Übungen zu Deskriptive Statistik und stetige Zufallsvariablen **Serie 8**

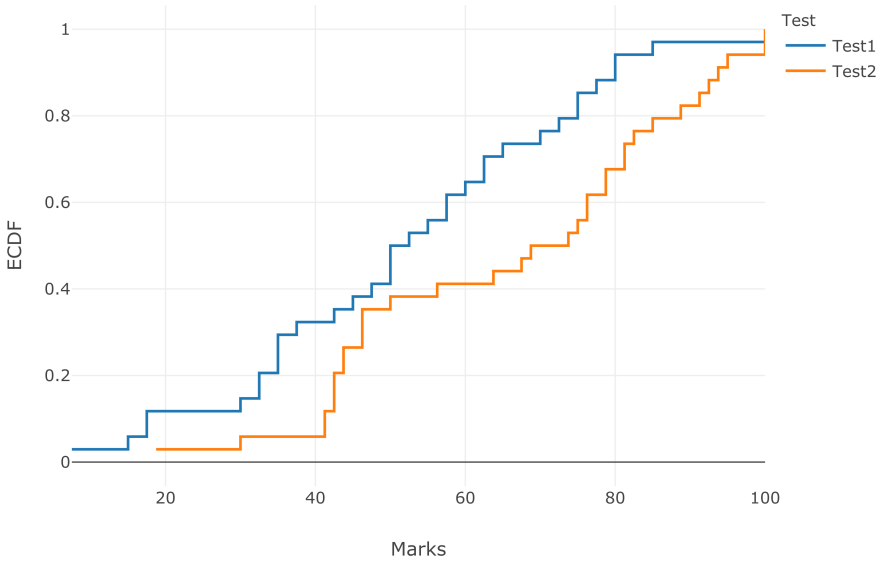
Aufgabe 1 Die Prüfungsergebnisse einer fiktiven Klassen sind auf Moodle unter der Dateiname `Exam_results.xlsx` zu finden. Die Klasse hat 2 Prüfungen geschrieben; beim ersten Test könnte man max. 20 Punkte und beim zweite Test könnte man max. 40 Punkte erreichen. Wir möchten mit Hilfe von deskriptiver Statistik untersuchen, ob das erste oder das zweite Test besser gelungen ist.

(i) Visualisierungen

Machen Sie einen Boxplot und einen Histogram der Daten. Probieren Sie verschiedene Werte der Bingrösse aus. Welche Darstellung gefällt Ihnen am Besten und warum (sowohl subjektiv als auch objektiv beantworten)?

(ii) Visualisierungen

Machen Sie einen Plot um die empirischen kumulativen Verteilungsfunktionen (ECDF) beider Tests zu zeigen (so etwas wie die folgende Graphik).



Einen ECDF Plot für einen Numpy-Array (Beispieldatensatz `data` hat nur eine Spalte) können Sie folgendermassen machen.

```
> from statsmodels.distributions.empirical_distribution import ECDF
> data = np.random.uniform(low=0, high=4, size= 10)
> ecdf = ECDF(data)
> df = pd.DataFrame({"x":ecdf.x, "y":ecdf.y})

> px.line(df, x = "x", y = "y") # smoothed ecdf

> px.line(df, x = "x", y = "y", line_shape="hv") # step ecdf
```

Überlegen Sie sich nun wie Sie die Funktion `ECDF` aus `statsmodels` auf alle Spalten eines Pandas Datensatzes anwenden können und event. müssen Sie danach die Daten in die nötige Form bringen. Ziel ist es die gewünschte Visualisierung mit möglichst wenig Code-Zeichen zu bekommen!

(iii) Quantitative Angaben (für jede Prüfung getrennt anzugeben)

Mit den Zufallsvariablen X und Y bezeichnen wir den prozentualen Punktzahlen der ersten bzw. der zweiten Prüfung. $\overline{x}$ ist der Mittelwert, $\tilde{x}$ der Median und s_x die Stichproben-Standardabweichung von X. Und $\overline{y}$ ist der Mittelwert, $\tilde{y}$ der Median s_y die Stichproben-Standardabweichung von Y.

