ФГАОУВПО «Казанский (Приволжский) федеральный университет»

Е.А.УТКИНА

ЭЛЕМЕНТЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ В СОЦИОЛОГИИ

Казань 2012

Печатается по решению учебно-методической комиссии института математики и механики им.Лобачевского Казанского федерального университета

УДК 303.4

Уткина Е.А. Элементы математической статистики в социологии. Казань: КФУ, 2012.- 50 с.

Пособие предназначено для студентов 1,2 курсов. В нем изложены необходимые для студентов нематематических специальностей разделы математической статистики, а также приведены варианты заданий для практических занятий, самостоятельной работы и контрольных работ.

Рецензенты: Е.А.Широкова, доктор физико-математических наук, доцент (КФУ)
А.Ф.Галимянов, кандидат физико-математических наук, доцент (КФУ),

© Уткина Е.А. 2012.

§1.Вариационные ряды.

Рассмотрим основные понятия, применяющиеся в математической статистике.

Определение. Объектом наблюдения называется совокупность предметов или явлений, обладающих каким-либо общим свойством или признаком качественного или количественного характера.

Объекты статистического наблюдения состоят из элементов, которые принято называть единицами наблюдения.

Результатом статистического наблюдения является числовая информация (данные). *С*ведения о том, какие значения принял признак, интересующий исследователя в статистической совокупности, называются статистическими данными. Признаки бывают качественными и количественными.

Признак называется количественным, если его значения выражаются числами.

Признак называется качественным, если он характеризуется некоторым состоянием или свойством элементов совокупности**.**

Определение. Генеральной называется статистическая совокупность, в которой исследованию подлежат все элементы совокупности (сплошное наблюдение).

Определение. Выборочной совокупностью или выборкой называется часть элементов генеральной совокупности, подлежащая исследованию**.** Она строится из генеральной совокупности с помощью случайного выбора, так чтобы каждый элемент выборки имел равные шансы быть отобранным.

Определение. Вариантами называются значения признака, которые при переходе от одного элемента совокупности к другому изменяются или варьируют. Они обычно обозначаются малыми латинскими буквами *x, y, z***.**

Итак, пусть в генеральной совокупности исследуется некоторый количественный признак. Из нее извлекается

случайным образом выборка объема n (это означает, что число элементов выборки равно n). Каждое значение в выборке x_i, $i = 1,...,k$, называется вариантой.

Число наблюдений значения x_i в выборке обычно обозначают n_i и называют частотами. Относительной частотой или частостью w_i называют отношение частоты n_i к объему выборки: $w_i = n_i / n$.

Вариационным рядом называется таблица следующего вида:

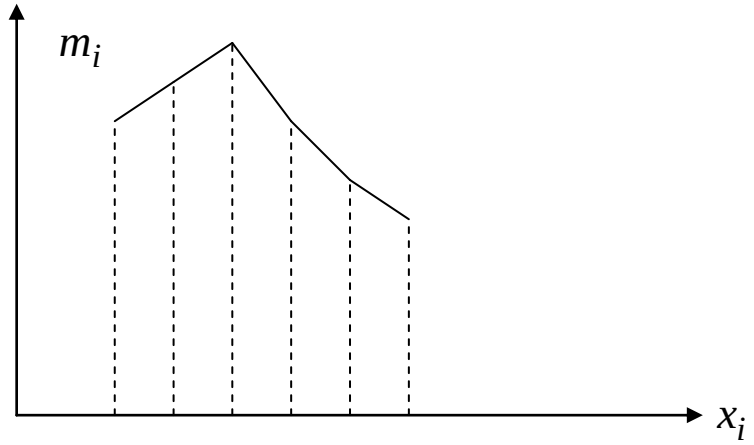
x_1	x_2	…	x_k
n_1	n_2	…	n_k

В ней варианты x_i расположены в порядке возрастания. Такая таблица называется еще дискретным вариационным рядом.

Графическое представление дискретного вариационного ряда.

1. Эмпирическая функция распределения $F_э(x) = n_x / n$, где n_x - число вариант, меньших x. Соединим расположенные рядом точки $(x_i, F_э(x_i))$ отрезками прямых, получим кумуляту.

2. Полигон распределения частот или частостей. Для этого строят точки с координатами (x_i, m_i) и соседние точки соединяют отрезками прямых.



Интервальный вариационный ряд

Если значения изучаемого признака сколь угодно мало отличаются друг от друга, строят вариационные ряды, называемые интервальными. Их общий вид приведен в таблице

Интервал	$x_0 - x_1$	$x_1 - x_2$	…	$x_{k-1} - x_k$
Частота	n_1	n_2	…	n_k

Здесь частота – это число вариант, попавших в соответствующий интервал.

Если все интервалы имеют одинаковую длину, то такие интервалы называются равновеликими. Во всех остальных случаях они называются неравновеликими. Часто первый и последний интервалы не имеют одной границы (соответственно нижней или верхней). Например, 1-й интервал может быть задан как «до 300», 2-й — «300-310», ... предпоследний — «390-400», последний — «400 и более» В этом случае считают длину 1-го интервала равной длине 2-го интервала, а длину последнего интервала – равной длине предпоследнего. При построении интервального вариационного ряда зачастую возникает необходимость выбрать величину интервалов (интервальную

разность). Для ряда с равной шириной интервалов применяют формулу Стэрджесса

$$k = \frac{x_{\max} - x_{\min}}{1 + 3{,}322 \lg n},$$

либо таблицу

Объем выборки n	Число классов	Объем выборки n	Число классов
6-11	4	188-377	9
12-22	5	378-755	10
23-46	6	756-1515	11
47-93	7	1516-3050	12
94-187	8		

Графическое представление интервального вариационного ряда

Графически интервальные вариационные ряды можно представить несколькими способами.

1. Гистограмма. Она представляет собой ступенчатую фигуру, образованную прямоугольниками. В их основаниях лежат интервалы (x_{i-1}, x_i), их высоты являются либо частотами (n_i), либо частостями ($w_i = n_i / n$). Во втором случае площадь i-го прямоугольника равна w_i, а всей гистограммы – 1. Если соединить середины верхних сторон прямоугольников, построим полигон.

2. Кумулянта. При ее построении по оси абсцисс откладывают значения признака (варианты), а по оси ординат — значения накопленных частоты или частостей. Строятся точки на пересечении значений признака (вариантов) и соответствующих им накопленных частот (частостей). Затем они соединяются отрезками ломаной. Эта ломаная (кривая) называется кумулятой или кумулятивной кривой. Абсциссами ее точек являются верхние границы интервалов. Ординатами являются накопленные частоты (частости) соответствующих интервалов. Иногда добавляют еще одну точку, абсциссой

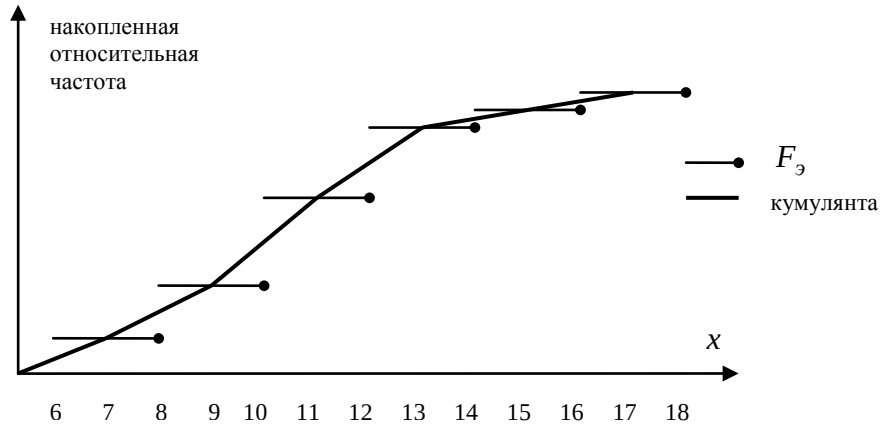
которой является нижняя граница первого интервала, а ордината равна нулю.

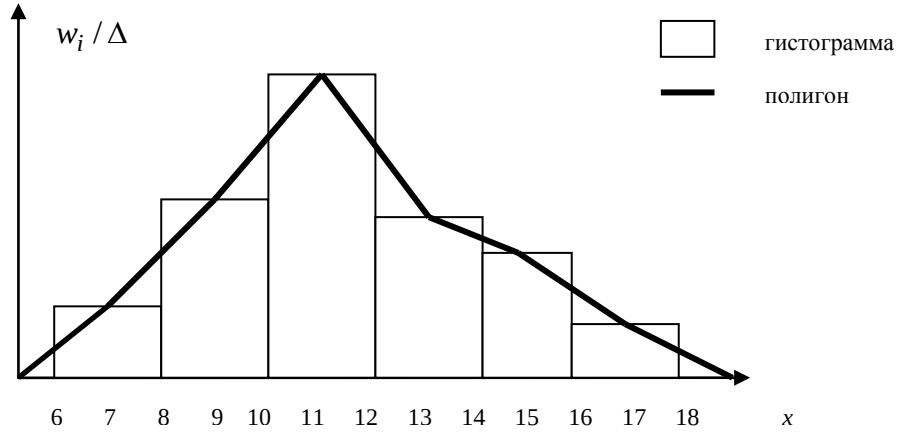
Пример 1. Получены данные об обращениях клиентов в автомойку

Интервал времени	До 8	8-10	10-12	12-14	14-16	Св. 16
Число клиентов	10	22	35	17	11	5

Построить функцию распределения эмпирическую, кумуляту, гистограмму, полигон. Длину первого и последнего открытых интервалов считаем равными соответственно длинам 2-го и предпоследнего интервалов. Для них $\Delta = 2$.

