Лекция 8

Сравнение средних по связанным парам и независимым выборкам

Сравнение дисперсий по независимым выборкам

Две задачи на сравнение средних

Пример 1 *(Newbold)*. A random sample of six salespersons who had attended a motivational course on sales techniques was monitored in the three months before and the three months after the course. The table shows the values of sales, in thousands of dollars, generated by these six salespersons in the two periods.

Salesperson	*Before the course*	*After the course*
1	212	237
2	282	291
3	203	191
4	327	341
5	165	192
6	198	180

Assuming that the population distributions are normal find an 80% confidence interval for the difference between two population means.

Пример 2 *(Демешев)*. Вашему вниманию представлены результаты прыжков в длину Васи Сидорова на двух соревнованиях. На первых среди болельщиц присутствовала Аня Иванова:

1.83 1.64 2.27 1.78 1.89 2.33 1.61 2.31

На вторых Аня среди болельщиц не присутствовала:

1.26 1.41 2.05 1.07 1.59 1.96 1.29 1.52 1.18 1.47

Влияет ли присутствие Ани Ивановой на результаты Васи Сидорова?

Пример 1 — связанные пары *Пример 2 — независимые выборки*

Связанные пары: сравнение средних

Выборка из наблюдений за двумя признаками:

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$$

В примере 1: X_i — объём продаж продавца *i* до прохождения курсов, Y_i — объём продаж того же продавца после курсов.

Часто нас интересуют не столько величины *X* и *Y* по отдельности, сколько различия между ними: в примере требовалось найти «difference between the population means» — оценить разницу между средними показателями в двух наборах данных, т. е. измерить эффективность курсов.

Попробуем рассчитать требуемый доверительный интервал, а заодно проверим гипотезу о том, что курсы неэффективны:

$H_0 : \mu_X = \mu_Y$ - курсы недействены;

$H_A : \mu_X < \mu_Y$ - курсы способствуют повышению продаж.

Выберем уровень значимости 10%.

Рассмотрим величины $d_i = X_i - Y_i$.

$$\mathrm{E}(d_i) = \mathrm{E}(X_i) - \mathrm{E}(Y_i).$$

Получается, оценить $\mu_X - \mu_Y$ — то же, что оценить μ_d, а это мы умеем.
Сделаем обычные предположения:

$$d_1, \ldots, d_n \text{ независимы}, \quad d_i \sim N(\mu_X - \mu_Y, \sigma_d^2).$$

Точечная оценка для разности средних:

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i = \frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i).$$

Точечная оценка для дисперсии разностей:

$$\hat{\sigma}_d^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} d_i^2 - n(\bar{d})^2\right).$$

Используем обычный доверительный интервал для м.о. и *t*-статистику:

$$\bar{d} - t_{n-1,\frac{\alpha}{2}}\frac{\hat{\sigma}_d}{\sqrt{n}} < \mu_X - \mu_Y < \bar{d} + t_{n-1,\frac{\alpha}{2}}\frac{\hat{\sigma}_d}{\sqrt{n}}$$

$$t = \frac{\bar{d} - \mu_0}{\hat{\sigma}_d/\sqrt{n}} \overset{H_0}{\sim} t_{n-1}$$

что это?

Решаем задачу про продавцов.

Рассчитаем разности d_i:

Salesperson	*Before the course*	*After the course*	*Diff* (d_i)
1	212	237	-25
2	282	291	-9
3	203	191	12
4	327	341	-14
5	165	192	-27
6	198	180	18

Средняя разность: $\bar{d} = -7.5.$ Оценка дисперсии: $\hat{\sigma}_d^2 = 352.3.$

Оценка станд. отклонения: $\hat{\sigma}_d = 18.77.$

Квантиль: $t_{n-1,\frac{\alpha}{2}} = t_{5,\frac{0.2}{2}} = 1.476.$

Нижняя граница дов. интервала: $-7.5 - \frac{1.476 \times 18.77}{\sqrt{5}} = -18.8.$

Верхняя граница дов. интервала: $-7.5 + \frac{1.476 \times 18.77}{\sqrt{5}} = 3.9.$

t-статистика: $\frac{-7.5-0}{18.77/\sqrt{5}} = -0.98.$

Критическое значение: $t_{5,0.1} = 1.476.$ *p*-значение: $P(U < -0.98) = 0.186,\ U \sim t_5.$

Выводы:

> нет оснований считать, что объём продаж в среднем повышается после курсов;

> интервальная оценка среднего изменения объёма продаж после курсов:

$$-18.8 < \mu_X - \mu_Y < 3.9$$

Не надо забывать о визуализации!

Можно было бы построить гистограмму для d_i (может, лучше брать $-d_i$ — прирост продаж после курсов), но у нас слишком мало наблюдений.

Другой вариант — ящик-с-усами (box-and-whiskers plot), но для него тоже наблюдений маловато.

Ещё пример: подбор сопоставимых пар (мэтчинг)

Задача 2. Исследователь изучает заработные платы выпускников кое-какого вуза. Из выпускников прошлого года он отобрал семь юношей и семь девушек и составил из них пары таким образом, чтобы в каждой паре выпускники имели примерно одинаковый средний балл в дипломе. Сведения о зарплатах (в тыс. руб.) он свёл в таблицу:

№	Юноша	Девушка
1	38	41
2	46	38
3	28	30
4	56	40
5	20	24
6	45	27
7	37	35

(a) Предположив, что разница заработных плат имеет нормальное распределение, определите, есть ли основания считать, что юношей быть выгоднее. Используйте уровень значимости 10%.

