STATISTIK – ÜBUNGEN TEIL I – DESKRIPTIVE STATISTIK

A) TECHNISCHE FERTIGKEITEN

A-1) Handelt es sich bei den vorliegenden statistischen Gesamtheiten um Bestands- oder Bewegungsgrößen?

a) Studierende an einer Hochschule.
b) Hochzeiten am Standesamt einer Gemeinde.
c) Bei der Behörde gemeldete Personenkraftwagen.
d) Maschinenausfälle in einer Werkstatt.
e) Wartende Kunden vor einem Abfertigungsschalter.

A-2) Im Servicecenter eines Unternehmens werden über einen Zeitraum eines Tages die eingehenden Anrufe aufgezeichnet. Gezählt wird die Anzahl der pro 10-Minuten-Zeitintervall eingehenden Anrufe. Für 40 derartige Zeitintervalle erhält man folgende Ergebnisse (die Liste kann mit Drag&Drop in R übernommen werden):

Liste = (0, 0, 1, 3, 4, 1, 2, 2, 1, 1, 1, 2, 3, 0, 2, 0, 1, 3, 1, 2, 2, 0, 1, 1, 6, 1, 0, 2, 3,
1, 1, 4, 2, 3, 2, 0, 3, 0, 1, 2)

a) Was stellt bei dieser Fragestellung die statistische Grundgesamtheit dar? Was sind die beobachteten Merkmale der statistischen Einheiten und wie sind sie skaliert?
b) Ermittle die absolute und relative Häufigkeitstabelle der eingehenden Anrufe und stelle die Häufigkeitsverteilung und Summenhäufigkeit grafisch dar.
c) Schätze das arithmetische Mittel aus der grafischen Darstellung für die Häufigkeit und den Median aus der grafischen Darstellung für die Summenhäufigkeit.
d) Berechne die möglichen Lageparameter (Zentralmaße und Streumaße)

Zur Erinnerung sei hier noch einmal die Tabelle aufgeführt:

	Zentralmaße	**Streumaße**
Extrem- u. Randgrößen	Modus	Min., Max, Spannweite
Anteilsgrößen	Median	Quantile
Gerichtete Summen	arithmetisches Mittel	mittlere absolute Abweichung Standardabweichung

e) Stelle die Daten in einem Boxplot dar.
(Siehe: https://www.youtube.com/watch?v=HsDeAoBOyS4)

A-3) Eine Anzahl von 1000 Kleinmotoren weist folgende Lebensdauer auf:

Lebensdauer [Jahren]	Anzahl der Motoren
[0, 2]	33
(2, 4]	276
(4, 6]	404
(6, 8]	237
(8, 10]	50

a) Stelle die Häufigkeitsverteilung und Summenhäufigkeit grafisch dar.

b) Schätze das arithmetische Mittel aus der grafischen Darstellung für die Häufigkeit und den Median aus der grafischen Darstellung für die Summenhäufigkeit.

c) Bestimme der Anteil der Motoren mit über 6 Jahren Lebensdauer.

d) Berechne die möglichen Lageparameter (Zentralmaße und Streumaße).

A-4) Ein technisches Servicecenter zeichnet an 100 Tagen die Häufigkeit der Einsätze auf. Es ergibt sich folgende Tabelle:

Anzahl der Einsätze pro Tag	Anzahl der Tage
[0, 10]	16
(10, 20]	48
(20, 30]	27
(30, 80]	9

a) Stelle die Häufigkeitsverteilung und Summenhäufigkeit grafisch dar.

b) Schätze das arithmetische Mittel aus der grafischen Darstellung für die Häufigkeit und den Median aus der grafischen Darstellung für die Summenhäufigkeit.

c) Bestimme der Anteil der Tage mit über 20 Einsätzen.

d) Berechne die möglichen Lageparameter (Zentralmaße und Streumaße). Welcher Aspekt könnte hier problematisch sein? Warum?

B) VERSTÄNDNISFRAGEN

B-1) Zeige, dass das arithmetische Mittel unter dem Schwerpunkt der Häufigkeitsfunktion liegt.

