Лекция 1. Доверительные интервалы и проверка гипотез для среднего и пропорций

1. Доверительный интервал если дисперсия известна
2. Определение размера выборки
3 Доверительные интервалы если σ неизвестна, но выборка нормальная (использование *t*-распределения).
4 Проверка гипотез для среднего
5 Доверительные интервалы для проверки гипотез для пропорцийervals and hypothesis tests for proportions.

Статистические выводы относительно данных выборки производятся в вероятностных терминах с оценкой доверия к данным выводам

1. Когда дисперсия, σ, известна

Пример 1: Производитель объявляет, что номинальный вес пакета равен 200 г. Как это можно проверить? Если предположить, что дисперсия известна и равна $\sigma = 5\text{gms}$ и распределение веса имеет нормальное распределение.

Берем выборку из 10 пакетов: выборочное среднее $\bar{x} = 203\,\text{gms}$.

Выборочное среднее $\bar{x}$ это естественная и несмещенная оценка среднего μ. Величина $\bar{x}$=203gms выглядит как разумная оценка μ. Но как надежна эта оценка? Вторая выборка естественно не даст 203г снова. В соответствии с нашими предположениями о нормальности $\bar{X} \sim N\left(\mu, \sigma/\sqrt{n}\right)$, для 95 % доверительного интервала имеем

$$P\left(\mu - \frac{1.96\sigma}{\sqrt{n}} \le \bar{X} \le \mu + \frac{1.96\sigma}{\sqrt{n}}\right) = 0.95.$$

Можно переписать это неравенство в виде

$$P\left(\bar{X} - \frac{1.96\sigma}{\sqrt{n}} \le \mu \le \bar{X} + \frac{1.96\sigma}{\sqrt{n}}\right) = 0.95.$$

В данном примере, $\bar{x} = 203, n = 10$. Две границы равны 203±(1.96 x 5)/√10).

Таким образом с 95% вероятностью истинное среднее лежит в диапазоне (199.9, 206.1).

CI confidence interval -доверительный интервал.

95% CI for

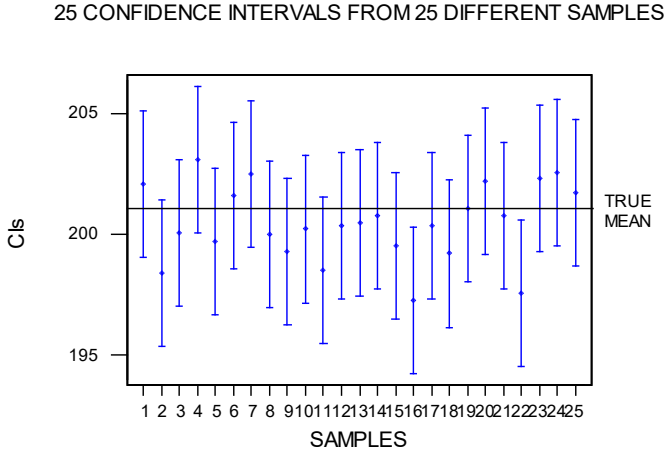
$$\mu \quad is \quad (\bar{x} - \frac{1.96\sigma}{\sqrt{n}}, \quad \bar{x} + \frac{1.96\sigma}{\sqrt{n}})$$

Продолжение Примера 1 :

В течение дня продукция выглядит например так:

Переменные	N	Среднее	Медиана	Стандартное отклонени
Пакетов сыра	1000	200.97	200.89	5.02

В течение дня 25 различных выборок по 10 было выбрано. Каждая имеет свое выборочное среднее, соответствующие доверительные интервалы выглядят так:



Центр каждого интервала $\bar{x}$ и меняется от выборки к выборке.

В предыдущем примере в выборке по 1000 образцов среднее равно μ=200.97. 23/25 из доверительных интервалов содержат истинное среднее. То есть из 25 испытаний 23 дадут доверительный интервал, содержащий истинное среднее, но 2 нет.

В среднем, 95% для 95% доверительных интервалов содержат μ.

90% CI for

$$\mu \quad is \quad (\bar{x} - \frac{1.645\sigma}{\sqrt{n}}, \quad \bar{x} + \frac{1.645\sigma}{\sqrt{n}})$$

Если мы хотим доверия 99%, то необходимо взять 0.5%=0.005 от каждого конца диапазона, это дает величину
$z_{0.005}$=2.5758.

99% CI
for

$$\mu \quad is \quad (\bar{x} - \frac{2.5758\sigma}{\sqrt{n}}, \quad \bar{x} + \frac{2.5758\sigma}{\sqrt{n}})$$

Это наиболее часто используемые доверительные интервалы, но мы можем построить интервалы для любого уровня доверия:

2. **Определение размера выборки**

Какой размер выборки необходим для получения заданного уровня доверия. Необходимо знать три вещи

a) Какой точности необходим доверительный интервал?
b) Каков уровень доверия?
c) Стандартное отклонение для выборки?

Границы доверительного интервала

$$\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \text{ для среднего}$$

равны $z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = d$, или для $\sqrt{n} = z_{\alpha/2}\cdot\frac{\sigma}{d}$

что дает необходимы размер выборки**:**

$$n = \left[\frac{z_{\alpha/2}\,\sigma}{d}\right]^2$$

В примере 2:

$z_{\alpha/2} = 1.645$ для 90% доверия

$\sigma = 4$ задано, и уровень ошибки d = 0.8 что дает:

$$n = \left(\frac{1.645 \times 4}{0.8}\right)^2 = 67.65.$$ то есть *68образцовв.*

3. Доверительный интервал для среднего если - σ неизвестна

Когда стандартное отклонение неизвестно используются распределения с длинными хвостами, ***t-распределения***. Они дают более широкие доверительные интервалы и с ними более проблематична проверка гипотез.

