QQ-Plot
Parameterschätzung

Peter Büchel

HSLU TA

Stat: Block 05

Beispieldatensatz Betondruckfestigkeit

k	$x_{(k)}$
1	24.4
2	27.6
3	27.8
4	27.9
5	28.5
6	30.1
7	30.3
8	31.7
9	32.2
10	32.8
11	33.3
12	33.5
13	34.1
14	34.6
15	35.8
16	35.9
17	36.8
18	37.1
19	39.2
20	39.7

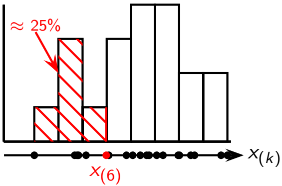
- Messung: Betondruckfestigkeit von $n = 20$ verschiedenen Proben
- Wie gut können die Daten mit einer Normalverteilung beschrieben werden?

Beispieldatensatz Betondruckfestigkeit

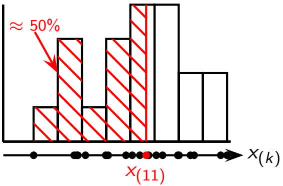
k	$x_{(k)}$	$\alpha_k = (k - 0.5)/n$	$\Phi^{-1}(\alpha_k)$	q_{α_k} **(für $\mathcal{N}(32.54, \sigma^2)$)**
1	24.4	0.025	-1.9600	24.53167
2	27.6	0.075	-1.4395	26.69133
3	27.8	0.125	-1.1503	27.89136
4	27.9	0.175	-0.9346	28.78670
5	28.5	0.225	-0.7554	29.53023
6	30.1	0.275	-0.5978	30.18445
7	30.3	0.325	-0.4538	30.78201
8	31.7	0.375	-0.3186	31.34273
9	32.2	0.425	-0.1891	31.88021
10	32.8	0.475	-0.0627	32.40478
11	33.3	0.525	0.0627	32.92522
12	33.5	0.575	0.1891	33.44979
13	34.1	0.625	0.3186	33.98727
14	34.6	0.675	0.4538	34.54799
15	35.8	0.725	0.5978	35.14555
16	35.9	0.775	0.7554	35.79977
17	36.8	0.825	0.9346	36.54330
18	37.1	0.875	1.1503	37.43864
19	39.2	0.925	1.4395	38.63867
20	39.7	0.975	1.9600	40.79833

QQ-Plot Betondruckfestigkeit

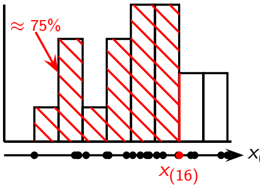
Empirisches 25% Quantil



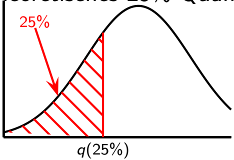
Empirisches 50% Quantil



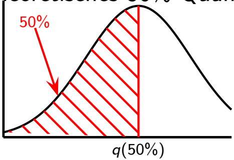
Empirisches 75% Quantil



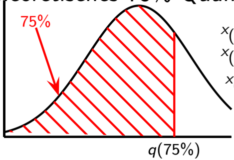
Theoretisches 25% Quantil



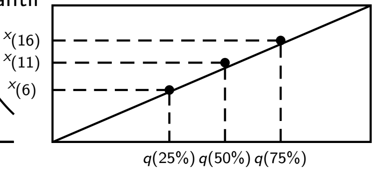
Theoretisches 50% Quantil



Theoretisches 75% Quantil



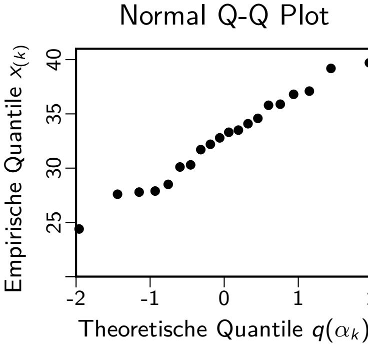
QQ-Plot



Normal-Plot

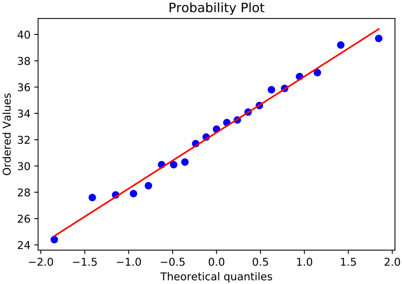
- Plotten theoretischen Quantile $q(\alpha_k) = \Phi^{-1}(\alpha_k)$ der Standardnormaverteilung gegen die empirischen Quantile $x_{(k)}$

