Naam: *Richting*:

Examen G0N34 Statistiek

8 september 2010

Enkele richtlijnen :

- Wie de vragen aanneemt en bekijkt, moet minstens 1 uur blijven zitten.
- Je mag gebruik maken van een rekenmachine, het formularium en statistische tabellen om dit examen op te lossen. Op het formularium en de tabellen mag niets geschreven worden! Berekeningen moeten altijd schriftelijk uitgevoerd worden tot het moment dat je de waarde zou kunnen opzoeken in een statistische tabel. Bijvoorbeeld: het uitrekenen van een kans onder een normale verdeling moet herleid worden tot een kans onder een standaardnormale verdeling. Een binomiale kans moet herleid worden tot een kans onder een normale verdeling (indien de CLS van toepassing is).
- Elk type rekentoestel is toegelaten, maar het geheugen moet gewist zijn. Alle communicatie-apparatuur is strikt verboden.
- Gebruik de voorziene ruimte om te antwoorden op de vragen. Tenzij anders vermeld, kan je telkens vóór- en achterkant van een blad gebruiken.
- Je hebt 3.5 uur tijd om het examen op te lossen.
- Schrijf op elk blad je naam!
- Bij het indienen van je examen, geef je ook kladpapier, formularium en tabellen af.

VEEL SUCCES !

Vraag 1

Beoordeel de volgende uitspraken. Als een uitspraak niet juist is of onvolledig, leg dan uit waarom en verbeter de uitspraak.

1. Stel dat X en Y twee toevalsvariabelen zijn die normaal verdeeld zijn. We nemen een steekproef (x_i, y_i) van omvang 100 en berekenen de Pearson correlatiecoëfficiënt. Deze is gelijk aan 0.02. Omdat dit kleiner is dan 0.05, mogen we besluiten dat op het 5% significantieniveau X en Y ongecorreleerd zijn.

2. We beschikken over twee steekproeven. Beiden bevatten 120 metingen van het lichaamsgewicht van volwassen vrouwen. Omdat de steekproefgroottes dezelfde zijn, zijn deze gegevens gepaard.

3. Wanneer we een lineaire regressie willen uitvoeren, moeten we nagaan of de verklarende variabele X en de responsvariabele Y normaal verdeeld zijn.

Vraag 2

Gegeven een reële toevalsvariabele X en een niet-lineaire afleidbare reële functie g. Toon volledig aan hoe je de verwachtingswaarde en de variantie van $g(X)$ kan benaderen a.d.h.v. de verwachtingswaarde en de variantie van X.

Vraag 3

Een onderzoekscentrum in de US wil onderzoeken of er een afhankelijkheid bestaat tussen het niveau van opleiding en religieuse overtuiging. Ze nemen een willekeurige steekproef van de Amerikaanse bevolking en bekomen de volgende aantallen:

		Religieuse Overtuiging			
		Fundamentalist	Gematigd	Vrijdenker	Totaal
Opleiding	Lager dan middelbaar	178	138	108	424
	Middelbaar	570	648	442	1660
	Universitair	138	252	252	642
	Totaal	886	1038	802	2726

1. Mag je op basis van deze gegevens besluiten dat de Amerikaanse bevolking uit meer Fundamentalisten bestaat dan uit Vrijdenkers?

2. Mag je op basis van deze gegevens besluiten dat er een afhankelijkheid bestaat tussen het niveau van opleiding en religieuse overtuiging?

Vraag 4

Gegeven Y_i $(i = 1, \ldots, n)$ n onafhankelijke variabelen die lognormaal verdeeld zijn met parameters μ_i en σ_i^2 (voor $i = 1, \ldots, n$).

1. Bepaal de verdeling van $\prod_{i=1}^{n} Y_i = Y_1 Y_2 \cdots Y_n$.
2. Toon aan dat de dichtheid van Y_i wordt gegeven door

$$f_{Y_i}(y) = \frac{1}{y\sigma_i\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\ln(y)-\mu_i}{\sigma_i}\right)^2} \qquad \text{voor } y > 0$$

3. Veronderstel dat $\sigma_i = \sigma$ en $\mu_i = \mu$ voor alle i. Bereken de maximum likelihoodschatters voor μ en σ^2.

Vraag 5

Gegeven een toevalsvariabele X met continue cumulatieve verdelingsfunctie F_X.

1. Toon aan dat $U = F_X(X)$ een uniforme verdeling heeft op $[0, 1]$. (hint: bereken de cumulatieve verdelingsfunctie van U)
2. Stel $Z = F_X^{-1}(U)$, met $U \sim \text{Uniform}[0, 1]$. Toon aan dat de cumulatieve verdelingsfunctie van Z gelijk is aan de cumulatieve verdelingsfunctie van X.

Vraag 6

Toon aan dat in enkelvoudige regressie $R^2 = r^2$, met R^2 de determinatiecoëfficiënt, en r de Pearson correlatie tussen de onafhankelijke en de afhankelijke variabele.

