Übung 9

1. Jemand möchte eine besonders exakte Messung vom Kupfergehalt einer Klärschlammprobe haben. Sie weiss, dass ihre Methode eine Standardabweichung von 30 mg/kg hat. Deshalb macht sie mehrere Bestimmungen und betrachtet dann den Durchschnitt.

a) Wieviele Messungen braucht sie, wenn sie höchstens einen Standardfehler von 10 mg/kg akzeptiert?

b) Angenommen, die Eichung (systematischer Fehler) sei auf 2 mg/kg genau vorgenommen worden. In der Praxis machen Genauigkeitsangaben (Standardfehler) keinen Sinn, die weniger als viermal so gross sind wie der systematische Fehler. Was ist die Höchstzahl von Messungen, die also sinnvollerweise gemacht werden sollten?

2. Von einem Gerät zur Messung von Luftschadstoffen möchte man wissen, ob es richtig geeicht ist. Dazu werden in einen geschlossenen Raum 20 ppm CO gespeist und dann vom Gerät 10 mal gemessen. Die folgenden Konzentrationen wurden abgelesen:

20.1, 20.3, 21.4, 19.2, 22.2, 20.1, 20.2, 20.4, 20.2, 20.3.

Kennzahlen: $\bar{x} = 20.440$, $s^2 = 0.660$, $s = 0.813$
Stelle eine geeignete Hypothese und Alternative auf:

a) Teste mit dem Vorzeichentest.

b) Benütze einen Überschlagstest mit der 2σ-Regel für $s(\overline{X})$ (Standardfehler des Mittels).

c) Verwende den t-Test.

d) Verwende den z-Test, falls die Genauigkeit des Messgeräts (d.h. der Einzelwerte) als $\sigma = 0.800$ ppm bekannt ist.

e) Vergleiche die Ergebnisse von a) bis d).

97.5%-Quantile der t-Verteilung mit ν Freiheitsgraden.

FG	1	2	3	4	5	6	7	8	9	10
$t_{\nu,0.975}$	12.70	4.303	3.182	2.776	2.571	2.447	2.365	2.306	2.262	2.228
FG	11	12	13	14	15	16	17	18	19	20
$t_{\nu,0.975}$	2.201	2.179	2.160	2.145	2.131	2.120	2.110	2.101	2.093	2.086
FG	30	40	50	60	70	80	90	100	1000	∞
$t_{\nu,0.975}$	2.042	2.021	2.008	2.000	1.994	1.990	1.987	1.984	1.962	1.96

3. Kacprzak und Chvojka (1976) untersuchten zwei Methoden zur Bestimmung des Quecksilbergehalts in Fischen. Sie verglichen eine neue Methode (selective reduction) mit einer etablierten Methode (permanganate method). Ein Vorteil der neuen Methode ist, dass damit das anorganische und das Methyl-Quecksilber gleichzeitig gemessen werden kann.
Der Quecksilbergehalt wurde bei 25 Jungfischen je mit beiden Methoden gemessen. Die Daten stehen Dir im File `mercury.dat` zur Verfügung. Die Variablen heissen `Fish`, `Select` (neue Methode) und `Perman` (etablierte Methode). Die Angaben sind in ppm.

a) Handelt es sich um verbundene (gepaarte) oder um unabhängige Stichproben?

b) Stelle diesen Daten in einem Streudiagramm dar und überlege Dir, wie die nachfolgenden Testergebnisse wohl ausfallen!

c) Führe den Vorzeichen-Test und den t-Test durch. Gib für jeden Test die Nullhypothese H_0, die Alternative H_A, die Teststatistik sowie den Verwerfungsbereich unter der Nullhypothese H_0 an.

d) Unterscheiden sich die beiden Messmethoden?

R-Anleitung:

b)
```
> mercury <-
read.table.url("http://stat.ethz.ch/Teaching/Datasets/mercury.dat")
                                          # Datensatz via Internet einlesen
> plot(mercury[,"Select"], mercury[,"Perman"])           # Streudiagramm
> abline(0,1)                                             # Gerade y = x
```

c)
```
> diff <- mercury[,"Select"] - mercury[,"Perman"]; diff
> mo <- ?                                               # Nullhypothese H0
> cat("Bindungen: ",sum(diff==mo),"\n")
> n <- ?                                     # Teststatistik K ~ Bin(?,0.5)
> cat("Annahmebereich: ",n/2 - sqrt(n), n/2 + sqrt(n),"\n")
> k <- sum(diff > mo)                                      # Teststatistik K
> cat("Teststatistik: ",k,"\n")
> cat("P-Wert: ",pbinom(k,n,0.5)*2,"\n")
>
> t.test(mercury[,"Select"],mercury[,"Perman"],paired=?)
      # Setze paired=T für gepaarte Stichproben oder paired=F für ungepaarte Stichproben
> qt(0.975,ν)                                                   # t_{ν,0.975}
```

(Quelle: Kacprzak, J., and Chvojka,R. (1976). Determination of methyl mercury in fish by flameless atomic absorption spectroscopy and comparison with an acid digestion method for total mercury. *J. Assoc. Offic. Anal. Chem., 59*, 153-157.)

