Matematisk statistik - repetition

⚠️ Ej i formelblad. Förväntad kunskap.

❗ Viktigt att kunna.

‼️ Mycket viktig att kunna. En "hint" om förekomst på tenta har förekommit.

Exempel finns i de kompletta anteckningarna.

Inför tentor

Vanligast uppgifter på tentor

Kommer troligtvis:

- Transformationer: $P(X \leq x), X \in N(m, \sigma)$
- Variabeltransformationer: $Y = X^2, F_Y(x) = ?, E(Y) = ?$
- ML-skattningar
- Konfidensintervall / normalapproximationer av binomialfördelnignar
- Regression med $y = a + bx$ eller $y = ae^{bx}$

Kommer antagligen:

- Minimum och maximum av n stokastiska variabler
- Förklaringsgraden R^2

Viktiga begrepp

Dessa står det mer om antingen i detta dokument eller i de kompletta anteckningarna.

- Väntevärde
- Varians
- Standardavvikelse
- Koppling mellan fördelningsfunktion och täthetsfunktion

Kapitel 1

Terminologi

Resultatet av ett statistiskt försök kallas för ett **utfall** (*outcome*).

Mängden av alla tänkbara utfall kallas **utfallsrummet** (*sample space*) och betecknas med Ω.

En delmängd av utfallsrummet kallas för **händelse** (*event*).

Mängdlära

Den tomma mängden betecknas $\emptyset$.

$A \subseteq B$ betecknar att A är en **delmängd** till B. Det vill säga att alla element i A finns i B.

$x \in A$ betecknar att x **finns i** A.

$|A|$ betecknar **kardinaliteten** hos A, det vill säga antalet element.

$A \cap B$ betecknar **snittet** av A och B. Det vill säga alla element som finns i både A och B. För **oberoende** händelser gäller att $P(A \cap B) = P(A) * P(B)$.

$A \cup B$ betecknar **unionen** av A och B. Det vill säga alla element som finns i antingen A eller B. Sannolikheten för unionen av två händelser: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

$A \setminus B$ betecknar mängden $a \in A | a \notin B$. Det vill säga alla element i A som inte finns i B, "A **inte** B".

u betecknar **universalmängden**. I den finns alla element i sammanhanget.

A^c eller $u \setminus A$ betecknar **komplementet** till A. Det vill säga alla element som inte finns i A.

Den klassiska sannolikhetsdefinitionen

Notera att sannolikheten för varje utfall måste vara samma för att definitionen ska fungera.

$$\text{sannolikheten} = \frac{\text{antal gynsamma utfall}}{\text{antal möjliga utfall}}$$

$$\iff$$

$$p = \frac{g}{m}$$

Kombinatorik

Antal sätt att ordna element

	Utan hänsyn till ordningen (*kombinationer*)	Med hänsyn till ordningen (*permutationer*)
Utan återäggning	$\binom{n}{k} = \frac{n}{k!(n-k)!}$	${}_nP_r = P(n, r) = \frac{n!}{(n-r)!}$
Med återläggning	$\binom{n+k-1}{k}$	n^r

❗ Stickprov (*sample*)

Med återläggning (*sample with replacement*)

Här används binomialfördelningen.

För beroende händelser:

$$p^x (1-p)^{n-x}$$

För oberoende händelser:

$$p^x(1-p)^{n-x}\binom{n}{x}$$

Utan återläggning (*sample without replacement*)

Kallas även för stickprov utan återläggning.

$$\frac{\text{gynnsamma utfall}}{\text{möjliga fall}} = \frac{\binom{\text{A}}{\text{tagna ur A}}\binom{\text{B}}{\text{tagna ur B}}\cdots}{\binom{A+B+\ldots}{\text{tagna}}}$$

Detta gäller även då vi har flera grupper, exempelvis A, B och C. Notera att summan av de tagna i täljaren måste vara lika med de tagna i nämnaren.

