Лабораторная работа МС

По данной выборке построить:

1) статистическое распределение, разбив выборку на $log_2 n + 1$ равных интервалов;
2) найти выборочное среднее, исправленное выборочное среднее квадратическое отклонение;
3) доверительный интервал для оценки с надёжностью γ неизвестного математического ожидания a, если генеральное среднее квадратическое отклонение σ равно «исправленному» среднему квадратическому отклонению s;
4) найти минимальный объём выборки при котором с надёжностью γ точность оценки математического ожидания a генеральной совокупности по выборочной средней равна $\delta = 0{,}5$;
5) построить полигон относительных частот;
6) выдвинуть гипотезу о законе распределения генеральной совокупности;
7) сравнить эмпирические и теоретические частоты с помощью критерия Пирсона при уровне значимости $\alpha = 0{,}05$ и $\alpha = 0{,}01$;
8) дать заключение по результатам анализа.

Краткая теоретическая справка

Пусть из генеральной совокупности извлечена выборка, причем x_1 наблюдалось n_1 раз, x_2 — n_2 раз, x_k — n_k и раз и $\sum n_i = n$ — объем выборки. Наблюдаемые значения x_i — называют *вариантами,* а последовательность вариант, записанных в возрастающем порядке, — *вариационным рядом.* Числа наблюдений называют *частотами,* а их отношения к объему выборки $n_i/n = W_i$ - *относительными частотами.*

Статистическим распределением выборки называют перечень вариант и соответствующих им частот или относительных частот. Статистическое распределение можно задать также в виде последовательности интервалов и соответствующих им частот (в качестве частоты, соответствующей интервалу, принимают сумму частот, попавших в этот интервал).

Полигоном частот называют ломаную, отрезки которой соединяют точки (x_1,n_1), $(x_2,n_2),\ldots,(x_k,n_k)$. Для построения полигона частот на оси абсцисс откладывают варианты x_i, а на оси ординат

— соответствующие им частоты n_i. Точки (x_i, n_i) соединяют отрезками прямых и получают полигон частот.

В зависимости от объёма выборки число группировки k берётся от 6 до 20. Для определения числа равных интервалов k, на которые следует разбить весь диапазон значений признака X [x_{min}, x_{max}], можно пользоваться формулой: $k = log_2 n + 1$, где n – объём статистической совокупности.

Полигоном относительных частот называют ломаную, отрезки которой соединяют точки (x_1, W_1), (x_2, W_2),…,(x_k, W_k).

Выборочной средней $\overline{x}_B$ называют среднее арифметическое значение признака выборочной совокупности.

Если все значения x_1, x_2,..., x_n признака выборки объема n различны, то

$\overline{x}_B = (x_1 + x_2 + \ldots + x_n)/n$. .

Если же значения признака x_1, x_2,..., x_k имеют соответственно частоты n_1, n_2, ..., n_k причем $n_1 + n_2 + \ldots + n_k = n$, то $\overline{x}_B = \left(\sum_{i=1}^{k} n_i x_i \right) \Big/ n$

Выборочной дисперсией D_B называют среднее арифметическое квадратов отклонений значений признака генеральной совокупности от их среднего значения $\overline{x}_B$.

Если все значения x_1, x_2,..., x_n признака выборки объема n различны, то

$$D_B = \left(\sum_{i=1}^{n} (x_i - \overline{x}_B)^2 \right) \Big/ n = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n} \sum_{i=1}^{n} x_i \right)^2 .$$

Если же значения признака x_1, x_2,..., x_k имеют соответственно частоты n_1, n_2,..., n_k, причем $n_1 + n_2 + \ldots + n_k = n$, то

$$D_B = \left(\sum_{i=1}^{k} n_i (x_i - \overline{x}_B)^2 \right) \Big/ n = \frac{1}{n} \sum_{i=1}^{k} n_i x_i^2 - \left(\frac{1}{n} \sum_{i=1}^{k} n_i x_i \right)^2 .$$

Исправленная дисперсия, которую обозначают через s^2:

$$s^2 = \frac{n}{n-1} D_B = \frac{\sum_{i=1}^{k} n_i (x_i - \overline{x}_B)^2}{n-1} .$$

Исправленная дисперсия является, несмещенной оценкой генеральной дисперсии.

Для оценки среднего квадратического отклонения генеральной совокупности используют «исправленное» среднее квадратиче-

ское отклонение, которое равно квадратному корню из исправленной дисперсии $\sigma = s$.

Асимметрия и эксцесс эмпирического распределения определяются равенствами $a_s = m_3/\sigma^3_в$, $e_k = m_4/\sigma^4_в - 3$;
здесь $\sigma_в$ – выборочное среднее квадратическое отклонение; m_3 и m_4 – центральные эмпирические моменты третьего и четвёртого порядков:

$$m_3 = \left(\sum n_i (x_i - \overline{x}_B)^3\right)/n,\ m_4 = \left(\sum n_i (x_i - \overline{x}_B)^4\right)/n,$$

или

$$m_3 = M_3 - 3M_2M_1 + 2M_1^3,\ m_4 = M_4 - 4M_3M_1 + 6M_2M_1^2 - 3M_1^4,$$

где $M_k = \left(\sum n_i x_i^k\right)/n$ – начальные моменты k-го порядка.