(b) Исследователь захотел проверить ту же гипотезу с помощью критерия знаков и критерия знаковых рангов. Рассчитайте соответствующие статистики. *Проверять до конца не надо.*

(c) По этим данным был рассчитан доверительный интервал для средней разности заработных плат: (-2.47; 12.47). Какой доверительной вероятности он соответствует?

Независимые выборки:

нормальное распределение, дисперсии равны.

Есть две выборки из разных генеральных совокупностей:

$$X_1, \ldots, X_{n_X}, \quad X_i \sim N\left(\mu_X, \sigma^2\right);$$
$$Y_1, \ldots, Y_{n_Y}, \quad Y_i \sim N\left(\mu_Y, \sigma^2\right).$$

Все наблюдаемые величины X_i и Y_i независимы между собой.

Мы хотим оценить разницу в генеральных средних $\mu_X - \mu_Y$ или проверить какую-либо гипотезу об этой разнице (например, что $\mu_X = \mu_Y$).

Точечная оценка очевидна: $\bar{X} - \bar{Y}$.

Кое-что о ней: $\mathrm{E}\left(\bar{X} - \bar{Y}\right) = \mathrm{E}\left(\bar{X}\right) - \mathrm{E}\left(\bar{Y}\right) = \mu_X - \mu_Y$.

$$\mathrm{D}\left(\bar{X} - \bar{Y}\right) = \mathrm{D}\left(\bar{X}\right) + \mathrm{D}\left(-\bar{Y}\right) + 2\,\mathrm{Cov}\left(\bar{X}, -\bar{Y}\right) = \frac{\sigma^2}{n_X} + \frac{\sigma^2}{n_Y} = \sigma^2 \frac{n_X + n_Y}{n_X n_Y}.$$

(т. к. Cov = 0 из-за независимости)

$$\bar{X} - \bar{Y} \sim N\left(\mu_X - \mu_Y, \sigma^2 \frac{n_X + n_Y}{n_X n_Y}\right).$$

(сумма и разность независимых нормальных величин тоже нормальны)

Нам пригодится оценка для дисперсии.

Мы можем оценить дисперсию по каждой из выборок:

$$\hat{\sigma}_X^2 = \frac{1}{n_X - 1} \sum_{i=1}^{n_X} (X_i - \bar{X})^2,$$

$$\hat{\sigma}_Y^2 = \frac{1}{n_Y - 1} \sum_{i=1}^{n_Y} (Y_i - \bar{Y})^2,$$

$$\mathrm{E}(\hat{\sigma}_X^2) = \mathrm{E}(\hat{\sigma}_Y^2) = \sigma^2.$$

Рассмотрим несмещённую оценку для дисперсии, опирающуюся на обе выборки:

$$\hat{\sigma}^2 = \frac{1}{n_X + n_Y - 2} \left(\sum_{i=1}^{n_X} (X_i - \bar{X})^2 + \sum_{i=1}^{n_Y} (Y_i - \bar{Y})^2 \right) = \frac{(n_X - 1)\hat{\sigma}_X^2 + (n_Y - 1)\hat{\sigma}_Y^2}{n_X + n_Y - 2}.$$

Покажем, что $\frac{(n_X + n_Y - 2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n_X + n_Y - 2}$:

$$\frac{(n_X + n_Y - 2)\hat{\sigma}^2}{\sigma^2} = \frac{(n_X - 1)\hat{\sigma}_X^2}{\sigma^2} + \frac{(n_Y - 1)\hat{\sigma}_Y^2}{\sigma^2}$$

- сумма независимых хи-квадрат величин.

Из определения распределения хи-квадрат видно, что сумма независимых хи-квадрат величин с $(n_X - 1)$ и $(n_Y - 1)$ степенью свободы будет и сама иметь распределение хи-квадрат с $(n_X - 1) + (n_Y - 1) = (n_X + n_Y - 2)$ степенями свободы.

Теперь докажем ключевое для построения интервалов и проверки гипотез утверждение:

$$U=\frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\hat{\sigma}\sqrt{\frac{n_X+n_Y}{n_X n_Y}}}\sim t_{n_X+n_Y-2}.$$

Доказательство аналогично тому, что было в лекции 3. Представим U в таком виде:

$$U=\frac{\left(\frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sigma\sqrt{\frac{n_X+n_Y}{n_X n_Y}}}\right)}{\left(\frac{\hat{\sigma}\sqrt{\frac{n_X+n_Y}{n_X n_Y}}}{\sigma\sqrt{\frac{n_X+n_Y}{n_X n_Y}}}\right)}=\frac{\left(\frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sigma\sqrt{\frac{n_X+n_Y}{n_X n_Y}}}\right)}{\sqrt{\frac{\hat{\sigma}^2}{\sigma^2}}}=\frac{\left(\frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sigma\sqrt{\frac{n_X+n_Y}{n_X n_Y}}}\right)}{\sqrt{\frac{1}{(n_X+n_Y-2)}\frac{(n_X+n_Y-2)\hat{\sigma}^2}{\sigma^2}}}=$$

$$=\frac{N(0,1)}{\sqrt{\frac{1}{(n_X+n_Y-2)}\chi^2_{n_X+n_Y-2}}}\sim t_{n_X+n_Y-2}.$$

Для последнего перехода (к распределению Стьюдента) нужна независимость. Откуда она?

