Министерство образования и науки Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
АМУРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
(ФГБОУ ВО «АмГУ»)

МЕТОДЫ ПРИКЛАДНОЙ СТАТИСТИКИ ДЛЯ СОЦИОЛОГОВ

сборник учебно-методических материалов

для направления подготовки 39.03.01 – Социология

2017 г.

Печатается по решению
редакционно-издательского совета
факультета математики и информатики
Амурского государственного
Университета

Составитель: Гришкина Т.Е.

Методы прикладной статистики для социологов: сборник учебно-методических материалов для направления подготовки 39.03.01 – Социология. – Благовещенск: Амурский гос. ун-т, 2017.

Рассмотрен на заседании кафедры общей математики и информатики 03.11.2017, протокол № 3.

ВВЕДЕНИЕ

Цель дисциплины: формирование культуры использования методов прикладной статистики для решения задач социологического характера и формирование инструментария для эффективного и своевременного получения качественных результатов социологических исследований на основании первичной статистической информации.

Задачи дисциплины:

- сбор и обработка «сырой» информации, необходимой для количественной и качественной оценки социальных процессов;
- анализ известных моделей с позиций их устойчивости к незначительным изменениям окружающей действительности;
- построение моделей путем анализа первичной информации, нахождение параметров моделей и верификация моделей.

В результате освоения дисциплины обучающийся должен демонстрировать следующие результаты образования:

1) знать: основные методы и модели прикладной статистики, применяемые в социологии;

2) уметь: применять методы моделирования социальных процессов, использовать средства дескриптивной статистики основные подходы к статистическому выводу; оценивать применимость средств формального представления для различных типов социально-экономических данных;

3) владеть: навыками научного анализа социальных проблем и процессов, навыками практического использования базовых знаний и методов математики и естественных наук; приемами прикладного статистического анализа социологической информации.

1. КРАТКОЕ ИЗЛОЖЕНИЕ ЛЕКЦИОННОГО МАТЕРИАЛА

Тема: Основы измерения и количественного описания данных

Ключевые вопросы

Выборка ее репрезентативность, нормальное распределение, правило «трех сигм», теорема Ляпунова и ее следствия, понятие статистической гипотезы, уровень значимости, доверительный интервал, зона неопределенности, алгоритм проверки статистических гипотез, описательные статистики.

Основные определения и методы

Математическая статистика занимается математическим описанием случайных явлений, т.е. построением вероятностных моделей, а также проверкой их пригодности. Поэтому выделяют два раздела: описательную статистику и статистику «проверяющую» (статистическую проверку гипотез); соответственно разделяется и методический аппарат. Понятия и методы описательной статистики создаются в теории вероятностей, а понятия и методы статистической проверки гипотез создаются в специальных теориях (например, в теории статистических решений) либо в приложениях теории вероятностей к конкретным наукам.

Генеральная совокупность – все множество имеющихся объектов. Выборка – набор объектов, случайно отобранных из генеральной совокупности. Объем генеральной совокупности N и объем выборки n – число объектов в рассматриваемой совокупности. Виды выборки: повторная – каждый отобранный объект перед выбором следующего возвращается в генеральную совокупность; бесповторная – отобранный объект в генеральную совокупность не возвращается.

Для того чтобы, по исследованию выборки можно было сделать выводы о поведении интересующего нас признака генеральной совокупности, нужно, чтобы выборка правильно представляла пропорции генеральной совокупности, то есть была репрезентативной (представительной). Учитывая закон больших чисел, можно утверждать, что это условие выполняется, если каждый объект выбран случайно, причем для любого объекта вероятность попасть в выборку одинакова.

В процессе статистического анализа иногда бывает необходимо сформулировать и проверить предположения относительно величины независимых параметров или закона распределения изучаемой генеральной совокупности (совокупностей). Такие предположения называются статистическими гипотезами.

Статистические гипотезы подразделяются на нулевые и альтернативные.

Выдвинутая гипотеза называется нулевой (основной). Ее принято обозначать H_0. Обычно нулевая гипотеза – это гипотеза об отсутствии различий.

По отношению к высказанной нулевой гипотезе всегда можно сформулировать альтернативную (конкурирующую), противоречащую ей. Альтернативную гипотезу принято обозначать H_1 .

Цель статистической проверки гипотез состоит в том, чтобы на основании выборочных данных принять решение о справедливости нулевой гипотезы H_0.

Так как проверка статистических гипотез осуществляется на основании выборочных данных, то такое решение неизбежно сопровождается некоторой, хотя возможно и очень малой, ошибкой.

Ошибка, состоящая в том, что мы отклонили нулевую гипотезу, в то время как она верна, называется ошибкой I рода, а ее вероятность – уровнем значимости α.

Ошибка, состоящая в том, что мы приняли нулевую гипотезу, в то время как она неверна, называется ошибкой II рода, а ее вероятность обозначают β. Величину равную $1 - \beta$ называют мощностью критерия.

Мощность критерия определяется эмпирическим путем, а уровень значимости задается исследователем. В психологических и социологических исследованиях низшим уровнем значимости принято считать α.=0,05 а достаточным α= 0,01.

Статистический критерий — это правило (формула), по которому определяется мера расхождения результатов выборочного наблюдения с высказанной гипотезой H_0.

Значение критерия, рассчитываемое по специальным правилам на основании выборочных данных, называется наблюдаемым значением критерия.

Значения критерия, определяемые на заданном уровне значимости α по таблицам распределения случайной величины, выбранной в качестве критерия, называются критическими точками.

В психолого-педагогических и социологических исследованиях принято определять значения критерия при α=0,01 и α=0,05. Полученные критические точки делят совокупность значений критерия на область допустимых значений или зону незначимости (область принятия нулевой гипотезы), критическую область или зону значимости (область принятия альтернативной гипотезы) и зону неопределенности.

Чаще всего критерии, используемые при психологических и социологических исследованиях, имеют положительные значения. Поэтому для простоты при решении прикладных задач изображают только неотрицательную часть оси значимости.

Основной принцип проверки статистических гипотез состоит в следующем:

- если наблюдаемое значение критерия принадлежит критической области, то нулевая гипотеза H_0 отклоняется и принимается конкурирующая H_1;

- если наблюдаемое значение критерия принадлежит области допустимых значений, то нулевую гипотезу H_0 нельзя отклонить;

-если наблюдаемое значение критерия принадлежит зоне неопределенности, то мы уже можем отклонить нулевую гипотезу H_0, но еще не можем принять конкурирующую H_1.

Критерии делятся на параметрические – включающие в формулу расчета параметры распределения (данные подчиняются нормальному закону распределения) и непараметрические – основанные на оперировании частотами или рангами.

Нормальным называется распределение вероятностей непрерывной случайной величины, которое описывается плотностью вероятности

$$f(x) = \frac{1}{\sigma_x \sqrt{2\pi}} e^{-\frac{(x-m_x)^2}{2\sigma_x^2}}$$

Нормальный закон распределения также называется законом Гаусса.

График плотности нормального распределения называется нормальной кривой или кривой Гаусса.

При рассмотрении нормального закона распределения выделяется важный частный случай, известный как правило трех сигм: вероятность того, что случайная величина отклонится от своего математического ожидания на величину, большую чем утроенное среднее квадратичное отклонение, практически равна нулю.

На практике считается, что если для какой – либо случайной величины выполняется правило трех сигм, то эта случайная величина имеет нормальное распределение.

Тема: Параметрические критерии проверки статистических гипотез

Ключевые вопросы

Критерии Стьюдента, Фишера для зависимых и независимых выборок, однофакторный и многофакторный дисперсионный анализ. Ограничения параметрических критериев

Основные определения и методы

1. t-критерий Стьюдента используется

а) для сравнения выборочной средней $\bar{x}$ с некоторым известным числовым значением a_0.

Возможны гипотезы:

H_0: $\bar{x} = a_0$ выборочная средняя генеральной совокупности равна заданному числу a_0.

H_1: $\bar{x} \neq a_0$ ($\bar{x} < a_0$, $\bar{x} > a_0$)выборочная средняя генеральной совокупности не равна (меньше, больше) заданному числу a_0.

