

Diskrete Verteilungen

Binomialverteilung: $X \sim \text{Bin}(n, p)$ $E(X) = n \cdot p$ $\text{Var}(X) = n \cdot p \cdot (1-p)$

$$F(x) = \sum_{i=0}^x \binom{n}{i} p^i (1-p)^{n-i}$$
$$P(X=x) = \binom{n}{x} p^x (1-p)^{n-x}$$

→ Verteilung der Häufigkeit X eines Ereignisses bei n unabhängigen Wiederholungen eines Versuchs.

Poisson-Verteilung: $X \sim \text{Pois}(\lambda)$ $E(X) = \lambda$ $\text{Var}(X) = \lambda$

$$F(x) = \sum_{i=0}^x e^{-\lambda} \cdot \frac{\lambda^x}{x!}$$
$$P(X=x) = e^{-\lambda} \cdot \frac{\lambda^x}{x!}$$

→ Verteilung der Anzahl Ereignisse in gegebenem Zeitintervall (unabhängig und in konstanter Rate, z.B. radioaktiver Zerfall)

Approximation der Binomialverteilung: für $n \rightarrow \infty$: $p \rightarrow 0$, $n \cdot p \rightarrow \lambda$

Python:

st. binom

st. poisson

Stetige Verteilungen ($F(x)$: kum. Verteilungsfunktion, $f(x)$: Wahrscheinlichkeitsdichte)

• Uniforme Verteilung: $X \sim \text{Unif}(a, b)$ $E(X) = \frac{a+b}{2}$ $\text{Var}(X) = \frac{(b-a)^2}{12}$

Breite: $b-a$ Fläche: $\frac{1}{b-a}$ Fläche: 1

$$G = \frac{b-a}{\sqrt{12}}$$

$$F(x) = \begin{cases} 0 & , x < a \\ \frac{x-a}{b-a} & , a \leq x \leq b \\ 1 & , x > b \end{cases}$$

$$f(x) = \begin{cases} \frac{1}{b-a} & , a \leq x \leq b \\ 0 & , \text{sowohl} \end{cases}$$

Python: $X \sim \text{Unif}(3, 7)$: $a=3$, $b=7$ $\text{loc}=a$, $\text{scale}=b-a$

st. uniform

$$P(4 \leq X \leq 6) = \text{uniform.cdf}(x=6, loc=3, scale=4) - \text{uniform.cdf}(x=4, loc=3, scale=4)$$

Exponentialverteilung: $X \sim \text{Exp}(\lambda)$

$$E(X) = \frac{1}{\lambda} \quad \text{Var}(X) = \frac{1}{\lambda^2} \quad \sigma = \frac{1}{\lambda}$$

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & , x \geq 0 \\ 0 & , \text{sonst} \end{cases}$$

$$f(x) = \begin{cases} \lambda \cdot e^{-\lambda x} & , x \geq 0 \\ 0 & , \text{sonst} \end{cases}$$

Approximation der Poisson-Verteilung: $T \sim \text{Exp}(\lambda) \rightarrow N \sim \text{Pois}(\lambda \cdot t)$

T : Aufzeit, N : Anzahl Auftreffer, t : Intervall (Zeit zw. Auftreffer)

Python: $X \sim \text{Exp}(3)$: $\lambda = 3$ $\text{scale} = 1/\lambda$

st. expon

$$P(0 \leq X \leq 4) = \text{expon.cdf}(x=4, scale=1/3)$$



• Normalverteilung: $X \sim N(\mu, \sigma^2)$ $E(X) = \mu$ $\text{Var}(X) = \sigma^2$

$$F(x) = \int_{-\infty}^x f(y) dy \quad f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Python: $IQ \sim N(100, 15^2)$: $\mu = 100$, $\sigma = 15$ $loc = 100$, $scale = 15$

st. norm

$$P(IQ > 130) = 1 - P(X \leq 130): 1 - \text{norm.cdf}(x=130, loc=100, scale=15)$$

• Standardnormalverteilung: $X \sim N(0, 1)$ $E(X) = \mu = 0$ $\text{Var}(X) = \sigma^2 = 1$

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad \Phi(x) = \int_{-\infty}^x \varphi(y) dy$$

Standardisierung ($N(\mu, \sigma^2) \rightarrow N(0, 1)$): $Z = \frac{x-\mu}{\sigma}$ $X \sim N(\mu, \sigma^2)$
 $Z \sim N(0, 1)$

Python:

st. norm $loc = 0$, $scale = 1$ (Standardwerte \rightarrow weglassen!)

