ДОВЕРИТЕЛЬНЫЕ ИНТЕРВАЛЫ

1. Основные положения

Основной метод математической статистики состоит в том, что для изучения генеральной совокупности объёма N из неё производится выборка, состоящая из n элементов, которая хорошо характеризует всю совокупность (свойство представительности). И на основании исследования этой выборочной совокупности мы с высокой достоверностью можем оценить генеральные характеристики. Чаще всего требуется выявить закон распределения генеральной совокупности и оценить его важнейшие числовые параметры, такие как генеральная средняя $\bar{x}_Г$, генеральная дисперсия $D_Г$ и среднее квадратическое отклонение $\sigma_Г$.

Очевидно, что для оценки этих параметров нужно вычислить соответствующие выборочные значения. Так, выборочная средняя позволяет нам оценить генеральную среднюю, причём оценить её точечно. Почему точечно? Потому что – это отдельно взятое, конкретное значение. Если из той же генеральной совокупности мы будем проводить многократные выборки, то в общем случае у нас будут получаться различные выборочные средние, и каждая из них представляет собой точечную оценку генерального значения. Таким образом, точечная оценка – число, которое используют для оценки параметра генеральной совокупности. Аналогично, несмещённой точечной оценкой генеральной дисперсии является исправленная выборочная дисперсия, и соответственно, стандартного отклонения – исправленное стандартное отклонение.

Недостаток точечных оценок состоит в том, что при небольшом объёме выборки (как оно часто бывает), мы можем получать выборочные значения, которые далеки от истины.

И в этих случаях логично потребовать, чтобы выборочная характеристика θ (средняя, дисперсия или какая-то другая) отличалась от генерального значения $\theta_Г$ менее, чем на некоторое положительное число δ:

$$|\theta_Г - \theta| < \delta \text{ или } \theta - \delta < \theta_Г < \theta + \delta$$

Интервал $(\theta - \delta;\ \theta + \delta)$ называется доверительным интервалом и представляет собой интервальную оценку генерального значения $\theta_Г$ по найденному выборочному значению θ. При этом важно помнить, что для разных выборок одной и той же генеральной совокупности могут получаться разные доверительные интервалы.

Но статистические методы не позволяют 100%-но утверждать, что рассчитанное значение будет удовлетворять указанному неравенству – ведь в статистике всегда есть место случайности. Таким образом, можно говорить лишь о вероятности γ, с которой это неравенство осуществится:

$$P(|\theta_Г - \theta| < \delta) = \gamma$$

Получается, что доверительный интервал с вероятностью γ «накрывает» истинное значение $\theta_Г$. Эта вероятность называется доверительной вероятностью (уровнем доверия) или надёжностью интервальной оценки. В свою очередь величина $\alpha = 1 - \gamma$ называется уровнем значимости. Обычно уровень значимости равен 0.01, 0.05, 0.1, что соответствует уровню доверия 0.99, 0.95, 0.9. Очень часто уровни значимости и доверия измеряются в процентах, то есть уровень доверия 0.99 и 99% — это одно и то же.

Таким образом, интервал со случайными концами $(\theta - \delta;\ \theta + \delta)$ называется доверительным интервалом для параметра θ с уровнем значимости α, если для любого $\theta_0 \in (\theta - \delta;\ \theta + \delta)$ выполняется:

$$P(\theta - \delta < \theta_0 < \theta + \delta) \geq 1 - \alpha$$

2. Доверительный интервал для среднего. Случай известной дисперсии

Важнейшей характеристикой генеральной совокупности является среднее значение. Что же необходимо сделать, чтобы построить для него доверительный интервал?

По ЦПТ среднее значение одинаково распределенных случайных величин стремится к нормальному распределению. Более того, верна следующая теорема:

Если распределение генеральной совокупности имеет конечные математическое ожидание и дисперсию, то при $n \to \infty$ основные выборочные характеристики (среднее, дисперсия, другие моменты) являются нормальными.

Рассмотрим случайную выборку объема n, вычислим среднее значение $\bar{x}$ по выборке и зададим уровень доверия γ. Доверительный интервал для среднего имеет вид $(\bar{x} - \Delta; \bar{x} + \Delta)$, где Δ — это точность интервальной оценки.

Пусть генеральная совокупность имеет нормальное распределение со стандартным отклонением σ. Тогда:

$$\Delta = \frac{\sigma}{\sqrt{n}} z_\alpha$$

где z_α – квантиль стандартного нормального распределения уровня 1-α/2.