Теперь ясно, как проверять гипотезу о разности средних.

$$H_0: \mu_X - \mu_Y = \delta_0.$$

$$H_A: \mu_X - \mu_Y \neq \delta_0, \text{ или } H_A: \mu_X - \mu_Y < \delta_0, \text{ или } H_A: \mu_X - \mu_Y > \delta_0.$$

Критическая статистика: $$t = \frac{(\bar{X} - \bar{Y}) - \delta_0}{\hat{\sigma}\sqrt{\frac{n_X + n_Y}{n_X n_Y}}} \overset{H_0}{\sim} t_{n_X + n_Y - 2}.$$

Критическая область и *p*-значение — как обычно.

Теперь вернёмся к Васе и Ане.

Пример 2. Вашему вниманию представлены результаты прыжков в длину Васи Сидорова на двух соревнованиях. На первых среди болельщиц присутствовала Аня Иванова:

1.83 1.64 2.27 1.78 1.89 2.33 1.61 2.31

На вторых Аня среди болельщиц не присутствовала:

1.26 1.41 2.05 1.07 1.59 1.96 1.29 1.52 1.18 1.47

Влияет ли присутствие Ани Ивановой на результаты Васи Сидорова?

Решение. $H_0: \mu_X = \mu_Y$

$H_A: \mu_X \neq \mu_Y$.

Выберем уровень значимости 5%.

	с Аней	*без Ани*
Выборочное среднее:	$\bar{X} = 1.9575$	$\bar{Y} = 1.48$
Оценка дисперсии:	$\hat{\sigma}_X^2 = 0.0907$	$\hat{\sigma}_Y^2 = 0.1018$

Оценка дисперсии по обеим выборкам: $\hat{\sigma}^2 = (7 \times 0.0907 + 9 \times 0.1018) : 16 = 0.0969.$

Оценка станд. отклонения: $\hat{\sigma} = \sqrt{0.0969} = 0.3114.$

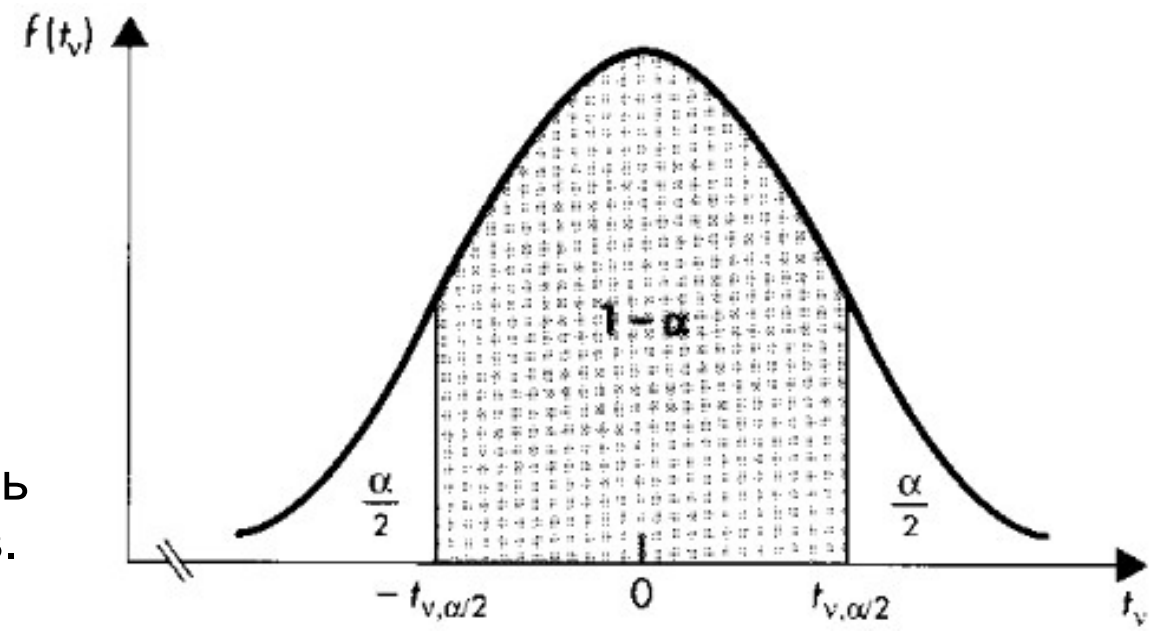
t-статистика: $\dfrac{1.9575 - 1.48}{0.3114 \times \sqrt{\dfrac{10 + 8}{80}}} = 3.233.$ Критическое значение: $t_{16, \frac{0.05}{2}} = 2.12.$

Вывод: $|t| > 2.12 \Rightarrow H_0$ отвергается, результаты Васи Сидорова зависят от присутствия Ани Ивановой.

А теперь построим доверительный интервал для разности средних с уровнем доверия (1-α).

