Tentamen Statistische Methoden
MST
10 juli 2015, 14.00–17.00 uur

Normering: MC vraag elk 0.4 punt. Open vragen allebei 2 punten.

1. Er is 3 keer gegooid met een dobbelsteen. De som van de worpen blijkt 14 te zijn, maar het resultaat van de afzonderlijke worpen is onbekend. Bereken de kans dat er geen enkele keer 6 is gegooid.

   a. 0.15 **b.** 0.20 **c.** 0.25 **d.** 0.30 **e.** 0.35 **f.** 0.40

2. De leugendetector wordt vooral in de VS veel ingezet bij het opsporen van misdadigers. We beschrijven dit in kansrekeningtermen: A is de gebeurtenis dat de verdachte schuldig is; B is de gebeurtenis dat de leugendetector de verdachte als schuldig aanwijst. Het is bekend dat $P(B \mid A) = 0.90$ en $P(B \mid A^c) = 0.10$ en $P(A) = 0.25$. Stel dat de leugendetector de verdachte aanwijst als schuldig. Hoe groot is dan de kans dat de verdachte daadwerkelijk schuldig is?

   a. 0.50 **b.** 0.90 **c.** 0.21 **d.** 0.40 **e.** 0.63 **f.** 0.75

3. De stochast X is uniform verdeeld $U(0, 1)$. Bereken de verwachting van X^4.

   a. 0.06 **b.** 0.16 **c.** 0.20 **d.** 0.30 **e.** 0.25 **f.** 0.50

4. Het kookpunt K van water bij een druk van 1 atmosfeer is normaal verdeeld met $\mu = 100$ en $\sigma^2 = 0.01$. Bereken de kans dat het water gaat koken bij een temperatuur ≤ 99.98 graden. Met andere woorden, bereken $P(K \leq 99.98)$ voor een $N(100, 0.01)$ stochast. Rond het antwoord af op twee decimalen:

   a. 0.42 **b.** 0.02 **c.** 0.00 **d.** 1.00 **e.** 0.50 **f.** 0.98

5. Gegeven X een continue stochast (random variable) die waarden aanneemt tussen 0 en 2. De kansdichtheidsfunctie f is gelijk aan:

$$f(x) = \frac{4 + 3x^2}{16}$$

   Bereken $\mathrm{P}(X \leq 1)$:

   a. $\frac{7}{16}$ **b.** $\frac{1}{8}$ **c.** $\frac{3}{16}$ **d.** $\frac{7}{8}$ **e.** $\frac{3}{4}$ **f.** $\frac{5}{16}$

6. De volgende dataset is normaal verdeeld:

4.0694	3.7768	4.8768	1.7955	1.9399
3.4538	−0.2941	2.6504	1.5171	1.6702
1.3931	−0.1377	0.4901	2.6384	3.2554
2.5877	0.3810	4.7406	2.6257	4.1865
0.4254	−3.8886	−1.4230	0.2702	4.2185

   De som van de data is 47.2 en de som van de kwadraten is 194.4 Welke verdeling $N(\mu, \sigma^2)$ past het best bij deze dataset:

   a. $N(2, 4)$ **b.** $N(2, 8)$ **c.** $N(2, 16)$
   d. $N(3, 4)$ **e.** $N(3, 8)$ **f.** $N(3, 16)$

7. De stochasten X en Y hebben beide verwachting $E[X] = E[Y] = 2$ en covariantie $\text{Cov}(X, Y) = 2$. Bereken de verwachting $E[XY]$ van het product van deze stochasten.

a. 0 **b.** 2 **c.** 4 **d.** 6 **e.** 8 **f.** 10

8. Volgens de ongelijkheid van Chebyshev geldt dat

$$P\left(\left|\frac{X-\mu}{\sigma}\right| \geq k\right) \leq \frac{1}{k^2}$$

Stel dat $E[X] = 2$ en $\text{Var}(X) = 5$. Hoe groot is de kans dat X in het interval $(-3, 7)$ ligt volgens de ongelijkheid van Chebyshev?

a. ≤ 0.80 **b.** ≤ 0.75 **c.** ≤ 0.5 **d.** ≥ 0.5 **e.** ≥ 0.75 **f.** ≥ 0.80

9. Het gemiddelde $\bar{x}$ en de mediaan m van een dataset zijn statistische schatters van de verwachting. Welke van de volgende uitspraken is correct?

