

Доверительный интервал

Доверительный интервал - это интервал значений, который охватывает неопределенность или вариабельность оценки параметра или статистики на основе имеющихся данных. Он предоставляет диапазон возможных значений, в котором находится истинное значение параметра с заданной вероятностью.

Доверительный интервал обычно выражается двумя значениями: нижней и верхней границей интервала. Например, доверительный интервал 95% для среднего значения означает, что существует 95% вероятность того, что истинное среднее значение находится в указанном интервале. Более узкий интервал указывает на более точную оценку, в то время как более широкий интервал указывает на большую неопределенность или вариабельность оценки.

Доверительный интервал зависит от уровня доверия, который определяет вероятность того, что истинное значение параметра находится в интервале. Наиболее распространенный уровень доверия - 95%, что означает, что в 95% случаев истинное значение будет попадать в доверительный интервал. Однако уровень доверия может быть выбран в зависимости от требуемой точности и надежности оценки.

Доверительные интервалы широко используются в статистике и исследованиях, чтобы оценить неопределенность оценок и сделать выводы на основе данных. Они позволяют учесть случайные флуктуации данных и предоставляют информацию о точности и надежности статистических оценок.

Доверительный интервал для математического ожидания

Доверительный интервал для математического ожидания (среднего значения) можно выразить с использованием следующей формулы:

Для случая, когда известна генеральная совокупность и известна её стандартная ошибка σ и размер выборки n , доверительный интервал для математического ожидания μ с заданным уровнем доверия $1 - \alpha$ вычисляется следующим образом:

$$\left(\bar{x} - z \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + z \cdot \frac{\sigma}{\sqrt{n}} \right)$$

где:

- \bar{x} - среднее значение выборки,
- z - критическое значение стандартного нормального распределения, связанное с уровнем доверия $(1 - \alpha)$,

- σ - стандартное отклонение генеральной совокупности,
- n - размер выборки.

```
def confidence_interval_for_expectation_t(data, confidence_prob):
    mean = np.mean(data)
    std = np.std(data, ddof=1)
    n = len(data)
    # t.ppf считает левостороннюю область, а нам требуется двусторонняя
    alpha = (1 - confidence_prob) / 2
    gamma = 1 - alpha
    t_value = st.t.ppf(gamma, n - 1)
    delta = t_value * std / np.sqrt(n)
    return (mean - delta, mean + delta)

confidence_interval_for_expectation_t(data, confidence_prob)
```

```
(29.554737570230277, 36.48970687421421)
```

```
interval_expectation = st.t.interval(confidence_prob, df=data_len - 1,
    loc=mean, scale=st.sem(data))
interval_expectation
```

```
(29.554737570230277, 36.48970687421421)
```

Для случая, когда известна только выборочная совокупность и неизвестна генеральная совокупность, доверительный интервал для математического ожидания μ с заданным уровнем доверия $(1 - \alpha)$ вычисляется следующим образом:

$$\left(\bar{x} - t(\gamma, n - 1) \cdot \frac{s}{\sqrt{n}}, \quad \bar{x} + t(\gamma, n - 1) \cdot \frac{s}{\sqrt{n}} \right)$$

где:

- \bar{x} - среднее значение выборки,
- t - критическое значение распределения Стьюдента, связанное с уровнем доверия $(1 - \alpha)$ и степенями свободы $(n - 1)$,
- s - стандартное отклонение генеральной совокупности,
- n - размер выборки.

Эти формулы позволяют определить диапазон значений, в котором с определенной вероятностью находятся истинные значения математического ожидания для выборочной и генеральной совокупности. Чем больше размер выборки и меньше стандартное отклонение, тем уже будет доверительный интервал и тем точнее будет оценка математического ожидания.

Доверительный интервал для стандартного отклонения

Доверительный интервал для среднего квадратического отклонения (стандартного отклонения) представляет диапазон значений, в котором с определенной вероятностью находится истинное значение стандартного отклонения генеральной совокупности. Формула для расчета доверительного интервала может быть различной в зависимости от метода и предположений о распределении данных.

Одним из наиболее распространенных методов для расчета доверительного интервала для среднего квадратического отклонения является использование распределения хи-квадрат (chi-square distribution).

Формула для доверительного интервала для среднего квадратического отклонения на основе распределения хи-квадрат с уровнем доверия $(1-\alpha)$ и $(n-1)$ степенями свободы выглядит следующим образом:

$$\left((n-1) \cdot \frac{S^2}{\chi_{left}^2}, (n-1) \cdot \frac{S^2}{\chi_{right}^2} \right) \text{ или } \left(\frac{S\sqrt{n-1}}{\chi_{left}}, \frac{S\sqrt{n-1}}{\chi_{right}} \right)$$

где:

- n - размер выборки,
- S^2 - выборочная дисперсия,
- S - выборочное стандартное отклонение,
- χ_{left}^2 , χ_{left}^2 и χ_{right}^2 , χ_{right}^2 - значения из распределения хи-квадрат, связанные с уровнем доверия $(1 - \frac{\alpha}{2})$ и $(\frac{\alpha}{2})$ соответственно.

```
def confidence_interval_std(data, confidence_prob):  
    std = np.std(data, ddof=1)  
    n = len(data)  
    chi_left = np.sqrt(st.chi2.ppf((1 + confidence_prob) / 2, n - 1))  
    chi_right = np.sqrt(st.chi2.ppf((1 - confidence_prob) / 2, n - 1))  
    return (std * np.sqrt(n - 1) / chi_left, std * np.sqrt(n - 1) / chi_right)  
  
confidence_interval_std(data, confidence_prob)
```