# Описание данных

** **

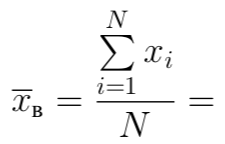
**Задача работы:** Исследовать распределение генеральной совокупности и оценить ее параметры распределения.

**Описание данных:** В качестве примера был выбран средне-душевой доход руб/мес по областям России. Для примера исследуем распределение чисел средне-душевого дохода

# 2. Вариационный ряд и вычисление его выборочных характеристик

Будем рассматривать полученные данные как реализации N независимых одинаково распределенных случайных величин. Составим вариационный ряд для полученных данных. По этому ряду найдем его характеристики

**Выборочное среднее** — начальный выборочный момент первого порядка, т.е. величина, равная среднему арифметическому членов ряда.



**Медиана** – значение, которое характеризует центр порядковой статистики. Для ряда из N элементов, медиана будет равна:



ЖОБАВИТЬ. ИЗ 2

**Мода** – характеристика процесса, показывающая его положение. Мода – наиболее части встречающееся значение из изучаемого множества.

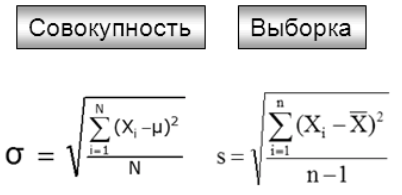
**Выборочное среднеквадратическое отклонение (Стандартная ошибка)** — величина, равная квадратному корню выборочной дисперсии



**Стандартное отклонение выборки** - это мера того, насколько широко разбросаны значения в выборке относительно их среднего.

**Стандартное отклонение (среднеквадратичное отклонение от среднего)** – мера вариации данных, важнейшая числовая характеристика распределения вероятностей случайной величины, обозначаемая символом σ (сигма). Стандартное отклонение является характеристикой процесса, показывающей меру его разброса. Стандартное отклонение может быть расчитана точно, только если известны значения всей генеральной совокупности. Ввиду того, что все элементы совокупности известны в редких случаях, то задача исследователя сводится к оценке стандартного отклонения по выборке из генеральной совокупности. Оценка стандартного отклонения по выборке, также называемая выборочным стандартным отклонением, обозначается символом  или s

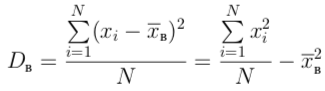
Формулы расчета стандартного отклонения:



где Хi - значения элементов совокупности или выбрки;  
µ - математическое ожидание;  
n - количество элементов в выборке;  
n - количество элементов в выборке;  
N - количество элементов в совокупности;  
 - среднее значение выборки.

**Выборочная дисперсия** — центральный выборочный момент второго порядка

**Дисперсия** – мера вариации данных, важнейшая числовая характеристика распределения вероятностей случайной величины, обозначаемая символом σ2 (сигма квадрат). Дисперсия является характеристикой процесса, показывающей меру его разброса. Также как и для стандартного отклонения дисперсия может быть расчитана для генеральной совокупности (тогда это будет точное значение дисперсии процесса) и для выборки (тогда это будет оценочное значение дисперсии процесса, обозначаемое символом  или s2).



Выборочная дисперсия и выборочное среднеквадратическое отклонение не являются несмещенными и эффективными оценками.



# 3. Интервальный вариационный ряд

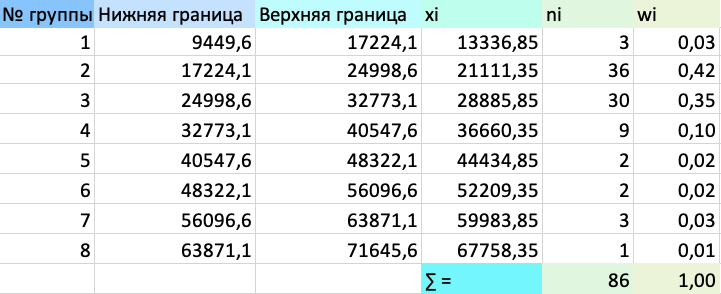
Составим интервальный ряд по данной выборке. Чтобы построить интервальный ряд, возьмем количество интервалов, рассчитанное по формуле Стерджесса :