B-2) Verschiebungssatz zur Berechnung der Standardabweichung:

Das arithmetische Mittel ist gegeben durch $\mu = \sum_{i=1}^{n} f(a_i) \cdot a_i$ und die Varianz ist definiert als $\sigma^2 = \sum_{i=1}^{n} f(a_i) \cdot (a_i - \mu)^2$.

Zeige durch direkte Umformung, dass daraus folgt (Verschiebungssatz):

$$\sigma^2 = \sum_{i=1}^{n} f(a_i) \cdot {a_i}^2 - \mu^2$$

B-3) Das komma-separierte File „**Fehlerquote.csv**“ enthält das Prüfergebnis von 50 Bauteilen auf Funktionstüchtigkeit. Dabei steht der Eintrag „**0**“ für ein fehlerfreies Bauteil und „**1**“ für ein fehlerhaftes Bauteil.

a) Welche Skalierung hat dieses Merkmal?

b) Stelle die Messergebnisse in einer Häufigkeitstabelle und grafisch dar.

c) Wie kann man in diesem Beispiel das arithmetische Mittel berechnen und wofür steht es in diesem Fall?

d) Wir groß ist die Standardabweichung σ? Leite eine Formel her und zeige, wie man sie in diesem Fall einfach aus dem arithmetischen Mittel errechnen kann.

e) Wie müsste die Verteilung in diesem Fall sein, damit die Streuung maximal bzw. minimal wird?

B-4) Das komma-separierte File „**Verteilungsvergleich.csv**“ enthält in 4 Spalten die Daten von folgenden Messreihen: Ergebnis von 40 Würfen mit einem Würfel (Annahme: gleichverteilt), die Zeitspannen (in Minuten) zwischen 41 vorbeifahrenden Autos (Annahme: exponentialverteilt), die Länge von 40 Telefongesprächen in Minuten (Annahme: normalverteilt) und die Länge von 40 Holzstiften in cm (Annahme: normalverteilt). Vergleiche die vier verschiedenen Verteilungen in den folgenden Fragen:

a) Erstelle für jede Messreihe eine Häufigkeitstabelle sowie eine grafische Darstellung der Häufigkeitsverteilung und der Summenhäufigkeit.

b) Schätze das arithmetische Mittel aus der grafischen Darstellung für die Häufigkeit und den Median aus der grafischen Darstellung für die Summenhäufigkeit.

c) Bestimme für jede Messreihe jene Merkmalsausprägung, unter welcher die kleinsten 25 % zu finden sind.

d) Berechne die möglichen Lageparameter (Zentralmaße und Streumaße). Versuche die Lage und Größe der Lageparameter aufgrund der Eigenschaften der Verteilungen zu verstehen. (Z.B.: Worauf deutet die verschiedenen Lage von Median und arithmetischem Mittel, wie verhält sich die Standardabweichung zu Streuparametern wie Spannweite oder Interquartilsabstand?)

e) Zeichne für jede Messreihe einen Boxplot

B-5) Lageparameter als Sicherheitskennzahlen. Eine Fluggesellschaft wirbt damit, dass pro 48 Millionen Passagierkilometer lediglich 1 Todesfall zu beklagen war. (Das klingt sehr gut, wenn man nur 800 km fliegen will). Mit dieser Statistik, so die Fluggesellschaft, ist die Reise mit ihr 10-mal sicherer als eine Autofahrt. (M.a.W.: Im Autoverkehr gibt es 10 Tote auf 48 Millionen Passagierkilometer, oder 1 Toten auf 4,8 Mio. Passagierkilometer.) Allerdings fliegt das Flugzeug im Durchschnitt auch 10-mal schneller als ein Auto fährt. Welche Kennzahl würde eine Pro-Auto-Initiative dieser Werbung entgegenstellen? (Beachte die effektiven Reisezeiten.)

A.K. Dewdney schreibt dazu in seinem Buch „200 Prozent von nichts“: Beim Fliegen gehen daher Sicherheitsingenieure weder von den Passagierstunden noch von den Passagierkilometern aus. Weil die allermeisten Flugunglücke beim Starten oder beim Landen geschehen, ist es sinnvoller, die Wahrscheinlichkeit eines Unfalls pro Flug zu betrachten. Denn schließlich muss jedes Flugzeug bei jedem Flug sowohl starten als auch landen.

