062_Folien

November 30, 2018

```
In [1]: import numpy as np                       # mathematical methods
        from scipy import stats                  # statistical methods
        from matplotlib import pyplot as plt     # plotting methods
        %matplotlib inline
```

1 Wahrscheinlichkeitstheorie

1.0.1 Zufallsvariable und Wahrscheinlichkeitsraum

1.0.2 Erwartungswert und Varianz

1.0.3 Diskrete Zufallsvariablen und Wahrscheinlichkeitsverteilungen

1.0.4 Kontinuierliche Zufallsvariable und Wahrscheinlichkeitsverteilungen

Zusammengesetzte kontinuierliche Verteilungen

1.0.5 Sätze der Statistik

1.1 (Wiederholung) Summe mehrerer *i.i.d.* Zufallsvariablen

Die Summe S_n ist eine Zufallsvariable

$$S_n = \sum_{i=1}^{n} X_i$$

1.1.1 Erwartungswert

$$\mathcal{E}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathcal{E}(X_i) = \sum_{i=1}^{n} \mu = n \cdot \mu$$

1.1.2 Varianz

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i) = \sum_{i=1}^{n} \sigma^2 = n \cdot \sigma^2$$

1.2 (Wiederholung) Mittelwert mehrerer *i.i.d.* Zufallsvariablen

Das arithemetische Mittel oder der durchschnittliche Wert von X nach n Versuchen ist eine Zufallsvariable