1 +3.2 log2 86 ≈ 8

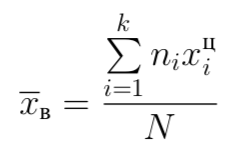
разобъём исходные данные по интервалам с шагом = 7774,46

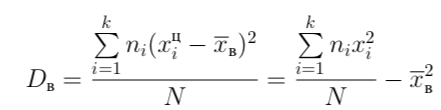
Начальная точка для интервалов : 9449,6

конечная точка : 71645,6



Выборочное среднеe, выборочная дисперсия и выборочное среднеквадратическое отклонение для интервального ряда вычисляются по формулам:

 = 28343,44302

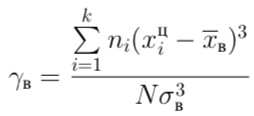
= 109346313,7

= 10456,8788

где ni — число элементов выборки, попавших в i-ый интервал, xцi — центр i-ого интервала, k — число интервалов.

**Коэффициент асимметрии** — это величина, определяющая степень несимметричности распределения случайной величины.

Выборочный коэффициент асимметрии вычисляется как отношение центрального выборочного момента 3-го порядка к кубу выборочного среднеквадратического отклонения, т.е.

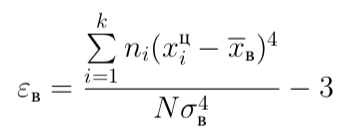


**Отрицательный коэффиициент** асимметрии показывает, что распределение более

плотное слева от центра. **Положительный** — что распределение более плотное справа.

**Коэффициент эксцесса** — это величина, которая характеризует степень остроты пика расределения случайной величины. Коэффициенты асимметрии и эксцесса для нормального распределения равны 0.

Выборочный коэффициент эксцесса определяется следующей формулой :



Вычитание 3 требуется для того, чтобы коэффициент эксцесса нормального распре- деления был равен 0.

**Положительный коэффициент** эксцесса показывает, что пик распределения более острый по сравнению с пиком нормального распределения.  
**Отрицательный̆** — пик более сглаженный̆.

# 4. Проверка гипотез о виде распределения

Для проверки гипотез о принадлежности выборки некоторому закону распределения воспользуемся критерием согласия Пирсона.

Данный критерий позволяет оценить значимость различий между наблюдаемыми данными и теоретическими ожиданиями, полученными в предположении о верности нулевой гипотезы, при заданном уровне значимости. Статистика критерия вычисляется по следующей формуле:



где mi = piN — теоретическая частота попаданий в i-ый интервал, ni — наблюдаемая частота, k — число интервалов.

Данная величина имеет распределение, близкое к распределению χ2 c (k – 1 степенями свободы.

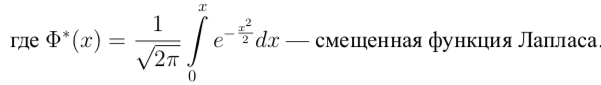
Вычислим теоретические вероятности попадания величин в заданные интервалы для трех видов распределений:

1. H0: распределение случайной величины ξ - нормальное

Вероятность попадания случайной величины, распределенной по нормальному

закону N (m, σ), в интервал (a, b) вычисляется по формуле :





Рассчитаем вероятности для каждого интервала по ф-ле

1)

0,1068

9,187461935

20,22976983

25,5406382

0,2149

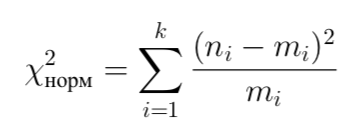
18,48919333

7,674491278

1,826529307

0,249258001

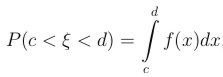
0,019503699

= 105,9699848

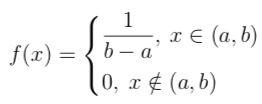
1. H0: распределение случайной величины ξ - равномерное