a. $\bar{x}$ is geen zuivere schatter van de verwachting en m is een zuivere schatter
b. $\bar{x}$ en m zijn geen van beide zuivere schatters van de verwachting
c. $\bar{x}$ is ongevoelig voor outliers en m is gevoelig voor outliers
d. $\bar{x}$ is gevoelig voor outliers en m is ongevoelig voor outliers

10. De Engelsman Tom Feeney is een crowdfundingsactie begonnen voor Griekenland. Dagelijks komt er een bedrag binnen van gemiddeld 96 duizend euro met een standaardafwijking van 4 duizend euro. Neem aan dat het dagelijkse bedrag van X duizend euro normaal verdeeld is met $\mu = 96$ en $\sigma = 4$. Neem ook aan dat de dagelijkse bedragen onafhankelijk zijn van elkaar. Hoe groot is de kans dat Tom Feeney na 100 dagen meer dan 10 miljoen euro heeft binnengehaald? Rond het antwoord af op twee decimalen.

a. 0.00 **b.** 0.01 **c.** 0.02 **d.** 0.03 **e.** 0.05 **f.** 0.10

11. Ik gooi 4 keer met een zuivere munt en tel het aantal keer dat ik Kop gooi. Hoe groot is de kans dat ik precies 2 keer Kop gooi, afgerond op twee decimalen?

a. 0.13 **b.** 0.25 **c.** 0.19 **d.** 0.11 **e.** 0.38 **f.** 0.50

12. Van 93 rokende vrouwen is bijgehouden hoeveel menstruatie cycli voorbijgingen voordat zij zwanger werden. De data staat in de tabel hieronder.

Cycles	1	2	3	4	5	6	7	8	9	10	11	12
Frequency	29	16	17	4	3	9	4	5	1	1	1	3

Source: C.R. Weinberg and B.C. Gladen. The beta-geometric distribution applied to comparative fecundability studies. *Biometrics*, 42(3):547–560, 1986.

Deze data wordt beschreven via een $Geo(p)$ stochast. Als we de waarde van p schatten via de kans $P(X = 1)$, wat is dan de schatting voor p voor deze dataset?

a. 0.27 **b.** 0.36 **c.** 0.21 **d.** 0.43 **e.** 0.31 **f.** 0.39

13. Het 98-procent betrouwbaarheidsinterval van een dataset $x_1, x_2, \ldots, x_{16}$ wordt gegeven door $(0.11, 0.75)$, via de student-t methode. Construeer het 99-procent betrouwbaarheidsinterval.

a. (0.10, 0.76) **b.** (0.09, 0.77) **c.** (0.08, 0.78)
d. (0.07, 0.79) **e.** (0.06, 0.80) **f.** (0.05, 0.81)

14. We voeren een t-test uit met nulhypothese $H_0 : \mu = 10$ met een dataset van 9 elementen met gemiddelde $\bar{x}_9 = 10.28$ en steekproef standaardafwijking $s_9 = 0.21$. De alternatieve hypothese is $\mu > 10$. Het significantieniveau is 0.01. De nulhypothese wordt verworpen indien het gemiddelde $\bar{x}_9$ ligt boven de kritieke waarde c. Hoe groot is c, afgerond op 5 honderste?

a. 10.10 **b.** 10.15 **c.** 10.20 **d.** 10.25 **e.** 10.30 **f.** 10.35

15. Gegeven de dataset

$$1, 1, 1, 1, 2, 3, 3, 3, 3, 5, 5, 5, 5, 6, 100$$

Bepaal de MAD.

a. 1 **b.** 2 **c.** 3 **d.** 5 **e.** 6 **f.** 100

OPEN VRAGEN

---

1. De volgende dataset komt uit een binomiale verdeling: $\text{Bin}(n, p)$ waarbij $n = 10$ en p onbekend is.

0	6	2	1
1	3	0	4
1	2	2	3
0	3	4	0

A. Motiveer waarom $p = 0.2$ een voor de hand liggende schatting is van de parameter p.

B. Er wordt getwijfeld aan de waarde 6 in deze dataset. Het kan een meetfout zijn. Voor een $\text{Bin}(10, 0.2)$ stochast geldt namelijk $P(X \geq 6) = 0.0064$. Bereken de kans dat het maximum van een dataset van 16 getallen groter of gelijk is aan 6.