1. Datensatz: $x_1, x_2, \ldots, x_n$
2. $\alpha_{(k)} = \frac{k-0.5}{n}, \; k = 1, 2, \ldots, n$
3. Theoretische Quantile: $q(\alpha_k) = \Phi^{-1}(\alpha_k)$
4. Empirische Quantile: $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$
5. $\left(q(\alpha_k), x_{(k)}\right)$ plotten

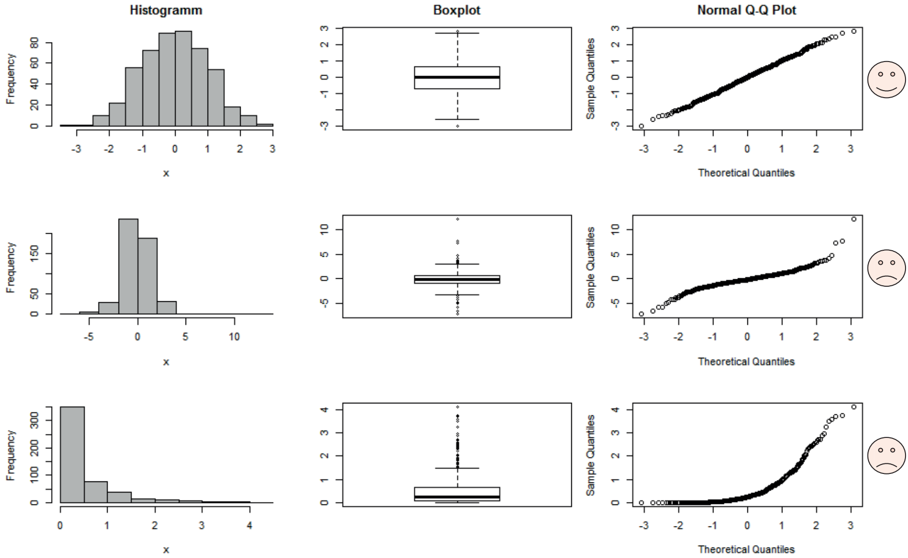


Normal-Plot mit Python

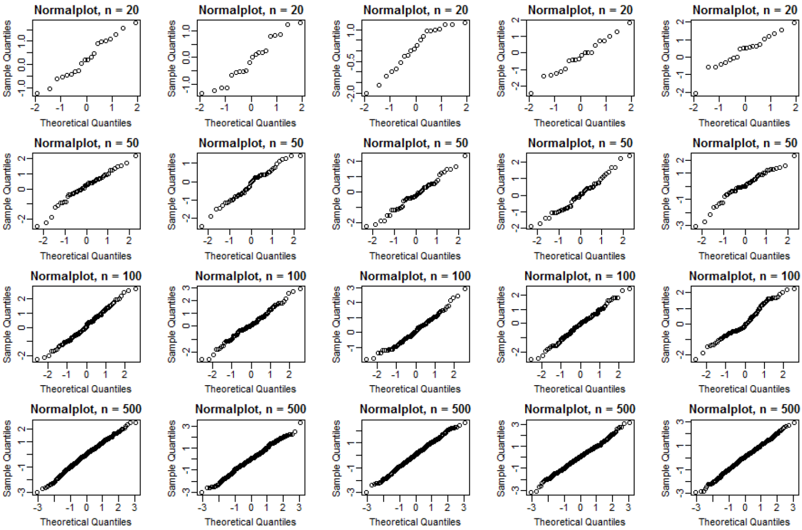
```
import numpy as np
import scipy.stats as st
import matplotlib.pyplot as plt

x = np.array([24.4, 27.6, 27.8, 27.9, 28.5, 30.1, 30.1, 30.3,
31.7, 32.2, 32.8, 33.3, 33.5, 34.1, 34.6, 35.8, 35.9, 36.8, 37.1,
39.2, 39.7])

st.probplot(x, plot=plt)
```