4. Die unten angegebenen Daten zeigen die (skalierte) Konzentration einer bestimmten chemischen Substanz in 10 geschnittenen Puffbohnenpflänzchen und in 10 Pflänzchen mit Wurzeln.

Daten:	geschnittene Pflanzen	53	58	48	18	55	42	50	47	51	45
	Pflanzen mit Wurzeln	36	33	40	43	25	38	41	46	34	29
Kennzahlen:	geschnittene Pflanzen	$\overline{x} = 46.7,\ s_x = 11.14$									
	Pflanzen mit Wurzeln	$\overline{y} = 36.5,\ s_y = 6.45$									

a) Handelt es sich um verbundene (gepaarte) oder um unverbundene Stichproben?

b) Stelle die Daten grafisch dar. Gibt es von Auge einen signifikanten Unterschied zwischen den beiden mittleren Konzentrationen?

c) Führe einen t-Test durch, um zu sehen, ob die beobachtete Differenz der Mittelwerte sich auf dem 5%-Niveau signifikant von 0 unterscheidet.
Gib auch die Nullhypothese H_0, die Alternative H_A, die Teststatistik sowie den Verwerfungsbereich unter der Nullhypothese H_0 an.

d) Führe den Wilcoxon-Test durch und gib auch die Nullhypothese H_0, die Alternative H_A, die Teststatistik sowie den Verwerfungsbereich unter der Nullhypothese H_0 an.

e) Der Computer liefert in c) den P-Wert 0.022 und in d) den P-Wert 0.0039. Interpretiere die beiden P-Werte. Welches weitere Vorgehen wäre naheliegend?

5. Die folgenden Daten stammen aus einem Experiment mit behandeltem und unbehandeltem Hafersaatgut. Das Saatgut wurde in drei Gruppen aufgeteilt. Die Gruppen 1 und 2 wurden separat mit demselben Wirkstoff gebeizt. Ein Teil blieb unbehandelt (Gruppe Check). Anschliessend wurden die Samen in je 7 Töpfen pro Gruppe zum Keimen gebracht. Am Ende des Versuchs wurde der Ertrag pro Topf [in Gramm] gemessen. (Wie die Töpfe im Gewächshaus angeordnet waren, lässt sich leider nicht mehr eruieren.)

	Treatment		
Replicate	Group 1	Group 2	Check
1	360	391	408
2	436	382	409
3	413	414	340
4	353	416	324
5	328	375	304
6	269	422	268
7	220	227	290

a) Stelle die Daten zuerst graphisch dar. Zeichne auch das arithmetische Mittel und den Median jeder Gruppe. Zwischen welchen Gruppen wären am ehesten (wenn überhaupt) signifikante Unterschiede zu erwarten?

b) Um die Daten zu analysieren verwenden wir jetzt das Varianzanalysemodell

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}\,, \quad \varepsilon_{ij} \text{ i.i.d. } \sim \mathcal{N}(0, {\sigma_\varepsilon}^2).$$

Vergleiche die drei Gruppen mit einer Einweg-Varianzanalyse (mit Treatment als Faktor). Führe den zugehörigen F-Test durch.
Kennzahlen: $SS_G = 6831.24$, $SS_E = 80743.71$
Gibt es signifikante Unterschiede auf dem 5%-Niveau? Lohnt sich das Beizen des Saatgutes?

c) (Freiwillig) Vermag das benutzte statistische Modell die gesamte Information aus den Daten zu schöpfen? Betrachte dazu die Replikate als Stufen eines zweiten Faktors B, mit Effekten β_j (und die Treatments als Stufen eines erstens Faktors A, mit Effekten α_i). Das Modell ohne Wechselwirkungen lautet jetzt

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}\,, \quad \sum_i \alpha_i = \sum_j \beta_j = 0\,, \quad \varepsilon_{ijk} \text{ i.i.d. } \sim \mathcal{N}(0, {\sigma_\varepsilon}^2).$$

Führe eine zweifaktorielle Varianzanalyse durch.
Kennzahlen: $SS_A = 6831.24$, $SS_B = 55615.62$, $SS_{AB} = SS_E = 25128.1$.
Liegt eventuell noch ein Schreibfehler vor?

(Quelle: Ostle, B. and R.W Mensing (1975), *Statistics in Research*, Iowa State University Press, Ames.)

Kurztabelle der kritischen Werte zum 5%-Niveau der F-Verteilungen mit ν_1 Freiheitsgraden im Zähler und ν_2 Freiheitsgraden im Nenner.

ν_2 \ ν_1 =	1	2	3	4	5	6	7	8	9	10	11	12
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	242.98	243.91
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.40	19.41
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76	8.74
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.94	5.91
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.70	4.68
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.60	3.57
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.31	3.28
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.10	3.07
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.94	2.91
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.82	2.79
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.72	2.69
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.63	2.60
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.57	2.53
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.51	2.48
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.46	2.42
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.41	2.38
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.34	2.31
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.31	2.28

6. Wir möchten untersuchen, ob erhöhter Ozongehalt in der Atmosphäre (a), grosse Feuchtigkeit im Boden (b) und erhöhter CO-Gehalt in der Atmosphäre (c) einen Einfluss auf die Netto-Photosynthese einer Fichte haben. Wir nehmen an, dass sich mögliche Effekte additiv überlagern.

