МС-14 Аудиторное задание

Характеристики связи двух признаков

1. Создайте совокупности sin вида sin(1), sin(2), ..., sin(100) и cos вида cos(1), cos(2), ..., cos(100). Найдите эмпирический коэффициент корреляции признаков sin и cos на совокупности натуральных чисел от 1 до 100.

$$\widehat{cov(X,Y)} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

$$\widehat{cov(X,Y)} = \overline{XY} - \bar{Y}\cdot\bar{Y}$$

$$\hat{\rho}_{xy} = \frac{\widehat{cov(X,Y)}}{\hat{\sigma}_X\hat{\sigma}_Y}$$

2. В группе Ω учатся студенты: $\omega_1, \omega_2, \dots, \omega_{30}$. Пусть X и Y - 100-балльные экзаменационные оценки по математическому анализу и теории вероятностей. Оценки студента ω_i обозначаются: $x_i = X(\omega_i)$ и $y_i = Y(\omega_i)$, $i = 1, \dots, 30$. Все оценки известны:

$x_1 = 71$, $y_1 = 71$, $x_2 = 52$, $y_2 = 58$, $x_3 = 72$, $y_3 = 81$, $x_4 = 87$, $y_4 = 92$, $x_5 = 81$, $y_5 = 81$, $x_6 = 100$, $y_6 = 94$, $x_7 = 90$, $y_7 = 96$, $x_8 = 54$, $y_8 = 46$, $x_9 = 54$, $y_9 = 60$, $x_{10} = 58$, $y_{10} = 62$, $x_{11} = 56$, $y_{11} = 49$, $x_{12} = 70$, $y_{12} = 60$, $x_{13} = 93$, $y_{13} = 86$, $x_{14} = 46$, $y_{14} = 48$, $x_{15} = 56$, $y_{15} = 61$, $x_{16} = 59$, $y_{16} = 52$, $x_{17} = 42$, $y_{17} = 40$, $x_{18} = 60$, $y_{18} = 60$, $x_{19} = 33$, $y_{19} = 37$, $x_{20} = 83$, $y_{20} = 92$, $x_{21} = 50$, $y_{21} = 57$, $x_{22} = 93$, $y_{22} = 93$, $x_{23} = 41$, $y_{23} = 42$, $x_{24} = 55$, $y_{24} = 64$, $x_{25} = 60$, $y_{25} = 59$, $x_{26} = 37$, $y_{26} = 30$, $x_{27} = 71$, $y_{27} = 71$, $x_{28} = 42$, $y_{28} = 44$, $x_{29} = 85$, $y_{29} = 82$, $x_{30} = 39$, $y_{30} = 39$.

Требуется найти следующие условные эмпирические характеристики: 1) ковариацию X и Y при условии, что одновременно $X \geq 50$ и $Y \geq 50$; 2) коэффициент корреляции X и Y при том же условии.

Ответ. 1) Ковариация = 209,02. 2) Коэффициент корреляции = 0,9315.

3. Значения признаков X и Y заданы на множестве $\Omega = \{1,2,\dots,100\}$ таблицей частот

	$Y = 3$	$Y = 5$	$Y = 8$
$X = 300$	17	19	18
$X = 600$	13	14	19

Из Ω без возвращения извлекаются 8 элементов. Пусть $\bar{X}$ и $\bar{Y}$ – средние значения признаков в выборочной совокупности. Найдите $\mathrm{Cov}(\bar{X}, \bar{Y})$.

