071_Folien

December 8, 2018

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

0.0.1 Beschreibende Statistik

0.0.2 Wahrscheinlichkeitstheorie

- Zufallsvariable und keitsverteilungen
- i.i.d
- Sätze der Statistik

0.0.3 Schließende Statistik

Punktschätzungen

- Stochastik: Statistik und Wahrscheinlichkeitstheorie
- Punktschätzer
  - Arithmetisches Mittel
  - Stichproben-Varianz
- Max-Likelihood-Prinzip
- Robuste Schätzung

Intervallschätzungen

0.0.4 Beschreibende Statistik

Charakteristische Kennzahlen von Daten

- Mittelwert, Median
- Varianz, Standardabweichung, Quantile ...
- Form

0.0.5 Wahrscheinlichkeitstheorie

Charakteristische Kennzahlen einer Wahrscheinlichkeits(dichte)-Verteilung

- Erwartungswert
- Varianz, Standardabweichung, Quantile ...
- Form der Verteilung
  - Parameter λ, μ, σ

0.1 Fragestellung:

0.1.1 Grundgesamtheit, wenn nur Stichprobe?

0.1.2 Parameter einer Verteilung, wenn nur Stichprobe?

0.2 Wahrscheinlichkeitstheorie

Zufallsvariable X

Charakteristische Parameter von Modellverteilungsfunktionen μ, σ, N, π, λ...
Daraus berechenbar:

- Erwartungswert $\mathcal{E}(X)$, Median, ...
- Varianz $Var(X)$, Standardabweichung, ...
- Wahrscheinlichkeiten für Bereiche ...
- ...

0.3 Wiederholung

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

$$F_n(z) \to \Phi(z)$$

$n \to \infty$! *Frequentistische Statistik*

1 Schließende Statistik

"*ars conjectandi*" Kombiniert empirische Daten mit Wahrscheinlichkeitstheorie

1.0.1 Stichprobe ⇒ Schlussfolgerung auf Grundgesamtheit

2 Fragen der Schließenden Statistik

- Welche Verteilung hat die Grundgesamtheit?
  - ⇒ Theorie, Ockhams Rasiermesser, Vergleich
- Welcher Parameterwert paßt am besten zu den Beobachtungen?
  - ⇒ Schätzungen
- Sind die Beobachtungen mit einem angenommenen Parameter vereinbar?
  - ⇒ Testen einer Nullhypothese
- Welche Parameterwerte sind mit den Beobachtungen vereinbar?
  - ⇒ Vertrauensintervall
- Wie kamen die Beobachtungen zustande?
  - ⇒ Versuchsplanung

3 Beispiel:

3.1 Grenzwertüberschreitung bei Asbestfasern

- Grenzwert liegt bei 1000 Fasern/m^3.
- Teure Messung $3\times$ durchführen mit jeweils $5l$ Raumluft
- Ergebnis: $x = (6, 4, 9)$
  - entspräche $(1200, 800, 1800)$ Fasern/m^3.

3.2 Welche Verteilung hat die Grundgesamtheit?

3.2.1 Modell

Poisson-Verteilung $\mathcal{P}(\lambda)$ für $x \in \{0, 1, 2, \dots\}$

$$P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

- Erwartungswert $\mathcal{E}(X) = \lambda$
- Varianz $\mathrm{Var}(X) = \lambda$

Grenzwert In $5l = \frac{1}{200}\ \mathrm{m}^3$ erwarten wir $\frac{1000}{200} = 5$ Fasern.
Die zum gerade noch erlaubten Grenzwert passende Verteilung der $5l$-Proben wäre daher

$$\mathcal{P}(\lambda{=}5): \quad P(x) = \frac{5^x}{x!} e^{-5}$$

mit in einem Kubikmeter erwarteten

$$\mathcal{E}\Big(\sum_{i=1}^{200} x_i\Big) = 200 \cdot \lambda = 200 \cdot 5 = 1000$$

Fasern

3.3 Welcher Parameterwert paßt am besten zu den Beobachtungen?

Die Stichprobe $(6, 4, 9)$ würde zu

$$\lambda = \mathcal{E}(X) \overset{n\to\infty}{\Leftarrow} \overline{x} = \frac{1}{3}\sum_{i=1}^{3} x_i = \frac{6+4+9}{3} = 6.33$$

am besten passen
$\Rightarrow$ Grenzwert überschritten. Sanierung ! ?

3.3.1 Ergebnis: *überschritten*

3.3.2 Ok?

Ist es möglich, die gleichen Stichprobenwerte auch mit dem (gerade noch unbedenklichen) $\lambda = 5$ zu erhalten?
Wie genau ist das "wahre" λ durch diese drei Messungen festgelegt?

```
In [3]: '''measured sample of asbestos fibres'''
        xsample = [6, 4, 9]                          # (random) sample
        xmean = np.asarray(xsample).mean()      # measured mean thereof
        print('expected fibres in 5l; within one standard deviation: {:.3f} .. {:.3f}'
              .format(xmean-np.sqrt(xmean), xmean+np.sqrt(xmean)))

expected fibres in 5l; within one standard deviation: 3.817 .. 8.850
```

---

4 Statistik - Modell Zufallsvariable

Aus der Grundgesamtheit werden n Werte gemessen.

