Лекция 1. Основы математической статистики.

27 февраля 2013 г.

Случайные выборки

Определение

Случайная выборка размера n, отвечающая случайной величине X с функцией распределения $F(x)$ - набор n независимых случайных величин $X_1, \ldots, X_n$, имеющих функцию распределения $F(x)$.

Определение

Статистика - функция случайной выборки.

Эмпирическая функция распределения

Определение

Эмпирическая функция распределения:

$$F_n(x) = \frac{1}{n}\sum_{l=1}^{n} \mathbf{1}_{(-\infty,x)}(X_l)$$

Теорема (Гливенко)

$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \xrightarrow[n\to\infty]{} 0$ *с вероятностью 1*

Использование в R

```
> set.seed(1);x <- rnorm(1000)
> ecdfx <- ecdf(x)
> plot(ecdfx)
```

Гистограмма

Определение

Эмпирическая плотность распределения(гистограмма):

$$f_{n,h}(x) = \frac{1}{nh} \sum_{l=1}^{n} \mathbf{1}_{(x_h, x_h+h)}(X_l)$$

где $x_h = [x/h]h$

Теорема

Если $n \to \infty$, $h \to 0$, $nh \to \infty$, *то* $\forall x$, *таких, что* $f(x)$ - *непрерывна,* $f_{n,h}(x) \to f(x)$ *по вероятности.*

Использование в R

```
> h <- hist(x)
```

Оценки параметров распределения

Определение (Выборочное среднее)

$$\overline{X}_n = \frac{1}{n}\sum_{l=1}^{n} X_l$$

Определение (Выборочная дисперсия)

$$S_n^2 = \frac{1}{n-1}\sum_{l=1}^{n} (X_l - \overline{X}_n)^2$$

Использование в R

```
> m <- mean(x)
> v <- var(x)
```

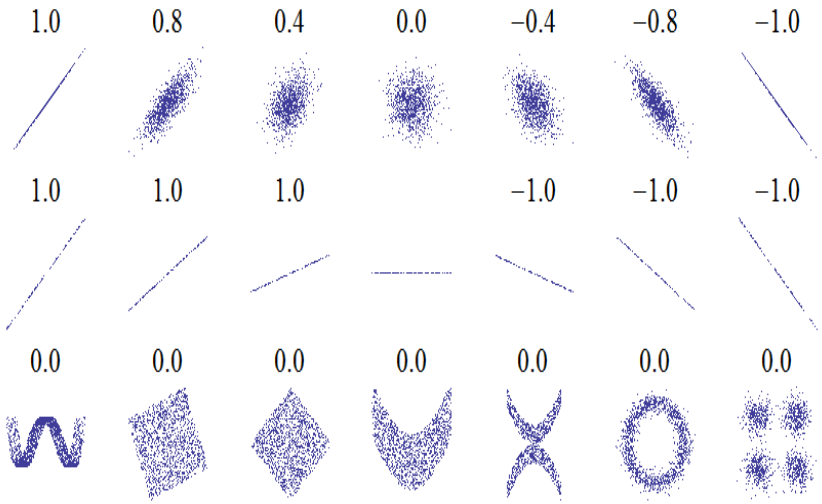
Коэффициент корреляции

Определение (Выборочный коэффициент корреляции)

$$R_n = \frac{\sum_{l=1}^{n} (X_l - \overline{X})(Y_l - \overline{Y})}{\sqrt{S_n^2(X) S_n^2(Y)}}$$

Использование в R

```
> corx <- cor(x)
```



Доверительные интервалы

Определение (Доверительный интервал)

Интервал $(\underline{\theta}(X_1, \ldots, X_n), \overline{\theta}(X_1, \ldots, X_n))$ *называется доверительным для параметра* θ *с уровнем значимости* α*, если*

$$P(\underline{\theta}(X_1, \ldots, X_n) < \theta < \overline{\theta}(X_1, \ldots, X_n)) \geq 1 - \alpha$$

