Verteilungsfunktionen und Quantile

April 26, 2016

1 Wichtige Verteilungsfunktionen

1.1 Normalverteilung (z-Verteilung)

Die Gaußverteilung ist gegeben durch

$$p(x) = \frac{1}{\sigma_x\sqrt{2\pi}} e^{-\frac{(x-\mu_x)^2}{2\sigma_x^2}}$$

Es gilt

$$E[x] = \mu_x$$

und

$$E[(x-\mu_x)^2] = \sigma_x^2$$

Substitution von $z = \frac{x-\mu_x}{\sigma_x}$ liefert die standardisierte Normalverteilung

$$p(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

für die gilt $E[x] = \mu_z = 0$ und $E[(x-\mu_z)^2] = \sigma_z^2 = 1$
Die Wahrscheinlichkeit P berechnet sich aus dem Integral

$$P(z_\alpha) = \int_{-\infty}^{z_\alpha} p(z)dz = Prob[z < z_\alpha] = 1 - \alpha$$

bzw.

$$1 - P(z_\alpha) = \int_{z_\alpha}^{\infty} p(z)dz = Prob[z > z_\alpha] = \alpha$$

1.2 Normalverteilung: scipy.stats.norm(x)

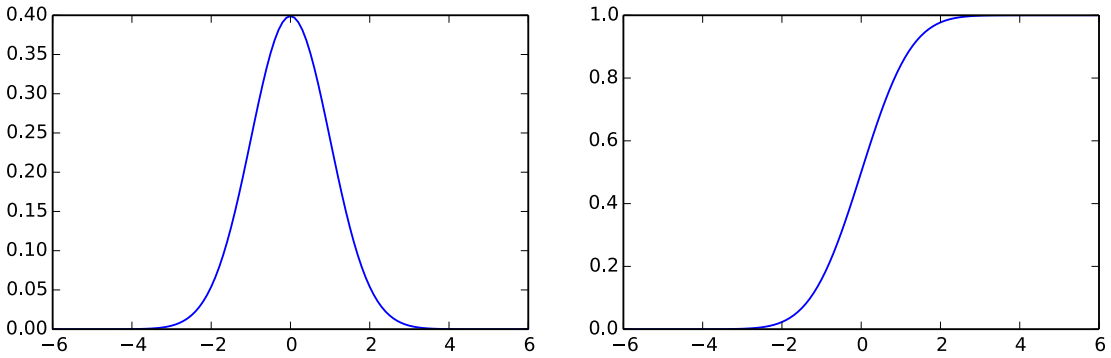
```
In [3]: %pylab inline
        %config InlineBackend.figure_format = 'svg'
```

```
Populating the interactive namespace from numpy and matplotlib
```

```
In [4]: import scipy.stats as stats

        x=linspace(-6,6,121)
        figure(figsize=(10,3))
        subplot(1,2,1)
        plot(x,stats.norm.pdf(x))
        subplot(1,2,2)
        plot(x,stats.norm.cdf(x))
```

```
Out[4]: [<matplotlib.lines.Line2D at 0x7fa546ba5f90>]
```




```
In [24]: #Wieviele Werte liegen innerhalb +-1, +-2 und +-3 Standardabweichungen?
         #beidseitig: *2

         z_cdf=stats.norm.cdf

         print (1-z_cdf(-1.0)*2)*100
         print (1-z_cdf(-2.0)*2)*100
         print (1-z_cdf(-3.0)*2)*100
```

```
68.2689492137
95.4499736104
99.7300203937
```

1.3 χ^2-Verteilung

Gegeben sind n unabhängige z-verteilte Zufallsvariablen $z_1, z_2, ..., z_n$. Die Summe bildet die Zufallsvariable

$$\chi^2 = z_1^2 + z_2^2 + ... + z_n^2$$

mit n-Freiheitsgraden. Die Wahrscheinlichkeitsdichteverteilung für $\chi^2 > 0$ ist gegeben durch

$$p(\chi) = \frac{1}{2^{n/2}\Gamma(n/2)} \chi^{(n/2-1)} e^{-\frac{x}{2}}$$

mit der Gamma-Funktion $\Gamma(n/2)$.

