ТЕОРИЯ ВЕРОЯТНОСТЕЙ и МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Часть III. Математическая статистика

1) Доверительные интервалы: версии КБЛ **[1]** и ФаЛеб **[6]**

2) «$\overline{xy}$–исчисление»

3) Регрессия Y на X и X на Y

3-4) Ковариация и коэффициент корреляции в ТВ (в «достатистическую» эпоху)

4) Выборочный коэффициент корреляции. Проверка значимости коэффициента корреляции: а) с помощью $t_{кр}$; б) с помощью $r_{кр}$

5) Выборочный коэффициент ранговой корреляции Спирмена

6) Задача 4 из *Моего Любимого Билета*: некоторые полезные формулы

7) Задача 4: *Alter Ego*

ЛИТЕРАТУРА

1. Карасев В.А., Богданов С.Н., Лёвшина Г.Д. Теория вероятностей и математическая статистика: Разд. 2. Математическая статистика: Учеб.-метод. Пособие. – М. МИСиС, 2005. – 117 с. № 1855. [печ.]

2. Карасев В.А., Лёвшина Г.Д. Теория вероятностей и математическая статистика: математическая статистика: практикум». – М. Изд. Дом МИСиС, 2016. № 2770. [электрон.]

3. Данченков И.В., Карасев В.А. Математическая статистика: проверка гипотезы о виде закона распределения: практикум. – М. : Изд. Дом НИТУ «МИСиС», 2017. – 54 с. № 2976

4. Гмурман В.Е. Теория вероятностей и математическая статистика: учебное пособие для вузов. – М.: Изд-во Юрайт, 2015. Работаем по: М.: Высшее образование, 2006. – 479 с.

5. Гмурман В.Е. Руководство к решению задач по теории вероятностей и математической статистике: учебное пособие для вузов. – М.: Изд-во Юрайт, 2015.

6. Лебедев А.В., Фадеева Л.Н. Теория вероятностей и математическая статистика. – М., 2018. – 480 с. [электрон.], Фадеева Л. Н., Лебедев А.В. 2011 [печ.]

7. Ефимов А.В., Поспелов А.С. и др. Сборник задач по математике для втузов. Специальные курсы (ТВ. МС. МО. УрЧП). – М., 1984. – 608 с. [печ.]; В 4 ч. – Ч. 4 (ТВ. МС). – М., 2003. – 432 с. [электрон., печ.]

8. Фёрстер Э., Рёнц Б. Методы корреляционного и регрессионного анализа. – М., 1983. – 304 с.

1) Доверительные интервалы: версии КБЛ [1] и ФаЛеб [6]

В **[6]** (изд. 2011 г., с. 293-294) строятся *доверительные интервалы* для параметров сл. в. $\xi \in N(a, \sigma^2)$ (это было в лекции V, с. 58). Мы хотим рассмотреть случай

2) Для неизвестного среднего a при неизвестной дисперсии σ^2:

$$\bar{x} - \frac{s}{\sqrt{n}}\hat{t}_\gamma < a < \bar{x} + \frac{s}{\sqrt{n}}\hat{t}_\gamma, \qquad \textbf{(*)}$$

где $\hat{t}_\gamma$ – критическая точка распределения Стьюдента (для двусторонней области) с $n-1$ степенью свободы и уровнем значимости $\alpha = 1-\gamma$. (s^2 – исправленная выборочная дисперсия, построенная по случайной выборке объема n из нормальной генеральной совокупности $N(a, \sigma^2)$.)

Единственное отличие от **[6]** – добавление значка ^ над t, выделяющего *двустороннюю* квантиль: действительно, при обосновании данного случая в **[6]** отмечено, что критическая точка определяется так, чтобы

$$P(|T| \geq t_{\text{кр}}) = \alpha = 1-\gamma.$$

Это равенство вместе с $P(|T| \geq t_{\text{кр}}) = 1 - P(|T| < t_{\text{кр}})$ приводит к соотношению

$$P(|T| < t_{\text{кр}}) = \gamma,$$

означающему, что $t_{\text{кр}} = \hat{t}_\gamma$ – *двусторонняя* γ–квантиль распределения Стьюдента.

В **[1]** (с. 21–22) тот же случай изложен по-другому. Чтобы не отвлекаться на детали, запишем **[1]**–аналог **[6]**–неравенств **(*)**:

$$\bar{x} - \frac{s}{\sqrt{n}} t_{1-\frac{\alpha}{2}} < a < \bar{x} + \frac{s}{\sqrt{n}} t_{1-\frac{\alpha}{2}}, \qquad \textbf{(**)}$$

где $t_{1-\frac{\alpha}{2}} = t_{1-\frac{\alpha}{2}}(n-1)$ – квантиль распределения Стьюдента с $n-1$ степенями свободы. Здесь важно то, что данная квантиль – *односторонняя*.

Итак, требуется доказать, что неравенства **(*)** и **(**)** эквивалентны, т.е. доказать равенство

$$\underset{\text{одност}}{t_{1-\alpha/2}} = \underset{\text{двуст}}{\hat{t}_{\gamma}}$$

Имеем $P(T < t_{1-\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2}$, $P(|T| < \hat{t}_{\gamma}) = \gamma$. Тогда

$$\gamma = P(-\hat{t}_{\gamma} < T < \hat{t}_{\gamma}) = F(\hat{t}_{\gamma}) - F(-\hat{t}_{\gamma}) = F(\hat{t}_{\gamma}) - (1 - F(\hat{t}_{\gamma})) = 2F(\hat{t}_{\gamma}) - 1 \Rightarrow$$

$$\Rightarrow F(\hat{t}_{\gamma}) = \frac{1+\gamma}{2} = \frac{1+1-\alpha}{2} = 1 - \frac{\alpha}{2}$$

$$\Downarrow$$

$$F(t_{1-\frac{\alpha}{2}}) = P(T < t_{1-\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2} \quad \Rightarrow \quad \hat{t}_{\gamma} = t_{1-\frac{\alpha}{2}},$$

что и требовалось…

2) «$\overline{xy}$–исчисление»

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i,\quad \bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i,\quad \overline{x^2}=\frac{1}{n}\sum_{i=1}^{n}x_i^2,\quad \overline{y^2}=\frac{1}{n}\sum_{i=1}^{n}y_i^2,\quad \overline{xy}=\frac{1}{n}\sum_{i=1}^{n}x_iy_i,$$

$$\sum(y_i-\bar{y})(x_i-\bar{x})=\sum x_iy_i-n\bar{x}\,\bar{y}=n(\overline{xy}-\bar{x}\,\bar{y}),$$

$$\sum(x_i-\bar{x})^2=\sum x_i^2-n\bar{x}=n(\overline{x^2}-\bar{x}^2),\quad \sum(y_i-\bar{y})^2=\sum y_i^2-n\bar{y}=n(\overline{y^2}-\bar{y}^2)$$

$\sigma_x^2=\frac{1}{n}\sum(x_i-\bar{x})^2=\overline{x^2}-\bar{x}^2=\mathrm{Var}(x)_{\text{неиспр}}$ –неисправленная выборочная дисперсия значений x признака X,

$\sigma_y^2=\frac{1}{n}\sum(y_i-\bar{y})^2=\overline{y^2}-\bar{y}^2=\mathrm{Var}(y)_{\text{неиспр}}$ –неисправленная выборочная дисперсия значений y признака Y,

$\sigma_{xy}=\frac{1}{n}\sum(y_i-\bar{y})(x_i-\bar{x})=\overline{xy}-\bar{x}\,\bar{y}=\mathrm{Cov}(x,y)_{\text{неиспр}}$ – неисправленная выборочная ковариация…,

$s_x^2=\frac{1}{n-1}\sum(x_i-\bar{x})^2=\frac{n}{n-1}(\overline{x^2}-\bar{x}^2)=\frac{n}{n-1}\sigma_x^2$ – исправленная выборочная дисперсия…,

$s_y^2=\frac{1}{n-1}\sum(y_i-\bar{y})^2=\frac{n}{n-1}(\overline{y^2}-\bar{y}^2)=\frac{n}{n-1}\sigma_y^2$ – исправленная выборочная дисперсия…,

$s_{xy}=\frac{1}{n-1}\sum(y_i-\bar{y})(x_i-\bar{x})=\frac{n}{n-1}(\overline{xy}-\bar{x}\,\bar{y})=\frac{n}{n-1}\sigma_{xy}$ – исправленная выборочная ковариация….