4. Ряд совместных наблюдений независимых нормально распределенных случайных величин X и Y, описывающих некоторый финансовый показатель двух фирм, задан двумерной выборкой:

{(167.9, -225.541); (133, -227.0618); (172.4, NA); (114.4, -187.947); (182.1, NA); (146.6, -238.1706); (NA, -195.6855); (157.5, -226.3498); (163.1, -232.4315); (164.6, -219.3768); (139.1, -205.4677); (112.9, NA); (149.6, -221.0258); (166.8, -190.341); (153.8, -219.2795); (NA, -198.8605); (87.5, -207.1957); (175.2, NA); (198.5, -277.6407); (147.7, -215.5379); (186, -209.1277); (150.9, -252.0035); (178.7, -221.1615); (143.3, -264.381); (148, -200.406); (NA, -291.3722); (184.8, -209.8789); (151.5, NA); (151.3, NA); (159.8, -261.9098); (124.5, -248.9302); (140, NA); (164.7, NA); (186.4, -255.7522); (154.5, -259.0014); (182.9, -222.0292); (112.8, -209.1327); (132.1, -224.1615); (180.5, -178.7437); (141.6, -261.1121); (157.8, -247.9286); (211, -209.7416); (136.9, -241.0031); (124.6, -276.8816); (109.4, -233.4274); (162.9, -235.5742); (130.8, NA); (187.5, -231.0311); (183.5, -232.3752); (193.9, -188.5517); (165, -257.8477); (184.5, -236.9394); (164.4, -225.4218); (166.1, -216.091); (241.3, -197.7659); (141.8, -219.751); (NA, -207.4731); (NA, -240.3647); (NA, -258.889); (136.6, -217.16); (194.5, -261.1401); (157.6, NA); (149.6, -213.3036); (152.5, -288.5258); (170.4, -241.9711); (NA, -243.0995); (133.6, -232.4539); (139.1, -214.5584); (111.7, NA); (138.1, -271.9439); (166.3, -204.7177); (185.6, NA); (160.4, -229.6342); (152.4, -237.8129); (197.6, -207.0127); (149.8, NA); (180.7, -215.8441); (156.1, -221.4436); (130.5, -286.4889); (140, -235.5511); (NA, -229.0371); (143.1, -257.7442); (177.6, -220.4417); (124.7, -256.3137); (142, -218.7544); (143.6, -260.6194); (121.3, -186.2013); (78.2, -173.376); (155.9, -261.1379); (137.6, -237.259); (170.8, -204.3441); (156.8, -212.3563); (128.4, -200.0559); (NA, -238.497); (129.3, -238.3039); (147.1, -257.0837); (117.9, -205.2149); (174.3, -247.1452); (163.2, -194.3524); (151.5, -219.2332); (153.3, -192.9653); (148.4, -215.8789); (174.8, -205.3518); (84.2, -197.7495); (163.6, -227.4809); (205.5, -250.75); (169.8, -211.6129); (NA, -188.3579); (116.9, NA); (205.5, -180.5642); (181.1, -195.1596); (137.4, -222.561); (140.5, -255.2292); (125, -221.2531); (212.9, -196.9889); (152.7, -200.074); (137.4, NA); (142.8, -201.6862); (178.4, -232.8285); (165.1, -208.838); (NA, -240.8741); (134.3, -224.8478); (180.5, -229.5657); (122.8, -204.9998); (179.7, -272.7181); (163.8, -239.3508); (182.2, -232.8887); (172.8, -220.529); (NA, -221.5642); (NA, -195.5116); (151, -222.4601); (NA, -256.248); (204.2, -230.9828); (182.9, -234.9166); (219.3, -198.5935); (153.4, NA); (85.1, -201.3523); (214.6, -226.9573); (96.2, -245.2855); (153, -261.5914); (112.8, -212.7011); (NA, -244.1466); (NA, -213.4919); (153.3, -239.8558); (177.6, -272.8503); (158.6, -314.0774); (NA, -249.3596); (162.3, -216.9371); (123.8, -197.6739); (158.3, -235.9429)}.

Скопируйте данную выборку и преобразуйте ее в Python или Excel в столбцы "A" и "B" соответственно для первой и второй фирмы. При этом связанные значения показателей должны располагаться в одной строке.

Очистите исходную выборку от пропущенных данных, обозначенных как "NA", и вычислите выборочный коэффициент корреляции Пирсона между X и Y.