(a) Berechnen Sie die Masszahlen $\overline{x}$, $\tilde{x}$, s_x, $\overline{y}$, $\tilde{y}$ und s_y.

(b) Die prozentuale Punktzahl von 50% der Studierenden liegt im welchen symmetrischen Bereich, um den Median der jeweiligen Prüfung herum?

(c) Die prozentuale Punktzahl von 80% der Studierenden liegt im welchen symmetrischen Bereich um den Median der jeweiligen Prüfung herum? Hinweis: verwenden Sie den Pandas Befehl `quantile`, z.B.folgendermassen um den Median eines Pandas-Datensatzes `df` zu berechnen

```
> df.quantile(0.5)
```

(d) Welche quantitative Aussagen können Sie über die Streung der Verteilung der Punktzahl jeder Prüfung machen, in dem Sie den jeweiligen Mittelwert und die Standardabweichung benützen?

Hinweis: benützen Sie `ECDF` um auszurechnen, wie viel Prozent der Verteilung im Bereich $\overline{x} \pm s_x$, $\overline{x} \pm 2 \cdot s_x$ und $\overline{x} \pm 3 \cdot s_x$ liegt (und analog für Y). Ein Beispiel als Hilfestellung

```
> data = np.random.uniform(low=0, high=4, size= 10)
> ecdf = ECDF(data)
# get cumulative probability for values
> print('P(X<3): %.3f' % ecdf(3))
```

Aufgabe 2 Die stetige Zufallsvariable X ist gleichverteilt zwischen 5 und 7, hat also die Wahrscheinlichkeitsdichte

$$f(x) = \begin{cases} \frac{1}{2}, & \text{für } x \in [5,7] \\ 0, & \text{sonst} \end{cases}$$

(i) Bestimmen Sie die kumulative Verteilungsfunktion $F(x)$ zu $f(x)$ und veranschaulichen Sie die beiden Funktionen in einem Graphen.

(ii) Berechnen Sie die Wahrscheinlichkeiten $P(X < 5.5)$ und $P(5.8 < X < 6.8)$.

(iii) Berechnen Sie den Median und das 90%-Quantil der Verteilung.

Aufgabe 3 Gegeben ist die Funktion

$$f(x) = \begin{cases} ax, & \text{falls } x \in [0,1] \\ 0, & \text{sonst} \end{cases}$$

mit $a \in \mathbb{R}_+$.

(i) Bestimmen Sie den Parameter a damit die Funktion f eine **Dichte** ist.

Hinweis: ignorieren Sie a und machen Sie eine kleine Skizze der Funktion $f(x)$ ohne a. Sie werden sehen, dass Sie einen Dreieck erhalten, dessen Flächeninhalt Sie problemlos berechnen können. Jetzt bestimme Sie a, so dass der Flächeninhalt unter der Kurve $f(x)$ gleich eins wird.

(ii) Berechnen Sie die zugehörige kumulative Verteilungsfunktion (CDF) und stellen Sie beide Funktionen graphisch dar.

Achtung: arbeiten Sie nur mit der normierten Funktion weiter! Wenn der Flächeninhalt unter der Kurve nicht eins ist, können wir keine Wahrscheinlichkeiten berechnen.

Hinweis: berechnen Sie von Hand oder mit Python $F(x) = P(X < x)$ für eine Reihe von x-Werten wie z.B. für $x_i \in \{0, 0.1, 0.2, 0.3, \ldots\}$. Sie erhalten für jedes x_i ein Dreieck und berechnen einfach dessen Fläche. Speichern Sie die Werte für x und $F(x)$ in einer Tabelle und stellen Sie den Zusammenhang graphisch dar. Erraten Sie nun, welche Funktion Ihre Daten am Besten beschreibt und testen Ihre Theorie aus.

(iii) Sei X eine Zufallsvariable mit der Dichtefunktion f. Berechnen Sie die folgenden Wahrscheinlichkeiten und zeichnen Sie diese ebenfalls in die erstellten Graphiken hinein.

$$P\left(\frac{1}{3} \leq X \leq \frac{3}{4}\right), \quad P\left(X \leq \frac{1}{2}\right), \quad P\left(X \geq \frac{3}{4}\right)$$