Теория вероятностей и математическая статистика

Доверительные интервалы.

Глеб Карпов

ФКН ВШЭ

Доверительные интервалы

- Использование только точечной оценки для оценки параметра — это как ловить рыбу в мутном озере гарпуном, а использование доверительного интервала — как ловить сетью. Мы можем бросить гарпун туда, где увидели рыбу, но скорее всего промахнемся. Если мы закинем сеть в эту область, у нас будет больше шансов, что рыбалка будет успешна.
- Если мы делаем точечную оценку, мы, вероятно, не попадем точно в неизвестный параметр. Если мы используем диапазон правдоподобных значений — доверительный интервал — у нас есть хороший шанс "поймать" параметр.
- Действительно, если наша точечная оценка $\hat{\theta}$ имеет непрерывное распределение, то $P_\theta\{\hat{\theta} = \theta\} = 0$.

$\overline{X}$; $E[\overline{X}] = E[X] = \mu$

$P(\overline{X} = \mu) = 0$

Доверительные интервалы

Definition

Доверительный интервал. Пусть $X_1, X_2, \dots, X_n$ — случайная выборка случайной величины X. Пусть задано $0 < \alpha < 1$. Пусть $L = L(X_1, X_2, \dots, X_n)$ и $U = U(X_1, X_2, \dots, X_n)$ — две статистики. Мы говорим, что интервал (L, U) является $(1-\alpha)\cdot 100\%$ доверительным интервалом для неизвестного параметра θ, если

$$1 - \alpha = P_\theta\{\theta \in (L, U)\}.$$

Вероятность того, что интервал включает θ, равна $1-\alpha$, которая называется **уровнем доверия** интервала.

Доверительные интервалы

Иллюстративный пример

L U

Для выборки $X_1, \dots, X_4$ из $\mathcal{N}(\mu, 1)$ интервальная оценка μ — это, например, $[\bar{X} - 1, \bar{X} + 1]$. Найдите вероятность того, что истинный параметр μ покрывается этим интервалом.

Решение

$E[\bar{X}] = E[X_i] = \mu$

$P\left(\bar{X} - 1 < \mu < \bar{X} + 1\right) = P\left(-1 < \mu - \bar{X} < 1\right) = P\left(-1 < \bar{X} - \mu < 1\right)$

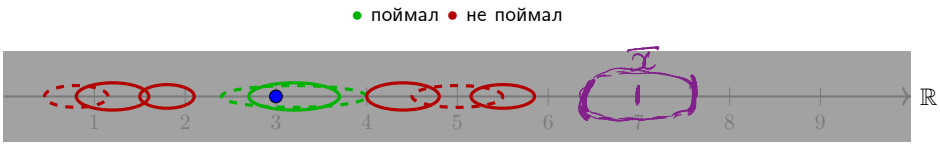
Знаем, что $\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \equiv \mathcal{N}\left(\mu, \frac{1}{4}\right)$, приводим к стандартному нормальному распределению:

$$P\left(-1 < \bar{X} - \mu < 1\right) = P\left(\frac{-1}{\frac{\sigma}{\sqrt{n}}} < \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < \frac{1}{\frac{\sigma}{\sqrt{n}}}\right) =$$

$$P\left(\frac{-1}{\frac{1}{2}} < Z < \frac{1}{\frac{1}{2}}\right) = P\left(-2 < Z < 2\right) \approx 0.9545$$

Действие доверительных интервалов

- Некоторые доверительные интервалы включают в себя θ, некоторые нет. Доверительный интервал поймает параметр с вероятностью $1 - \alpha$.



Доверительные интервалы для среднего генеральной совокупности

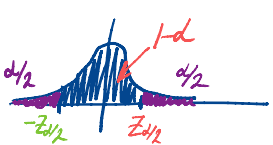
Если дисперсия генеральной совокупности известна

- Нам нужно: случайная выборка размера n, дисперсия $\sigma^2 \equiv Var[X_i]$ известна a priori.
- Утверждение состоит в том, что мы хотим, чтобы наш доверительный интервал (L, U) покрывал неизвестное среднее генеральной совокупности μ с вероятностью $1 - \alpha$:

$$1 - \alpha = P(L(X) < \mu < U(X)) = P(-U < -\mu < -L)$$

- Внедряем в центральную часть неравенства известную случайную величину путём одновременного изменения всех частей неравенства:

$$1 - \alpha = P\left(\frac{\bar{X} - U}{\frac{\sigma}{\sqrt{n}}} < \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < \frac{\bar{X} - L}{\frac{\sigma}{\sqrt{n}}} \right)$$