Наблюдаемое значение t-критерия рассчитывается по формуле:

- если дисперсия генеральной совокупности неизвестна

$$t_{набл} = \frac{\bar{x} - a_0}{S}\sqrt{n}$$

- если дисперсия генеральной совокупности известна

$$t_{набл} = \frac{\bar{x} - a_0}{\sigma_{ген}}\sqrt{n}$$

где $\bar{x}$ - выборочная средняя;

a_0 — числовое значение генеральной средней;

S — исправленное среднее квадратическое отклонение;

$\sigma^2_{ген}$ - известная дисперсия генеральной совокупности;

n — объем выборки.

Критическое значение $t_{кр}$ следует находить с помощью таблиц распределения Стьюдента по уровню значимости α и числу степеней свободы $k = n - 1$.

b) для обнаружения различия между средними значениями $\bar{x}$, $\bar{y}$ двух выборок.

Возможны гипотезы:

H_0: $\bar{x} = \bar{y}$ средние значения двух выборок равны,

H_1: $\bar{x} \neq \bar{y}$ средние значения двух выборок не равны.

Наблюдаемое значение t-критерия рассчитывается по формуле:

- для независимых выборок

$$t_{набл} = \frac{\bar{x} - \bar{y}}{\sqrt{(n_1 - 1)S_x^2 + (n_2 - 1)S_y^2}}\sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}}$$

- для зависимых выборок

$$t_{эмп} = \frac{\sum d}{\sqrt{\frac{n\sum d^2 - (\sum d)^2}{n-1}}}$$

где $S_x^2 = \frac{1}{n_1 - 1}\sum\left(x - \bar{x}\right)^2$ - выборочная дисперсия 1 выборки;

$S_y^2 = \frac{1}{n_2 - 1}\sum\left(y - \bar{y}\right)^2$ - выборочная дисперсия 2 выборки;

$\bar{x}$ - среднее значение признака для 1 выборки;

$\bar{y}$ - среднее значение признака для 2 выборки;

n_1 – объем 1 выборки;

n_2 – объем 2 выборки;

d — разность между результатами в каждой паре («после»минус «до»);

n —число пар данных в зависимых выборках

Критическое значение $t_{кр}$ следует находить с помощью таблиц распределения Стьюдента по уровню значимости α и числу степеней свободы $k = n_1 + n_2 - 2$.

t-критерий для независимых выборок можно использовать для сравнения средних показателей экспериментальной группы с контрольной группой.

t-критерий для зависимых выборок очень полезен в тех ситуациях, когда две сравниваемые группы основываются на одной и той же совокупности наблюдений (субъектов), которые тестировались дважды (например, до и после эксперимента).

2. *F* — критерий Фишера-Снедекора используют

а) для сравнения разброса значений двух выборок, т.е. для проверки гипотезы о равенстве дисперсий.

Возможны гипотезы:

H_0 : $S_x^2 = S_y^2$ - разброс значений признака относительно среднего одинаковый в обеих выборках.

H_1: $S_x^2 \neq S_y^2$ - разброс значений признака не совпадает.

Наблюдаемое значение F — критерия рассчитывается по формуле:

$$F_{набл} = \frac{S_x^2}{S_y^2},$$

где S_x^2 — большая (по величине) выборочная дисперсия;

S_y^2 — меньшая (по величине) выборочная дисперсия.

Критическое значение $F_{крит}$ следует находить с помощью таблицы распределения Фишера-Снедекора по уровню значимости α и числу степеней свободы $k_1 = n_1 - 1$ и $k_2 = n_2 - 1$,

где k_1 — число степеней свободы большей (по величине) дисперсии;

k_2 — число степеней свободы меньшей (по величине) дисперсии;

n_1 — объем выборки большей (по величине) дисперсии;

n_2 — объем выборки меньшей (по величине) дисперсии.

Тема: Непараметрические критерии проверки статистических гипотез

Ключевые вопросы

Ранжирование переменных; критерии различий (Q-Розенбаума, U-Манна-Уитни, H-Крускала-Уоллиса, S-тенденций Джонкира); критерии изменения (G-знаков, T-Вилкоксона, χ2-Фридмана, L-тенденций Пейджа); алгоритмы, сходства, различия и ограничения критериев; виды задач, решаемых с помощью данных критериев

Основные определения и методы

а) Q-критерий Розенбаума

Критерий используется для оценки различий между двумя выборками по уровню какого-либо признака, количественно измеренного.

Возможны гипотезы:

H_0: Уровень признака в выборке 1 не превышает уровня признака в выборке 2.

H_1: Уровень признака в выборке 1 превышает уровень признака в выборке 2.

Ограничения критерия Q

1) В каждой из сопоставляемых выборок должно быть не менее 11 наблюдений. При этом объемы выборок должны примерно совпадать. Если в обеих выборках меньше 50 наблюдений, то абсолютная величина разности между объемами выборок n_1 и n_2, соответственно, не должна быть больше 10 наблюдений. Если в каждой из выборок больше 51 наблюдения, но меньше 100, то абсолютная величина разности между объемами выборок n_1 и n_2, соответственно, не должна быть больше 20 наблюдений. Если в каждой из выборок больше 100 наблюдений, то допускается, чтобы одна из выборок была больше другой не более чем в 1,5-2 раза.

2) Диапазоны разброса значений в двух выборках должны не совпадать между собой, в противном случае применение критерия бессмысленно.

3) Измерение может быть проведено по шкале порядка, интервалов или отношений.

4) Выборки должны быть независимыми.

Эмпирическое значение критерия подсчитывается по формуле:

$$Q_{эмп} = S_1 + S_2,$$

где S_1 – количество наблюдений в выборке 1, которые выше максимального значения в выборке 2;

S_2 - количество наблюдений в выборке 2, которые ниже минимального значения выборки 1.

Значения в выборках должны быть упорядочены по возрастанию признака.

Критические значения Q-критерия определяются по таблице для данных n_1 и n_2 и для выбранного уровня значимости. Если $Q_{эмп}$ не меньше $Q_{кр}$, то H_o отвергается.

б) U- критерий Манна-Уитни

Критерий предназначен для оценки различий между двумя выборками по уровню какого-либо признака, количественно измеренного.

Возможны гипотезы:

H_0: Уровень признака в группе 2 не ниже уровня признака в группе 1.

H_1: Уровень признака в группе 2 ниже уровня признака в группе 1.

Ограничения критерия U

1) В каждой выборке должно быть не менее 3 наблюдений. Допускается, чтобы в одной выборке было 2 наблюдения, но тогда во второй их должно быть не менее 5.
2) В каждой выборке должно быть не более 60 наблюдений.
3) Измерение может быть проведено по шкале интервалов или отношений.
4) Выборки должны быть несвязными.

Наблюдения обеих выборок необходимо объединить и проранжировать по степени нарастания признака.

Эмпирическое значение критерия рассчитывается по формуле:

$$U_{эмп} = (n_1 \cdot n_2) + \frac{n_x \cdot (n_x + 1)}{2} - T_x,$$

где n_1, n_2 - количество испытуемых в выборках 1 и 2 соответственно,

T_x – большая из ранговых сумм,

n_x – количество испытуемых в группе с большей суммой рангов.

Критическое значение критерия определяется по таблице для данных n_1 и n_2 и выбранного уровня значимости. Если $U_{эмп}$ больше $U_{кр}$, то принимается H_o.

в) H - критерий Крускала – Уоллиса

Критерий предназначен для оценки различий между тремя и более выборками по уровню какого-либо признака. Он позволяет установить, что уровень признака изменяется при переходе от группы к группе, но не указывает направление этих изменений.

Возможны гипотезы:

H_o: Между выборками 1, 2, 3 и т. д. существуют лишь случайные различия по уровню исследуемого признака.

H_1: Между выборками 1, 2, 3 и т. д. существуют неслучайные различия по уровню исследуемого признака.

Ограничения критерия H

1) Измерение может быть проведено по шкале интервалов или отношений.
2) Выборки должны быть независимыми.
3) При сопоставлении 3-х выборок допускается, чтобы в одной из них $n=3$, а в двух других $n=2$. Но при таких численных составах установить различия можно лишь на низшем уровне значимости.
4) При большем 5 количестве выборок и испытуемых в каждой выборке необходимо пользоваться таблицей критических значений критерия χ^2. Число степеней свободы при этом определяется как ν=c-1, где c –количество сопоставляемых выборок.

Наблюдения всех выборок необходимо объединить и проранжировать по степени нарастания признака.

