Лекция 18. Оценяване на хипотези

Емпиричното оценяване на точността на хипотези е един фундаментален въпрос в машинно самообучение. В следващите две лекции ще разгледаме базови статистически методи за оценка на точността на хипотези, фокусирайки се на следните три въпроса:

1. При зададена наблюдавана точност на някоя хипотеза върху ограничена извадка от данни, до колко добре тази оценка преценява точността на хипотезата върху други допълнителни данни?
2. Ако една хипотеза превъзхожда друга върху определена извадка от данни, доколко е вероятно, че същата хипотеза ще бъде по-точна от другата върху всякакви данни?
3. Когато данните са ограничени, какъв е най-добрият начин да ги използваме, за да научим една хипотеза и да оценим нейната точност?

Тъй като ограниченият размер на използваните извадки от данни може да изопачи общото разпределение на данни, оценяването на истинската точност на базата на подобни извадки може да бъде грешно. Статистическите методи, заедно с определени предположения за характера на разпределение на данни, позволяват да бъде ограничена разликата между наблюдаваната точност върху извадката от наличните данни и истинската точност върху цялото разпределение на данни.

18.1. Мотивация

Често е много важно да оценим колкото може по-точно поведение на научените хипотези. Една от причините е просто за да разберем, дали да използваме тези хипотези или не. Например, при обучението от една ограничена по размер база от данни, описваща ефективността от прилагане на различни медикаменти, е важно да разберем колкото може по-точно точността на научените хипотези. Като втория аргумент за необходимостта от методи за оценяване на поведението на хипотези можем да изтъкнем, че подобни методи лежат в основата на множество от методите за индуктивно самообучение. Например, при допълнителното подрязване на класификационни дървета, целящо избягване на преспецификацията на данни, трябва да преценяваме влияние на възможни подрязвания върху точността на получените класификационни дървета. По тази причина е много важно да бъдат разбрани възможни грешки, наследени от оценяването на точността на подрязани и неподрязани дървета.

Оценяването на точността на хипотези е сравнително лесно, когато разполагаме с достатъчно много данни. Обаче, когато трябва да научаваме хипотези и да преценяваме тяхната бъдеща точност само на база на ограничени по обем данни, възникват следните затруднения:

1. *Пристрастие при оценяване.* Наблюдаваната точност на научената хипотеза върху обучаващите примери е често само едно много лошо приближение на нейната точност върху бъдещите примери. Тъй като научената хипотеза е била изведена от тези примери, обикновено те предлагат едно оптимистично, пристрастно приближение за точността на хипотезата върху бъдещите примери. Особено това е вероятно, когато алгоритмът за обучение използва много богато пространство от хипотези, позволяващо му да преспецифицира (да покрива прекалено точно) обучаващите примери. За да получим една безпристрастна оценка за бъдещата точност на хипотезата, ние обикновено я тестваме върху определено множество от тестови примери, избрани независимо от обучаващите примери и хипотезата.
2. *Различия при оценяване.* Дори когато точността на хипотеза се оценява върху едно безпристрастно множество от тестови примери, което е независимо от обучаващите примери, измерената стойност на точността може да се отличава от истинската точност в зависимост от конкретното подреждане на тестовите примери. Колко по-малко е множеството от тестовите примери, толкова е по-голямо очакваното отклонение.

18.2. Оценяване на точността на хипотезата

При оценяването на някоя научена хипотеза ние сме най-заинтересовани в преценка на точността, с която хипотезата ще класифицира бъдещите примери. В същото време ние бихме искали да знаем вероятната грешка при тази оценка.

В тази лекция ще разглеждаме следната постановка на задачата за самообучение. Има определено пространство от възможни примери X (например, множество от всички хора), върху които могат да бъдат дефинирани различни целеви функции (например, хора, които планират да купят нови ски тази година). Ще предполагаме, че различните примери в X могат да се наблюдават с различни честоти за срещане. Един удобен начин за моделиране на това предположение е да смятаме, че съществува определено, неизвестно за нас разпределение на вероятност $\mathcal{D}$, което определя вероятността за срещане на всеки пример x в X (например, $\mathcal{D}$ може да назначава по-голяма вероятност да срещнем 19 годишен човек от колко 109 годишен). Обърнете внимание, че $\mathcal{D}$ не казва нищо за това, дали x е положителен или отрицателен пример; то само определя вероятността за срещане на x. Задачата за самообучение е да научим целевото понятие или целевата функция f чрез разглеждане на пространството H от възможни хипотези. Обучаващите примери на целевата функция f са предоставени от учителя, който избира всеки пример независимо един от друг съгласно разпределението $\mathcal{D}$, и след това предоставя на обучаемия всеки пример x заедно с неговото правилно значение на целевата функция $f(x)$.