Es gilt

$$E[\chi_n^2] = \mu_{\chi^2} = n$$

und

$$E[(\chi_n^2 - \mu_{\chi^2})^2] = \sigma_{\chi^2}^2 = 2n$$

Die Gamma-Funktion ist definiert als

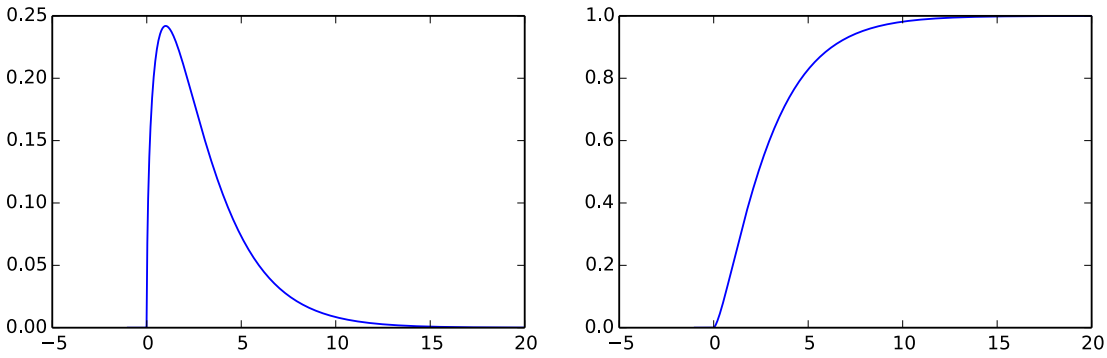
$$\Gamma(z) = \int_0^\infty t^{n-1} e^{-t}\, dt$$

Für positive Integer n ist das Ergebnis $\Gamma(n) = (n-1)!$.

1.4 χ^2-Verteilung: scipy.stats.chi2(x,df)

```
In [3]: x=linspace(-1,20,1000)
        n=3# Freiheitsgrade
        figure(figsize=(10,3))
        subplot(1,2,1)
        plot(x,stats.chi2.pdf(x,n))
        subplot(1,2,2)
        plot(x,stats.chi2.cdf(x,n))
```

```
Out[3]: [<matplotlib.lines.Line2D at 0x7f9fa2dcc050>]
```



1.5 Γ-Funktion: scipy.special.gamma(x)

```
In [45]: from scipy.special import gamma
         for i in range(1,20):
             print gamma(i)
```

```
1.0
1.0
2.0
6.0
24.0
120.0
720.0
5040.0
40320.0
362880.0
3628800.0
39916800.0
479001600.0
6227020800.0
87178291200.0
1.307674368e+12
2.0922789888e+13
3.55687428096e+14
6.40237370573e+15
```

1.6 F-Verteilung

Gegeben seien die Zufallsvariablen y_1 und y_2 mit χ^2-Verteilung und n_1 bzw. n_2 Freiheitsgraden. Die F-Verteilung ergibt sich aus dem Verhältnis

$$F_{n_1,n_2} = \frac{y_1/n_1}{y_2/n_2} = \frac{y_1 n_2}{y_2 n_1}$$

1.7 Student t-Verteilung

Gegeben sind zwei unabhängige Zufallsvariablen y und z. y folgt einer χ^2_n-Verteilung und z ist normalverteilt mit Freiheitsgrad n. Die Student $t-$Verteilung errechnet sich aus

$$t_n = \frac{z}{\sqrt{y/n}}$$

Es gilt

$$E[t_n] = \mu_{\chi^2} = 0$$

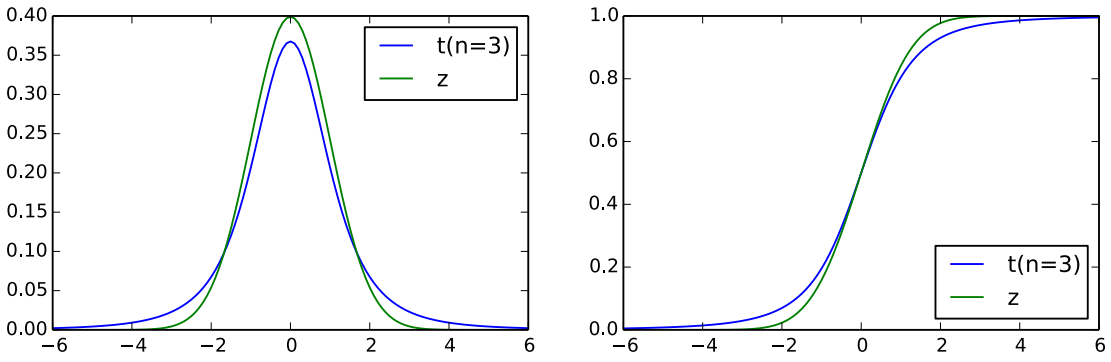
und

$$E[(t_n - \mu_t)^2] = \sigma_t^2 = \frac{n}{n-2}$$

für $n > 2$.

```
In [4]: x=linspace(-6,6,121)
        n=3 # Freiheitsgrade
        figure(figsize=(10,3))
        subplot(1,2,1)
        plot(x,stats.t.pdf(x,n),label='t(n='+str(n)+')')
        plot(x,stats.norm.pdf(x),label='z')
        legend()
        subplot(1,2,2)
        plot(x,stats.t.cdf(x,n),label='t(n='+str(n)+')')
        plot(x,stats.norm.cdf(x),label='z')
        legend(loc=4)
```

```
Out[4]: <matplotlib.legend.Legend at 0x7f9f9be8b5d0>
```