Тогда доверительный интервал для среднего с известной дисперсией имеет вид:

$$(\bar{x} - \frac{\sigma}{\sqrt{n}} z_\alpha;\ \bar{x} + \frac{\sigma}{\sqrt{n}} z_\alpha)$$

Задача 1. Дана выборка 9, 5, 7, 7, 4, 10, дисперсия $\sigma^2 = 1$. Постройте 99% доверительный интервал.

Решение.

Ищем среднее значение выборки: $\bar{x} = 7$.

В нашей задаче α = 0.01. По таблице нормального распределения находим 1-α/2=0.995 и и определяем квантиль $z_\alpha = 2.58$.

Теперь можем найти точность: Δ = 1.05. Тогда доверительный интервал с уровнем значимости 0.01 имеет вид: (7 − 1.05; 7 + 1.05) = (5.95; 8.05).

```
import numpy as np
import scipy.stats as stats
x = np.array([9, 5, 7, 7, 4, 10])
n = len(x)
sigma = 1
alpha = 0.01
x0 = x.mean()
z = abs(stats.norm.ppf(alpha/2))
print(f'Доверительный интервал: ({round(x0-sigma/n**0.5*z,2)} ; {round(x0+sigma/n**0.5*z,2)})')
```

Задача 2. Пусть для выборки объема n = 25 вычислено среднее $\bar{x} = 130$. Из предыдущих исследований известно стандартное отклонение σ = 12. Постройте 98% доверительный интервал для среднего значения.

Решение.

```
import numpy as np
import scipy.stats as stats
n = 25
sigma = 12
alpha = 0.02
x0 = 130
z = abs(stats.norm.ppf(alpha/2))
print(f'Доверительный интервал: ({round(x0-sigma/n**0.5*z,2)} ; {round(x0+sigma/n**0.5*z,2)})')
```

Благодаря тому, что мы знаем формулу для доверительного интервала, можно решить задачу: найти минимальный необходимый объем выборки для того, чтобы с заданной точностью Δ и уровнем значимости α найти среднее значение. Для этого достаточно применить формулу:

$$n = \left(\frac{z_\alpha \sigma}{\Delta}\right)^2$$

Фактически это правило, которое определить требуемый объем выборки при проведении исследований.

3. Доверительный интервал для среднего. Случай неизвестной дисперсии и большой выборки (n>30)

Если выборка больше 30, но стандартное отклонение нам неизвестно, то вместо σ мы будем использовать выборочное стандартное отклонение:

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

То есть решение аналогично случаю известной дисперсии, только вместо отклонения по генеральной совокупности используется отклонение по выборке, поскольку для больших выборок (n>30) разность между этими отклонения становится пренебрежимо малой.

Доверительный интервал для среднего при неизвестной дисперсии, но большой выборке (n > 30), имеет вид:

$$(\bar{x} - \frac{s}{\sqrt{n}} z_\alpha;\ \bar{x} + \frac{s}{\sqrt{n}} z_\alpha)$$

Задача 1. Дана выборка: 9, 5, 7, 7, 4, 10, 2, 4, 6, 8, 3, 6, 2, 1, 3, 6, 3, 4, 2, 6, 4, 5, 3, 5, 4, 5, 6, 7, 5, 6, 8, 3, 4, 7, 1, 0, 5, 9, 7, 4. Постройте 98% доверительный интервал.

Решение.

```
import numpy as np
import scipy.stats as stats
import pandas as pd

x = np.array([9, 5, 7, 7, 4, 10, 2, 4, 6, 8, 3, 6, 2, 1, 3, 6, 3, 4, 2, 6, 4, 5, 3, 5, 4, 5, 6, 7, 5, 6, 8, 3, 4, 7, 1, 0, 5, 9, 7, 4])
df = pd.DataFrame(x)
df.hist(bins=11)
n = len(x)
s_not_bias = (x.var()*n/(n-1))**0.5
s_bias = x.var()**0.5
alpha = 0.01
x0 = x.mean()
z = abs(stats.norm.ppf(alpha/2))
print(f'Доверительный интервал (несмещенная дисперсия): ({round(x0-s_not_bias/n**0.5*z,2)} ; {round(x0+s_not_bias/n**0.5*z,2)})')
print(f'Доверительный интервал (смещенная дисперсия): ({round(x0-s_bias/n**0.5*z,2)} ; {round(x0+s_bias/n**0.5*z,2)})')
```

4. Доверительный интервал для среднего. Случай неизвестной дисперсии и малой выборки

Самый проблемный случай для любого исследователя, когда выборка маленькая и про её параметры ничего неизвестно. Если дисперсия неизвестна и объем выборки небольшой ($n \leq 30$), тогда вместо нормального распределения теперь используется t-распределение.