⚠️‼️ Betingad sannolikhet (Bayes sats)

Om $B_1, .., B_n$ är en partition av Ω och $P(B_i) \neq 0, \forall i$ gäller för varje händelse A att

$$P(B_j \mid A) = \frac{P(B_j) * P(A \mid B_j)}{P(A)} = \frac{P(B_j) * P(A \mid B_j)}{\sum_{i=1}^{n} P(B_i) * P(A \mid B_i)}$$

Kapitel 2

Fördelningsfunktion (*probability function*)

$$F_X(x) = P(X \leq x)$$

Notera att $F_X(\text{övre gräns}) = 1$.

Median

För en exponentialfördelning gäller följande:

$$\begin{aligned} F_X(x_{0.50}) = 1 - e^{-\lambda x_{0.50}} &= 0.50 \\ e^{-\lambda x_{0.50}} &= 0.50 \Rightarrow \\ -\lambda x_{0.50} &= ln(0.50) \Rightarrow \\ x_{0.50} &= -\frac{1}{\lambda} ln(0.50) = \frac{ln(2)}{\lambda} \end{aligned}$$

⚠️ Väntevärde / genomsnitt (*mean*)

$$\begin{cases} \mu = E(X) = \sum_x x p_X(x) & \text{om kontinuerlig} \\ \mu = E(X) = \int_{-\infty}^{\infty} x f_X(x) dx & \text{om diskret} \end{cases}$$

Varians

$$V(X) = \sigma^2 = E((X-m)^2) = E(X^2) - (E(X))^2$$

Anmärkning: alltid större än 0.

Standardavvikelse

$$\sigma = \sqrt{V(X)}$$

Anmärkning: alltid positiv.

Kapitel 3

Diskreta fördelningar

För diskreta fördelningar betecknas vanligtvis täthetsfunktionen $f_X(x)$ och för kontinuerliga $p_X(x)$.

Binomialfördelningen

Fördelningen används då vi har n oberoende försök och sannolikheten för att lyckas $P(\text{lyckas}) = p$ är konstant. $X \in Bin(n, p)$.

$$f_X(x) = \binom{n}{x} p^x (1-p)^{n-x}, x = 0, 1, \ldots, n$$
$$E(X) = np$$
$$V(X) = npq = np(1-p)$$

Poissionfördelningen

Fördelningen för sannolikheten under ett intervall, exempelvis tid. $X \in Po(m)$.

$$f_X(x) = e^{-m} * \frac{m^x}{x!}$$
$$E(X) = m$$
$$V(X) = m$$

Där m är medelvärdet per intervalls-enhet. Exempelvis fyra samtal per minut. Notera att väntevärdet och variansen har samma värde.

$$X_1 \in Po(m_1)$$
$$X_2 \in Po(m_2)$$
$$X_1 + X_2 \in Po(m_1 + m_2)$$

Kontinuerliga fördelningar

Den likformiga fördelningen

Det är samma chans att få alla tal. Exempelvis har en tärning en likformig fördelning där varje uppstättning ögon har $\frac{1}{6}$ chans att slås.

$$P_X^k = \frac{1}{n}$$
$$F_X(x) = \sum_{k \leq x} P_X(k)$$

Exponentialfördelningen

Används vanligtvis för väntetider. Exempelvis "tiden till första bilen kör förbi". Aldrig negativa värden på x. λ står för intensiteten / händelse per tidsenhet.

$$f_X(x) = \lambda e^{-\lambda x}$$
$$F_X(x) = \int_0^x \lambda e^{-\lambda y} dy = 1 - \lambda e^{-\lambda x}, x \geq 0$$

Normalfördelningen

Normalfördelningen är den vanligaste fördelningen. Exempelvis medelvärdet av många mätningar, så som "längden av 18-åringar".