- Stichprobe besteht aus Zufallsvariablen $\{X_1, \dots, X_n\}$
- endlicher Erwartungswert, endliche Varianz
- *unabhängige* Messungen
- aus ein und derselben Grundgesamtheit, *identische Wiederholung*
- Stochastisches Modell

4.1 Schätzungen sind Zufallsvariable

Punktschätzer (Intervallschätzer, siehe später)

für Kennzahlen

- Erwartungswert
- Varianz
- Korrelation
- ...

oder für Parameter

- λ einer Poissonverteilung
- μ und σ^2 einer Normalverteilung
- π bei Binomialverteilung
- ...

4.2 Bekannte Punktschätzer

$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ für den Erwartungswert μ, $\mathcal{E}(X) = \mu$ von $\mathcal{N}(\mu, \sigma^2)$
$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ für die Eintrittswahrscheinlichkeit π, $\mathcal{E}(X) = \pi$ eines Bernoulli-Experiments $\mathcal{B}(\pi)$
$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ für die Stichprobenvarianz $\mathrm{Var}(X) = \sigma^2$
$\tilde{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ für die empirische Varianz $\mathrm{Var}(X) = \sigma^2$

4.3 Allgemeine Schätzfunktion

Schätzfunktion $T_\theta = g(X_1, \ldots, X_n)$ mit Schätzwert $t = g(x_1, \ldots, x_n)$ für den Parameter θ der Grundgesamtheit.

Realisierung Realisierung der Stichprobe $\{x_1, \ldots, x_n\}$ bestimmt den Wert des Schätzers.

Schätzstatistik Wissen um die Verteilung der Zufallsvariable *Schätzer*.

4.4 Eigenschaften von Schätzstatistiken

4.4.1 Erwartungstreue

Erwartungswert der Schätzstatistik = Erwartungswert der Grundgesamtheit - weder über- noch unterschätzen - $\mathcal{E}_\theta(T) = \theta$

Restfehler *Bias* $Bias_\theta(T) = \mathcal{E}_\theta(T) - \theta$

- idealerweise $Bias = 0$
- zumindest klein

4.4.2 Beispiel Stichprobenmittel

Das *Stichprobenmittel*

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

ist erwartungstreu für den Erwartungswert $\mathcal{E}(X) = \mu$ der Grundgesamtheit.
$\hat{\mu}$ ist ein erwartungstreuer Schätzer:

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

Beweis $\mathcal{E}(\hat{\mu}) = \mathcal{E}(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{E}(X) = \mu$

4.4.3 Beispiel Stichprobenvarianz

Die *Stichprobenvarianz*

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

ist erwartungstreu für die Varianz $\text{Var}(X) = \sigma^2$ der Grundgesamtheit.

$\hat{\sigma}^2$ ist ein erwartungstreuer Schätzer:

$$\hat{\sigma}^2 = \frac{1}{n-1}(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2)$$

Beweis Sei o.b.d.A $\mathcal{E}(X) = 0$, dann folgt mit $\mathcal{E}(\overline{X}^2) = \text{Var}(\overline{X}) = \frac{\sigma^2}{n}$:

$$\mathcal{E}(\hat{\sigma}^2) = \mathcal{E}\Big(\frac{1}{n-1}[\sum_{i=1}^{n} X_i^2 - n\overline{X}^2]\Big)$$

$$= \frac{1}{n-1}(\sum_{i=1}^{n}\mathcal{E}(X_i^2) - n\mathcal{E}(\overline{X}^2)) = \frac{1}{n-1}(n\sigma^2 - n\frac{1}{n}\sigma^2) = \sigma^2 = \text{Var}(X) = S^2$$

Bemerkung: **Hier** liegt der tiefere Grund für den Faktor $\frac{1}{N-1}$ der (empirischen) Stichproben-Varianz.

4.4.4 Ergebnis: Eigenschaften Stichprobenvarianz

- Die Stichproben-Varianz S^2 ist für alle Verteilungen ein erwartungstreuer Schätzer der Varianz σ^2
- Die Verteilung von $\hat{\sigma}^2$ hängt von der Verteilung von X_i ab
- (Wenn X normalverteilt ist, dann hat $(n-1)\hat{\sigma}^2/\sigma^2$ eine χ^2-Verteilung mit $m = n-1$ Freiheitsgraden)

4.4.5 Gegenbeispiel Empirische Varianz

Für die empirische Varianz $\tilde{S^2}$ gilt

$$\mathcal{E}(\tilde{S^2}) = \frac{n-1}{n}\sigma^2$$

- es bleibt eine systematische Messabweichung *Bias*

$$\text{Bias}_{\sigma^2}(\tilde{S}^2) = \mathcal{E}_{\sigma^2}(\tilde{S}^2) - \sigma^2 = -\frac{1}{n}\sigma^2$$

- womit die Varianz tendentiell unterschätzt wird
- aber sie ist *asymptotisch erwartungstreu* für $n \to \infty$

---

5 Schätzfunktion

Woher bekommen wir eine Schätzfunktion T für einen Parameter θ ? **Bisher:** Plausibel und nachgewiesen: - Arithmetischer Mittelwert für Erwartungswert $\mathcal{E}(X)$ - Stichprobenvarianz für $\text{Var}(X)$

Die Wahrscheinlichkeitsverteilung $F(x)$ bzw. Wahrscheinlichkeitsdichte $f(x)$ hänge von Parameter(n) θ ab.