Вероятность попадания случайной величины, распределенной равномерно на

интервале (c, d), в интервал (a, b) вычисляется по формуле :



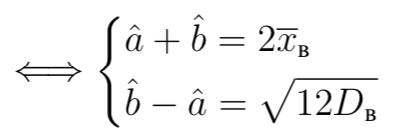
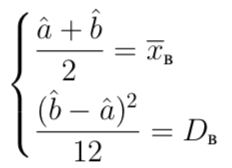
где f(x) — функция плотности вероятности равнормерного распределения



Оценки параметров равномерного распределения найдём методом моментов.

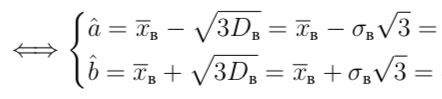
Метод моментов заключается в нахождении состоятельных оценок параметров распределения путем приравнивания теоретических начальных и центральных моментов случайной̆ величины к выборочным моментам. Теоретические значения математического ожидания и дисперсии равномерного распределения равны и

Их выборочными оценками являются выборочное среднее и несмещенная выборочная дисперсия. Получим систему:



a = 10231,5977

B = 46455,28835



Теперь мы можем вычислить точный вид функции плотности вероятности равномерного распределения для нашей выборки:

Вычислим вероятности для каждого из интервалов:

16,8

16,8

18,662

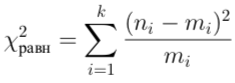
18,662

14,19

0

0

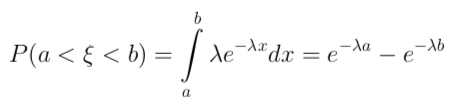
0

 = 56,4057961

1. H0: распределение случайной величины ξ является показательным

Вероятность попадания случайной величины, распределенной показательно, в

интервал (a, b) вычисляется по формуле :



По методу моментов найдем оценку параметра λ. Математическое ожидание

показательного распределения равно Тогда

 = 0.0000352815

Вычислим вероятности для каждого из интервалов:

39,1638792

0,13064725

11,2356639

0,09930588

8,54030525

0,07548308

6,49154464

0,05737521

4,93426764

0,04361129

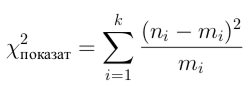
3,7505707

0,03314924

2,85083453

0,10503412

9,03293412

 = 152,58205

ТАБЛ

# 5. Критерий согласия Пирсона

Мы получили оценку χ2 для рассматриваемых в данной работе распределений, теперь необходимо принять решение о распределении нашей выборки. Начнем с критерия согласия Пирсона. Для его проверки нам необходимо задать уровень значимости α - вероятность отвергнуть верную нулевую гипотезу, а также вычислить количество степеней свободы df = k − l − 1, где k - количество непересекающихся интервалов, l - количество оцененных по выборке параметров.

Определим уровень α = 0.05:

**Применим критерий к нормальному распределению:**

Рассчитаем df = 8−2−1 = 6

В табличных критических значений Пирсона при данном уровне значимости находим χ20.05 = 12.6 < 105,96 - гипотеза H0 отклоняется

**Применим критерий к равномерному распределению:**

Рассчитаем df = 8−2−1 = 6

В таблице распределений при данном уровне значимости находим χ20.05 = 12.6 < 56,4057961 - гипотеза H0 отвергается

**Применим критерий к показательному распределению:**

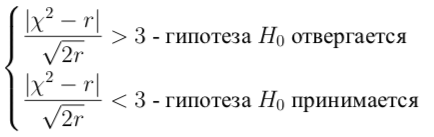
Рассчитаем df = 8−1−1 = 7

В таблице распределений при данном уровне значимости находим χ20.05 = 14.1 < 152,58205 - гипотеза H0 отвергается

Вывод : Поскольку значения статистики не попадает в область принятия, то согласно критерию 3 гипотезы отвергается с уровнем доверия 0,95. Значит, по критерию согласия Пирсона для нашей выборки точно не подходит ни одно распределение.