Anmärkning: $\Phi(-x) = 1 - \Phi(x)$

Standardiserad normalfördelning (standardized normal distribution)

Anta att $m = 0$, $\sigma = 1$, det vill säga x tillhör normalfördelningen mellan 0 och 1.

$$\Phi(x) = f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$
$$\Phi(y) = \int_{-\infty}^y \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

Notera termen $\frac{1}{\sqrt{2\pi}}$, denna kallas normeringskonstant då den får integralens värde att bli ett (krav per definition). För $\Phi(x)$ finns ingen enkel funktion.

Kapitel 4

Kovarians

$$C(X, Y) = Cov(X, Y) = E(X * Y) - E(X) * E(Y)$$

Korrelation

$$\rho = \frac{C(X, Y)}{\sigma_X * \sigma_Y}$$

Där $C(X,Y)$ är kovariansen / samvariationen av X och Y. Notera att ρ inte har någon sort eller enhet, det är enbart en konstant. Notera också att ρ alltid är i intervallet $-1 \leq \rho \leq 1$.

Då ρ är ± 1 är korrelationen exakt. Är $\rho > 0$ talar man om en positiv korrelation. Om ρ är 0 är värdena helt oberoende. Är $\rho < 0$ talar man om en negativ korrelation.

Kapitel 5

‼️ Transformationer

Speciellt transformationer av normalfördelningar är av intresse. (Exempelvis $X \in N(179,7)$).

Centralt exempel

18-åringar mönstrar. Längden är normalfördelad likt $X \in N(179,6)$. Beräkna sannolikheten att en 18-åring som mönstrar är 173cm eller kortare.

$$P(X \leq 173) = P(\underbrace{\frac{X-179}{6}}_{Z} \leq \frac{173-179}{6})) = \underbrace{P(Z \leq -1)}_{N(0,1)} = 1 - P(Z \leq 1) \approx 0.1587$$

Alternativt exempel

$Y = kX^2$

$$F_Y(u) = P(Y \leq u) = P(kX^2 \leq u) =$$
$$P(X^2 \leq \frac{u}{k}) = P(|X| \leq \sqrt{\frac{u}{k}}) =$$
$$P(-\sqrt{\frac{u}{k}} \leq X \leq \sqrt{\frac{u}{k}}) = F_X(\sqrt{\frac{u}{k}}) - F_X(-\sqrt{\frac{u}{k}})$$

Generell metod för transformationer

Gäller då invers till $g(X)$ finns och $g(X)$ är växande ($y' = positivt$).

$Y = g(X)$

$$F_Y(x) = P(Y \leq x) = P(g(X) \leq x) = P(\underbrace{g^{-1}(g(X))}_{X} \leq g^{-1}(x)) = F_X(g^{-1}(x))$$

⚠️ Fördelning för minimum (seriekoppling)

$$Y = min(X_1, X_2, \ldots, X_n)$$
$$F_Y(x) = 1 - (1 - F_X(x))^n$$

⚠️ Fördelning för maximum (maximum)

$$Z = max(X_1, X_2, \dots, X_n)$$
$$F_Z(x) = (F_X(x))^n$$

Kapitel 6

Normalapproximation av binomialfördelningen

Givet att $X \in Bin(n, p)$ där n är antalet försök och p är sannolikheten för ett lyckat försök gäller att om variansen $npq \geq 10$ kan binomialfördelningen skrivas om som normalfördelning likt $X \underset{\sim}{\in} N(np, \sqrt{npq})$.

Kapitel 7

Intensitet (*Intensity*)

Antalet fel per tidsenhet.

$$\lambda(t) = \lim_{h \to 0} \frac{P(Y \leq t + h \mid Y > t)}{h} = -\frac{R'_Y(t)}{R_Y(t)}$$

Funktionssannolikhet (*Reliability*)

$$R(t) = R(0) * e^{-\int_0^t \lambda(y)du} = R(0)e^{-\lambda t} \underbrace{=}_{\text{ofta}} e^{-\lambda t}$$

För övrigt är $R(0)$ sannolikheten att systemet fungerar vid början, exempelvis $p = 0.98 \Rightarrow R(0) = 0.98, R(t) = 0.98e^{-\lambda t}$.

