Föreläsning 1

Thomas Önskog

30/10 2017

Introduktion till sannolikhetsteori och statistikteori

Ämnet matematisk statistik som denna kurs ger en introduktion till har många användningsområden. Matematisk statistik behövs för att göra opinionsundersökningar och väderprognoser, för att analysera resultatet i läkemedelstester, för att bestämma priset på optioner och för att göra kvalitetskontroller av en tillverkningsprocess med mera. Den matematiska statistiken har två delar: **sannolikhetsteori** och **statistikteori**.

Första halvan av kursen ägnas åt sannolikhetsteori och vi kommer strax att börja studera denna ifrån grunden, men vi börjar med en kort diskussion av statistikteori, vilket andra halvan av kursen ägnas åt. Statistikteori handlar om att analysera **slumpmässiga datamängder** med matematiska metoder. Säg att vi har gjort ett antal mätningar av en fysikalisk konstant. Naturliga frågeställningar är då exempelvis:

1. Vilket värde på konstanten stämmer bäst överens med datat? Detta är ett exempel på **punktskattning** (kapitel 11).

2. Ange ett intervall som med 95% sannolikhet innehåller det korrekta värdet på konstanten? Detta är ett exempel på ett **konfidensintervall** (kapitel 12).

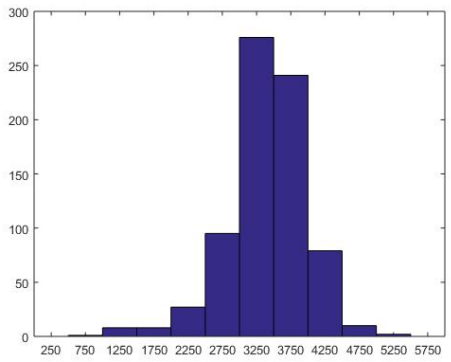
3. Kan vi med 99% sannolikhet utesluta att konstanten är noll? Detta är ett exempel på **hypotesprövning** (kapitel 13).

Ett viktigt moment i analysen av datamängder är att visualisera datat, så kallad **deskriptiv statistik**. Vi demonstrerar detta med hjälp av datafilen birth.dat på kurshemsidan som innehåller födelsevikten (avrundad till närmaste tiotal gram) hos $n = 747$ nyfödda spädbarn. Denna datafil kommer även att användas i datorlaboration 2. Datat är ogrupperat, och ser ut som följer

$$3050,\ 3180,\ 3880,\ 2750,\ 4040,\ 2990,\ 3100,\ 2900,\ 3200, \ldots, 3670.$$

Om vi bestämmer de **absoluta frekvenserna** f_i för de olika förekommande födelsevikterna y_i eller de **relativa frekvenserna** $p_i = f_i/n$, så får vi grupperade data som kan åskådliggöras i en frekvenstabell och/eller ett stolpdiagram.

y_i	f_i
⋮	⋮
3500	11
3510	7
3520	4
3530	8
3540	4
3550	7
⋮	⋮



Eftersom y_i kan anta många olika värden är det praktiskt att dela in datat i klasser och göra ett **histogram** där höjden av varje stapel är proportionell mot totala frekvensen för vikterna i klassen.

Det är ofta av intresse att bestämma ett "genomsnittligt värde" eller ett **lägesmått** för en datamängd. Det vanligaste valet av lägesmått är (aritmetiska) **medelvärdet**. Om $x_1, x_2, ..., x_n$ betecknar observationerna så ges medelvärdet av

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

I exemplet ovan är medelvärdet $\bar{x} = 3.40$ kg. Det är även intressant att veta hur mycket datat varierar, dvs att bestämma ett **spridningsmått** för datat. De vanligaste spridningsmåtten är **variansen** och **standardavvikelsen** som ges av

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \text{respektive} \quad s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

I exemplet ovan är standardavikelsen $s = 0.57$ kg.

Sannolikhetsteorins grunder

Sannolikhetsteori handlar om att formulera och analysera **slumpmodeller**, dvs matematiska modeller som, till skillnad från deterministiska modeller, innehåller någon form av slumpmässighet. Från en slumpmodell kan vi aldrig dra helt säkra slutsatser utan bara uttala oss i termer av sannolikheten för att olika händelser inträffar, exempelvis att en tärning visar sex vid tio tärningskast i rad eller att det snöar i Stockholm i morgon. En slumpmodell kan användas för att beskriva ett **slumpförsök**, dvs ett försök där resultatet inte kan bestämmas på förhand utan varierar från gång till gång. Vi inleder med tre definitoner.

