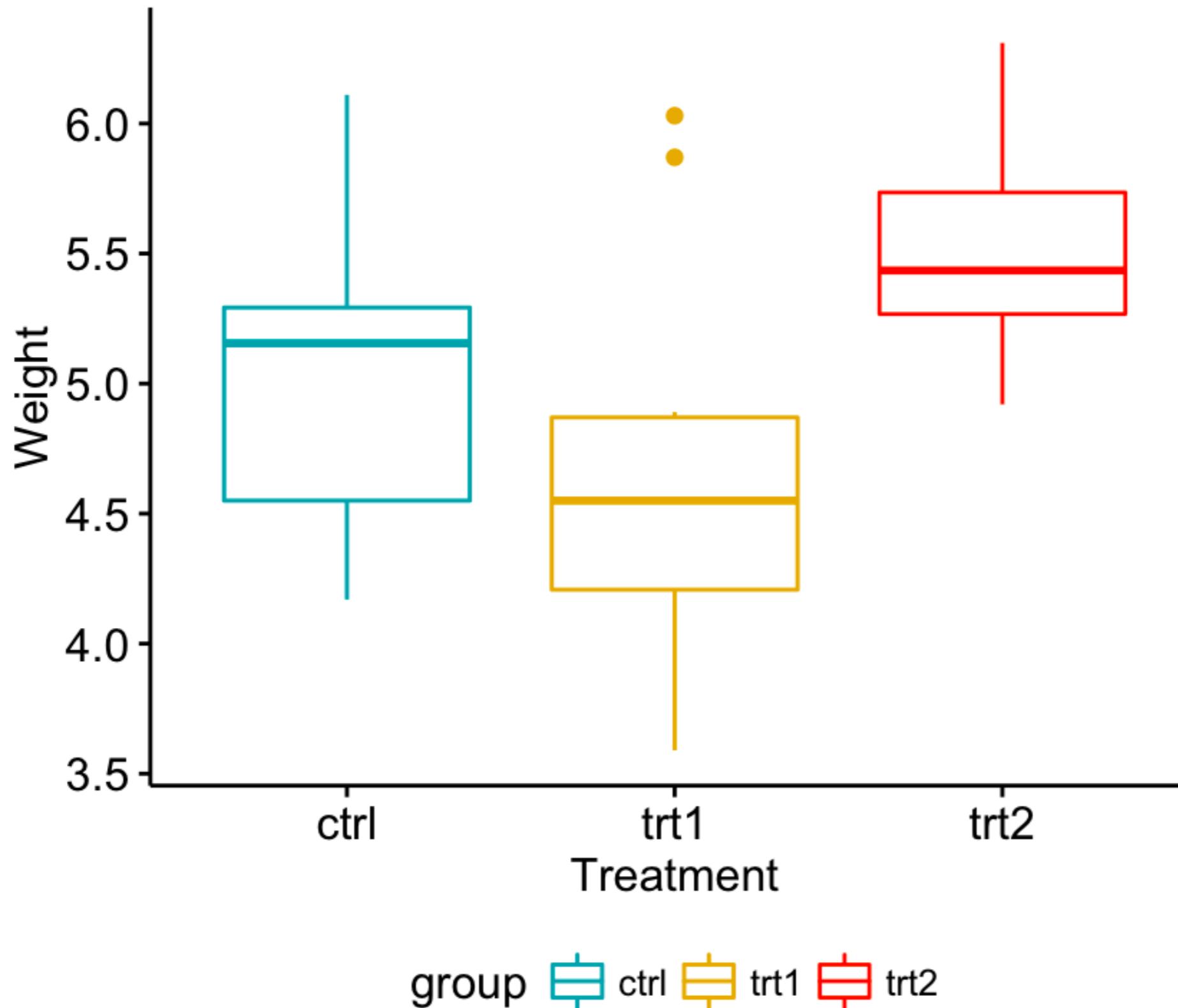
Signif. codes:	0 ‘***’	0.001 ‘**’	0.01 ‘*’	0.05 ‘.’
	0.1 ‘ ’	1		

Residual standard error: 0.3303 on 98 degrees of freedom

Multiple R-squared: 0.4694, Adjusted R-squared: 0.464

F-statistic: 86.7 on 1 and 98 DF, p-value: 3.78e-15

1-way ANOVA



1-way ANOVA

FactorA	A1	A2	A3
	y_{11}	y_{12}	y_{13}
	y_{21}

	y_{n1}		y_{nm}

1-way ANOVA

$$SST = SSX + SSE$$

$$SST = SSA + SSE$$

$$SST = SS_{among} + SS_{within}$$

$$SST = \sum_i \sum_j (y_{ij} - \bar{y})^2$$

$$SSA = \sum_i \sum_j (\bar{y}_{-j} - \bar{y})^2 = n \cdot \sum_j (\bar{y}_{-j} - \bar{y})^2$$

$$SSE = \sum_i \sum_j (y_{ij} - \bar{y}_{-j})^2$$

1-way ANOVA

$$SST = SSX + SSE$$

$$SST = SSA + SSE$$

$$SST = SS_{among} + SS_{within}$$

$$SST = \sum_i \sum_j (y_{ij} - \bar{y})^2 \text{ Сумма квадратов Y}$$

$$SSA = n \cdot \sum_j (\bar{y}_{-j} - \bar{y})^2 \text{ Сумма квадратов Y, объясняемая фактором A}$$

Фактически - сколько мы объясним, если будем предсказывать у только по j (значению фактора A)

$$SSE = \sum_i \sum_j (y_{ij} - \bar{y}_{-j})^2 \text{ Сумма квадратов Y, не объясняемая фактором A}$$

FactorA	A1	A2	A3
	\bar{y}_{-1}	\bar{y}_{-2}	\bar{y}_{-3}
	\bar{y}_{-1}

	\bar{y}_{-1}	\bar{y}_{-2}	\bar{y}_{-3}

1-way ANOVA

$$H_0 : \mu_{A_1} = \mu_{A_2} \dots = \mu_{A_n} = \mu$$

H_1 : Не все средние равны

FactorA	A1	A2	A3
	y_{11}	y_{12}	y_{13}
	y_{21}

	y_{n1}		y_{nm}

1-way ANOVA

$$H_0 : \mu_{A_1} = \mu_{A_2} \dots = \mu_{A_n} = \mu$$

Если H_0 верна, то фактически, у нас имеется n выборок (по 1 выборке на значение фактора А).

В каждой мы можем оценить дисперсию. Что мы можем потом сделать?

1-way ANOVA

$$H_0 : \mu_{A_1} = \mu_{A_2} \dots = \mu_{A_n} = \mu$$

1) Если H_0 верна, то фактически, у нас имеется n выборок (по 1 выборке на значение фактора A).

В каждой мы можем оценить дисперсию. Что мы можем потом сделать?

Подсчитать pooled variance

$$s_{pooled}^2 = \frac{\sum_j (n_j - 1)s_j^2}{\sum_j (n_j - 1)} = \frac{SSE}{N - a} = MSE$$

$$s_j^2 = \frac{1}{n_j - 1} \sum_i (y_{ij} - \bar{y}_{-j})^2 \quad SSE = \sum_i \sum_j (y_{ij} - \bar{y}_{-j})^2$$

1-way ANOVA

$$H_0 : \mu_{A_1} = \mu_{A_2} \dots = \mu_{A_n} = \mu$$

1) Если H_0 верна, то фактически, у нас имеется n выборок (по 1 выборке на значение фактора A)

В каждой мы можем оценить дисперсию. Что мы можем потом сделать?