Интервал	Середина интервала	Частота n_i	$w_i = n_i / n$	w_i / Δ	Накопленная относительная частота
6-8	7	10	0,1	0,02	0,1
8-10	9	22	0,22	0,044	0,32
10-12	11	35	0,35	0,07	0,67
12-14	13	17	0,17	0,034	0,84
14-16	15	11	0,11	0,022	0,95
16-18	17	5	0,05	0,01	1
Сумма	-	100	-	-	-





Задача 1. Получены данные об обращениях клиентов в автомойку

Интервал времени	До 8	8-10	10-12	12-14	14-16	Св. 16
Число клиентов	11	21	35	17	7	7

Построить эмпирическую функцию распределения, кумуляту, гистограммы, полигон.

§2. Сводные характеристики выборки.

Чтобы исследовать параметр теоретического распределение генеральной совокупности, по результатам выборки вычисляется его точечная оценка. В силу случайности результатов выборки, полученная оценка является случайной величиной.

Оценка называется несмещенной, если математическое ожидание вычисленной оценки равно теоретическому значению параметра генеральной совокупности для любого объема выборки. В противном случае оценку называют смещенной.

Если наблюдаются варианты $x_1,...,x_k$ с соответствующими частотами $n_1,...,n_k$, то выборочная средняя $\bar{x}_в = \sum_{i=1}^{k} x_i n_i / n$ является несмещенной оценкой генеральной средней $\overline{x_c}$, поскольку $M(\overline{x_в}) = \overline{x_c}$.

Смещенной оценкой генеральной дисперсии $D_г$ является выборочная дисперсия $D_в = \sum_{i-1}^{k} n_i (x_i - \overline{x_в})^2 / n = \sum_{i=1}^{k} n_i x_i^2 / n - (\overline{x_в})^2$, поскольку $M(D_в) = (n-1)D_г/n$. Чтобы вычислить несмещенную оценку генеральной дисперсии вводят поправочный коэффициент. С учетом этого исправленная выборочная дисперсия $s^2 = nD_в/(n-1)$.

Пусть варианты являются равноотстоящими, то есть разность между любыми соседними вариантами равна постоянной Δ. Перейдем к условным вариантам $u_i = (x_i - c)/\Delta$. Здесь с- ложный нуль (это варианта, расположенная в середине вариационного ряда; если их две, то выбирают из них варианту с наибольшей частотой).

Пусть варианты неравноотстоящие. Разобьем весь вариационный ряд на 8-10 равновеликих интервалов длины Δ,

возьмем затем середины интервалов и получим случай равноотстоящих вариант.

Условным эмпирическим моментом порядка р называется величина $M_p = \sum_{i=1}^{k} n_i u_i^p / n$. С учетом этого $\overline{x_в} = M_1 \Delta + c$, $D_в = (M_2 - M_1^2)\Delta^2$. Если вместо интервала рассматривать его середину, возникает систематическая ошибка при расчете выборочной дисперсии. Для ее уменьшения, вводят поправку Шеппарда и находят уточненное значение выборочной дисперсии: $D_в^* = D_в - \Delta^2 /12.$.

Пример 2. Получены данные о числе пассажиров, перевозимых автобусом 43 маршрута, по часам.

Время/час	До 7	7-10	10-13	13-16	16-19	Св.19
Число пассажиров	10	22	35	17	11	5

Найти сводные характеристики выборки.

Первый и последний интервалы не имеют нижней и верхней границы соответственно. Поэтому будем считать длины 1-го и 2-го интервала, а также последнего и предпоследнего равными. Заполним таблицу.

Интервал	Середина интервала x_i	Частота n_i	u_i	$n_i u_i$	$n_i u_i^2 = n_i u_i \times u_i$
4-7	5,5	10	-2	-20	40
7-10	8,5	22	-1	-22	22
10-13	11,5	35	0	0	0
13-16	14,5	17	1	17	17
16-19	17,5	11	2	22	44

19-22	20,5	5	3	15	45
Сумма	-	100	-	12	168

Между любыми соседними вариантами разность x_i постоянна и равна Δ=3. В середине вариационного ряда расположены 11,5 и 14,5. Ложный нуль с=11,5, поскольку частота 35>17. Вычислим $M_1 = \sum_{k=1}^{6} n_i u_i / 100 = 0{,}12$. Тогда выборочная средняя $\overline{x_в} = M_1 \Delta + c = = 0{,}12 \cdot 3 + 11{,}5 = 11{,}86$, $M_2 = \sum_{k=1}^{6} n_i u_i^2 / 100 = 1{,}68$, Тогда выборочная дисперсия $D_в = (M_2 - M_1^2)\Delta^2 = (1{,}68 - 0{,}12^2) \times 3^2 = 14{,}99$. Уточненное значение выборочной дисперсии $D_в^* = D_в - \Delta^2 / 12 = 14{,}99 - 0{,}75 \approx 14{,}24.$

Задача 2. Получены данные о числе пассажиров, перевозимых автобусом 43 маршрута, по часам.

Время/час	До 7	7-10	10-13	13-16	16-19	Св.19
Число пассажиров	11	21	34	18	8	8

Найти сводные характеристики выборки.

Замечание. Статистические функции пакета Excel позволяют определить сводные характеристики выборки.

Функция СРЗНАЧ вычисляет $\overline{x_в}$.

Функция СТАНДОТКЛОН вычисляет генеральное стандартное отклонение $\sqrt{D_г}$ по выборке.

Функции ДИСП и ДИСПР вычисляют соответственно s^2 и $D_в$.

§3. Мода и медиана

Определение. Модой называется значение, появляющееся чаще всего у единиц совокупности.

Пример 3. Определить моду совокупности 4; 4; 3; 4; 2; 1.

Мода равна 4.

Задача 3. Определить моду для совокупности 2; 6; 7; 6; 6; 2; 5.

Пример 4. Определить моду вариационного ряда, приведенного в таблице

Значение	2	3	4	6
Частота	11	17	21	12

Мода равна 4

Задача 4. Определить моду для вариационного ряда.

Значение	11	16	17	19
Частота	19	25	17	12

Мода M_0 интервального вариационного ряда с равновеликими интервалами определяется по формуле:

$M_0 = x_{\min} + h_i(n_{M_0} - n_{M_0-1})/((n_{Mo} - n_{Mo-1}) + (n_{Mo} - n_{Mo+1}))$, где $x_{\min}$ - нижняя граница модельного интервала; n_{M_0} - частота модального интервала - интервала, содержащего моду; n_{M_0-1} - частота интервала, предшествующего модальному; n_{M_0+1} - частота интервала, следующего за модальным; h_i – длина модального интервала.

Пример 5. Определить моду вариационного ряда;

Значение	0-5	5-10	10-15	15-20	20-25	25-30
Частота	9	22	35	17	10	4

Длина интервалов здесь одинакова и равна h=5. Модальным является интервал 10-15, а его нижняя граница $x_{\min}$ =10. Частота модального интервала n_{M_0}=35, предшествующего модальному n_{M_0-1}=22, последующего за модальным n_{M_0+1}=17. Отсюда мода равна:

M_0= 10+5(35-22)/((35-22)+(35-17)) ≈ 12,1.

Задача 5. Определить моду для следующего вариационного ряда:

Значение	0-5	5-10	10-15	15-20	20-25	25-30
Частота	9	22	35	19	7	8

Отметим, что некоторые распределения не имеют моды или имеют несколько мод.

Замечание. Функция МОДА пакета Excel возвращает значение моды множества данных.

Определение. Медианой называется значение наблюдения, находящееся в середине распределения.

Чтобы определить медиану, варианты должны быть упорядочены либо по возрастанию, либо по убыванию.

В том случае, когда число вариант n нечетно, медиана равна варианте под номером $(n+1)/2$. Если число вариант n четно, медиана определяется как полусумма серединных вариант: $M_e = 0{,}5 \times (x_{n/2} + x_{n/2+1})$.

Пример 6. Определить медиану совокупности 3; 4; 6; 7; 10; 12; 15.

Здесь n=7 (нечетно). Медиана равна варианте под номером (n+1)/2= (7+1)/2=4, это 7.

Задача 6. Определить медиану совокупности №1: 25; 22; 20; 18; 7 и

для совокупности №2: 35; 32; 30; 28; 27; 25; 24;22.

Для интервального вариационного ряда с интервалами одинаковой величины медианный интервал - это первый интервал, сумма накопленных частот которого больше

полусуммы всех частот $\sum_{i=1}^{k} n_i/2$. Медиана Me при этом определяется по формуле: $M_e = x_{\min} + h(\sum_{i=1}^{k} n_i/2 - A)/n_{M_e}$. где n_{Me} - частота медианного интервала; A- накопленная частота интервала, предшествующего медианному; $x_{\min}$ - нижняя граница медианного интервала; h - ширина интервалов.

Пример 7. Вычислить значение медианы для вариационного ряда из примера 5.

Интервал	0-5	5-10	10-15	15-20	20-25	25-30
Частота	9	22	35	17	10	4
Накопленная частота	9	31	66	83	93	97

Эта таблица заполняется так. Первые две строки взяты из условия. Каждый элемент 3-й строки равен сумме предыдущего элемента 3-й строки и числа из этого же столбца 2-й строки.

В нашем случае $\sum_{i=1}^{6} n_i/2$ =97/2=48,5, поэтому медианным является интервал – это интервал 10-15. Нижняя граница медианного интервала $x_{\min} = 10$, накопленная частота интервала, предшествующего медианному A=31, частота медианного интервала n_{Me}=35. Поэтому медиана M_e=10+5(48,5-31)/35=12,5.

Задача 7. Вычислить значение медианы для вариационного ряда из задачи 5.

Замечание. Для вычисления значения медианы в пакете Excel можно применять функцию МЕДИАНА.

§4. Процентиль, дециль, квартиль.

Процентиль P_m применяется для вычисления точки, ниже которой находится m% вариант. Чтобы найти процентиль P_m, нужно упорядочить варианты в возрастающем или убывающем порядке и умножить общее число наблюдений $\sum_{i=1}^{k} n_i$ на процент m. Вычисленное значение показывает номер нужного процентиля.

Пример 8. Определим в примере 7 процентиль P_{15}.