Kapitel 11

⚠️‼️ Maximum likelihood-metoden (ML-metoden)

På tentan erfodras lösning steg för steg.

Egenskaper:

- Mindre varians än någon annan skattning.
- Asymptotiskt normalfördelad då $n \to \infty$. Det vill säga, mer och mer normalfördelad desto mer data. Lätt att räkna med.
- Att skatta en funktion. Funktionen av ML-skattningen $g(\theta) \Rightarrow g(\theta^*)$. Exempelvis $x^2 \Rightarrow (x^*)^2$
- Den är asymptotiskt värderiktig då $n \to \infty$. Den är inte alltid detta, men med tillräckligt högt n ger skattningen alltså väntevärdesriktighet.

Steg 1

Ställ upp likelihood-funktionen $L(\theta)$ och förenkla den så långt som möjligt.

$$L(\theta) = \begin{cases} p_X(x_1,\theta) * p_X(x_2,\theta) * ... * p_X(x_n,\theta) & \text{om diskret} \\ f_X(x_1,\theta) * f_X(x_2,\theta) * ... * f_X(x_n,\theta) & \text{om kontinuerlig} \end{cases}$$

Steg 2

Beräkna logaritmen $ln(L(\theta))$.

$$(f * g)' = fg' + f'g$$
$$(ln(f * g))' = (lnf + lng)' = \frac{f'}{f} + \frac{g'}{g}$$

Steg 3

Beräkna derivatan av $ln(L(\theta))$ med avseende på θ. Maximera sedan.

$$\frac{d}{d\theta} ln(L(\theta)) = 0$$

Lös sedan ut θ. Detta ger ML-skattningen av θ, θ^*.

Minsta kvadrat-metoden (MK-metoden)

Minimera variansen som funktion av θ.

$$\sum (x_i - m(\theta))^2 = 0$$

Kapitel 12

‼️ Intervallskattning / konfidensintervall

Vanligtvis har man en gräns, en sannolikhet med vilken man vill att svaret ska stämma. Av tradition är denna *konfidensgrad* 95%. Denna konfidensgrad anges vanligtvis i uppgiften på en tenta, annars utgår man från att det är 95% som gäller. Notera att "bredden" / "längden" innebär att konfidensgraden täcker hela fördelningen. Därför multipleras värdet som läses ur tabellen med 2. Se uppgift 183 och 184 för exempel.

$$\begin{cases} \overline{x} \pm \lambda_{\alpha/2} \frac{\sigma}{\sqrt{n}} & \text{känt } \sigma, \\ \overline{x} \pm t_{\alpha/2}(n-1) * \frac{s}{\sqrt{n}} & \text{okänt } \sigma \end{cases}$$

λ_α och t_α slås upp i formelbladet där α är $1 - \textbf{konfidensgrad}$, exempelvis $\lambda_{0.025}$ för konfidensgrad 95%.

För binomialfördelningen gäller genom normalapproximation följande:

$$p^* \pm 1.96\sqrt{\frac{p^*q^*}{n}}$$

För poissionfördelningen gäller genom normalapproximation följande:

$$m^* \pm 1.96\sqrt{\frac{m^*}{n}}$$

‼️ Centrala exempel

Binomialfördelning

Vi undersöker andelen miljöpartister (mp). Av 1000 personer är 50 mp.

$$p^* = \frac{50}{1000} = 0.05$$
$$npq = 1000 * 0.05 * 0.95 = 47.5 > 10 \Rightarrow \textbf{norm. approx. ok!}$$
$$p^* \pm 1.96\frac{\sqrt{47.5}}{1000} = p^* \pm \underbrace{1.96\sqrt{\frac{0.05 * 0.095}{1000}}}_{0.014} \Rightarrow (0.036; 0.064)$$

Det vill säga att miljöpartiets röstandel kommer att ligga mellan 3.6 till 6.4 procent.

