Майнор "Статистический анализ и его применение"

Модели и первичный анализ статистических данных

Лекция 3. Основные понятия теории вероятностей и прикладной статистики

Случайные события и случайные величины. Числовые характеристики случайной величины. Функции распределения для дискретных и непрерывных случайных величин. Математическое ожидание, дисперсия и другие статистические характеристики. Нормальный, биномиальный, пуассоновский и другие законы распределения вероятностей. Примеры функций распределения в различных задачах прикладной статистики.

Случайность и детерминизм

Для изучения явлений окружающего мира необходимо строить различные математические модели. Некоторые явления и события описываются детерминированными моделями. Но в силу сложности многих объектов и в силу случайности окружающего мира необходимо также знать и уметь использовать закономерности, свойственные моделям, в которых учитывается элементы случайности, неопределенности, непредсказуемости. Для этого развиваются такие направления математики как теория вероятности, математическая статистика, теория случайных процессов.

Примеры - подбрасывание монеты, бросание кубика, рулетка, различные игры с непредсказуемыми исходами, спортивные состязания, различные атмосферные явления, биологическое разнообразие, демографические закономерности и т.п. Во всем есть доля детерминизма и есть доля случайности.

Теория вероятности, статистика и анализ данных

В теории вероятности вводятся понятия случайных исходов, случайных событий и случайных величин, которые количественно характеризуют случайные явления. Теория вероятностей изучает модели случайных величин, свойства этих моделей и то, какие выводы можно сделать о том, какие события будут нас ожидать при проведении испытаний со случайными величинами в будущем.

Статистика и анализ данных действуют в обратном направлении. Набор реализаций случайной величины называется выборкой из нее. По конкретным наблюдениям (по конечным выборкам) надо определить параметры модели, которая генерирует случайные величины. Отфильтровать существенную часть и избавиться от несущественной информации. Предсказать, если возможно, как эта случайная величина будет себя вести в будущих испытаниях.

Вероятности случайных событий

Каждому случайному событию A можно поставить в соответствие число $P(A)$, которое будет являться мерой возможности его появления - вероятность события A.

Свойства вероятности

1) $0 \le P(A) \le 1$.

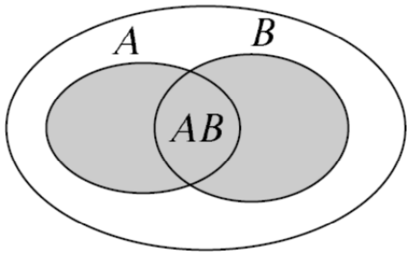
2) Вероятность невозможного события равна нулю: $P(\varnothing)=0$.

3) Вероятность достоверного события равна единице: $P(\Omega)=1$.

4) $A+\overline{A}=\Omega \quad \Rightarrow \quad P(A)+P(\overline{A})=1$.

5) $A \subseteq B \quad \Rightarrow \quad P(A) \le P(B)$

6) $P(A+B)=P(A)+P(B)-P(AB)$



7) Независимые события: $P(AB)=P(A)P(B)$.

8) Условная вероятность: $P(A|B)=\dfrac{P(AB)}{P(B)}$, если $P(B)>0$.

$$P(AB)=P(A|B)P(B)=P(B|A)P(A).$$

9) Формула полной вероятности:

$$P(A)=P(A|B)P(B)+P(A|\overline{B})P(\overline{B}).$$

10) Формула Байеса:

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)}, \text{ если } P(B)>0.$$

Случайной величиной называют числовую величину, значение которой зависит от того, какое именно событие произошло.

Дискретные случайной величины

Дискретная случайная величина X принимает счетное множество значений $\{x_1, x_2, x_3, \ldots\}$ с вероятностями $\{p_1, p_2, p_3, \ldots\}$, где $p_i \equiv P(X = x_i) \geq 0$ $\forall i$ и выполняется свойство нормировки: $\sum_{i=1}^{\infty} p_i = 1$

Дискретные случайные величины описываются с помощью таблиц распределения вероятности

Значения	x_1	x_2	x_3	$\ldots$	x_n
Вероятности	p_1	p_2	p_3	$\ldots$	p_n

с помощью графиков распределения вероятности.

Примеры дискретных случайных величин

1. Дискретная случайная величина ξ с двумя исходами (схема Бернулли, распределение Бернулли):

$$P(\xi = 1) = p, \quad P(\xi = 0) = q = 1 - p.$$

$$\xi \sim Ber(p).$$

2. Биномиальная случайная величина $X = \sum_{i=1}^{n} \xi_i$, где ξ_i взаимно независимые бинарные случайные величины с исходами 0 и 1:

$$P(X = k) = C_n^k p^k (1-p)^{n-k}, \quad k = 0, 1, 2, \ldots, n.$$

$$C_n^k = \frac{n!}{k!(n-k)!}.$$

$$X \sim Bin(n, p).$$

3. Пуассоновская случайная величина $X \sim Pois(\lambda)$:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad \lambda > 0, \quad k = 0, 1, 2, \ldots.$$

Число автобусов за 1 час, число определенных слов в единице текста, число опечаток и т.п. (закон редких событий).

