Vorgaben **Abgabe bis Do 25.4.24 um 24:00 Uhr**

- Lösungswege auf dieses Blatt. Abgabe als PDF in Moodle. Evtl.
- Achten Sie auf klar nachvollziehbare Lösungswege.
- Sofern nichts anderes gesagt wird, rechnen Sie W. mit 4 Nachkommastellen.

Vorgabe: Wo bekannte W-Modelle anwendbar sind, führen Sie die Berechnung von W. mit Python durch. Dokumentieren Sie im PDF den Methodenaufruf inkl. Parameterwerte und das Ergebnis.

Aufgabe 3.1

PA-Pkt

a) Wahrscheinlichkeitsbegriffe 1 P

Wie kommt der Wert zustande, den eine Wetterapp als Regenwahrscheinlichkeit für einen bestimmten Tag anzeigt? Welcher Wahrscheinlichkeits**begriff** (Typ) wird dabei verwendet? (Recherchieren Sie ggf.)

- Niederschlagswahrscheinlichkeit
- Wird in % angegeben, → 50% = bei 5 von 10 Fällen für den vorhersage zeitraum hat es geregnet

b) Formulieren Sie die folgenden W. 1-3 mit Hilfe von Ereignis-Bezeichnern: 3 P

Welche Ereignis-Bezeichner nutzen Sie? Legende:

W = Wort W
kW = kein Wort W
S = Spam

1. W., dass eine E-Mail mit dem Wort W eine Spam-Mail ist: $P(W \cap S)$
2. W., dass eine Spam-Mail das Wort W beinhaltet: $P(W \cup S)$
3. W., dass eine E-Mail nicht das Wort W beinhalten, aber dennoch Spam ist: $P(kW \cap S)$

c) Bedingte W. und Frage nach Unabhängigkeit 2 P

In einem gewissen Land wird eine Erhebung mit sehr hohem Umfang durchgeführt zur Frage:
Ist die W., dass man mit Hochschulabschluss eine Arbeit zum Niedriglohn hat, geringer als generell die W., einen Niedriglohn zu erhalten?
Aufgrund des hohen Umfangs darf man die relativen Häufigkeiten als Wahrscheinlichkeiten auffassen.
1.

Abschluss \ Lohnsegment	A: Niedriglohn	$\bar{A}$: Kein Niedriglohn	Gesamt
$\bar{B}$: Ohne Hochschulabschluss	7.328.163	22.282.037	29.610.200
B: Mit Hochschulabschluss	110.497	6.389.303	6.499.800
Gesamt	7.438.660	28.671.340	36.110.000

1. Formulieren Sie die Frage mit der Notation für W. und Ereignisse A, B. Antwort herleiten:

$P(A) = \frac{7438660}{36110000} = 0{,}206 = 20{,}6\%$

$P(B) = \frac{110.497}{7438660} = 0{,}0149 = 1{,}49\%$

Die Wahrscheinlichkeit einen Mindestlohn Job zu bekommen ist geringer mit einem Hochschulabschluss.

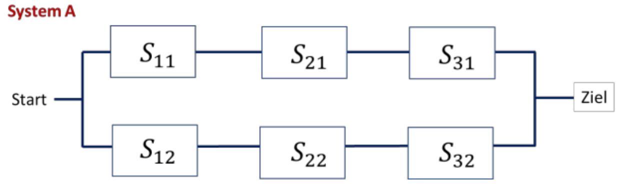
2. Sind die Ergeignisse A, B stochastisch unabh.?
Sie sind abhängig, da beide auf den Mindestlohn abhängen

d) **Zuverlässigkeit von Systemen bzw. Prozessen**

Für ein Teilsystem S_{ij} ist die W. für einen Ausfall gegeben durch p_i. Die W. für G_{ij} = Teilsystem S_{ij} fuktioniert gut, ist also $q_i = 1 - p_i$. Alle Teilsysteme fallen unabhängig voneinander aus. Berechnen Sie die W., dass folgende Systeme A, B, C funktionieren.
Achten Sie auf Nachvollziehbarkeit des Lösungsweges. Ergebnisse mit den Symbolen p_i bzw. q_i.