3) Регрессия Y на X и X на Y

$\rho_{yx} = \hat{b} = \dfrac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2} = \dfrac{\mathrm{Cov}(x,y)_{\text{неиспр}}}{\mathrm{Var}(x)_{\text{неиспр}}} = \dfrac{\sigma_{xy}}{\sigma_x^2} = \dfrac{s_{xy}}{s_x^2}$ – выборочный коэффициент регрессии Y на X ,

$\hat{a} = \overline{y} - \hat{b}\overline{x}$ – свободный член уравнения регрессии Y на X ,

$\hat{y} = \hat{a} + \hat{b}x = \overline{y} + \hat{b}(x - \overline{x})$ – уравнение прямой линии регрессии Y на X ,

$\rho_{xy} = \hat{\beta} = \dfrac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{y^2} - \overline{y}^2} = \dfrac{\mathrm{Cov}(x,y)_{\text{неиспр}}}{\mathrm{Var}(y)_{\text{неиспр}}} = \dfrac{\sigma_{xy}}{\sigma_y^2} = \dfrac{s_{xy}}{s_y^2}$ – выборочный коэффициент регрессии X на Y,

$\hat{\alpha} = \overline{x} - \hat{\beta}\overline{y}$ – свободный член уравнения регрессии X на Y,

$\hat{x} = \hat{\alpha} + \hat{\beta}y = \overline{x} + \hat{\beta}(y - \overline{y})$ – уравнение прямой линии регрессии X на Y.

3-4) Ковариация и коэффициент корреляции в ТВ ([ФаЛеб], с. 97-98, 99-100)

В рамках МС эти объекты иногда называются теоретическими (см., напр., **[ФаЛеб]**, с. 393).

5.7. Ковариация. Коэффициент корреляции

Ковариацией случайных величин ξ и η называется число

$$\operatorname{cov}(\xi, \eta) = M[(\xi - M\xi)(\eta - M\eta)]$$

(в предположении существования всех математических ожиданий).

Из определения ковариации вытекают следующие ее свойства:

1. Ковариацию можно вычислить по формуле

$$\operatorname{cov}(\xi, \eta) = M(\xi\eta) - (M\xi)(M\eta).$$

2. Если ξ и η — независимые случайные величины, то $\operatorname{cov}(\xi, \eta) = 0$.

3. $\operatorname{cov}(\xi, \eta) = \operatorname{cov}(\eta, \xi)$.

4. $\operatorname{cov}(C\xi, \eta) = C\operatorname{cov}(\xi, \eta)$; $\operatorname{cov}(\xi, C\eta) = C\operatorname{cov}(\xi, \eta)$.

5. $\operatorname{cov}(\xi_1 + \xi_2, \eta) = \operatorname{cov}(\xi_1, \eta) + \operatorname{cov}(\xi_2, \eta)$; $\operatorname{cov}(\xi, \eta_1 + \eta_2) = \operatorname{cov}(\xi, \eta_1) + \operatorname{cov}(\xi, \eta_2)$.

6. $\operatorname{cov}(\xi, \xi) = D\xi$.

5. $\operatorname{cov}(\xi_1 + \xi_2, \eta) = \operatorname{cov}(\xi_1, \eta) + \operatorname{cov}(\xi_2, \eta)$; $\operatorname{cov}(\xi, \eta_1 + \eta_2) = \operatorname{cov}(\xi, \eta_1) + \operatorname{cov}(\xi, \eta_2)$.

6. $\operatorname{cov}(\xi, \xi) = D\xi$.

7. Если случайные величины ξ и η имеют конечные дисперсии $D\xi$ и $D\eta$, то дисперсия суммы этих случайных величин существует и равна

$$D(\xi + \eta) = D\xi + D\eta + 2\operatorname{cov}(\xi, \eta).$$

8. $-\sqrt{D\xi D\eta} \leq \operatorname{cov}(\xi, \eta) \leq \sqrt{D\xi D\eta}$.

9. Равенство $\operatorname{cov}(\xi_1, \xi_2) = \pm\sqrt{D\xi_1 D\xi_2}$ достигается тогда и только тогда, когда случайные величины ξ_1 и ξ_2 линейно зависимы.

Итак, ковариацию можно считать мерой зависимости случайных величин, так как для независимых случайных величин ковариация равна нулю. Существенным недостатком ковариации является то, что ее размерность совпадает с произведением размерностей случайных величин. Естественно, желательно иметь безразмерную характеристику зависимости. Таковой является коэффициент корреляции.

Теоретический коэффициент корреляции

Коэффициентом корреляции случайных величин ξ и η (с положительными дисперсиями) называется число

$$\rho = \frac{\operatorname{cov}(\xi, \eta)}{\sqrt{D\xi}\sqrt{D\eta}}.$$

Коэффициент корреляции является одной из важных мер зависимости случайных величин. Как следует из свойств ковариации, он принимает значения от -1 до $+1$, отражая как силу зависимости (по абсолютной величине), так и характер (положительная или отрицательная).

Чем ближе $|\rho|$ к единице, тем с большим основанием можно считать, что ξ и η находятся в линейной зависимости, т.е. коэффициент корреляции характеризует не всякую зависимость, а только так называемую *линейную вероятностную зависимость*, которая заключается в том, что при возрастании одной случайной величины другая имеет тенденцию изменяться по линейному закону.

Можно сказать, что коэффициент корреляции ρ отражает степень линейной зависимости случайных величин. С возрастанием ξ случайная величина η имеет тенденцию к увеличению при $\rho > 0$ и к уменьшению при $\rho < 0$. Поэтому при $\rho > 0$ говорят о *положительной корреляционной зависимости* ξ и η, при $\rho < 0$ — об *отрицательной.*

Для независимых случайных величин коэффициент корреляции равен нулю. Если $\rho = 0$, то случайные величины называются *некоррелированными.* Из независимости случайных величин следует их некоррелированность, но наоборот — не всегда.

Для линейно зависимых величин, т.е. в случае $\eta = a\xi + b$, где a и b — константы, коэффициент корреляции равен $+1$ при $a > 0$ и -1 при $a < 0$.

4) Выборочный коэффициент корреляции.