Beispiele Normalplots für 3 Datensätze mit $n = 500$



Normalplots von simulierten Standardnormalverteilungen



Parameterschätzung - Maximum-Likelihood

- $X \sim \mathrm{Pois}(\lambda)$ mit unbekanntem Parameter λ, welchen wir mit der Maximum-Likelihood Methode schätzen wollen.

- Datensatz : $x_1, \ldots, x_n$

- **Likelihood-Funktion**: Wahrscheinlichkeit, dass $X_1 = x_1$ **und** $X_2 = x_2$ etc. **und** $X_3 = x_3$ beobachtet werden, falls $X_i \sim$Pois(λ) und alle Beobachtungen x_i unabhängig voneinander sind:

$$
\begin{aligned}
P(X_1 = x_1 \cap \ldots \cap X = x_n) &= P(X_1 = x_1) \cdot P(X_2 = x_2) \cdots P(X = x_n) \\
&= e^{-\lambda} \frac{\lambda^{x_1}}{x_1!} \cdot e^{-\lambda} \frac{\lambda^{x_2}}{x_2!} \cdots e^{-\lambda} \frac{\lambda^{x_n}}{x_n!} \\
&= \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} \\
&= L(\lambda)
\end{aligned}
$$

- **log-Likelihood-Funktion** ist:

$$\begin{aligned} l(\lambda) &= \log(L(\lambda)) \\ &= \log\left(\prod_{i=1}^{n} e^{-\lambda}\frac{\lambda^{x_i}}{x_i!}\right) \\ &= \sum_{i=1}^{n} \log\left(e^{-\lambda}\frac{\lambda^{x_i}}{x_i!}\right) \end{aligned}$$

- Leitet man $l(\lambda)$ nach λ ab und setzt $l'(\lambda) = 0$, so erhält man die Gleichung

$$\frac{1}{\lambda}\sum_{i=1}^{n} x_i - n = 0\,.$$

- Maximum-Likelihood Schätzer : $\hat{\lambda} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

Maximum-Likelihood-Schätzer für Poisson-Verteilung

Maximum-Likelihood Schätzer

Sind die Datenpunkte x_i unabhängige Realisierungen der Zufallsvariablen $X_i \sim \text{Poisson}(\lambda)$, so ergibt der Maximum-Likelihood-Schätzer:

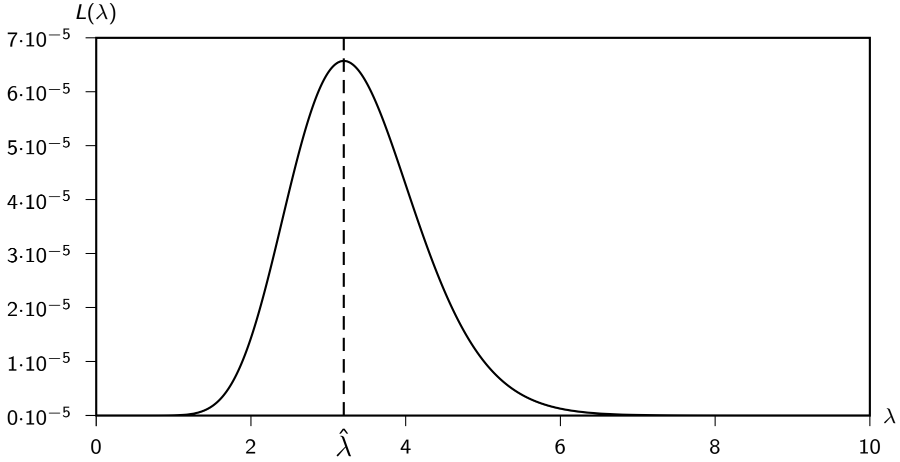
$$\hat{\lambda} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- **Beispiel:** Hersteller für Isolationsmaterialen misst die Anzahl kanzerogener Fasern pro mm^2 in 5 Proben. Gemessenen Werte $x_1, \ldots, x_5$ seien:

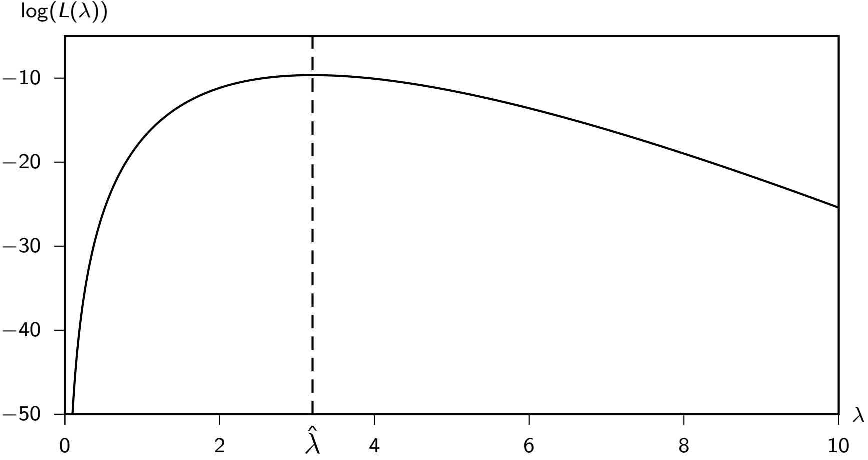
$$4, 1, 2, 3, 6$$

- Anzahl kanzerogener Fasern folgt Poisson-Verteilung: Beobachtungen sind $x_1, \ldots, x_5$ Realisierungen von $X_1, \ldots, X_5 \sim \text{Poisson}(\lambda)$ sind. Maximum-Likelihood-Schätzer: $\hat{\lambda} = \bar{x} = 3.2$

Likelihood-Funktion für Datensatz kanzerogener Fasern



Log-Likelihood Funktion für Datensatz kanzerogener Fasern



Momentenmethode

- Maximum-Likelihood Schätzer bei Poisson-Verteilung:

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

  d.h. $\hat{\lambda} = \widehat{E(X)} = \bar{x}$. Geschätzter Erwartungswert ist arithmetisches Mittel.

- Maximum-Likelihood Schätzer bei Binomial-Verteilung, z.B. x_1 Gewinne von n_1 Losen beim 1. Versuch, etc. (siehe Serie 5):

$$\hat{\pi} = \frac{x_1 + \ldots + x_n}{n_1 + \ldots + n_2} = \frac{x}{n} \ .$$

  Da $\mathrm{E}(X) = n\pi$ und $\hat{\pi} = \frac{\widehat{E(X)}}{n}$, ist der geschätzte Erwartungswert die beobachtete Anzahl Gewinne $\widehat{E(X)} = x$

Momentenmethode

- X Binomial-verteilt: $\mathrm{E}(X) = n\pi$
- Also $\pi = \mathrm{E}(X)/n$
- n (Anzahl unabhängiger Versuche) wird als bekannt vorausgesetzt
- Pragmatisch motivierte Schätzung ist dann:
$$\widehat{\mathrm{E}(X)} = x = \text{ beobachtete Anzahl Gewinne}$$
- Man schätzt den Erwartungswert also durch die Beobachtung.

Momentenmethode

Somit ergibt sich aufgrund der **Momentenmethode** die relative Häufigkeit

$$\hat{\pi} = x/n$$

Zusammenfassung Parameterschätzer

- Für unsere beobachteten Daten nehmen wir ein **Modell** an (z.B. Binomialverteilung)
- Das Modell enthält unbekannte Parameter (z.B. π)
- Basierend auf den beobachteten Daten versuchen wir, die Parameter zu schätzen (z.B. mit Momentenmethode, Maximum-Likelihood Methode).
- **Das heisst, dass wir basierend auf den beobachteten Daten versuchen, Rückschlüsse über den datengenerierenden Mechanismus zu ziehen!**