Напоминаю: речь о независимых выборках из нормального распределения с одинаковой дисперсией.

$$U=\frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\hat{\sigma}\sqrt{\frac{n_X+n_Y}{n_X n_Y}}}\sim t_{n_X+n_Y-2}.$$



Берём число $t_{n_X+n_Y-2,\frac{\alpha}{2}}$ - квантиль расп-я Стьюдента с n_X+n_Y-2 ст. св. порядка $1-\frac{\alpha}{2}$).

$$1-\alpha=P\left(-t_{n_X+n_Y-2,\frac{\alpha}{2}}<\frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\hat{\sigma}\sqrt{\frac{n_X+n_Y}{n_X n_Y}}}<t_{n_X+n_Y-2,\frac{\alpha}{2}}\right)=$$

$$=P\left(\bar{X}-\bar{Y}-t_{n_X+n_Y-2,\frac{\alpha}{2}}\hat{\sigma}\sqrt{\frac{n_X+n_Y}{n_X n_Y}}<\mu_X-\mu_Y<\bar{X}-\bar{Y}+t_{n_X+n_Y-2,\frac{\alpha}{2}}\hat{\sigma}\sqrt{\frac{n_X+n_Y}{n_X n_Y}}\right).$$

Получили выражение для доверительного интервала:

$$\bar{X}-\bar{Y}-t_{n_X+n_Y-2,\frac{\alpha}{2}}\hat{\sigma}\sqrt{\frac{n_X+n_Y}{n_X n_Y}}<\mu_X-\mu_Y<\bar{X}-\bar{Y}+t_{n_X+n_Y-2,\frac{\alpha}{2}}\hat{\sigma}\sqrt{\frac{n_X+n_Y}{n_X n_Y}}$$

Применим его к задаче про Васю и Аню: рассчитаем 90% доверительный интервал для разности ожидаемых длин прыжков с Аней и без неё.

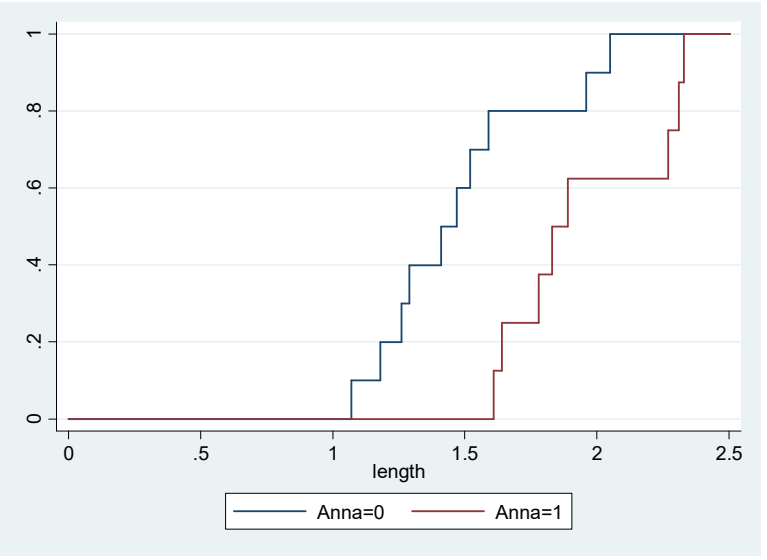
Табличное значение: $t_{16,\frac{0.1}{2}}=1.746.$

$$1.9575-1.48-1.746\times 0.3114\sqrt{\frac{8+10}{80}}<\mu_X-\mu_Y<1.9575-1.48+1.746\times 0.3114\sqrt{\frac{8+10}{80}}$$

$$0.22<\mu_X-\mu_Y<0.74$$

Итак, в присутствии Ани Вася прыгает дальше в среднем на 22 — 74 см.

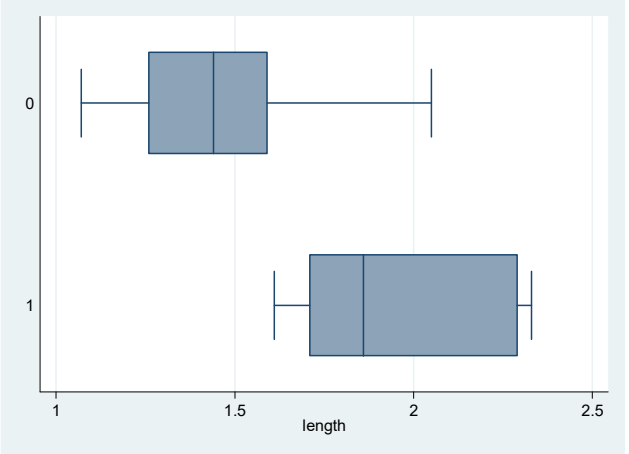
Визуализация, функции распределения:



Усатые ящики:

Ани нет

Аня есть



Как проверить, что дисперсии в двух генеральных совокупностях равны?
В этом нам поможет

распределение Снедекора-Фишера

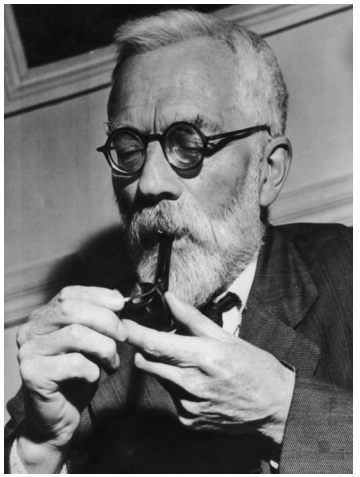
Пусть $Y_1 \sim \chi^2_m$, $Y_2 \sim \chi^2_n$, Y_1 и Y_2 независимы, $W = \dfrac{\frac{1}{m} Y_1}{\frac{1}{n} Y_2}$.

Распределение величины W называется распределением Снедекора-Фишера с m и n степенями свободы.

Обозначается $W \sim F_{m,n}$.