$$\overline{X_n} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

1.2.1 Erwartungswert

$$\mathcal{E}(\overline{X_n}) = \mu$$

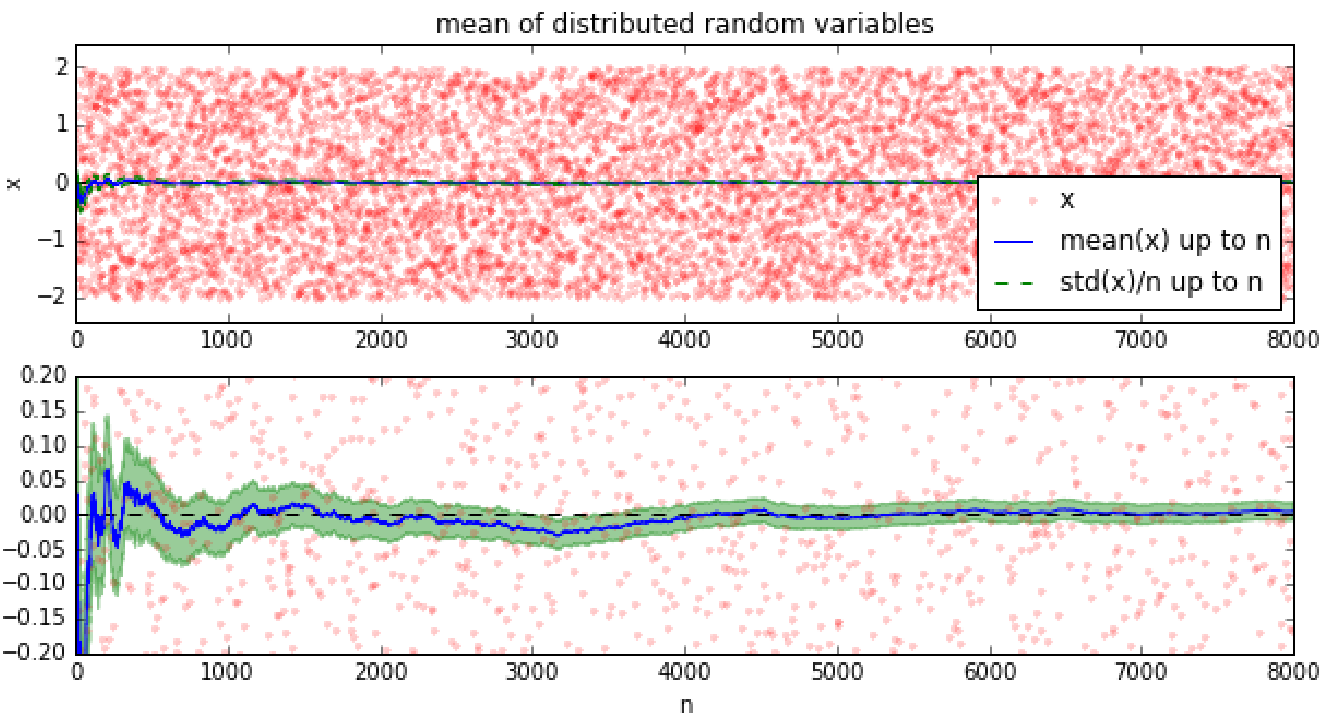
Varianz

$$\mathrm{Var}(\overline{X_n}) = \frac{1}{n}\sigma^2$$

```
In [42]:
```

```
After n=   5 samples mean is -0.20689 and variance/n is  0.18172
After n=  10 samples mean is -0.28676 and variance/n is  0.09550
After n= 100 samples mean is -0.04933 and variance/n is  0.01145
After n=1000 samples mean is -0.01287 and variance/n is  0.00135
After n=8000 samples mean is  0.00444 and variance/n is  0.00016
```

```
<matplotlib.figure.Figure at 0x7f63c44f1860>
```



2 Gesetz der großen Zahlen

Das arithmetische Mittel $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$ konvergiert *nach Wahrscheinlichkeit* gegen den Erwartungswert $\mathcal{E}(\overline{X_n}) = \mathcal{E}(X) = \mu$:
Für eine beliebig kleine Konstante $c > 0$ gilt

$$P(|\overline{X}_n - \mu| \leq c) \xrightarrow{n\to\infty} 1$$

Beweis Mit Hilfe der **Ungleichung von Tschebyscheff**

$$P(|\tilde{X} - \tilde{\mu}| < c) \geq 1 - \frac{\tilde{\sigma}^2}{c^2}$$

und $\tilde{X} = \overline{X}_n$ sowie $\tilde{\sigma}^2 = \frac{\sigma^2}{n} \xrightarrow{n\to\infty} 0$

3 Theorem von Bernoulli

Sei X eine diskrete Zufallsvariable mit möglichen Ereignissen x_j mit Wahrscheinlichkeit $p_j = p(X = x_j)$.
Dann gilt für die relative Häufigkeit h_j, mit der das Ereignis x_j eintritt:

$$h_j \xrightarrow{n\to\infty} p(X = x_j)$$

```
In [3]: '''Bernoulli theoreme for discrete random variables
           one die, N repetitions, only "6" counts as a win'''
        np.random.seed(98765)
```

```
Ns = (100, 1000, 10000, 100000)
for N in Ns:
    x = [np.random.choice([1, 2, 3, 4, 5, 6]) for _ in range(N)]
    sixs = x.count(6)
    print('in {:6d} throws, "six" appeared {:6d} times = {:8.3f}% = (16.666{:+6.3f})%'
          .format(N, sixs, 100*sixs/N, 100*(sixs/N-1/6)))
```

```
in    100 throws, "six" appeared     18 times =   18.000% = (16.666+1.333)%
in   1000 throws, "six" appeared    154 times =   15.400% = (16.666-1.267)%
in  10000 throws, "six" appeared   1703 times =   17.030% = (16.666+0.363)%
in 100000 throws, "six" appeared  16824 times =   16.824% = (16.666+0.157)%
```