- При использовании t-распределения мы предполагаем что исходная популяция нормальна. (Однако t-распределение достаточно робастно по отношению к ненормальности.)
- Поскольку σ неизвестна, мы должны оценить ее через выборочную дисперсию, ***s.***
- Число степей свободы *t* – распределения берется равным (n-1).

Если размер выборки равен *n* из нормального распределения дает $\bar{x}, s$,
то 95% доверительный интервал для среднего μ равен

$$(\bar{x} - t_{0.025,n-1}\cdot\frac{s}{\sqrt{n}} \quad , \quad \bar{x} + t_{0.025;n-1}\cdot\frac{s}{\sqrt{n}})$$

Пример 3:
7 образцов дает

$\bar{x}$=186.43, s=3.99. Квантиль $t_{0.025;6} = 2.4469$

То есть доверительный интервал для среднего равен

186.43-2.4469 x 3.99/√7 , 186.43+2.4469 x 3.99/√7 = **(182.73,190.12).**

Заметим, что *t*-интервал шире, поскольку мы имеем меньше информации.

4. Проверка гипотез

Использует данные из выборки для принятия решения: допускаем ли мы данное утверждение относительно выборки или отвергаем его.

Логика проверки гипотез аналогична британской судебной практике: лицо считается невиновным пока не доказано обратное.
С точки зрения теории вероятностей, если нулевая гипотеза относительно популяции справедлива, то какова вероятность наблюдения данной выборки?

Пример 1:
Поток из 200 студентов получил контрольное задание. По ошибке 10 из них получили другой вариант, но варианты были случайным образом перемешаны Правильный вариант средняя оценка μ=65 стандартное отклонение $\sigma = 5$ (190 студентов)

Ошибочный вариант $\bar{x}_{10}$=60

Студенты жалуются, что неверный вариант был труднее.

Вопрос: Предположим, что неверный вариант был труднее. Какова вероятность что 10 из 190 студентов получат среднюю оценку 60 или ниже.

Нулевая гпотеза (H_0):

H_0: задания были одинаково трудными

Альтернатива (H_a):

H_a : неверное задание более трудное

Принимая H_0, $P(\bar{X} \le 60) = P(Z \le -3.16) = 0.0008.$

Есть две возможности:

1. Задания одинаковой трудности, но произошло редкое событие.→ Не отвергаем H_0.

ИЛИ

2. Неверный вариант более трудный→ Отвергаем H_0.

Однако в данном примере заключение таково

Отвергаем H_0 в пользу H_a с вероятностью P-value = 0.0008.

***P*-value** обычно если **p-value is < 0.05 мы отвергаем нулевую гипотезу.**

5. Доверительный интервал и проверка гипотез для пропорций

Пример 1 из британской парламентской практики: Заместитель лидера партии утверждает, что из-за серии скандалов, в которые был вовлечен лидер партии, поддержка партии значительно упала со времени последних выборов, когда партия получила 55% голосов. По последнему опросу только 33 из случайной выборки в 80 человек сказали, что готовы поддержать эту партию на выборах. На что Лидер заявил, заявление заместителя не верно, так как это всего лишь случайная выборка.
Пусть p – это уровеь поддержки партии и две гипотезы

$$H_0: p=0.55, \quad H_a: p<0.55.$$

p это неизвестный параметр относительно которого мы хотим сделать заключение. Наиболее естественная оценка p это $\hat{p}=33/80,$ то есть наблюдаемая пропорция из случайной выборки.
Если n велико, например, $(\geq 80, \text{say})$, то

$$95\% \text{ C.I. для } p \text{ равен } \hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$\text{Консервативный, приближенный 95\% C.I. для } p \quad \hat{p} \pm \frac{1}{\sqrt{n}}.$$

Продолжение примера:
Найдем 90% CI для данного $p = \hat{p}=33/80=0.41$ и $\alpha=0.1$ то есть so $z_{0.05}=1.645$.
Для большой выборки имеем

$$0.41 \pm 1.645\sqrt{\frac{0.41\times 0.59}{80}} = (0.32,\ \ 0.50).$$

Малая выборка

$$0.41 \pm \frac{1.645}{2\sqrt{80}} = (0.32,\ \ 0.50).$$

Отметим, что $p(1-p)$ не слишком отличается от 1/4 для p между 0.4 и 0.6.

Если $\hat{p}$ наблюдаемая пропорция, то Z – соответствующая тестовая статистика:

Для проверки $H_0 : p = p_0$, используем $Z = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}} \sim N(0,1).$

Если Z велико (больше 2) нулевая гипотеза отвергается.

Шаг 1:Гипотезы $H_0: p = 0.55, \quad H_a: p < 0.55.$

Шаг 2: Тестовая статистика $$z = \frac{0.41 - 0.55}{\sqrt{\frac{0.55 \times 0.45}{80}}} = -2.52$$

Шаг 3 & 4: находим P-value & заключение **Отвергаем H_0 с вероятностью P-value 0.0059.**