За илюстрацията да разгледаме целевата функция "хора, планиращи да купят нови ски тази година", като е зададена една извадка от обучаващите данни, събрана чрез анкетиране на хора, които пристигнали в Боровец. В този случай пространството на

примери X е пространството от всички хора, които могат да бъдат описани чрез такива атрибути като възраст, място живеене, място на работата, колко пъти ходили на ски миналата година и т.н. Разпределението $\mathcal{D}$ задава за всеки човек x вероятността, че x ще се срещне измежду хората, пристигнали в Боровец. Целевата функция $f: X \to \{0,1\}$ класифицира всеки човек съгласно това, дали той планира или не да купи новите ски тази година.

При тази обща постановка на задачата нас ще ни интересува следните два въпроса:

1. При зададена хипотеза h и извадката от данни, съдържаща n примера, избрани по случаен начин съгласно разпределението $\mathcal{D}$, каква ще бъде най-добрата оценка за точността на h върху новите примери, избрани от същото разпределение?
2. Каква е възможната грешка при тази оценка на точността?

18.3. Извадкова грешка и истинска грешка

За да отговорим на поставените въпроси трябва да правим разлика между двете понятия за точността или, съответното, за грешка. Първото е относителният брой на грешки, които една хипотеза прави върху наличната извадка от данни. Второто е относителния брой на грешки, които хипотезата прави върху цялото неизвестно разпределение $\mathcal{D}$ на примери. Ще ги наричаме, съответно, *извадкова грешка* и *истинска грешка*.

Извадковата грешка на някоя хипотеза по отношение на определената извадка S от примери, избрани от X, е тази част от S, които хипотезата класифицира неправилно:

> **Определение**: *Извадковата грешка* (означавана чрез $error_S(h)$) на хипотезата h по отношение на целевата функция f и извадката от данни S е
>
> $$error_S(h) \equiv \frac{1}{n} \sum_{x \in S} \delta(f(x), h(x))$$
>
> където n е броя на примери в S, а функцията $\delta(f(x), h(x))$ приема стойност *1*, когато $f(x) \neq h(x)$ и *0* – в противен случай.

Истинската грешка на някоя хипотеза е вероятността, че тя ще класифицира неправилно някой пример, който случайно е избран от разпределението $\mathcal{D}$.

> **Определение**: *Истинската грешка* (означавана чрез $error_{\mathcal{D}}(h)$) на хипотезата h по отношение на целевата функция f и разпределението $\mathcal{D}$ е вероятността, че h ще класифицира неправилно някой пример x, избран случайно съгласно разпределението $\mathcal{D}$.

$$error_D(h) \equiv \Pr_{x \in D}[f(x) \neq h(x)]$$

Където $\Pr_{x \in D}$ означава, че вероятността се изчислява върху разпределението на примери $\mathcal{D}$.

Обикновено ние искаме да знаем истинската грешка на хипотезата, тъй като това е грешката, която можем да очакваме при прилагането на хипотезата към бъдещите примери. Обаче, единственото, което ние можем да измерим, е извадковата грешка на хипотезата върху определена извадка от данни, с които разполагаме. Основният въпрос, който ще разгледаме в този раздел е: *"До колко добре извадковата грешка дава оценка за истинската грешка?*

18.4. Оценяване на грешка и биномното разпределение

Как точно разликата между извадковата грешка и истинската грешка зависи от размера на извадката? Този въпрос е пример на един добре проучен проблем в математическата статистика – проблемът за оценяването на част от генералната съвкупност, която има определено свойство, по зададена наблюдавана част от някоя случайна извадка от същата съвкупност. В нашият случай, интересуващото ни свойство е, че *h* грешно класифицира примера.

Ключът към отговора на този въпрос е наблюдението, че когато измерваме извадковата грешка, ние изпълняваме един експеримент със случаен резултат. Първо, ние конструираме някоя случайна извадка *S* от *n* примери, независимо избрани от разпределението $\mathcal{D}$, а след това измерваме извадковата грешка $error_S(h)$. Ако повтаряме този експеримент много пъти, избирайки всеки път различни случайни извадки S_i с размер *n*, ще очакваме да наблюдаваме различни стойности на $error_{S_i}(h)$ в зависимост от случайните разлики в подреждане на различни S_i. В тези случаи можем да кажем, че резултатът $error_{S_i}(h)$ от всеки i-ти експеримент е една *случайна величина*. В общия случай можем да мислим за една случайна величина като за името на експеримента със случаен резултат. Стойността на тази случайна величина е наблюдаваният резултат от случайния експеримент.