# 6. Критерий Романовского

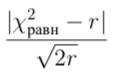
Критерий Романовского позволяет принять или отвергнуть гипотезу о распределении, основываясь на следующих формулах:



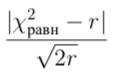
Где r = df = k – l −1 - число степеней свободы χ2

**Критерий Романовского для нормального распределения:**

r = df = 8 − 2 − 1 = 6

 = 28,8588488 > 3

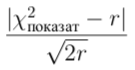
**Критерий Романовского для равномерного распределения:**

r = df = 8 − 2 − 1 = 6

= 14 > 3

**Критерий Романовского для показательного распределения:**

r = df = 8 − 1 − 1 = 7

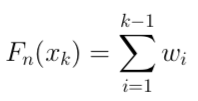
 = 150,711222 > 3

Значит , по критерию Романовского для нашей выборки точно не подходят АПОИАПЛИО распределения.

# 7. Критерий Колмогорова

Критерий согласия Колмогорова, как и критерий Пирсона или критерий Романовского, используется для проверки гипотез о виде распределения. Статистикой критерия является величина

- эмпирическая функция распределения, F ∗ — функция распределения в предположении о верности нулевой гипотезы,



**Для нормального распределения:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| № группы | wi | Fn(x) | pi | Fn\*(x) | |Fn(x)-Fn\*(x)| |
| 1 | 0,03 | 0,03 | 0,1068 | 0,1068 | 0,08 |
| 2 | 0,42 | 0,45 | 0,2352 | 0,342 | 0,11 |
| 3 | 0,35 | 0,80 | 0,2969 | 0,6389 | 0,16 |
| 4 | 0,10 | 0,90 | 0,2149 | 0,8538 | 0,05 |
| 5 | 0,02 | 0,92 | 0,08923 | 0,94303 | 0,02 |
| 6 | 0,02 | 0,94 | 0,0212 | 0,96423 | 0,02 |
| 7 | 0,03 | 0,97 | 0,00289 | 0,96712 | 0,00 |
| 8 | 0,01 | 0,98 | 0,000226 | 0,967346 | 0,01 |

Значит , Dn(x) = supx|Fn(x) − F ∗(x)| = 0,16

= 1,48

(из табл берем) = 0,15 < 0,16 - гипотеза H0 не принимается

**Для показательного распределения:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| № группы | wi | Fn(x) | pi | F\*(x) | | Fn(x) - F\*(x) | |
| 1 | 0,03 | 0,03 | 0,45539394 | 0,45539394 | 0,43 |
| 2 | 0,42 | 0,45 | 0,13064725 | 0,5860412 | 0,14 |
| 3 | 0,35 | 0,80 | 0,09930588 | 0,68534707 | 0,11 |
| 4 | 0,10 | 0,90 | 0,07548308 | 0,76083015 | 0,14 |
| 5 | 0,02 | 0,92 | 0,05737521 | 0,81820536 | 0,10 |
| 6 | 0,02 | 0,94 | 0,04361129 | 0,86181664 | 0,08 |
| 7 | 0,03 | 0,97 | 0,03314924 | 0,89496588 | 0,08 |
| 8 | 0,01 | 0,98 | 0,10503412 | 1 | 0,02 |

Значит , Dn(x) = supx|Fn(x) − F ∗(x)| = 0,43

= 3,98

Найдем критическое значение D- критерия Колмогорова для данного распределения с α = 0.05:   
 = 0,15 < 0,43 - гипотеза H0 не принимается