Определение. Выборочным (эмпирическим) коэффициентом корреляции для выборки вида (x_1, y_1), $(x_2, y_2), \ldots, (x_n, y_n)$ называется выборочная характеристика

$$r = r_{\text{в}} = r_{xy} = \frac{s_{xy}}{s_x s_y} = \rho_{yx}\frac{s_x}{s_y} = \rho_{xy}\frac{s_y}{s_x} = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\sqrt{\overline{x^2} - \bar{x}^2}\sqrt{\overline{y^2} - \bar{y}^2}} = \text{и т.д.}$$

Немного скрининга (**[ФаЛеб]**, с. 392-393):

18.3. Выборочные коэффициенты корреляции

Основными характеристиками, описывающими степень связи между составляющими X и Y двумерной случайной величины (X, Y) являются ковариация $\operatorname{cov}(X, Y) = \mu_{xy} = M(X - a_x)(Y - a_y)$ и коэффициент корреляции $\rho_{xy} = \dfrac{\mu_{xy}}{\sigma_x \sigma_y}$, который является мерой линейной зависимости между X и Y.

Так как в случае статистических данных закон распределения двумерной случайной величины неизвестен, для оценки тесноты связи применяется эмпирическая (или выборочная) ковариация $\hat{\mu}_{xy} = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ и эмпирический (или выборочный) коэффициент корреляции.

Выборочным коэффициентом корреляции для выборки вида (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) называется выборочная характеристика

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}.$$

Выборочный коэффициент корреляции может принимать значения от -1 до $+1$. Он сходится к теоретическому коэффициенту корреляции соответствующих случайных величин, если тот существует. По абсолютной величине и знаку коэффициента можно судить о степени зависимости (сильная или слабая) и характере (положительная или отрицательная).

Проверка гипотезы о существовании линейной зависимости ≈ ≈ Проверка значимости коэффициента корреляции

а) с помощью $t_{кр}$

Выборочный коэффициент корреляции обычно используется в предположении нормального закона распределения данных (нормальность данных). Как известно, в этом случае из равенства нулю теоретического коэффициента ρ следует независимость случайных величин (в более общем случае это неверно). В случае нормального распределения можно проверить гипотезу $\rho = 0$. Пусть

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}.$$

Если гипотеза $\rho = 0$ верна, то T имеет распределение Стьюдента с $n - 2$ степенями свободы. При уровне значимости α выберем критическую точку $t_{кр} = t_{кр}(\alpha; n - 2)$ для двусторонней области. Если $|T| < t_{кр}$, гипотеза $\rho = 0$ принимается, иначе — отвергается.

Теперь **[8]**, с. 194–195:

Проверяя значимость коэффициента парной корреляции, устанавливают наличие или отсутствие корреляционной связи между исследуемыми явлениями. При отсутствии связи коэффициент корреляции генеральной совокупности равен нулю (ρ = 0). Процедура поверки начинается с формулировки нулевой и альтернативной гипотез:

H_0: различие между выборочным коэффициентом корреляции *r* и ρ = 0 незначимо,

H_1: различие между *r* и ρ = 0 значимо, и, следовательно, между переменными *y* и *x* (величинами *Y* и *X*) имеется существенная связь <позволяющая говорить о линейной зависимости>. Из альтернативной гипотезы следует, что нужно воспользоваться двусторонней критической областью.

Вычисленная по результатам выборки статистика

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

сравнивается с критическим значением, определяемым по таблице распределения Стьюдента при заданном уровне значимости **α** и $\boldsymbol{k = n - 2}$ степенях свободы.

Правило применения критерия.

Если $|T| > t_{кр} = t_{кр}(\alpha; n-2)$, то гипотеза **$H_0$** на уровне значимости α отвергается, т.е. связь между переменными значима. Согласно Гмурману **[4]**, с. 327, это означает, что выборочный коэффициент корреляции значимо отличается от нуля (кратко говоря, значим), а *X* и *Y* коррелированы, т.е. связаны линейной зависимостью.

Если $|T| \le t_{кр}$, то гипотеза **H_0** на уровне значимости α принимается.

Следует отметить, что в одних пособиях осторожно говорится о наличии связи и значимости в противовес стохастической независимости, в других категорично противопоставляются линейные зависимость и независимость. Пример такого противопоставления – **[1]**, с. 69, где рассматриваются две гипотезы (с уровнем значимости $\alpha = 1 - P = 1 - \gamma$):

H_0: линейная зависимость между величинами X и Y не существует ($\rho = 0$ принадлежит доверительному интервалу для истинного коэффициента корреляции),

H_1: линейная зависимость между величинами X и Y существует ($\rho = 0$ не принадлежит найденному доверительному интервалу).

Примиряет обе точки зрения – осторожную и категоричную – доверительный интервал для ρ, но этот интервал в данном случае трудно вычислять. На этот случай были созданы **номограммы** (см. ниже на след. странице), из которых видно, что $\rho = 0$ не принадлежит доверительному интервалу для ρ_{xy} только при значениях *r*, по абсолютной величине близких к 1.

Номограммы для коэффициента корреляции

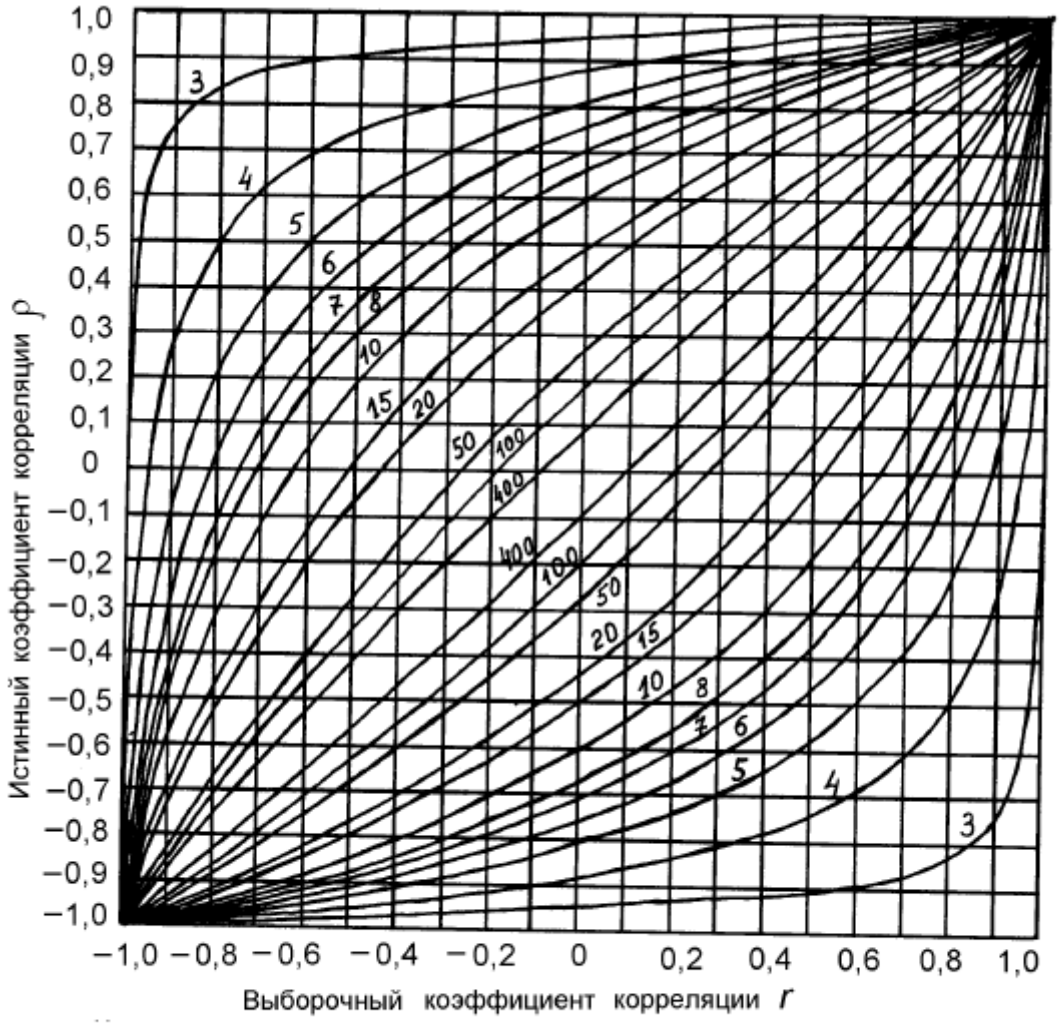


Рис. П1. Номограммы для нахождения доверительного интервала коэффициента корреляции при доверительной вероятности $P = 0{,}95$

Трудности использования очевидны…

Приближенный доверительный интервал для коэффициента корреляции

Вначале – несколько фрагментов из **[ФаЛеб]**-2011, с. 302-303:

14.4. Интервальная оценка коэффициента корреляции

Пусть (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) — независимые наблюдения над двумерной нормальной случайной величиной. Построим асимптотический доверительный интервал для коэффициента корреляции ρ, соответствующий надежности γ.