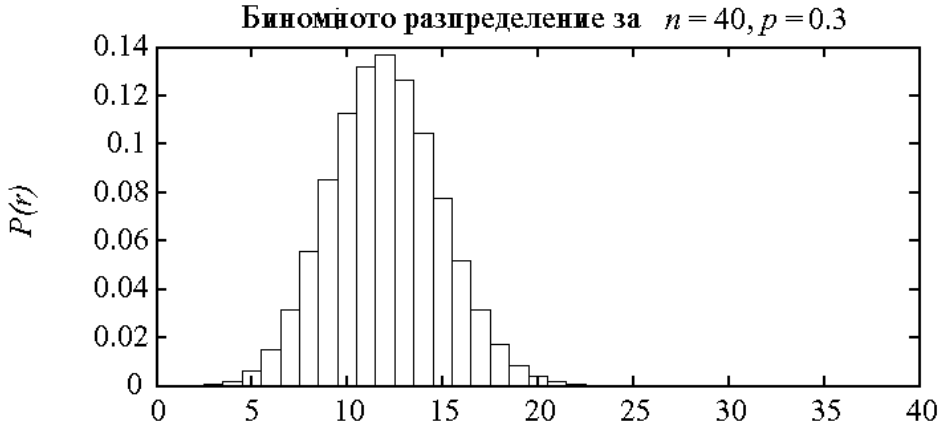
Биномното разпределение е едно от широко използваните вероятностни разпределения на дискретна случайна величина. То описва ситуацията, когато всяко изпитание (експеримент) има два възможни изхода (събития), като техните вероятности при всеки експеримент остават постоянни (*p* – на едното събитие, *q* – на другото и $p + q = 1$), а самите експерименти – независими. Класическият пример за биномния експеримент е подхвърлянето на монета няколко пъти. В този случай са възможни само два събития – монетата да падне по герб или по лице. Нека неизвестната за нас вероятност, че монетата ще падна по лице, е *p*. Да предположим, че ние хвърлихме монетата *n* пъти, от които *r* пъти тя падна по лице. В този случай разумната оценка за *p* е *r/n*. Ако направим нов експеримент от *n* хвърляния, естествено е да се очаква, че броят на падания по лице *r* в този случай

ще бъде различен и, следователно, ще получим другата оценка за p. Биномното разпределение описва за всяка възможна стойност на r (*0, 1, 2*, и т.н.) вероятността *Pr(r)* да наблюдаваме точно r падания на монетата по лице при n независими изпитания (хвърляния на монетата), при условие, че истинската вероятност да падне монетата по лице е p:

$$\Pr(r) = C_n^r p^r (1-p)^{n-r}, \text{ където } C_n^r = \frac{n!}{r!(n-r)!} \quad (18.1)$$

Оценяването на p по една случайна извадка от хвърляния на монета е еквивалентно на оценяването на истинската грешка $error_D(h)$ чрез тестване на h върху една случайна извадка S от примери. Едно хвърляне на монетата отговаря на избор на един случаен пример от $\mathcal{D}$ и определяне, дали той е класифициран грешно от h. Вероятността p, че при едно случайно хвърляне монетата ще падна по лице, съответства на вероятността, че един случайно избран пример ще бъде класифициран грешно (т.е p съответства на $error_D(h)$). Броят r на наблюдаваните падания на монетата по лице при извършване на n експеримента – съответства на броя на неправилните класификации върху n случайно избрани примера. Следователно, r/n съответства на $error_S(h)$. Задачата за оценяването на p за монетите е идентична на задачата за оценяването на $error_D(h)$ за хипотезите.

И така, биномното разпределение характеризира вероятността монетата да падне по лице r пъти при n хвърляния, както и вероятността да наблюдаваме r грешки при класифициране на извадката от данни, съдържаща n случайно избрани примера. Графика (хистограма) на биномното разпределение (при конкретни стойности на n и p) е показана долу.



18.3.1. Средно и дисперсия на биномното разпределение

Математическото очакване (или *средната стойност*) на случайна величина *Y*, приемаща възможните стойности $y_1, ..., y_n$, се определя като $E[Y] = \sum_i y_i \Pr(Y = y_i)$.

Например, ако *Y* приема стойност *1* с вероятност *0.7*, стойност *2* с вероятност *0.2* и стойност *3* с вероятност *0.1*, то *E[Y] =1*07 + 2*0.2 + 3*0.1 = 1.3*. С други думи, изчисляване на математическото очакване става по начина, аналогичен на изчисляването на средната аритметична величина, като разликата е само в теглата, които при математичното очакване са вероятностите, чиято сума е *1*.

Ако *Y* се подчинява на биномното разпределение, то може да се покаже, че:

$$E[Y] = np \quad (18.2)$$

където *n* и *p* са параметри на биномното разпределение (18.1).