В нашем случае m=15%. Тогда $m\sum_{i=1}^{6} n_i = 0{,}15 \times 97 = 14{,}55$.

Первый интервал, накопленная частота которого больше 14,55 это интервал 5-10. Поэтому

процентиль P_{15} равен $5+5\dfrac{14{,}55-9}{31-9} \approx 6{,}26$.

Задача 8. Вычислить в условиях задачи 7 процентиль P_{22}.

Децидилями называются процентили $P_{10}, P_{20}, \ldots, P_{80}, P_{90}$. Они обозначаются $D_1, D_2, \ldots, D_8, D_9$ соответственно.

Квартилями называются процентили P_{25}, P_{50}, P_{75}, обозначаются Q_1, Q_2, Q_3.

Децильным коэффициентом дифференциации называется отношение D_9/D_1. Он применяется при изучении распределения многих социально-экономических показателей для характеристики дифференциации.

Пример 9. Определить децильный коэффициент дифференциации в примере 7.

В условиях задачи $0{,}1\sum_{i=1}^{6} n_i = 0{,}1 \times 97 = 9{,}7$, то $D_1 = 5 + 5\frac{9{,}7 - 9}{31 - 9} = 5{,}16$. Так как $0{,}9\sum_{i=1}^{6} n_i = 0{,}9 \times 97 = 87{,}3$, то $D_9 = 20 + \frac{87{,}3 - 83}{93 - 83}(25\text{-}20) \approx 22{,}15$.

Отсюда следует, что децильный коэффициент дифференции равен $D_9 / D_1 = 22{,}7/5 = 4{,}29$.

Задача 9. Определить децильный коэффициент дифференциации в задаче 7.

§5. Показатели вариации.

Размах вариации

Определение. Размахом вариации *R* называется разность между наибольшим и наименьшим наблюдаемыми значениями $R = x_{max} - x_{min}$.

Размах вариации полезен для оценки изменчивости при сравнении большого количества выборок. Но поскольку практически любая выборка содержит нетипично большие и малые значения, размах вариации может привести к неверным выводам. Отметим, что по размаху вариации невозможно определенно сказать о значениях между двумя крайними.

Пример 10. Определить размах вариации для вариационного ряда.

Значение	3	5	6
Частота	11	13	17

Размах вариации $R = x_{max} - x_{min} = 6\text{-}3 = 3$.

Задача 10. Определить размах вариации вариационного ряда.

Значение	7	9	10
Частота	11	8	14

Коэффициент вариации

Это распространенный показатель колеблемости, вычисляется по формуле $V=\sigma/\bar{x}\times 100\%$, где σ – стандартное отклонение. Используется для оценки типичности средних величин. Чем меньше значение коэффициента вариации, тем однороднее совокупность по изучаемому признаку и типичнее средняя. Совокупности с коэффициентом вариации более 30-35% принято считать неоднородными.

Пример 11. Определим коэффициент вариации в примере 2.

Здесь стандартное отклонение $\sigma=\sqrt{D_{\hat{a}}}=\sqrt{14{,}99}\approx 3{,}87$. Тогда коэффициент вариации равен $V=\sigma/\bar{x}\times 100\%=3{,}87/11{,}86\times 100\%\approx 32{,}63\%$.

Задача 11. Определить коэффициент вариации в задаче 2.

§6.Асимметрия и эксцесс.

Для оценки асимметричности распределения применяется показатель асимметрии As, вычисляемый по формуле:

$As=\mu_3/\sigma^3$, где $\mu_3=\sum_{i=1}^{k} n_i(x_i-\bar{x})^3/\sum_{i=1}^{k} n_i$ -центральный момент 3-го порядка, σ – стандартное отклонение.

Если As >0, то асимметрии распределения правосторонняя (вытянутость вправо). Если As<0, то асимметрии распределения левосторонней (вытянутость влево).

Выборочная средняя всегда смещена в сторону экстремальных значений. Если в распределении присутствует несколько нетипично больших значений (то есть $As>0$), то медиана больше выборочной средней. Если в распределении содержится несколько нетипично маленьких значений (то есть $As<0$), то медиана меньше выборочной средней. Это означает,

что сравнение выборочной средней и медианы укажет, каково направление асимметрии.

Пример 12. Определить показатель асимметрии в примере 10. Заполним таблицу.

Номер	x_i	n_i	$x_i n_i$	$(x_i - \bar{x})^2$	$n_i(x_i - \bar{x})^2$	$n_i(x_i - \bar{x})^3 = n_i(x_i - \bar{x})^2 \times (x_i - \bar{x})$
1	3	11	33	3,53	38,80	-6,62
2	5	13	65	0,01	0,19	0,00
3	6	17	102	1,26	21,40	1,41
Сумма	-	41	200	4,8	60,39	-5,21

Здесь выборочная средняя равна $\bar{x} = \sum_{i=1}^{k} n_i x_i / \sum_{i=1}^{k} n_i = 200/41 \approx 4{,}88$.

Тогда стандартное отклонение $\sigma = \sqrt{1{,}47} \approx 1{,}21$. Центральный момент 3-го порядка равен $\mu_3 = \sum_{i=1}^{k} n_i (x_i - \bar{x})^3 / \sum_{i=1}^{k} n_i = -5{,}21/41 \approx -0{,}127$.

Показатель асимметрии равен As = $\mu_3 / \sigma^3 = -0{,}127/1{,}21^3 \approx -0{,}07 < 0$. Наблюдается левосторонняя асимметрия.

Задача 12. Определить показатель асимметрии в задаче 10.

Большую роль в анализе вариационных рядов, определении типа кривой распределения и при выравнивании вариационных рядов играет показатель эксцесса Ex, вычисляемый по формуле: Ex = $\mu_4 / \sigma^4 - 3$, где $\mu_4 = \sum_{i-1}^{k} n_i (x_i - \bar{x})^4 / \sum_{i=1}^{k} n_i$ - центральный момент 4-го порядка, σ - стандартное отклонение.

В случае Ex>0 ряд островершинен, а когда Ex<0, ряд низковершинен.

Пример 13. Определим показатель эксцесса в примере 12. заполним таблицу.

Номер	$(x_i - \bar{x})^2$	$n_i (x_i - \bar{x})^2$	$n_i (x_i - \bar{x})^4 = n_i (x_i - \bar{x})^2 \times (x_i - \bar{x})^2$
1	3,53	38,80	136,96
2	0,01	0,19	0,0019
3	1,26	21,40	26,96
Сумма	4,8	60,39	163,93

Тогда $\mu_4 = 163{,}93/41 \approx 4$, $Ex = 4/2{,}71 - 3 = -1{,}52$

§7.Доверительные интервалы.

Зная выборочную среднюю и выборочную дисперсию можно оценить с некоторой вероятностью, называемой доверительной, интервал, в котором содержится параметр генеральной совокупности. Этот интервал называется доверительным интервалом.

Доверительный интервал для оценки генеральной средней a нормально распределенного количественного признака *X* по выборочной средней $\overline{X}$ при известном среднем квадратическом отклонении σ генеральной совокупности (на практике — *при объеме выборки* $n \geq 30$) определяется соотношением

$$\overline{X} - t\sigma/\sqrt{n} < a < \overline{X} + t\sigma/\sqrt{n},$$

где t определяется с помощью таблицы распределения Лапласа из уравнения $2\Phi_0(t) = p$, где p - доверительная вероятность.

Пример 14. Руководство фирмы провело выборочное обследование 800 служащих. Средний стаж работы в фирме равен 9,1 года, а среднеквадратическое отклонение – 1,3 года. Считая стаж работы служащих распределенным по нормальному закону, определить с вероятностью 95%

доверительный интервал, в котором окажется средний стаж работы всех служащих фирмы.

Решение. По условию $\overline{X} = 9{,}1$, $\sigma = 1{,}3$, n=800, p=0,95. Для нахождения t используем формулу $2\Phi_0(t) = p$, $2\Phi_0(t) = 0{,}95$, $\Phi_0(t) = 0{,}475$, следовательно, $t = 1{,}96$. Тогда $t\sigma/\sqrt{n} \approx 0{,}09$, а значит, $9{,}0099 < a < 9{,}19$.

Задача 14. Строительная компания хочет оценить возможности бизнеса на рынке строительных работ. Было опрошено 600 домовладельцев. Средняя стоимость строительных работ составляет 5000 у.е. Среднеквадратическое отклонение составляет 10 у.е. Считая стоимость работ распределенной по нормальному закону, определить доверительный интервал, в котором окажется средняя плата за услуги строителей с доверительной вероятностью 99%.

Доверительный интервал для генеральной средней при неизвестной генеральной дисперсии определяется так

$$\overline{X} - t_{\alpha/2,n-1}s/\sqrt{n-1} < a < \overline{X} + t_{\alpha/2,n-1}s/\sqrt{n-1}.$$

Здесь $\overline{X}$ - выборочная средняя, n - объем выборки, $\alpha = 1 - p$, p - доверительная вероятность, s - выборочное стандартное отклонение, $t_{\alpha/2,n-1}$ определяется с помощью таблицы распределения Стьюдента. Кроме того, это значение можно вычислить с помощью функции =СТЬЮДРАСПОБР(α; $n-1$) пакета Excel.

Пример 15.Средний вес опрошенных $\overline{X} = 51$килограмм, выборочное стандартное отклонение $s = 0{,}4$ кг. Объем выборки $n = 50$ человек. Определить с доверительной вероятностью $p = 98\%$ доверительный интервал для веса людей в генеральной совокупности.

Так как $p = 0{,}98$, $\alpha = 1 - 0{,}98 = 0{,}02$, $\alpha/2 = 0{,}01$. Следовательно, $t_{0{,}01;49} = 2{,}679952$.

$\overline{X} \pm t_{\alpha/2,n-1} s/\sqrt{n-1} = 51 \pm 0{,}15314$. Искомым является интервал $(50{,}84686; 51{,}15314)$.