Эмпирическое значение критерия H подсчитывается по формуле:

$$H_{эмп}=\frac{12}{N(N+1)}\cdot\sum_{i=1}^{c}\frac{T_i^2}{n_i}-3\cdot(N+1),$$

где N – общее количество испытуемых в объединенной выборке,

n_i – количество испытуемых в каждой выборке,

T_i^2 – квадраты сумм рангов по каждой i–ой выборке.

Если эмпирическое значение критерия меньше критического значения, то H_o принимается.

г) S – критерий тенденций Джонкира

Критерий S предназначен для выявления тенденций изменения признака при переходе от выборки к выборке при сопоставлении трех и более выборок.

Гипотезы:

H_o: Тенденция возрастания значений признака при переходе от выборки к выборке является случайной.

H_1: Тенденция возрастания значений признака при переходе от выборки к выборке не является случайной.

Ограничения критерия

1) Измерение может быть проведено по шкале интервалов или отношений.
2) Выборки должны быть независимыми.
3) Количество наблюдений в каждой выборке должно быть одинаковым.
4) Нижняя граница применимости критерия: не менее трех выборок и не менее двух элементов в каждом наблюдении. Верхняя граница определяется таблицей приложения: не более 6 выборок и не более 10 наблюдений в каждой выборке.

Выборки необходимо располагать по возрастанию суммы значений признака слева на право.

Эмпирическое значение критерия рассчитывается по формуле:

$$S_{эмп}=2A-B,$$

где A – общая сумма инверсий,

$\mathrm{B}=\frac{\mathrm{c}\cdot(\mathrm{c}-1)}{2}\cdot\mathrm{n}^2$ – максимально возможное значение величины А.

Под числом инверсий понимается число значений признака, больших каждого конкретного значения рассматриваемой выборки и расположенных правее от нее.

Критические значения критерия определяются по таблице, в соответствии с выбранным уровнем значимости, количеством выборок (c) и числом наблюдений (n) в каждой выборке.

Если эмпирическое значение критерия меньше критического значения, то принимается гипотеза H_o.

д) G-критерий знаков

Критерий знаков предназначен для установления общего направления сдвига исследуемого признака. Он позволяет установить, в какую сторону в выборке в целом изменяются значения признака при переходе от первого измерения ко второму: изменяются ли показатели в сторону улучшения, повышения или усиления или, наоборот, в сторону ухудшения, понижения или ослабления.

Возможны гипотезы:

H_0: Преобладание типичного направление сдвига является случайным.

H_1: Преобладание типичного направление сдвига не является случайным.

Ограничения критерия G

1) Измерение может быть проведено по шкале порядка, интервалов или отношений.
2) Выборка должна быть однородной и связной.
3) Объем выборки должен быть равным от 5 до 300.
4) При равенстве типичных и нетипичных сдвигов критерий знаков неприменим.

Эмпирическое значение критерия $G_{эмп}$ принимают равным числу нетипичных сдвигов, т. е. не преобладающих сдвигов в сторону увеличения или уменьшения показателя.

Критическое значение критерия $G_{кр}$ определяют по таблице в соответствии с выбранным уровнем значимости и объемом выборки без учета нулевых сдвигов. Если $G_{эмп}$ не превосходит $G_{кр}$, то гипотеза H_0 отвергается.

ж) Парный критерий Т – Вилкоксона

Критерий применяется для сопоставления показателей, измеренных в двух разных условиях на одной и той же выборке испытуемых. Он позволяет установить не только направленность изменений, но и их выраженность. С его помощью мы определяем, является ли сдвиг показателей в каком-то одном направлении более интенсивным, чем в другом.

Возможны гипотезы:

H_0: Интенсивность сдвигов в типичном направлении не превосходит интенсивности сдвигов в нетипичном направлении.

H_1: Интенсивность сдвигов в типичном направлении превышает интенсивности сдвигов в нетипичном направлении.

Ограничения критерия Т

1) Измерение может быть проведено по любой шкале, кроме номинальной.
2) Выборка должна быть связной.
3) Объем выборки должен быть равным от 5 до 50.

Эмпирическое значение критерия подсчитывают по формуле:

$$T_{эмп} = \sum R_r ,$$

где R_r – ранговые значения сдвигов с более редким знаком.

Критическое значение критерия $T_{кр}$ определяется для данного объема выборки и выбранного уровня значимости по таблице. Если $T_{эмп}$ не превосходит $T_{кр}$, то гипотеза H_0 отвергается.

з) Критерий χ_r^2 Фридмана

Критерий применяется для сопоставления показателей, измеренных в трех или более условиях на одной и той же выборке испытуемых. Он позволяет установить, что величины показателей от условия к условию изменяются, но при этом не указывает на направление изменений.

Возможны гипотезы:

H_0: Между показателями, полученными (измеренными) в разных условиях, существуют лишь случайные различия.

H_1: Между показателями, полученными (измеренными) в разных условиях, существуют неслучайные различия.

Ограничения критерия χ_r^2

1) Измерение может быть проведено по шкале интервалов или отношений.
2) Выборка должна быть связной.
3) В выборке должно быть не менее двух испытуемых, каждый из которых имеет не менее трех показателей. Количество измерений не может превышать 100.

Эмпирическое значение критерия вычисляется по формуле:

$$\chi^2_{r\ эмп} = \left[\frac{12}{n \cdot c \cdot (c+1)} \cdot \sum_{i=1}^{c} T_i^2 \right] - 3 \cdot n \cdot (c+1) ,$$

где c – количество условий,

n – количество испытуемых,

T_i – суммы рангов по каждому из условий.

Критическое значение критерия $\chi^2_{r\ кр}$ определяем при выбранном уровне значимости и данном объеме выборки по правилам:

1) При $c=3$ и $n \leq 9$, критические значения определяются по таблице.
2) При $c=4$ и $n \leq 4$, критические значения определяются по таблице.
3) При большем числе измерений и испытуемых критические определяются по таблице для критерия χ^2. В этом случае число степеней свободы определяется по формуле $v = c - 1$.

Если $\chi_r^2{}_{эмп}$ не меньше $\chi_r^2{}_{кр}$ то гипотеза H_0 отклоняется.

и) L – критерий тенденций Пейджа.

Критерий L Пейджа применяется для сопоставления показателей, измеренных в трех и более условиях на одной и той же выборке испытуемых. Критерий позволяет выявить тенденции в изменении величин признака при переходе от условия к условию, а также указывает на направление этих изменений.

Возможны гипотезы:

H_0: Увеличение индивидуальных показателей при переходе от первого условия ко второму, а затем к третьему и далее, случайно.

H_1: Увеличение индивидуальных показателей при переходе от первого условия ко второму, а затем к третьему и далее, неслучайно.

Ограничения критерия L

1) Измерение может быть проведено по ранговой шкале, шкале интервалов или отношений.

2) Выборка должна быть связной.

3) В выборке должно быть не менее двух и не больше 12 испытуемых, каждый из которых имеет не менее трех показателей. Максимальное число условий – 6.

Эмпирическое значение критерия определяется по формуле:

$$L_{эмп} = \sum_{i=1}^{c} (T_i \cdot i),$$

где c – количество условий,

T_i – суммы рангов по каждому из условий,

i –порядковый номер, приписанный каждому условию, после упорядочения по возрастанию сумм рангов.

Критическое значение критерия $L_{кр}$ определяем при выбранном уровне значимости, данном объеме выборки и данном количестве условий по таблице. Если $L_{эмп}$ не меньше $L_{кр}$, то гипотеза H_0 отклоняется.

Тема: Критерии согласия

Ключевые вопросы

Эмпирические и теоретические частоты; критерии согласия χ2-Пирсона, λ-Колмогорова-Смирнова, φ-Фишера, их алгоритмы, сходства и различия; примеры задач.

Основные определения и методы

1. χ^2 – критерий Пирсона.

Основная расчетная формула эмпирического значения χ^2:

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_э - f_m)^2}{f_m},$$

где k – количество разрядов признака,

f_m – теоретическая частота,

$f_э$ – эмпирическая частота.

Если количество разрядов равно двум (принимает минимально возможное значение), то в расчетную формулу вносится поправка на непрерывность:

$$\chi^2 = \sum_{i=1}^{k} \frac{(|f_э - f_m| - 0{,}5)^2}{f_m}.$$

Критическое значение $\chi^2_{кр}$ определяется по таблице в соответствии с определенным числом степеней свободы и уровнем значимости.

Используется в двух вариантах:

а) для сопоставления эмпирического распределения с теоретическим; в этом случае проверяется нулевая гипотеза H_0 об отсутствии различий между теоретическим и эмпирическим распределением.