5. Рассмотрим некоторые данные официальной статистики за 2010 г. для восьми федеральных округов РФ. Имеются сведения о численности зрителей театров на 1000 человек населения и средних уровнях безработицы в этих округах (https://rosstat.gov.ru/folder/210/document/13204). Безработица определяется как отношение численности незанятого населения к общей численности трудоспособного населения (%). Численность зрителей определяется как отношением численности зрителей к среднегодовой численности населения федерального округа.

В результате имеем таблицу:

Федеральный округ	Численность зрителей театров на 1000 человек населения, X	Уровень безработицы, %, Y
Центральный	262	4,7
Северо-Восточный	279	6,2
Южный	123	7,7
Северо-Кавказский	90	16,9
Приволжский	210	7,6
Уральский	209	8,0
Сибирский	239	8,7
Дальневосточный	185	8,7

Построить диаграмму рассеивания точек (наблюдений). Можно ли считать предложенные в задачах социальные характеристики коррелированными? Рассчитать коэффициент линейной корреляции Пирсона в Python или Excel непосредственно по формуле, используя функцию **Excel КОРРЕЛ( )** и инструмент пакета Анализ данных «**Корреляция**».

6. Найти матрицу корреляций всех столбцов из файла **Corr.xlsx** Лист <**Данные**>.

Домашнее задание

1. Эмпирическое распределение признаков X и Y на генеральной совокупности $\Omega = \{1,2,\ldots,100\}$

	$Y=1$	$Y=3$	$Y=5$
$X=100$	13	16	20
$X=300$	12	28	11

Из Ω случайным образом без возвращения извлекаются 10 элементов. Пусть $\bar{X}$ и $\bar{Y}$ – средние значения признаков на выбранных элементах. Требуется найти: 1) математическое ожидание $E(\bar{X})$; 2) дисперсию $Var(\bar{Y})$; 3) коэффициент корреляции $\rho(\bar{X},\bar{Y})$.

2. Эмпирическое распределение признаков X и Y на генеральной совокупности $\Omega=\{1,2,...,100\}$ задано таблицей частот

	$Y=1$	$Y=3$	$Y=5$
$X=200$	10	15	16
$X=300$	14	11	34

Из Ω случайным образом без возвращения извлекаются 10 элементов. Пусть $\bar{X}$ и $\bar{Y}$ – средние значения признаков на выбранных элементах. Требуется найти: 1) математическое ожидание $E(\bar{Y})$; 2) стандартное отклонение $\sigma(\bar{X})$; 3) ковариацию $\mathrm{Cov}(\bar{X},\bar{Y})$.

3. Ряд совместных наблюдений независимых нормально распределенных случайных величин X и Y, описывающих некоторый финансовый показатель двух фирм, задан двумерной выборкой:

{(-214.4, -196.1222); (-256.9, -175.173); (-202.3, NA); (-239.3, -208.7948);
(-287.3, -238.7711); (-196.5, -274.856); (-224, -278.7524); (-240.2, -229.0939);
(-231.9, -237.8394); (-179.2, -208.9465); (NA, -186.388); (-258.8, NA);
(-237.8, -278.6498); (-273.5, -220.0173); (-206, -217.5404); (NA, NA);
(-227, -204.1577); (-241.6, -240.6529); (-234.6, NA); (-235.2, -266.0953);
(-202.4, -221.0894); (-183.1, -206.8865); (-220, -274.9425); (-208.2, -232.9762);
(NA, -211.8263); (-245.4, -240.4984); (-254.2, -230.6192); (-216.6, -260.0378);
(-221.7, NA); (-192.8, -242.5521); (-229.3, NA); (-212.9, -216.6312); (-230.5, NA);
(-220.7, -231.6356); (-201.2, NA); (-220.6, -270.0291); (-237.4, -222.8705);
(-248.9, -282.398); (-249.5, -168.73); (-196.7, -243.2915); (-241.7, -228.9693);
(-254.8, -243.7941); (-217, -169.2123); (-209.8, -186.8498); (-279, -266.3833);
(NA, NA); (-197.2, -201.7408); (NA, -235.5852); (-202.5, -252.863);
(-273.1, -220.8998); (-220.5, -239.6719); (-252.8, -272.5675); (-235.4, -218.1677);
(-206.6, -190.3638); (-213.1, -252.4642); (-207.9, -229.7951); (-272.4, -187.7126);
(-224.2, -224.0721); (-169.8, -231.0987); (-216.3, -187.6854); (-250, -243.1863);
(-227.2, -212.2725); (-229.2, -258.8585); (-251.3, -247.0714); (-236.6, -227.2609);
(-232.5, -195.074); (-234.7, -281.0113); (-240.6, -235.828); (-245.5, -217.1208);
(-223.4, -204.1562); (-236.2, -199.0068); (NA, -202.9738); (-254.3, NA);
(-259.1, -227.7556); (-279, NA); (-224.2, -231.7992); (-201.4, NA);
(-244.3, -239.6067); (-179.6, -216.5177); (-165.3, -263.392); (-229.9, -204.9858);
(-246.4, -172.9445); (-202.1, -196.6104); (-231.7, -210.8363); (NA, -222.951);
(-209.2, NA); (-200.1, -198.9821); (-232.9, -237.748); (-229.6, -254.527);
(-219.1, -252.8629); (-201.5, -252.3746); (-229.8, -235.0515); (-248.5, -214.7245);
(-182.7, -241.9679); (-236.3, -185.3818); (-285.3, -241.1946); (-262.6, -259.8436);
(-213.6, -207.5926); (-201, -267.2625); (-224.3, -247.7756); (NA, -185.1585);
(-226.7, -264.6949); (-216.4, -295.7108); (-190, -202.274); (-265.5, -246.4795);
(-213, -260.683); (-232.3, -246.545); (-217.2, -263.4495); (-199.2, -230.0708);
(-208.6, -268.45); (-201.2, -214.3047); (-262.2, -212.7205); (-188.2, -274.0026);
(-155.2, -270.0219); (-288, -219.8755); (-226.4, -248.257); (-252.9, -234.2434);
(-238.5, NA); (-226.2, -223.539); (-160.6, -254.5109); (-243.6, -237.9364);

(-219.6, -196.5018); (-208.3, -224.02); (-218.6, -261.5601); (-228.3, -238.0651); (-263, -227.8186); (-265, -194.2366); (-206.4, -240.3734); (-239.6, -226.9095); (NA, -218.4492); (-227.2, -236.1292); (-232.9, -266.8274); (-242.1, -218.763); (-217.3, -317.6571); (NA, -213.1899); (-213.3, -239.8124); (-266.6, -233.759); (-237.4, -244.3302); (-313.6, -246.9523); (-250.8, -215.8164); (-210.3, -258.3312); (-250.9, -210.9057); (-240.6, -230.8002); (-242.8, -274.1018); (-235.2, -234.8576); (-291.4, -244.8539); (NA, NA); (-207, -229.9618); (-206.7, NA); (-152.7, NA)}.

Скопируйте данную выборку и преобразуйте ее в Python или Excel в столбцы "A" и "B" соответственно для первой и второй фирмы. При этом связанные значения показателей должны располагаться в одной строке.
Очистите исходную выборку от пропущенных данных, обозначенных как "NA", и вычислите выборочный коэффициент корреляции Пирсона между X и Y.

4. Имеется следующая статистика по ресторанам-закусочным быстрого обслуживания (в выборку включены рестораны, рассчитанные на большое число посетителей и имеющие просторные обеденные залы), расположенным на автомобильной магистрали:

Ресторан-закусочная	**№1**	**№2**	**№3**	**№4**	**№5**	**№6**
Средняя цена тарелки супа, руб.	40	30	120	50	140	100
Процент занятых столиков в дневное время	90	90	70	85	65	80

Вычислите коэффициент линейной корреляции Пирсона между указанными двумя показателями и дайте его содержательную интерпретацию.

5. Найти и раскрасить матрицу корреляций для столбцов A и B из файла **Corr.xlsx**