3 P

System A

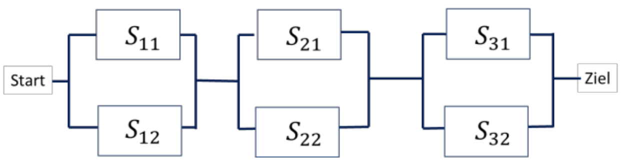


$P(p_1) = P((S_{11} \cap S_{21} \cap S_{31}) \cup (S_{12} \cap S_{22} \cap S_{32}))$

$P(p_1) = P(S_{11} \cap S_{21} \cap S_{31}) + P(S_{12} \cap S_{22} \cap S_{32}) - P(S_{11} \cap_{21} \cap S_{31} \cap S_{12} \cap S_{22} \cap S_{31})$

$P(q_1) = 1 - P(p_1)$

System B



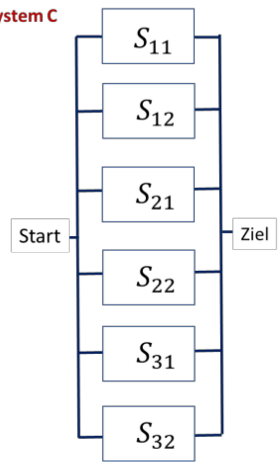
$P(p_2) = P(S_{11} \cap S_{12}) \cup P(S_{21} \cap S_{22}) \cup P(S_{31} \cap S_{32})$ **3 P**

$P(p_2) = (P(S_{11}) + P(S_{12}) - P(S_{11} \cap S_{12})) \cdot$
$(P(S_{21}) + P(S_{22}) - P(S_{21} \cap S_{22})) \cdot$
$(P(S_{31}) + P(S_{32}) - P(S_{31} \cap S_{32}))$

$P(q_2) = 1 - P(p_2)$

System C

3 P



$P(p_3) = P(S_{11}) \cup P(S_{12}) \cup P(S_{21}) \cup P(S_{22}) \cup P(S_{31}) \cup P(S_{32})$

$= P(S_{11}) + P(S_{12}) + P(S_{21}) + P(S_{22}) + P(S_{31}) + P(S_{32}) -$
$(P(S_{11}) \cdot P(S_{12}) \cdot P(S_{21}) \cdot P(S_{22}) \cdot P(S_{31}) \cdot P(S_{32}))$

$P(q_3) = 1 - P(p_3)$

e) Skript-Aufgabe 10.1 **7 P**

Nutzen Sie A_i = Paket i wird korrekt übertragen
X = Anzahl korrekt übertragener Pakete bei 100 versendeten Paketen
Formulieren Sie damit die gefragten P(Ereignis) mit Hilfe der Symbole der Mengenlehre bzw. Zufallsvariablen.

Für eine Datenübertragung in Form von IP-Telefonie ermittelt man aufgrund einer längeren Beobachtung der Paketverlustrate (Verhältnis Anzahl verlorengegangener zu Anzahl gesendeter Datenpakete, wie z. B. IP-Pakete) die W. für den Verlust eines Paketes zu 3 %. Gehen Sie davon aus, dass der Verlust eines Paketes **unabhängig** ist vom Verlust anderer Pakete.

$X = 97$

Wie groß ist die W., dass von 100 nacheinander gesendeten Paketen…

1) das erste und das 100. Paket ankommen?

$P(A_1 \cap A_{100}) = 0{,}97 \cdot 0{,}97$
$= 0{,}9409$
$= 94{,}09\%$

2) alle Pakete ankommen?