Дисперсията определя, как се очаква да варира случайната величина относно своята средна стойност: $Var(Y) \equiv E[(Y - E[y])^2]$. Дисперсията описва очакваната квадратична грешка от използването на резултата от едно единствено наблюдение на *Y* за оценяването на *E[Y]*.

Квадратен корен на дисперсията се нарича *стандартното отклонение* на *Y*: $\sigma_Y = \sqrt{E[(Y - E[Y])^2]}$.

Ако *Y* се подчинява на биномното разпределение, то може да се покаже, че:

$$\begin{aligned} Var(Y) &= np(1-p) \\ \sigma_y &= \sqrt{np(1-p)} \end{aligned} \quad (18.3)$$

18.3.2. Оценители, пристрастие и дисперсия

След като показахме, че случайната величина $error_S(h)$ се подчинява на биномното разпределение, да се върнем към нашия основен въпрос: “Каква е вероятната разлика между $error_S(h)$ и истинската грешка $error_D(h)$?”.

Да опишем $error_S(h)$ и $error_D(h)$ в термините на уравнение (18.1), задаващо биномното разпределение:

$$error_S(h) = \frac{r}{n}$$
$$error_D(h) = p$$

където n е броя на примери в извадката S, а r – броя на тези примери от S, които класифицирани неправилно от h; p е вероятността за неправилното класифициране на един пример, избран по случаен начин от $\mathcal{D}$.

В статистиката $error_S(h)$ се нарича *оценителят* на истинската грешка $error_D(h)$. В общия случай, оценителят е една случайна величина, използвана за оценяване на определен параметър на генералната съвкупност, от която е избрана дадената извадка. Очевидният въпрос за всеки оценител е дали в средно той дава правилното приближение (оценка). Ще дефинираме о*ценъчното предпочитание (субективната оценка)* на един оценител като разликата между очакваната стойност на оценителя и истинското значение на оценявания параметър.

Определение: *Субективната оценка* на оценителя Y за произволен параметър p е

$$E[Y] - p.$$

Ако субективната оценка е нула, то ще казваме че Y е *обективен (не субективен)* оценител на p.

Дали $error_S(h)$ е обективния оценител на $error_D(h)$? Да, тъй като при биномното разпределение очакваната стойност за r е равна на np. Тъй като n е константа, от тук следва, че очакваната стойност е r/n и е равна на p.

Трябва да се отбележи, че когато споменахме, че тестването на хипотези върху обучаващите примери дава оптимистичната оценка за грешката на хипотезите, ние сме имали предвид точно понятие за оценъчното предпочитание. За да бъде $error_S(h)$ един обективен оценител на $error_D(h)$, хипотезата h и извадката S трябва да се избират *независимо*.

Освен това, не трябва да се бърка оценъчното пристрастие с понятието *индуктивното пристрастие*, което беше въведено по-рано. Оценъчното пристрастие е една *числена стойност*, докато индуктивното пристрастие е едно *множество от утвърждения*.

Другото важно свойство на един оценител е неговата дисперсия. При възможността за избор сред няколко обективни оценителя, има смисъл да се избере този с най-малката дисперсия. По определението на дисперсия този избор ще даде най-малката очаквана квадратична грешка между оценителя и истинската стойност на оценявания параметър.

Да разгледаме като пример случай, когато тестваме някоя хипотеза и установихме, че тя прави $r = 12$ грешки върху една извадка от $n = 40$ примера. Обективната оценка за $error_D(h)$ се дава от $error_S(h) = r/n = 12/40 = 0.3$. Дисперсията на тази оценка е предизвикана напълно от вариацията на r, тъй като n е константа. Тъй като r е биномно разпределена, нейната дисперсия се изчислява като $np(1-p)$. Обаче, тъй като p е неизвестно, можем да го заменим с нашата оценка за p като r/n. По този

начин, оценъчната дисперсия на r е $40*0.3(1 - 0.3) = 8.4$, а стандартното отклонение е $\sqrt{8.4} \approx 2.9$. От тук следва, че стандартното отклонение на случайната величина $error_S(h) = r/n$ приблизително равно на $2.9/40 = 0.07$. С други думи, $error_S(h) = 0.3 \pm 0.07$.

В общия случай, когато хипотезата h прави r грешки в извадката от n независими тестови примери, стандартното отклонение на $error_S(h)$ се изчислява по формулата:

$$\sigma_{error_S(h)} = \frac{\sigma_r}{n} = \sqrt{\frac{p(1-p)}{n}} \quad (18.4)$$

която може да бъде приближена чрез замяната на p с $r/n \equiv error_S(h)$:

$$\sigma_{error_S(h)} \approx \sqrt{\frac{error_S(h)(1-error_S(h))}{n}} \quad (18.5)$$

18.4. Доверителни интервали и нормалното разпределение

Често използваният начин за описание на несигурността, асоциирана с оценка на някой параметър, е да се укаже интервал, в който се очаква да лежи истинската стойност на оценявания параметър, заедно с вероятността, с която се очаква тя да попадне в този интервал. Подобни оценки са наричат *доверителни интервали*.