2 Stichproben-Verteilungsfunktionen

2.1 Stichproben-Mittel bei bekannter Varianz

N unabhängige Messungen x (normalverteilt) mit Mittelwert μ_x und Varianz σ_x^2

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Der Mittelwert beträgt

$$\mu_{\bar{x}} = \mu_x$$

mit Varianz

$$\sigma_{\bar{x}}^2 = \frac{\sigma_x^2}{N}$$

Der sogenannte Standardfehler (gegeben durch die Standardabweichung σ) reduziert sich proportional zu $\frac{1}{\sqrt{N}}$

2.2 Stichproben-Varianz

Die Stichproben-Varianz ist gegeben als

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

Für N unabhängige Messungen x (normalverteilt) mit Mittelwert μ_x und Varianz σ_x^2 gilt

$$\sum_{i=1}^{N}(x_i - \bar{x})^2 = \sigma_x^2 \chi_n^2$$

mit der χ_n^2-Verteilung mit $n = N - 1$ Freiheitsgraden. Es folgt für die Wahrscheinlichkeit

$$Prob[s^2 > \frac{\sigma_x^2 \chi_{n,\alpha}^2}{n}] = \alpha$$

3 Auswahl spezieller Prüfverfahren

3.1 Vergleich zweier Mittelwerte

Die Stichproben-Mittelwerte $\bar{a}$ und $\bar{b}$ sollen auf signifikanten Unterschied geprüft werden. Bei gleichem Stichprobenumfang $n_a = n_b$ errechnet sich die Wahrscheinlichkeit aus den Quantilen der t-Verteilung

$$\hat{t} = \frac{|\bar{a} - \bar{b}|\sqrt{n}}{s_a^2 + s_b^2}$$

mit der Anzahl der Freiheitsgrade beträgt $\Phi = 2n - 2$ ($n = n_a = n_b$)

3.2 Vergleich Stichproben-Mittel mit bekanntem Mittelwert

Der Stichproben-Mittelwert $\bar{a}$ soll auf signifikanten Unterschied hinsichtlich des bekannten Mittelwertes μ geprüft werden.

$$\hat{t} = |\bar{a} - \mu|\frac{\sqrt{n}}{s^2}$$

Die Anzahl der Freiheitsgrade beträgt $\Phi = n - 1$.

3.2.1 Beispiel

Beispiel aus Schönwiese Seite 138.

Ein geowissenschaftlicher zu untersuchender Prozess folge in guter Näherung der Normalverteilung mit μ=10 und σ=2.7 Maßeinheiten. Es ist zu prüfen, ob die Stichprobe mit $\bar{a} = 7.5$ und s=1.8 damit vereinbar ist, wobei der Stichproben-Umfang $n = 50$ betragen soll. Der Tabellenwert für das t-Quantil beträgt $t(\Phi = 49, \alpha = 0.05) \approx 1.68$.

```
In [14]: mu=10.0
         s=2.7

         a=7.5
         sa=1.8

         n=50.0
         df=n-1

         t=abs(a-mu)*sqrt(n)/s
         print t
```

```
6.54728501099
```

```
In [16]: a=0.05
         df=n-1
         x=linspace(-10,0,10000)
         print abs(x[argmax(stats.t.cdf(x,df)>a)])
         print abs(x[argmax(stats.norm.cdf(x)>a)]) # Für große n ist das t-Quantil ähnlich der Normalve
```

```
1.67616761676
1.64416441644
```

3.2.2 Folgerung

Es ist 6.54> 1.67

Daraus folgt, dass der Messwert nicht mit dem Prozess vereinbar ist.