Vraag 7

Men wil onderzoeken of het gemiddeld aantal wittebloedcellen in het plasma van rokende donoren significant lager is dan bij niet rokende donoren. Men meet daarom het aantal witte bloedcellen in het plasma van 15 rokers en 12 niet-rokers. De resulterende data set bevat twee variabelen:

- **witte blcellen :** het aantal witte bloedcellen in plasma (decimaal getal $\times 10^9$).
- **groep :** 'Roker' of 'Geen Roker'.

We analyseren de gegevens m.b.v. R. Maak gebruik van de bijgevoegde output om de volgende vragen te beantwoorden.

1. Formuleer de nulhypothese en de alternatieve hypothese van de gestelde onderzoeksvraag.

2. Welke teststatistiek ga je gebruiken? Ga na of alle veronderstellingen die je hiervoor aanneemt ook voldaan zijn.

3. Welke waarde neemt de teststatistiek aan? Wat is de p-waarde?

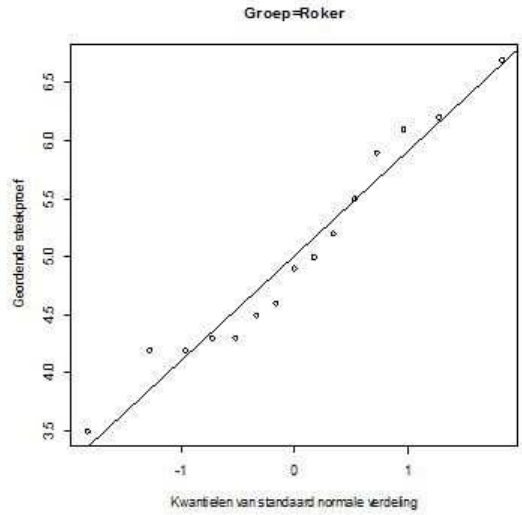
4. Maak een schets waarop je de p-waarde aanduidt.

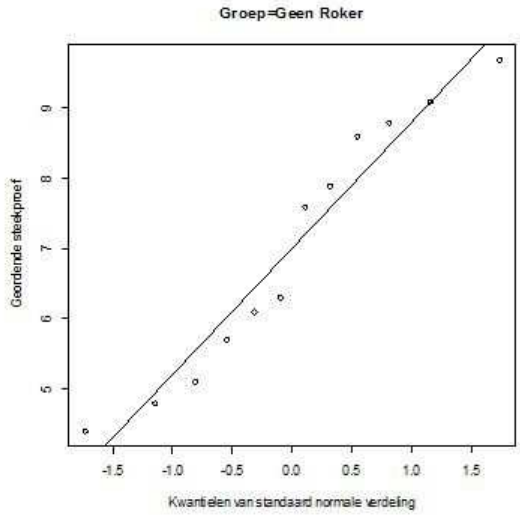
5. Wat is de betekenis van deze p-waarde?

6. Formuleer je besluit.

7. Wat is de betekenis van een type II fout bij deze hypothesetest?

```
> summary(witte_blcellen[groep=="Roker"])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.500   4.300   4.900   5.007   5.700   6.700

> summary(witte_blcellen[groep=="Geen Roker"])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.400   5.550   6.950   7.008   8.650   9.700

> shapiro.test(witte_blcellen[groep=="Roker"])

        Shapiro-Wilk normality test

data:  witte_blcellen[groep == "Roker"]
W = 0.9556, p-value = 0.617

> shapiro.test(witte_blcellen[groep=="Geen Roker"])

        Shapiro-Wilk normality test

data:  witte_blcellen[groep == "Geen Roker"]
W = 0.9332, p-value = 0.4157
```






```
> var.test(witte_blcellen[groep=="Roker"], witte_blcellen[groep=="Geen Roker"])

        F test to compare two variances

data:  witte_blcellen[groep == "Roker"] and witte_blcellen[groep == "Geen Roker"]
F = 0.2466, num df = 14, denom df = 11, p-value = 0.01616
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.07340476 0.76297929
sample estimates:
ratio of variances
         0.2465526


> t.test(witte_blcellen[groep=="Roker"], witte_blcellen[groep=="Geen Roker"],
 var.equal=TRUE)

        Two Sample t-test

data:  witte_blcellen[groep == "Roker"] and witte_blcellen[groep == "Geen Roker"]
t = -3.713, df = 25, p-value = 0.001031
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.1119681 -0.8913652
sample estimates:
mean of x mean of y
 5.006667  7.008333
```

```
> t.test(witte_blcellen[groep=="Roker"], witte_blcellen[groep=="Geen Roker"],
 var.equal=FALSE)

        Welch Two Sample t-test

data:  witte_blcellen[groep == "Roker"] and witte_blcellen[groep == "Geen Roker"]
t = -3.4614, df = 15.3, p-value = 0.003402
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.232135 -0.771198
sample estimates:
mean of x mean of y
 5.006667  7.008333

> wilcox.test(witte_blcellen[groep=="Roker"], witte_blcellen[groep=="Geen Roker"])

        Wilcoxon rank sum test with continuity correction

data:  witte_blcellen[groep == "Roker"] and witte_blcellen[groep == "Geen Roker"]
W = 31.5, p-value = 0.004634
alternative hypothesis: true location shift is not equal to 0
```