Modelle für Messdaten



Wahrscheinlichkeitsdichte: $P(a < X \leq b) = F(b) - F(a) = \int_a^b f(x) dx ; \int_{-\infty}^{\infty} f(x) dx = 1$

Erwartungswert: $E(X) = \mu_x = \int_{-\infty}^{\infty} x \cdot f(x) dx \quad E(X^2) = \int_{-\infty}^{\infty} x^2 \cdot f(x) dx$

Varianz: $\text{Var}(X) = \sigma_x^2 = E[(X - E(X))^2] = \int_{-\infty}^{\infty} (x - E(X))^2 \cdot f(x) dx$

$$\text{Var}(X) = E(X^2) - [E(X)]^2 \Leftrightarrow E(X^2) = \text{Var}(X) + [E(X)]^2$$

Eigenschaften von Erwartungswert & Varianz:

$$E(a + bX) = a + b \cdot E(X)$$

$$\text{Var}(a + bX) = b^2 \cdot \text{Var}(X)$$

$$E(X+Y) = E(X) + E(Y)$$

$$\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) \quad (X \& Y \text{ i.i.})$$

$$E(X \cdot Y) = E(X) \cdot E(Y) \quad (X \& Y \text{ i.i.})$$

$$E(g(X)) = \sum_i g(x_i) \cdot p(x_i) \quad \text{bzw. } \int_{-\infty}^{\infty} g(x) \cdot f(x) dx \quad (\text{distinct \& sh.})$$

Quantile: Umkehrung der kumulativen Verteilungsfunktion $cdf \leftrightarrow ppf$

$$P(X \leq q(\alpha)) = \alpha \quad F(q(\alpha)) = \alpha \Leftrightarrow q(\alpha) = F^{-1}(\alpha)$$

$$\alpha\text{-Quantile f\"ur } Y = a + bX: q_Y(\alpha) = a + b \cdot q_X(\alpha)$$

$$f_Y(y) = \frac{1}{b} f_X\left(\frac{y-a}{b}\right)$$

Lineare Transformation: $X \sim N(\mu, \sigma^2), Y = a + bX \sim N(a + b\mu, b^2\sigma^2)$

Beispiel: Temperatur Grad Celsius \rightarrow Fahrenheit:

$$T_F = \frac{9}{5} \cdot T_c + 32, \quad a = \frac{9}{5}, \quad b = 32, \quad \sigma_F = b \cdot \sigma_c = \frac{9}{5} \sigma_c$$



(X & Y i.i.): X und Y unabhängig!

Gesetz der grossen Zahlen (GGZ): je grösser n , desto nähern \bar{X}_n am Erwartungswert; $n \rightarrow \infty \Rightarrow \bar{X}_n \rightarrow \mu$

Summe & rel. Häufigkeit:

$$E(S_n) = E(X_1 + X_2 + \dots + X_n) = \sum_{i=1}^n E(X_i) = n \cdot \mu$$

$$\text{Var}(S_n) = \text{Var}(X_1 + X_2 + \dots + X_n) = \sum_{i=1}^n \text{Var}(X_i) = n \cdot \sigma_x^2$$

$$\sigma(S_n) = \sqrt{n} \cdot \sigma_x$$

$$E(\bar{X}_n) = E\left(\frac{X_1 + X_2 + \dots + X_n}{n}\right) = \frac{1}{n} \sum_{i=1}^n (E(X_i)) = \frac{1}{n} \cdot n \mu = \mu$$

$$\text{Var}(\bar{X}_n) = \text{Var}\left(\frac{X_1 + X_2 + \dots + X_n}{n}\right) = \frac{1}{n} \sum_{i=1}^n (\text{Var}(X_i)) = \frac{1}{n} \cdot n \cdot \sigma_x^2 = \frac{\sigma_x^2}{n}$$

$$\text{Standardfehler: } \sigma(\bar{X}_n) = \frac{\sigma_x}{\sqrt{n}} \quad \text{relative Fehl.: } \bar{x}_n \pm \frac{\sigma(\bar{X}_n)}{\bar{x}_n} \cdot 100\%$$

Zentraler Grenzwertsatz: $X_i \sim N(\mu, \sigma^2)$, X_1, X_2, \dots, X_n i.i.d. "je grösser n ,

$$S_n \xrightarrow{d} N(n \cdot \mu, n \cdot \sigma_x^2) \quad \bar{X}_n \xrightarrow{d} N\left(\mu, \frac{\sigma_x^2}{n}\right) \quad \text{desto besser die Approximation!}$$