Beispiel Bernoulli-Experiment:

$$f(x|\pi) = P(X{=}x|\pi) = \pi^x(1-\pi)^{1-x} \quad \text{für } x \in \{0,1\}$$

Beispiel Normalverteilung:

$$f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{für } x \in \mathbb{R}$$

5.1 entweder 1. Methode der kleinsten Quadrate

Anfitten der parametrischen Verteilung an die Daten der Stichprobe durch Minimieren der Fehlerquadratsumme (*sum of squared residua*): $\text{SSR} = \sum(x_i - \mu)^2$

$$\underset{\mu}{\text{argmin}} \sum(x_i - \mu)^2$$

5.2 oder 2. Satz von Bayes

- Gegeben: Modellverteilung $f(x)$
- mit Parameter θ
- Satz von Bayes

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{f(x)}$$

- Gesuchter Parameter θ:

$$\underset{\theta}{\operatorname{argmax}}\big(f(\theta|x)\big)$$

Siehe *Angewandte Statistik II*

5.3 oder 3. Maximum-Likelihood-Prinzip

Likelihood-Funktion

$$L(\theta) = f(x|\theta)$$

5.3.1 Bekannte Verteilung, unbekannter Parameter, realisierte Stichprobe

Ist der Wert der Stichprobe x gegeben (das Zufallsexperiment X also durchgeführt), dann ist

$$L(\theta) = f(x|\theta)$$

eine Funktion von θ

6 Maximum Likelihood Prinzip

Der Schätzer $\widehat{\theta}$ zum realisierten Messwert x ergibt sich aus der Maximierung der Likelihood-Funktion $L(\theta)$:

$$L(\widehat{\theta}) = \max_\theta L(\theta)$$

6.1 Ein Beispiel: geometrische Verteilung:

$P(X{=}x) = p_\pi(x) = (1-\pi)^{x-1} \cdot \pi^1$

Experiment: Messung $x = 3$

Frage: Welche Verteilung p_π bzw. welcher Parameter π paßt am besten zur Messung?

Likelihood: Betrachte $p_\pi(x)$ als Funktion von π

$$L(\pi) = p_\pi(x) = (1-\pi)^{x-1} \cdot \pi$$

Likelihood: $L(\pi) = p_\pi(x) = (1-\pi)^{x-1} \cdot \pi$

Maximieren Notwendige Bedingung: Ableitung

$$\frac{\partial L(\pi)}{\partial \pi} = -\pi(x-1)(1-\pi)^{x-2} + 1 \cdot (1-\pi)^{x-1} = (1-\pi)^{x-2}[1-\pi \cdot x]$$

Nullsetzen

$$\frac{\partial L(\pi)}{\partial \pi} = (1-\pi)^{x-2}[1-\pi \cdot x] = 0$$

Auflösen nach π maximiert $L(\pi)$ bzw. $p_\pi(x)$ für $\widehat{\pi}$

$$\widehat{\pi} = \frac{1}{x} = \frac{1}{3}$$

Zweite Ableitung kleiner Null?
Randwerte?

Ergebnis: Für das Beispiel, nach 3 Stunden den ersten Rechnerabsturz gesehen zu haben, erhält man als plausibelsten *Likelihood*-Parameter für die geometrische Verteilung den Wert $\widehat{\pi} = \frac{1}{3}$

Der Erwartungswert für $X \sim \mathcal{G}\rceil\wr\Updownarrow(\frac{1}{3}) = \ldots$

```
In [4]: pi = 1./3
        print('expectation value of geom({:.5f}) is {:.5f}'.format(pi, stats.geom(pi).expect()))
```

```
expectation value of geom(0.33333) is 3.00000
```

6.2 Bisher: Stichprobe *ein* Wert

6.3 Jetzt: *mehrere* i.i.d. Zufallsvariable

Experiment X wird n-mal durchgeführt: $x_1, x_2, \ldots x_n$

Bei gegebenem θ sind die x_i gemäß $f(x)$ verteilt, die Verbund-Wahrscheinlichkeit diese n Werte zu erhalten beträgt also:

$$f(x_1, x_2, \ldots x_n|\theta) = f(x_1|\theta) \cdot f(x_2|\theta) \cdots f(x_n|\theta)$$

6.4 Likelihood-Funktion

6.4.1 Bekannte Verteilung, unbekannter Parameter, realisierte Stichprobe

Sind die Werte x_i gegeben (das Experiment X also n-mal *i.i.d.* durchgeführt), dann ist die Likelihood-Funktion

$$L(\theta) = f(x_1, x_2, \ldots x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

6.5 Log-Likelihood-Funktion

Mit Hilfe der streng monotonen Logarithmus-Funktion ergibt sich aus der Likelihood-Funktion die *Log-Likelihood-Funktion* $l = log(L)$

$$l(\theta) = \log\, L(\theta) = \log \prod_{i=1}^{n} f(x_i|\theta) = \sum_{i=1}^{n} \log\, f(x_i|\theta)$$

die sich leichter optimieren läßt und trotzdem das selbe Maximum für θ liefert.

6.6 Maximum-Log-Likelihood-Prinzip

Die Zufallsvariable X habe die Wahrscheinlichkeitsverteilung $f(x|\theta)$ mit unbekanntem zu bestimmenden Parameter θ.