Определение: *N% доверителен интервал* за параметър p е интервалът, който с вероятност $N\%$ се очаква да съдържа p.

Как можем да изведем някой доверителен интервал за $error_D(h)$? Отговорът се базира върху факта, че оценителят на тази случайна величина $error_S(h)$ се подчинява на биномното разпределение. Средната стойност на това разпределение е равна на $error_D(h)$, а стандартното отклонение се задава от ф-ла (18.5). Следователно, за да получим $N\%$ доверителен интервал трябва само да намерим интервал, центриран около средната стойност $error_D(h)$, който е достатъчно широк да съдържа $N\%$ от общата вероятност за това разпределение. Как да намерим размера на този интервал?

За съжаление за случая на биномното разпределение подобните изчисления са твърде трудоемки. За щастие обаче, в повечето случаи може да се намери едно лесно за изчисляване добро приближение към него, тъй като за достатъчно големи извадки биномното разпределение може да бъде добре апроксимирано от така наречено *нормално разпределение*. Нормалното разпределение е едно от най-важните и най-добре изучените непрекъснати разпределения. Неговото изучаване се датира от 18 век, когато се изследва характера на грешки при измерванията. Тогава е било установено, че несъответствията (различията) между повтарящи се

измервания на едно и също физическо количество имат неочакваната степен на постоянство. Било е установено, че разпределението на несъответствията може да се апроксимира с помощта на определена крива на непрекъснато разпределение, известна като "нормална крива на грешките".

Определение: Разпределението на непрекъсната случайна величина *X*, плътността на вероятност на която се задава чрез формулата:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (18.6)$$

се нарича *нормално (или Гаусово) разпределение*; μ и σ са параметри, напълно определящи нормалното разпределение.

Ако непрекъсната случайна величина *X* следва нормалното разпределение, то вероятността, че *X* попада в интервал *(a, b)* се задава от формула:

$$\Pr[X \in (a,b)] = \int_a^b p(x)dx$$

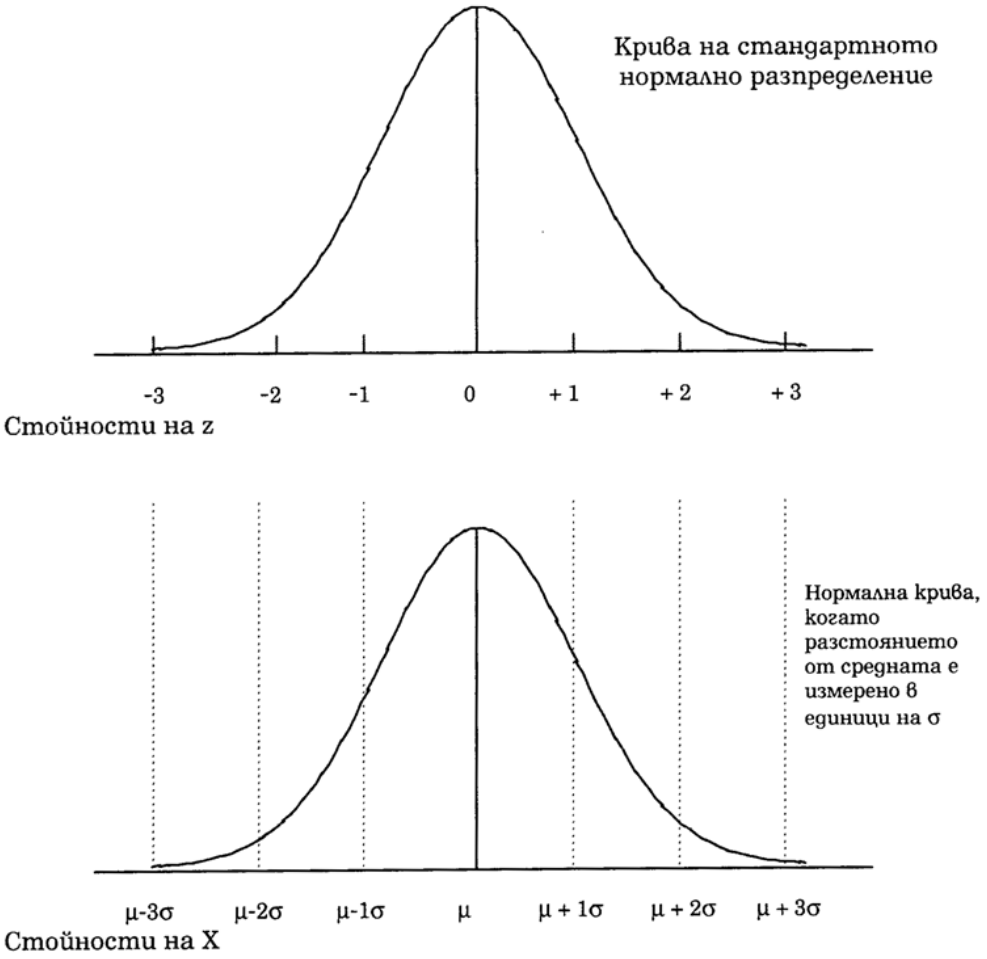
Математическото очакване на нормално разпределена случайна величина *X*:

$$E[X] = \mu$$

а дисперсията и стандартното отклонение:

$$Var(X) = \sigma^2; \quad \sigma_X = \sigma$$

Една от най-често срещаните задачи при нормалното разпределение е търсене на вероятността, че случайната величина ще попадне в определен интервал *(a, b)*. Изчисленията се правят с помощта на интегралната функция на нормалното разпределение: $\int_a^b p(x)dx$, за което обикновено са необходими много изчисления. Проблемът се решава чрез въвеждане на новата променлива $z = \frac{x-\mu}{\sigma}$ и се работи с функцията $P(z) = \frac{2}{\sqrt{2\pi}} \int_0^z e^{-\frac{u^2}{2}} du$. За стойностите на тази функция са разработени таблиците, които съдържат площи (вероятности) под кривата само за положителните стойности на *z*, тъй като нормалната крива е симетрична. Чрез заместването на оригиналните стойности на случайната величина *X* със *z* стойностите, нормалното разпределение се превръща в *стандартното нормално разпределение* със средното равно на нула и стандартното отклонение равно на единица (виж Фиг. 18-1).

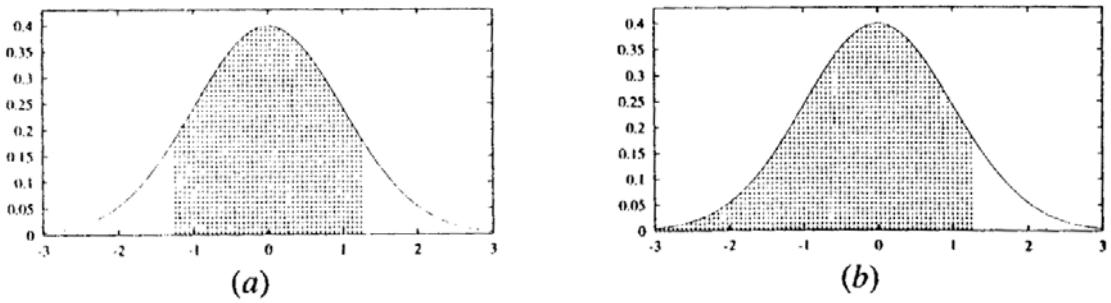


Фиг. 18-1

И така, таблиците определят размера на интервала около средната стойност, който съдържа *N%* вероятност при нормалното разпределение – т.е. точно това, което ни трябва за изчисляване на *N%* доверителен интервал. Като пример, да разгледаме Таблица 18-1:

Доверителен интервал N%	50%	68%	80%	90%	95%	98%	99%
Константа z_N	0.67	1.00	1.28	1.64	1.96	2.33	2.58

Константата z_N определя ширина на най-малкия интервал около средната стойност, който включва *N%* от общата вероятност при нормалното разпределение (по точно половина от интервала), измерена в единиците на стандартното отклонение. Фиг. 18-2(а) илюстрира такъв интервал за $z_{0.80}$.



Фиг. 18-2. Стандартното нормално разпределение. а) С достоверност 80% стойността на случайната величина ще лежи в двустранен интервал [-1.28, 1.28] ($z_{0.80}$ = 1.28). С достоверност 10% тя ще попада вдясно от този интервал, а с достоверност 10% - вляво от него. b) С достоверност 90% случайната величина ще попада в едностранен интервал $[-\infty, 1.28]$.

С други думи, ако една случайна величина Y се подчинява на нормалното разпределение със средното μ и стандартното отклонение σ, то измерената случайна стойност y на Y ще попада в интервал

$$\mu \pm z_N \sigma \quad (18.7)$$

в *N%* от случаи, когато се прави измерването.

Еквивалентно, средната стойност μ ще попада в интервала

$$y \pm z_N \sigma \quad (18.8)$$

в *N%* от случаи, когато се прави измерването.