Приближенный доверительный интервал для ρ имеет вид

$$\text{th}(\text{arth}\,\hat{\rho}_n - \frac{1}{\sqrt{n-3}}u_\gamma) < \rho < \text{th}(\text{arth}\,\hat{\rho}_n + \frac{1}{\sqrt{n-3}}u_\gamma),$$

где

$$\hat{\rho}_n = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]\left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2\right]}}.$$

– выборочный коэффициент корреляции, а u_γ определяется из $\Phi_0(u_\gamma) = \gamma / 2$.

Найденный приближенный доверительный интервал настолько мало отличается от истинного, что может применяться уже для выборок объема $n \geq 10$.

Вернемся к критерию значимости коэффициента корреляции с помощью $t_{кр}$. Запишем

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \qquad (*_T)$$

и напомним

Правило применения критерия.

Если $|T| > t_{кр} = t_{кр}(\alpha; n-2)$, то гипотеза **$H_0$** на уровне значимости α **отвергается**, т.е. связь между переменными значима.

Если $|T| \leq t_{кр}$, то гипотеза **H_0** на уровне значимости α **принимается**.

Проверка гипотезы о существовании линейной зависимости ≈ ≈ Проверка значимости коэффициента корреляции

б) с помощью $r_{кр}$

Пусть $r = r_{xy}$ – выборочный коэффициент корреляции между величинами X и Y и

$$r_{кр} = \frac{t_{кр}}{\sqrt{t_{кр}^2 + n - 2}}, \qquad (*_{кр})$$

где $t_{кр} = \hat{t}_{\gamma=1-\alpha}(n-2) = t_{1-\frac{\alpha}{2}}(n-2)$ (см. первую тему настоящей лекции). Данная связь позволяет составить таблицу значений $r_{кр}$ по таблице квантилей $t_{кр}$ распределения Стьюдента.

Правило применения $r_{кр}$ –критерия (см., напр., **[8]** (с. 195), **[1]**, (с. 69).

Если $|r| > r_{кр}$, то гипотеза $\mathbf{H_0}$ на уровне значимости α **отвергается** (= связь между переменными значима = связь между переменными существенна = линейная зависимость между величинами существует).

Если $|r| < r_{кр}$, то гипотеза $\mathbf{H_0}$ на уровне значимости α **принимается** (= результаты наблюдений считаем непротиворечащими гипотезе об отсутствии связи = линейная зависимость между величинами не существует).

Связь между $t_{кр}$ –критерием и $r_{кр}$ –критерием основана на импликации

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \quad \Leftrightarrow \quad r = \frac{T}{\sqrt{T^2 + n - 2}}$$ (кстати, **Упр.**).

Пример из [8] на применение $t_{кр}$– и $r_{кр}$–критериев

Проверим гипотезу о независимости производительности труда от уровня механизации работ при α = 0,05 на n = 14 народных предприятиях ГДР конца 1970-х гг. По данным, приведенным в рабочей таблице, было вычислено, что $r = r_{xy} = 0{,}9687$. По **(*$_T$)** получаем

$$T = \frac{0{,}9687\sqrt{14-2}}{\sqrt{1-0{,}9687^2}} = 13{,}52.$$

По таблице распределения Стьюдента для α = 0,05 и f (= k) =12 находим критическое значение этой статистики: $t_{12;\,0{,}05}$ = 2,179. Поскольку $T > t_{12;\,0{,}05}$, нулевую гипотезу отвергаем, допуская ошибку лишь в 5% случаев.

Мы получим тот же результат, если будем сравнивать $r = r_{xy} = 0{,}9687$ с критическим значением коэффициента корреляции $r_{кр} = r_{12;\,0{,}05}$ = 0,5324, *найденным по соответствующей таблице* при α = 0,05 и f (= k) =12. «Соответствующая таблица» может быть построена по таблице распределения Стьюдента с использованием **(*$_{кр}$)**. В **[8]** такая таблица построена (таблица 6 на с. 289).

5) Выборочный коэффициент ранговой корреляции Спирмена

По **[ФаЛеб]**-2011, с. 393-394:

Наблюдения всегда можно упорядочить по возрастанию какой-либо переменной (x или y). **Рангом наблюдения** называется его номер в таком ряду. Если какое-то значение переменной встречается несколько раз, ему приписывается *средний* ранг. Обозначим ранги наблюдений по возрастанию x и y через r_i и s_i соответственно. Пусть

$$S = \sum_{i=1}^{n} (r_i - s_i)^2 .$$

Выборочным коэффициентом ранговой корреляции Спирмена называется величина

$$r_S = 1 - \frac{6S}{n^3 - n} .$$

Этот коэффициент также может принимать значения от -1 до $+1$. Аналогичным образом он отражает силу и характер зависимости между величинами. Для проверки гипотезы о независимости случайных величин существуют специальные таблицы критических точек. Однако при больших n можно проверять гипотезу так же, как для обычного выборочного коэффициента корреляции.

Заметим, что с помощью коэффициента Спирмена можно анализировать и ситуации, когда некоторый признак объекта («качество», «привлекательность» и т.п.) нельзя строго выразить численно, но можно упорядочить объекты по его возрастанию или убыванию, т.е. **проранжировать** их.

По **[8]**, с. 162-163:

Пример

Определим тесноту связи между производительностью труда и уровнем механизации работ на 10 промышленных предприятиях. Данные приведены в табл. 12.