Подсчитать pooled variance

$$\sigma^2 \approx s_{pooled}^2 = \frac{\sum_j (n_j - 1)s_j^2}{\sum_j (n_j - 1)} = \frac{SSE}{N - a} = MSE$$

$$s_j^2 = \frac{1}{n_j - 1} \sum_i (y_{ij} - \bar{y}_{-j})^2 \quad SSE = \sum_i \sum_j (y_{ij} - \bar{y}_{-j})^2$$

1-way ANOVA

$$H_0 : \mu_{A_1} = \mu_{A_2} \dots = \mu_{A_n} = \mu$$

2) Если H_0 верна, то средние для значений фактора распределены следующим образом

$$\bar{y}_{-j} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Мы опять можем оценить наши дисперсию, но уже через эти средние!

$$\frac{\sigma^2}{n} \approx \frac{1}{a-1} \sum_{i=1}^a (\bar{y}_{-j} - \bar{y})^2$$

$$\sigma^2 \approx \frac{1}{a-1} \cdot n \cdot \sum_{i=1}^a (\bar{y}_{-j} - \bar{y})^2$$

1-way ANOVA

$$H_0 : \mu_{A_1} = \mu_{A_2} \dots = \mu_{A_n} = \mu$$

2) Если H_0 верна, то средние для значений фактора распределены следующим образом

$$\bar{y}_{-j} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Мы опять можем оценить наши дисперсию, но уже через эти средние!

$$\sigma^2 \approx \frac{1}{a-1} \cdot n \cdot \sum_{i=1}^a (\bar{y}_{-j} - \bar{y})^2 \quad SSA = \sum_i \sum_j (\bar{y}_{-j} - \bar{y})^2 = n \cdot \sum_j (\bar{y}_{-j} - \bar{y})^2$$

$$\sigma^2 \approx \frac{SSA}{a-1} = MSA$$

1-way ANOVA

3) Если H_0 верна, то

$$\sigma^2 \approx \frac{1}{a-1} \cdot n \cdot \sum_{i=1}^a (\bar{y}_{-j} - \bar{y}) = MSA$$

а наблюдений
(наших средних),
считали общее
среднее - $df = a - 1$

$$\sigma^2 \approx s_{pooled}^2 = \frac{\sum_j (n_j - 1)s_j^2}{\sum_j (n_j - 1)} = \frac{SSE}{N - a} = MSE$$