Комбинирайки тези факти с известните по-рано, можем да изведем общата формула за изчисляване на *N%* доверителен интервал за хипотези с дискретни стойности. Първо, знаем че $error_S(h)$ следва биномното разпределение със средното $error_D(h)$ и стандартното отклонение, зададено от формула (18.5). Второ, знаем, че за достатъчно голям размер на извадката n това биномно разпределение добре се апроксимира с нормалното разпределение. Трето, уравнение (18.8) показва, как да намерим *N%* доверителен интервал за оценяване на средното на нормалното разпределение. Следователно, подставяйки средното и стандартното отклонение на $error_S(h)$ в (18.8), получаваме формулата за *N%* доверителен интервал за хипотези с дискретни стойности:

$$error_S(h) \pm z_N \sqrt{\frac{error_S(h)(1-error_S(h))}{n}} \quad (18.9)$$

За илюстрация да предположим, че извадката S съдържа *40* примера, а хипотезата h прави $r = 12$ грешки върху тези данни. В този случай, извадковата грешка $error_S(h) = 12/40 = 0.30$. *95%* доверителен интервал за оценяване на $error_D(h)$ в нашия конкретен пример е $0.30 \pm (1.96 * 0.07) = 0.30 \pm 0.14$.

По същия начин, както сме изчислили 95% доверителен интервал за $error_D(h)$, ние можем да изчислим и 68% доверителен интервал, който в този случай ще бъде

$0.30 \pm (1.00 * 0.07) = 0.30 \pm 0.07$. Обърнете внимание, че 68% доверителен интервал е по-малък от 95% доверителен интервал – това има интуитивен смисъл, тъй като ние сме намалили вероятността, с която изискваме $error_D(h)$ да попада в съответния интервал.

Уравнението (18.9) показва как да се изчисляват доверителните интервали за оценяване на $error_D(h)$, базирайки се на $error_S(h)$. При използването на тази формула трябва да се отчита, че:

- тя е приложима само за хипотези, приемащи дискретни стойности и че се предполага, че извадката S е взета по случаен начин от същото разпределение, от които ще бъдат взети и бъдещите данни,
- данните са независими от тестваните хипотези.
- Освен това, трябва да се има предвид, че формулата задава само приблизителен доверителен интервал, макар че това приближение е достатъчно добро, когато извадката съдържа най-малко *30* примера, а стойността на $error_S(h)$ не е много близка до 0 или 1. По-точната емпирична формула, описваща кога това приближение работи ефективно е: $n * error_S(h)(1 - error_S(h)) \geq 5$

  За по-малки стойности на n е по-добре да се ползват точните стойности от биномното разпределение.

18.4.1. Двустранни и едностранни доверителни интервали

Обърнете внимание, че доверителният интервал е ограничен от двете страни, т.е. той ограничава оценяваната стойност и отгоре, и отдолу. В някои случаи ние сме заинтересувани в ограничение само от едната страна. Например, може да ни е важно да знаем отговора на въпроса: "Каква е вероятността, че $error_D(h)$ не е по-голяма от U?" Това е един естествен въпрос, тъй като ние се интересуваме от ограничаване на максималната граница на грешката на h и не се интересуваме от случая, когато истинската грешка е значително по-малка от нашата оценка.

Тази едностранна оценка може лесно да бъде получена. Процедурата следва от факта, че нормалното разпределение е симетрично около своята средна стойност. По тази причина всеки двустранен доверителен интервал, базиран на нормалното разпределение, лесно може да се превърне в съответния едностранен доверителен интервал (виж Рис 18-2 (b)). Така $100(1 - \alpha)\%$ доверителен интервал с долната граница L и горната U може да се превърне в $100(1 - \alpha/2)\%$ доверителен интервал с долната граница L и без горна граница или в $100(1 - \alpha/2)\%$ доверителен интервал с горната граница U и без долна граница. α означава вероятността, че правилната стойност лежи извън указания интервал. С други думи α е вероятността, че стойността ще попадне в незащрихованата област от Рис. 18-2 (a), а $\alpha/2$ – вероятността за попадане в незащрихованата област от Рис. 18-2 (b).

За илюстрация да разгледаме пример, когато h прави $r = 12$ грешки върху извадката от $n = 40$ независимо избрани примера. Както вече е показано по-рано,

двустранен 95% доверителен интервал за този случай е 0.30 ± 0.14. Т.е. *100(1 – α) = 95%,* следователно *α = 0.05*. Прилагайки изведеното по-горе правило, можем да кажем, че с достоверност *100(1 – α/2)% = 97.5%* истинската грешка $error_D(h)$ не е по-голяма от 0.30 + 0.14 = 0.44, не правейки никакви заключения за долната ѝ граница. По този начин получаваме едностранното ограничение на грешката $error_D(h)$ *с* двойно по-голяма достоверност, отколко при двустранното ограничение. (Не забравяйте, че *вероятността за грешка* в нашата оценка намаляла двойно – от α – към α/2, т.е. двойно се увеличи достоверността на нашата оценка).