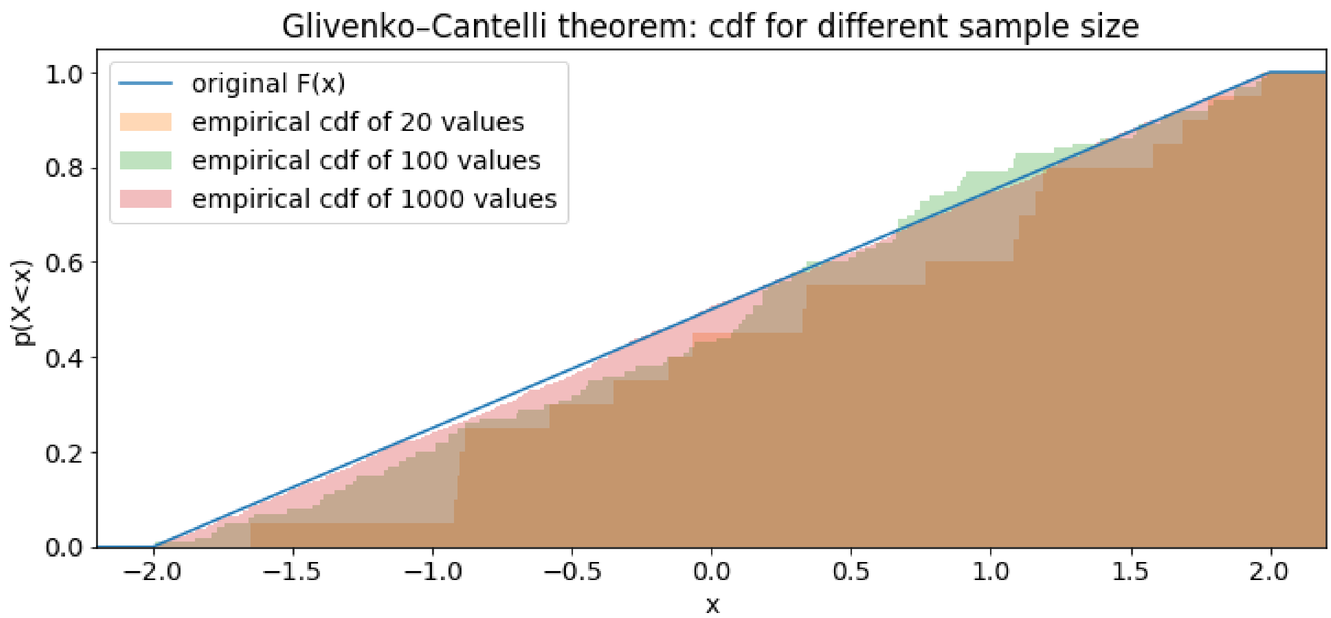
4 Hauptsatz der Statistik, Satz von Gliwenko-Cantelli

Sei X eine Zufallsvariable mit Verteilungsfunktion $F(x)$ und $\{X_i\}$ mit $i \in \{1 \ldots N\}$ *i.i.d.* Wiederholungen.
Dann konvergiert die relative Häufigkeit $F_n(x)$, daß $X_i \leq x$ gegen $F(x)$ nach Wahrscheinlichkeit:

$$P(sup|F_n(x) - F(x)| \leq c) \xrightarrow{n\to\infty} 1$$

```
In [5]: '''Glivenko-Cantelli theorem: cdf for different sample size'''
        np.random.seed(98765)
        f = plt.figure(figsize=(12,5))
        x = np.linspace(-3, 3, 601)
        distrib = stats.uniform(loc=-2, scale=4.)
        F = distrib.cdf(x)
        plt.plot(x, F, label='original F(x)')
        bins = np.linspace(-3, 3.01, 602)
        for N in (20, 100, 1000):
            Xi = distrib.rvs(size=N)
            h = plt.hist(Xi, bins=bins, cumulative=True, density=True, edgecolor="none",
                         alpha=.3, label='empirical cdf of {} values'.format(N));
            sup = np.abs(F-h[0]).max()
            print('for N={:4d} sup(F-Fn)={:9.5f}'.format(N, sup))
        plt.axis((-2.2, 2.2, 0, 1.05))
        plt.legend(loc='upper left')
        plt.xlabel('x')
        plt.ylabel('p(X<x)')
        plt.title('Glivenko-Cantelli theorem: cdf for different sample size');
```

```
for N=  20 sup(F-Fn)=  0.21750
for N= 100 sup(F-Fn)=  0.07500
for N=1000 sup(F-Fn)=  0.01550
```