п наблюдений (наших
средних), считали а
средних - потому $df = n$
- a

$$\frac{MSA}{MSE} \sim F(a - 1, n - a)$$

Распределение Фишера

X - выборка размера n

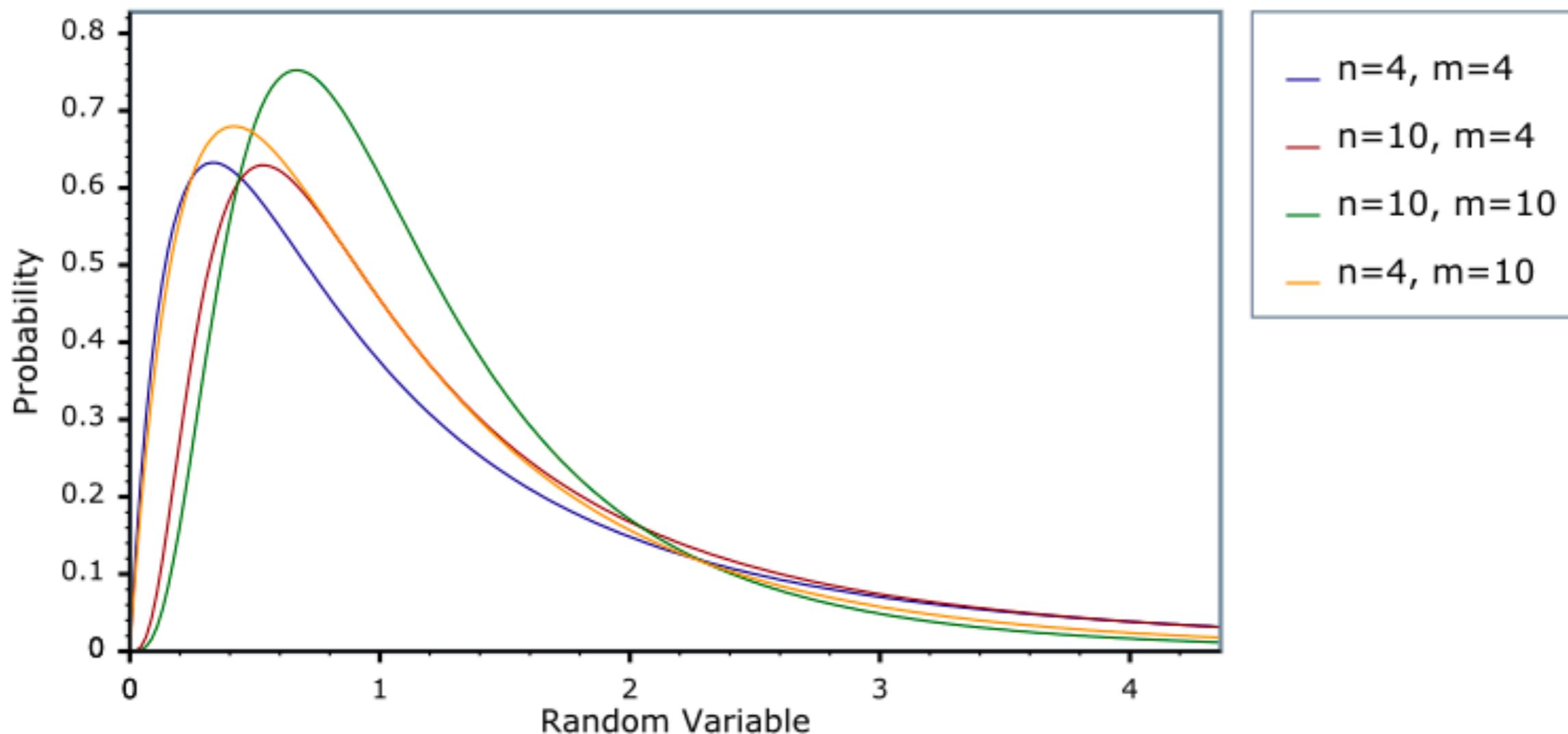
Y - выборка размера m

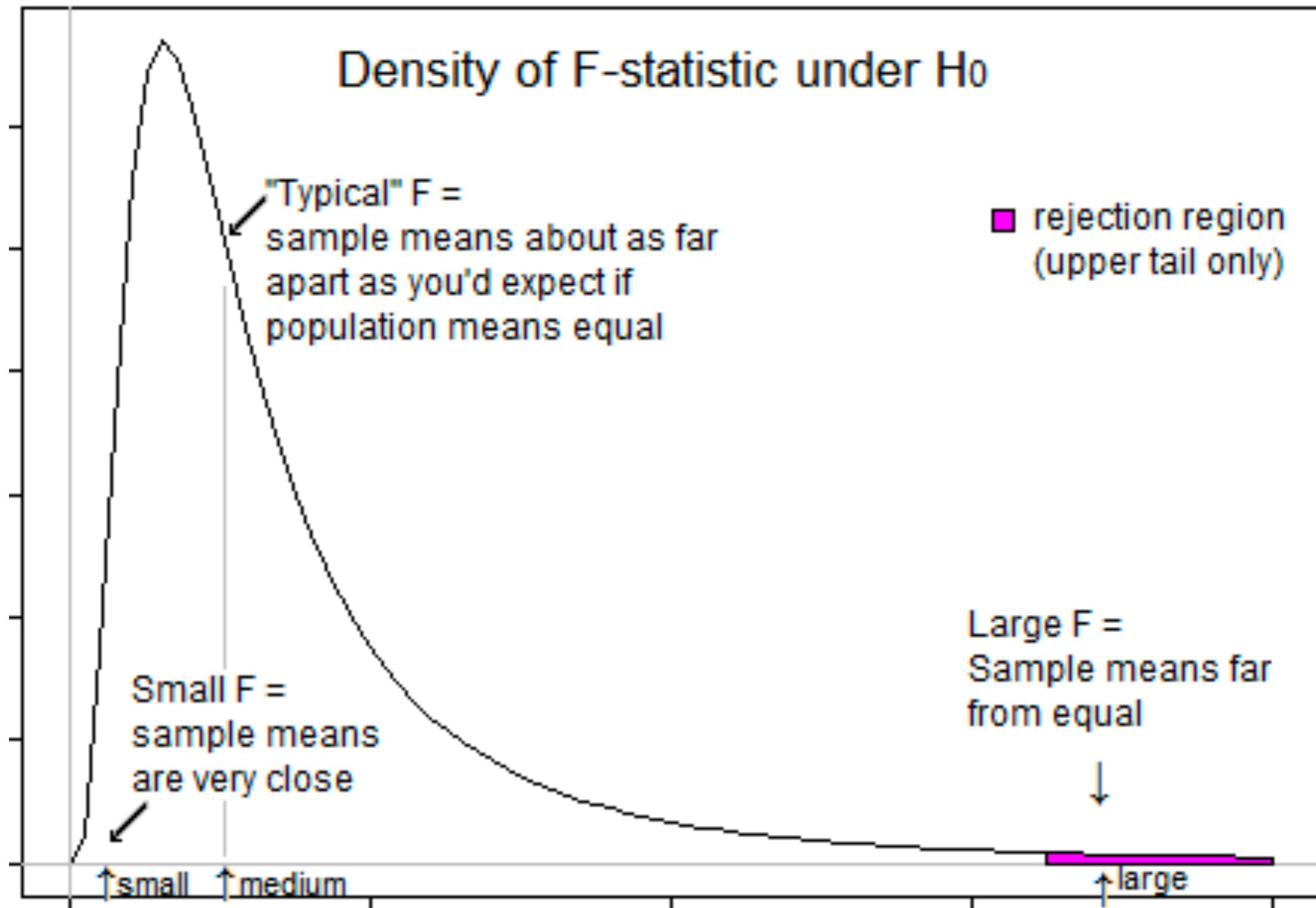
$$X \sim N(\mu_x, \sigma^2)$$

$$Y \sim N(\mu_y, \sigma^2)$$

$$\frac{S_x^2}{S_y^2} \sim F(n, m)$$

F Distribution PDF





**Нам нужен только правый хвост, так как значения F в левом хвосте говорят нам,
что средние еще более похожи, чем мы обычно ожидаем. Мы не считаем такие
случаи основанием отвергать гипотезу**

Замечание

- Строго говоря, для ANOVA не обязательно иметь выборки одинакового размера, хотя сбалансированная выборка и желательна

Предположения one-way ANOVA

- Независимость наблюдений
- Нормальность остатков
- Гомоскедастичность (однородность дисперсий)

Задача

Проверить гипотезу о равенстве средних в группах

Group1	Group2	Group3
51	23	56
45	43	76
33	23	74
45	43	87
67	45	56

Надо заполнить такую таблицу

	SS	df	MS	F	P-value
Factor	SSA	k-1	SSA/(k-1)	MSA/MSE	$P(F > \dots)$
Error	SSE	N-k	SSE/(N-k)		
Total	SST	N-1			

Group1	Group2	Group3
51	23	56
45	43	76
33	23	74
45	43	87
67	45	56

$$SST = \sum_i \sum_j (y_{ij} - \bar{y})^2 =$$

$$SSA = \sum_i \sum_j (\bar{y}_{-j} - \bar{y})^2 = n \cdot \sum_j (\bar{y}_{-j} - \bar{y})^2 =$$

$$SSE = \sum_i \sum_j (y_{ij} - \bar{y}_{-j})^2 =$$

Надо заполнить такую таблицу

	SS	df	MS	F	P-value
Factor	SSA	k-1	SSA/(k-1)	MSA/MSE	$P(F > \dots)$
Error	SSE	N-k	SSE/(N-k)		
Total	SST	N-1			

Group1	Group2	Group3
51	23	56
45	43	76
33	23	74
45	43	87
67	45	56

$$SST = \sum_i \sum_j (y_{ij} - \bar{y})^2 = 4884$$

$$\bar{y} = 51.13$$

Group1	Group2	Group3
51	23	56
45	43	76
33	23	74
45	43	87
67	45	56

$$\bar{y} = 51.13 \quad \bar{y}_{-1} = 48.2 \quad \bar{y}_{-2} = 35.4 \quad \bar{y}_{-3} = 69.8$$

$$SSE = \sum_i \sum_j (y_{ij} - \bar{y}_{-j})^2 =$$

$$SSA = \sum_i \sum_j (\bar{y}_{-j} - \bar{y})^2 = n \cdot \sum_j (\bar{y}_{-j} - \bar{y})^2 =$$

Group1	Group2	Group3
51	23	56
45	43	76
33	23	74
45	43	87
67	45	56

$$\bar{y} = 51.13 \quad \bar{y}_{-1} = 48.2 \quad \bar{y}_{-2} = 35.4 \quad \bar{y}_{-3} = 69.8$$

$$SSE = \sum_i \sum_j (y_{ij} - \bar{y}_{-j})^2 = 1860.8$$

$$\sum_i \sum_j (\bar{y}_{-j} - \bar{y})^2 = n \cdot \sum_j (\bar{y}_{-j} - \bar{y})^2 = 604.6 * 5 = 3022.9$$

	SS	df	MS	F	P-value
Factor	SSA	k-1	SSA/(k-1)	MSA/MSE	$P(F > \dots)$
Error	SSE	N-k	SSE/(N-k)		
Total	SST	N-1			

$$SST = \sum_i \sum_j (y_{ij} - \bar{y})^2 = 4884$$

$$SSE = \sum_i \sum_j (y_{ij} - \bar{y}_{-j})^2 = 1860.8$$

$$SSA = \sum_i \sum_j (\bar{y}_{-j} - \bar{y})^2 = n \cdot \sum_j (\bar{y}_{-j} - \bar{y})^2 = 3022.9$$

	SS	df	MS	F	P-value
Factor	3022.9				
Error	1860.8				
Total	4884				

	SS	df	MS	F	P-value
Factor	SSA	k-1	SSA/(k-1)	MSA/MSE	$P(F > \dots)$
Error	SSE	N-k	SSE/(N-k)		
Total	SST	N-1			

$$SST = \sum_i \sum_j (y_{ij} - \bar{y})^2 = 4884$$

$$SSE = \sum_i \sum_j (y_{ij} - \bar{y}_{-j})^2 = 1860.8$$

$$SSA = \sum_i \sum_j (\bar{y}_{-j} - \bar{y})^2 = n \cdot \sum_j (\bar{y}_{-j} - \bar{y})^2 = 3022.9$$