Доверительный интервал в этом случае имеет вид:

$$(\bar{x} - \frac{s}{\sqrt{n}} t_\alpha(n-1);\ \bar{x} + \frac{s}{\sqrt{n}} t_\alpha(n-1))$$

Здесь $t_\alpha(n-1)$ — это квантиль распределения Стьюдента уровня $1-\alpha/2$ с $n-1$ степенью свободы (мы ищем это число в таблице t-распределения).

Стоит отметить, что распределение Стьюдента стремится к нормальному распределению при $n \to \infty$, поэтому при больших выборках доверительные интервалы для среднего, посчитанные по любой из наших формул, будут почти совпадать.

Число степеней свободы зависит от того, сколько имеется связей между наблюдениями. Так как мы знаем среднее, то наблюдения связаны одним равенством и степеней свободы становится на одну меньше.

Задача 1.

По выборке объёма n = 17 вычислено выборочное среднее $\bar{x}$ = 20.5 мм и выборочная дисперсия s^2 = 16 мм^2 диаметра валиков. Постройте доверительные интервалы уровня надёжности γ = 0.9 для среднего значения диаметра валика. Предполагается, что диаметры валиков имеют нормальное распределение.

Решение.

```
import numpy as np
import scipy.stats as stats
n = 17
```

```
s_not_bias = (16*n/(n-1))**0.5
s_bias = 16**0.5
alpha = 0.1
x0 = 20.5
t = abs(stats.t.ppf(alpha/2, df=n-1))
print(f'Доверительный интервал (несмещенная дисперсия): ({round(x0-s_not_bias/n**0.5*t,2)} ; {round(x0+s_not_bias/n**0.5*t,2)})')
print(f'Доверительный интервал (смещенная дисперсия): ({round(x0-s_bias/n**0.5*t,2)} ; {round(x0+s_bias/n**0.5*t,2)})')
```

Задача 2.

По группе семей с доходом 154 руб./чел. зафиксированы следующие цифры потребления молока за месяц (на одного человека): 8.3, 8.6, 8.7, 8.8, 9.1, 9.3, 9.4, 13.4, 13.5, 13.8, 13.9, 14.1, 14.3. Найти доверительный интервал для математического ожидания с надежностью γ = 0.95. Выборка произведена из нормальной совокупности.

Решение

```
import numpy as np
import scipy.stats as stats
x = np.array([8.3, 8.6, 8.7, 8.8, 9.1, 9.3, 9.4, 13.4, 13.5, 13.8, 13.9, 14.1, 14.3])
n = len(x)
x0 = x.mean()
alpha = 0.05
s_not_bias = (x.var()*n/(n-1))**0.5
s_bias = x.var()**0.5
t = abs(stats.t.ppf(alpha/2, df=n-1))
print(f'Доверительный интервал (несмещенная дисперсия): ({round(x0-s_not_bias/n**0.5*t,2)} ; {round(x0+s_not_bias/n**0.5*t,2)})')
print(f'Доверительный интервал (смещенная дисперсия): ({round(x0-s_bias/n**0.5*t,2)} ; {round(x0+s_bias/n**0.5*t,2)})')
```

5. *Доверительный интервал для доли*

Следующим популярным параметром, который часто требует оценивания, является доля признака p в генеральной совокупности.

По выборке мы можем определить долю $\hat{p}$ того или иного признака, просто посчитав число объектов m с этим признаком и поделив на объем выборки n, то есть $\hat{p} = m/n$. Долю объектов, не обладающих этим признаком, обозначают $\hat{q} = 1 - \hat{p}$.

Асимптотический доверительный интервал для доли имеет вид:

$$(\hat{p} - \sqrt{\frac{\hat{p}\hat{q}}{n}} z_\alpha \; ; \; \hat{p} + \sqrt{\frac{\hat{p}\hat{q}}{n}} z_\alpha)$$

Важно, что для использования этой формулы требуется выполнение условий $n\hat{p} \geq 5$ и $n\hat{q} \geq 5$.