QQ-Plot „Passen Messdaten zu bestimmter Verteilung?“ (hier: Normalverteilung)

$x = \text{pd.Series}([\dots])$ # Messwerte

$\text{alpha_k} = (\text{np.arange}(1, x.size + 1) - 0.5) / x.size$

$\text{quant_theor} = \text{st.norm.ppf}(q=\text{alpha_k}, loc=x.mean(), scale=x.std())$

$\text{quant_empir} = \text{np.sort}(x)$

$\text{plt.plot}(\text{quant_theor}, \text{quant_empir}, 'o')$

alternativ

$\text{st.probplot}(x, plot=plt)$



Parameterschätzung

- Momentenmethode: Verteilung bekannt, Parameter unbekannt

z.B. Exponentialverteilung: unbekannten Parameter λ anhand $E(X)$ berechnen

$$E(X) = \frac{1}{\lambda} \Leftrightarrow \lambda = \frac{1}{E(X)} \quad (\text{theoretischer Parameter } \rightarrow \text{durch empirischen ersetzen})$$

$$\lambda \rightarrow \hat{\lambda}, E(X) \rightarrow \bar{x} \Rightarrow \hat{\lambda} = \frac{1}{\bar{x}}$$

- Maximum-Likelihood-Methode

n Beobachtungen x_1, x_2, \dots, x_n i.i.d., z.B. $X \sim \text{Bin}(n, \pi)$ mit $n=100$, $X=58$

gesucht: π , das möglichst gut zu Beobachtung passt

Idee: Ergebnis $X=58$ gilt als das Wahrscheinlichste: $P(X=58) = \binom{100}{58} \pi^{58} (1-\pi)^{42}$

muss maximal werden \rightarrow nach π ableiten und =0 setzen (kompliziert!)

Vereinfachung: beide Seiten Logarithmieren, da Extremum von $\log(f(x))$ auch ein Extremum von $f(x)$ ist!

- Likelihood-Funktion

n Beobachtungen x_1, x_2, \dots, x_n gegeben, Parameter p (von $f(x)$) gesucht

$L(p) = f(x_1) \cdot f(x_2) \cdot f(x_3) \cdots f(x_n) \rightarrow$ Logarithmieren (einfach!!)

- Logarithmen-Gesetze

$$y = \log_a(x) \Leftrightarrow a^y = x \quad a^{\log_a(x)} = x \quad \log_a(a) = 1 \quad \log_a(1) = 0$$

$$\log(uv) = \log(u) + \log(v) \quad \log\left(\frac{u}{v}\right) = \log(u) - \log(v)$$

$$\log(u^r) = r \cdot \log(u) \quad \log_a(x) = \frac{\ln(x)}{\ln(a)} = \frac{\log_b(x)}{\log_b(a)}$$



Hypothesentest

Passt Messreihe zu Größe? Passt Mittelwert der Messreihe zu wahren Mittelwert?

1. Modell: X_1, X_2, \dots, X_n i.i.d. $\sim N(\mu, \sigma^2)$

2. Nullhypothese: $H_0: \mu = \mu_0$

3. Alternativhypothese: $H_A: \mu \neq \mu_0$ (beidseitig), $\mu < \mu_0$, $\mu > \mu_0$ (einseitig)

4. Teststatistik unter H_0 : $X_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

5. Signifikanzniveau: $\alpha = 0.05 = 5\%$

6. Verwerfungsbereich: $K = (-\infty, x_{\min}] \cup [x_{\max}, +\infty)$ (beidseitig) $ppf(q = \dots)$

7. Testentscheid: $\bar{X}_n \in K: H_0$ wird verworfen! $p \leq \alpha$

$\bar{X}_n \notin K: H_0$ wird beibehalten! $p > \alpha$

$$p = cdf(x = \dots)$$

P-Wert: W'keit, dass unter H_0 ein mindestens so extremes Ergebnis in Richtung der Alternative beobachtet wird wie das bereits beobachtete.

σ_x (Standardabweichung der Population) bekannt?