B-6) **Will Rogers Phänomen.** Die X12-Sensoren werden oft und in vielfältiger Weise in technische Geräte eingebaut: In Geräte, die sich in der harten Witterung der Alpen oder in der Arktis bewähren müssen ebenso, wie in zivile Gebrauchs- und Unterhaltungsgeräte. Diese harten Einsätze werden als „Risikoeinsätze“ bezeichnet und man weiß, dass die Lebensdauer der Sensoren bei Risikoeinsätzen mit durchschnittlich 5 Jahren um 4,5 Jahre kürzer ist als jene im zivilisatorischen Alltag mit durchschnittlich 9,5 Jahren. **Im Jahr 2006 hat man aufgrund einer Studie auch die Gruppe der Senioren und Kinder zu den Risikoeinsätzen zugeordnet**, da es hier öfter zu Fehlbedienungen mit Folgeschäden kommt.

Die nachfolgende Tabelle zeigt die durchschnittliche Lebensdauer der Sensoren in Abhängigkeit von der Verwendungsgruppe:

	durchschnittliche Lebensdauer des Sensors in Jahren	
Jahr	im Risikoeinsatz	im Alltagseinsatz
2005	5,0	9,5
2010	5,3	9,7

Im September 2010 publiziert das Unternehmen, dass es ihm gelungen ist, die durchschnittliche Lebensdauer ihres X12-Sensors weiter anzuheben (Siehe obige Tabelle). Die Lebensdauer stieg in beiden Bereichen um 2 % bis 6 %.

Dabei stand 2006 eine Anzahl von 30000 Testgeräten im überwachten Risikoeinsatz und 40000 Testgeräte (10000 davon bei Senioren und Kindern) im überwachten Alltagseinsatz. Eine genauere interne Untersuchung zeigte Jahre später den Umstand, dass die o.g. Tabelle trügerisch ist: Die Lebensdauer der X12-Sensoren ist vielmehr in allen Bereichen gesunken:

	durchschnittliche Lebensdauer in Jahren		
Jahr	im Risikoeinsatz (30000 Geräte)	Senioren und Kinder (10000 Geräte)	im Alltagseinsatz (30000 Geräte)
2005	5,0	8,0	10,0
2010	4,5	7,8	9,7

Wie ist es möglich, dass, obwohl die durchschnittlichen Lebensdauern in allen Gruppen zwischen 2005 und 2010 gesunken sind, die erste Datenauswertung einen durchgehenden Anstieg der Lebensdauern aufzeigt? Handelt es sich um eine Datenmanipulation, oder hängt es vielleicht davon ab, welcher Gruppe die Senioren und Kinder zugeschlagen werden?

B-7) Simpsons Paradoxon. In einem Studienversuch wird ermittelt, ob eine neue Produktionsmethode tatsächlich den Ausschuss verringert. An zwei Produktionsstandorten werden dazu vergleichende Versuche der beiden Produktionsverfahren durchgeführt. In Marginalstadt will man aufgrund der starken Auftragslage nach dem Produkt nicht viel Risiko eingehen und produziert nebenbei nur ein kleines Los im neuen Verfahren. Am erst vor kurzem errichteten Standort in Avantstadt will man das neue Verfahren gleich bei einer größeren Losgröße testen. Nach der Qualitätsprüfung ergibt sich die nachfolgend wiedergegebene Situation: Das getestete neue Produktionsverfahren hat an beiden Standorten mehr Ausschuss produziert als das herkömmliche:

Standort	Avantstadt		Marginalstadt	
Produktionsverfahren	herkömmlich	neu	herkömmlich	neu
Losgröße	250	1050	1050	250
Ausschuss abs.	7	42	63	18
Ausschuss rel.	2,8 %	4,0 %	6,0 %	7,2 %

Das neue Produktionsverfahren erhöht offenbar an beiden Produktionsstandorten den Ausschuss um unliebsame 1,2 Prozentpunkte. Die Entwickler des neuen Verfahrens sind über das Ergebnis bestürzt, denn sie haben viel Geld und Arbeit in die Entwicklung gesteckt.