В качестве теоретических распределений могут выступать, например, равномерное или нормальное распределения.

В случае равномерного распределения теоретические частоты подсчитываются по формуле:

$$f_m = \frac{n}{k},$$

где *n* – количество наблюдений.

Число степеней свободы *v=k-1*.

В случае нормального распределения теоретическая частота подсчитывается по формуле:

$$f_m = \frac{nh}{\sigma} \cdot \varphi(u_i),$$

где *h* – шаг (разность между двумя соседними значениями признака),

$$u_i = \frac{x_i - \bar{x}}{\sigma},$$

φ(u) – нормированная дифференциальная функция Лапласа.

Число степеней свободы рассчитывают как *k-3*.

б) как расчет однородности двух и более независимых экспериментальных выборок; в этом случае проверяется гипотеза H_0 об отсутствии различий между эмпирическими (экспериментальными) распределениями.

в) для сравнения показателей внутри одной выборки; в этом случае проверяется гипотеза H_0: сравниваемые признаки не влияют друг на друга.

Исходные данные удобно представлять в виде таблицы сопряженности:

Разряды	Эмпирические частоты					
	первое распределение	второе распределение	…	j-ое распределение	…	c-ое распределение
1	n_{11}	n_{12}	…	n_{1j}	…	n_{1c}
2	n_{21}	n_{22}	…	n_{2j}	…	n_{2c}
…	…	…	…	…	…	…
k	n_{k1}	n_{r2}	…	n_{kj}	…	n_{kc}

Где n_{ij} – эмпирическая частота, соответствующая *i*-ому разряду *j*-ого распределения.

Для каждой ячейки таблицы, соответствующая теоретическая частота рассчитывается по формуле:

$$f_{mij} = \frac{\left(\sum_{j=1}^{c} n_{ij}\right) \cdot \left(\sum_{i=1}^{k} n_{ij}\right)}{\sum_{i=1}^{k}\sum_{j=1}^{c} n_{ij}}$$

или $$f_{mij} = \frac{(\text{сумма частот по строке})(\text{сумма частот по столбцу})}{\text{общее количество наблюдений}}.$$

Число степеней свободы определяется по формуле*: v=(k-1)(c-1).*

В случае, когда число переменных в двух сравниваемых выборках велико, можно использовать для вычисления $\chi^2_{эмп}$ следующую формулу:

$$\chi^2_{эмп} = 4\sum_{i=1}^{k} \frac{(f_{k1})^2}{f_{k1} + f_{k2}} - 2n,$$

где f_{k1}- частоты первого распределения,

f_{k2}- частоты второго распределения,

n- число элементов в каждой выборке.

Если число значений в выборках различно, то используют формулу:

$$\chi^2_{эмп} = \frac{(n_1 + n_2)^2}{n_1 n_2}\left(\sum_{i=1}^{k} \frac{(f_{k1})^2}{f_{k1} + f_{k2}} - \frac{n_1^2}{n_1 + n_2}\right).$$

Для применения критерия хи-квадрат необходимо соблюдать следующие условия: измерение может быть проведено по любой шкале; выборки должны быть случайными и независимыми; желательно, чтобы объем выборки был больше 20. С увеличением объема выборки точность критерия повышается; теоретическая частота для каждого выборочного интервала не должны быть меньше 5; сумма наблюдений по всем интервалам должна быть равна общему количеству наблюдений.

2. λ- критерий Колмогорова-Смирнова

Критерий λ предназначен для сопоставления двух распределений:

а) эмпирического распределения с теоретическим;

б) одного эмпирического распределения с другим эмпирическим распределением.

Для применения λ- критерия необходимо соблюдать следующие условия: измерение может быть проведено по шкале интервалов или отношений; выборки должны быть случайными и независимыми; желательно, чтобы суммарный объем двух выборок был большим или равным 50. С увеличением объема выборки точность критерия повышается; эмпирические данные должны допускать возможность упорядочения по возрастанию или убыванию какого-либо признака и обязательно отражать какое-то его однонаправленное изменение.

Возможны гипотезы:

H_0: Различия между двумя распределениями не достоверны.

H_1: Различия между двумя распределениями достоверны.

Эмпирическое значение критерия для сопоставления эмпирического распределения с теоретическим определяется по формуле:

$$d_{эмп} = \max \frac{\left|f_э^* - f_m^*\right|}{n},$$

где $f_э^*$ - накопленные эмпирические частоты,

f_m^* - накопленные теоретические частоты.

Критическое значение критерия определяется по таблице, при определенном уровне значимости и данном объеме выборки. Если число элементов выборки больше 100, то величина критических значений вычисляется по формуле:

$$d_{кр} = \begin{cases} 1{,}36/\sqrt{n} \\ 1{,}63/\sqrt{n} \end{cases}.$$

Эмпирическое значение критерия для сопоставления эмпирического распределения с другим эмпирическим распределением определяется по формуле:

$$\lambda_{эмп} = d_{\max} \cdot \sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}},$$

где n_1-количество наблюдений в первой выборке,

n_2- количество наблюдений во второй выборке,

d_{max}- наибольшая абсолютная величина разности между накопленными частостями по каждому разряду.

Уровень значимости соответствующий полученному значению λ определяется по таблице.

3. Критерий Фишера – φ

Критерий Фишера предназначен для сопоставления двух рядов выборочных значений по частоте встречаемости какого-либо признака. Этот критерий можно применять для оценки различий в любых двух выборках зависимых или независимых. С его помощью можно сравнивать показатели одной и той же выборки, измеренные в разных условиях.

Эмпирическое значение критерия подсчитываются по формуле:

$$\varphi_{эмп} = (\varphi_1 - \varphi_2) \cdot \sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}},$$

где φ_1 - величина, определяемая по таблице, соответствующая большей процентной доле,

φ_2 - величина, определяемая по таблице, соответствующая меньшей процентной доле.

Уровень значимости соответствующий полученному значению φ определяется по таблице. Для применения критерия Фишера φ необходимо соблюдать следующие условия: измерение может быть проведено по любой шкале; характеристики выборок могут быть любыми; нижняя граница – в одной из выборок может быть только два наблюдения, при этом во второй должно быть не менее 30 наблюдений (верхняя граница не определена); нижние границы двух выборок должны содержать не меньше 5 наблюдений в каждой.

Тема: Корреляционно-регрессионный анализ

Ключевые вопросы

Коэффициенты корреляции Пирсона, Спирмена, Кендалла, ассоциации, рангово-бисериального, бисериального, корреляционного отношения. Построение моделей регрессии различного вида. Анализ их надежности и устойчивости к изменениям внутри системы и внешней среды.

Основные определения и методы

Задача корреляционного анализа сводится к установлению направления и формы между варьирующими признаками, измерению тесноты, и, наконец, к проверке значимости коэффициентов корреляции.

Переменные *x* и *y* могут быть измерены в разных шкалах. Это обстоятельство определяет выбор соответствующего коэффициента линейной корреляции.

Тип шкалы		Мера связи
Переменная *x*	Переменная *y*	
Интервальная или отношений	Интервальная или отношений	Коэффициент Пирсона r_{xy}
Ранговая, интервальная или отношений	Ранговая, интервальная или отношений	Коэффициент Спирмена g_{xy}
Ранговая	Ранговая	Коэффициент τ Кендалла
Дихотомическая	Дихотомическая	Коэффициент ассоциации φ
Дихотомическая	Ранговая	Рангово-бисериальный $R_{гв}$

Тип шкалы		Мера связи
Дихотомическая	Интервальная или отношений	Бисериальный $R_{бис}$

Все коэффициенты по абсолютной величине не могут превосходить 1.

а) Коэффициент корреляции Пирсона вычисляется по формуле:

$$r_{xy} = \frac{\sum_i (x_i - \bar{x}) * (y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 * \sum_i (y_i - \bar{y})^2}} = \frac{n * \sum_i x_i y_i - \sum_i x_i * \sum_i y_i}{\sqrt{(n * \sum_i x_i^2 - (\sum_i x_i)^2) * (n * \sum_i y_i^2 - (\sum_i y_i)^2)}}$$

где x_i – значения, переменных принимаемые переменной x,

y_i - значения, переменных принимаемые переменной y,

$\bar{x}$ – средняя по x,

$\bar{y}$ - средняя по y.

Оценка значимости осуществляется при числе степеней свободы $k=n-2$.