18.5. Общ подход към извеждане на доверителните интервали

В предишния раздел подробно разгледахме как се извеждат доверителните интервали за един конкретен случай – оценяване на истинската грешка $error_D(h)$ за хипотеза *h* с дискретни стойности на базата на извадката от *n* независимо избрани примера. В този раздел ще покажем един общ подход за извеждане на доверителните интервали, който се използва при различни проблеми, свързани с оценяването на определени параметри и, в частност, при задачата за оценяването на средната (очакваната) стойност на една съвкупност на базата на измерената средна стойност на една случайно избрана от нея извадка с размер *n*. Този общ процес се състои от следните стъпки:

1. Идентификация на параметъра *p* на съвкупност, който ще бъде оценяван (например $error_D(h)$).
2. Определяне на оценителя *Y* (например $error_S(h)$). Желателно е да бъде избран обективен оценител с най-малката дисперсия.
3. Определяне на вероятностното разпределение $\mathcal{D}_Y$, на което се подчинява оценителят *Y*, включително неговите средна стойност и дисперсия.
4. Определяне на *N%* доверителен интервал чрез намирането на прагове *L* и *U*, такива че *N%* от вероятността от разпределение $\mathcal{D}_Y$ попада между *L* и *U*.

18.5.1. Централна пределна теорема

Изводът на доверителния интервал е съществено улеснен от една фундаментална за теорията на оценяването теорема, наречена "*централна пределна теорема*". Да разгледаме отново общата постановка на задачата, когато наблюдаваме значения на *n* независимо избрани случайни величини $Y_1, \ldots, Y_n$, които имат едно и също, неизвестно за нас, вероятностно разпределение (например *n* хвърляния на една и съща монета). Нека μ означава средното на това неизвестно разпределение, управляващо всяка случайна величина Y_i, а σ – стандартното му отклонение. Казваме, че тези Y_i са *независими, еднакво разпределени случайни величини*, тъй като те описват независими експерименти, всеки от които се подчинява на едно и също вероятностно разпределение. В опита да оценим средното на вероятностното

разпределение на Y_i, ние изчисляваме средното на извадката $\bar{Y}_n \equiv \frac{1}{n}\sum_{i=1}^{n} Y_i$ (например честота на наблюдаване на лице при n хвърляния на монетата). Централната пределна теорема утвърждава, че вероятностното разпределение, на което се подчинява случайната величина $\bar{Y}_n$, се приближава до нормалното разпределение при $n \to \infty$, *независимо от разпределението, на което се подчиняват основните случайни величини* Y_i. Освен това, средното на вероятностното разпределение, определящо $\bar{Y}_n$, се приближава към μ, а неговото стандартно отклонение – към $\frac{\sigma}{\sqrt{n}}$. По формално:

> ***Централна пределна теорема:*** Нека $Y_1, \ldots, Y_n$ е множеството от независими, еднакво разпределени случайни величини, подчиняващи се на някое произволно вероятностно разпределение със средното μ и ограничена дисперсия σ^2. Нека $\bar{Y}_n \equiv \frac{1}{n}\sum_{i=1}^{n} Y_i$ е извадковата средна стойност. Тогава, при $n \to \infty$ разпределението на случайната величина $\frac{\bar{Y}_n - \mu}{\frac{\sigma}{\sqrt{n}}}$ се приближава до нормалното разпределение със средното 0 и стандартното отклонение 1.

Това е доста учудващ факт, тъй като той утвърждава, че ние знаем вида на разпределението, на което се подчинява средната на извадката $\bar{Y}_n$, дори когато не знаем нищо за вида на разпределението, на което се подчиняват отделни наблюдаваните случайни величини Y_i! Освен това, централната пределна теорема описва, как средното и дисперсията на $\bar{Y}_n$ могат да се използват за определяне на средното и дисперсията на отделни Y_i.

Централната пределна теорема е много полезна, тъй като утвърждава, че когато ние определяме някой оценител, който е средното на определената извадка (например $error_S(h)$), разпределението, определящо този оценител, може да се апроксимира с нормалното разпределение при достатъчно голяма стойност на n. Ако също така ни е известна и дисперсията на това (приблизително) нормално разпределение, то чрез формулата (18.8) можем да изчислим необходимия доверителен интервал. Едно общоприето емпирично правило гласи, че апроксимацията чрез нормалното разпределение може да се ползва при $n \geq 30$. Да припомня, че в предишния раздел ние сме ползвали такова нормалното разпределение за апроксимация на биномното разпределение, което по-точно описва $error_S(h)$.