Таблица 12

Производительность труда и уровень механизации работ на 10 предприятиях

Предприятие	Средняя выработка продукции в единицу рабочего времени, изд./ч	Коэффициент механизации работ, %	Ранги значений переменных		Разности рангов	
i	y_i	x_i	w_i	v_i	$(v_i - w_i)$	$(v_i - w_i)^2$
1	127	43	1	4	+3	9
2	120	51	2	1	−1	1
3	125	55	3	2	−1	1
4	126	57	4	3	−1	1
5	133	60	5	7	+2	4
6	129	62	6	5	−1	1
7	132	65	7	6	−1	1
8	135	68	8	8,5	+0,5	0,25
9	135	70	9	8,5	−0,5	0,25
10	140	74	10	10	0	0
Сумма	1 302	605	55	55	0	18,5

Например, ранг $v_5 = 7$ означает, что предприятие 5 по уровню механизации работ стоит на седьмом месте при расположении предприятий в порядке возрастания соответствующего показателя. По данным табл. 12 вычисляем коэффициент ранговой корреляции:

$$r_s = 1 - \frac{6 \cdot 18{,}5}{10(10^2 - 1)} = 0{,}888.$$

Величина r_s свидетельствует о тесной положительной связи между производительностью труда и уровнем механизации работ. Коэффициент парной корреляции, вычисленный непосредственно по исходным данным, равен: $r_{yx} = 0{,}833$. Сравнивая r_s и r_{yx}, убеждаемся, что они мало отличаются друг от друга. Коэффициент ранговой корреляции в общем служит довольно хорошей характеристикой степени связи исследуемых переменных. Его достоинство заключается в том, что он не связан с предпосылкой нормальности распределения исходных данных. Но не следует упускать из вида, что при переходе от первоначальных значений к рангам происходит определенная потеря информации. Коэффициент ранговой корреляции тем больше приближается к коэффициенту парной корреляции, чем меньше корреляционная связь между изучаемыми переменными отлична от линейной и чем сильнее эта связь. Для нормально распределенной генеральной совокупности и при достаточно большом объеме выборки ($n \geqslant 30$) между обоими коэффициентами существует следующее асимптотическое соотношение:

$$r_{yx} = 2 \sin\left(\frac{\pi}{6} r_s\right). \qquad (7.4)$$

Метод ранговой корреляции не требует линейной корреляции между переменными. Но, однако, необходимо, чтобы функция регрессии, отражающая эту связь, была монотонной.

Особенно полезной оказывается ранговая корреляция при исследовании связей между явлениями, не поддающимися количественной оценке. В таких случаях исследователь на основе своего опыта, или производя сравнение с каким-либо эталоном, приписывает элементам выборки ранги по каждому из изучаемых качественных признаков. Например, ранговую корреляцию можно использовать при исследовании зависимости между сортностью продукции, ее сроком службы и производственными затратами. При изучении качества изделий их часто классифицируют по следующим уровням: «отличное, очень хорошее, хорошее, среднее, плохое». Аналогично можно прошкалировать и другие признаки.

Упр. Проверить, что $r_S = 0{,}888$ и $r_{xy} = 0{,}833$.