Im Executive Summary an die Konzernzentrale werden der Kürze und Lesbarkeit wegen die Daten der beiden Standorte zusammengefasst – und lösen prompt Aufregung aus! Warum, das sieht man an nachfolgender Tabelle:

Produktionsverfahren	herkömmlich	neu
Losgröße	1300	1300
Ausschuss abs.	70	60
Ausschuss rel.	5,4 %	4,6 %

Plötzlich erscheint das neue Produktionsverfahren doch besser als das alte zu sein! Je nachdem, ob wir die Ergebnisse getrennt oder gemeinsam betrachten, ergibt sich für das neue Produktionsverfahren im Vergleich zum herkömmlichen entweder eine Verschlechterung oder eine Verbesserung. Dabei hat niemand geschummelt — die Zahlen sind völlig korrekt. Sie stammen aus ein und demselben Datenmaterial, abseits von unterschiedlichen statistischen Schwankungen. Woran kann dieser Widerspruch liegen?

B-8) Von den 583 Bewerberinnen und Bewerbern auf einen Studienplatz an der Hochschule für angewandte Wissenschaften haben nur 145 einen Studienplatz erhalten. Das entspricht einer Aufnahmequote von 24,9 % (sie nachfolgende Tabelle). Es ist allgemein bekannt, dass es nicht einfach ist, an dieser Hochschule einen Studienplatz zu ergattern.

	angenommen	abgelehnt	Erfolgsquote
Studenten	111	269	29,2 %
Studentinnen	34	169	16,7 %
Summe	145	438	24,9 %

Für Aufregung sorgt ein Leserbrief, welcher in der Lokalzeitung darauf hinweist, dass Frauen an dieser Hochschule im Bewerbungsverfahren eklatant benachteiligt werden: Die Erfolgsquote von Bewerberinnen liegt nicht nur unter dem Durchschnitt sondern

überdies
auch mehr als 12 Prozentpunkte unter der ihrer männlichen Kollegen. Der Fall beschäftigt noch einige Wochen die Medien.

In der QM-Abteilung der Hochschule sieht man sich die Zahlen etwas näher an und schlüsselt die nach den 4 vorhandenen Studiengängen auf:

	Studentinnen		Studenten	
Studiengang	beworben	angenommen	beworben	angenommen
STG 1	10	8	80	50
STG 2	3	2	60	38
STG 3	40	14	40	13
STG 4	150	10	200	10
Summe	203	34	380	111

Das Bild ist überraschend! Berechnet man die Aufnahmequoten in die einzelnen Studienprogramme so zeigt sich, dass die Studentinnen in jedem einzelnen Studiengang sogar besser abschneiden als ihre männlichen Kollegen. Wie ist das möglich?

C) OFFENE UNTERSUCHUNG

C-1) Offene Untersuchung: Das komma-separierte File „**Unternehmensumsaetze.csv**“ enthält Daten zu den 97 weltweit größten und börsennotierten Konzernen. In dieser Tabelle sind die Umsätze und Gewinne in Mrd. $ angegeben.

Untersuche die Daten mit den bekannten Methoden. Beantworte damit Fragen, wie z.B.: „Wie verteilen sich die Unternehmen auf Länder und Branchen?“, „Wie verteilen sich Gewinne, Umsätze und Mitarbeiter?“, „Welche Branchen generieren besonders viele Umsätze oder Gewinne pro Mitarbeiter?“, und andere mehr. Verwende dazu geeignete Häufigkeitsdarstellungen und Lageparameter.

Wiederhole die Erstellung einer Regressionsgeraden mit der Methode der kleinsten Quadrate. Siehe dazu https://www.youtube.com/watch?v=btsd-7AGDjc und weiters

https://www.youtube.com/watch?v=Ekbw28n6IX0 .

Lege eine passende Regressionsgerade durch die Merkmale „Mitarbeiter“ und „Gewinn“. Wo gibt es Ausreißer.

Erkundige Dich nach der Lorenzkurve und wende sie auf die Werte der Merkmale „Umsätze“ und „Mitarbeiter“ an.