Задача 15.Средний вес опрошенных $\overline{X} = 68$ кг, выборочное стандартное отклонение $s = 0{,}7$ кг. Объем выборки $n = 41$ человек. Определить с доверительной вероятностью $p = 95\%$ доверительный интервал для веса людей в генеральной совокупности.

Для дальнейших вычислений нам понадобится таблица

α	0,4	0,25	0,2	0,15	0,1	
z_α	0,253	0,675	0,842	1,036	1,282	
α	0,1	0,05	0,025	0,01	0,005	0,001
z_α	1,282	1,645	1,960	2,326	2,576	3,090

Ею можно пользоваться, если объем выборки $n \geq 30$. Эти значения можно получить с помощью пакета Excel, применяя функцию =НОРМСТОБР(1-α)

Пример 16. Находясь в условиях примера 14 вычислить объем выборки, зная, что ширина доверительного интервала $\pm 0{,}1$ кг.

Воспользуемся формулой $z_{\alpha/2} s/\sqrt{n-1} \leq 0{,}1$. Тогда $n \geq 1 + (10 z_{\alpha/2} s)^2 = 1 + (10 \times 2{,}326 \times 0{,}4)^2 \approx 87{,}58$. Значит, минимальный объем выборки равен 88 человек.

Задача 16. Находясь в условиях задачи 14 вычислить объем выборки, зная, что ширина доверительного интервала $\pm 0{,}3$ кг.

Доверительный интервал для генеральной доли.

Часто требуется определить доверительный интервал для генеральной доли – доли объектов генеральной совокупности, обладающих некоторым свойством. Он вычисляется по следующему правилу. Выполняется выборка объема n, из которой n_1 объектов обладают нужным свойством. Затем вычисляется выборочная доля $\hat{p} = n_1/n$. Если выполняются условия $n\hat{p} \geq 5$, $n(1-\hat{p}) \geq 5$, доверительный интервал для генеральной доли задается формулой $\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$.

Пример 17. О работе ЖКХ микрорайона опросили n=2100 человек. $n_1 = 300$ человек оказались недовольны работой ЖКХ. Найти доверительный интервал доли недовольных работой ЖКХ в генеральной совокупности, если доверительная вероятность $p = 98\%$.

Вычислим $\hat{p} = 300/2100 \approx 0{,}14$. Проверим выполнение условий $n\hat{p} = 300 \geq 5$, $n(1-\hat{p}) = 1806 \geq 5$. Они выполнены. $\alpha = 1 - 0{,}98 = 0{,}02$, значит $\alpha/2 = 0{,}01$, т.е. $z_{\alpha/2} = 2{,}326$, $\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} \approx 0{,}14 \pm 2{,}326 \times \sqrt{0{,}14 \times 0{,}86/2100} \approx 0{,}14 \pm 0{,}018$. Значит, искомым интервалом является (0,122;0,158).

Задача 17. О работе ЖКХ микрорайона опросили n=1500 человек. $n_1 = 400$ человек оказались недовольны работой ЖКХ. Найти доверительный интервал доли недовольных работой ЖКХ в генеральной совокупности, если доверительная вероятность $p = 95\%$.

Рассмотрим пример об отыскании объема выборки при известной ширине интервала.

Пример 18. В условиях примера 17 требуется определить объем выборки, если ширина доверительного интервала $\pm 0{,}004$.

$$z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} \leq 0{,}004 \Rightarrow$$

$$(z_{\alpha/2})^2\,\hat{p}(1-\hat{p})/n \leq (0{,}004)^2 = 0{,}000016 \Rightarrow$$

$$n \geq (z_{\alpha/2})^2\,\hat{p}(1-\hat{p})/0{,}000016 \approx 2{,}326^2 \times 0{,}14 \times 0{,}86/0{,}000016 \approx 40712$$

Задача 18. В условиях задачи 17 требуется определить объем выборки, если ширина доверительного интервала $\pm 0{,}007$.

Интервал предсказания.

Интервал предсказания позволяет использовать данные выборки, чтобы предсказать с заданной вероятностью значения нового наблюдения, считая, что новое наблюдение получено так же, как и прочие. Он определяется формулой

$$(\overline{X} - t_{\alpha/2,n-1} s\sqrt{(n+1)/(n-1)}; \overline{X} + t_{\alpha/2,n-1} s\sqrt{(n+1)/(n-1)}).$$

Здесь $\overline{X}$ - выборочная средняя, n –объем выборки, $\alpha = 1 - p$, p - доверительная вероятность, s – выборочное стандартное отклонение.

Пример 19. Результат замеров температуры в Казани в 12 часов дня на Кремлевской в течении 6 дней таковы: $\overline{X} = 11^0 C$, выборочное стандартное отклонение $s = 3$. Предполагая, что результаты измерения температуры распределены нормально, определить с вероятностью p=98% интервал предсказания для результатов замера на 7 день.

Так как p=98%, то $\alpha = 0{,}02 \Rightarrow \alpha/2 = 0{,}01$. $t_{0{,}01;5} = 4{,}03$. $\overline{X} \pm t_{\alpha/2,n-1} s\sqrt{(n+1)/(n-1)} = 11 \pm 4{,}03 \times 3\sqrt{7/5} \approx 11 \pm 14{,}31$. То есть интервал имеет вид:
(-3,31;25,31).

Задача 19. Результат замеров температуры в Казани в 12 часов дня на Кремлевской в течении 10 дней таковы: $\overline{X} = 15^0 C$, выборочное стандартное отклонение $s = 2$. Предполагая, что результаты измерения температуры распределены нормально, определить с вероятностью p=99% интервал предсказания для результатов замера на 11 день.

Контрольная работа №1

Задание 1. Получены данные об обращениях клиентов в автомойку

Интервал времени	До 8	8-10	10-12	12-14	14-16	Св. 16
Число клиентов	A	b	c	d	e	f

Построить функцию распределения эмпирическую, кумуляту, гистограмму, полигон. Найти математическое ожидание и дисперсию.

Параметры определяются из таблицы

	1	2	3	4	5
a	20	23	17	16	19
b	15	16	11	18	18
c	13	18	3	11	21
d	25	19	5	19	30
e	66	26	7	23	25
f	17	34	19	17	26

Задание 2.

а)Средний вес опрошенных $\overline{X}$ килограмм, выборочное стандартное отклонение s кг. Объем выборки n человек. Определить с доверительной вероятностью p доверительный интервал для веса людей в генеральной совокупности.

б)Вычислить объем выборки, зная, что ширина доверительного интервала $\pm m$ кг.

	1	2	3	4	5
$\overline{X}$	56	62	75	70	64
s	0,6	0,8	1	0,6	0,9
n	100	120	140	110	123
p	0,95	0,98	0,99	0,95	0,98
m	0,1	0,2	0,3	0,4	0,5

Задание 3.

а) О работе ЖКХ микрорайона опросили n человек, n_1 из них оказались недовольны работой ЖКХ. Найти доверительный интервал доли недовольных работой ЖКХ в генеральной совокупности, зная доверительную вероятность p.

б) определить объем выборки, если ширина доверительного интервала $\pm m$

	1	2	3	4	5
n	1300	1500	1100	1325	1600
n_1	100	48	200	500	140
p	0,95	0,98	0,99	0,95	0,98

m	0,01	0,02	0,03	0,04	0,05

Задание 4.

Результат замеров температуры в Казани в 13 часов дня на Кремлевской в течении d дней $\overline{X}$, выборочное стандартное отклонение s. Предполагая, что результаты измерения температуры распределены нормально, определить с вероятностью p интервал предсказания для результатов замера на d+1-ый день.

	1	2	3	4	5
$\overline{X}$	14	15	12	13	17
s	0,4	0,32	0,61	0,15	0,8
p	0,95	0,98	0,99	0,95	0,98
m	0,1	0,2	0,3	0,4	0,5

§8.Испытание гипотез.

Часто требуется узнать, подчиняется ли заданным ограничениям генеральная совокупность. Для этого проводят испытание гипотез. Сначала из генеральной совокупности выбирают n элементов (выборку объема n), для которых вычисляют нужные характеристики. Далее формулируют две гипотезы – основную, которую обозначают H_0, и альтернативную, обозначаемую H_1. Гипотеза H_0 является утверждением, подлежащим проверке. Пусть, например, гипотеза

H_0: средний балл на экзамене $a = 65$.

Альтернативная гипотеза задается одним из трех способов:

H_1: $a > 65$ (правосторонняя проверка);

H_1: $a < 65$ (левосторонняя проверка);

H_1: $a \neq 65$ (двусторонняя проверка).

При этом первые две проверки называются односторонними. Задается доверительная вероятность p. Так называется величина, отражающая степень уверенности исследователя в результате испытания. Вычисляется уровень значимости, равный $\alpha = 1 - p$

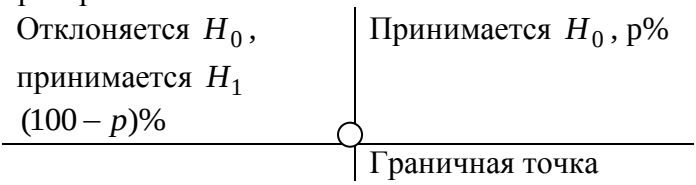
для односторонней проверки, и $\alpha = \frac{1-p}{2}$ для двусторонней проверки. По заданным значениям α и n в зависимости от задачи по таблицам (или с помощью Excel) находят граничные точки, которые затем наносят на координатную ось. Затем по заданным параметрам находят значение, которое называется статистикой. Его тоже наносят на координатную ось. В зависимости от расположения статистики и граничных точек возможны варианты:

1) принимается H_0;

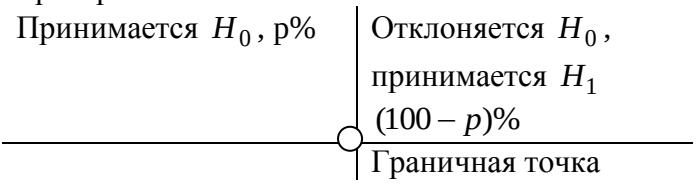
2) отклоняется H_0 и без дополнительной проверки принимается H_1;

3) недостаточно данных для приема гипотез.