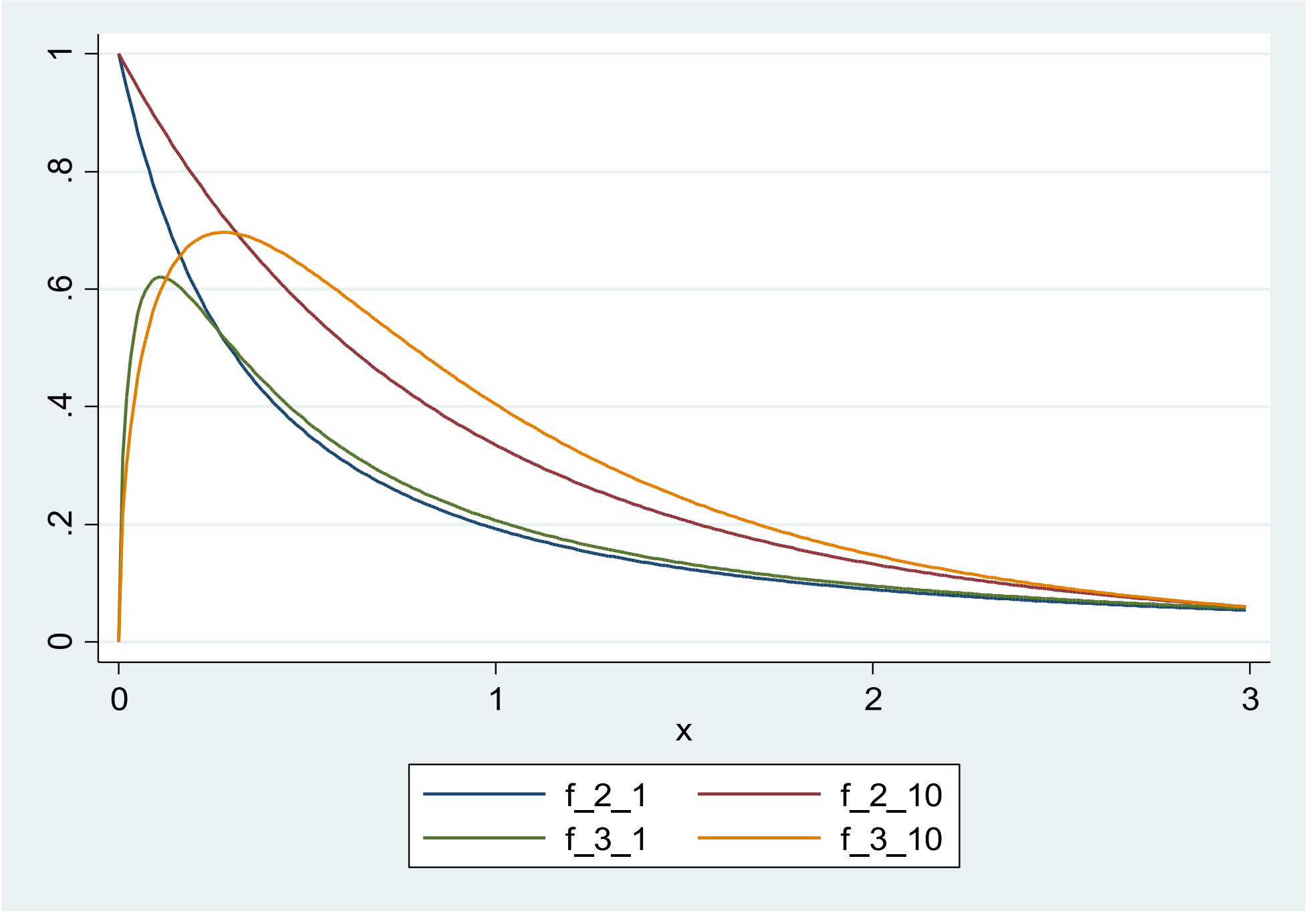
$$\mathrm{E}(W) = \frac{n}{n-2}, \quad n > 2; \qquad \mathrm{D}(W) = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}, \quad n > 4.$$

