G0N34a Statistiek: Examen 7 juni 2010 (review)

1 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 hypotheses is de kans op een type-II fout altijd groter dan de kans op een type-I fout.

   → **Fout!**

   - Kans op type-II fout hangt af van het significantie niveau α (kans op type-I fout).
   - Kans op type-II fout hangt af van het alternatief μ_1 (i.e. de *echte* waarde van μ). Hoe verder de vooropgestelde waarde μ_0 ligt van μ_1, hoe kleiner de kans op een type-II fout.

2. Om te testen of de correlatie tussen continue variabelen significant is, moeten deze variabelen normaal verdeeld zijn.

   → **Onvolledig!**

   - **Lineair verband**: via de Pearson-correlatiecoëfficiënt.
     Voorwaarde: variabelen zijn bivariaat normaal verdeeld.
   - **Monotoon verband**: via de Spearman-rangcorrelatie.
     Voorwaarde: geen. Niet parametrische test!

3. Een kwantielplot is een krachtige grafische methode om normaliteit van een steekproef na te gaan.

   → **Onvolledig!**

   - Een **normale** kwantielplot ...
     Vb. via een exponentiële kwantielplot kan je normaliteit natuurlijk niet nagaan, maar wel of de steekproef uit een exponentiële verdeling komt.
   - Kwantielplots geven enkel een indicatie. Moet nog formeel getest worden via een hypothesetest.

2 Vraag 2

1. Waarvoor dient een Average Shifted Histogram?

   → Schatting van de onderliggende dichtheidsfunctie van een steekproef.

2. Leg beknopt en hoofdzakelijk in woorden uit hoe dit ASH wordt bekomen voor een willekeurige steekproef.

   → Zie .pdf omtrent ASH.

3 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?

Oplossing:

	Bevolking			Kans op cariës (C)
Snoepen (S)	85%	↗ Af en toe (A):	× 65%	55%
		↘ Vaak (V):	× 35%	80%
Niet snoepen (S^c)	15%			19%

Noteer:
S : patiënt snoept $\Rightarrow$ S^c : patiënt snoept niet.
C : patiënt heeft cariës.

We zoeken: kans dat patiënt snoept, indien hij cariës heeft $= \mathrm{P}[S\,|C]$. Aangezien we $\mathrm{P}[C\,|S] = \mathrm{P}[C\,|A]\,\mathrm{P}[A] + \mathrm{P}[C\,|V]\,\mathrm{P}[V]$ kunnen berekenen uit het gegeven, is het nuttig om via de regel van Bayes te werken. Dus

$$
\begin{aligned}
\mathrm{P}[S\,|C] &\overset{Bayes}{=} \frac{\mathrm{P}[C\,|S]\cdot\mathrm{P}[S]}{\mathrm{P}[C]} \\
&\overset{WTK}{=} \frac{\mathrm{P}[C\,|S]\cdot\mathrm{P}[S]}{\mathrm{P}[C\,|S]\cdot\mathrm{P}[S] + \mathrm{P}[C\,|S^c]\cdot\mathrm{P}[S^c]} \\
&\overset{WTK}{=} \frac{(\mathrm{P}[C\,|A]\,\mathrm{P}[A] + \mathrm{P}[C\,|V]\,\mathrm{P}[V])\cdot\mathrm{P}[S]}{(\mathrm{P}[C\,|A]\,\mathrm{P}[A] + \mathrm{P}[C\,|V]\,\mathrm{P}[V])\cdot\mathrm{P}[S] + \mathrm{P}[C\,|S^c]\cdot\mathrm{P}[S^c]} \\
&= \frac{(0.55\cdot 0.85 \times 0.65 + 0.80\cdot 0.85 \times 0.35)\cdot 0.85}{(0.55\cdot 0.85 \times 0.65 + 0.80\cdot 0.85 \times 0.35)\cdot 0.85 + 0.19\cdot 0.15} \\
&\approx 0.9417.
\end{aligned}
$$

Een kansboom had hier ook gekund.

4 Vraag 4

Gegeven is de gezamenlijke dichtheidsfunctie $f_{X,Y}$ van de bivariate stochastische vector (X, Y) met (Pearson) correlatiecoëfficiënt ρ:

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

1. Bepaal ρ.
2. Veronderstel dat X en Y Normaal verdeeld zijn. Bereken vervolgens de voorwaardelijke dichtheid van X in het punt $x = 2$, gegeven dat $Y = 1$, i.e. $f_X(2\,|Y=1)$.