$P(A_{alle}) = 0{,}97^{100}$
$= 0{,}04756$
$= 4{,}756\%$

3) das erste Paket nicht ankommt?

$P(\bar{A}_1) = 3\%$

4) das erste oder das 100. Paket ankommt?

$P(A_1 \cup A_{100}) = P(A_1) + P(A_{100}) - P(A_1) \cdot P(A_{100})$
$= 0{,}97 + 0{,}97 - 0{,}97 \cdot 0{,}97$
$= 0{,}9991 = 99{,}91\%$

5) entweder das erste oder das 100. Paket ankommt?

$P(A_1 \cap \bar{A}_{100}) \cup P(\bar{A}_1 \cap A_{100})$
$= 0{,}97 \cdot 0{,}03 + 0{,}03 \cdot 0{,}97 - (0{,}97 \cdot 0{,}03 \cdot 0{,}03 \cdot 0{,}97)$
$= 0{,}0574 = 5{,}74\%$

6) genau 90 Pakete ankommen?

$P(A_{1-90}) = \binom{100}{10} \cdot 0{,}03^{10} \cdot 0{,}97^{90}$

$= \frac{100!}{(100-10)! \cdot 10!} \cdot 0{,}03^{10} \cdot 0{,}97^{90} = 6{,}5973 \cdot 10^{-4}$

7) mindestens 95 Pakete ankommen?

$P(x \leq 5) =$ stats.binom.cdf(k=5, n=100, p=0,03)

$P(x \leq 5) = 0{,}9192 = 91{,}92\%$

Bedingte W., Satz von Bayes, (Diskrete) W. über W-Modelle berechnen

Aufgabe 3.2 **Spam-E-Mails:** Einkommende E-Mails werden auf Spam geprüft. Aus Erfahrung weiß man, dass eine zufällig einkommende E-Mail mit W. 0,2 % mindestens einmal das Spam-verdächtige Wort W1, mit W. von 0.3 % mindestens einmal das Wort W2 und mit einer W. von 1 % **zwar nicht diese beiden, aber andere** Spam-verdächtige Wörter beinhaltet. Wir nehmen an, dass die kritischen Wörter unabhängig voneinander in derselben E-Mail auftreten können.

a) Berechnen Sie (mit 6 Nachkommastellen) die W., dass eine E-Mail kein Spam-verdächtiges Wort beinhaltet. (Achtung: Überlegen Sie mittels VENN-Diagramms, welche Ereignisse disjunkt sind, welche nicht). **4 P**

W1 = 0,2 %
W2 = 0,3 %
W3 = 1 %

Rechnen Sie im Folgenden weiter, als ob die Antwort in (a) die W. 98 % sei.
Rechenwege mit mindestens 4 Nachkommastellen. Ergebnis in % mit 2 Nachkommastellen.

b) Wie groß ist die W., dass sich unter 1000 E-Mails…
b.1) … höchstens 10 E-Mails mit mindestens einem Spam-verdächtigen Wort befinden?
Zufallsvariable und deren W-Verteilung? Lösung: **3 P**

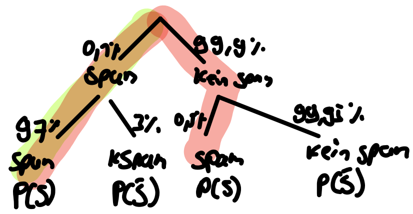
$P(x \le 10) = \text{stats.binom.cdf}(10, 1000, \overset{1-0{,}98}{0.02}) = 0{,}0102 = 1{,}02\%$

b.2) … zwischen 10 und 30 solcher E-Mails befinden (Grenzen einschließlich) **2 P**