4.0.1 Mittelwert

Sei X eine Zufallsvariable mit Erwartungswert μ und Varianz σ^2.
Die Zufallsvariable *Mittelwert* mehrere *i.i.d* X_i

$$M_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

hat den Erwartungswert

$$\mu_M = \mathcal{E}(M_n) = \mu$$

und die Varianz

$$\sigma_M^2 = \mathrm{Var}(M_n) = \frac{1}{n}\cdot\sigma^2$$

Standardisieren ergibt eine neue Zufallsvariable Z_n

$$Z_n = \frac{M_n - \mu_M}{\sigma_M} = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{n}{n}\mu}{\frac{1}{\sqrt{n}}\sigma} = \frac{\sum_{i=1}^{n}(X_i-\mu)}{\sqrt{n}\cdot\sigma}$$

5 Zentraler Grenzwertsatz

Seien X_i unabhängig identisch verteilte Zufallsvariablen mit Erwartungswert

$$\mathcal{E}(X_i) = \mu$$

und Varianz

$$\mathrm{Var}(X_i) = \sigma^2$$

Dann konvergiert die Verteilungsfunktion

$$F_n(z) = P(Z_n \leq z)$$

der standardisierten Summe

$$Z_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{X_i-\mu}{\sigma}$$

für $n \to \infty$ an jeder Stelle $z \in \mathbb{R}$ gegen die Verteilungsfunktion der **Standardnormalverteilung**

$$F_n(z) \xrightarrow{n\to\infty} \Phi(z)$$

Kurz:

$$Z_n \overset{a}{\sim} \mathcal{N}(0,1)$$

5.1 Zentraler Grenzwertsatz: Beispiel Binomialverteilung

Binomialverteilte Zufallsvariable $X \sim \mathcal{B}(N, \pi)$ mit $N = 10$ und $\pi = \frac{1}{2}$ #### Kennwerte

$$\mathcal{E}(X) = \mu_B = N \cdot \pi \quad = 5$$

$$\text{Var}(X) = \sigma_B^2 = N \cdot \pi \cdot (1 - \pi) = 2,5$$

$$Std(X) = \sqrt{\text{Var}(X)} = \sqrt{\frac{5}{2}} \approx 1,6$$

Ein Versuch $n = 100$ malige Durchführung des Zufallsexperiments X_i mit $i \in \{1 \ldots n\}$ ergibt den Mittelwert

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Erwartungswert von $\overline{X}_n$

$$\mu_n = \mathcal{E}(\overline{X}_n) = \frac{n}{n} \mathcal{E}(X) = \mu_B = N \cdot \pi = 10 \cdot \frac{1}{2} = 5$$

Varianz von $\overline{X}_n$

$$\sigma_n^2 = \text{Var}(\overline{X}_n) = \frac{n \cdot \text{Var}(X)}{n^2} = \frac{\sigma^2}{n} = \frac{N \cdot \pi \cdot (1 - \pi)}{n} = 0.025$$

Standardabweichung

$$\sigma_n = \sqrt{\text{Var}(\overline{X}_n)} \approx 0,16$$

Standardisieren

$$Z_n = \frac{\overline{X}_n - \mu_n}{\sigma_n} = \sqrt{n} \frac{\overline{X}_n - \mu_B}{\sigma_B}$$

Mehrfache Durchführung *rep*-malige Wiederholung dieses Versuchs ergibt eine (empirische) Verteilung der Z_n

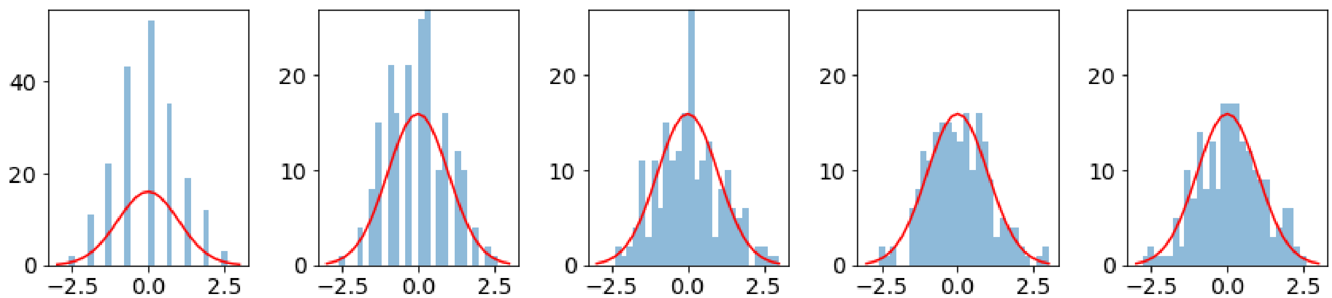
Zentraler Grenzwertsatz für die Verteilung (aus der Theorie) der Z_n