Oplossing:

Zie slide 21 *Multivariate kansmodellen*: de dichtheidsfunctie van de bivariate normale verdeling heeft de vorm

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sqrt{\det(\mathbf{\Sigma})}} e^{-\frac{1}{2}\mathbf{z}^t\mathbf{\Sigma}^{-1}\mathbf{z}}, \tag{2}$$

met $\mathbf{z} = (x-\mu_X, y-\mu_Y)^t$ en $\mathbf{\Sigma} = \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix} = \begin{pmatrix} \sigma_X^2 & \rho_{XY}\sigma_X\sigma_Y \\ \rho_{XY}\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}$, vermits $\sigma_{XY} = \mathrm{Cov}[X,Y] = \rho_{XY}\sigma_X\sigma_Y$.

Wanneer we het rechterlid van (2) volledig uitwerken, na substitutie van $\mathbf{z}$ en $\mathbf{\Sigma}$, dan vinden we

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho_{XY}^2}} e^{-\frac{1}{2\left(1-\rho_{XY}^2\right)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2 - \frac{2\rho_{XY}(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}\right]}. \tag{3}$$

Wanneer we formule (3) vergelijken met de opgave (1) dan zien we onmiddellijk dat

- $\mu_X = 1$
- $\sigma_X = 2$
- $\mu_Y = 0$
- $\sigma_Y = 3$
- $2\pi\sigma_X\sigma_Y\sqrt{1-\rho_{XY}^2} = 6\pi\sqrt{3} \Rightarrow \rho_{XY}^2 = \frac{1}{4} \Rightarrow \rho_{XY} = \pm\frac{1}{2}$.

En vermits moet gelden dat $-\frac{2\rho_{XY}(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} = +\frac{(x-1)y}{6}$, volgt dat $\rho = -\frac{1}{2}$.

Om de voorwaardelijke dichtheid $f_{X|Y}\,(2\,|1\,)$ te berekenen, kan je gebruik maken van de eigenschap: $f_{X|Y}\,(x\,|y\,) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$. Bovendien, vermits de gezamenlijke dichtheid (1) uit de opgave overeenkomt met de dichtheidsfunctie van de bivariate normale verdeling met parameters $\mu_X = 1, \sigma_X = 2, \mu_Y = 0, \sigma_Y = 3$ en $\rho_{XY} = -1/2$, mogen we aannemen dat de vector (X, Y) uit de opgave bivariaat normaal verdeeld is en bijgevolg dat de variabelen X en Y normaal verdeeld zijn. Dit impliceert dat

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma_Y} e^{-\frac{1}{2}\left(\frac{y-\mu_Y}{\sigma_Y}\right)^2}.$$

Dus:

$$f_{X|Y}\,(2\,|1\,) = \frac{\frac{1}{6\pi\sqrt{3}}\exp\left(-\frac{2}{3}\left[\left(\frac{2-1}{2}\right)^2 + \left(\frac{1}{3}\right)^2 + \frac{(2-1)1}{6}\right]\right)}{\frac{1}{\sqrt{2\pi}3} e^{-\frac{1}{2}\left(\frac{1-0}{3}\right)^2}} \approx 0.1713$$

5 Vraag 5

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

Diploma	n	Percentiel				
		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 X = *Jaarlijks inkomen* (per diploma)?

   → In elk van de 3 gevallen is de variabele *Jaarlijks inkomen* rechtsscheef verdeeld. Dit kan je o.a. afleiden uit het feit dat $p_{90} - p_{50} > p_{50} - p_{10}$, in elk van de gevallen; hierbij is p_x het x-percentiel.

   Hoe groter de afstand tussen het 10-percentiel en het 50-percentiel (de mediaan), hoe zwaarder de linker staart. Hoe groter de afstand tussen de mediaan en het 90-percentiel, hoe zwaarder de rechter staart. In ons geval is de rechter staart duidelijk systematisch zwaarder dan de linker (vb. $158 - 80 > 80 - 52$), wat aangeeft dat de verdeling rechtsscheef is.

2. Loont het de moeite om te doctoreren? Test of de mediaan van het *Jaarlijks inkomen* van mensen met een doctoraat (PhD) significant hoger is dan 115, i.e. de mediaan van het *Jaarlijks inkomen* van mensen met een master diploma.