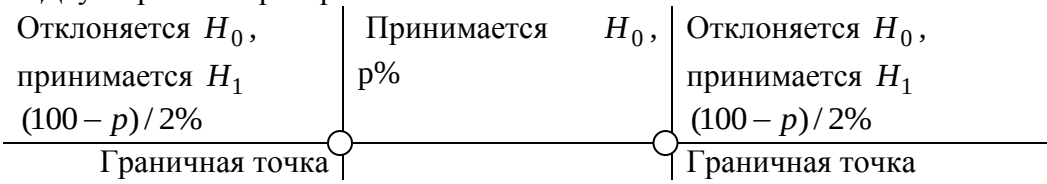
Левосторонняя проверка:



Правосторонняя проверка:



Двусторонняя проверка:



Область принятия шире, если выше доверительная вероятность.

§9.Испытание гипотез на основе выборочной средней при неизвестной генеральной дисперсии

Вычислим выборочную среднюю $\overline{X}$ и выборочное стандартное отклонение s для выборки объема n. Пусть a – предполагаемое значение генеральной средней. По таблице t-распределения Стьюдента найдем $t_{\alpha,n-1}$. Граничными точками являются: для правосторонней проверки $t_{\alpha,n-1}$, для левосторонней проверки $-t_{\alpha,n-1}$, для двусторонней проверки $\pm t_{\alpha,n-1}$. Статистика вычисляется по формуле $t=\frac{\overline{X}-a}{s}\sqrt{n-1}$. Значения граничных точек могут быть определены с помощью Excel. Для двусторонней проверки $t_{\alpha,n-1}$=СТЬЮДРАСПОБР(1-p;n-1), для односторонней - $t_{\alpha,n-1}$=СТЬЮДРАСПОБР(2(1-p);n-1).

Пример 20. Производитель утверждает, что средний вес мотка пряжи не меньше a=50г. Инспектор отобрал 10 мотков пряжи и взвесил. Их вес был 48, 49, 50, 49, 47, 45, 51, 48, 51, 45 соответственно. Не противоречит ли это утверждению производителя? Предполагается, что вес мотков пряжи распределен нормально. Доверительная вероятность p=99%.

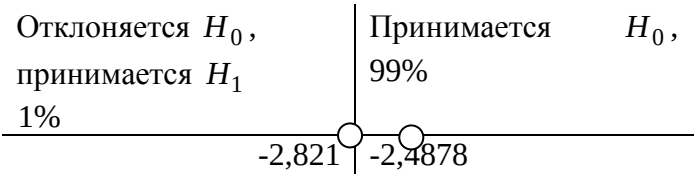
Решение.

H_0: генеральная средняя нормальной совокупности a=50г.

H_1: a<50г.

Проведем левостороннюю проверку. $\alpha=1-p=0{,}01\Rightarrow t_{\alpha,n-1}=2{,}821$, а значит, граничной точкой является -2,821. Вычислим $\overline{X}=\sum_{i=1}^{n} x_i/n=48{,}3$, $s=\sqrt{\sum_{i=1}^{n}(x_i-\overline{X})^2/n}\approx$ 2,05г. Вычислим статистику t=-2,4878. Можно использовать функцию СТАНДОТКЛОНП пакета Excel.

Отметим значения:



Принимается гипотеза H_0 на уровне значимости 1%. Выборка инспектора не противоречит утверждению производителя.

Задача 20. Производитель утверждает, что средний вес мотка пряжи не меньше а=100г. Инспектор отобрал 12 мотков пряжи и взвесил. Их вес был 101, 99, 98, 95, 101, 99, 96, 99, 101, 94, 101, 102г. соответственно. Не противоречит ли это утверждению производителя? Предполагается, что вес мотков пряжи распределен нормально. Доверительная вероятность р=97%.

§10. Испытание гипотез на основе выборочной доли.

Вычислим выборочную долю по правилу. Выполним выборку объема n, из которой n_1 объектов обладают нужным свойством. Затем вычислим выборочная доля $\hat{p} = n_1 / n$. Сравним ее с генеральной долей $\bar{p}$. Для правосторонней проверки вычислим граничную точку z_α, для левосторонней $-z_\alpha$, для двусторонней $\pm z_\alpha$. Статистика определяется по формуле

$$z = \frac{\hat{p} - \bar{p}}{\sqrt{\bar{p}(1-\bar{p})}}\sqrt{n}.$$

Пример 20. Производитель утверждает, что доля бракованных изделий не превосходит 4%. В случайной выборке объема n=100 изделий оказалось 7 бракованных изделий. Не противоречит ли это утверждению производителя? Доверительная вероятность р=99%.

Решение.

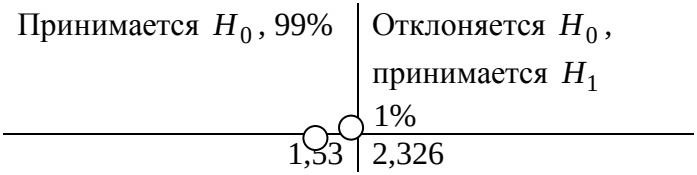
H_0 :доля бракованных изделий $\bar{p}$ =0,04

H_1: $\bar{p}$ >0,04.

Проведем правостороннюю проверку. $\alpha = 1 - p = 1 - 0{,}99 = 0{,}01 \Rightarrow z_\alpha = 2{,}326$ Генеральная доля $\hat{p} = 0{,}07$. Тогда статистика

$$z = \frac{\hat{p} - \overline{p}}{\sqrt{\overline{p}(1-\overline{p})}}\sqrt{n} = \frac{0{,}07 - 0{,}04}{\sqrt{0{,}04(1-0{,}04)}}\sqrt{100} \approx 1{,}53$$

Отметим значения на числовой оси



Принимается гипотеза H_0.

Задача 21. Производитель утверждает, что доля бракованных изделий не превосходит 5%. В случайной выборке объема n=120 изделий оказалось 10 бракованных изделий. Не противоречит ли это утверждению производителя? Доверительная вероятность p=95%.

§11. Испытание гипотез о двух генеральных дисперсиях.

Пусть для двух независимых выборок объема n_1 и n_2 соответственно требуется узнать, принадлежат ли они нормальным генеральным совокупностям с одинаковой дисперсией. Найдем для каждой выборки выборочную дисперсию s_1^2 и s_2^2 соответственно. По первой выборке оценка генеральной дисперсии $\sigma_1^2 = n_1 s_1^2/(n_1 - 1)$, по второй - $\sigma_2^2 = n_2 s_2^2/(n_2 - 1)$. Статистика $F = \max(\sigma_1, \sigma_2)/\min(\sigma_1, \sigma_2)$. Обозначим через n_A объем выборки с большей генеральной дисперсией, n_B - соответственно, с меньшей. Граничная точка задается с помощью таблицы F-распределения Фишера $F_{\alpha; n_A - 1; n_B - 1}$. Кроме того, ее можно вычислить с помощью функции FРАСПОБР($\alpha; n_A - 1; n_B - 1$) пакета Excel.

Пример 22. Инвестиция 1 рассчитана на $n_1 = 14$ лет, дисперсия ежегодных прибылей $s_1^2 = 25\%^2$. Инвестиция 2 рассчитана на $n_2 = 11$ лет, дисперсия ежегодных прибылей $s_2^2 = 20\%^2$. Предполагается, что распределение ежегодных прибылей на инвестиции подчинено нормальному закону распределения. Проверить, равны ли риски инвестиций 1 и 2. Доверительная вероятность p=99%.

Решение.

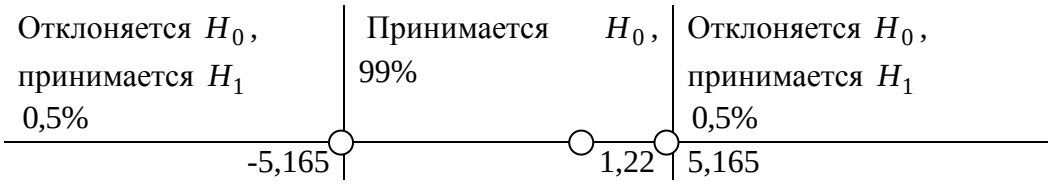
H_0: $\sigma_1^2 = \sigma_2^2$,

H_1: $\sigma_1^2 \neq \sigma_2^2$.

Вычислим оценку генеральной дисперсии по первой выборке $\sigma_1^2 = n_1 s_1^2 /(n_1 - 1) = 14 \times 25/(14-1) \approx 26{,}92$, по второй - $\sigma_2^2 = n_2 s_2^2 /(n_2 - 1) = 11 \times 20/10 = 22$,

$F = \max(\sigma_1, \sigma_2) / \min(\sigma_1, \sigma_2) = 26{,}92/22 \approx 1{,}22$. $n_A = 14$, $n_B = 11$. Проведем двустороннюю проверку. $\alpha = (1-p)/2 = (1-0{,}99)/2 = 0{,}005 \Rightarrow$

$F_{0{,}005;13;11} \approx 5{,}165$, а значит, граничными точками являются $\pm 5{,}165$.



Принимается гипотеза H_0.

Задача 22. Инвестиция 1 рассчитана на $n_1 = 17$ лет, дисперсия ежегодных прибылей составляет $s_1 = 10\%^2$, Инвестиция 2 рассчитана на $n_2 = 14$ лет, дисперсия ежегодных прибылей составляет $s_2 = 10\%^2$. Предполагается, что распределение ежегодных прибылей на инвестиции подчинено

нормальному закону распределения. Проверить, равны ли риски инвестиций 1 и 2. Доверительная вероятность p=95%.

§12. Испытание гипотезы по выборочным средним с неизвестными генеральными дисперсиями.