	SS	df	MS	F	P-value
Factor	3022.9	2			
Error	1860.8	12			
Total	4884	14			

	SS	df	MS	F	P-value
Factor	SSA	k-1	SSA/(k-1)	MSA/MSE	$P(F > \dots)$
Error	SSE	N-k	SSE/(N-k)		
Total	SST	N-1			

$$SST = \sum_i \sum_j (y_{ij} - \bar{y})^2 = 4884$$

$$SSE = \sum_i \sum_j (y_{ij} - \bar{y}_{-j})^2 = 1860.8$$

$$SSA = \sum_i \sum_j (\bar{y}_{-j} - \bar{y})^2 = n \cdot \sum_j (\bar{y}_{-j} - \bar{y})^2 = 3022.9$$

	SS	df	MS	F	P-value
Factor	3022.9	2	1511.5	9.75	
Error	1860.8	12	155		
Total	4884	14			

$$F_{(2,12)}(0.95) = 3.88$$

Отвергаем H_0

	SS	df	MS	F	P-value
Factor	SSA	k-1	SSA/(k-1)	MSA/MSE	$P(F > \dots)$
Error	SSE	N-k	SSE/(N-k)		
Total	SST	N-1			

$$SST = \sum_i \sum_j (y_{ij} - \bar{y})^2 = 4884$$

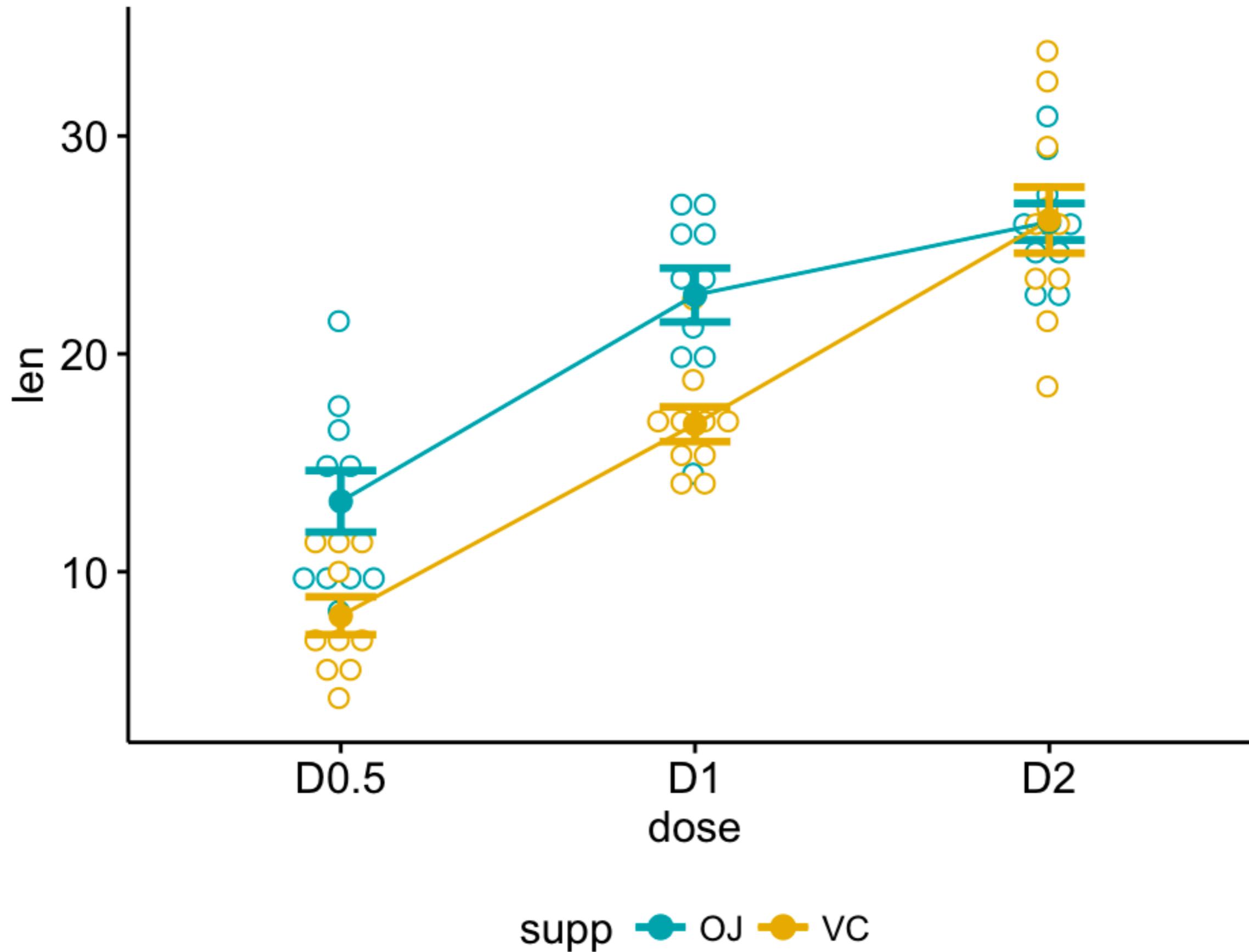
$$SSE = \sum_i \sum_j (y_{ij} - \bar{y}_{-j})^2 = 1860.8$$

$$SSA = \sum_i \sum_j (\bar{y}_{-j} - \bar{y})^2 = n \cdot \sum_j (\bar{y}_{-j} - \bar{y})^2 = 3022.9$$

	SS	df	MS	F	P-value
Factor	3022.9	2	302.3	1.95	0.003
Error	1860.8	12	155.1		
Total	4884	14			

**Отвергаем H_0
(фактор влияет)**

2-way ANOVA



2-way ANOVA (без взаимодействия)

На среднее могут влиять два фактора, взаимодействие факторов нет

$$SST = SSA + SSB + SSE \quad (1)$$

$$SST = \sum_i \sum_j (y_{ij} - \bar{y})^2$$

$$SSA = n \cdot \sum_j (\bar{y}_{\cdot j} - \bar{y})^2$$

m - число уровней фактора A

n - число уровней фактора B

$$SSB = m \cdot \sum_j (\bar{y}_{i \cdot} - \bar{y})^2$$

$$SSE = \sum_i \sum_j (y_{ij} - \bar{y}_{\cdot j} - \bar{y}_{i \cdot} + \bar{y})^2$$

(тут удобнее считать SSE через (1))

2-way ANOVA (без взаимодействия)

	SS	df	MS	F	P-value
Factor A	SSA	a-1	SSA/(a-1)	MSA/MSE	$P(F > \dots)$
Factor B	SSB	b-1	SSB/(b-1)	MSB/MSE	$P(F > \dots)$
Error	SSE	N-a-b+1	SSE/(N-a-b+1)		
Total	SST	N-1			

	Машина1	Машина2	Машина3	Машина 4
Работник1	21.20	23.30	20.25	21
Работник2	20.10	24.30	19.10	22
Работник3	19.42	25.42	18	21.4
Работник4	23.20	23.20	20	22.4
Работник5	20.20	23.20	19.35	20.47