   → We testen

$$\begin{aligned} H_0 &: med(X_{\text{PhD}}) \leq 115; \\ H_1 &: med(X_{\text{PhD}}) > 115. \end{aligned}$$

   Mediaantest: de hypothese H_0 is plausibel indien de steekproefmediaan niet *teveel groter is dan 115*, of m.a.w. **indien het aantal PhDs met een *Jaarlijks inkomen* $\geq$ 115 duizend \$ niet *teveel kleiner* is dan 642/2 = 321**.

   zij A = aantal PhDs met een *Jaarlijks inkomen* $\geq$ 115 duizend \$. Merk op dat in het geval van PhDs een inkomen van 115 duizend \$ overeenkomt met het 25 percentiel. M.a.w. 1/4 van alle PhDs verdiend minstens 115 duizend \$ $\Rightarrow A = 642/4 = 160.5 \rightarrow 160$.

Onder H_0 geldt dat $A \sim \text{Bin}\,(n = 642, p = 0.5)$. De p-waarde, i.e. de kans -onder H_0- dat de teststatistiek A nog extremer is dan de experimentele waarde (160) -in de richting (*teveel kleiner:* $\leq$) van het alternatief- bedraagt dan:

$$\begin{aligned} p &= \text{P}\,[A \leq 160] \\ &\overset{CLS}{\approx} \text{P}\,[Y \leq 160 + 0.5]\,, \text{ met } Y \sim \text{N}\,(\mu = np = 321, \sigma = np(1-p) = 160.5) \\ &\approx 0, \end{aligned}$$

dus we verwerpen H_0, i.e. het loont de moeite om te doctoreren.

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 te werk zou gaan om deze hypothese te testen.

   $\rightarrow$ We willen testen

$$\begin{aligned} H_0 &: med(X_{\text{PhD}}) \leq med(X_{\text{Master}}); \\ H_1 &: med(X_{\text{PhD}}) > med(X_{\text{Master}}). \end{aligned}$$

   Mogelijkheden:

   (a) via transformatie:
   - transformatie tot normaliteit (data zijn rechtsscheef, dus mogelijks log-normaal verdeeld);
   - test op verschil in gemiddeldes van de getransformeerde data (via t-test voor ongepaarde gegevens, dus test eerst op gelijkheid van de varianties);
   - formuleer een besluit m.b.t. de medianen van de oorspronkelijke data (indien de data log-normaal zouden zijn, dan geldt dat de medianen van de getransformeerde data gelijk zijn aan de gemiddeldes van de getransformeerde data + dat de medianen van de oorspronkelijke data gelijk zijn aan de exp van de medianen van de getransformeerde data).

   (b) via Wilcoxon:
   - merk op dat de gegevens niet uit een normale verdeling komen en je daarom een niet-parametrische tet zal uitvoeren;
   - let erop dat de oorspronkelijke H_0 en H_1 moet worden aangepast naar die voor Wilcoxon;
   - formuleer een besluit omtrent gelijkheid of verschil in verdelingen en leidt daaruit informatie af omtrent de verhouding tussen de medianen van de oorspronkelijke data.

6 Vraag 6

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

1. Toon aan dat $f_{X_i}(x) = p^x(1-p)^{1-x}$, voor $x = 0, 1$.

   $\rightarrow$ 2 manieren:

   (a) Invullen: $f_{X_i}(1) = p = 1 - f_{X_i}(0)$.

   (b) Via de Binomiaal verdeling. Als $X \sim$ Bernoulli(p), dan geldt tevens dat $X \sim$ Binomiaal(1, p). Bijgevolg is $f_X(x) = \binom{1}{x} p^x (1-p)^{1-x}$, voor $x = 0, 1$. Tevens geldt, voor elk willekeurig natuurlijk getal n, dat $\binom{n}{0} = \binom{n}{n} = 1$, zodat $f_X(x) = p^x (1-p)^{1-x}$.