Джордж Снедекор

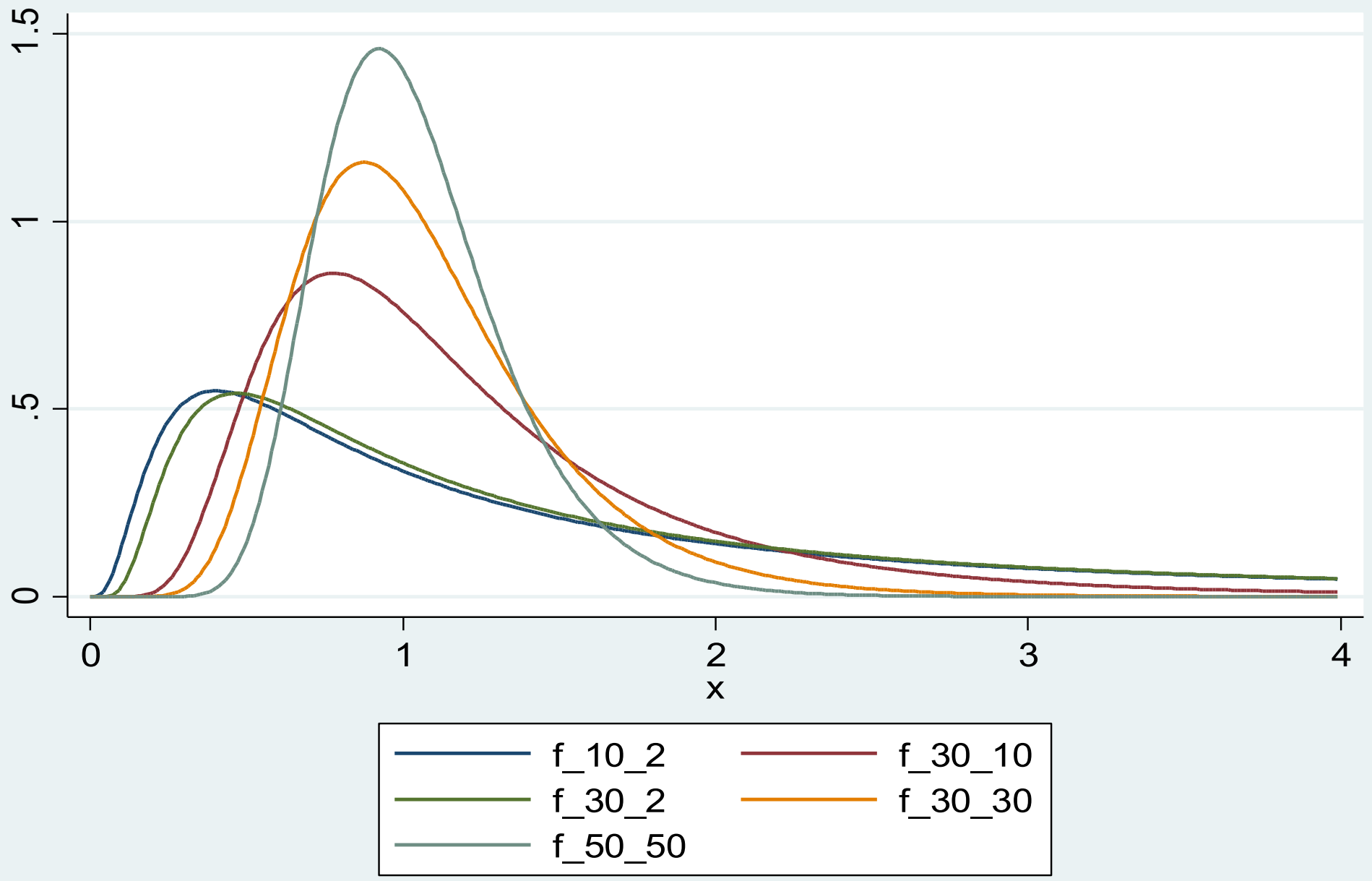


Рональд Эйлмер Фишер

Функция плотности Снедекора-Фишера для степеней свободы (2, 1), (3, 1), (2,10), (3,10).



Функция плотности Снедекора-Фишера
для степеней свободы (10, 2), (30, 2), (30,10), (30,30), (50, 50).



Проверка гипотезы о равенстве дисперсий

в двух нормальных генеральных совокупностях — введение

Пусть $X_1, \ldots, X_{n_X}, \quad X_i \sim N(\mu_X, \sigma^2);$

$Y_1, \ldots, Y_{n_Y}, \quad Y_i \sim N(\mu_Y, \sigma^2).$

Все эти величины независимы.

Тогда $\dfrac{\hat{\sigma}_X^2}{\hat{\sigma}_Y^2} \sim F_{n_X-1,\, n_Y-1}.$

Докажем: ◄ $\dfrac{\hat{\sigma}_X^2}{\hat{\sigma}_Y^2} = \dfrac{\dfrac{1}{n_X-1}\dfrac{(n_X-1)\hat{\sigma}_X^2}{\sigma^2}}{\dfrac{1}{n_Y-1}\dfrac{(n_Y-1)\hat{\sigma}_Y^2}{\sigma^2}} \sim \dfrac{\dfrac{1}{n_X-1}\chi^2_{n_X-1}}{\dfrac{1}{n_Y-1}\chi^2_{n_Y-1}}.$

Числитель и знаменатель независимы => $\dfrac{\hat{\sigma}_X^2}{\hat{\sigma}_Y^2} \sim F_{n_X-1,\, n_Y-1}.$ ►

Как это нам поможет?

Проверка гипотезы о равенстве дисперсий
в двух нормальных генеральных совокупностях — суть

Пусть $X_1, \ldots, X_{n_X}, \quad X_i \sim N(\mu_X, \sigma_X^2);$

$Y_1, \ldots, Y_{n_Y}, \quad Y_i \sim N(\mu_Y, \sigma_Y^2).$

Все эти величины независимы.

$H_0: \sigma^2_X = \sigma^2_Y$

Выбираем одну из трёх альтернатив:

$H_A: \sigma^2_X \neq \sigma^2_Y$

$H_A: \sigma^2_X > \sigma^2_Y$

$H_A: \sigma^2_X < \sigma^2_Y$

Критическая статистика: $F = \dfrac{\hat{\sigma}_X^2}{\hat{\sigma}_Y^2} \overset{H_0}{\sim} F_{n_X-1,\, n_Y-1}.$

Какие значения F-статистики согласуются, а какие противоречат основной гипотезе?