6) Задача 4 из *Моего Любимого Билета*: некоторые полезные формулы

Если успеем – на доске

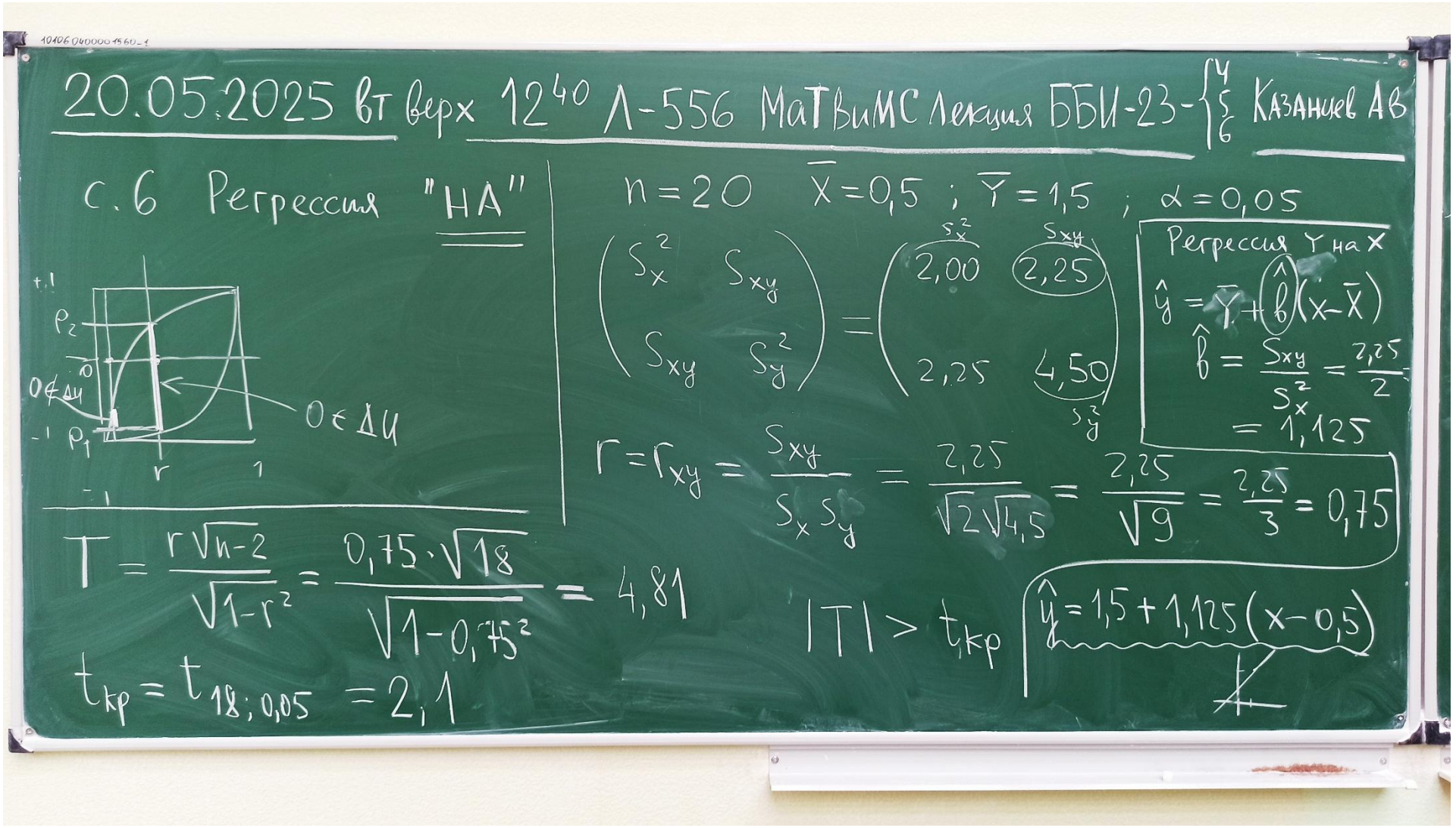
7) Задача 4: *Alter Ego*

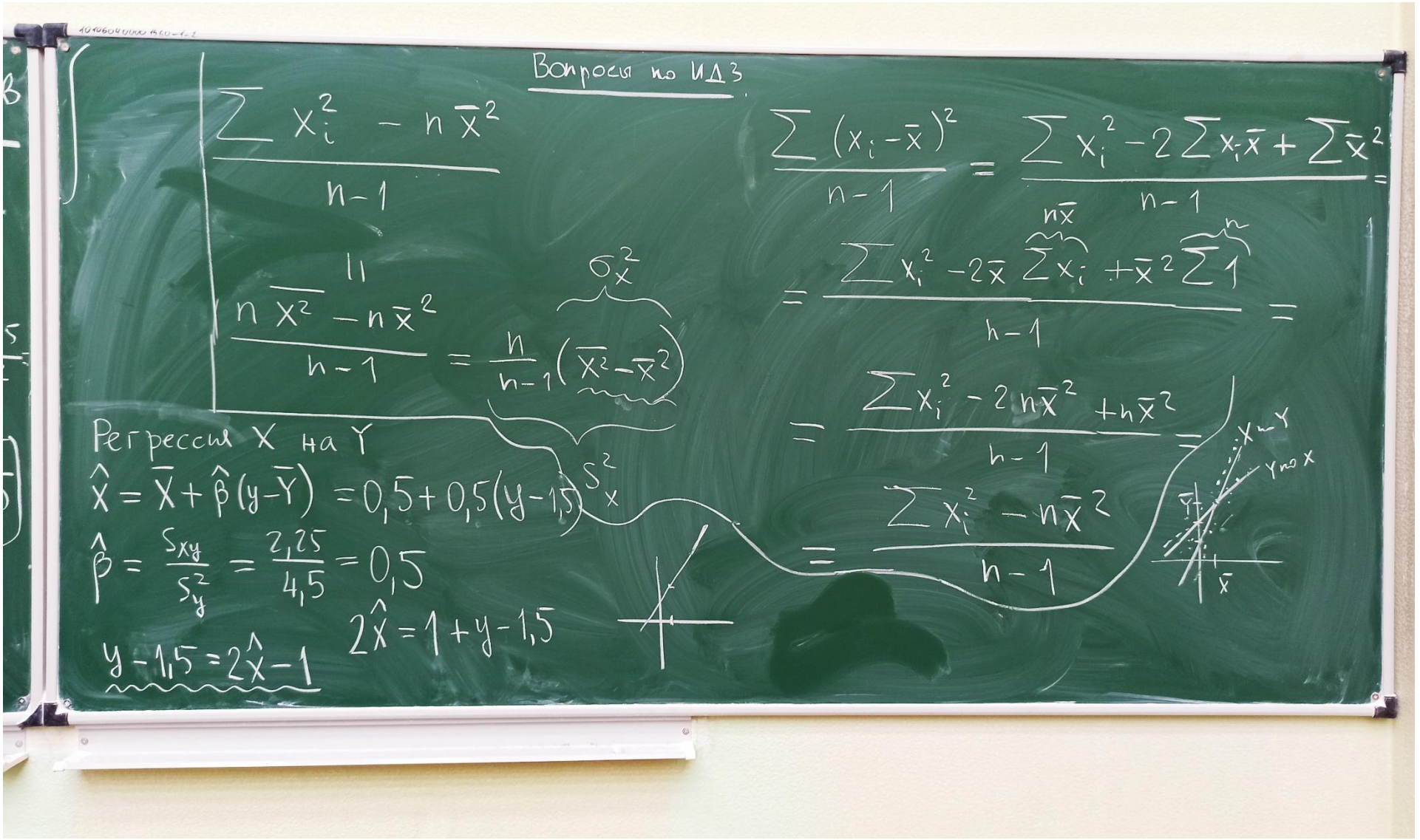
Задача. При изучении зависимости величин X и Y по серии из $n = 20$ измерений получена эмпирическая матрица ковариаций $\begin{pmatrix} 2,00 & 2,25 \\ 2,25 & 4,50 \end{pmatrix}$; $\overline{X} = 0,5$; $\overline{Y} = 1,5$. Найти выборочный коэффициент корреляции. Проверить гипотезу о существовании линейной зависимости между X и Y с уровнем значимости $\alpha = 0,05$. Написать уравнения прямых регрессии (Y на X и X на Y), построить их графики.

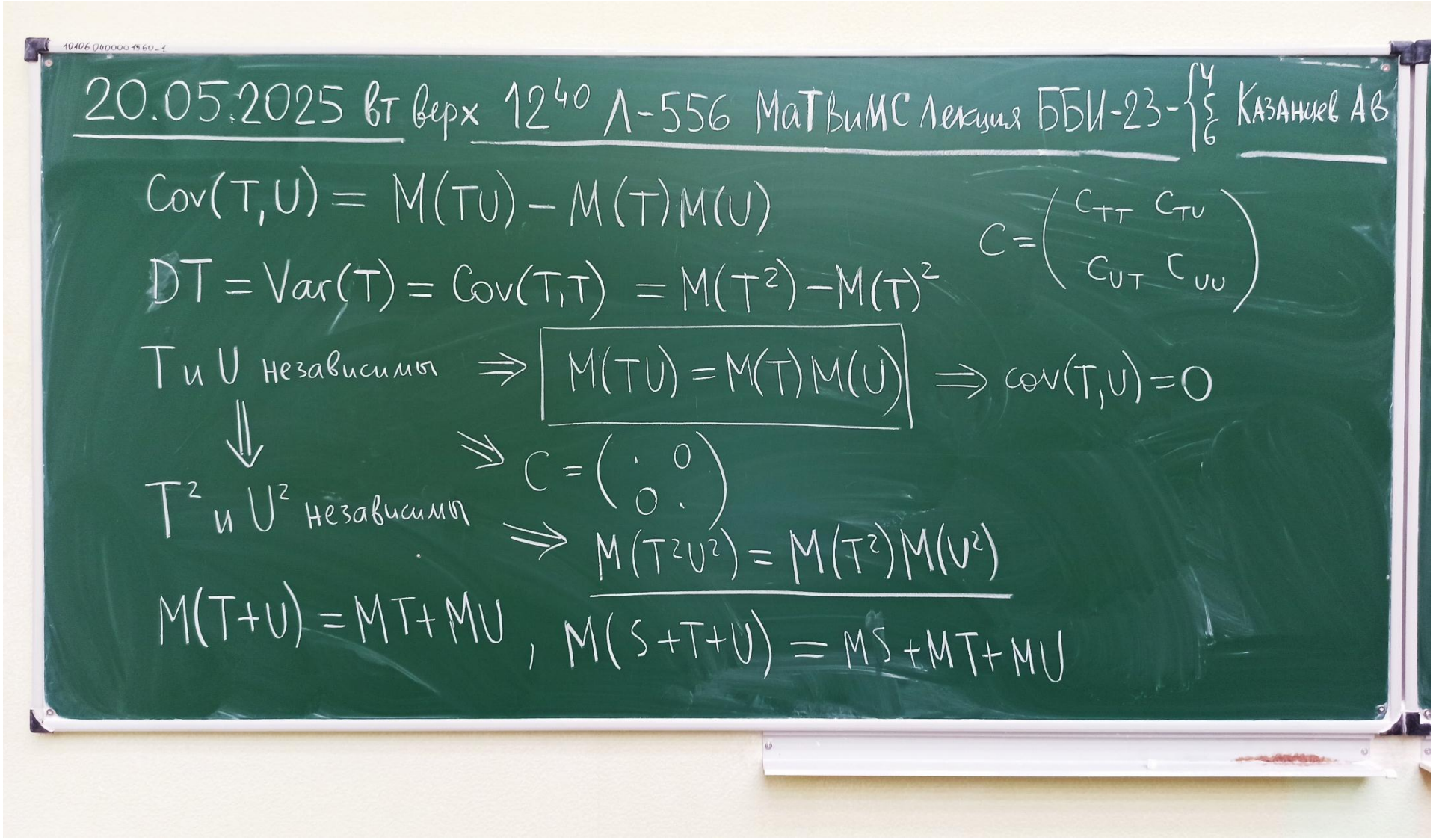
Она же:

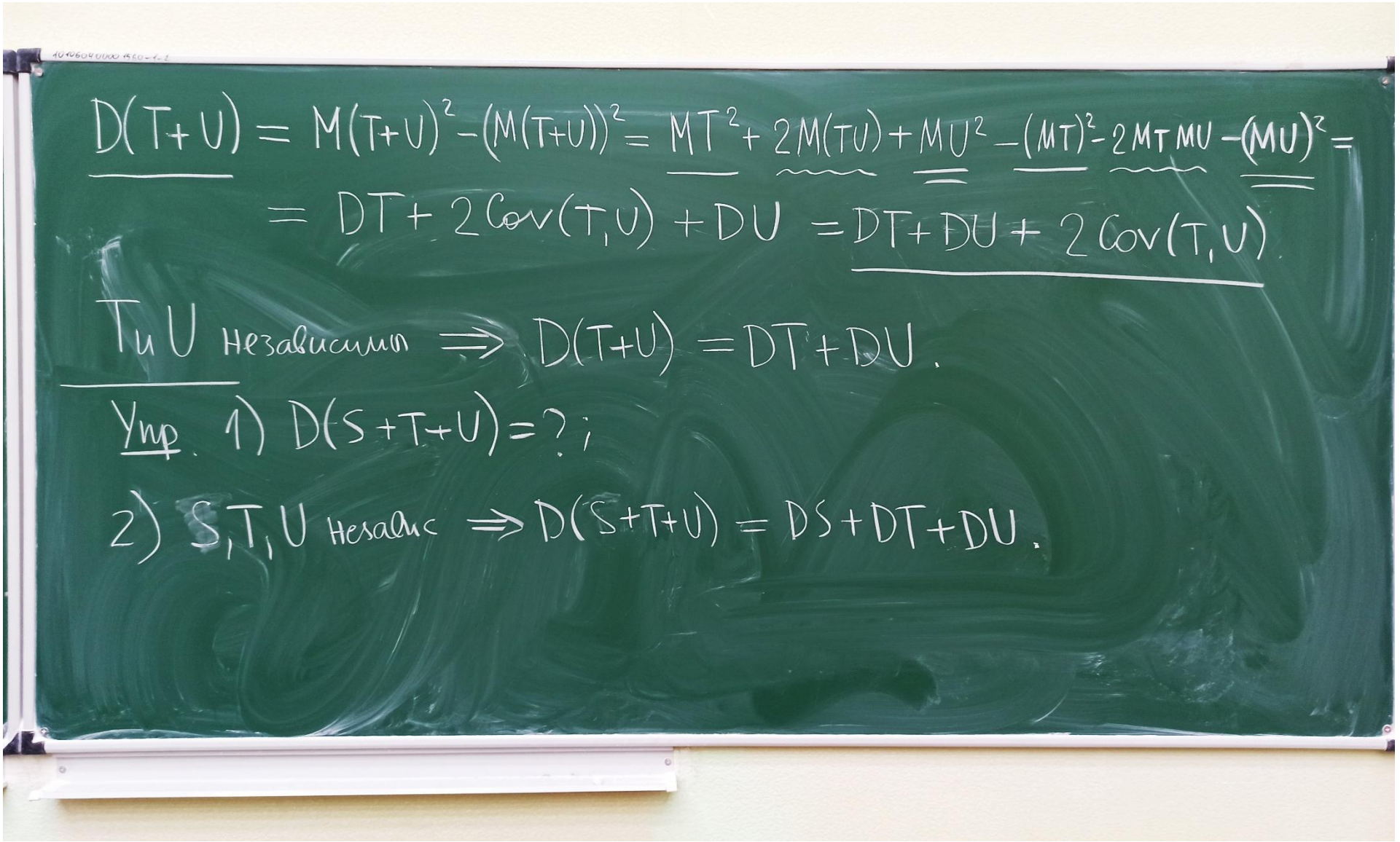
Задача. При изучении зависимости величин X и Y по серии из $n = 20$ измерений получена эмпирическая матрица ковариаций $\begin{pmatrix} 2,00 & 2,25 \\ 2,25 & 4,50 \end{pmatrix}$; $\overline{X} = 0,5$; $\overline{Y} = 1,5$. Найти выборочный коэффициент корреляции и проверить гипотезу о его значимости с уровнем $\alpha = 0,05$. Написать уравнения прямых регрессии (Y на X и X на Y), построить их графики.