б) Коэффициент корреляции рангов Спирмена вычисляется по формуле:

$$\rho = 1 - \frac{6 \cdot \sum_i (d_i)^2}{n \cdot (n^2 - 1)},$$

где n – количество ранжируемых признаков

d_i – разность между рангами по двум переменным для каждого испытуемого.

При наличии одинаковых рангов в числитель добавляются поправки на одинаковые ранги:

$$D_1 = \frac{n^3 - n}{12},$$

$$D_2 = \frac{k^3 - k}{12},$$

где n – число одинаковых рангов в первом столбце,

k – число одинаковых рангов во втором столбце.

По каждой группе одинаковых рангов вводится своя поправка.

Критически значения определяются при уровне значимости равном числу значений признака, по таблице критических значений ρ Спримена.

в) Коэффициент ассоциации φ вычисляется по формуле:

$$\varphi = \frac{p_{xy} - p_x \cdot p_y}{\sqrt{p_x \cdot (1 - p_x) \cdot p_y \cdot (1 - p_y)}},$$

где p_x – частота или доля признака, имеющего 1 по x, $(1-p_x)$ – частота или доля признака, имеющего 0 по x, p_y – частота или доля признака, имеющего 1 по y, $(1-p_y)$ – частота или доля признака, имеющего 0 по y, p_{xy} – частота или доля признака, имеющих 1 и по x и по y.

г) Коэффициент корреляции τ Кендалла вычисляется по формуле:

$$\tau = 1 - \frac{4 \cdot Q}{N \cdot (N - 1)},$$

где Q – число инверсий (подсчет инверсий осуществляется суммированием числа рангов второго признака меньше каждого из рангов второго признака, при условии, что ранги первого признака упорядочены по возрастанию);

N – число ранжируемых признаков.

д) Бисериальный коэффициент корреляции вычисляется по формуле:

$$R_{бисэмп} = \frac{\bar{x}_1 - \bar{x}_0}{\sigma_y} * \sqrt{\frac{n_1 * n_0}{N*(N-1)}},$$

где $\bar{x}_1$ - среднее по тем элементам переменной *y*, которым соответствует признак 1 в переменной *x*,

$\bar{x}_0$ - среднее по тем элементам переменной *y*, которым соответствует признак 0 в переменной *x*,

n_1 – число единиц в переменной *x*,

n_0 – число нулей в переменной *x*,

$N=n_1+n_0$,

σ_y - среднее квадратическое отклонение переменной *y*.

е) Рангово-бисернальный коэффициент корреляции вычисляется по формуле:

$$R_{rb\ эмп} = \frac{(\bar{x}_1 - \bar{x}_0)^2}{N},$$

где $\bar{x}_1$ - средний ранг по тем элементам переменной *y*, которым соответствует признак 1 в переменной *x*; $\bar{x}_0$ - средний ранг по тем элементам переменной *y*, которым соответствует признак 0 в переменной *x*; *N* – количество элементов в переменной *x*.

Взаимосвязь между переменными величинами может быть описана разными способами. Например, эту связь можно описать с помощью различных коэффициентов корреляции (линейных, частных, корреляционного отношения и т. п.). В то же время эту связь можно выразить по-другому: как зависимость между аргументом (величиной *X*) и функцией *Y*. В этом случае задача будет состоять в нахождении зависимости вида *X*=*F*(*Y*) или, напротив, в нахождении зависимости вида *X*=*F*(*Y*). При этом изменение функции в зависимости от изменений одного или нескольких аргументов называется регрессией.

Графическое выражение регрессионного уравнения называют линией регрессии. Линия регрессии выражает наилучшее предсказание зависимой переменной (*Y*) по независимым переменным (*X*). Эти независимые переменные, а их может быть много, носят название *предикторов*.

Регрессию выражают с помощью двух уравнений регрессии, которые в самом простом случае выглядят, как уравнения прямой, а именно так:

$$Y = a0 + a1 \cdot X$$
$$X = b0 + b1 \cdot Y$$

В уравнении 1 *Y* – зависимая переменная, а *X* – независимая переменная, *a*0 свободный член, а *a*1 – коэффициент регрессии, или угловой коэффициент, определяющий наклон линии регрессии по отношению к осям координат.

В уравнении 2 *X* – зависимая переменная, а *Y* – независимая переменная, *b*0 свободный член, а *b*1 – коэффициент регрессии, или угловой коэффициент, определяющий наклон линии регрессии по отношению к осям координат.

Линии регрессии пересекаются в точке *O* ($\bar{x}$, $\bar{y}$), с координатами, соответствующими средним арифметическим значениям корреляционно связанных между собой переменных *X* и *Y*. Линия *AB*, проходящая через точку *O*, соответствует линейной функциональной зависимости между переменными величинами *X* и *Y* равен r_{xy}= 1. При этом наблюдается такая закономерность: чем сильнее связь между *X* и *Y* , тем ближе обе линии регрессии к прямой *AB*. При отсутствии связи между *X* и *Y* линии регрессии оказываются под прямым углом по отношению друг к другу и в этом случае r_{xy}=0.

Количественное представление связи (зависимости) между *X* и *Y* (между *Y* и *X*) называется регрессионным анализом. Главная задача регрессионного анализа заключается , собственно гово-

ря, в нахождении коэффициентов $a0$, $b0$, $a1$ и $b1$ и определении уровня значимости полученных аналитических выражений, связывающих между собой переменные X и Y .

При этом коэффициенты регрессии $a1$ и $b1$ показывают, насколько в среднем величина одной переменной изменяется при изменении на единицу меры другой. Коэффициент регрессии $a1$ можно подсчитать по формуле:

$$a1 = r_{xy} \cdot \frac{S_y}{S_x}$$

а коэффициент $b1$:

$$b1 = r_{xy} \cdot \frac{S_x}{S_y}$$

где r_{yx} – коэффициент корреляции между переменными X и Y;

S_x – среднеквадратическое отклонение, подсчитанное для переменной X;

S_y – среднеквадратическое отклонение, подсчитанное для переменной Y.

Тема: Кластерный анализ

Ключевые вопросы

Постановка задачи кластерного анализа, построение дендрограммы, иерархические и неиерархические структуры, агломеративные и дивизитивные методы.

Основные определения и методы

Кластерный анализ объединяет различные процедуры, используемые для проведения классификации. В результате применения этих процедур исходная совокупность объектов разделяется на кластеры или группы (классы) схожих между собой объектов. Под кластером обычно понимают группу объектов, обладающую свойством плотности (плотность объектов внутри кластера выше, чем вне его), дисперсией, отделимостью от других кластеров, формой (например, кластер может иметь очертания гиперсферы или эллипсоида), размером.

Наиболее часто методы кластерного анализа используют в социологии, биологии, медицине, археологии, маркетинговых исследованиях, экономике.

Сложность задач кластерного анализа состоит в том, что реальные объекты являются многомерными, то есть описываются не одним, а несколькими параметрами, и объединение объектов в группы проводится в пространстве многих измерений, что весьма непросто. Кроме того, данные могут носить нечисловой характер. В целом методы кластеризации делятся на агломеративные (агломерат – скопление) и итеративные дивизивные (division – деление, разделение).

В агломеративных, или объединительных, методах происходит последовательное объединение наиболее близких объектов в один кластер. Процесс такого последовательного объединения можно показать на графике в виде дендограммы, или дерева объединения.

Исходными данными для анализа могут быть собственно объекты и их параметры, например, марки машин и их параметры: цена, время для разгона до скорости 60 миль/час и другие.

Данные для анализа могут быть представлены матрицей расстояний между объектами, в которой на пересечении строки с номером i и столбца с номером j записано расстояние между объектами. Переход от объектов к расстоянию между объектами – важный момент.

Расстояние между объектами – одна из мер сходства. Интуитивно понятно, что чем меньше расстояние между объектами, тем они более схожи.

Часто используют евклидову метрику, например, если объект описывается двумя параметрами, то он может быть изображён точкой на плоскости, а расстояние между объектами – это расстояние между точками, вычисленное по теореме Пифагора.

Если вы не будете возводить в квадраты координатные расстояния, а просто возьмёте их абсолютные значения и просуммируете, то получите так называемое манхэттенское расстояние, или «расстояние городских кварталов». Такое расстояние связано с перемещением человека по улицам

города, а не с движением по ровной местности (перемещаться можно только по линиям, параллельным осям координат в декартовой системе координат).