2. De stochasten X en Y zijn onafhankelijk en $Exp(2)$ verdeeld.

A. Bereken de verwachting van $X + Y^2$.

B. Bereken $P(X + Y > 1)$.

Formula sheet for exams PROBSTAT

HAND IN AFTER THE EXAM!!

NOTE: THIS SHEET IS NOT A SUMMARY OR OVERVIEW, AND ONLY SERVES AS MEANS OF AID.

Probability Distributions

1. **Bernoulli distribution:** $Ber(p)$.
   $\mathrm{P}(X=1)=p$ and $\mathrm{P}(X=0)=1-p$. $\mathrm{E}[X]=p$; $\mathrm{Var}(X)=p(1-p)$.
2. **Binomial distribution:** $Bin(n,p)$.
   $\mathrm{P}(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$ for $k=0,1,\ldots,n$. $\mathrm{E}[X]=np$; $\mathrm{Var}(X)=np(1-p)$.
3. **Geometric distribution :** $Geo(p)$.
   $\mathrm{P}(X=k)=p(1-p)^{k-1}$ for $k=1,2,\ldots$. $\mathrm{E}[X]=1/p$; $\mathrm{Var}(X)=(1-p)/p^2$.
4. **Poisson-distribution:** $Pois(\mu)$.
   $\mathrm{P}(X=k)=\frac{\mu^k}{k!}e^{-\mu}$ for $k=0,1,\ldots$. $\mathrm{E}[X]=\mu$; $\mathrm{Var}(X)=\mu$.
5. **Exponential distribution:** $Exp(\lambda)$.
   $f(x)=\lambda e^{-\lambda x}$ and $F(x)=1-e^{-\lambda x}$ for $x\geq 0$. $\mathrm{E}[X]=1/\lambda$; $\mathrm{Var}(X)=1/\lambda^2$.
6. **Uniform distribution** on $[a,b]$: $U(a,b)$.
   $f(x)=\frac{1}{b-a}$ and $F(x)=\frac{x-a}{b-a}$ for $a\leq x\leq b$. $\mathrm{E}[X]=\frac{1}{2}(a+b)$; $\mathrm{Var}(X)=\frac{1}{12}(b-a)^2$.
7. **Normal distribution:** $N(\mu,\sigma^2)$.
   $f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$ and $F(x)=\int_{-\infty}^{x}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{t-\mu}{\sigma})^2}dt$. $\mathrm{E}[X]=\mu$; $\mathrm{Var}(X)=\sigma^2$.
8. **Cauchy distribution** $Cau(\beta,\alpha)$:
   $f(x)=\frac{\alpha}{\pi(\alpha^2+(x-\beta)^2)}$, $F(x)=\frac{1}{2}+\frac{1}{\pi}\arctan\left(\frac{x-\beta}{\alpha}\right)$. $\mathrm{E}[X]$ and $\mathrm{Var}(X)$ do not exist.
9. **Pareto distribution** $Par(\alpha)$:
   $f(x)=0$ for $x<1$, and $f(x)=\frac{\alpha}{x^{\alpha+1}}$, and $F(x)=1-x^{-\alpha}$ for $x\geq 1$.
   $\mathrm{E}[X]=\infty$ for $0<\alpha\leq 1$, and $\mathrm{E}[X]=\frac{\alpha}{\alpha-1}$ for $\alpha>1$. $\mathrm{Var}(X)=\frac{\alpha}{(\alpha-1)^2(\alpha-2)}$ for $\alpha>2$.

Jensen's inequality

If g is a convex function and X a random variable, then: $g(\mathrm{E}[X])\leq \mathrm{E}[g(X)]$.

Covariance and correlation

1. Definition covariance: $\mathrm{Cov}(X,Y)=\mathrm{E}[(X-\mathrm{E}[X])(Y-\mathrm{E}[Y])]$.
   Properties: $\mathrm{Cov}(X,Y)=\mathrm{E}[XY]-\mathrm{E}[X]\cdot\mathrm{E}[Y]$, $\mathrm{Cov}(rX+s,tY+u)=rt\,\mathrm{Cov}(X,Y)$.
2. Definition correlation coefficient: $\rho(X,Y)=\frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}$.
3. $\mathrm{Var}(X+Y)=\mathrm{Var}(X)+\mathrm{Var}(Y)+2\,\mathrm{Cov}(X,Y)$.