Доверительные интервалы

Пример

Легко показать, что для выборки из нормального распределения $\mathrm{N}(\theta, \sigma^2)$

$$\sqrt{n} \cdot \frac{\overline{X}_n - \theta}{\sigma} \sim \mathrm{N}(0, 1)$$

Таким образом, доверительный интервал для параметра θ *будет*

$$(\overline{X}_n - \frac{z_{1-\frac{\alpha}{2}}\sigma}{\sqrt{n}}, \overline{X}_n + \frac{z_{1-\frac{\alpha}{2}}\sigma}{\sqrt{n}})$$

Проверка статистических гипотез

Определение

$X_1, \ldots, X_n \sim F(x, \theta)$, $\theta \in \Theta = \Theta_0 \cup \Theta_1$, $\Theta_0 \cap \Theta_1 = \emptyset$

Нулевая гипотеза H_0: $\theta \in \Theta_0$

Альтернативная гипотеза H_1: $\theta \in \Theta_1$

Статистический критерий - правило, позволяющее отвергнуть или принять нулевую гипотезу.

Определение

Вероятность ошибки первого рода α *- вероятность отвергнуть гипотезу* H_0 *при условии, что она верна.*

Определение

Вероятность ошибки второго рода - вероятность принять гипотезу H_0 *при условии, что верна гипотеза* H_1.

Пример

Есть выборка $X_1, \ldots, X_n \sim \mathrm{N}(\theta, \sigma^2)$, *где* σ^2 - *известный параметр, а* θ - *неизвестный.*
Хотим проверить гипотезу $\theta = \theta_0$.
Посчитаем следующую статистику:

$$\sqrt{n} \cdot \frac{\overline{X}_n - \theta_0}{\sigma}$$

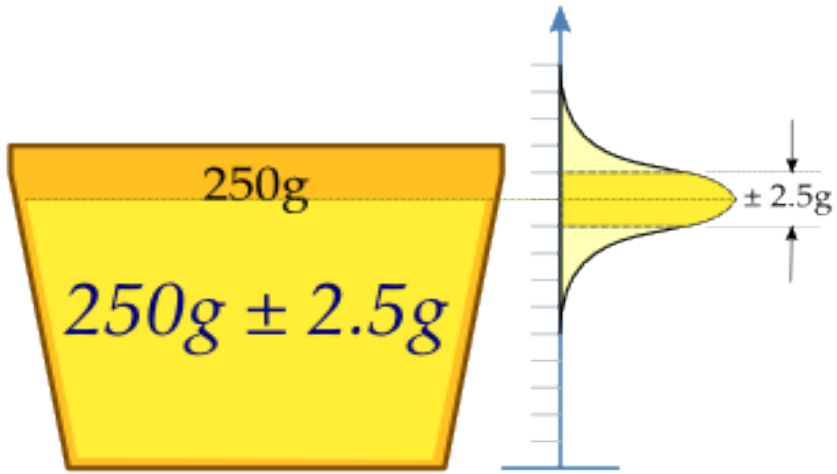
Если $\theta = \theta_0$, *то эта статистика имеет распределение* $\mathrm{N}(0, 1)$.
Таким образом, если в качестве статистического критерия рассмотреть попадание в доверительный интервал

$$(\overline{X}_n - \frac{z_{1-\frac{\alpha}{2}}\sigma}{\sqrt{n}}, \overline{X}_n + \frac{z_{1-\frac{\alpha}{2}}\sigma}{\sqrt{n}}),$$

то вероятность ошибки первого рода будет α.

Пример

Машина должна заполнять упаковку веществом на 250 грамм. Задача: имея выборку из 25 упаковок, проверить с уровнем значимости 0.05, правильно ли откалибрована машина. Ранее было оценено, что распределение заполнения вещества нормально со средним отклонением 2.5 грамма.



Правильно ли откалибрована машина, если среднее значение выборки 251 грамм?

Определение (P-значение)

Пусть есть случайная выборка и некоторая статистика с известным распределением.
Обычно (!), P-значение - это вероятность того, что случайная величина, имеющая то же распределение, что и статистика, примет значение, большее(в случае двустороннего распределения, возможно, и меньшее) фактического значения статистики на данной выборке.
P-значение следует сравнивать с каким-нибудь заранее выбранным порогом, например 0.05. Если оно меньше, то нулевая гипотеза отвергается.