- Обычно интервалы хотят делать симметричными, поэтому делаем симметричную замену переменных:

$$1 - \alpha = P\left(-z_{\alpha/2} < Z < z_{\alpha/2} \right)$$

Доверительные интервалы для среднего генеральной совокупности

Если дисперсия генеральной совокупности известна

- После нахождения критической точки $z_{\alpha/2}$, такой, что $P(Z > z_{\alpha/2}) = \frac{\alpha}{2}$, выполняем обратную замену и восстанавливаем нужные границы:

$$z_{\alpha/2} = \frac{\bar{X} - L}{\frac{\sigma}{\sqrt{n}}} \to L = \bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$-z_{\alpha/2} = \frac{\bar{X} - U}{\frac{\sigma}{\sqrt{n}}} \to U = \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

- В итоге получаем теоретический $(1-\alpha)100\%$ доверительный интервал для среднего генеральной совокупности μ:

$$\mu \in \left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

Доверительные интервалы для среднего генеральной совокупности

Если дисперсия генеральной совокупности известна

- Однако на практике $(1-\alpha)100\%$ доверительный интервал для среднего генеральной совокупности μ записывается как:

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

- Отличие лишь в одной букве: $\bar{x}$ вместо $\bar{X}$ — конечно, потому что на практике у нас есть именно **реализация** выборочного среднего, те данные, что нам удалось собрать. Вокруг них и строим интервал.

Напоминание: распределения Бернулли и биномиальное

- Случайный эксперимент Бернулли — эксперимент с двумя исходами, часто представляемый как случайная величина со значениями 0 и 1.
- Вероятность успешного исхода $(X = 1)$ обозначается как p, поэтому функция вероятности имеет вид: $P(X = 1) = p$, $P(X = 0) = 1 - p = q$. Математическое ожидание равно $E[X] = p$.
- Если случайная величина Бернулли независимо реализуется в последовательности с фиксированной вероятностью p, это называется **процессом Бернулли**. Пример: подбросить одинаковую монету 7 раз подряд.
- Биномиальное распределение: распределение случайной величины, обозначающей количество успешных результатов в процессе Бернулли. Пример: каждая новая последовательность из 7 подбрасываний монеты, вероятно, будет иметь разное количество орлов.
- Если Y обозначает число успешных результатов, её функция вероятности записывается как:

$$P(Y = k) = C_n^k p^k q^{(n-k)}$$

$$Y = \sum_{i=1}^{n} X_i$$

- Математическое ожидание $E[Y] = np$, а дисперсия $Var[Y] = npq$.

Доля генеральной совокупности: мотивация

- - Нравится ли вам качество бренда?
  - Нравится ли вам этот новый тип кузова автомобиля?
  - Вы курите?
  - Является ли этот продукт бракованным?

  Ответы на все эти и подобные вопросы в основном бинарные или могут быть сделаны бинарными.
- Большая группа людей или объектов (генеральная совокупность) часто обладает некоторым бинарным признаком или атрибутом. Можно предположить, что существует некоторая доля p людей или объектов в генеральной совокупности, обладающих одним и тем же конкретным бинарным признаком.
- Для бизнеса, независимой социологии, медицины и некоторых естественных наук часто важно знать эту долю. Однако, чтобы сделать это честно, нужно исследовать всех людей или объектов, что обычно буквально невозможно.
- Новый вопрос в статистике: как приблизительно оценить эту p только по выборке бинарных ответов?
- **Важная идея**: величину p можно также интерпретировать как вероятность случайно взять элемент, обладающий искомым признаком, из генеральной совокупности

Точечная оценка для истинной доли

- Предположим, у нас есть выборка $\{X_1, ..., X_n\}$ из распределения Бернулли с $P(X_i = 1) = p$. Вся выборка тогда может рассматриваться как процесс Бернулли длины n.
- Введём Y — случайную величину, показывающую количество положительных ответов в выборке, она имеет биномиальное распределение с $E[Y] = np$ и $Var[Y] = np(1-p)$.
- Тогда $\hat{p} = \frac{Y}{n}$ — случайная величина, называемая **выборочной долей** и показывающая долю положительных ответов к размеру выборки.
- Её свойства могут быть выведены из Y, а именно:

$$E[\hat{p}] = \frac{E[Y]}{n} = p, \quad Var[\hat{p}] = \frac{Var[Y]}{n^2} = \frac{p(1-p)}{n}$$