Zwei verschiedene Versuchsanordnungen mit jeweils 4 Einzelversuchen stehen zur Auswahl. Die Versuche werden in einer Klimakammer durchgeführt. Als Zielgrösse X_i wird jeweils die Netto-Photosynthese an einem Fichtenzweig gemessen.

	Anordnung I		**Anordnung II**	
Versuch	Behandlung	Modell	Behandlung	Modell
1	keine	$X_1 = \mu + \varepsilon_1$	mit (a), (b) und (c)	$X_1 = \mu + a + b + c + \varepsilon_1$
2	mit (a)	$X_2 = \mu + a + \varepsilon_2$	mit (a)	$X_2 = \mu + a + \varepsilon_2$
3	mit (b)	$X_3 = \mu + b + \varepsilon_3$	mit (b)	$X_3 = \mu + b + \varepsilon_3$
4	mit (c)	$X_4 = \mu + c + \varepsilon_4$	mit (c)	$X_4 = \mu + c + \varepsilon_4$

Die Beobachtungsfehler ε_i seien unabhängig, identisch verteilt mit Erwartungswert 0 und Varianz σ^2.

a) Versuche die beiden Versuchspläne 3-dimensional als Ecken von zwei perspektivisch gezeichneten Würfeln darzustellen (3 Faktoren, je zwei Niveaus: „hoch“, „tief“).

b) Gib eine naheliegende („naive“) Schätzung für μ, a, b und c an, falls die Versuchsanordnung I verwendet wird. Bestimme die Erwartungswerte und die Varianzen der Schätzer.

c) Führe Teilaufgabe b) mit der Versuchsanordnung II durch. (Hinweis: Versuche a oder ein Vielfaches davon durch Addition (mit geeigneten Vorzeichen) aller vier Beobachtungen zu isolieren. Gehe analog für b, c und μ vor.)

d) Welchen Einfluss haben die Versuchspläne auf Erwartungswerte und Varianzen der Schätzer?

7. (Freiwillig) Im Rahmen einer grossangelegten Umweltstudie wurde der Benzinverbrauch verschiedener Autotypen in verschiedenen Situationen untersucht: In drei Städten der USA (Variable `STADT`, 1=Los Angeles, 2=San Francisco, 3=Portland) wurde der Benzinverbrauch von 5 Automobil-Typen (Variable `AUTO`, Werte von 1 bis 5) ermittelt. Pro Kombination wurden 3 Testfahrten durchgeführt. Die Zielgrösse (Variable `KMP4L`) ist die Strecke in km, welche mit 4 Litern Benzin zurückgelegt werden konnte. Die Daten befinden sich im File `automob.dat`.

a) Stelle die Daten graphisch dar (wenn möglich mit unterschiedlichen Symbolen und Farben).

R-Anleitung:

```
> automob <-
   read.table.url("http://stat.ethz.ch/Teaching/Datasets/NDK/automob.dat")
> attach(automob)
> plot(as.numeric(AUTO), KMP4L, type='n')
                                   # Mit type='n' werden nur die Achsen gezeichnet
> legend(1, 42, c('Los Angeles','San Francisco','Portland'), pch='+o*',
      col=c(1,2,3))                           # Schreibt die Legende, siehe Help
> text(as.numeric(AUTO), KMP4L, labels=c('+','o','*')[STADT], col=STADT)
             # Mit labels wird Stadt=1 mit '+', Stadt=2 mit 'o' und Stadt=3 mit '*' bezeichnet.
               # text zeichnet das Symbol von STADT[i] an die Stelle (AUTO[i], KMP4L[i]).
```

b) Analysiere die Zielgrösse `KMP4L` mit einer zweifaktoriellen Varianzanalyse. Verwende das volle Modell mit Interaktion. Welche Effekte sind signifikant?

R-Anleitung:

```
> auto.aov <- aov(KMP4L ~ STADT*AUTO)
> summary(auto.aov)                                    # Varianzanalysetabelle
```

c) Führe für jede Stadt eine einfaktorielle Varianzanalyse durch.

R-Anleitung:

```
> auto1.aov <- aov(KMP4L~AUTO, subset=STADT==1)   # Varianzanalyse für Stadt 1
> summary(auto1.aov)                                   # Varianzanalysetabelle
> ...
```

d) Wiederhole Aufgabe b) ohne die Werte von San Francisco. Welche Effekte sind jetzt signifikant?

R-Anleitung:

```
> auto13.aov <- aov(KMP4L ~ STADT*AUTO,subset=!(STADT==2))
> summary(auto13.aov)
> detach('automob')
```

(Quelle: Dixon & Massey, *Introduction to statistical analysis*, 1983, p 191)

Abgabe: Montag, 12. Juni, in der Pause der Vorlesungsstunde