

Statistik 1 hemtentamen

Det viktiga i statistikinläringen är att tillämpa kunskaperna i praktiken – helst på ens eget ämnesområde. Detta bidrar till att skapa en god intuition för vad statistiska begrepp and analyser innebär i en kontext som är bekant för en själv.

Som en bas för inläring av grundläggande statistik kan man använda sig av boken: **Biometri** av **Ulf Olsson** m. fl. (2005). Kursbiblioteket har ett rätt stort antal exemplar av boken, men får man ej tag på den, kan andra liknande böcker eller kompendier användas utan vidare. Exempelvis finns det över 10 exemplar av kompendiet **Statistik: en introduktion** av **Östen Widjeskog** i ÅA-biblioteket. Kurshemsidan erbjuder också en hel del länkar till bakgrundsmaterial om sannolikhet och statistiska analyser.

I hemtentamen skall deltagarna kunna kritiskt tillämpa statistisk metodik på diverse datamaterial för att ta ställning till frågor och hypoteser som är typiska i forskningssammanhang. Resultat från dessa analyser samt egna reflexioner kring dem sammanställs i en rapport som skickas till föreläsaren, antingen per mejl eller per internpost. En godkänd rapport returneras med kommentarer till studiekansliet för deltagares huvudämne eller direkt till deltagaren ifall rapporten sänts per mejl. Eventuella icke-godkända rapporter returneras till deltagarna med beskrivning av de ändringar som måste utföras för att studieprestationen skall godkännas.

För samtliga uppgifter i hemtentamen uppmanas deltagarna själva att skaffa lämpliga datamaterial för de analyser som nämns nedan. Detta är en viktig del av inlärningsprocessen, eftersom datamaterialets karaktär påverkar kraftigt vilka analyser är meningsfulla att utföra. Samtidigt bekantar man sig med typiska frågeställningar och mätmetoder som används inom ett visst ämne. Data kan hämtas t ex från internet, från vetenskapliga publikationer (helst på ens eget ämnesområde), från forskarna på eget ämne eller från längre hunna studerande på samma ämne som redan håller på med graduprojekt. Vid sökning av vetenskaplig litteratur är Google Scholar och Web of Science utmärkta hjälpmedel. Flera deltagare kan givetvis dela på ett och samma datamaterial, men rekommendationen är att skaffa data som reflekterar ens egna intressen inom studierna.

Vid rapportering av resultaten bör datakälla och materialets karaktär beskrivas för samtliga uppgifter! Själva datafilerna behöver ej bifogas med rapporten. Notera att data från en och samma undersökning/experiment kan användas vid flera uppgifter ifall att det är lämpligt med tanke på frågeställningen. Rapporterna för hemtentamen bör returneras senast vid början av studieperiod 4 under läsåret 2010-11.

Uppgift 1:

Beskrivande numerisk och grafisk statistik. Grundbegrepp inom sannolikhet.

Förklara följande begrepp, du kan gärna varva formella beskrivningar med egna ord.

1. Medelvärde, varians, standardavvikelse, median, typvärde

2. Olika noggrannhetsskalor vid mätning. nominell-, ordinal-, intervall- och kvotskala. Kontinuerlig variabel, kategorisk variabel. Beskriv i tur och ordning vilka sammanfattningar eller storheter är meningsfulla att beräkna från data som har observerats enligt en viss mätskala. Beskriv även vad kan gå fel ifall du beräknar något som datat inte tillåter.
3. Histogram, scatterplot (spridningsdiagram)
4. Sannolikhet, sannolikhet baserad på frekvenstolkning, subjektiv sannolikhet
5. Betingad sannolikhet
6. Frekvenstabell, marginalfördelning, betingad fördelning (conditional distribution)

Beräkna utifrån data de kvantiteter som nämndes vid punkt 1,3,5,6 och tolka resultaten.

Uppgift 2:

Elementär modellbaserad statistisk analys och presentation av slutsatser

Förklara följande begrepp, du kan gärna varva formella beskrivningar med egna ord.

1. Statistisk inferens
2. Statistisk modell, parameter
3. Skattning av en parameter
4. Konfidensintervall
5. Binomialmodell(n,p), skattning och konfidensintervall för binomialmodellens parameter p
6. Normalmodell med ett okänt medelvärde μ och standardavvikelse σ , skattning av μ och σ , samt konfidensintervall för μ
7. Odds, oddskvot/relative risk, log-oddskvot, konfidensintervall för oddskvot

Beräkna utifrån data de kvantiteter som nämndes vid punkt 5-7 och tolka resultaten.

Uppgift 3:

Statistisk inferens och hypoteser

Förklara följande begrepp, du kan gärna varva formella beskrivningar med egna ord.

1. Hypotesprövning
2. Nollhypotes, mothypotes, dubbelsidig hypotes, enkelsidig hypotes
3. Icke-parametriska test, så som Mann-Whitney U och Kruskal-Wallis testen
4. p-värde, signifikansnivå

Beräkna utifrån data testen vid punkt 3 och tolka resultaten. Formulera lämpliga hypoteser enligt punkt 2 för data som behandlades vid punkt 5-7 i uppgift 2 ovan. Utför hypotesprövningen och tolka resultaten.

Uppgift 4:

Samband och kausalitet mellan variabler

Förklara följande begrepp, du kan gärna varva formella beskrivningar med egna ord.

1. Korrelation, kausalitet (orsak-verkan samband)
2. Konfidensintervall för korrelationskoefficient
3. Chi-square test av oberoende för två dikotoma variabler.
4. Simpson's paradox

Betrakta begreppen vid punkt 1-4 i samband med data och tolka resultaten

Uppgift 5:

Regressionsanalys och envägs ANOVA

Förklara följande begrepp, du kan gärna varva formella beskrivningar med egna ord.

1. Linjär regression
2. Variansanalys (ANOVA), speciellt sk envägs ANOVA (one-way ANOVA)
3. Post hoc test

Utför analyser av den typ som nämns vid punkt 1-3 och tolka resultaten

Uppgift 6:

Hantering av multipla hypotesprövningar.

Utgå från ett datamaterial som innehåller ett relativt stort antal (t ex 100 stycken) kontinuerliga eller diskreta variabler (eller både och). Skapa antingen A) en kategorisk variabel som indikerar en slumpmässig indelning av observationerna till två ungefär lika stora 'fiktiva' grupper eller B) en kontinuerlig variabel vars värden är slumpmässigt fördelade enligt en normalfördelning.

Lägg den nya variabeln till datamaterialet och testa för samband mellan den och de riktiga variablerna. Låt X vara en kontinuerlig och Z en kategorisk variabel bland de riktiga variablerna. Ifall du valt A är det meningsfullt att testa för skillnad mellan medelvärden för X hos de två grupperna. På motsvarande sätt kan du testa för samband mellan Z och de två fiktiva grupperna. Ifall du valt B kan du testa för korrelation mellan X och den fiktiva kontinuerliga variabeln.

Undersök hur många test resulterar i ett signifikant samband på 5%-nivån då de individuella testen inte justeras för multipla test. Hur skulle situationen ändras om man använde sig av Bonferroni-korrektion? Det bör dock påpekas att bättre och modernare metoder än Bonferroni-korrektion finns tillgängliga, t ex False Discovery Rate (FDR) kontroll enligt metoden av Benjamini och Hochberg.