Требуется определить, принадлежат ли выборки объема n_1 и n_2 соответственно нормальным генеральным совокупностям с одинаковыми средними. Для этого проверяем гипотезу

$H_0: a_1 = a_2$.

Дальнейшая проверка зависит от того, равны ли неизвестные генеральные дисперсии.

Случай 1. Пусть неизвестные генеральные дисперсии равны. По таблице t-распределения Стьюдента находим $t_{\alpha;n_1+n_2-2}$. Граничными точками являются: для правосторонней проверки $t_{\alpha;n_1+n_2-2}$, для левосторонней проверки $-t_{\alpha;n_1+n_2-2}$, для двусторонней проверки $\pm t_{\alpha;n_1+n_2-2}$. Статистика определяется с помощью формулы

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

Пример 23. На обработку каждой из $n_1 = 20$ анкет первым способом затрачено в среднем $\overline{X_1} = 25c$, выборочная дисперсия $s_1^2 = 2c^2$. На обработку каждой из $n_2 = 18$ анкет вторым способом затрачено в среднем $\overline{X_2} = 29c$, выборочная дисперсия $s_2^2 = 1c^2$. Следует ли из этого, что на обработку одной анкеты вторым способом требуется в среднем больше времени? Доверительная вероятность p=95%.

Решение.

Применим результаты §11, чтобы проверить гипотезу о совпадении неизвестных генеральных дисперсиях.

H_0: $\sigma_1^2 = \sigma_2^2$,

H_1: $\sigma_1^2 \neq \sigma_2^2$.

Вычислим оценку генеральной дисперсии по первой выборке $\sigma_1^2 = n_1 s_1^2 /(n_1 - 1) = 20 \times 2/(20-1) \approx 2{,}11$, по второй - $\sigma_2^2 = n_2 s_2^2 /(n_2 - 1) = 18 \times 1/17 = 1{,}06$,

$F = \max(\sigma_1, \sigma_2) / \min(\sigma_1, \sigma_2) = 2{,}11/1{,}06 \approx 1{,}99$. $n_A = 20$, $n_B = 18$.

Проведем двустороннюю проверку. $\alpha = (1-p)/2 = (1-0{,}95)/2 = 0{,}025 \Rightarrow$

$F_{0{,}025;19;17} \approx 2{,}63$, а значит, граничными точками являются $\pm 2{,}63$.

Отклоняется H_0, принимается H_1 2,5%	Принимается H_0, 95%	Отклоняется H_0, принимается H_1 2,5%
-2,63	1,99	2,63

Таким образом, мы получили, что неизвестные генеральные дисперсии равны. Теперь выдвинем гипотезы

H_0: $a_1 = a_2$,

H_1: $a_1 < a_2$.

Проведем левостороннюю проверку, p=0,95. $\alpha = 1 - p = 1 - 0{,}95 = 0{,}05 \Rightarrow$ $t_{\alpha, n_1+n_2-2} = t_{0{,}05;20+18-2} \approx 1{,}688$.

Найдем теперь значение статистики

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{25 - 29}{\sqrt{\frac{20 \cdot 2 + 18 \cdot 1}{20 + 18 - 2}\left(\frac{1}{20} + \frac{1}{18}\right)}} = -9{,}69966$$

Отклоняется H_0, принимается H_1 5%	Принимается H_0, 95%
-9,69966	-1,688

Отклоняем гипотезу H_0 и принимаем гипотезу H_1 на уровне значимости 5%.

Задача 23. На обработку каждой из $n_1 = 40$ анкет первым способом затрачено в среднем $\overline{X_1} = 22c$, выборочная дисперсия $s_1^2 = 7c^2$. На обработку каждой из $n_2 = 35$ анкет вторым способом затрачено в среднем $\overline{X_2} = 26c$, выборочная дисперсия $s_2^2 = 4c^2$. Следует ли из этого, что на обработку одной анкеты вторым способом требуется в среднем больше времени? Доверительная вероятность p=99%.

Случай 2. Неравенство генеральных дисперсий.

Если $n_1 \geq 30$, $n_2 \geq 30$, граничными будут точки: z_α для правосторонней проверки, $-z_\alpha$ для левосторонней проверки, $\pm z_\alpha$ для двусторонней проверки. Статистика определяется с помощью формулы $z = \dfrac{\overline{X_1} - \overline{X_2}}{\sqrt{\dfrac{s_1^2}{n_1 - 1} + \dfrac{s_2^2}{n_2 - 1}}}$, где $\overline{X_1}$, $\overline{X_2}$ - есть средние заданных выборок.

Пример 24. На обработку каждой из $n_1 = 42$ анкеты первым способом затрачено в среднем $\overline{X_1} = 40c$, выборочная дисперсия $s_1^2 = 8c^2$, а каждой из $n_2 = 50$ анкет вторым способом - $\overline{X_2} = 36c$, выборочная дисперсия $s_2^2 = 4c^2$. Можно ли сделать вывод, что при обработке анкет первым способом на обработку одной анкеты в среднем требуется больше времени. Доверительная вероятность p=99%.

Решение.

Применим опять результаты §11, для проверки гипотезы о совпадении неизвестных генеральных дисперсиях.

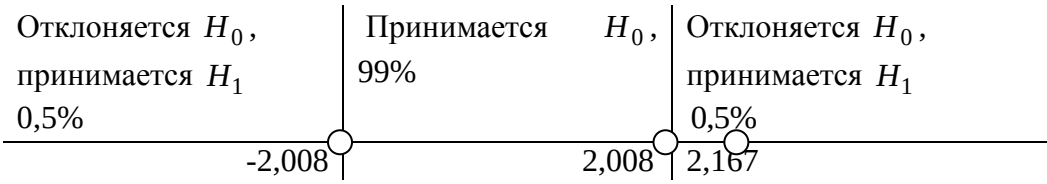
$H_0: \sigma_1^2 = \sigma_2^2$,

$H_1: \sigma_1^2 \neq \sigma_2^2$.

Вычислим оценку генеральной дисперсии по первой выборке $\sigma_1^2 = n_1 s_1^2 /(n_1 - 1) = 42 \times 8/(42-1) \approx 8{,}195$, по второй - $\sigma_2^2 = n_2 s_2^2 /(n_2 - 1) = 50 \times 4/49 \approx 4{,}082$,

$F = \max(\sigma_1, \sigma_2) / \min(\sigma_1, \sigma_2) = 8{,}195/4{,}082 \approx 2{,}008$. $n_A = 42$, $n_B = 50$. Проведем двустороннюю проверку. $\alpha = (1-p)/2 = (1-0{,}99)/2 = 0{,}005 \Rightarrow$

$F_{0{,}01;41;49} \approx 2{,}008$, а значит, граничными точками являются $\pm 2{,}008$.



Таким образом, мы получили, что неизвестные генеральные дисперсии различны.

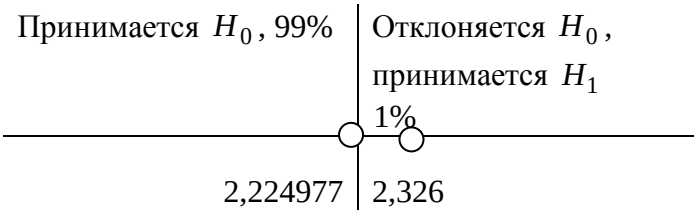
Теперь выдвинем гипотезы

$H_0: a_1 = a_2$,

$H_1: a_1 > a_2$.

Проведем правостороннюю проверку. p=99%, значит, $\alpha = 1 - p = 1 - 0{,}99 = 0{,}01 \Rightarrow z_\alpha = 2{,}326$. Статистика

$$z = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}}} = \frac{40 - 36}{\sqrt{\frac{8}{42-1} + \frac{4}{36-1}}} = 2{,}224977$$



Принимаем гипотезу H_0 на уровне значимости 1%.

Задача 24. На обработку каждой из $n_1 = 48$ анкеты первым способом затрачено в среднем $\overline{X_1} = 24c$, выборочная дисперсия $s_1^2 = 10c^2$, а каждой из $n_2 = 60$ анкет вторым способом - $\overline{X_2} = 29c$, выборочная дисперсия $s_2^2 = 5c^2$. Можно ли сделать вывод, что при обработке анкет первым способом на обработку одной анкеты в среднем требуется меньше времени. Доверительная вероятность p=95%.

§13. Испытание гипотез на основе выборочной доли.

Пусть требуется сделать вывод о двух выборках объема $n_1 \geq 30$, $n_2 \geq 30$ с выборочными долями $\hat{p}_1$ и $\hat{p}_2$, взяты ли они из генеральных совокупностей с одинаковой генеральной долей. Проверяем гипотезу

$H_0: p_1 = p_2$ - генеральные доли равны.

Для правосторонней проверки граничной точкой будет z_α, $-z_\alpha$ для левосторонней проверки, $\pm z_\alpha$ для двусторонней проверки. Статистика определяется формулой

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\overline{p}(1-\overline{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$, где $\overline{p}$ - выборочная доля в объединенной выборке.

Пример 25. Проводились испытания новой вакцины. В эксперименте участвовали $n_1 = 5000$ мужчин и $n_2 = 5100$ женщин. Побочные эффекты возникли у 100 мужчин и 110 женщин. Можно ли утверждать, что побочные эффекты после использования вакцины возникают чаще у женщин. Доверительная вероятность p=98%

Решение. Выборочные доли равны $\hat{p}_1 = 100/5000 = 0{,}02$, $\hat{p}_2 = 110/5100 = 0{,}022$.

$H_0: p_1 = p_2$

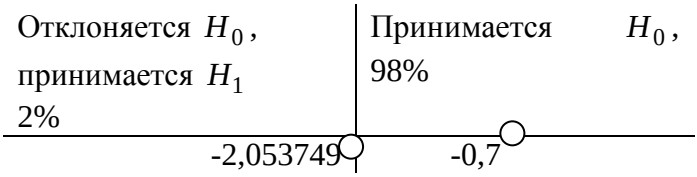
$H_1: p_1 < p_2$.