**Для равномерного распределения:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| № группы | wi | Fn(x) | pi | Fn\*(x) | |Fn(x)-Fn\*(x)| |
| 1 | 0,03 | 0,03 | 0 | 0,19579 | 0,17 |
| 2 | 0,42 | 0,45 | 0,1958 | 0,39159 | 0,06 |
| 3 | 0,35 | 0,80 | 0,2177 | 0,60929 | 0,19 |
| 4 | 0,10 | 0,90 | 0,2176 | 0,82689 | 0,07 |
| 5 | 0,02 | 0,92 | 0,1654 | 0,99229 | 0,07 |
| 6 | 0,02 | 0,94 | 0 | 0,99229 | 0,05 |
| 7 | 0,03 | 0,97 | 0 | 0,99229 | 0,02 |
| 8 | 0,01 | 0,98 | 0 | 0,99229 | 0,01 |

Значит , Dn(x) = supx|Fn(x) − F ∗(x)| = 0,19

= 1,76

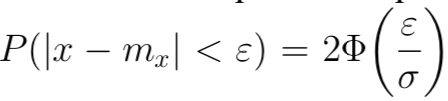
= 0,15 < 0,19 - гипотеза H0 не принимается

На основании полученных результатов мы можем сделать вывод, что по критерию Колмогорова отвергаются все гипотезы.

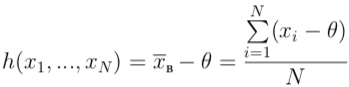
# 8. Доверительные интервалы для параметров распределения

Перейдем от точечных к интервальным оценкам. Доверительный интервал для параметра θ — это такой числовой интервал (a, b), в котором с заданной вероятностью γ содержится θ, т.е. P (θ ∈ (a, b)) = γ.

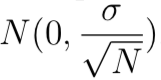
Доверительный интервал для математического ожидания выборки, распределенной по нормальному закону с известным среднеквадратическим отклонением σ, определяется из соотношения



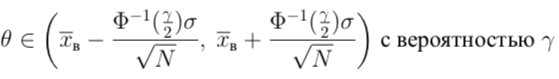
Вводится опорная случайная величина:

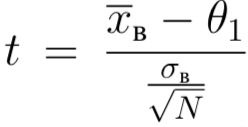


где θ — неизвестное математическое ожидание.

Распределение опорной величины не зависит от параметра θ и подчиняется нормальному закону 

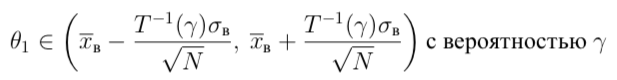
Приравнивая вероятность попадания в интервал к γ, получим выражение для границ доверительного интервала для θ:

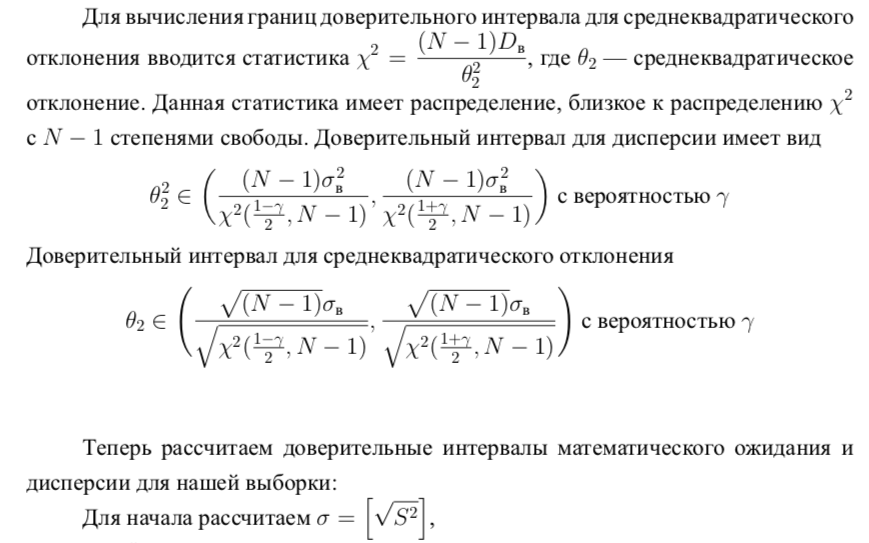


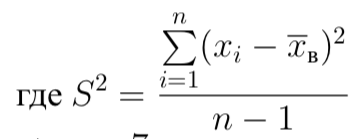
Если среднеквадратическое отклонение неизвестно, в качестве опорной величины выбирается статистика , близкая к распределению Стьюдента с σ

N − 1 степенями свободы.

