Проверка статистических гипотез

Проверка статистических гипотез является важным инструментом в анализе данных и принятии решений на основе эмпирических наблюдений. Она позволяет нам делать выводы о параметрах или законах распределения генеральной совокупности на основе выборочных данных.

Нулевая и альтернативная гипотезы

Статистическая гипотеза представляет собой утверждение о параметрах или законе распределения генеральной совокупности, которое мы хотим проверить с использованием доступных данных. В процессе проверки гипотезы формулируются нулевая и альтернативная гипотезы. Нулевая гипотеза H_0 предполагает, что никаких значимых различий или эффектов нет, тогда как альтернативная гипотеза H_1 предполагает наличие какого-то эффекта или различия.

Проверка гипотезы основывается на сборе выборочных данных и применении статистического теста. Статистический тест использует выборку для вычисления статистической меры (например, t-статистики, z-статистики, или других мер), которая позволяет сделать вывод о согласии или несогласии выборки с нулевой гипотезой. Результаты теста представляются в виде p-значения, которое указывает на вероятность получить такие или более экстремальные данные, если нулевая гипотеза верна.

В зависимости от полученного p-значения и уровня значимости (обычно обозначается α), мы принимаем решение о отвержении или принятии нулевой гипотезы. Если p-значение меньше или равно α, то мы отвергаем нулевую гипотезу в пользу альтернативной гипотезы, что указывает на наличие статистически значимого эффекта или различия. В противном случае, если p-значение больше α, мы не отвергаем нулевую гипотезу, и делаем вывод о недостаточной статистической значимости.

Ошибки первого и второго рода

Ошибки первого и второго рода являются двумя видами ошибок, которые могут возникнуть при статистической проверке гипотез. Они связаны с принятием или отвержением нулевой гипотезы на основе выборочных данных.

Ошибки первого рода, также известные как ложноположительные ошибки, возникают, когда мы отвергаем нулевую гипотезу, хотя она на самом деле верна. То есть мы

делаем вывод о наличии эффекта или различия, когда его на самом деле нет. Вероятность ошибки первого рода обозначается символом α (уровень значимости). Чем ниже уровень значимости, тем меньше вероятность ошибки первого рода. Ошибки первого рода могут иметь серьезные последствия, особенно если они приводят к неправильным решениям или выводам.

Ошибки второго рода, также известные как ложноотрицательные ошибки, возникают, когда мы принимаем нулевую гипотезу, хотя она на самом деле неверна. То есть мы не обнаруживаем эффект или различие, когда они действительно существуют. Вероятность ошибки второго рода обозначается символом β. Чем ниже вероятность ошибки второго рода, тем выше мощность статистического теста, то есть способность обнаружить реальные различия или эффекты.

Ошибки первого и второго рода тесно связаны и обычно сопряжены между собой. При увеличении мощности статистического теста (уменьшении вероятности ошибки второго рода), вероятность ошибки первого рода может увеличиться. Поэтому в статистических исследованиях важно находить баланс между ошибками первого и второго рода, а также принимать во внимание практическую значимость результатов при принятии решений.

Проверка гипотезы о том, что результаты получены из распределения Пуассона с помощью критерия Пирсона

Сформулируем 2 гипотезы H_0 и H_1

- H_0: Случайная величина имеет распределения Пуассона.
- H_1: Случайная величина не имеет распределения Пуассона.

Вычислим ожидаемые частоты в каждом интервале, если бы выборка имела распределение Пуассона.

Сначала найдем λ

$$\lambda \approx \bar{x}$$

$$\overline{x} = \frac{1}{n} \sum x_i n_i$$

```
lmbda = (data["X"] * data["n"]).sum() / n
lmbda
```

```
1.0099667774086378
```

Формула для расчета вероятности появления значения x в распределении Пуассона:

$$P(i) = \frac{\lambda^i \cdot e^{-\lambda}}{i!}$$

где:

- e - базисное число экспоненты (примерно равно 2.71828),
- λ - параметр распределения Пуассона,
- i - значение переменной.