Poissonfördelningen

Till en viss telefonväxel kommer i genomsnitt 80 samtal på en 2-minutersintervall. Gör ett 95% konfidensintervall för väntevärdet av antalet samtal på två minuter.

$$80 \pm 1.97\sqrt{\frac{80}{1}}$$
$$80 > 15 \Rightarrow \textbf{norm. approx. ok :-)}$$
$$80 \pm \underbrace{1.96\sqrt{80}}_{17.5} \Rightarrow (62.5; 97.5)$$

Svar: $(62.5, 97.5)$.

Binomialfördelningen

"Hur många behöver man fråga för felmarginal på en viss procent?"

$$p^* \pm 1.96\sqrt{\frac{p^* q^*}{n}}$$
$$\underbrace{0.02}_{\text{felmarginal}} = 1.96\sqrt{\frac{p^* q^*}{n}}$$
$$0.0004 = 1.96^2 \frac{p^*(1-p^*)}{n}$$
$$n = 2500 * 1.96^2 \underbrace{p^*(1-p^*)}_{\text{graf max. vid } p^*=0.5}$$

Det vill säga worst case $p^* = 0.5 \Rightarrow n = 2401$. Utgår man från worstcase är man alltid på den säkra sidan.

Jämför med fallet då $p^* = 0.1 \Rightarrow n = 864$. Dessa siffror avrundas uppåt (större chans).

Binomialfördelningen

En teknolog vill undersöka hur stor andel som vill köpa en viss produkt. Konstanten p är helt okänd. Hur många personer måste tillfrågas om felmarginalen $\leq 5\%$?

Lösning:

Okänt $p \Rightarrow$ worst case. $p^* = 0.5$.

$$
\begin{aligned}
0.05 &= 1.96\sqrt{\frac{p^*(1-p^*)}{n}} \\
0.0025 &= 1.96^2\frac{p^*(1-p^*)}{n} \\
n &= \frac{1.96^2}{0.0025} * 0.25 = 1.96^2 * 100 = 384
\end{aligned}
$$

Det vill säga att teknologen behöver fråga 384 personer.

Stickprov i par

Bit	Före	Efter	Differens
2	...	...	$z_1 = y_1 - x_1$
1	...	...	$z_2 = y_2 - x_2$
n	...	...	$z_n = y_n - x_n$

Då vi tidigare hade två stickprov (före och efter) har vi nu ett stickprov, differensen.

$$\overline{z} \pm t_{0.025}(n-1)\frac{s_z}{\sqrt{n}}$$

Oberoende stickprov

$x_1, x_2, \ldots, x_{n_1}$ och $y_1, y_2, \ldots, y_{n_2}$.

$$\overline{x} - \overline{y} \pm 1.96\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Kapitel 13

Hypotestest

För normalfördelning gäller

$$\begin{cases} u = \frac{\overline{x}-m}{\sigma/\sqrt{n}} & \text{känt } \sigma \\ t = \frac{\overline{x}-m}{s/\sqrt{n}} & \text{okänt } \sigma \end{cases}$$

För binomalfördelning gäller genom normalapproximation följande:

$$u = \frac{p^* - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

För poissonfördelning gäller genom normalapproxmiation följande:

$$u = \frac{\overline{x} - m_0}{\sqrt{\frac{m_0}{n}}}$$

För känt σ gämförs u med $\lambda_{a/2}$. För okänt σ (eller vid normalapproximation) jämförs t med $t_{\alpha/2}(n-1)$. Här är α exempelvis 0.95 för 95%-intervall.

$$\rho = \frac{C(X,Y)}{\sigma_X * \sigma_Y}$$

Där $C(X,Y)$ är kovariansen / samvariationen av X och Y och ρ är korrelationen. Notera att ρ inte har någon sort eller enhet, det är enbart en konstant. Notera också att ρ alltid är i intervallet $-1 \leq \rho \leq 1$.