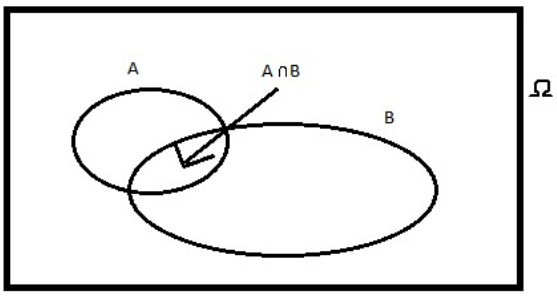
Definition. *Resultatet av ett slumpförsök kallas för ett* ***utfall*** *(betecknas ω).*

Definition. *Mängden av alla möjliga utfall kallas för* ***utfallsrummet*** *(betecknas Ω).*

Definition. *En* ***händelse*** *(betecknas $A, B, ...$) är en mängd av utfall, dvs $A \subset \Omega$.*

Exempel. *Ett träningskast med en sexsidig tärning är ett exempel på ett slumpförsök. Utfallen ω för slumpförsöket är i detta fall antalet prickar som kommer upp vid tärningskastet och utfallsrummet är $\Omega = \{1, 2, 3, 4, 5, 6\}$. Ett exempel på en händelse A för detta slumpförsök är $A =$ "udda antal prickar" $= \{1, 3, 5\}$.*

Eftersom händelser är delmängder av ett utfallsrum, så kan vi använda mängdoperationer såsom snitt, union och komplement för att uttrycka nya händelser och vi kan även visualisera händelser med hjälp av Venndiagram. Om exempelvis $A = \{1, 3, 5\}$ och $B = \{4, 5, 6\}$, så är händelsen $A \cap B = \{5\}$. Om två händelser inte kan inträffa samtidigt, så sägs de vara **disjunkta**.



Inom sannolikhetsteorin är vi intresserade av att bestämma **sannolikheten** för att en händelse inträffar. För födelseviktdatat gäller att 69% av de nyfödda hade en födelsevikt på mellan 3 och 4 kg. Vi förväntar oss därför att ett nyfött barn med ca 69% sannolikhet väger mellan 3 och 4 kg. Denna sannolikhet bygger på observationer (relativ frekvens). Om vi tar med fler observationer (exempelvis alla nyfödda i Sverige 2017), så kommer den relativa frekvensen att variera men så småningom stabilisera sig på en viss nivå. För att undvika svårigheten att bestämma sannolikheter exakt från relativa frekvenser, så postulerar vi att det till varje händelse A finns ett tal $\mathbb{P}(A)$ som vi kallar för sannolikheten för A. Exempelvis sätter vi $\mathbb{P}$("tärningen visar sex prickar") $= 1/6$. Vi väljer sannolikheten så att den relativa frekvensen för A hamnar mycket nära $\mathbb{P}(A)$ då antalet observationer är mycket stort. På detta sätt får vi en entydig definition av sannolikheter som inte beror på empiriska observationer. Kolmogorov tog 1933 fram följande axiomsystem för sannolikheter.

Axiom. *Ett* ***sannolikhetsmått*** $\mathbb{P}$ *är en funktion av händelser sådan att*

1. $0 \leq \mathbb{P}(A) \leq 1$ *för alla händelser* $A \subset \Omega$.
2. $\mathbb{P}(\Omega) = 1$.
3. *Om* $A_1, A_2, \ldots$ *är parvis disjunkta händelser, så gäller* $\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$

Notera att det sista villkoret implicerar att om $A \cap B = \emptyset$, dvs om A och B är disjunkta händelser, så gäller $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$. Från axiomen kan vi härleda många andra räkneregler för sannolikheter. Exempelvis gäller

$$\begin{aligned} 1 &= [\text{Axiom 2}] = \mathbb{P}(\Omega) = [\text{Definition av komplement}] = \mathbb{P}(A \cup A^*) \\ &= [\text{Axiom 3, } A \cap A^* = \emptyset] = \mathbb{P}(A) + \mathbb{P}(A^*), \end{aligned}$$

Detta bevisar följande sats.

Sats (**Komplementsatsen**). *För alla händelser A gäller* $\mathbb{P}(A^*) = 1 - \mathbb{P}(A)$.

En följdsats till komplementsatsen är att $\mathbb{P}(\emptyset) = \mathbb{P}(\Omega^*) = 1 - \mathbb{P}(\Omega) = 0$. Från axiomen kan vi också visa följande sats (se detaljer i boken).

Sats (**Additionssatsen**). *För alla händelser A och B, så gäller*

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

Satsen inses intuitivt från Venndiagrammet på föregående sida eftersom $\mathbb{P}(A) + \mathbb{P}(B)$ "mäter" händelsen $A \cap B$ dubbelt.

Antag nu att utfallsrummet Ω består av m möjliga utfall $\omega_1, \omega_2, ..., \omega_m$ som alla är lika sannolika. Vi sätter då $\mathbb{P}(\omega_i) = 1/m$ och säger att sannolikhetsmåttet är **likformigt** (jfr. tärningskast). Betrakta en händelse A som inträffar för g av de möjliga utfallen. Enligt den **klassiska sannolikhetsdefinitionen** gäller det då att $\mathbb{P}(A) = g/m$.

Kombinatorik används ofta för att bestämma antalet gynnsamma utfall g och antalet möjliga utfall m. Exempelvis säger **multiplikationsprincipen** att om det finns n_1 respektive n_2 sätt att utföra åtgärd ett respektive två, så kan båda åtgärderna utföras på totalt $n_1 n_2$ sätt. Vi inleder nästa föreläsning med att undersöka fler metoder för att bestämma antalet gynnsamma och möjliga utfall.