	Машина1	Машина2	Машина3	Машина 4	
Работник1	21.20	23.30	20.25	21	21.44
Работник2	20.10	24.30	19.10	22	21.38
Работник3	19.42	25.42	18	21.4	21.06
Работник4	23.20	23.20	20	22.4	22.2
Работник5	20.20	23.20	19.35	20.47	20.81
	20.82	23.88	19.34	21.45	21.38

$$SST = \sum_i \sum_j (y_{ij} - \bar{y})^2 = 71.70$$

	Машина1	Машина2	Машина3	Машина 4	A
Работник1	21.20	23.30	20.25	21	21.44
Работник2	20.10	24.30	19.10	22	21.38
Работник3	19.42	25.42	18	21.4	21.06
Работник4	23.20	23.20	20	22.4	22.2
Работник5	20.20	23.20	19.35	20.47	20.81
B	20.82	23.88	19.34	21.45	21.38

$$SST = \sum_{i} \sum_{j} (y_{ij} - \bar{y})^2 = 71.70$$

$$SSA = n \cdot \sum_j (\bar{y}_{-j} - \bar{y})^2 = 10.75 * 5 = 53.73$$

$$SSB = m \cdot \sum_j (\bar{y}_{-j} - \bar{y})^2 = 1.10 * 4 = 4.40$$

	Машина1	Машина2	Машина3	Машина 4	A
Работник1	21.20	23.30	20.25	21	21.44
Работник2	20.10	24.30	19.10	22	21.38
Работник3	19.42	25.42	18	21.4	21.06
Работник4	23.20	23.20	20	22.4	22.2
Работник5	20.20	23.20	19.35	20.47	20.81
B	20.82	23.88	19.34	21.45	21.38

$$SST = \sum_i \sum_j (y_{ij} - \bar{y})^2 = 71.70$$

$$SSA = n \cdot \sum_j (\bar{y}_{-j} - \bar{y})^2 = 10.75 * 5 = 53.73$$

$$SSB = m \cdot \sum_j (\bar{y}_{-j} - \bar{y})^2 = 1.10 * 4 = 4.40$$

$$SSE = SST - SSA - SSB = 71.70 - 53.73 - 4.40 = 13.57$$

$$SST = \sum_i \sum_j (y_{ij} - \bar{y})^2 = 71.70$$

$$SSA = n \cdot \sum_j (\bar{y}_{-j} - \bar{y})^2 = 10.75 * 5 = 53.73$$

$$SSB = m \cdot \sum_j (\bar{y}_{-j} - \bar{y})^2 = 1.10 * 4 = 4.40$$

$$SSE = SST - SSA - SSB = 71.70 - 53.73 - 4.40 = 13.57$$

	SS	df	MS	F	P-value
FactorA	53.73				
FactorB	4.40				
Error	13.57				
Total	71.70				

$$SST = \sum_i \sum_j (y_{ij} - \bar{y})^2 = 71.70$$

$$SSA = n \cdot \sum_j (\bar{y}_{-j} - \bar{y})^2 = 10.75 * 5 = 53.73$$

$$SSB = m \cdot \sum_j (\bar{y}_{-j} - \bar{y})^2 = 1.10 * 4 = 4.40$$

$$SSE = SST - SSA - SSB = 71.70 - 53.73 - 4.40 = 13.57$$

	SS	df	MS	F	P-value
FactorA	53.73	3			
FactorB	4.40	4			
Error	13.57	12			
Total	71.70	19			

$$SST = \sum_i \sum_j (y_{ij} - \bar{y})^2 = 71.70$$

$$SSA = n \cdot \sum_j (\bar{y}_{-j} - \bar{y})^2 = 10.75 * 5 = 53.73$$

$$SSB = m \cdot \sum_j (\bar{y}_{-j} - \bar{y})^2 = 1.10 * 4 = 4.40$$

$$SSE = SST - SSA - SSB = 71.70 - 53.73 - 4.40 = 13.57$$

	SS	df	MS	F	P-value
FactorA	53.73	3	17.91		
FactorB	4.40	4	1.10		
Error	13.57	12	1.13		
Total	71.70	19			

$$SST = \sum_i \sum_j (y_{ij} - \bar{y})^2 = 71.70$$

$$SSA = n \cdot \sum_j (\bar{y}_{-j} - \bar{y})^2 = 10.75 * 5 = 53.73$$

$$SSB = m \cdot \sum_j (\bar{y}_{-j} - \bar{y})^2 = 1.10 * 4 = 4.40$$

$$SSE = SST - SSA - SSB = 71.70 - 53.73 - 4.40 = 13.57$$

	SS	df	MS	F	P-value	$F_{(3,12)}(0.95) = 3.49$
FactorA	53.73	3	17.91	16.28		$F_{(4,12)}(0.95) = 3.26$
FactorB	4.40	4	1.10	0.97		
Error	13.57	12	1.13			
Total	71.70	19				

**На уровне значимости 0.05 отвергаем равенство средних
для машин**

**На уровне значимости 0.05 не имеем оснований
отвергнуть равенство средних для работников**

	SS	df	MS	F	P-value
FactorA	53.73	3	17.91	16.28	
FactorB	4.40	4	1.10	0.97	
Error	13.57	12	1.13		
Total	71.70	19			

$$F_{(3,12)}(0.95) = 3.49$$

$$F_{(4,12)}(0.95) = 3.26$$

**На уровне значимости 0.05 отвергаем равенство средних
для машин (машины влияют)**

**На уровне значимости 0.05 не имеем оснований
отвергнуть равенство средних для работников**

	SS	df	MS	F	P-value
FactorA	53.73	3	17.91	16.28	0.0002
FactorB	4.40	4	1.10	0.97	0.46
Error	13.57	12	1.13		
Total	71.70	19			

2-way ANOVA (без взаимодействия)

- Независимость наблюдений
- Нормальность остатков
- Гомоскедастичность (однородность дисперсий)
- Аддитивный вклад факторов

2-way ANOVA (с взаимодействием)

		Strain			
		B	C	K-12	W
Temp	Low	4, 4	7, 3	7, 5	6, 4
	Medium	7, 5	9, 7	7, 7	7, 7
	High	5, 5	5, 5	9, 7	6, 6

На среднее могут влиять два фактора и их взаимодействие

Часто ли обоснованно исключать взаимодействие факторов?

2-way ANOVA (с взаимодействием)

		Strain			
		B	C	K-12	W
Temp	Low	4, 4	7, 3	7, 5	6, 4
	Medium	7, 5	9, 7	7, 7	7, 7
	High	5, 5	5, 5	9, 7	6, 6

На среднее могут влиять два фактора и их взаимодействие

Часто ли обоснованно исключать взаимодействие факторов?

Чаще всего - нет.

2-way ANOVA (с взаимодействием)

На среднее могут влиять два фактора и их взаимодействие

$$SST = SSA + SSB + SSAB + SSE \quad (1)$$

$$SST = \sum_i \sum_j \sum_k (y_{ijk} - \bar{y})^2$$

m - число уровней фактора A

$$SSA = r \cdot n \cdot \sum_j (\bar{y}_{-j-} - \bar{y})^2$$

n - число уровней фактора B

$$SSB = r \cdot m \cdot \sum_i^j (\bar{y}_{i-} - \bar{y})^2$$

k - число значений в каждой ячейке таблицы (число повторностей)

$$SSAB = r \cdot \sum_i \sum_j \sum_k (y_{ij-} - \bar{y}_{-j-} - \bar{y}_{i-} + \bar{y})^2$$

(тут удобнее считать SSAB через (1))

$$SSE = \sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{ij-})^2$$

	SS	df	MS	F	P-value
FactorA	SSA	$dfa=m-1$	MSA	MSA/MSE	$F(dfa, dfe)$
FactorB	SSB	$dfb=n-1$	MSB	MSB/MSE	$F(df_b, dfe)$
FactorAB	SSAB	$df_{ab} = (n - 1) * (m - 1)$	MSAB	MSAB/MSE	$F(df_{ab}, dfe)$
Error	SSE	$dfe = n * m * (r - 1)$	MSE		
Total	SST	$dft = n * m * r - 1$			

Биолог хочет сравнить скорости роста четырех штаммов бактерии *E. coli*. (K-12, B, C, and W) при трех температурах (низкая, средняя и высокая). Она берет 24 одинаковые пробирки с одинаковым количеством среды и одинаковым исходным числом бактерий, по две пробирки для каждого штамма и значения температуры, и измеряет оптическую плотность (OD) после 12-часовой инкубации.