Доверительный интервал покрывающий неизвестный параметр a с надёжностью γ: $(\overline{x}_B - t\sigma/\sqrt{n}; \overline{x}_B + t\sigma/\sqrt{n})$; точность оценки $\delta = t\sigma/\sqrt{n}$.

Число t определяется из равенства $2\Phi(t) = \gamma$, или $\Phi(t) = \gamma/2$; по таблице функции Лапласа находят аргумент t, которому соответствует значение функции Лапласа, равное $\gamma/2$.

Замечание. Если требуется оценить математическое ожидание с наперед заданной точностью δ и надежностью γ, то минимальный объем выборки, который обеспечит эту точность, находят по формуле $n = t^2\sigma^2/\delta^2$.

Эмпирическими частотами называют фактически наблюдаемые частоты n_i.

Пусть имеются основания предположить, что изучаемая величина X распределена по некоторому определенному закону. Чтобы проверить, согласуется ли это предположение с данными наблюдений, вычисляют частоты наблюдаемых значений, т. е. находят теоретически частоту n_i' каждого из наблюдаемых значений в предположении, что величина X распределена по предполагаемому закону.

Выравнивающими (теоретическими) в отличие от фактически наблюдаемых эмпирических частот называют частоты n_i', найденные теоретически (вычислением). Выравнивающие частоты находят с помощью равенства $n_i' = nP_i$, где n — число испытаний; P_i — вероятность наблюдаемого значения x_i, вычисленная при допущении, что X имеет предполагаемое распределение.

В частности, если имеются основания предположить, что случайная величина X (генеральная совокупность) распределена нормально, то выравнивающие частоты могут быть найдены по формуле $n_i' \approx \frac{nh}{\sigma_B}\varphi(u_i)$, где n — число испытаний (объем выборки), h — длина частичного интервала, σ_B — выборочное среднее квадратическое отклонение, $u_i = (x_i - \overline{x}_B)/\sigma_B$ (x_i — середина i-го частичного интервала), $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$.

χ^2 - критерий Пирсона

Назначения критерия

Критерий χ^2 применяется в двух целях;

1) для сопоставления *эмпирического* распределения признака с *теоретическим* - равномерным, нормальным или каким-то иным;
2) для сопоставления *двух, трех или более эмпирических* распределений одного и того же признака.

Описание критерия

Критерий χ^2 отвечает на вопрос о том, с одинаковой ли частотой встречаются разные значения признака в эмпирическом и теоретическом распределениях или в двух и более эмпирических распределениях.

При сопоставлении эмпирического распределения с теоретическим мы определяем степень расхождения между эмпирическими и теоретическими частотами.

Чем больше расхождение между двумя сопоставляемыми распределениями, *тем больше* эмпирическое *значение* χ^2.

Гипотезы **(**для сопоставления *эмпирического* распределения признака с *теоретическим***)**

H_0: Полученное эмпирическое распределение признака не отличается от теоретического (например, нормального) распределения.

H_1: Полученное эмпирическое распределение признака отличается от теоретического распределения.

Ограничения критерия

1. Объем выборки должен быть достаточно большим: $n \geq 30$. При
   $n < 30$ критерий χ^2 дает весьма приближенные значения. Точность критерия повышается при больших n.
2. Теоретическая частота для каждой ячейки таблицы не должна быть меньше 5: $n_i' \geq 5$. Это означает, что если число разрядов задано заранее и не может быть изменено, то мы не можем применять метод χ^2, не накопив определенного минимального числа наблюдений.
3. Выбранные разряды должны «вычерпывать» все распределение, то есть охватывать весь диапазон вариативности признаков. При этом группировка на разряды должна быть одинаковой во всех сопоставляемых распределениях.
4. Разряды должны быть неперекрещивающимися: если наблюдение отнесено к одному разряду, то оно уже не может быть отнесено ни к какому другому разряду.
5. Сумма наблюдений по разрядам всегда должна быть равна общему количеству наблюдений.

Алгоритм расчета критерия χ^2

1. Занести в таблицу данные статистического распределения: наименования разрядов и соответствующие им эмпирические частоты (n_i).
2. Рядом с каждой эмпирической частотой записать теоретическую частоту (n_i').
3. Подсчитать разности между эмпирической и теоретической частотой по каждому разряду (строке) и записать их в третий столбец.
4. Определить число степеней свободы по формуле: $v=k-3$, где k - количество разрядов признака (при сравнении с нормальным распределением).
5. Возвести в квадрат полученные разности и занести их в четвертый столбец.
6. Разделить полученные квадраты разностей на теоретическую частоту и записать результаты в пятый столбец.