- ja: z-Test (norm.ppf / norm.cdf)

- nein: t-Test ($t.ppf / t.cdf$) mit $n-1$ Freiheitsgraden

$$T = \frac{\bar{X}_n - \mu_0}{\hat{\sigma}_x / \sqrt{n}} \quad \text{oder} \quad T = \frac{\bar{D}_n - \mu_0}{\hat{\sigma}_0 / \sqrt{n}} \quad \text{für Differenzen (gepaarte Stichproben)}$$

Python:

z-Test: `norm.cdf(x=[sample mean], loc=[pop. mean], scale=[pop std] / n ** 0.5)`

(einseitig, zweiseitig: 1/2 - Pfeilspitze durch 2)

t-Test: `t.cdf(x=[sample mean], loc=[pop. mean], scale=[sample std] / sqrt(n), df=n-1)`

oder mit obiger Teststatistik T: `t.cdf(x=[sample mean], df=n-1)`

oder: `tr = tttest_1samp(a=[data], popmean=[mu_0]).statistic`

$p = t.cdf(tr, df=n-1)$



Vertrauensintervall

Angabe, mit welchen W'keit sich das wahre μ in welchem Bereich befindet.

$$I_{95\%} = [\bar{x}_n - z_{\frac{\alpha}{2}} \cdot \sigma_x / \sqrt{n}, \bar{x}_n + z_{\frac{1-\alpha}{2}} \cdot \sigma_x / \sqrt{n}] \quad \alpha=0.05$$

norm.interval(alpha=0.95, loc=[sample mean], scale=[sample std] / sqrt(n))

t.interval(... df=n-1)

$$\mu \text{ bekannt: } I = [\mu - z_{\frac{\alpha}{2}} \cdot \sigma, \mu + z_{1-\frac{\alpha}{2}} \cdot \sigma]$$

Weitere Testverfahren

Stichproben	gepaart	ungepaart
normalverteilt	ttest_rel(x, y)	ttest_ind(x, y, equal_var=False)
nicht normalverteilt	Vorzeichentest ¹	mannwhitneyu(x, y) ²

wilcoxon(d, correct=True)²

¹ Vorzeichentest: binom_test(x=[Anzahl > 0], n=[Anzahl], p=0.5)

² Wilcoxon: symmetrische Verteilung vorausgesetzt, i. d. R. mächtiger als t-Test

Varianzanalyse (ANOVA)

H0: alle Gruppen folgen gleichem Modell, g Gruppen mit m Beobachtungen

$Y_{ij} = \mu + T_i + \epsilon_{ij}$, Y_{ij} : j-te Beobachtung i-ter Gruppe, μ : gemeinsamer (globaler) Mittelwert,

T_i : behandlungsspezifische Abweichung, ϵ_{ij} : Fehlerterm pro Gruppe und Behandlung

$\mu = \mu_1$ (1. Gruppe als Referenz) $\rightarrow T_1 = 0, T_2 = \mu_2 - \mu, \dots$

from statsmodels.formula.api import ols

from statsmodels.stats.anova import anova_lm

fit = ols('Zielvariable ~ Faktor(en)', data=df).fit() # fit.summary(), fit.params

fit.pred = fit.get_prediction()

fit.pred.conf_int() # 95%-Vertrauensintervall

anova_lm(fit)

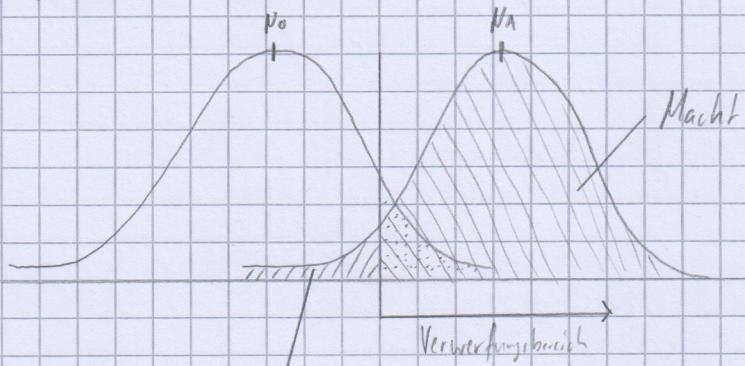
¹ Zv n F1+F2 (additiv)

Zv n ((F1+F2))

Zv n F1 * F2 (multiplikativ)

Zv n ((F1 * F2))

Fehler, Signifikanzniveau & Macht



Fehler 2. Art

Fehler 1. Art: H_0 falschlicherweise verworfen $\hat{=}$ Signifikanzniveau α

Fehler 2. Art: H_0 falschlicherweise beibehalten $\hat{=}$ 1 - Macht von μ_A

Bootstrap: durch Resampling einer Testreihen Informationen darüber gewinnen

Verfahren: pro Messpunkt eine Kugel in die Ume legen, Ziehen mit Zurücklegen

`s = np.random.choice(x, n * nboot, replace=True)` # n: Größe der Probe, nboot: Parameter