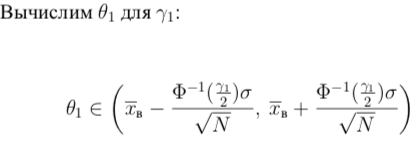
Доверительный интервал имеет вид :

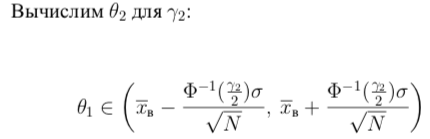


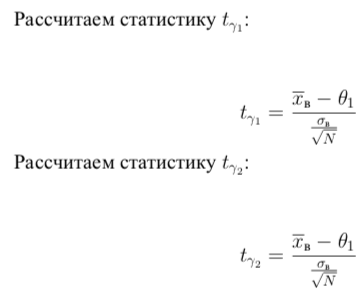


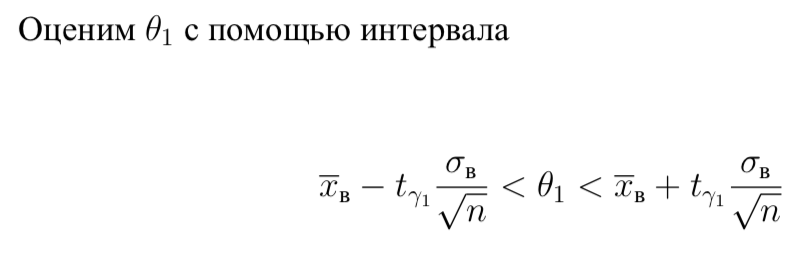
 среднеквадратичное отклонение, ⇒ σ =

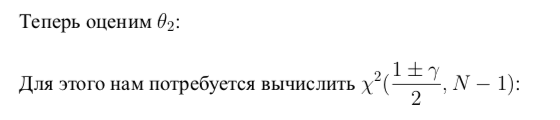
Далее

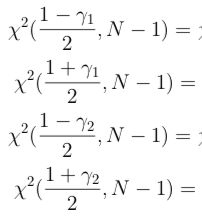




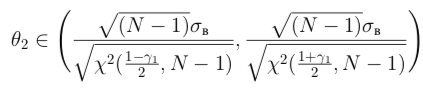


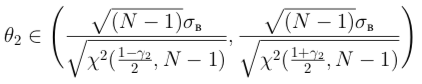






Вычислим значения границ доверительного интервала для разных γ:





|  |  |
| --- | --- |
| S^2 |  |
| 89811239,44 |  |
| S |  |
| 9476,879203 |  |
| интервалы для тета 1 при гамма 1: |  |
| 26254,10319 | 30260,0224 |
| интервалы для тета 1 при гамма 2: |  |
| 25620,51392 | 30893,61166 |
| интервал для статистики t гамма1: |  |
| 26053,81255 | 30460,31303 |
| интервал для статистики t гамма2: |  |
| 25356,86604 | 31157,25954 |
|  |  |
| теперь разберемся с тета 2: |  |
|  |  |
| вот эти 4 разных Хи2: |  |
|  |  |
| 113,5435976 |  |
| 62,23862642 |  |
| 123,5217042 |  |
| 55,97270327 |  |
| интервалы для тета 2 при гамма1: |  |
| 9019,55515 | 12182,50067 |
| интервалы для тета 2 при гамма2: |  |
| 8647,584335 | 12846,30718 |

# 9. Вывод

Итак, проведя данное исследование, можно предположить, что распределение очень похоже на нормальное, но предполагаемая функция , описывающая это распределение как нормальное, дает очень большие ошибки - поэтому эту функцию нельзя рассматривать как нормально распределенную