Критические области для F-статистики при уровне значимости α и p-значения:

$H_A: \sigma^2_X \neq \sigma^2_Y$ $\quad$ $отвергаем\ H_0\ при\ F > F_{n_X-1,n_Y-1,\frac{\alpha}{2}}\ или\ F < F_{n_X-1,n_Y-1,1-\frac{\alpha}{2}};$

$$p-значение = 2 \times min\left[P\left(W > F\right), P\left(W < F\right)\right].$$

$H_A: \sigma^2_X > \sigma^2_Y$ $\quad$ $отвергаем\ H_0\ при\ F > F_{n_X-1,n_Y-1,\alpha};$

$$p-значение = P\left(W > F\right).$$

$H_A: \sigma^2_X < \sigma^2_Y$ $\quad$ $отвергаем\ H_0\ при\ F < F_{n_X-1,n_Y-1,1-\alpha};$

$$p-значение = P\left(W < F\right).$$

Здесь $W \sim F_{n_X-1,n_Y-1}$, F — рассчитанное значение F-статистики.

Всё похоже на проверку гипотезы о дисперсии, только вместо хи-квадрат — F.

Полезное свойство *F*-распределения: $F_{n_X-1,n_Y-1,\alpha} = 1/F_{n_Y-1,n_X-1,1-\alpha}.$

почему так?

С помощью него можно рассчитывать левосторонние критические значения (часто в таблицах приводятся только правосторонние).

Возвращаемся к Васе и Ане

Оценённая дисперсия длины Васиного прыжка:

в присутствии Ани $\hat{\sigma}^2_X = 0.0907$,

в отсутствие Ани $\hat{\sigma}^2_Y = 0.1018$.

H_0: $\sigma^2_X = \sigma^2_Y$

H_A: $\sigma^2_X \neq \sigma^2_Y$

Возьмём уровень значимости 10%.

Статистика: $$F = \frac{\hat{\sigma}^2_X}{\hat{\sigma}^2_Y} = \frac{0.0907}{0.1018} = 0.8901.$$

Критические значения: $F_{7,9,\frac{0.1}{2}} = 3.29$; $F_{7,9,1-\frac{0.1}{2}} = 0.27$.

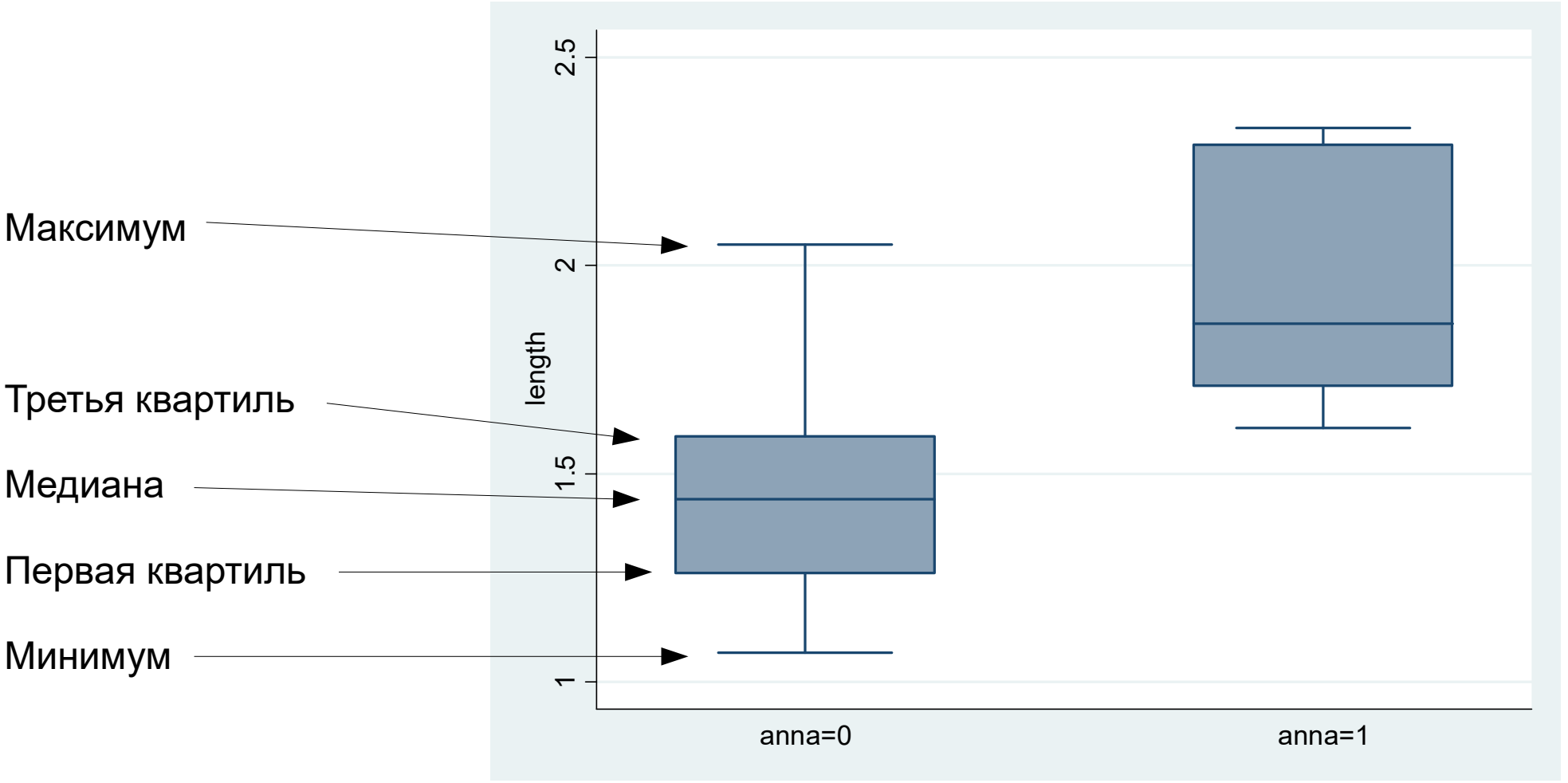
$0.27 < F < 3.29$ — в критическую область не попадаем.