Проведем левостороннюю проверку. $\alpha = 1 - p = 1 - 0{,}98 = 0{,}02 \Rightarrow z_\alpha = 2{,}053749$. Граничной точкой является -2,053749. Выборочная доля объединенной выборки $\overline{p} = (100 + 110)/(5000 + 5100) \approx 0{,}021$.

Статистика

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\overline{p}(1-\overline{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0{,}02 - 0{,}022}{\sqrt{0{,}021(1-0{,}021)\left(\frac{1}{5000} + \frac{1}{5100}\right)}} \approx -0{,}7 .$$

Тогда



Принимаем гипотезу H_0 на уровне значимости 2%. Побочные эффекты от нового лекарства у женщин и мужчин возникают одинаково.

Задача 25. Проводились испытания новой вакцины. В эксперименте участвовали $n_1 = 3000$ мужчин и $n_2 = 3800$ женщин. Побочные эффекты возникли у 50 мужчин и 110 женщин. Можно ли утверждать, что побочные эффекты после использования вакцины возникают чаще у женщин. Доверительная вероятность p=95%

§14. Испытание гипотез по спаренным данным.

В ряде случаев выборки являются зависимыми. В этом случае элементы группируют попарно (по одному из каждой выборки), затем проводят испытание гипотезы для средней разности между парными измерениями.

В данном случае применим алгоритм.

Определим граничные точки с помощью таблицы t-распределения Стьюдента: $t_{\alpha;n-1}$(для правосторонней проверки), $-t_{\alpha;n-1}$(для левосторонней проверки), $\pm t_{\alpha;n-1}$(для двусторонней проверки). Обозначим n – объем парной выборки. Затем найдем в каждой паре разность значений d. Для полученных разностей определим $\overline{X_d}$, вычислим выборочное стандартное отклонение s_d . Определим значение статистики по формуле $t = \dfrac{\overline{X_d}\sqrt{n-1}}{s_d}$.

Пример 26. Можно ли утверждать, что приборы учета, выпускаемые заводами 1 и 2 имеют различные сроки служб. Доверительная вероятность p=98%.

Номер прибора	X – срок службы приборов завода 1, мес.	Y - срок службы приборов завода 2, мес.
1	60	58
2	62	54
3	61	62
4	63	60
5	66	64

Решение

Выдвенем гипотезы

$H_0: a_d = 0$ (срок службы приборов одинаков)

$H_1: a_d \neq 0$ (срок службы приборов различен).

Проведем двустороннюю проверку.

Номер прибора	X – срок службы приборов завода 1, мес.	Y - срок службы приборов завода 2, мес.	$d = X - Y$	d^2
1	60	58	2	4
2	62	54	8	64

3	61	62	-1	1
4	63	60	3	9
5	66	64	2	4
Сумма	-	-	14	82

$\overline{X_d} = \sum d / n = 14/5 = 2{,}8$,

$s_d^2 = \sum d^2 / n - \overline{X_d^2} = 82/5 - 2{,}8^2 = 8{,}56$.

$\alpha = (1-p)/2 = 0{,}01 \Rightarrow t_{\alpha;n-1} = t_{0{,}01;4} = 3{,}746947$.

Статистика $t = \dfrac{\overline{X_d}\sqrt{n-1}}{s_d} = \dfrac{2{,}8\sqrt{4}}{8{,}56} \approx 0{,}654$.

Итак, гипотеза H_0 принимается на уровне значимости 2%.

Задача 26. Можно ли утверждать, что приборы учета, выпускаемые заводами 1 и 2 имеют различные сроки служб. Доверительная вероятность p=95%

Номер прибора	X – срок службы приборов завода 1, мес.	Y - срок службы приборов завода 2, мес.
1	64	61
2	61	66
3	64	60
4	57	59
5	59	62

§15. Испытание гипотезы о принадлежности нового наблюдения генеральной совокупности.

Когда проводится новое наблюдение, иногда требуется проверить, принадлежит ли оно к той же нормальной совокупности, что и выборка. Для определения граничной точки

находим по таблице t – распределения Стьюдента $t_{\alpha;n-1}$ - для правосторонней проверки, $-t_{\alpha;n-1}$ - для левосторонней проверки, $\pm t_{\alpha;n-1}$ - для двусторонней проверки. Теперь определим статистику. Для выборки объема n вычислим выборочную среднюю $\overline{X}$ и выборочное стандартное отклонение s. Обозначим $X_{нов}$ результат нового наблюдения. Статистика определяется формулой $t = \frac{X_{нов} - \overline{X}}{s\sqrt{n+1}}\sqrt{n-1}$.

Пример 27. Для выборки объема n=30 средний годовой доход на человека составил $\overline{X}$ = 200 тыс.руб., выборочное стандартное отклонение s=0,6 тыс.руб. Годовой доход очередного респондента $X_{нов}$= 189 тыс.руб. Можно ли утверждать, что он принадлежит той же целевой группе. Считаем, что годовой доход распределен нормально. Доверительная вероятность p=99%

Решение.

H_0: новый респондент принадлежит той же целевой группе.

H_1: новый респондент не принадлежит той же целевой группе.

Проведем двустороннюю проверку.

$p = 0{,}99 \Rightarrow \alpha = (1-p)/2 = 0{,}005$,

$t_{\alpha;n-1} = t_{0{,}005;29} = 2{,}756 \Rightarrow$ граничные точки $\pm 2{,}756$. Статистика

$t = \frac{X_{нов} - \overline{X}}{s\sqrt{n+1}}\sqrt{n-1} = = \frac{189-200}{0{,}6\sqrt{31}}\sqrt{29} \approx -10{,}6392$.

Отклоняется H_0, принимается H_1 0,5%	Принимается H_0, 99%	Отклоняется H_0, принимается H_1 0,5%
-10,6392 -2,756		2,756

Отклоняется H_0, принимается H_1 на уровне значимости 0,5%.

Задача 27. Для выборки объема n=20 средний годовой доход на человека составил $\overline{X}$ = 150 тыс.руб., выборочное стандартное отклонение s=0,8 тыс.руб. Годовой доход очередного респондента $X_{нов}$= 160 тыс.руб. Можно ли утверждать, что он принадлежит той же целевой группе. Считаем, что годовой доход распределен нормально. Доверительная вероятность p=95%.

§16. Непараметрические испытания.

Ранее мы предполагали нормальное распределение генеральных совокупностей. Теперь будем проверять гипотезу о наличии связи между значениями двух величин.

H_0: Связи между значениями двух величин нет

H_1: Связь между значениями двух величин есть.

Пусть задана выборка объема n элементов x_1, x_2,…, x_n. И пусть известно, что все элементы выборки обладают двумя признаками: A и B. A именно, признак A принимает значения A_1,…, A_m, а признак B – значения B_1,…, B_k. Обозначим n_{ij} - число элементов выборки, обладающих одновременно признаками A_i и B_j. С помощью этих данных строим таблицу наблюдаемых частот. Обозначим $n_{\cdot j} = \sum_{i=1}^{m} n_{ij}$, $n_{i\cdot} = \sum_{j=1}^{k} n_{ij}$. Тогда статистика определяется с помощью формулы $\chi^2 = n\left(\sum_{i,j=1}^{m,k} \frac{n_{ij}^2}{n_{i\cdot} n_{\cdot j}} - 1\right)$. Граничная точка определяется с помощью таблицы χ^2 - распределения (ее называют также «хи-квадрат») $\chi^2_{\alpha,r}$, где $\alpha = 1 - p$ - уровень значимости, $r = (m-1)\times(k-1)$. Ее можно определить с помощью статистической функции ХИ2ОБР($\alpha; r$) пакета Excel.

Пример 28. При отборе интервьюеров для полевых работ на собеседовании оценивались умение устанавливать первоначальный контакт и зондирование. Результат собеседования – в таблице. Оценки находятся в диапазоне от 2 до 5 (максимум). Определить, есть ли связь между оценками. Доверительная вероятность 95%.

Умение устанавливать контакт	Зондирование			
	Пять	Четыре	Три	Два
Пять	24	40	11	3
Четыре	13	22	18	10
Три	18	21	13	11
Два	16	10	18	4

Решение.

Выдвинем гипотезы

H_0: связи между оценками нет

H_1: связь между оценками есть.

Проведем одностороннюю проверку, поскольку применяется критерий χ^2. Построим на основе заданной таблицу

Умение устанавливать контакт	Зондирование				
	Пять	Четыре	Три	Два	Сумма
Пять	24	40	11	3	78
Четыре	13	22	18	10	63
Три	18	21	13	11	63
Два	16	10	18	4	48
Сумма	71	93	60	28	252

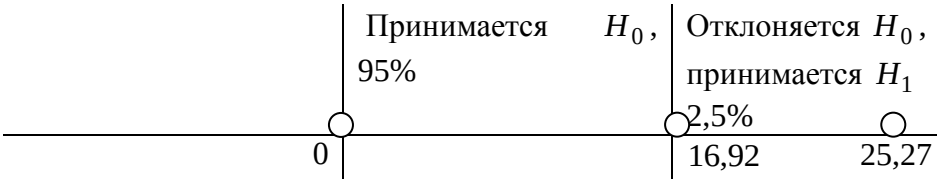
Вычислим значение статистики

$$\chi^2 = n\left(\sum_{i,j=1}^{m,k}\frac{n_{ij}^2}{n_{i\cdot}n_{\cdot j}} - 1\right) = 252\cdot\left(\frac{24^2}{71\cdot 78} + \frac{40^2}{93\cdot 78} + \frac{11^2}{60\cdot 78} + \frac{3^2}{28\cdot 78} + \right.$$

$$+\frac{13^2}{71\cdot 63} + \frac{22^2}{93\cdot 63} + \frac{18^2}{60\cdot 63} + \frac{10^2}{28\cdot 63} + \frac{18^2}{71\cdot 63} + \frac{21^2}{93\cdot 63} + \frac{13^2}{60\cdot 63} +$$

$$\left.+\frac{11^2}{28\cdot 63} + \frac{16^2}{71\cdot 48} + \frac{10^2}{93\cdot 48} + \frac{18^2}{60\cdot 48} + \frac{4^2}{28\cdot 48} - 1\right) \approx 25{,}27.$$

Доверительная вероятность p=0,95, уровень значимости $\alpha = 1 - p = 0{,}05$, $r = (m-1)\cdot(n-1) = (4-1)\cdot(4-1) = 9$. По таблице χ^2 распределения находим $\chi^2_{0{,}05;9} = 16{,}92$.