$$Z_n \sim \mathcal{N}(0, 1)$$

5.2 Zentraler Grenzwertsatz: Beispiel Binomialverteilung

In [6]:
```python
'''central limit theoreme - binomial distribution'''
nplots = 5
ns = [4**i for i in range(nplots)]          # size of sums = {1, 4, 16, 64, 256}
np.random.seed(9876543)
rep = 200                                   # number of repetitions to get x-sum-distribution
binoN, binoPi = (10, 0.5)                   # N and pi of binomial distributions
# ... resulting in characteristics mean and var
binoMu, binoVar = (binoN*binoPi, binoN*binoPi*(1-binoPi))
bins = np.linspace(-3.0, 3.0, 31)           # bins for histogram plotting
factor = rep*(bins[-1]-bins[0])/(bins.shape[0]-1)    #
f = plt.figure(figsize=(12,3))
distrib = stats.binom(binoN, binoPi)        # Starter: of distribution binomial(N=10, pi=0.5)
for i, n in enumerate(ns):
    X = distrib.rvs(size=(rep, n))          # draw n times rep values of distribution
    means = X.mean(axis=1)                  # the rep=200 means of Xi with i=1..n
    Z = np.sqrt(n) * (means-binoMu) / np.sqrt(binoVar)  # standardize distribution
    f.add_subplot(1,nplots,i+1)
    plt.hist(Z, bins=bins, edgecolor="none", alpha=.5)
    if i>0:
        plt.ylim(0, 27)
    plt.plot(bins, factor*stats.norm.pdf(bins), 'r-');
    print('X_{:3d} has mean={:.3f} and var={:.5f}'.format(n, means.mean(), means.var()))
plt.tight_layout()
print('standardized probability distribution of random variable X:')
```

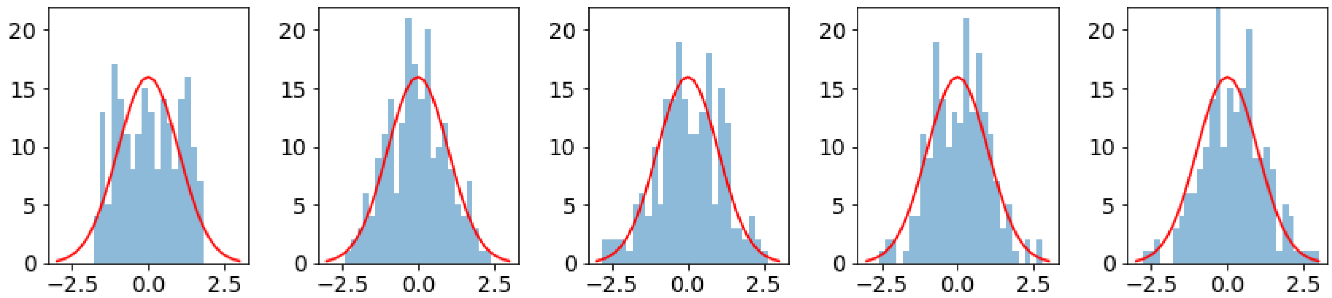
```
X_  1 has mean=4.965 and var=2.64377
X_  4 has mean=5.008 and var=0.62057
X_ 16 has mean=5.021 and var=0.18809
X_ 64 has mean=5.004 and var=0.04315
X_256 has mean=5.012 and var=0.01170
standardized probability distribution of random variable X:
```



