

Naam:

Richting:

Examen G0N34 Statistiek

7 juni 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. Bij het testen van hypothesen is de kans op een type II fout altijd groter dan de kans op een type I fout.
2. Om te testen of de correlatie tussen twee continue variabelen significant is, moeten deze variabelen normaal verdeeld zijn.
3. Een kwantielplot is een krachtige grafische methode om normaliteit van een steekproef na te gaan.

Vraag 2

1. Waarvoor dient een Average Shifted Histogram?
2. Leg beknopt en hoofdzakelijk in woorden uit hoe dit ASH bekomen wordt voor een willekeurige steekproef. Gebruik maximaal de voorzijde van deze pagina.

Vraag 3

Uit een onderzoek door het Verbond der Vlaamse Tandartsen is gebleken dat wie vaak snoept 80% kans heeft op cariës. Bij mensen die nooit snoepen bedraagt deze kans 19%. Tevens is geweten dat 15% van de bevolking nooit snoept en dat 65% van de mensen die snoepen, slechts af en toe snoepen. Bij deze laatste categorie bedraagt de kans op cariës 55%.

Indien een tandarts bij een patiënt cariës vaststelt, wat is dan de kans dat deze patiënt snoept?

Vraag 4

Gegeven zijn twee toevalsvariabelen X en Y met gezamenlijke dichtheidsfunctie (voor alle $x, y \in \mathbb{R}$):

$$f_{X,Y}(x,y) = \frac{1}{6\pi\sqrt{3}} \exp\left(-\frac{2}{3} \left[\left(\frac{x-1}{2}\right)^2 + \left(\frac{y}{3}\right)^2 + \frac{(x-1)y}{6} \right]\right)$$

1. Bepaal de (Pearson) correlatie tussen X en Y .
2. Bereken de voorwaardelijke dichtheid van X in het punt $x = 2$, gegeven dat $Y = 1$, i.e. $f_{X|Y}(2|1)$.

Vraag 5

Onderstaande tabel toont ons informatie over het jaarlijks inkomen (in 1000\$) van Amerikaanse statistici die tewerkgesteld zijn in de private sector, volgens hun diploma (bachelor, master, doctoraat).

		Percentiel				
Diploma	n	10	25	50	75	90
bachelor	25	52	65	80	127	158
master	514	80	95	115	138	169
doctoraat	642	97	115	140	175	219

1. Wat kan je afleiden over de verdeling van de variabele Inkomen (per diploma)?
2. Loont het de moeite om te doctoreren? Test of het mediaan inkomen van mensen met een doctoraat significant hoger is dan 115 (wat in deze data set overeenkomt met het mediaan inkomen van statistici met een master diploma.)

3. Veronderstel dat je beschikt over alle gegevens waarop de voorgaande tabel gebaseerd is (dus de inkomens van alle mensen die aan de studie deelnamen). Je wil testen of het mediaan inkomen van Amerikaanse statistici met een doctoraat significant hoger is dan het mediaan inkomen van Amerikaanse statistici met een master diploma. Leg zo volledig mogelijk uit hoe je tewerk zou gaan om deze hypothese te testen.

Vraag 6

Gegeven zijn $X_1, X_2, \dots, X_n$ onafhankelijke toevalsvariabelen die Bernoulli verdeeld zijn met kans op succes p .

1. Toon aan dat

$$f_{X_i}(x) = p^x(1-p)^{(1-x)} \text{ voor } x = 0, 1.$$

2. Toon aan dat de maximum likelihoodschatter van p gegeven wordt door $\bar{X}_n$.
3. Men wil een schatter opstellen voor $\theta = \text{Var}(X_i)$ en gebruikt hiervoor $\hat{\theta} = \bar{X}_n(1 - \bar{X}_n)$. Is dit een zuivere schatter voor $\text{Var}(X_i)$?
4. Bepaal een benadering voor $\text{Var}(\hat{\theta})$ via de Delta methode.

Vraag 7

In België werd een onderzoek uitgevoerd naar de vraag of de opbrengst van witloof beïnvloed wordt door de schade toegebracht door de witloofmineervlieg. Voor 30 verschillende proefvelden werd de hoeveelheid geplukt witloof per week (in kg) gemeten (variabele **opbrengst**) alsook het percentage aan planten dat door het insect beschadigd werd (variabele **schade**).

In R worden de twee variabelen bestudeerd, en er wordt een regressie analyse uitgevoerd. Dit levert de volgende output:

```
> summary(schade)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 8.00  13.00   15.50   19.87  26.50   47.00
```

```
> summary(opbrengst)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
34.28  37.59   38.81   39.09  40.62   43.38
```

```
> shapiro.test(schade)
      Shapiro-Wilk normality test
```

```
data:  schade W = 0.8922, p-value = 0.005436
```

```
> summary(opbrengst)
      Shapiro-Wilk normality test
```

```
data:  opbrengst W = 0.9786, p-value = 0.7868
```

```
> Results <- lm(opbrengst ~ schade)
> summary(Results)
Call: lm(formula = opbrengst ~ schade)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-2.96542 -1.63783  0.01662  1.11708  3.55421
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 42.12987    0.74161      ?      ?
schade      -0.15296    0.03322      ?      ?
```

```
Residual standard error:  ? on  ? degrees of freedom
Multiple R-squared:  ?,      Adjusted R-squared: 0.4106
F-statistic: 21.21 on 1 and 28 DF,  p-value: 8.161e-05
```

```
> anova(Results)
```

```
Analysis of Variance Table
```

```
Response: opbrengst
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
schade	1	72.848	?	?	?
Residuals	28	96.188	?	?	?

Bijhorende figuren staan achteraan.

1. Vul de ontbrekende gegevens in de tabellen in (op de 13 plaatsen waar een vraagteken staat). Je mag de ontbrekende waarden rechtstreeks invullen in de tabellen. Bewerkingen om deze waarden te bekomen mag je hieronder schrijven.

2. Is de regressie zinvol? Leg volledig uit a.d.h.v. een geschikte hypothesetest.

3. Maak een figuur waarop je de p-waarde aangeeft die je in voorgaande test krijgt.

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

5. Construeer een 95% betrouwbaarheidsinterval voor de richtingscoëfficiënt van de regressierechte. Geef de betekenis van dit betrouwbaarheidsinterval.