Отклоняется H_0, принимается H_1 на уровне значимости 5%.

Задача 28. При отборе интервьюеров для полевых работ на собеседовании оценивались умение устанавливать первоначальный контакт и зондирование. Результат собеседования – в таблице. Оценки находятся в диапазоне от 2 до 5 (максимум). Определить, есть ли связь между оценками. Доверительная вероятность 95%.

Умение устанавливать контакт	Зондирование			
	Пять	Четыре	Три	Два
Пять	20	18	13	7
Четыре	23	14	16	6
Три	22	18	21	13
Два	10	7	6	19

§17. Порядковые испытания.

Ранее мы имели дело с данными, для которых можно было провести измерение. В настоящем параграфе рассматриваются порядковые испытания, данные в них называются порядковыми. Составляются два набора длины $n \geq 10$ и требуется проверить, существует ли связь между ними. При этом задается доверительная вероятность p, уровень значимости $\alpha = 1 - p$. Формулируются гипотезы

H_0: связи между наборами нет

H_1: связь между наборами есть.

Находим граничную точку z_α по таблице. Затем по данным наборов вычисляем ранговый коэффициент корреляции Спирмена $r_s = 1 - 6\dfrac{\sum d^2}{n(n^2-1)}$. Тогда статистика определяется формулой $z = r_s\sqrt{n-1}$.

Пример 29. Два человека дегустируют 10 сортов кофе. Эти сорта каждый расположил в порядке убывания предпочтений по 10 балльной системе. Проверить наличие связи между результатами. Доверительная вероятность p=98%.

Сорт кофе	Дегустатор 1	Дегустатор 2
1	6	7
2	5	6
3	10	8
4	5	6
5	2	1
6	7	7
7	4	3
8	8	9
9	9	10
10	6	5

Решение.

Озаглавим столбец, в который записывается разность между результатами первого и второго дегустатора через d. Каждый элемент четвертого столбца возведем в квадрат и обозначим столбец d^2. Выдвинем гипотезы

H_0: связи между результатами исследований нет

H_1: связь между результатами исследований есть.

Сорт кофе	Дегустатор 1	Дегустатор 2	d	d^2
1	6	7	-1	1
2	5	6	-1	1
3	10	8	2	4
4	5	6	-1	1
5	2	1	1	1
6	7	7	0	0
7	4	3	1	1
8	8	9	-1	1
9	9	10	-1	1
10	6	5	1	1
Сумма				12

Ранговый коэффициент корреляции Спирмена равен

$$r_s = 1 - 6\frac{\sum d^2}{n(n^2-1)} = 1 - 6\frac{12}{10(10^2-1)} \approx 0{,}93.$$ Статистика

$z = r_s\sqrt{n-1} = 0{,}93 \cdot 3 = 2{,}79$.

Теперь определим граничную точку. p=0,98, $\alpha = 1 - p = 1 - 0{,}98 = 0{,}02 \Rightarrow z_\alpha = z_{0{,}02} = 2{,}053749$

Отклоняется H_0, принимается H_1 на уровне значимости 2%. Между результатами исследований существует связь.

Пример 29. Два человека дегустируют 10 сортов кофе. Эти сорта каждый расположил в порядке убывания предпочтений по 10 балльной системе. Проверить наличие связи между результатами. Доверительная вероятность p=97%.

Сорт кофе	Дегустатор 1	Дегустатор 2
1	1	4
2	5	8
3	7	3
4	6	1
5	4	10
6	4	8
7	6	4
8	7	3
9	5	3
10	8	10

Контрольная работа №2

Задание 1.

Производитель утверждает, что доля бракованных изделий не превосходит $\hat{p}$ %. В случайной выборке объема n изделий оказалось m бракованных изделий. Не противоречит ли это утверждению производителя? Доверительная вероятность p%.

	1	2	3	4	5
$\hat{p}$	3	5	2	6	8
n	100	120	110	105	97
m	4	6	5	8	9
p	95	97	99	98	95

Задание 2.

Для выборки объема n средний годовой доход на человека составил $\overline{X}$ тыс.руб., выборочное стандартное отклонение s тыс.руб. Годовой доход очередного респондента $X_{нов}$тыс.руб. Можно ли утверждать, что он принадлежит той же целевой группе. Считаем, что годовой доход распределен нормально. Доверительная вероятность p%

	1	2	3	4	5
$\overline{X}$	250	300	350	400	450
s	0,6	0,5	0,7	0,4	0,8
$X_{нов}$	220	305	345	402	440
p	95	97	99	98	95

Задание 3.

Проводились испытания новой вакцины. В эксперименте участвовали n_1 мужчин и n_2 женщин. Побочные эффекты возникли у m_1 мужчин и m_2 женщин. Можно ли утверждать, что побочные эффекты после использования вакцины возникают чаще у женщин. Доверительная вероятность p%

	1	2	3	4	5
n_1	2000	2500	3000	3500	3300
n_2	1800	2700	3200	3600	3700
m_1	40	70	100	110	120
m_2	45	75	110	120	125
p	95	97	99	98	96

Задание 4.

На обработку каждой из n_1 анкет первым способом затрачено в среднем $\overline{X_1}c$, выборочная дисперсия $s_1^2c^2$, а каждой из n_2 анкет вторым способом - $\overline{X_2}c$, выборочная дисперсия $s_2^2c^2$. Можно ли сделать вывод, что при обработке анкет первым способом на обработку одной анкеты в среднем требуется больше времени. Доверительная вероятность p%.

	1	2	3	4	5
n_1	50	55	60	65	70
n_2	40	48	55	60	68
$\overline{X_1}$	20	22	24	25	30
$\overline{X_2}$	18	23	21	24	29
s_1^2	1	2	1,5	3	4

s_2^2	2	1	3	2	4
p	95	97	99	98	96

§18. λ - критерий Колмогорова – Смирнова.

Этот критерий применяется для проверки гипотезы о распределении непрерывной случайной величины. А именно, сравниваются функции распределения эмпирическая $F_э(x)$ и предполагаемая $F(x)$. Для этого

1.Произведем выборку объема $n \geq 50$.

2.Построим эмпирическую функцию $F_э(x)$.

3.По данным выборки построим предполагаемую функцию распределения.

4.Определи значение статистики по формуле $\lambda = \max_{x_i} |F(x_i) - F_э(x_i)| \sqrt{n}$.

5.По уровню значимости $\alpha = 1 - p$ по таблице

α	0,20	0,10	0,05	0,02	0,01	0,001
λ_α	1,073	1,224	1,358	1,520	1,627	1,950

6. Если $\lambda < \lambda_\alpha$, различия между эмпирическим и предполагаемым распределениями несущественны. В противном случае различия существенны.

Пример 29. Респондентам задали вопрос, сколько телевизионных программ они внимательно смотрели за последнюю неделю. Результат приведен в таблице

Число программ	1	2	3	4	5	6	7	8	9
Частота	20	10	11	8	9	12	15	11	13

Определить с помощью λ - критерия Колмогорова-Смирнова на уровне значимости $\alpha = 0{,}05$, согласуются ли данные выборки с равномерным распределением на отрезке $[0{,}10]$.

Решение. Выдвинем две гипотезы

H_0 : различия несущественны

H_1: различия существенны.

Известно, что функция распределения случайной величины, равномерно распределенной на отрезке $[0,10]$, имеет вид $F(x)=\begin{cases}0, x\le 0\\ x/10, 0<x\le 10\\ 1, x>10\end{cases}$.

Тогда таблица будет иметь вид

x_i	n_i	n_i/n	$F_э(x_i)$	$F(x_i)=0{,}1x_i$	$\lvert F(x_i)-F_э(x_i)\rvert$
1	20	0,18	0,18	0,1	0,08
2	10	0,09	0,27	0,2	0,07
3	11	0,1	0,37	0,3	0,07
4	8	0,07	0,44	0,4	0,04
5	9	0,08	0,52	0,5	0,02
6	12	0,11	0,63	0,6	0,03
7	15	0,14	0,77	0,7	0,07
8	11	0,1	0,87	0,8	0,07
9	13	0,13	1	0,9	0,1
Сумма	109				

Найдем теперь наибольшее значение в последнем столбце $\max\limits_{x_i}\lvert F(x_i)-F_э(x_i)\rvert=0{,}1$. Тогда статистика $\lambda=\max\limits_{x_i}\lvert F(x_i)-F_э(x_i)\rvert\sqrt{n}=0{,}1\cdot\sqrt{109}=1{,}044$. Зная уровень значимости $\alpha=0{,}05$, определим $\lambda_\alpha=1{,}358$. Получили, что $\lambda<\lambda_\alpha$, поэтому принимаем гипотезу H_0 на уровне значимости $\alpha=0{,}05$. Это означает, что данные выборки равномерно распределены на отрезке $[0,10]$.

Задача 29. Респондентам задали вопрос, сколько телевизионных программ они внимательно смотрели за последнюю неделю. Результат приведен в таблице

Число программ	1	2	3	4	5	6	7	8	9
Частота	18	11	12	11	8	13	14	10	10

Определить с помощью λ- критерия Колмогорова-Смирнова на уровне значимости $\alpha = 0{,}05$, согласуются ли данные выборки с равномерным распределением на отрезке $[0{,}10]$.

Оглавление

Список литературы

1. Просветов Г.И. Социологические исследования: задачи и решения: Учебно-практическое пособие.- М.: Изд-во «Альфа- Пресс», 2009.- 208 с.