Der Schätzer $\widehat{\theta}$ zu n *i.i.d.* Messwerten x_i maximiert die Log-Likelihood-Funktion $l(\theta) = \sum_{i=1}^{n} \log\, f(x_i|\theta)$:

$$l(\widehat{\theta}) = \max_\theta\, l(\theta)$$

Ableitung vereinfacht sich mit

$$\Psi(x, \theta) := \frac{\partial}{\partial\theta} \log\; f(x|\theta)$$

zu

$$\frac{\partial l(x_i, \theta)}{\partial\theta} = \sum_i \Psi(x_i, \theta)$$

[ÜA] Schätzer für μ und σ der Normalverteilung

6.7 Maximum-Log-Likelihood-Prinzip

6.7.1 Anmerkungen

- Log-Likelihood-Prinzip ist allgemein anwendbar
- kann oft geschlossen gelöst werden.
- eignet sich **nicht** für Abschätzung der *Verteilung* von θ

6.7.2 Vergleich zu Kleinste-Quadrate

- Ergebnis ist meist dasselbe
- Log-Likelihood benötigt Gesamt-Wahrscheinlichkeitsverteilung
  - Kleinste-Quadrate kommt mit Mittelwert (und Varianz/Kovarianz) aus
- Log-Likelihood kann manchmal nur numerisch simuliert werden

6.8 Zurück zum Beispiel Asbestfasern

Grenzwert $x_G = 5.0$

$n = 3$ Messwerte $x_1 = 6,\ x_2 = 4,\ x_3 = 9$ entstammen einer Poissonverteilung mit unbekanntem Parameter λ.

Die Likelihood-Funktion dafür ist

$$L(\lambda) = f(x_1|\lambda) \cdot f(x_2|\lambda) \cdot f(x_3|\lambda) = e^{-\lambda}\frac{\lambda^6}{6!} \cdot e^{-\lambda}\frac{\lambda^4}{4!} \cdot e^{-\lambda}\frac{\lambda^9}{9!} = e^{-3\lambda}\frac{\lambda^{19}}{6!4!9!}$$

Die Log-Likelihood-Funktion ist

$$l(\lambda) = \ln L(\lambda) = -3\lambda + 19 \ln \lambda - \ln 6!4!9!$$

$$l(\lambda) = \ln L(\lambda) = -3\lambda + 19 \ln \lambda - \ln 6!4!9!$$

Ableitung nullsetzen

$$\frac{\partial\, l(\lambda)}{\partial \lambda} = -3 + \frac{19}{\lambda} = 0$$

ergibt Extremwert für

$$\widehat{\lambda} = \frac{19}{3} = 6.33$$

Zweite Ableitung bestätigt *Maximum*

$$\frac{\partial^2\, l(\lambda)}{\partial \lambda^2} = -\frac{19}{\lambda^2} < 0$$

und damit $\hat{\lambda} = \overline{x}$

6.9 Zwischenergebnis: Poissonverteilung

Der Maximum-Likelihood-Schätzer (*MLE*) für den Parameter λ der Poissonverteilung ist das Stichprobenmittel

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

7 Zusammenfassung Maximum-Log-Likelihood-Prinzip

1. Modellverteilung mit Parameter θ für die Zufallsvariable X

$$f_\theta(X)$$

2. Daraus Likelihood für Parameter θ bei Meßwerten x angeben

$$L_X(\theta) = f_\theta(\mathbf{x}) = f_\theta(x_1) \cdot f_\theta(x_2) \cdots f_\theta(x_n)$$

beziehungsweise Log-Likelihood-Funktion

$$l_X(\theta) = \ln f_\theta(\mathbf{x}) = \sum_{i=1}^{n} \ln f_\theta(x_i)$$

und diese 3. maximieren:

$$\frac{\partial l(\theta)}{\partial \theta} = 0$$

$$\frac{\partial^2 l(\theta)}{\partial \theta^2} < 0$$

keine Randwerte > Maximum

4. Messung der Stichprobe

$$\mathbf{x} = \{x_i\}$$

5. Berechnen des Maximum-Likelihood-Parameters

$$\widehat{\theta}$$

7.0.1 Antwort auf Frage

7.0.2 Welcher Parameterwert paßt am besten zu den Beobachtungen?

- Bestimme Schätzer
- Max-Log-Likelihood-Prinzip

7.1 Ist die Beobachtung mit einem gegebenen Parameterwert vereinbar?

Insbesondere hier vereinbar mit dem Grenzwert?
$\Rightarrow$ nächstes Kapitel "Statistische Tests"

7.2 Welche Parameterwerte sind mit der Beobachtung vereinbar?

$\Rightarrow$ nächstes Kapitel "Vertrauensintervalle"