- Как следствие ИТМЛ, если $n > 30$, то:

$$\hat{p} \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right), \quad \text{или} \quad \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim \mathcal{N}(0, 1)$$

Доверительные интервалы для истинной доли признака в генеральной совокупности

- Если выполнены условия ИТМЛ, то действуем знакомым способом:

$$1-\alpha = P(L<p<U) = P(-U<-p<-L) =$$

$$P\left(\frac{\hat{p}-U}{\sqrt{\frac{p(1-p)}{n}}} < \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} < \frac{\hat{p}-L}{\sqrt{\frac{p(1-p)}{n}}}\right) = P\left(-z_{\alpha/2} < Z < z_{\alpha/2}\right)$$

N(0,1)

- Находим из таблицы или иным образом критическую точку $z_{\alpha/2}$, такую, что $P\left(Z > z_{\alpha/2}\right) = \frac{\alpha}{2}$, и выполняем обратное преобразование:

$$L = \hat{p} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}, \quad U = \hat{p} + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

- Однако, эта формула над нами смеется. Мы не знаем истинное значение p, хотим его поймать в доверительный интервал, а оно присутствует в формуле. :clown_face:

Доверительные интервалы для истинной доли признака в генеральной совокупности

Поэтому на практике $(1-\alpha)100\%$ доверительный интервал для истинной доли признака p записывается как:

$$p \in \left(\tilde{p} - z_{\alpha/2}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}}, \tilde{p} + z_{\alpha/2}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}} \right),$$

где $\tilde{p}$ — **реализация** выборочной доли, т.е. реальные, окончательные числа в нашем распоряжении, показывающие долю бинарного признака в имеющейся выборке.

Пример: выборка размера 200 содержит 15 бракованных продуктов, в выборке из 80 студентов 64 выполняют домашние задания самостоятельно, и т.д.

Иллюстративные задачи

Пример 1: Анализ рынка

Маркетинговая команда премиальной кофейной сети хочет понять предпочтения клиентов относительно нового сезонного напитка. Они провели опрос среди 400 случайно выбранных клиентов, и 156 из них выразили заинтересованность в новом напитке.

1. Постройте 90% доверительный интервал для истинной доли всех клиентов, которые были бы заинтересованы в новом сезонном напитке.
2. Маркетинговая команда хочет быть уверена на 95%, что их оценка доли генеральной совокупности находится в пределах ± 0.03 (3 процентных пункта) от истинной доли. Какой размер выборки им потребуется для достижения этого уровня точности?
3. Основываясь на доверительном интервале из пункта 1, порекомендовали бы вы запуск нового напитка, если компания требует, чтобы по крайней мере 35% клиентов были заинтересованы для того, чтобы запуск был прибыльным? Объясните ваши рассуждения.

2. $U = \tilde{p} + z_{\alpha/2}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}}$

0.03

$X \cdot (1-X)$ $(0, 1)$

$X^* = 1/2$

Иллюстративные задачи

$X \sim Ber(p)$ $P(X=1)=p$ $P(X=0)=1-p$ | $1-\alpha = P(L < p < U)$

Решение

1. $\tilde{p} = \frac{156}{400} = 0.39$, $z_{0.05} = 1.645$:

$$p \in \left(0.39 - 1.645\sqrt{\frac{0.39 \cdot 0.61}{400}}, 0.39 + 1.645\sqrt{\frac{0.39 \cdot 0.61}{400}}\right) = (0.350, 0.430)$$

2. $z_{0.025} = 1.96$. Полуширина интервала $= E = 0.03$.
Так как мы не знаем, какая доля получится в новой выборке, используем самую консервативную оценку доли $\tilde{p} = 0.5$, при ней интервал получается наиболее широким при прочих фиксированных переменных:

$$E = z_{\alpha/2}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}} \qquad \tilde{p} = 1/2$$

$$n = \frac{z_{\alpha/2}^2 \tilde{p}(1-\tilde{p})}{E^2} = \frac{1.96^2 \cdot 0.5 \cdot 0.5}{0.03^2} \approx 1068$$

3. Нижняя граница интервала 0.35, совпадает с минимальной границей, нужной для запуска. Даже в худшем случае (левая граница интервала) доля заинтересованных клиентов как раз нужная. Можно рекомендовать продукт к запуску.

Ex: $(0.3, 0.38)$ X