7. Просуммировать значения пятого столбца. Полученную сумму обозначить как $\chi^2_{эмп}$

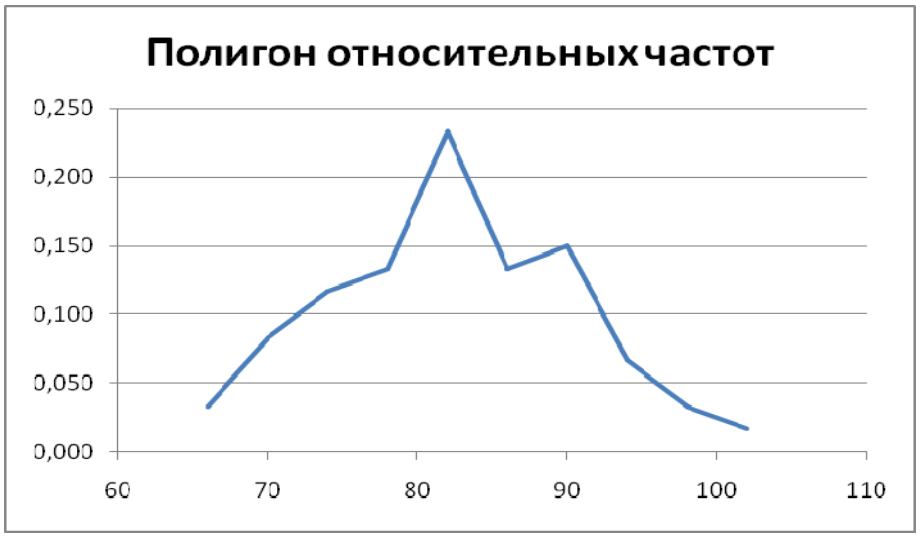
$$\chi^2 = \sum_{j=1}^{k} \frac{(n_j - n'_j)^2}{n'_j}$$

8. Определить по таблице критических точек распределения χ^2 для данного числа степеней свободы v $\chi^2_т$. Если $\chi^2_{эмп}$ меньше критического значения, расхождения между распределениями статистически недостоверны. Различия между двумя распределениями могут считаться достоверными, если $\chi^2_{эмп}$ достигает или превышает $\chi^2_{0.05}$- и тем более достоверными, если $\chi^2_{эмп}$ достигает или превышает $\chi^2_{0.01}$.

Критические значения $\chi^2_{теор}$ могут быть найдены с помощью стандартной функции Ms Excel: ХИ2ОБР(вероятность; число степеней свободы).

Пример нахождения точечных и интервальных оценок с помощью MS Excel

86	81	76	80	84	85			
95	77	85	95	89	83			
82	77	76	71	87	68			
89	64	81	90	72	97			
91	75	80	79	85	83			
78	94	87	103	70	87			
90	70	82	99	81	89			
84	79	78	74	81	75			
81	76	73	81	89	93			
89	85	83	92	84	72			
X_{min}=	64	64						
X_{max}=	103	104						
R=	39	40						
h=		4						
$X_л$	$X_{пр}$	X_c	n_i	v_i	$X_c n_i$	$X^2{}_c n_i$	$X^3{}_c n_i$	$X^4{}_c n_i$
64	68	66	2	0,033	132	8712	574992	37949472
68	72	70	5	0,083	350	24500	1715000	120050000
72	76	74	7	0,117	518	38332	2836568	209906032
76	80	78	8	0,133	624	48672	3796416	296120448
80	84	82	14	0,233	1148	94136	7719152	632970464
84	88	86	8	0,133	688	59168	5088448	437606528
88	92	90	9	0,150	810	72900	6561000	590490000
92	96	94	4	0,067	376	35344	3322336	312299584
96	100	98	2	0,033	196	19208	1882384	184473632
100	104	102	1	0,017	102	10404	1061208	108243216
			60		**4944**	**411376**	**3,5E+07**	**2,93E+09**



$X_{выб}$=	82,40	M_1=	82,4	m_3=	41,728		
$D_{выб}$=	66,51	M_2=	6856,27	m_4=	11176		
S^2=	67,63	M_3=	575958,4				
$\sigma_{выб}$=	8,16	M_4=	48835156,3				
S=	8,22						
A_s=	0,08						
E_s=	-0,47						
γ=	0,90						
t=	1,64		Доверительный интервал				
δ=	1,75	(	80,65	84,15	)		
γ=	0,95						
t=	1,96						
δ=	0,50						
N=	1039,25						

Минимальный объем выборки, обеспечивающий заданную точность: 1040

Замечание. Для нахождения абсолютных частот в Ms Excel можно воспользоваться функцией =ЧАСТОТА(A1:F10;B18:B27), в которой первый аргумент – блок в котором располагаются исходные данные, второй аргумент – правые границы частичных интервалов. Функция матричная, для корректного её использования необходимо воспользоваться следующим алгоритмом: а) выделяется блок, в котором должны располагаться результаты; б) вводится формула; в) завершается операция нажатием [Ctrl]+[Shift]+[Enter]. Найти решение уравнения $\Phi(t) = \gamma/2$ относительно переменной t можно найти по таблице или воспользовавшись функцией =НОРМОБР(0,5+НАДЕЖНОСТЬ/2;0;1), где НАДЕЖНОСТЬ – значение γ.