---

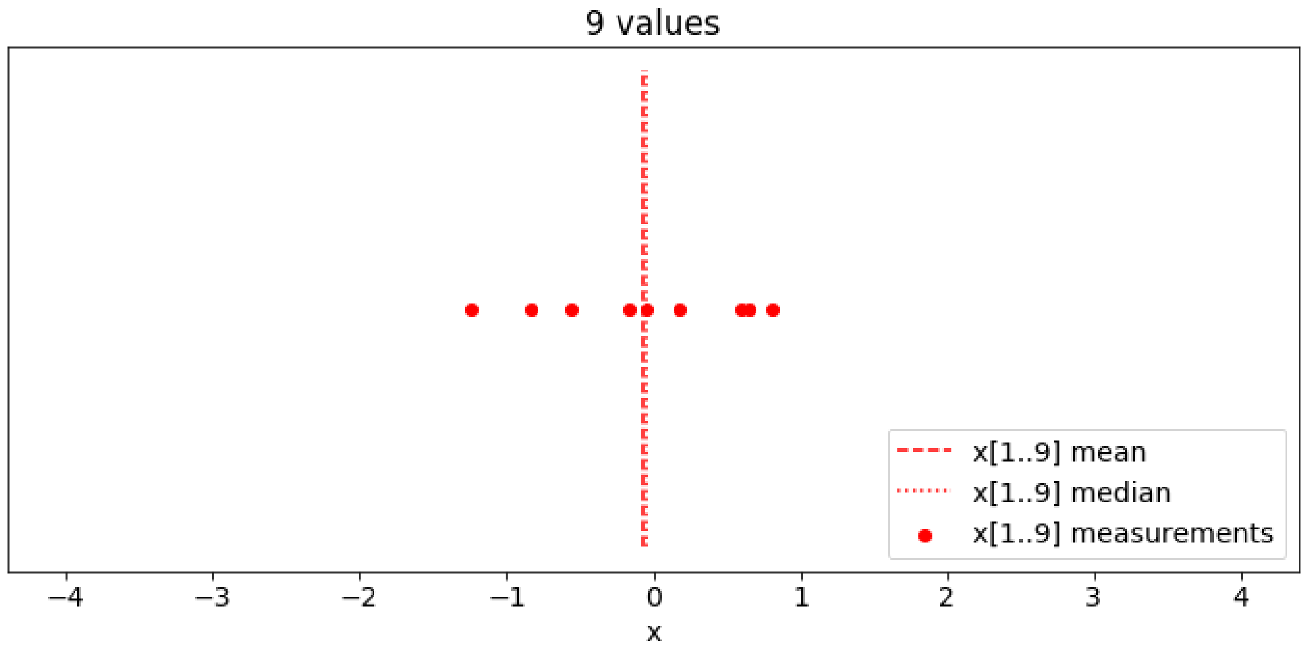
8 Robuste Schätzung

Problem Schätzungen gelten nur, wenn die Voraussetzungen der Modellverteilung die *richtige* ist.

Umgehung des Problems Oft wird (zu Recht) die Normalverteilung angenommen.
Jedoch ist die Schätzung einer Normalverteilung empfindlich gegen *Ausreisser*.

8.0.1 Ein Beispiel:

In [5]:
```
'''outliers influence mean (and median)'''
np.random.seed(56789)                       # obtain the same, but random result again
x9 = stats.norm.rvs(size=9)                 # nine standard-normally distributed values
plt.figure(figsize=(12,5))
plt.scatter(x9,np.zeros_like(x9), color='r',  label='x[1..9] measurements')
plt.plot(2*[x9.mean()],     (-.5, .5), 'r--', label='x[1..9] mean')
plt.plot(2*[np.median(x9)], (-.5, .5), 'r:',  label='x[1..9] median')
plt.xlabel('x')
plt.title('9 values')
plt.plot((-4, 4), (0, 0) , 'w,')
plt.yticks([])
plt.legend(loc='lower right');
```



Arithmetisches Mittel

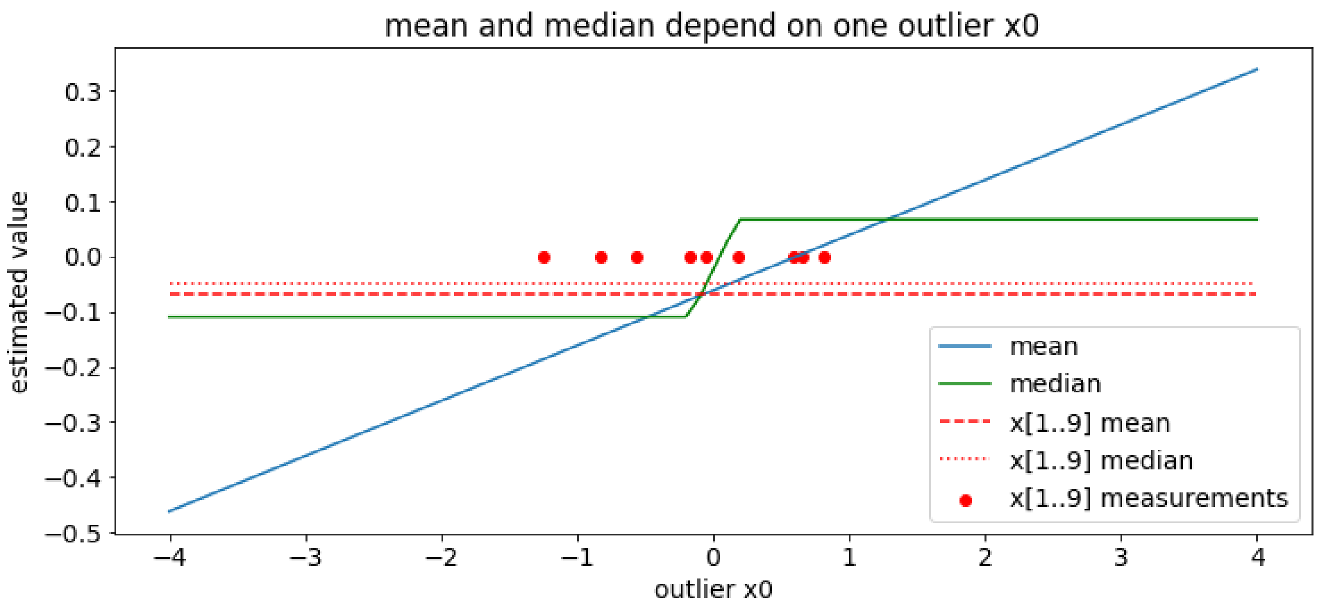
$$\overline{x_9} = \frac{1}{9}\sum_{i=1}^{9} x_i$$

ein neuer Messwert dazu:

$$\overline{x_{10}} = \frac{1}{10}\sum_{i=0}^{9} x_i$$
$$= \frac{9}{10}\overline{x_9} + \frac{1}{10}x_0$$