ФОТО НА ЛЕКЦИИ

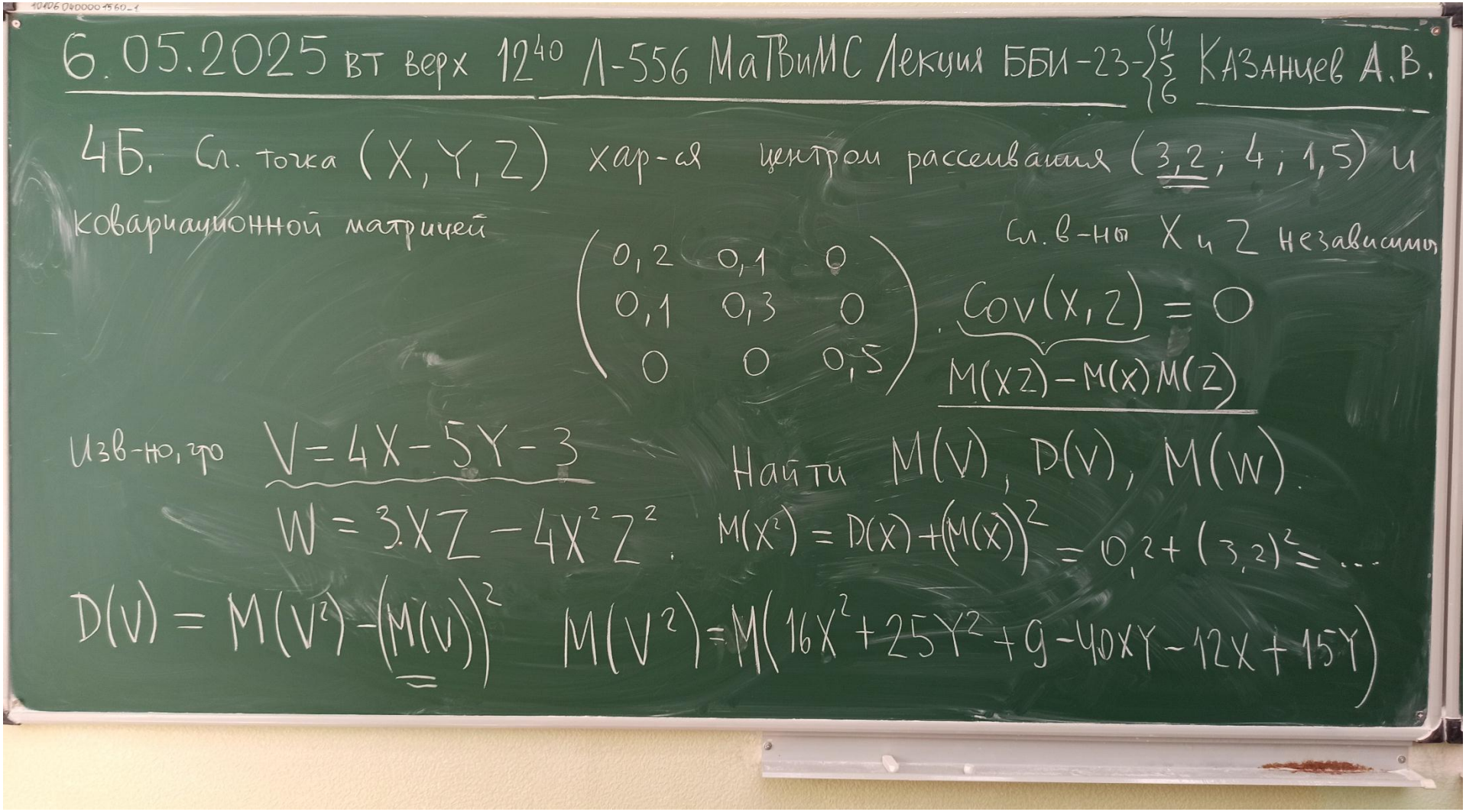


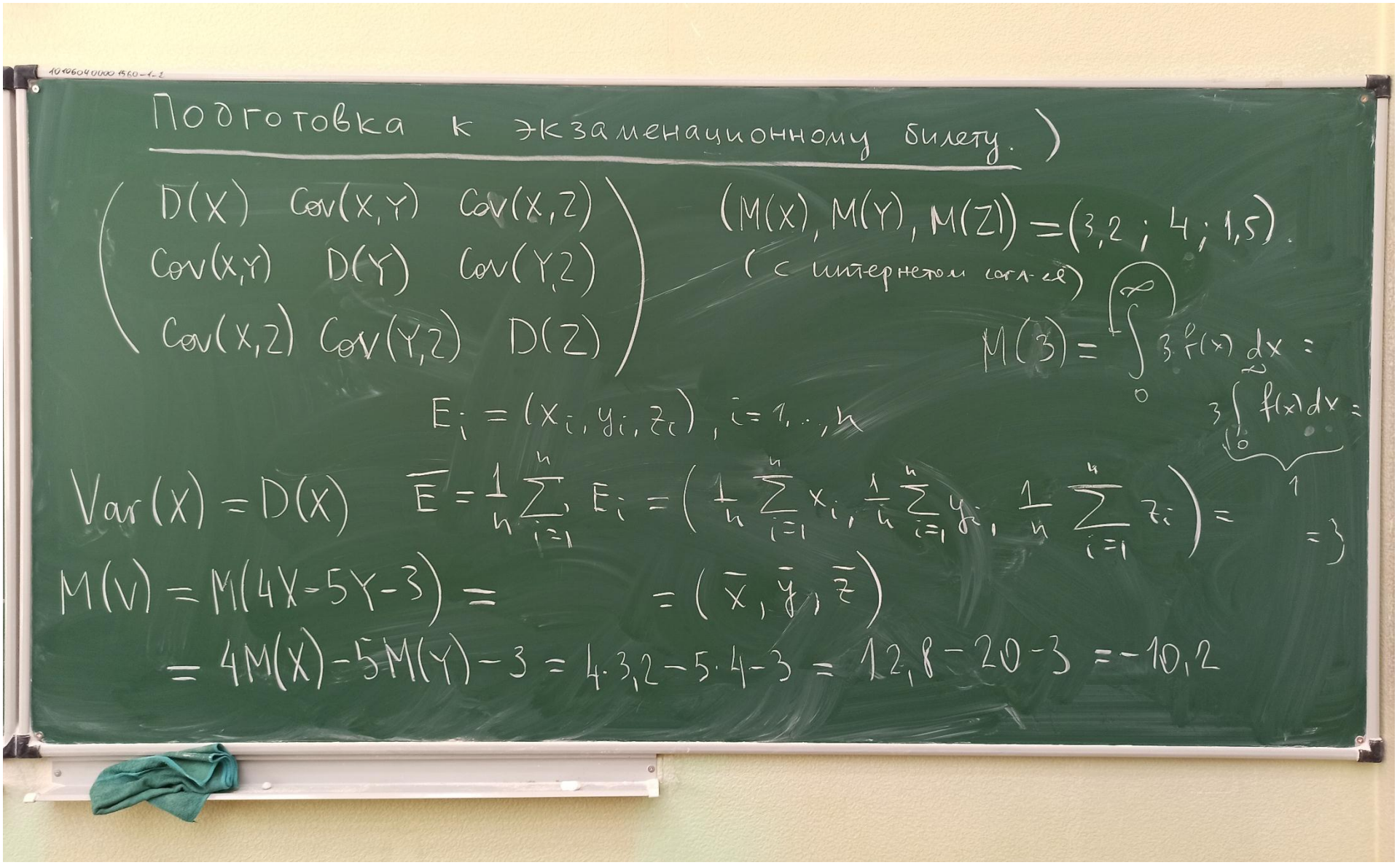

Вопросы по ИДЗ
$\frac{\sum x_i^2 - n\bar{x}^2}{n-1}$
$\frac{n\overline{x^2} - n\bar{x}^2}{n-1} = \frac{n}{n-1}(\overline{x^2} - \bar{x}^2)$
σ_x^2
s_x^2
$\frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{\sum x_i^2 - 2\sum x_i\bar{x} + \sum \bar{x}^2}{n-1} =$
$= \frac{\sum x_i^2 - 2\bar{x}\sum x_i + \bar{x}^2 \sum 1}{n-1} =$
$= \frac{\sum x_i^2 - 2n\bar{x}^2 + n\bar{x}^2}{n-1} =$
$= \frac{\sum x_i^2 - n\bar{x}^2}{n-1}$
Регрессия X на Y
$\hat{X} = \bar{X} + \hat{\beta}(y - \bar{Y}) = 0,5 + 0,5(y - 1,5)$
$\hat{\beta} = \frac{s_{xy}}{s_y^2} = \frac{2,25}{4,5} = 0,5$
$2\hat{x} = 1 + y - 1,5$
$y - 1,5 = 2\hat{x} - 1$
X на Y
Y на X
$\bar{Y}$
$\bar{x}$


20.05.2025 вт верх 12^40 Л-556 МатВиМС лекция ББИ-23-{4 5 6 Казанцев АВ
Cov(T,U) = M(TU) − M(T)M(U)
DT = Var(T) = Cov(T,T) = M(T²) − M(T)²
C = (C_TT C_TU ; C_UT C_UU)
T и U независимы ⇒ M(TU) = M(T)M(U) ⇒ cov(T,U) = 0
⇒ C = (· 0 ; 0 ·)
T² и U² независимы ⇒ M(T²U²) = M(T²)M(U²)
M(T+U) = MT + MU, M(S+T+U) = MS + MT + MU


$D(T+U) = M(T+U)^2 - (M(T+U))^2 = MT^2 + 2M(TU) + MU^2 - (MT)^2 - 2MT\,MU - (MU)^2 =$
$= DT + 2\,Cov(T,U) + DU = DT + DU + 2\,Cov(T,U)$.
T и U независимы $\Rightarrow D(T+U) = DT + DU$.
Упр. 1) $D(S+T+U) = ?$;
2) S, T, U независ $\Rightarrow D(S+T+U) = DS + DT + DU$.

С прошлой лекции




Подготовка к экзаменационному билету.)
$\begin{pmatrix} D(X) & Cov(X,Y) & Cov(X,Z) \\ Cov(X,Y) & D(Y) & Cov(Y,Z) \\ Cov(X,Z) & Cov(Y,Z) & D(Z) \end{pmatrix}$
$(M(X), M(Y), M(Z)) = (3,2; 4; 1,5)$.
(с интернетом согл-ся)
$M(3) = \int_0^{+\infty} 3 f(x)\,dx =$
$E_i = (x_i, y_i, z_i),\ i = 1, \dots, n$
$3\int_0^{+\infty} f(x)\,dx = 3 \cdot 1 =$
$Var(X) = D(X)$
$\overline{E} = \frac{1}{n}\sum_{i=1}^{n} E_i = \left(\frac{1}{n}\sum_{i=1}^{n} x_i, \frac{1}{n}\sum_{i=1}^{n} y_i, \frac{1}{n}\sum_{i=1}^{n} z_i\right) =$
$= (\bar{x}, \bar{y}, \bar{z})$
$M(V) = M(4X - 5Y - 3) =$
$= 4M(X) - 5M(Y) - 3 = 4 \cdot 3{,}2 - 5 \cdot 4 - 3 = 12{,}8 - 20 - 3 = -10{,}2$