4 Quantile

```
In [102]: # Quantile (Verteilunfsfunktion) der Student-Verteilung (tV)
          # Vergleiche Schönwiese A.3 (S. 287)
          # oder http://de.wikipedia.org/wiki/Studentsche_t-Verteilung

          a=0.05
          x=linspace(-10,0,10000)
          for df in range(1,10): #Freiheitsgrade
              print df," %1.3f " % abs(x[argmax(stats.t.cdf(x,df)>a)])
          print '...'
          df=49
          print df," %1.3f " % abs(x[argmax(stats.t.cdf(x,df)>a)])
          print '-------------------------------------------------------------------------'
          print 'Vergleich von t-Verteilung und Normalverteilung für große Anzahl Freiheitsgrade'
          df=10000
          print df," %1.3f " % abs(x[argmax(stats.t.cdf(x,df)>a)])

          print df," %1.3f " % abs(x[argmax(stats.norm.cdf(x)>a)])
```

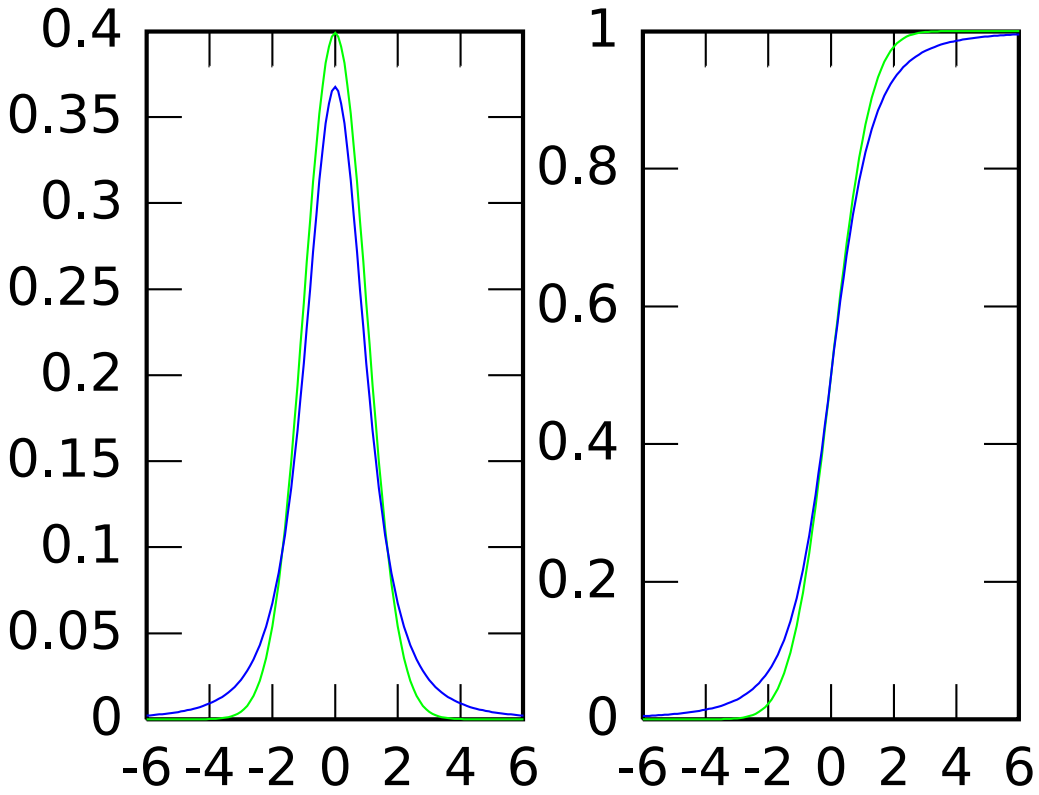
```
1  6.314
2  2.919
3  2.353
4  2.131
5  2.014
6  1.942
7  1.894
8  1.859
9  1.832
...
49  1.676
--------------------------------------------------------------------------------
Vergleich von t-Verteilung und Normalverteilung für große Anzahl Freiheitsgrade
10000  1.644
10000  1.644
```

4.1 Prüfung auf Daten-Unabhängigkeit

Vorsicht ist geboten, wenn die Daten nicht unabhängig sind. Dies ist der Fall, wenn z.B. Trends oder Autokorrelationen vorliegen. Weisen die Daten Autokorrelationen auf, sind sie nicht mehr unabhängig! Die Anzahl der Freiheitsgrade reduziert sich. Dies muss bei der Berechnung der Wahrscheinlichkeit berücksichtigt werden.

4.2 Octave-Beispiele

```
In [12]: x=linspace(-6,6,121);
         n=3 # Freiheitsgrade
         subplot(1,2,1);
         plot(x,normpdf(x,0,1),'g-');
         hold;
         plot(x,tpdf(x,n),'b-');
         subplot(1,2,2);
         plot(x,normcdf(x,0,1),'g-');
         hold;
         plot(x,tcdf(x,n),'b-');
```




```
Out[12]: n =  3
```

Quellen

- Bendat und Piersol, Random Data, Wiley 1986
- Schönwiese, Praktische Statistik, 2013