Вывод: нет оснований отвергнуть основную гипотезу и считать, что дисперсия длины Васиных прыжков связана с присутствием Ани.

p-значение (рассчитано в Stata): 0.8983.

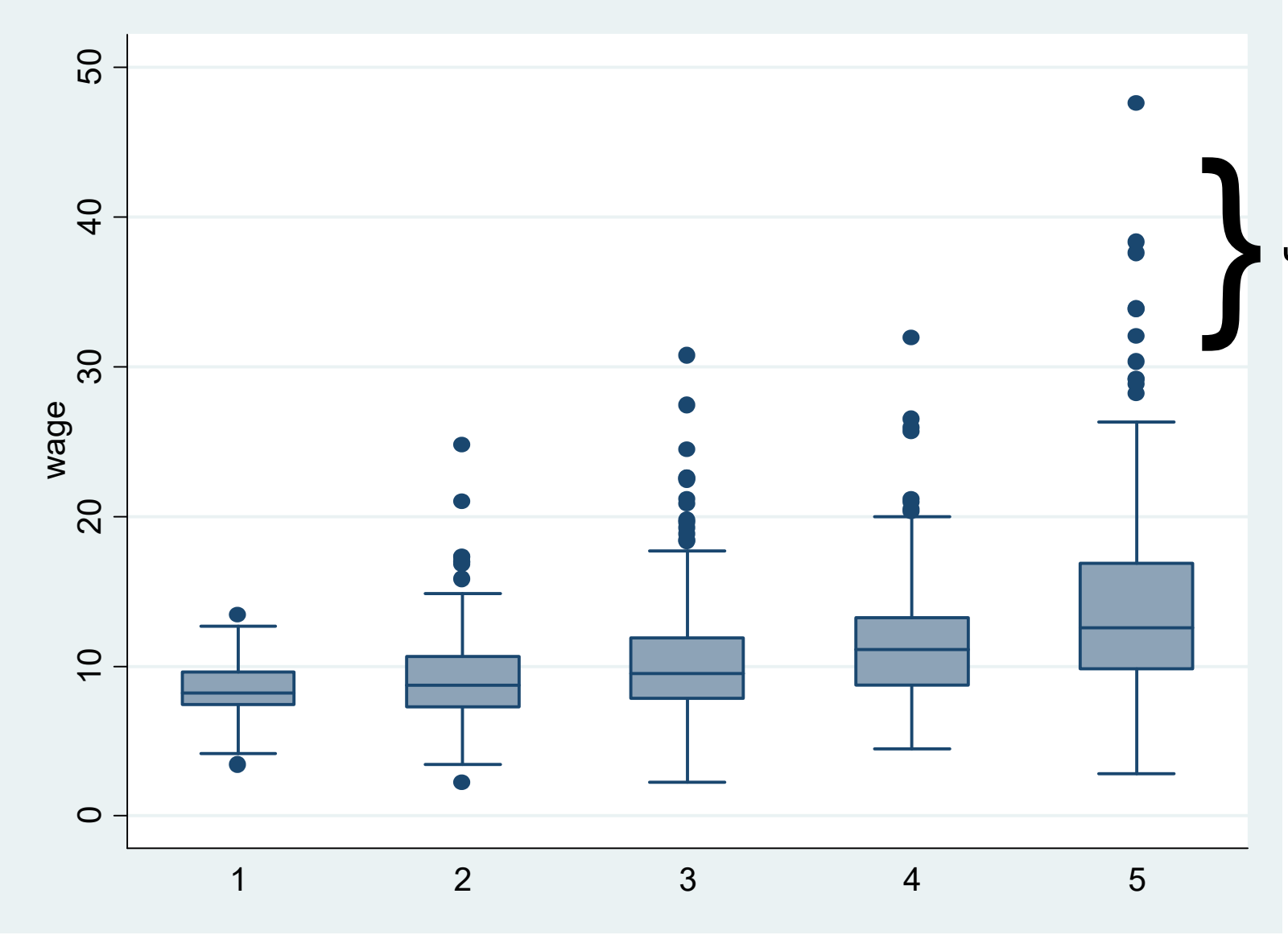
И ещё про визуализацию: графики «ящик-с-усами» (box-and-whiskers)

Очень полезная штука, позволяющая сопоставить как среднее, так и разброс в нескольких выборках. Так выглядят усатые ящики для данных про Васю и Аню:



Наблюдений маловато. На следующем слайде — пример для больших выборок.

Заработные платы в Бельгии 1994 года в зависимости от уровня образования (1 — низший, 5 — высший), франки в час:



Как сравнивать средние при разных дисперсиях,
почитайте сами у Ньюболда или в Devore-Berk.

Про сравнение долей тоже говорить не буду.

В следующий раз:

Анализ связанных пар и независимых выборок: непараметрические критерии.

(что делать, если распределение не нормальное, а выборки малы?)

Полезно повторить тему:
схема Бернулли, биномиальное распределение.