5.3 Zentraler Grenzwertsatz: Beispiel Gleichverteilung

In [7]:
```
'''central limit theoreme - uniform distribution'''
nplots = 5
ns = [4**i for i in range(nplots)]         # size of sums = {1, 4, 16, 64, 256}
np.random.seed(98765432)
rep = 200                           # number of repetitions to get x-sum-distribution
unia, unib = (-1.5, 1.5)            # a..b of uniform distributions
bins = np.linspace(-3.0, 3.0, 31)# bins for histogram plotting
factor = rep*(bins[-1]-bins[0])/(bins.shape[0]-1)
f = plt.figure(figsize=(12,3))
distrib = stats.uniform(loc=unia, scale=unib-unia) # distribution uniform a..b
# ... results in characteristic mean and var
uniMu, uniVar = (distrib.expect(), distrib.var())
for i, n in enumerate(ns):
    X = distrib.rvs(size=(rep, n))                # draw n times rep values of distribution
    means = X.mean(axis=1)                        # the rep=200 means of Xi with i=1..n
    Z = np.sqrt(n)*(means-uniMu)/np.sqrt(uniVar)# standardize distribution
    f.add_subplot(1,nplots,i+1)
    plt.hist(Z, bins=bins, edgecolor="none", alpha=.5)
    plt.ylim(0, 22)
    plt.plot(bins, factor*stats.norm.pdf(bins), 'r-');
    print('X_{:3d} has mean={:6.3f} and var={:.5f}'
                                      .format(n, means.mean(), means.var()))
plt.tight_layout()
print('standardized probability distribution of random variable X:')
```

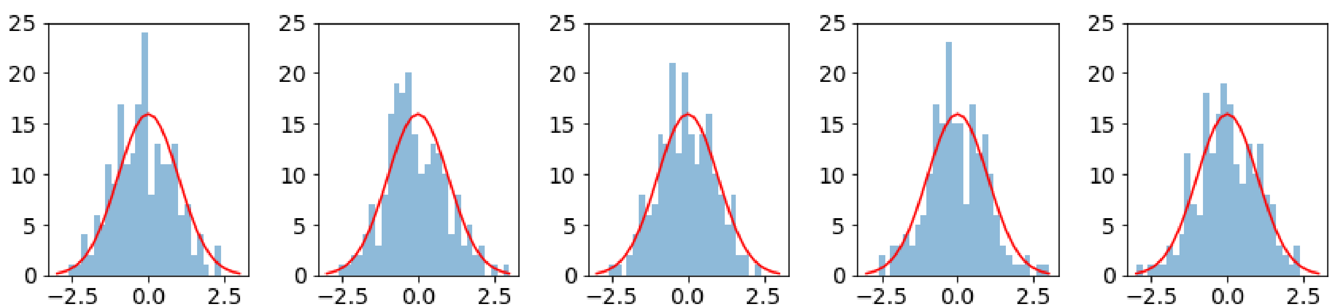
```
X_  1 has mean= 0.037 and var=0.72586
X_  4 has mean=-0.021 and var=0.17273
X_ 16 has mean= 0.006 and var=0.05273
X_ 64 has mean= 0.008 and var=0.01135
X_256 has mean= 0.009 and var=0.00296
standardized probability distribution of random variable X:
```



5.4 Zentraler Grenzwertsatz: Beispiel Normalverteilung

In [8]:
```
'''central limit theoreme - Normal distribution'''
nplots = 5
ns = [4**i for i in range(nplots)]     # size of sums = {1, 4, 16, 64, 256}
np.random.seed(9876543)
rep = 200                              # number of repetitions to get x-sum-distribution
mu, sigma = 1.5, 2                     # mu, sigma of a Gaussian normal distributions
bins = np.linspace(-3.0, 3.0, 31)      # bins for histogram plotting
factor = rep*(bins[-1]-bins[0])/(bins.shape[0]-1)
f = plt.figure(figsize=(12,3))
distrib = stats.norm(loc=mu, scale=sigma) # Gaussian distribution (mu, sigma)
for i, n in enumerate(ns):
    X = distrib.rvs(size=(rep, n))     # draw n times rep values of distribution
    means = X.mean(axis=1)             # the rep=200 means of Xi with i=1..n
    Z = np.sqrt(n)*(means-mu) / sigma # standardize distribution
    f.add_subplot(1,nplots,i+1)
    plt.hist(Z, bins=bins, edgecolor="none", alpha=.5)
    plt.ylim(0, 25)
    plt.plot(bins, factor*stats.norm.pdf(bins), 'r-');
    print('X_{:3d} has mean={:.3f} and var={:.5f}'
                                    .format(n, means.mean(), means.var()))
plt.tight_layout()
print('standardized probability distribution of random variable X:')
```