Для вычисления ожидаемых частот в каждом интервале при предположении распределения Пуассона, используется следующая формула:

$$n' = nP_i$$

где:

- n' - ожидаемая частота,
- n - общее количество наблюдений или сумма наблюдаемых частот,
- P_i - вероятность появления значения x в распределении Пуассона.

```
def expected_freq(x):
p = st.poisson.pmf(x, mu=lmbda)
return n * p

data["n'"] = data["X"].apply(expected_freq)
data["n'"]
```

```
0    109.633555
1    110.726248
2     55.914916
3     18.824069
4      4.752921
5      0.960058
Name: n', dtype: float64
```

Теперь вычислим значение статистики критерия Пирсона

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - n_i^*)^2}{n_i^*} = \frac{(\text{наблюдаемая частота} - \text{ожидаемая частота})^2}{\text{ожидаемая частота}}$$

```
chi2_value = ((data["n"] - data["n'"])**2 / data["n'"]).sum()
chi2_value
```

```
1.2568488786645524
```

Определение критического значения

```
chi2_critical = st.chi2.ppf(1-alpha, df=df)
chi2_critical
```

```
7.814727903251179
```

Проверка гипотезы о том, что закон распределения генеральной совокупности является нормальным при уровне значимости α

Сформулируем 2 гипотезы H_0 и H_1

- H_0: закон распределения генеральной совокупности является нормальным.
- H_1: закон распределения генеральной совокупности не является нормальным.

Для проверки гипотезы о нормальности распределения воспользуемся критерием согласия хи-квадрат. Нужно вычислить теоретические (ожидаемые) частоты попадания значений в каждый интервал, если бы распределение было нормальным. Для этого воспользуемся формулой:

$$P_i = P(x_i < X < x_{i+1}) = \Phi(\frac{x_{i+1} - \bar{x}}{S}) - \Phi(\frac{x_i - \bar{x}}{S})$$

$$f' = fP_i$$

```
p = st.norm.cdf((data["x_(i+1)"] - mean) / std) - st.norm.cdf((data["x_i"]
- mean) / std)

assert np.isclose(p.sum(), 1, rtol=.01), "Сумма теоретических оснований
должна быть равна 1."

data["f'"] = p * data["f"].sum()
data["f'"]
```

```
0      1.989322
1      9.051664
2     28.600693
3     62.783364
4     95.779742
5    101.564606
6     74.862071
7     38.350836
```

```
8     13.650848
9      3.374717
Name: f', dtype: float64
```

Вычислим значение статистики критерия χ^2:

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - f_i^*)^2}{f_i^*} = \frac{(\text{наблюдаемая частота} - \text{ожидаемая частота})^2}{\text{ожидаемая частота}}$$

```
chi2_value = ((data["f"] - data["f'"])**2 / data["f'"]).sum()
chi2_value
```

```
1.2613069812716673
```

Определение критического значения

```
chi2_critical = st.chi2.ppf(1-alpha, df=df)
chi2_critical
```

```
11.070497693516351
```

Проверка гипотезы о том, что закон распределения генеральной совокупности является показательным при уровне значимости α

Для проверки гипотезы о показательном распределения воспользуемся критерием согласия хи-квадрат. Нужно вычислить теоретические (ожидаемые) частоты попадания значений в каждый интервал, если бы распределение было равномерным. Для этого воспользуемся формулой:

$$P_i = P(x_i < X < x_{i+1}) = e^{-\lambda x_i} - e^{-\lambda x_{i+1}}$$

$$f' = fP_i$$

$$\lambda = 1/\bar{x}$$

```
lmbda = 1 / mean
lmbda
```

```
0.3826530612244898
```

```
p = st.expon.cdf(data["x_(i+1)"], scale=1/lmbda) -
st.expon.cdf(data["x_i"], scale=1/lmbda)

assert np.isclose(p.sum(), 1, rtol=.01), "Сумма теоретических оснований
должна быть равна 1."

data["f'"] = p * data["f"].sum()
data["f'"]
```

```
0     152.848836
1      92.944306
2      56.517565
3      34.367196
4      20.898001
5      12.707655
6       7.727269
7       4.698797
8       2.857244
9       1.737433
10      1.056498
Name: f', dtype: float64
```

Вычислим значение статистики критерия χ^2:

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - f_i^*)^2}{f_i^*} = \frac{(\text{наблюдаемая частота} - \text{ожидаемая частота})^2}{\text{ожидаемая частота}}$$

```
chi2_value = ((data["f"] - data["f'"])**2 / data["f'"]).sum()
chi2_value
```

```
12.292391902659396
```

Определим критическое значение

```
chi2_critical = st.chi2.ppf(1-alpha, df=df)
chi2_critical
```

```
11.070497693516351
```