`s = np.reshape(s, (n, nboot))` # n Spalten, nboot Zeilen

`s = s.mean(axis=0)` # spaltenweise

`np.sqrt(s)`

`np.percentile(s, q=[2.5, 97.5])` # 95%-Vertrauensintervall

„echter“ Bootstrap:

`s = np.random.choice(x, n * nboot, replace=True)`

`table = np.reshape(s, (n, nboot))`

`col_mean = np.mean(table, axis=0)`

`delta_mean = col_mean - x.mean()`

`bounds = np.percentile(delta_mean, q=[2.5, 97.5])`

`interval = x.mean() - [bounds[1], bounds[0]]`

Zeitreihen



Muster: Trend (langfristig), Saisonalität (wiederholend), serielle Korrelation (benachbarte Werte)

- DataFrame: Datums Spalte als Zeitindex:

`df['date-col'] = pd.DatetimeIndex(df['date-col'])`

`df.set_index('date-col', inplace=True)`

- Einschränkung auf Zeitfenster Jan 2001 - Jun 2015:

`df.loc['2001-01': '2015-06']`

- Spaltenweise plotten: `df.plot(subplots=True)`

- Box-Cox-Transformation zur Korrektur von Schiefe & Varianz

$$g(x) = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{für } \lambda \neq 0 \\ \log(x) & \text{für } \lambda = 0 \end{cases}$$

```
def boxcox(x, l): # x: NumPy-Array, l: Lambda
    return np.log(x) if (l == 0)
    else (x**l - 1) / l
```

↳ Parameter λ (l) so wählen, dass Kurve möglichst linear / konstante Varianz!

- Zeitverschiebung (shifting) mit Lag von k : $g(x_i) = x_{i-k}$ ($k > 0$: rückwärts) !

`df['col'].shift(9) # k=-9` ($k > 0$: vorwärts)

Spezialfall $k=1$: backshift $B(x_i) = x_{i-1}$ → Anwendung: für Differenzen

- Boxplot: `df.boxplot('col', by='index-col')`

- Lag-Plot: from pandas.plotting import lag_plot `lag_plot(df['col'])`

- Zerlegung von Zeitreihen: $X_k = m_k + s_k + z_k$ bzw. $X_k = m_k \cdot s_k \cdot z_k$ (additiv/multipl.)

m_k : Trendkom., s_k : saisonaler Effekt, z_k : Fehlterm

- Moving-Average-Filte: Trend unter saisonalem Effekt abschätzen (Fensterbreite wählen)

`df['trend'] = df['x'].rolling(window=12).mean()`

`df['seasonal'] = df['x'] - df['trend'] # $\hat{s}_k = x_k - \hat{m}_k$`

- STL: seasonal decomposition of time from statsmodels.tsa.seasonal import decompose

- Vergleich von Darstellungen und Gr.ordnung: `decompose(np.log(df['x'])), period=12).plot()`

$$x = np.log(a.astype('float')) - np.log(a.shift(1).astype('float')) \quad \left. \right\} \text{plot}(x, y)$$

$$y = \quad b \quad b$$

• Durchschnittliche Saisonalität (Beispiel: 12 Monate, 20 Jahre)

```
df2 = df['seasonal'].values.reshape((12, 20)) # 12 cols m, 20 rows y
```

```
avg = np.nanmean(df2, axis=0) # column-wise (along col1)
```

```
df['seasonal_avg'] = np.tile(A=avg, reps=12) # original length
```

```
residue = df['seasonal'] - df['seasonal_avg']
```

• Dekomposition (keine Ausreißer, saison. konstant)

```
from statsmodels.tsa.seasonal import seasonal_decompose
```

```
seasonal_decompose(df['x'], model='additive', freq=12).plot()
```

```
seasonal_decompose(np.log(df['x']), model='additive').resid().plot()
```

Ableitungs- & Integrationssregeln

$$f(x) = x^n \quad f'(x) = n \cdot x^{n-1} \quad F(x) = \frac{1}{n+1} \cdot x^{n+1} + C$$

$$f(x) = e^x \quad f'(x) = e^x \quad F(x) = e^x + C$$

$$f(x) = a^x \quad f'(x) = a^x \cdot \ln(a) \quad F(x) = \frac{1}{\ln(a)} \cdot a^x + C, \quad a > 0, a \neq 1$$

$$f(x) = \ln(x) \quad f'(x) = \frac{1}{x} \quad F(x) = x \cdot \ln(x) - x + C$$

$$f(x) = \frac{1}{x} = x^{-1} \quad f'(x) = -\frac{1}{x^2} = -x^{-2} \quad F(x) = \ln(|x|) + C, \quad x \neq 0$$