В реальных задачах евклидова метрика может оказаться вовсе не подходящей. В этих задачах используется понятие «меры сходства» объектов (расстояние – одна из мер сходства). Часто эта мера сходства является эмпирической, сходство измеряется непосредственно.

Важной мерой сходства, которая традиционно используется в социальных науках, являются статистические коэффициенты корреляции, например, коэффициент корреляции Пирсона.

Для бинарных данных часто просто вычисляют количество параметров, которые совпадают у объектов. Далее это число делят на общее число параметров и получают меру сходства. Меры, построенные таким образом, называют коэффициентами ассоциативности.

Важную роль играют иерархические и партиционные методы, причём последние применяются в подавляющем большинстве случаев. В иерархических методах каждое наблюдение образовывает сначала свой отдельный кластер. На первом шаге два соседних кластера объединяются в один; этот процесс может продолжаться до тех пор, пока не останутся только два кластера. расстояние между кластерами является средним значением всех расстояний между всеми возможными парами точек из обоих кластеров.

Дистанционные меры и меры подобия зависят от вида переменных, участвующих в анализе, то есть выбор меры зависит от типа переменной и шкалы, к которой она относится: интервальная переменная, частоты или бинарные (дихотомические) данные.

Тема: Факторный анализ

Ключевые вопросы

Выделение латентных переменных (факторов), интерпретация факторных нагрузок и факторных весов, моделирование значений наблюдаемых переменных на основе выделенных латентных факторов

Основные определения и методы

Факторный анализ – статистический метод, который используется при обработке больших массивов экспериментальных данных. Задачами факторного анализа являются: сокращение числа переменных (редукция данных) и определение структуры взаимосвязей между переменными, т.е. классификация переменных, поэтому факторный анализ используется как метод сокращения данных или как метод структурной классификации.

Важное отличие факторного анализа от всех описанных выше методов заключается в том, что его нельзя применять для обработки первичных, или, как говорят, «сырых», экспериментальных данных, т.е. полученных непосредственно при обследовании испытуемых. Материалом для факторного анализа служат корреляционные связи, а точнее – коэффициенты корреляции Пирсона, которые вычисляются между переменными (т.е. психологическими признаками), включенными в обследование. Иными словами, факторному анализу подвергают корреляционные матрицы, или, как их иначе называют, матрицы интеркорреляций. Наименования столбцов и строк в этих матрицах одинаковы, так как они представляют собой перечень переменных, включенных в анализ. По этой причине матрицы интеркорреляций всегда квадратные, т.е. число строк в них равно числу столбцов, симметричные, т.е. на симметричных местах относительно главной диагонали стоят одни и те же коэффициенты корреляции.

Элементы факторной матрицы называются «факторными нагрузками, или весами»; и они представляют собой коэффициенты корреляции данного фактора со всеми показателями, использованными в исследовании. Факторная матрица очень важна, поскольку она показывает, как изучаемые показатели связаны с каждым выделенным фактором. При этом факторный вес демонстрирует меру, или тесноту, этой связи.

2. МЕТОДИЧЕСКИЕ УКАЗАНИЯ К ПРАКТИЧЕСКИМ ЗАНЯТИЯМ

Практические занятия сопровождают лекционный курс дисциплины. Практические занятия должны проводиться в логичном единстве с теоретическим курсом, подкрепляя и уточняя понятийный аппарат.

Каждое практическое занятие начинается с теоретического опроса необходимого материала и проверки домашнего задания. Далее на конкретных примерах рассматриваются пути и способы применения тех математических методов, которые не требуют использования электронных вычислительных машин. При этом необходимо активизировать самостоятельную работу студентов. Задания и методические указания к ним выдаются студентам, каждый из которых выбирает оптимальный для себя темп работы. Преподавателю отводится роль консультанта и помощника. Задания, вызвавшие трудности у большинства студентов, разбираются на доске.

В конце занятия выдается домашнее задание, состоящее из теоретических вопросов, уяснение которых необходимо для следующего занятия и практических заданий по пройденному материалу.

При выполнении домашнего задания решать задачи удобнее поэтапно, в той последовательности, в какой эти задания сформулированы. В этом случае при возникновении трудностей будет легче обратиться к анализу тех тем, которые изложены в лекции и задач, разобранных на практическом занятии.

После выполнения практической части задания следует найти ответы на теоретические вопросы, заданные преподавателем и, таким образом, подготовится к осознанному восприятию следующего материала.

Активная, регулярная самостоятельная работа над домашним заданием – путь к успешному усвоению дисциплины.

Тема: Основы измерения и количественного описания данных

Типовые задания

1. Студенты группы из 30 человек написали контрольную работу. Каждый студент получил определённое количество баллов:75, 145, 150, 180, 125, 150, 150, 165, 95, 135, 130, 70, 130, 105, 135, 135, 100, 160, 60, 85, 120, 60, 145, 150, 135, 132, 140, 65, 170,155.

Требуется:

1. Построить сгруппированный вариационный ряд
2. Построить эмпирическую функцию распределения
3. Построить гистограмму и полигон относительных частот
4. Найти выборочные точечные характеристики: выборочную среднюю, выборочную дисперсию, эксцесс, асимметрию, моду.
5. Выдвинуть гипотезу относительно близости распределения к нормальному.

2. Как известно, почерк человека, в том числе наклон букв, тесно связан с его характером. Низкий наклон ($30\text{-}40^0$) свидетельствует о вспыльчивости и возбудимости человека, излишней прямоте и торопливости в поступках; наклон в $40\text{-}50^0$ характеризует гармоничное развитие натуры; наклон $50\text{-}90^0$ свидетельствует о самообладании, узком диапазоне увлечений. Среди студентов АмГУ выборочно был исследован почерк 50 человек. Оказалось, что у 30% присутствующих низкий наклон, у 50%-наклон $40\text{-}50^0$ и у 20% наклон $50\text{-}90^0$. Найти распределение частот. Относительных частот, построить полигон и гистограмму. Вычислить числовые характеристики сгруппированного ряда.

3. Курс «Социальная психология» прослушало 50 человек. Полученные студентами на экзамене оценки представляют собой следующий набор цифр: 3,4,5,4,3,3,5,…(составить самостоятельно). Построить вариационный статистический ряд. Вычислить моду, медиану, выборочную среднюю.

4. Результаты выборочных наблюдений над непрерывной случайной величиной X приведены ниже в виде интервалов одинаковой длины и соответствующих им частот:

$x_i - x_{i+1}$	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50
n_i	12	18	45	15	10

Построить: а) гистограмму частот и б) гистограмму относительных частот случайной величины X.

Тема: Параметрические критерии проверки статистических гипотез

Типовые задания

1. Построить частотное распределение, подсчитать стандартное отклонение для следующих рядов значений: (результаты тестирования интеллекта по тесту Векслера 2-х групп испытуемых, численностью 50 человек):

Результаты тестирования 1-ой группы испытуемых:

85,93,93,99,101,105,109,110,111,115,115,116,116,117,117,117,118,119,121,121,122,124,124,124,124,125,125,125,127,127,127,127,127,128,130,131,132,132,133,134,134,135,138,138,140,143,144,146,150,158.

Результаты тестирования 2-й группы:

70,75,76,78,78,81,82,82,83,84,84,84,85,86,86,86,89,89,90,91,91,91,91,92,92,93,93,95,95,95,96, 96,98,98,100,101,103,103,103,105,108,110, 115,118,119,123,124,125,127,129.

Подсчитать t- критерий Стьюдента для двух выборок.

Проверить результаты на статистическую значимость.

2. Подсчитайте критерий Фишера для следующих результатов измерения тревожности 2-х различных групп

x	90	29	39	79	88	88	72	76	91	78
y	41	49	56	64	72	58	59	62	68	48

Тема: Непараметрические критерии проверки статистических гипотез

Типовые задания

1. Подсчитайте критерий Крускала -Уоллиса для следующих данных

23	45	34	21
20	12	24	22
34	34	25	26
35	11	40	27

2. Подсчитайте критерий Манна-Уитни для следующих данных:

x	y
6	
	8
25	
25	
30	
	31
	32
38	
39	
	41
	41
43	

x	y
44	
	45
	46
	50
	55

3. Две группы испытуемых решали техническую задачу. Показателем успешности было время решения задачи. Испытуемые 1 группы получали за выполнение денежное вознаграждение, испытуемые 2 группы – нет. Психолога интересует вопрос: влияет ли денежное вознаграждение на успешность решения задачи?