Проверка гипотезы нормальности

Определение (Статистика Шапиро-Уилка)

$$W = \frac{1}{s^2}\left[\sum_{i=1}^{n} a_{n-i+1}(x_{n-i+1} - x_i)\right]^2$$

Использование в R

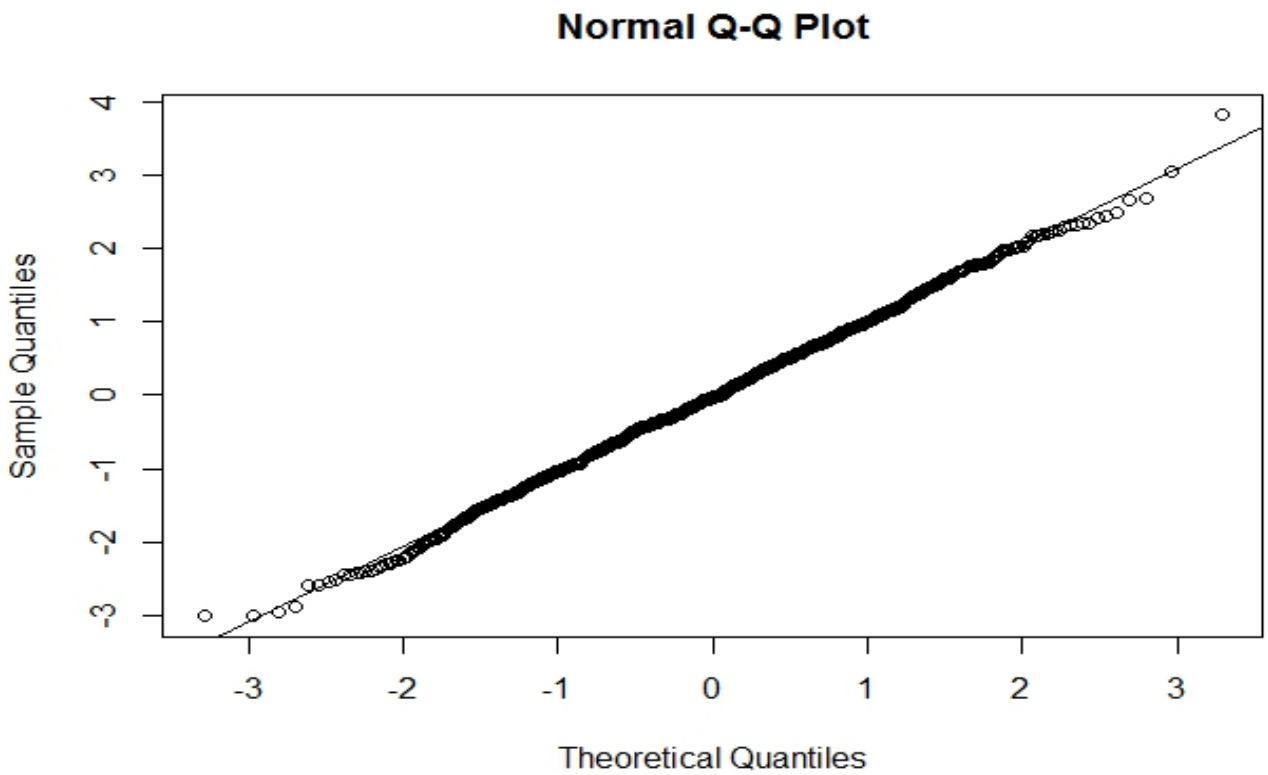
```
> shapiro.test(x)

    Shapiro-Wilk normality test
data:  x
W = 0.9988, p-value = 0.7258
```

Проверка гипотезы нормальности

Использование в R (График квантиль-квантиль)

```
> qqnorm(x)
> qqline(x)
```



Метод максимального правдоподобия

Определение

$X_1, \ldots, X_n$ - случайная выборка с плотностью распределения $f(x, \theta)$. Оценкой максимального правдоподобия для θ называется

$$\theta_n = \underset{\theta}{argmax}(f(X_1, \theta) \cdot \ldots \cdot f(X_n, \theta))$$

Использование в R

```
> library("MASS")
> f <- fitdistr(x,"normal")
      mean             sd
  -0.01164814    1.03439825
 ( 0.03271054) ( 0.02312985)
```