2. Toon aan de de MLE van p wordt gegeven door $\overline{X}_n$.

   $\rightarrow$ De MLE vinden we door de (log-)likelihood functie af te leiden naar de parameter(s) waarvoor we een schatter zoeken. We moeten dus allereerst de (log-)likelihood functie bepalen. In dit geval hebben we

$$L(p; x_1, x_2, ..., x_n) = \prod_{i=1}^{n} f_{X_i}(p; x_i) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i},$$

   zodat

$$logL(p; x_1, x_2, ..., x_n) = \sum_{i=1}^{n} \log\left[f_{X_i}(p; x_i)\right] = \log(p) \sum_{i=1}^{n} x_i + \log(1-p) \left(n - \sum_{i=1}^{n} x_i\right),$$

   dus

$$\frac{\partial}{\partial p} logL(p; x_1, x_2, ..., x_n) = \frac{1}{p} \sum_{i=1}^{n} x_i - \frac{1}{1-p} \left(n - \sum_{i=1}^{n} x_i\right).$$

   de MLE $\hat{p}_{MLE}$ voor p voldoet aan

$$\frac{1}{\hat{p}_{MLE}} \sum_{i=1}^{n} x_i - \frac{1}{1-\hat{p}_{MLE}} \left(n - \sum_{i=1}^{n} x_i\right) = 0,$$

   waaruit volgt dat

$$\hat{p}_{MLE} = \frac{1}{n} \sum_{i=1}^{n} x_i = \overline{X}_n.$$

3. Men wil een schatter opstellen voor $\theta = \text{Var}[X_i]$ en gebruikt hiervoor $\hat{\theta} = \overline{X}_n\left(1 - \overline{X}_n\right)$. Is dit een zuivere schatter voor $\text{Var}[X_i]$?

   $\rightarrow$ We moeten nagaan of $\text{E}\left[\hat{\theta}\right] = \theta = \text{Var}[X_i] = p(1-p)$.

$$\begin{aligned}
\text{E}\left[\hat{\theta}\right] &= \text{E}\left[\overline{X}_n\right] - \text{E}\left[\overline{X}_n^2\right] \\
&= \mu_X - \left(\text{Var}\left[\overline{X}_n\right] + \text{E}\left[\overline{X}_n\right]^2\right) \\
&= \mu_X - \left(\frac{\sigma_X^2}{n} + \mu_X^2\right) \\
&= p - \left(\frac{p(1-p)}{n} + p^2\right) \\
&= p\left[1 - \frac{1-p}{n} - p\right] \\
&= p(1-p)\left(1 - \frac{1}{n}\right) \\
&\neq p(1-p).
\end{aligned}$$

   De voorgestelde schatter is dus niet zuiver (wel asymptotisch, i.e. voor $n \rightarrow \infty$).

4. Bepaal een benadering voor $\text{Var}\left[\hat{\theta}\right]$ via de Delta methode.

   $\rightarrow$ De Delta methode kan worden gebruikt om een **benadering te bepalen voor de verwachte waarde, de variantie, ... van een functie g van een toevalsvariabele**. De techniek bestaat erin om eerst de functie g te ontwikkelen in een Taylorreeks rond een goed gekozen punt en vervolgens de verwachte waarde, variantie, ... te bepalen van de reeksontwikkeling. Voor de verwachte waarde volstaat een ontwikkeling t.e.m. de tweede orde. Voor de variantie volstaat een ontwikkeling t.e.m. de eerste orde.

   In ons geval geldt: $\hat{\theta} = g\left(\overline{X}_n\right) = \overline{X}_n\left(1 - \overline{X}_n\right) = \overline{X}_n - \overline{X}_n^2$. De beschouwde functie is dus $g(x) = x - x^2$, zodat $g'(x) = 1 - 2x$.

   Merk op dat $\hat{\theta}$ een functie is van de variabele $\overline{X}_n$. Het is duidelijk dat $\overline{X}_n$ in de buurt zal liggen van $\text{E}\left[\overline{X}_n\right] = \text{E}[X] = p$, dus zullen we de functie g ontwikkelen rond het punt p. Dit geeft (tot de eerste orde):

$$g(x) \approx g(p) + (x-p)g'(p).$$

Bijgevolg hebben we

$$\begin{aligned}
\operatorname{Var}\left[\hat{\theta}\right] &= \operatorname{Var}\left[g\left(\overline{X}_n\right)\right] \\
&\approx \operatorname{Var}\left[g(p) + \left(\overline{X}_n - p\right) g'(p)\right] \\
&= \operatorname{Var}\left[g(p)\right] + \operatorname{Var}\left[\left(\overline{X}_n - p\right) g'(p)\right] \\
&= 0 + \left(g'(p)\right)^2 \operatorname{Var}\left[\left(\overline{X}_n - p\right)\right] \\
&= \left(g'(p)\right)^2 \operatorname{Var}\left[\overline{X}_n\right] \\
&= (1-2p)^2 \frac{p(1-p)}{n}
\end{aligned}$$