Результаты получены следующие (в секундах):

1 группа: 39, 38, 46, 25, 25, 30, 43.

2 группа: 46, 50, 45, 32, 41, 41, 31, 55.

Подсчитайте критерий Манна – Уитни.

Тема: Критерии согласия

Типовые задания

1. Психолог решает задачу: будет ли удовлетворенность работой на данном предприятии распределена равномерно по следующим альтернативам (градациям):1 - Работой вполне доволен; 2 - Скорее доволен, чем не доволен; 3-Трудно сказать, не знаю, безразлично; 4-Скорее недоволен, чем доволен; 5 - Совершенно недоволен работой. Для решения этой задачи производится опрос случайной выборки из 65 респондентов (испытуемых) об удовлетворенности работой: «В какой степени Вас устраивает Ваша теперешняя работа?», причем ответы должны даваться согласно вышеозначенным альтернативам. Полученные ответы представлены в таблице.

Альтернативы	1	2	3	4	5
Частота выбора	8	22	14	9	12

2. По данным аттестационной комиссии из 53 студентов энергетического факультета 23 справились на отлично, а из 45 студентов экономического факультета с тем же заданием справились 23 человека. Можно ли утверждать, что различия в успешности решения аттестационной работы экономистами и энергетиками достоверны.

Тема: Корреляционно-регрессионный анализ

Типовые задания

1. Был проведен опрос о количестве времени (в часах), которое тратит каждый студент на изучение учебной литературы и на просмотр художественных фильмов. Можно ли сделать вывод о существовании связи между 2-мя этими переменными?

№	Время на изучение учебной литературы	Время на просмотр фильмов
1	7	0
2	8	4
3	5	10
4	1	7
5	15	0
6	5	1
7	2	10
8	4	6
9	1	9
10	1	8

2. Зависимость между величинами выражается в виде экспериментально полученной таблицы. Определить коэффициент корреляции Пирсона. Сделать выводы.

X	0,5	1	1,5	2	2,5	3
У	0,01	0,11	0,35	0,6	1,58	2,31

3. Подсчитайте корреляцию между физической привлекательностью студенток и их академическими достижениями.

№	Ранг по красоте	Ранг по академическим достижениям
1	3	7.5
2	2	2
3	6	9
4	8	4
5	4	4
6	10	10
7	7	6
8	1	1
9	9	7.5
10	5	4

4. Существует ли связь между лидерством и дружелюбностью? Исследователи отмечали наличие или отсутствие у человека лидерской позиции в группе, одновременно относя его либо к дружелюбным, либо к недружелюбным людям. Результаты показаны в таблице

Лидерская позиция	Дружелюбные	Недружелюбные	Всего
Лидер	2	4	6
Не явл. лидером	10	4	14
Всего	12	8	20

Что можно сказать о связи между этими переменными? Рассчитайте коэффициент ассоциации.

Тема: Кластерный анализ

Типовые задания

1. Используя не менее двух методов кластер – процедур провести классификацию точек (3; 4) , (-3; 8), (-2;-6), (5; 7), (6; 0), (8; 4), (-3; 9), (2;-6), (5; -7), (5; 0). В качестве расстояний выбрать евклидово расстояние.

Тема: Факторный анализ

Типовые задания

1. Определите удовлетворенность, какой стороной жизни преобладает у разных респондентов. Пусть в эксперименте участвовало 10 человек, которые отвечали на вопросы, представленные в таблице

Возраст	Образование	Стаж непрерывной работы	Работа по специальности	Зарплата, тыс.руб	Кол-во человек в семье	Жилплощадь	Наличие хобби	Посещение учреждение вне работы

Образование: 2 – высшее, 1 – специальное, 0 – среднее. Работа по специальности: 1 – да, 0 – нет. Хобби: 1 – да, 0 – нет. Посещение учреждение вне работы: 3 – часто, 2 – иногда, 1 – нет.

Проведите стандартизацию данных и определите оптимальное количество факторов, рассчитайте матрицы факторных нагрузок и факторных весов, сделайте вывод.

3. МЕТОДИЧЕСКИЕ УКАЗАНИЯ К ЛАБОРАТОРНЫМ ЗАНЯТИЯМ

Лабораторные работы предназначены для получения практических навыков студентами при изучении дисциплины. Предлагаемые задания охватывают основные вопросы, рассматриваемые в рамках данного курса.

К выполнению лабораторной работы следует приступать после ознакомления с теоретической частью соответствующей темы. Результаты всех лабораторных работ необходимо представить в письменном виде.

Тема: Основы измерения и количественного описания данных

Типовые задания

Задание 1. Даны наблюдавшиеся значения некоторой случайной величины:

34	25	29	34	12	28	13	28	28	17
23	31	32	23	16	22	34	22	25	28
24	24	25	28	26	19	29	21	30	18
30	30	21	30	19	20	30	34	20	36
28	36	27	17	27	26	26	19	29	24
37	28	31	25	23	33	35	31	22	30
25	26	22							

Требуется:

1) построить сгруппированный статистический ряд;

2) построить гистограмму и полигон относительных частот;

3) найти выборочные точечные характеристики: среднюю, дисперсию, среднее квадратическое отклонение.

Задание 2. Даны наблюдаемые значения некоторой случайной величины.

34	25	29	34	12	28	13	28	28	17
23	31	32	23	16	22	34	22	25	28
24	24	25	28	26	19	29	21	30	18
30	30	21	30	19	20	30	34	20	36
28	36	27	17	27	26	26	19	29	24
37	28	31	25	23	33	35	31	22	30
25	26	22							

Требуется:

1) найти выборочные точечные характеристики: асимметрию, эксцесс, моду, коэффициент вариации;

2) проверить гипотезу относительно близости распределения к нормальному.

Тема: Параметрические критерии проверки статистических гипотез

Типовые задания

Задание 1. Два университета (А и В) готовят специалистов аналогичных специальностей. Министерство образования решило проверить качество подготовки в обоих университетах, организовав для этого объемный тестовый экзамен для студентов пятого курса. Отобранные случайным образом студенты показали следующие результаты:

A: 41, 50, 35, 45, 53, 30, 57, 20, 50, 44, 36, 48, 55, 28, 40, 50;

B: 40, 57, 52, 38, 20, 25, 47, 52, 48, 55, 48, 53, 39, 49, 46, 45, 55, 43, 51, 55, 40.

Можно ли утверждать при уровне значимости $\alpha = 0{,}05$, что один из университетов обеспечивает лучшую подготовку.

Задание 2. В условиях предыдущей задачи определите, есть ли основания считать, что разброс оценок у студентов одного университета больше чем у другого.

Тема: Непараметрические критерии проверки статистических гипотез

Типовые задания

Задание 1. Можно ли утверждать, что студенты-психологи превосходят студентов-физиков по уровню невербального интеллекта. Данные, полученные с помощью методики Д. Векслера, приведены в таблице.

Использовать критерий Манна-Уитни. Сформулировать гипотезы.

Данные внести в один столбец. Показатели физиков обозначить во втором столбце цифрой 1, психологов – 2.

Студенты-физики		Студенты - психологи	
Код имени испытуемого	Показатель невербального интеллекта	Код имени испытуемого	Показатель невербального интеллекта
1. И.А.	111	1. Н.Т.	113
2. К.А.	104	2. ОБ.	107
3. К.Е.	107	3. Е.В.	123
4. ПА.	90	4. Ф.О.	122
5. С.А.	115	5. И.Н.	117
6. Ст.А.	107	6. И.Ч.	112
7. Т.А.	106	7. И.В.	105
8. ФА.	107	8. К.О.	108
9. Ч.И.	95	9. Р.Р.	111
10. Ц.А.	116	10. Р.И.	114
11. См.А.	127	11. О.К.	102
12. К.Ан.	115	12. Н.К.	104
13. Б.Л.	102		
14. Ф.В.	99		

Задание 2. Психолог проводит с младшими школьниками коррекционную работу по формированию навыков внимания, используя для оценки результатов корректурную пробу. Задача состоит в том, чтобы определить, будет ли уменьшаться количество ошибок внимания у младших школьников после специальных коррекционных упражнений? Для решения этой задачи психолог у 19 детей определяет количество ошибок при выполнении корректурной пробы до и после коррекционных упражнений. В таблице приведены соответствующие экспериментальные данные.