		Strain			
		B	C	K-12	W
Temp	Low	4, 4	7, 3	7, 5	6, 4
	Medium	7, 5	9, 7	7, 7	7, 7
	High	5, 5	5, 5	9, 7	6, 6

- (a) (3 pts) Постройте таблицу ANOVA для этих данных. Объясните Ваши вычисления.
- (b) (2 pts) Протестируйте на 5% уровне значимости гипотезу о том, что скорость роста зависит от температуры, штамма, или комбинации этих двух факторов. Какие предположения при этом нужно сделать?
- (c) (1 pt) Основываясь на результатах предыдущего пункта, биолог записала в своем журнале, что скорость роста увеличивается с увеличением температуры. Прокомментируйте это утверждение и, если оно неверно, предложите статистический тест для его проверки.

	B	C	K-12	W
Low	4, 4	7, 3	7, 5	6, 4
Medium	7, 5	9, 7	7, 7	7, 7
High	5, 5	5, 5	9, 7	6, 6

	B	C	K-12	W
Low	4, 4	7, 3	7, 5	6, 4
Medium	7, 5	9, 7	7, 7	7, 7
High	5, 5	5, 5	9, 7	6, 6

$$\bar{y} = 6$$

$$SST = \sum_i \sum_j \sum_k (y_{ijk} - \bar{y})^2 = 54$$

Считаем средние в ячейках

	B	C	K-12	W
Low	4, 4	7, 3	7, 5	6, 4
Medium	7, 5	9, 7	7, 7	7, 7
High	5, 5	5, 5	9, 7	6, 6

	B	C	K-12	W
Low	4	5	6	5
Medium	6	8	7	7
High	5	5	8	6

$$SSE = \sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{ij-})^2 = 18$$

Считаем средние в ячейках

	B	C	K-12	W
Low	4, 4	7, 3	7, 5	6, 4
Medium	7, 5	9, 7	7, 7	7, 7
High	5, 5	5, 5	9, 7	6, 6

	B	C	K-12	W
Low	4	5	6	5
Medium	6	8	7	7
High	5	5	8	6

$$SSE = \sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{ij-})^2 = 18$$

Считаем средние в ячейках

	B	C	K-12	W	(A)
Low	4	5	6	5	5
Medium	6	8	7	7	7
High	5	5	8	6	6
(B)	5	6	7	6	6

Считаем, почти забыв про то, что были повторности (только домножить надо)

$$SSA = r \cdot n \cdot \sum_j (\bar{y}_{-j-} - \bar{y})^2 = 2 \cdot 3 \cdot 2 = 12$$

$$SSB = r \cdot m \cdot \sum_j (\bar{y}_{-j-} - \bar{y})^2 = 2 \cdot 4 \cdot 2 = 16$$

$$SST = SSA + SSB + SSAB + SSE$$

$$SST = \sum_i \sum_j \sum_k (y_{ijk} - \bar{y})^2 = 54$$

$$SSA = r \cdot n \cdot \sum_j (\bar{y}_{-j-} - \bar{y})^2 = 2 \cdot 3 \cdot 2 = 12$$

$$SSB = r \cdot m \cdot \sum_j (\bar{y}_{-j-} - \bar{y})^2 = 2 \cdot 4 \cdot 2 = 16$$

$$SSE = \sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{ij-})^2 = 18$$

$$SSAB = SST - SSA - SSB - SSE = 54 - 46 = 8$$

	SS	df	MS	F	P-value
FactorA	12		MSA	MSA/MSE	F(dfa, dfe)
FactorB	16		MSB	MSB/MSE	F(df _b , dfe)
FactorAB	8		MSAB	MSAB/MSE	F(df _{ab} , dfe)
Error	18		MSE		
Total	54				

$$SST = \sum_i \sum_j \sum_k (y_{ijk} - \bar{y})^2 = 54$$

$$SSA = r \cdot n \cdot \sum_j (\bar{y}_{-j-} - \bar{y})^2 = 2 \cdot 3 \cdot 2 = 12$$

$$SSB = r \cdot m \cdot \sum_j (\bar{y}_{-j-} - \bar{y})^2 = 2 \cdot 4 \cdot 2 = 16$$

$$SSE = \sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{ij-})^2 = 18$$

$$SSAB = SST - SSA - SSB - SSE = 54 - 46 = 8$$

	SS	df	MS	F	P-value
FactorA	12	3	MSA	MSA/MSE	$F(dfa, dfe)$
FactorB	16	2	MSB	MSB/MSE	$F(df_b, dfe)$
FactorAB	8	6	MSAB	MSAB/MSE	$F(df_{ab}, dfe)$
Error	18	12	MSE		
Total	54	23			

	SS	df	MS	F	P-value
FactorA	12	3	4	MSA/MSE	$F(dfa, dfe)$
FactorB	16	2	8	MSB/MSE	$F(df_b, dfe)$
FactorAB	8	6	1.33	MSAB/MSE	$F(df_{ab}, dfe)$
Error	18	12	1.50		
Total	54	23			

	SS	df	MS	F	P-value
FactorA	12	3	4	2.67	F(3, 12)
FactorB	16	2	8	5.33	F(2, 12)
FactorAB	8	6	1.33	0.89	F(6, 12)
Error	18	12	1.50		
Total	54	23			

(b) (2 pts) Протестируйте на 5% уровне значимости гипотезу о том, что скорость роста зависит от температуры, штамма, или комбинации этих двух факторов. Какие предположения при этом нужно сделать?

$$F_{(2,12)}(0.95) = 3.88 > 2.67 \text{ - нет основание предполагать, что рост бактерий зависит от штамма}$$

$$F_{(3,12)}(0.95) = 3.49 < 5.33 \text{ - есть основание предполагать, что рост бактерий зависит от температуры}$$

$$F_{(4,12)}(0.95) = 3.26 > 0.89 \text{ - нет основания отвергнуть гипотезу о том, что взаимодействия факторов нет}$$

(c) (1 pt) Основываясь на результатах предыдущего пункта, биолог записала в своем журнале, что скорость роста увеличивается с увеличением температуры. Прокомментируйте это утверждение и, если оно неверно, предложите статистический тест для его проверки.

(c) (1 pt) Основываясь на результатах предыдущего пункта, биолог записала в своем журнале, что скорость роста увеличивается с увеличением температуры. Прокомментируйте это утверждение и, если оно неверно, предложите статистический тест для его проверки.

Биолог неправа, т.к результат ее анализа дает основания считать, что рост зависит от температуры, но не дает понимания направления изменения роста

(c) (1 pt) Основываясь на результатах предыдущего пункта, биолог записала в своем журнале, что скорость роста увеличивается с увеличением температуры. Прокомментируйте это утверждение и, если оно неверно, предложите статистический тест для его проверки.