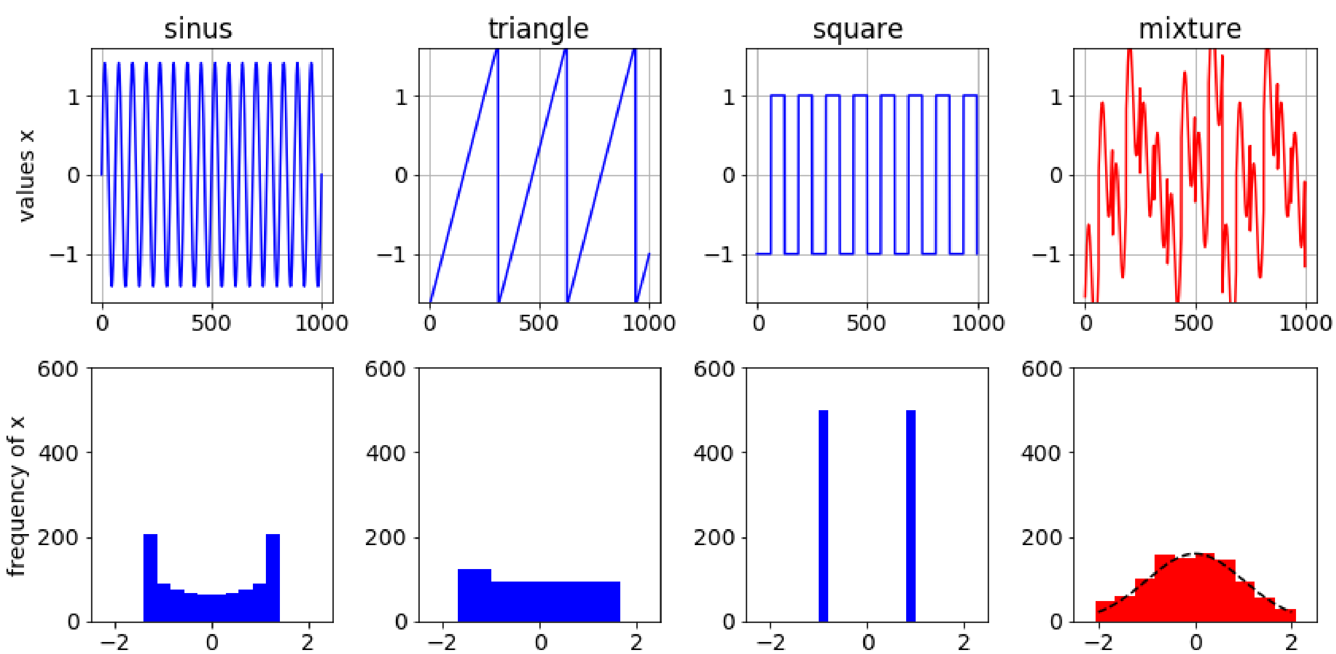
```
X_  1 has mean=1.206 and var=3.96231
X_  4 has mean=1.452 and var=1.05583
X_ 16 has mean=1.468 and var=0.20611
X_ 64 has mean=1.486 and var=0.06509
X_256 has mean=1.501 and var=0.01527
standardized probability distribution of random variable X:
```