№	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
До	24	12	42	30	40	55	50	52	50	22	33	78	79	25	28	16	17	12	25
После	22	12	41	31	32	44	50	32	21	34	56	78	23	22	12	16	17	18	25

Для решения задачи использовать критерий Т Вилкоксона

Тема: Критерии согласия

Типовые задания

Задание 1. Используя критерий Пирсона, при уровне значимости 0,05, проверить, согласуется ли гипотеза о нормальном распределении генеральной совокупности X с эмпирическим распределением выборки объема n=200, представленным в таблице

x_i	5	7	9	11	13	15	17	19	21
n_i	15	26	25	30	26	21	24	20	13

Задание 2. Используя критерий Пирсона, при уровне значимости 0,05, проверить, согласуется ли гипотеза о нормальном распределении генеральной совокупности X с заданным эмпирическим распределением.

x_i; x_{i+1}	-20;-10	-10;0	0;10	10;20	20;30	30;40	40;50
n_i	20	47	80	89	40	16	8

Тема: Корреляционно-регрессионный анализ

Типовые задания

Задание 1. В плане комплексного исследования личности у студентов психологического факультета определялся социометрический статус на курсе (У) и в своей учебной группе (Х). Результаты представлены в таблице. Имеется ли связь между Х и У.

Х	1	-9	22	-11	-6	-0,5	5	9	0	2	11	15	-8	0
У	48	-9	100	-1	14	14	14	11	-28	46	42	27	19	83

Задание 2. Супругам предложили проранжировать семь личностных черт, имеющих определяющее значение для семейного благополучия. Задача заключается в том, чтобы определить, в какой степени совпадают оценки супругов по отношению к ранжированию качеств. Данные представлены в таблице.

Черты личности	муж	жена
ответственность	7	1
общительность	1	5
сдержанность	3	7
выносливость	2	6
жизнерадостность	5	4
терпеливость	4	3
решительность	6	2

Задание 3. Даны результаты эксперимента

X	7,5	7	8,3	8,4	6,9	7,7	8,1	7,6	7,9	8,2
Y	26	27	22	21	27	25	21	21	23	22

Требуется:

1) в предположении, что между x и y существует линейная зависимость, определить ее эмпирическое уравнение;

2) в предположении, что между x и y существует квадратичная зависимость, определить ее эмпирическое уравнение;

3) найти сумму квадратов отклонений для найденных зависимостей сравнить качество приближений.

Задание 4. Дана выборка

X	51	50	33	40	42	51	52	51	55	36	53
У	70	56	31	48	73	72	40	66	76	34	63

По заданной выборке:

1) найти уравнение прямой линии регрессии У на Х;

2) оценить тесноту линейной связи, вычислив выборочный коэффициент корреляции;

3) проверить гипотезу о значимости коэффициента корреляции при уровне значимости 0,1.

Тема: Кластерный анализ

Типовые задания

Задание 1. Объединить 8 фирм, занимающихся производством и установкой окон в несколько схожих совокупностей путем построения дендрограммы

Фирма	Возраст фирмы	Кол-во сотруд-ников	Количество установленных окон за месяц	Средняя цена окна (тыс.руб)	Количество филиалов	Скидки(%)	Популярность (0-нет, 1да)
1	11	200	30	17	20	25	1
2	3	14	12	13,2	2	33	1
3	3	50	8	14,8	5	30	1
4	5	80	14	15	10	10	1
5	8	35	17	14,2	12	15	1
6	2	10	5	11,5	1	5	0
7	7	20	3	13,5	1	0	0
8	3	20	1	12,9	1	0	0

Задание 2. Провести классификации пяти точек: (1,2), (4, 3), (-1, -1), (-1, 0), (-3, 3).

Тема: Факторный анализ

Типовые задания

Задание 1. Используя данные, полученные у случайной выборки учащихся первого класса (x1 – вес тела, x2 – рост, x3 – количество слов, читаемых в минуту, x4 – оценка по чтению за год, x5 – длина руки, x6 – количество книг прочитанных за год, x7 – количество выученных стихотворений) провести факторный анализ данных. Определить факторные нагрузки переменных, собственные значения факторов. Данные представлены в таблице. Сделать выводы.

	X1	X2	X3	X4	X5	X6	X7
1	20	120,00	22	3	10	12	12
2	31	122,00	26	3	13	13	13
3	22	123,00	45	5	12	23	19
4	37	126,00	38	5	14	19	19
5	42	127,00	44	5	14	24	19
6	45	123,00	32	4	15	17	17
7	38	123,00	31	4	13	16	16
8	44	124,00	33	4	13	18	18
9	50	126,00	39	5	14	19	19
10	20	127,00	35	4	10	18	18
11	26	125,00	27	3	11	11	11
12	30	129,00	28	3	12	13	13
13	50	124,00	26	3	13	13	13
14	26	125,00	27	3	12	12	12
15	32	124,00	36	4	13	18	18
16	28	125,00	42	5	11	19	19
17	20	122,00	35	4	10	18	18
18	25	127,00	34	4	12	16	16
19	20	122,00	26	3	11	13	13
20	30	123,00	15	6	15	15	15

4. МЕТОДИЧЕСКИЕ УКАЗАНИЯ ДЛЯ САМОСТОЯТЕЛЬНОЙ РАБОТЫ СТУДЕНТОВ

Самостоятельная работа студентов предназначена для углубления сформированных знаний, умений, навыков. Самостоятельная работа развивает мышление, позволяет выявить причинно-следственные связи в изученном материале, решить теоретические и практические задачи. Самостоятельная работа студентов проводится с целью: систематизации и закрепления полученных теоретических знаний и практических умений студентов; углубления и расширения теоретических знаний; формирования умений использовать справочную документацию и специальную литературу; развития познавательных способностей и активности студентов: творческой инициативы, самостоятельности, ответственности и организованности.

Виды и формы самостоятельных работ по дисциплине «Методы прикладной статистики для социологов».

Студентами практикуется два вида самостоятельной работы:

- аудиторная;
- внеаудиторная.

Аудиторная самостоятельная работа по дисциплине выполняется на учебных занятиях под непосредственным руководством преподавателя и по его заданию. В этом случае студенты обеспечиваются преподавателем необходимой учебной литературой, дидактическим материалом, в т. ч. методическими пособиями и методическими разработками.

Внеаудиторная самостоятельная работа выполняется студентом по заданию преподавателя, но без его непосредственного участия. Видами заданий для внеаудиторной самостоятельной работы могут быть:

- для овладения знаниями: чтение текста (учебника, методической литературы); составления плана текста; графическое изображение структуры текста, графическое изображение последовательности выполнения работы, выполнение лабораторных работ; конспектирование текста; выписки из текста; работа со словарями и справочниками; ознакомление с нормативными документами; учебно-исследовательская работа; использование компьютерной техники, интернета и др.;
- для закрепления систематизации знаний: работа с конспектом лекции (обработки текста); повторная работа над учебным материалом (учебника, первоисточника, дополнительной литературы); составление плана выполнения работы в соответствие с планом, предложенным преподавателем; изучение ГОСТов; ответы на контрольные вопросы; тестирование, выполнение упражнений и лабораторных работ;
- для формирования умений: решение задач и упражнений по образцу; решение вариативных задач и упражнений; выполнение чертежей, схем.

Основное содержание самостоятельной работы составляет выполнение домашних заданий, подготовка к самостоятельным работам, выполнение индивидуального комплексного задания, подготовка к экзамену.

Прежде чем приступать к выполнению лабораторной работы, необходимо ознакомиться с содержанием теоретических вопросов по представленному списку литературы и по лекциям.

Каждый учебный семестр заканчивается аттестационными испытаниями: зачетно - экзаменационной сессией.

Подготовка к экзаменационной сессии и сдача зачетов и экзаменов является ответственейшим периодом в работе студента. Серьезно подготовиться к сессии и успешно сдать все экзамены - долг каждого студента. Рекомендуется так организовать свою учебу, чтобы перед первым днем начала сессии были сданы и защищены все лабораторные работы, сданы все зачеты, выполнены другие работы, предусмотренные графиком учебного процесса.

Основное в подготовке к сессии - это повторение всего материала, курса или предмета, по которому необходимо сдавать зачет или экзамен. Только тот успевает, кто хорошо усвоил учебный материал.

СОДЕРЖАНИЕ

Татьяна Евгеньевна Гришкина,
старший преподаватель кафедры общей математики и информатики АмГУ