In [5]:
```
'''outliers influence mean (and median)'''
np.random.seed(56789)                           # obtain the same, but random result again
x9 = stats.norm.rvs(size=9)                     # nine standard-normally distributed values
x0 = np.linspace(-4., 4., 81)                   # one of these additional values to the nine x_1..9
xmeans = np.asarray( [np.append(x9, x).mean() for x in x0] )     # mean - depending on x0
print('means:    ', xmeans[:3], ' - ', xmeans[-3:])
plt.figure(figsize=(12,5))
plt.plot(x0, xmeans, label='mean')
xmeds = np.asarray( [np.median(np.append(x9, x)) for x in x0] ) # median - depending on x0
print('medians: ', xmeds[:3], ' - ', xmeds[-3:])
plt.plot(x0, xmeds, 'g-', label='median')
plt.scatter(x9,np.zeros_like(x9), color='r', label='x[1..9] measurements')
plt.plot((-4, 4), 2*[x9.mean()], 'r--', label='x[1..9] mean')
plt.plot((-4, 4), 2*[np.median(x9)], 'r:', label='x[1..9] median')
plt.xlabel('outlier x0')
plt.ylabel('estimated value')
plt.title('mean and median depend on one outlier x0')
plt.legend(loc='lower right');
```

```
means:     [-0.46192852 -0.45192852 -0.44192852]  -  [ 0.31807148  0.32807148  0.33807148]
medians:   [-0.10973318 -0.10973318 -0.10973318]  -  [ 0.06628869  0.06628869  0.06628869]
```



8.1 Robuste Verfahren

Wichtigste Kenngrößen

- Lage
- Streuung einer Verteilung.

Problem

- Die Lage wird durch Ausreisser stark beeinflußt.
- Die Streuung wird durch Ausreisser stark beeinflußt.

Ausweg Gib Vorgabe *Normalverteilung* auf, zugunsten einer langschwänzigen (*high kurtosis*) Verteilung.

Vorteil

- berechnet arithmetisches Mittel besser, wenn Ausreisser in den Beobachtungen.
- stört wenig, wenn doch normalverteilt

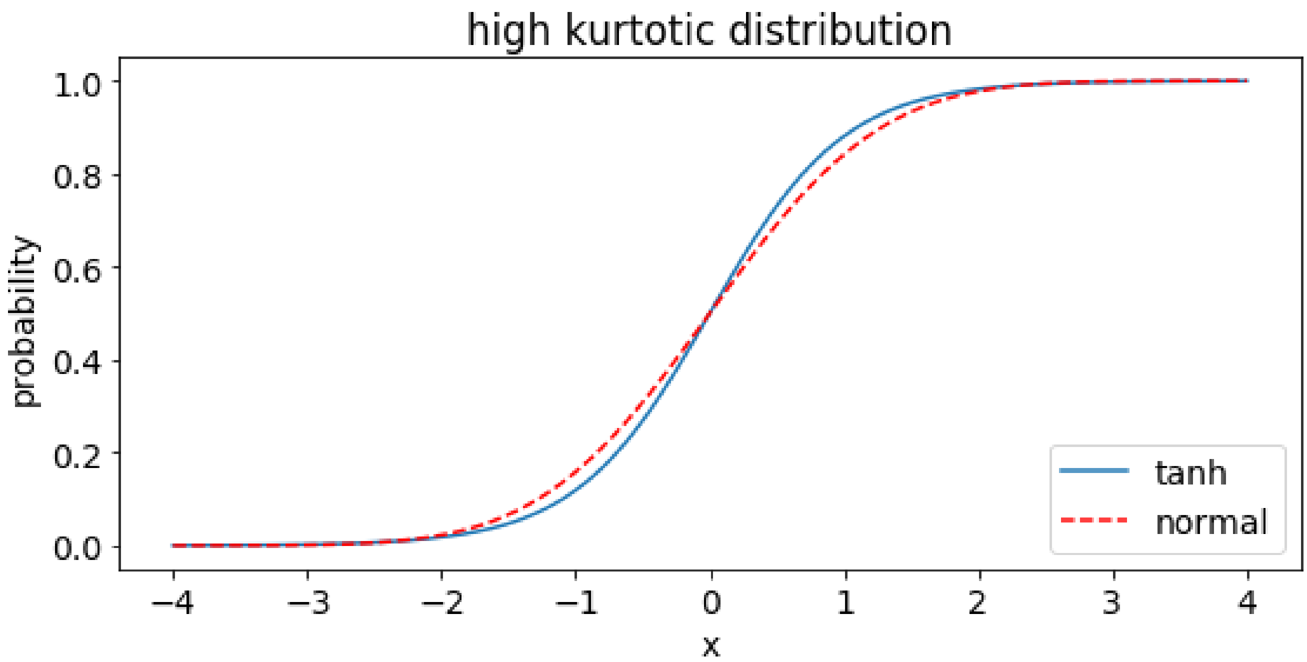
8.1.1 Beispiel für langschwänzige Verteilungsfunktion: tanh(x)

```
In [6]: '''create a complete distribution by just defining the cdf'''
        class tanh_gen(stats.rv_continuous):
            def _cdf(self, x):
                return .5+.5*np.tanh(x)         # the tanh-function as cdf
```