5.5 Zentraler Grenzwertsatz für beliebige Mischungen

In [9]:
```
'''central limit theoreme - even for non i.i.d. mixtures - but many'''
x = np.linspace(0, 16, 1000)
f = plt.figure(figsize=(12, 6))
sinus = 1.41*np.sin(2.*np.pi*x)                      # sine wave distribution
triangle=(x%5-2.5)/1.5                               # sawtooth distribution
square = 2*((x%2)>1.0)-1.                            # rectangular distribution
mixture=1./np.sqrt(3)*(sinus+triangle+square)        # normalize
cols = ['b', 'b', 'b', 'r']                          # colors for plotting
names = ['sinus   ', 'triangle', 'square  ', 'mixture ']
for i, curve in enumerate([sinus, triangle, square, mixture]):
    f.add_subplot(2, 4, 1+i)                         # i_th plot upper row
    plt.title(names[i])
    plt.grid(True)
    plt.plot(curve, cols[i])                         # plot of i-th curve
    plt.ylim(-1.6, 1.6)
    if i==0:
        plt.ylabel('values x')
    f.add_subplot(2, 4, 5+i)                         # i_th plot lower row
    #plt.title(names[i]+' histogram')
    plt.axis((-2.5, 2.5, 0, 600))
    plt.hist(curve, color=cols[i])                   # histogram of i-th curve
    print('{} has mean={:6.3f} and std={:6.3f}'
                          .format(names[i], curve.mean(), curve.std()))
    if i==0:
        plt.ylabel('frequency of x')
x = np.linspace(-2, 2, 401)
# and for comparison of mixture: add standard normal to last plot
plt.plot(x, 400*stats.norm.pdf(x), 'k--')
plt.tight_layout();
```

```
sinus    has mean= 0.000 and std= 0.997
triangle has mean=-0.083 and std= 0.988
square   has mean=-0.002 and std= 1.000
mixture  has mean=-0.049 and std= 0.915
```



5.5.1 Zentraler Grenzwertsatz gilt näherungsweise auch für

- unabhängig summierte Ursachen
  - zwar non i.i.d.
  - aber gleiche Größenordnung der Streuung σ_i

Beispiele: - Körpergröße beeinflußt durch Genetik: mehrer Wachstumsschübe - Rauschen in einer Anlage durch viele Komponenten - Meßfehler, die mehrere unabhängige Ursachen haben - Verhalten von Populationen

6 Zusammenfassung

- Zufallsvariable mit stetiger Wahrscheinlichkeitsverteilung
  - $x \in \mathbb{R}$
- Wahrscheinlichkeitsdichte $f(x)$
  - Punktwahrscheinlichkeit $\to 0$

$$P(a \le x \le b) = \int_a^b f(x)\,\mathrm{d}x = F(b) - F(a) \ \le 1$$

  - subjektiv "ars conjectandi" (Theorie, Interpretation)
  - Normierung 1
  - Verteilungsfunktion $F(x)$
- Kennzahlen
  - Erwartungswert

$$\mathcal{E}(X) = \mu = \int_{-\infty}^{\infty} x \cdot f(x)\,\mathrm{d}x$$

  - Varianz

$$\mathrm{Var}(X) = \sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 \cdot f(x)\,\mathrm{d}x = \mathcal{E}((X-\mu)^2)$$

  - Schiefe und Kurtosis
  - Kovarianz

7 ...

- Stetige Verteilungen
  - Normalverteilung
  - Standardnormalverteilung

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

  - Exponentialverteilung

$$f(x) = \lambda e^{-\lambda x} \qquad x \ge 0$$

  - Pareto-Verteilung
  - Cauchy/Lorentz-Verteilung

$$Y = \frac{X_1}{X_2} \ \to \ Y \sim \mathcal{Cauchy}$$

  - Quadrat-Verteilung X^2
- Zusammengesetzte Verteilungen
  - Summe von Zufallsvariablen
  - Mittelwert von Zufallsvariablen

8 ...

- Satz von Bernoulli
$$h_j \to p(X = x_j)$$
- Hauptsatz der Statistik
$$F_n(x) \to F(x)$$
- Gesetz der großen Zahlen
$$\overline{X}_n \to \mu$$
- Zentraler Grenzwertsatz
$$F_n(z) \to \Phi(z)$$

9 Ausblick

- χ^2 -Verteilung
- Student-t -Verteilung

10 Fragen?