Биолог неправа, т.к результат ее анализа дает основания считать, что рост зависит от температуры, но не дает понимания направления изменения роста

К примеру, можно провести регрессионный анализ

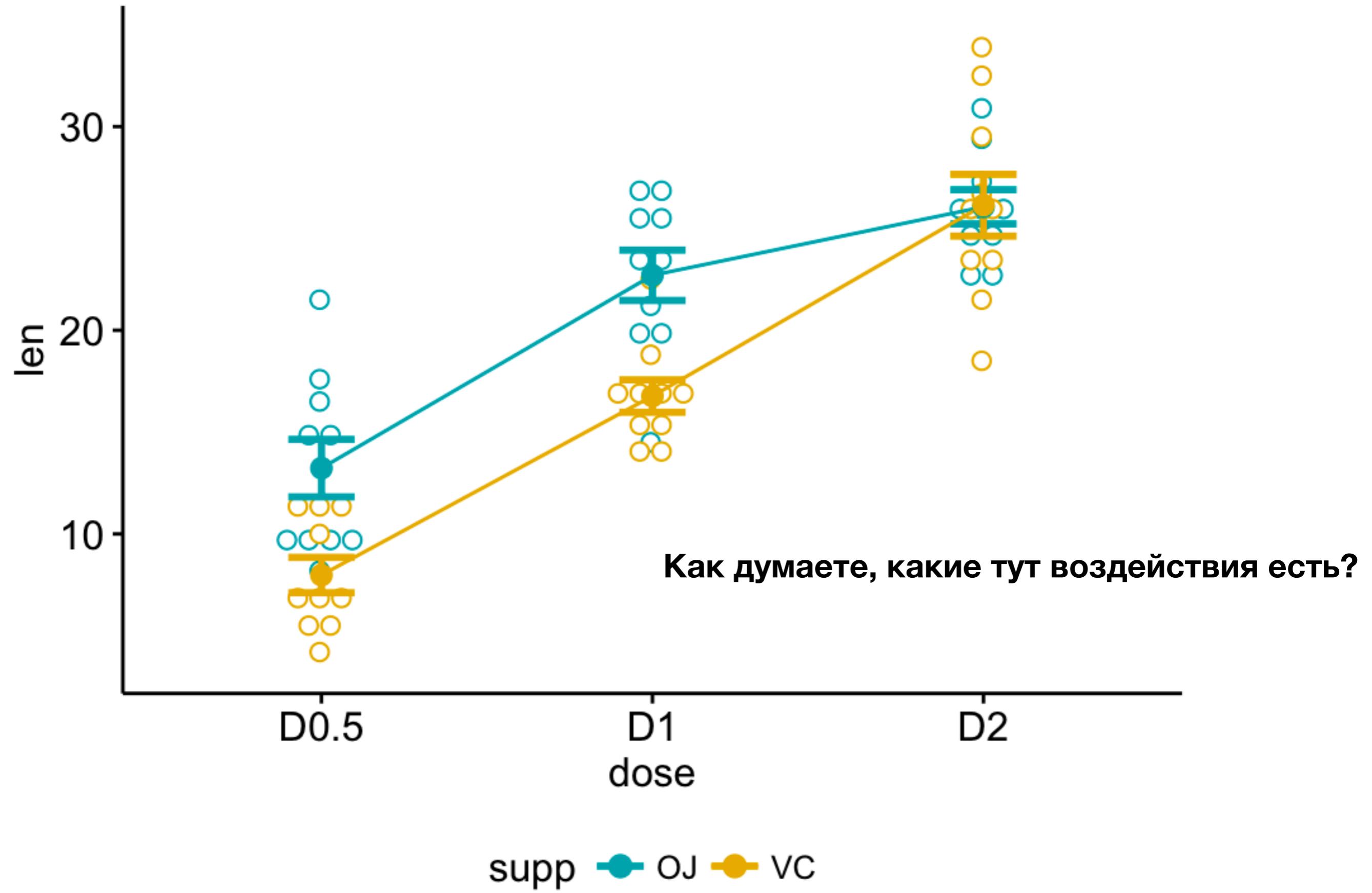
2-way ANOVA (с взаимодействием)

- Независимость наблюдений
- Нормальность остатков
- Гомоскедастичность (однородность дисперсий)
- Аддитивный вклад факторов ?

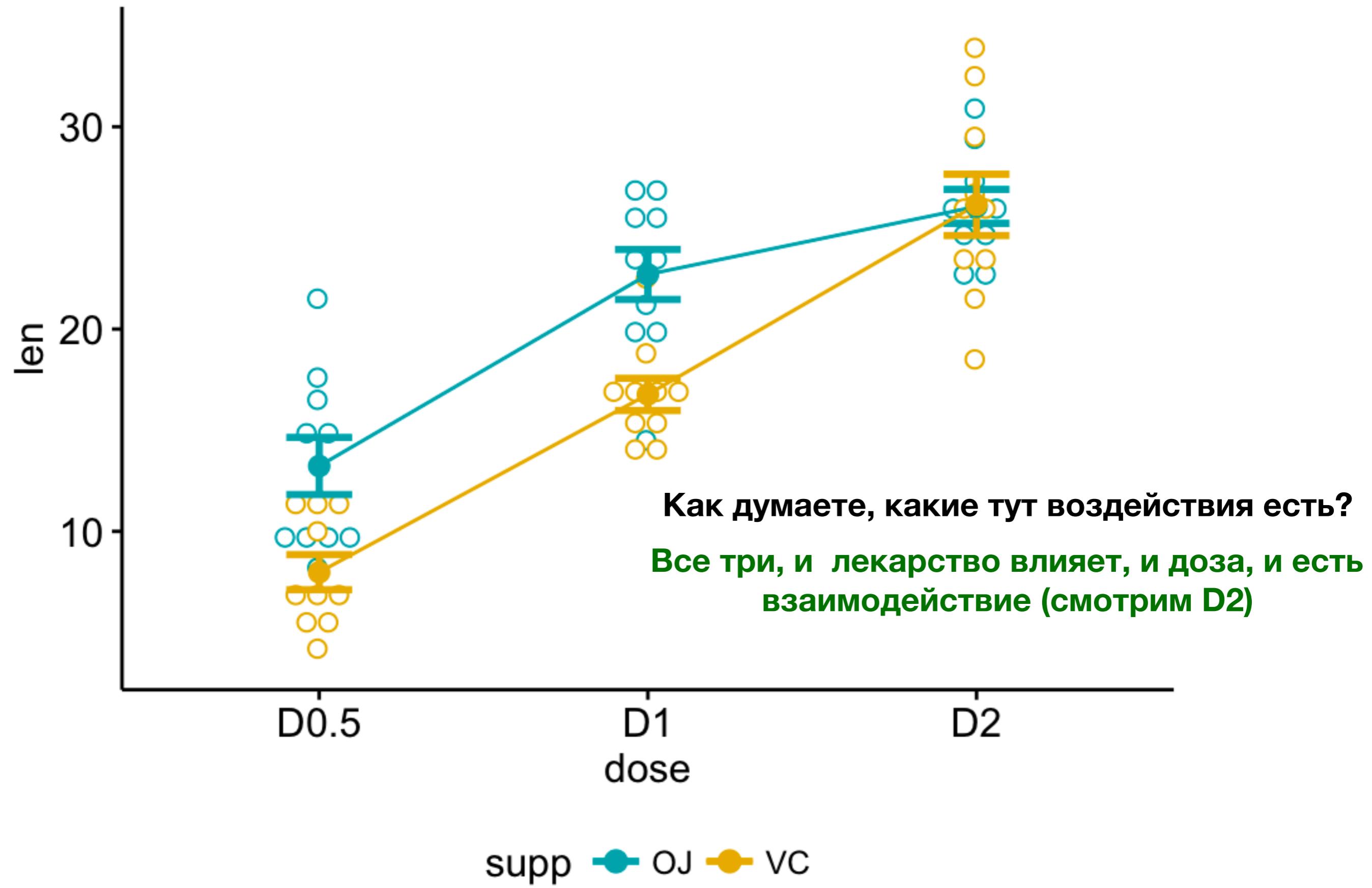
2-way ANOVA (с взаимодействием)

- Независимость наблюдений
- Нормальность остатков
- Гомоскедастичность (однородность дисперсий)
- **Аддитивный вклад факторов** ! Да, просто есть еще вклады взаимодействий факторов, но они тоже просто прибавляются

2-way ANOVA



2-way ANOVA



- Допустим, ANOVA дала для какого-то фактора значимые результаты. Как дальше определить, какой именно уровень/уровни фактора отличаются?