Если мы хотим узнать минимально необходимый объем выборки для того, чтобы с заданными точностью и уровнем доверия оценить долю признака в генеральной совокупности, то сделать это можно по формуле:

$$n = \hat{p}\hat{q}\left(\frac{z_\alpha}{\Delta}\right)^2$$

Если выборочная доля $\hat{p}$ неизвестна, то в таких случаях её кладут равной 0.5, потому что при этом данное выражение принимает наибольшее значение.

Задача 1. С целью размещения рекламы опрошено 420 телезрителей, из которых данную передачу смотрят 170 человек. С доверительной вероятностью γ=0,91 найти долю телезрителей, охваченных рекламой в лучшем случае.

Решение.

```
import numpy as np
```

```
import scipy.stats as stats
n = 420
m = 170
alpha = 0.09
p = m/n
q = 1-p
z = abs(stats.norm.ppf(alpha/2))
print(f'Доверительный интервал: ({round(p-(p*q/n)**0.5*z,2)} ; {round(p+(p*q/n)**0.5*z,2)})')
print(f'Лучший вариант охвата: {round(p+(p*q/n)**0.5*z,2)}')
```

Задача 2. Чтобы партия прошла в парламент, ей требуется набрать не менее 7% голосов избирателей. Служба мониторинга опросила 1000 избирателей, среди которых оказалось 68 избирателей, собирающихся голосовать за партию А. Постройте 95% доверительный интервал для доли избирателей в процентах, собирающихся голосовать за партию А.

Решение.

```
import numpy as np
import scipy.stats as stats
n = 1000
m = 68
alpha = 0.05
p = m/n
q = 1-p
z = abs(stats.norm.ppf(alpha/2))
print(f'Доверительный интервал: ({round((p-(p*q/n)**0.5*z)*100,2)} %; {round((p+(p*q/n)**0.5*z)*100,2)} %)')
```

Дополнительные статьи по теме: https://habr.com/ru/articles/857470/

6. *Доверительный интервал для дисперсии*

Доверительный интервал для дисперсии генеральной совокупности по выборочной дисперсии s^2 от выборки размером n имеет вид:

$$\left(\frac{(n-1)s^2}{\chi_r^2(\alpha)} ; \frac{(n-1)s^2}{\chi_l^2(\alpha)}\right)$$

Здесь значения $\chi_r^2(\alpha)$ и $\chi_l^2(\alpha)$ находятся по таблицам χ^2-распределения с n − 1 степенью свободы, причем в таблице мы ищем α/2 и 1 – α/2 (они не равны).

Задача 1. По данным выборки объема n = 12 было найдено, что $\sum x_i = 216$, $\sum x_i^2 = 4046$. Постройте 90% доверительный интервал для теоретической дисперсии.

Решение.

Выборочное среднее равно: $\bar{x} = \frac{1}{n}\sum x_i = 18$.

Выборочная дисперсия равна: $s^2 = \frac{1}{n-1}\left(\sum x_i^2 - n\bar{x}^2\right) = \frac{1}{11}(4046 - 12 \cdot 18^2) = 14.36$.

По таблице χ^2-распределения находим α/2 = 0.05, 1 − α/2 = 0.95, число степеней свободы n − 1 = 12 − 1 = 11 и определяем критические точки $\chi^2_l = 4.57$, $\chi^2_r = 19.675$.

Искомый 90%-доверительный интервал имеет вид: (11*14.36/19.675 ; 11*14.36/4.57) = (8; 34.56).

```
import numpy as np
import scipy.stats as stats
n = 12
sum_x = 216
sum_x_2 = 4046
alpha = 0.1
x_mean = sum_x/n
```

```
D = 1/(n-1)*(sum_x_2 - n*x_mean**2)
chi2_r = stats.chi2.ppf(1-alpha/2, df=n-1)
chi2_l = stats.chi2.ppf(alpha/2, df=n-1)
print(f'Доверительный интервал: ({round((n-1)*D/chi2_r,2)}; {round((n-1)*D/chi2_l,2)})')
```

Задача 2. По выборке объёма n = 17 вычислено выборочное среднее $\bar{x} = 20.5$ мм и выборочная дисперсия $s^2 = 16$ мм^2 диаметра валиков. Постройте доверительные интервалы уровня надёжности $\gamma = 0.9$ для дисперсии диаметра валика. Предполагается, что диаметры валиков имеют нормальное распределение.

Решение.