Wnioskowanie statystyczne

21 marca 2014

Powtórzenie

x - zmienna losowa ~ $N(x_0,\sigma)$

Jakie jest prawdopodobieństwo, że

$$x_0 - z\,\sigma \;\square\; x_{śr} \;\square\; x_0 + z\,\sigma$$

Powtórzenie

x - zmienna losowa ~ $N(x_0,\sigma)$

Jakie jest prawdopodobieństwo, że

$$x_0 - z\,\sigma \square x_{śr} \square x_0 + z\,\sigma$$

Takie same jak prawdopodobieństwo, że

$$-z \square (x_{śr} - x_0)/\sigma \square z$$

N(0,1)

P = poziom ufności

```
P = 2 * norm.cdf(z) - 1
```

Powtórzenie

x - zmienna losowa ~ $N(x_0,\sigma)$

Jakie jest prawdopodobieństwo, że

$$x_0 - z\,\sigma \;\square\; x_{śr} \;\square\; x_0 + z\,\sigma$$

Takie same jak prawdopodobieństwo, że

$$-z \;\square\; (x_{śr} - x_0)/\sigma \;\square\; z$$

N(0,1)

P = poziom ufności

```
P = 2 * norm.cdf(z) - 1
```

Jeżeli oznaczymy α = 1−P = poziom istotności

```
α = 2 - 2 * norm.cdf(z)
z = norm.ppf(1 - α / 2)
```

Wartość oczekiwana (1)

Założenia:

1) badana zmienna ma rozkład normalny
2) znamy wariancję σ^2 populacji

   $x_{śr}$ - zmierzona wartość średniej (lub `numpy.mean(...)`)
   n - liczba pomiarów
   σ - znana wariancja

α - poziom istotności (np. 0.05)

$$z_{\alpha/2} = \texttt{scipy.stats.norm.ppf}(1-\alpha/2)$$

przedział ufności dla wartości oczekiwanej populacji:

$$x_{śr} \pm z_{\alpha/2}\ \sigma / \sqrt{n}$$

Dygresja: rozmiar próby

"Promień" przedziału ufności, czyli maksymalna odległość pomiędzy wartością zmierzoną i prawdziwą wynosi

$$D = z_{\alpha/2}\ \sigma\ /\ \sqrt{n}$$

Rozmiar próby dla ustalonego $D \square D_0$

$$z_{\alpha/2}\ \sigma\ /\ \sqrt{n} \square D_0$$

$$\sqrt{n} \square z_{\alpha/2}\ \sigma\ /\ D$$

```
z_{α/2} = scipy.stats.norm.ppf(1-α/2)
  n = math.ceil( (z_{α/2}*σ/D)**2 )
```

Wartość oczekiwana (2)

Założenia:

1) badana zmienna ma rozkład normalny
2) **nie** znamy wariancji populacji

$x_{śr}$ - zmierzona wartość średniej (lub `numpy.mean(`*`pomiary`*`)`)
n - liczba pomiarów

α - poziom istotności (np. 0.05)

```
s = numpy.std(pomiary)
t_α/2 = scipy.stats.t(n-1).ppf(1-α/2)
```

przedział ufności dla wartości oczekiwanej populacji:

$$x_{śr} \pm t_{\alpha/2}\, s / \sqrt{(n-1)}$$

Wartość oczekiwana (3)

Założenia:

1) badana zmienna ma rozkład normalny
2) **nie** znamy wariancji populacji
3) **pomiarów jest bardzo dużo** (n > 30)

$x_{śr}$ - zmierzona wartość średniej (lub `numpy.mean(pomiary)`)
n - liczba pomiarów

α - poziom istotności (np. 0.05)

```
s = numpy.std(pomiary)
```

$z_{\alpha/2}$ = `scipy.stats.norm.ppf(1-α/2)`

przedział ufności dla wartości oczekiwanej populacji:

$$x_{śr} \pm z_{\alpha/2}\, s / \sqrt{n}$$

Wariancja

Założenia:

1) badana zmienna ma rozkład normalny
2) **nie** znamy wariancji populacji

   $x_{śr}$ - zmierzona wartość średniej (lub `numpy.mean(`*`pomiary`*`)`)
   n - liczba pomiarów

α - poziom istotności (np. 0.05)

$$s = \texttt{numpy.std}(\textit{pomiary})$$

$$x^2_L = \texttt{scipy.stats.chi2(n-1).ppf}(\alpha/2)$$

$$x^2_P = \texttt{scipy.stats.chi2(n-1).ppf}(1-\alpha/2)$$

przedział ufności dla **wariancji** populacji:

$$(n-1)s^2/x^2_L \le \sigma^2 \le (n-1)s^2/x^2_P$$

Co, jeżeli

- zmienna losowa **nie** jest z rozkładu normalnego?
- chcemy zbadać przedział ufności
  np. dla **mediany** ?
- chcemy numerycznie zweryfikować wyniki?
- nie znamy wzorów?

Bootstrap

Jedna seria pomiarów → jeden wynik
Wiele serii pomiarów → wiele wyników

Jak zrobić wiele serii z jednej?

Bootstrap

Jedna seria pomiarów → jeden wynik

Wiele serii pomiarów → wiele wyników

Jak zrobić wiele serii z jednej?

Losujemy ze zwracaniem.

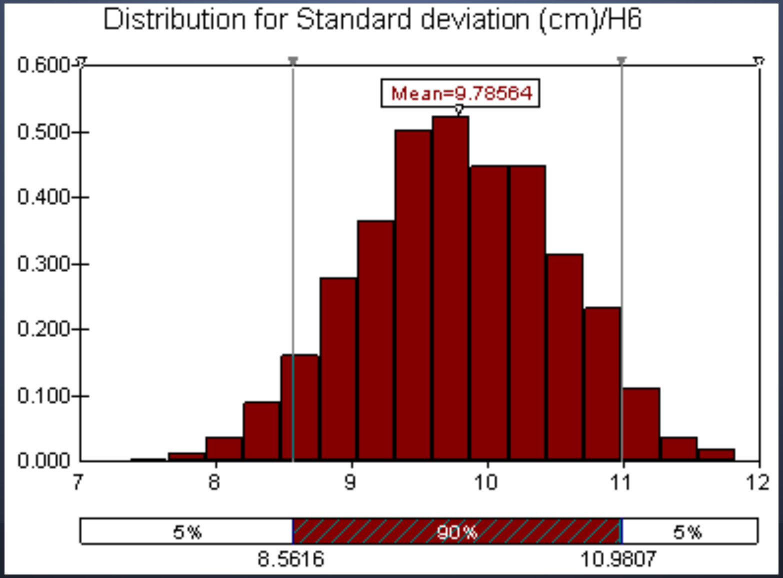
```
def randsample(old_sample) :
    n = len(old_sample)
    ind = numpy.random.randint(n, size = n)
    y = old_sample[ind]
    return y
```

lub np.

```
n = len(old_sample)
numpy.random.choice(old_sample, size = n)
```

Bootstrap

1. Generujemy **wiele** serii pomiarów i dla każdej z nich obliczamy żądaną wartość.
2. Wyniki tworzą próbę z jakiegoś (nieznanego) rozkładu.
3. Wyciągamy z wyników odpowiedni kwantyl.



Bootstrap

1. Generujemy **wiele** serii pomiarów i dla każdej z nich obliczamy żądaną wartość.
2. Wyniki tworzą próbę z jakiegoś (nieznanego) rozkładu.
3. Wyciągamy z wyników odpowiedni kwantyl.

Przykład (przedział ufności dla średniej przy α=0.1 czyli 10%):

1. Generujemy wiele serii pomiarów i dla każdej obliczamy średnią.
2. Bierzemy wyniki (tablicę średnich) i obliczamy kwantyle dla α/2 i 1-α/2
   - x_L `= scipy.stats.scoreatpercentile(wyniki, 5.0)`
   - x_P `= scipy.stats.scoreatpercentile(wyniki, 95.0)`
3. Przedziałem ufności jest [x_L ; x_P]