- Допустим, ANOVA дала для какого-то фактора значимые результаты. Как дальше определить, какой именно уровень/уровни фактора отличаются?
 - 1) можно сделать кучу парных t-тестов (не забывая про поправку на множественное тестирование)**
 - 2) можно использовать специально созданные под это тесты, которые позволяют делать более мягкую поправку или делать не все тесты в специфических случаях**

Профессор преподает статистику студентам магистерской и докторской программ. Он заметил, что оценки студентов зависят от того, где раньше учился студент, а также различаются между студентами магистерской и докторской программ. Профессор свел оценки по статистике (по шкале от 0 to 10) в следующую таблицу.

	Университет А	Университет В	Университет С
Магистры	4 5 6 5	4 6 7 7	7 6 7 8
PhD	5 7 8 8	6 8 9 9	8 9 10 9

- (3 pts) Постройте таблицу ANOVA для этих данных.
- (1 pt) На 1% уровне значимости проверьте гипотезу о том, что оценки зависят от типа программы.
- (1 pt) На 1% уровне значимости проверьте гипотезу о том, что оценки зависят от предыдущего университета. Что можно сказать о взаимодействии этих факторов?

Solution: First we compute the average in each group

	University A	University B	University C	Mean
Master	5	6	7	6
phD	7	8	9	8
Mean	6	7	8	7

$$SS(\text{within}) = (4 - 5)^2 + (5 - 5)^2 + \dots + (9 - 9)^2 = 24$$

$$SS(\text{level}) = ((6 - 7)^2 + (8 - 7)^2) * 3 * 4 = 24$$

$$SS(\text{school}) = ((6 - 7)^2 + (7 - 7)^2 + (8 - 7)^2) * 2 * 4 = 16$$

$$SS(\text{total}) = (4 - 7)^2 + (5 - 7)^2 + \dots + (9 - 7)^2 = 64$$

$$SS(\text{interaction}) = 64 - 24 - 24 - 16 = 0$$

Source	SS	df	MS	F-Value	$F_{0.05}$
Level	24	1	24	18.04	8.28
School	16	2	8	6.015	6.012
Level*School	0	2	0	0	6.012
Error	24	18	1.33		
Total	64	23			

Есть основания отвергнуть гипотезу о равенстве средних для разных программ

Есть слабые основания отвергать гипотезу о равенстве средних для университетов

Нет оснований отвергнуть гипотезу об отсутствии взаимодействия факторов друг с другом

6. Кофеин (в частности, содержащийся в кофе) является одним из наиболее широко распространенных стимуляторов. Более 90% взрослых американцев употребляют напитки с кофеином почти ежедневно (O'Callaghan *et al*, Risk Manag Healthcare Policy 2018, 11:263–271). Несмотря на то, что кофеин очевидно улучшает производительность труда, есть опасения, что он плохо влияет на сон. Поэтому некоторые люди предпочитают не употреблять напитки с кофеином во второй половине дня.

В одном исследовании производилось изучение потребления кофеина мужчинами и женщинами среди сотрудников небольшой ИТ компании. Входящих в здание сотрудников спрашивали: “В какое время дня вы больше всего любите пить кофе?”

	Утро	День	Вечер
Женщины	17	4	8
Мужчины	20	6	15

- (a) (1 pt) Можно ли использовать χ^2 -тест для того, чтобы оценить степень ассоциации между полом и потреблением кофе в разное время суток? Объясните свой ответ.
- (b) (1 pt) К какому выводу Вы приходите относительно ассоциации между полом и потреблением кофе в разное время суток на 5% уровне значимости? Четко сформулируйте свой вывод.
- (c) (1 pt) Как изменится Ваш вывод, если объединить категории “День” и “Вечер” в категорию “Вторая половина дня”?
- (d) (1 pt) Можно заметить, что мужчины больше женщин предпочитают пить кофе во второй половине дня. Поддерживает ли это утверждение ваш результат из пункта (c)? Если да, то на каком уровне значимости?
- (e) (1 pt) Как вы думаете, можно ли распространить результаты этого исследования на всех любителей кофе? Объясните почему да или почему нет.

6

a) надо подсчитать таблицу ожидаемых значений, если больше 5 - используем хи-квадрат, иначе точный тест Фишера

c) опять считаем таблицу ожидаемых значений, тут больше 5 - используем хи-квадрат

d) считаете доли мужчин и женщин, предпочитающих пить кофе во второй половине дня. Делаете тест пропорций **ИЛИ** используете тот факт, что p-value теста хи-квадрат равно p-value двустороннего теста пропорций. То есть надо разделить p-value предыдущего пункта на 2, чтобы получить односторонний тест (но до этого надо проверить, какая из долей больше)

e) selection bias - выбираем только людей из IT компании.

удобная выборка - просто набрали людей, входящих в здание

Non-response bias - в выборке нет людей, работающих удаленно

Точно **НЕВЕРНЫЕ** ответы на e:

“Если для IT-компаний, то можно” - нет, есть ошибка неответа и опрошена одна компания

“Нельзя судить, так как у нас выборка, а судим о генеральной совокупности” - задача статистики в этом и состоит

“Нельзя, так как выборка маленькая” - маленькая выборка тоже может хорошо охарактеризовать генеральную совокупность, если ее верно составить.

И точной ошибкой будет в любом из пунктов написать “принимаем H_0 ”.

3. Если Вы тестируете нулевую гипотезу о том, что выборка 0.52, 0.01, 0.22, 0.35, 0.28, 0.03, 0.15, 0.5, 0.14, 0.44 получена из равномерного распределения на интервале [0, 1], то чему равно значение тестовой статистики?
- A. 0.26
 - B. 0.30
 - C. 0.36
 - D. 0.42
 - E. 0.48

Solution: One-sample KS-test.

x	eCDF	D	x	eCDF	D
0.01	0.10	0.09	0.28	0.60	0.32
0.03	0.20	0.17	0.35	0.70	0.35
0.14	0.30	0.16	0.44	0.80	0.36
0.15	0.40	0.25	0.50	0.90	0.40
0.22	0.50	0.28	0.52	1.00	0.48

$$\max\{D\} = 0.48$$

5. Директор программы биоинформатики в некотором университете сообщает, что 60% её студентов — биологи, 20% — медики, а оставшиеся 20% — специалисты из других областей. Случайная выборка из 50 студентов состоит из 25 биологов, 17 медиков и 8 специалистов других специальностей. Достаточно ли оснований считать, что заявление директора программы не соответствует действительности?

- A. Да, на 0.5% уровне значимости
- B. Да, на 1%, но не на 0.5% уровне значимости
- C. Да, на 5%, но не на 1% уровне значимости
- D. Да, на 10%, но не на 5% уровне значимости
- E. Нет, даже на 10% уровне значимости

Solution:

χ^2 – test of Goodness of Fit

Expected: Biology=30, Medicine=10, Other=10

Observed: Biology=25, Medicine=17, Other=8

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{5^2}{30} + \frac{7^2}{10} + \frac{2^2}{10} = 6.13, \chi^2_{0.05}(2) = 5.99 < 6.13 < \chi^2_{0.01}(2) = 9.21,$$

hence H_0 is rejected at the 5%, but not at 1% significance level.