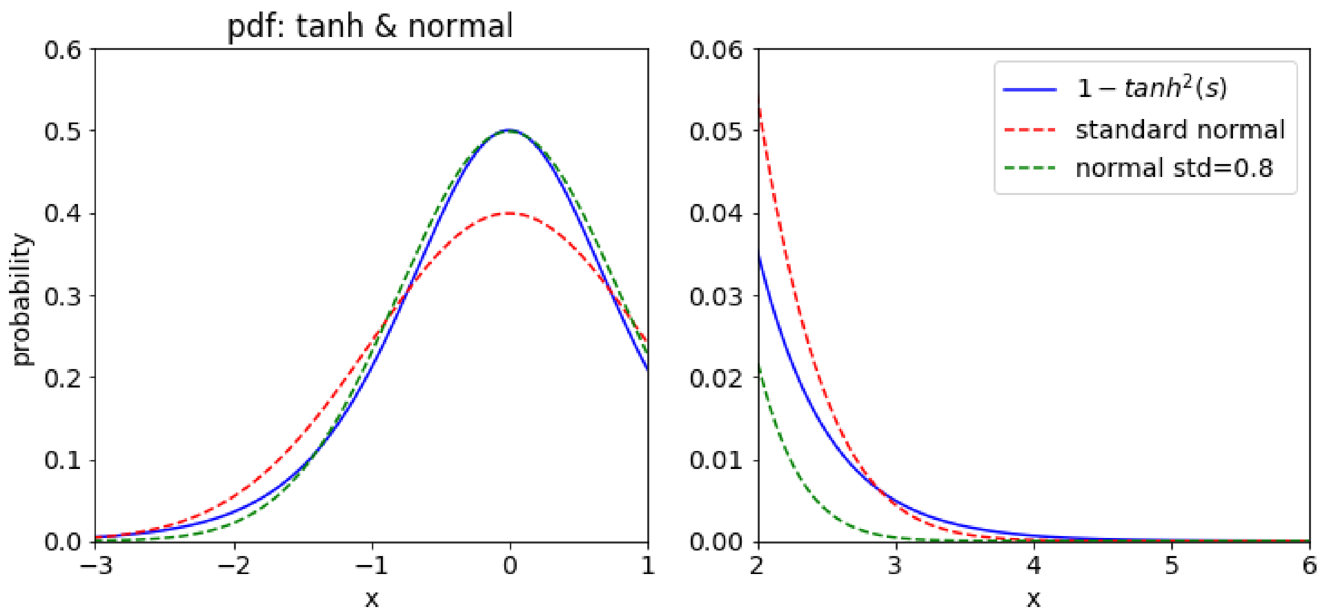
```
In [7]: highkurtosis = tanh_gen(name="high kurt tanh") # call once to establish
```

```
In [8]: '''show cumulative distribution function of high kurtotic tanh'''
        fig = plt.figure(figsize=(9, 4))
        xs = np.linspace(-4., 4., 601)
        plt.plot(xs, highkurtosis.cdf(xs), label='tanh')
        plt.plot(xs, stats.norm.cdf(xs), 'r--', label='normal')     # compare with standard normal distribution
        plt.title('high kurtotic distribution')
        plt.xlabel('x')
        plt.ylabel('probability')
        plt.legend(loc='lower right');
        kurtosis = highkurtosis.stats(moments='mvsk')[3]            # gives mean, var, skew and kurt
        print('kurtosis of tanh-distribution is {:.3f}'.format(np.float(kurtosis)))
```

```
kurtosis of tanh-distribution is 1.200
```




```
In [9]: '''show probability density function of "high kurt tanh" - despite not defined'''
        x = np.linspace(-10., 10., 1001)          # x in 0.01 resolution
        p_n = stats.norm.pdf(x)                   # comparison: standard normal distribution
        p_s = stats.norm(0, .8).pdf(x)            # comparison: tighter normal distribution
        p_t = highkurtosis.pdf(x)                 # the defined high kurtotic distribution
        f=plt.figure(figsize=(12, 5))
        f.add_subplot(1, 2, 1)                    # main plot: x=-3 to 1
        plt.xlim(-3.0, 1.0)
        plt.ylim(0., 0.6)
        plt.plot(x, p_t, 'b-')
        plt.plot(x, p_n, 'r--')
        plt.plot(x, p_s, 'g--')
        plt.title('pdf: tanh & normal')
        plt.xlabel('x')
        plt.ylabel('probability');
        f.add_subplot(1, 2, 2)                    # 10x magnified: x=1 to 6
        plt.xlim(2.0, 6.0)
        plt.ylim(0., 0.06)
        plt.plot(x, p_t, 'b-', label='$1-tanh^2(s)$')      # pdf of tanh.cdf
        plt.plot(x, p_n, 'r--', label='standard normal')  # standard normal pdf
        plt.plot(x, p_s, 'g--', label='normal std=0.8')   # tighter normal pdf
        plt.xlabel('x')
        plt.legend(loc='upper right');
```