$P(10 \le x \le 30) = \text{stats.binom.cdf}(30, 1000, 0.02) = 0.9874$

$= 0{,}9874 - \text{stats.binom.cdf}(9, 1000, 0.02) = 0{,}9827$

c) Ein Spam-Filter erkennt 97 % der Spam-E-Mails, er legt jedoch auch 0.5 % der Nicht-Spam-E-Mails in den Spam-Ordner ab. Aus Erfahrung weiß man, dass 0.1% aller eingehenden E-Mails tatsächlich Spam-Mails sind. Betrachten Sie Fragestellungen. Erstellen Sie zunächst einen passenden W-Baum. An den Ästen sowohl W. als Zahlenwerte als P(Ereignisbezeichner). **2 P**



c.1) Wie groß ist die W., dass eine E-Mail im Spam-Ordner landet? **2 P**

$P(S) = 0{,}999 \cdot 0{,}05 + 0{,}01 \cdot 0{,}97$
$= 0{,}0596 = 5{,}96\%$

c.2) Wie groß ist die W., dass eine E-Mail vom Spam-Filter korrekt behandelt wird? **2 P**

$P(S\text{-korrekt}) = 0{,}97 = 97\%$

c.3) Wie groß ist die W., dass eine E-Mail im Spam-Ordner tatsächlich Spam ist?
und wie nennt man diese W. mit dem Fachbegriff (binärer Klassifikatoren)? **2 P**

$P(S) = 0{,}9705 - 0{,}0999 \cdot 0{,}5 = 0{,}471$
$= 47{,}1\%$

c.4) Wie groß ist Trennfähigkeit (Segreganz) des Spam-Filters? **2 P**

$P(S) = 97\% - 0{,}5\% = 96{,}95\%$

<u>Aufgabe 3.3</u> Stichproben: Wir betrachten eine Personengruppe A, B, C mit 1000, 3000, 6000 Mitgliedern. Innerhalb jeder Gruppe handelt bzw. antwortet jede Person unabhängig von den anderen. Wir betrachten eine Eigenschaft, z.B. ob man Sport mag. Die W., dass eine Person aus Grupp A Sport mag, ist 80 %, für Gruppe B 60 % und für Gruppe C 30 %. Wir wählen für die folgenden Teilaufgaben jeweils eine zufällige Stichprobe von n Personen aus. Wie groß ist die W., ...

a) … W.,, dass bei $n = 100$ höchstens 40 Personen aus Gruppe A oder B stammen?

Zufallsvariable: $\frac{4000}{100} = 40$

$P(x \leq 40) = $ stats.poisson.cdf(40,40)

$= 0{,}5419$

Welche Hypergeometrische Verteilung passt exakt oder näherungsweise? Berechne die W.!

$P(x \leq 40) = $ stats.hypergeom.cdf(40, 10000, 10, 4000) $= 0{,}5435$

3 P

Welche Binomialverteilung passt exakt oder näherungsweise? Berechne die W. nach diesem Modell!

$P(x \leq 40) = $ stats.binom.cdf(40,100,0.4) $= 0{,}5433$

3 P

Falls die Ergebnisse relativ nahe beieinander liegen: Wie kann man dies begründen?

Es werden die selben Variablen benutzt, nur die Rechnungen ändern sich

1 P

b) …W., dass bei $n = 100$ der Gruppe C mindestens 70 keinen Sport mögen?

$6000 \cdot 0{,}7 = 4200$ kein Sport

3 P

$P(x \geq 70) = 1 - $ stats.hypergeom.cdf(70, 6000, 100, 4200)

$P(x \geq 70) = 0{,}4678$

<u>Aufgabe 3.4</u> Ein Scanner soll durch Produktionsfehler von Werkstücken erkennen, welche auf einem Fließband an seinen Sensoren vorbeilaufen. Im Durchschnitt wird 1 Fehler pro 5 Minuten gemeldet. Wie groß ist die W., dass in einer zufällig ausgewählten Minute mindestens 2 Fehler gemeldet werden?

Zufallsvariable: $\frac{1}{5} = 0{,}2$

Lösung: $P(x \geq 2) =$

4 P