Непрерывные случайные величины

Описываются с помощью функции распределения:

$$F_X(x) = P(X \leq x)$$

и с помощью плотности распределения (плотности вероятности, дифференциальной функции распределения):

$$f_X(x): \quad \int_a^b f_X(x) dx = P(a \leq X \leq b).$$

Свойства функций распределения

1) $F(x)=\int\limits_{-\infty}^{x} f(t)dt$; $f(x)=\dfrac{dF(x)}{dx}\ge 0$.

2) $0\le F(x)\le 1$.

3) $F(-\infty)=0$.

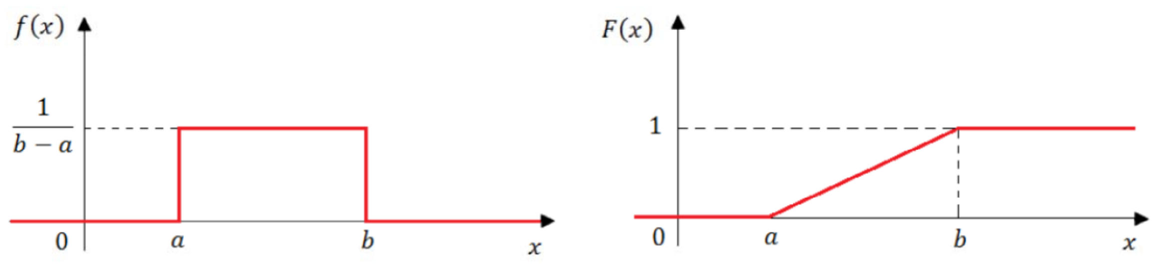
4) Свойство нормировки: $F(\infty)=\int\limits_{-\infty}^{+\infty} f(t)dt=P(-\infty<X<+\infty)=1$.

Примеры непрерывных случайных величин

1. Равномерно распределенная случайная величина: $X\sim U(a,b)$.

   Плотность вероятности равномерная на отрезке от *a* до *b*:

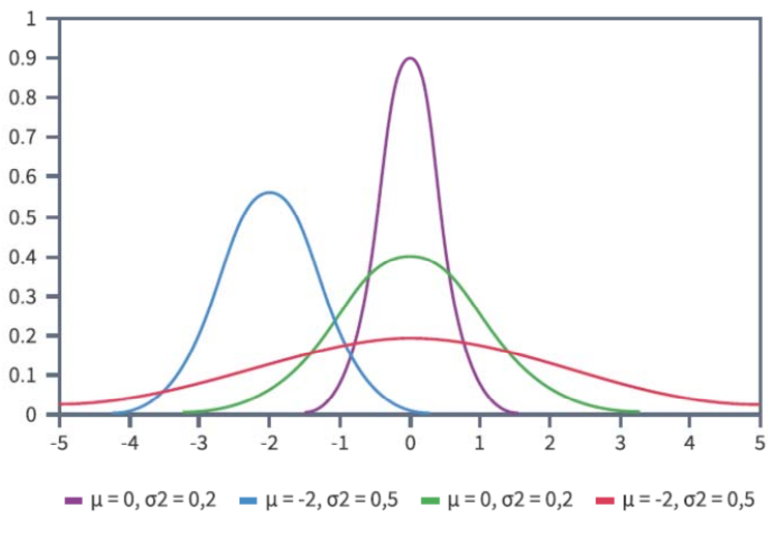
$$f(x)=\begin{cases}\dfrac{1}{b-a}, & x\in[a,b]\\ 0, & x\notin[a,b]\end{cases}$$



2. Нормальное распределение (гауссовское распределение): $X\sim N(\mu,\sigma^2)$.

   Плотность вероятности распределена от -∞ до +∞:

$$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$



Математическое ожидание и дисперсия случайных величин

Для дискретных случайных величин X с распределением

Значения	x_1	x_2	x_3	$\ldots$	x_n
Вероятности	p_1	p_2	p_3	$\ldots$	p_n

математическое ожидание (среднее значение) $E(X)$ и дисперсия $\operatorname{var}(X)=\sigma_X^2=E\left[\left(X-E(x)\right)^2\right]=E\left(X^2\right)-\left(E(x)\right)^2$ определяются с помощью формул суммирования

$$E(X)=\sum_{i=1}^{n} p_i x_i = p_1 x_1 + p_2 x_2 + \ldots + p_n x_n\,;$$

$$\sigma_X^2=\sum_{i=1}^{n} p_i\left(x_i-E(X)\right)^2=\sum_{i=1}^{n} p_i x_i^2-\left(\sum_{i=1}^{n} p_i x_i\right)^2.$$

Для непрерывных случайных величин X с функцией распределения вероятности $F(x)$ и плотностью вероятности $f(x)$ среднее значение $E(X)$ и дисперсия $\operatorname{var}(X)=\sigma_X^2=E\left[\left(X-E(x)\right)^2\right]=E\left(X^2\right)-\left(E(x)\right)^2$ определяются с помощью интегрирования

$$E(X)=\int_{-\infty}^{\infty} x\,dF(x)=\int_{-\infty}^{\infty} xf(x)\,dx\,;$$

$$\sigma_X^2=\int_{-\infty}^{\infty}\left[x-E(X)\right]^2 f(x)\,dx=E\left(X^2\right)-\left[E(X)\right]^2=$$

$$=\int_{-\infty}^{\infty} x^2 f(x)\,dx-\left[\int_{-\infty}^{\infty} xf(x)\,dx\right]^2.$$