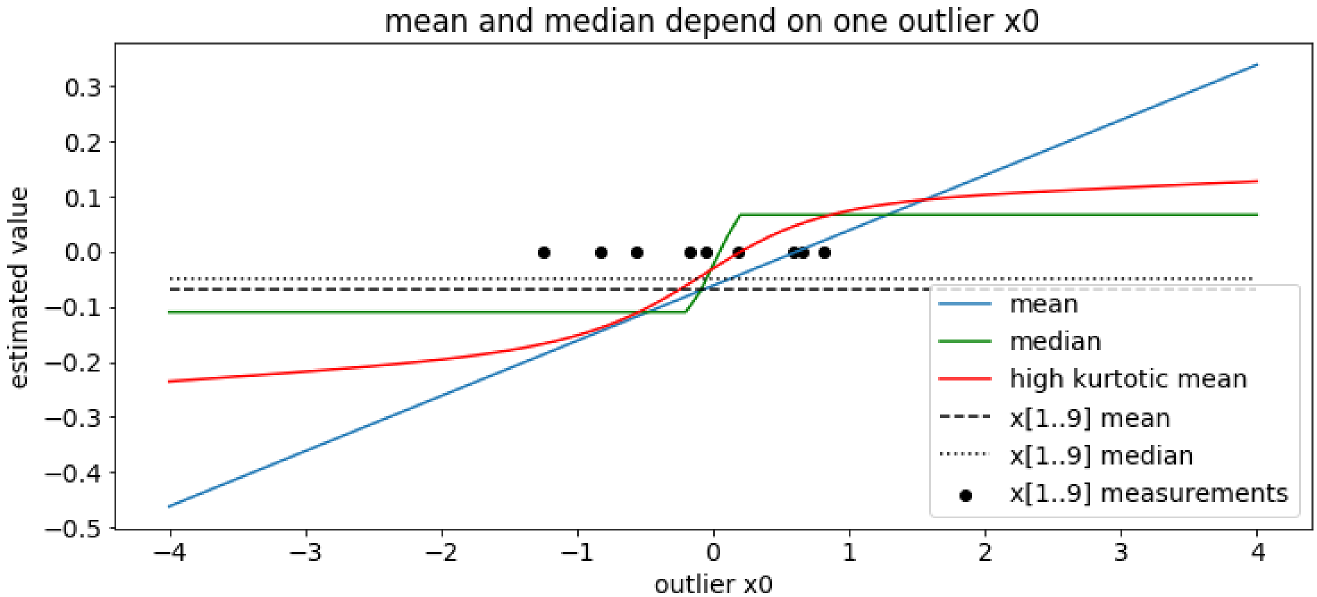



```
In [10]: '''outliers influence mean but less high kurtotic distribution'''
         # same as above
         np.random.seed(56789)
         x9 = stats.norm.rvs(size=9)
         x0 = np.linspace(-4., 4., 81)
         xmeans = np.asarray( [np.append(x9, x).mean() for x in x0] )
         print('means:    ', xmeans[:3], ' - ', xmeans[-3:])
         plt.figure(figsize=(12,5))
         plt.plot(x0, xmeans, label='mean')
         xmeds = np.asarray( [np.median(np.append(x9, x)) for x in x0] )
         print('medians: ', xmeds[:3], ' - ', xmeds[-3:])
         plt.plot(x0, xmeds, 'g-', label='median')

         hkms = [highkurtosis.fit(np.append(x9, x))[0] for x in x0]  # means of high-kurtotic fit
         plt.plot(x0, hkms, 'r-', label='high kurtotic mean')        # the new plot here

         plt.scatter(x9, np.zeros_like(x9), color='k', label='x[1..9] measurements')
         plt.plot((-4, 4), 2*[x9.mean()], 'k--', label='x[1..9] mean')
         plt.plot((-4, 4), 2*[np.median(x9)], 'k:', label='x[1..9] median')
         plt.xlabel('outlier x0')
         plt.ylabel('estimated value')
         plt.title('mean and median depend on one outlier x0')
         plt.legend(loc='lower right');
```

```
means:     [-0.46192852 -0.45192852 -0.44192852]  -  [ 0.31807148  0.32807148  0.33807148]
medians:  [-0.10973318 -0.10973318 -0.10973318]  -  [ 0.06628869  0.06628869  0.06628869]
```



8.2 Ausblick

Testen, wie viele Ausreisser die Berechnung verträgt, um dennoch sinnvolle Werte zu erhalten:

- Bruchpunkt, Anteil m an Ausreissern an gesamten Beobachtungen n.
- Spätestens bei $m > \frac{n}{2}$ bricht die Schätzung zusammen.

Siehe auch Kapitel "Bootstrap".

9 Zusammenfassung Punktschätzer

Stochastik

- Zufallsvariable mit Wahrscheinlichkeitsverteilung(sdichte) $f(x)$
- **i.i.d.** als Voraussetzung

Punktschätzer

- Für Parameter
  - μ und σ^2 der Normalverteilung
  - λ der Poissonverteilung
  - ...
- Für Kenngrößen
  - Arithmetisches Mittel schätzt Erwartungswert, Mittelwert der Grundgesamtheit
  - Stichproben-Varianz schätzt Varianz der Grundgesamtheit
  - ...

10 ...

Punkschätzer finden

- Kleinste-Quadrate-Methode
- Max-Log-Likelihood-Prinzip
- Bayes-Statistik

Robuste Schätzung

- verringert Fehler des Lageparameters gegenüber Ausreißern
- gelockerte Verteilung, akzeptiert Ausreißer mit höherer Wahrscheinlichkeit
- obwohl nur näherungsweise zur Theorie passend ($\mathcal{N}$)

11 Ausblick

Intervallschätzer

- Konfidenzintervalle

Teststatistik

- Unterschied von Werten
- Verhältnis von Werten

12 Fragen?