Empirical distribution function

For a dataset of n elements: $F_n(x)=\frac{\text{number of elements in the dataset} \leq x}{n}$.

Law of large numbers and central limit theorem

When $X_1,X_2,\ldots$ is a sequence of independent random variables, all having the same distribution, expectation μ, and variance σ^2, then it holds that (where Φ denotes the distribution function of the $N(0,1)$ distribution):

Law of large numbers:
for any $\varepsilon>0$ it holds that $\lim_{n\to\infty}\mathrm{P}\left(|\bar{X}_n-\mu|>\varepsilon\right)=0$.

Central limit theorem:

$$\lim_{n\to\infty}\mathrm{P}\left(\sqrt{n}\frac{\bar{X}_n-\mu}{\sigma}\leq a\right)=\Phi(a) \text{ and } \lim_{n\to\infty}\mathrm{P}\left(\frac{X_1+\cdots+X_n-n\mu}{\sigma\sqrt{n}}\leq a\right)=\Phi(a).$$

Estimators

The bias of an estimator T for θ: $\mathrm{E}[T]-\theta$.
When T_1 and T_2 are unbiased estimators, then T_2 is more efficient than T_1 if $\mathrm{Var}(T_2)<\mathrm{Var}(T_1)$.
The mean squared error of an estimator T for θ: $\mathrm{MSE}(T)=\mathrm{E}[(T-\theta)^2]$.
Property: $\mathrm{MSE}(T)=\mathrm{Var}(T)+(\mathrm{E}[T]-\theta)^2$.

Bootstrap simulation for $\bar{X}_n-\mu$

Given: a dataset $x_1,x_2,\ldots,x_n$ from a distribution with distribution function F. Determine an estimate $\hat{F}$ of F by means of the data and let μ^* be the expectation corresponding to $\hat{F}$. Repeat the following steps a large number of times:

1. Generate a bootstrap dataset $x_1^*,x_2^*,\ldots,x_n^*$ from $\hat{F}$.
2. Compute the centered sample mean $\bar{x}_n^*-\mu^*$ of the bootstrap dataset.

The empirical bootstrap simulation corresponds with the choice $\hat{F}=F_n$, the parametric bootstrap simulation with the choice $\hat{F}=F_{\hat{\theta}}$.

Confidence intervals

Given: a dataset $x_1,x_2,\ldots,x_n$, α a number between 0 and 1, and sample statistics L_n and U_n such that $\mathrm{P}(L_n<\theta<U_n)=1-\alpha$ for any value of θ. Then (l_n,u_n), with l_n and u_n determined from the data, is a $100(1-\alpha)\%$ confidence interval for θ.

Standardized and studentized mean

When $X_1,\ldots,X_n$ are independent with a $N(\mu,\sigma^2)$ distribution, then $\frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}}$ and $\frac{\bar{X}_n-\mu}{S_n/\sqrt{n}}$ have a $N(0,1)$-distribution and a $t(n-1)$–distribution, respectively. Here $t(n-1)$ is the t-distribution with $n-1$ degrees of freedom.

Variance estimators

1. Regression model: $Y_i=\alpha+\beta x_i+U_i$ with U_i independent, with $\mathrm{E}[U_i]=0$, and $\mathrm{Var}(U_i)=\sigma^2$.
For σ^2: $\hat{\sigma}^2=\frac{1}{n-2}\sum_{i=1}^{n}(Y_i-\hat{\alpha}-\hat{\beta}x_i)^2$.

For the slope: $S_b^2=\frac{n}{n\sum x_i^2-(\sum x_i)^2}\hat{\sigma}^2$, for the intercept: $S_a^2=\frac{\sum x_i^2}{n\sum x_i^2-(\sum x_i)^2}\hat{\sigma}^2$.

2. For the two-sample t-test:

Pooled variance: $S_p^2=\frac{(n-1)S_X^2+(m-1)S_Y^2}{n+m-2}\left(\frac{1}{n}+\frac{1}{m}\right)$, non-pooled: $S_d^2=\frac{S_X^2}{n}+\frac{S_Y^2}{m}$.