Statistiska metoder, kompendium

1 Sannolikhet och uppräkning

1.1 Stickprov och population

I statistik är en *population* den totala, teoretiska, mängden av händelser eller objekt som studeras. Varje delmängd, som är datan som faktiskt erhålls, är ett *stickprov*. I ovanliga statistiska fall kan det vara så att hela populationen finns tillgänglig– det rör sig då av nödvändighet om ett ändligt utfallsrum. Om vi har tillgång till hela populationen innebär det att vi aldrig kommer se okänd data och det finns då ingen anledning att använda maskininlärningstekniker.

I maskininlärning kallas egenskaperna hos populationen för en *grundsanning* (*ground truth*). Det är denna vi försöker lära maskinen, det vill säga representera numeriskt. Den numeriska representationen sägs vara en *modell* av grundsanningen.

Resonemang kring population och stickprov kräver subtilt olika definitioner och metoder för att vara statistiskt säkerställda. Egenskapen hos en skattning att den gäller för både stickprov och population kallas *rättvis* (*unbiased*).

Till stöd har vi två huvudsakliga sätt att definera sannolikhet.

1.2 Relativ frekvens (relative frequency, frequentist view)

Antalet gånger en händelse inträffar delat med totala antalet observationer. Av särskild vikt för resonemang kring *stickprov*.

$$P[A] = \frac{f}{n} \tag{1}$$

1.3 Klassisk sannolikhet (classical probability)

Antalet sätt en händelse kan inträffa delat med antalet möjliga utfall. Av särskild vikt för resonemang kring *populationen*.

$$P[A] = \frac{n(A)}{n(S)} \tag{2}$$

Sannolikheten att en viss händelse inträffar skrivs

$$P[X = x]$$

Ex. Om X är utfallet av en sex-sidig tärning så är $P[X = 1] = 1/6$. Notera att om vi byter till relativ frekvens och räknar på en ändlig uppsättning tärningsslag så fluktuerar värdet. Det blir bara exakt 1/6 av slumpen eller när antalet försök går mot oändligheten.

För att räkna klassisk sannolikhet behöver vi kunna räkna upp händelser och antalet möjliga händelser.

1.4 Permutationer och kombinationer

En *permutation* är ett arrangemang av objekt med hänsyn till ordning, dvs ett val utan återläggning (upprepning) men med hänsyn till ordning. För n typer av distinkta objekt valda r åt gången gäller

$$ {}_nP_r = \frac{n!}{(n-r)!} \tag{3} $$

Ex. Låt $n = 4$ och $r = 2$. Låt oss säga att vi har fyra olika frukter: banan, äpple, apelsin och nektarin. Antalet sätt att ge en frukt vardera, till två personer, är då $\frac{4!}{(4-2)!} = 12$.

Ex. Låt A vara en lista av tal $A = [1, 2, 3, 4]$. Listan $A' = [2, 3, 1, 4]$ kallas en permutation av A. Antalet sådana som kan formas är alltså ${}_4P_4 = \frac{4!}{(4-4)!} = 24$. Längden av listan är fyra och det är därmed också fyra objekt i listan. Detta särksilda fall, antalet sätt att arrangera n-st objekt, är alltså alltid $n!$.

En *kombination* är ett arrangemang av objekt utan hänsyn till ordning, dvs ett val utan återläggning och utan hänsyn till ordning. För n typer av distinkta objekt, valda r åt gången, gäller

$$ {}_nC_r = \binom{n}{r} = \frac{n!}{r!(n-r)!} \tag{4} $$

Ex. Låt $A = \{\texttt{B}, \texttt{A}, \texttt{SELECT}, \texttt{START}\}$ vara en mängd av knapptryckningar. Antalet delmängder som har storlek 2, dvs antalet samtidiga knapptryckningar med två knappar, är då ${}_4C_2 = \frac{4!}{2!(2!)} = 6$.

1.4.1 Kombinatorik

Kombinatorik är just studiet av alla sätt saker kan arrangeras och väljas och för utförlighetens skull sammanfattas här några formler.

n^k	Val med återläggning och med hänsyn till ordning
${}_nP_k = \frac{n!}{(n-k)!}$	Val utan återläggning men med hänsyn till ordning
${}_nC_k = \binom{n}{k} = \frac{n!}{k!(n-k)!}$	Val utan återläggning och utan hänsyn till ordning
$\binom{k+n-1}{n-1} = \binom{n+k-1}{k}$	Val med återläggning men utan hänsyn till ordning

2 Fördelningar

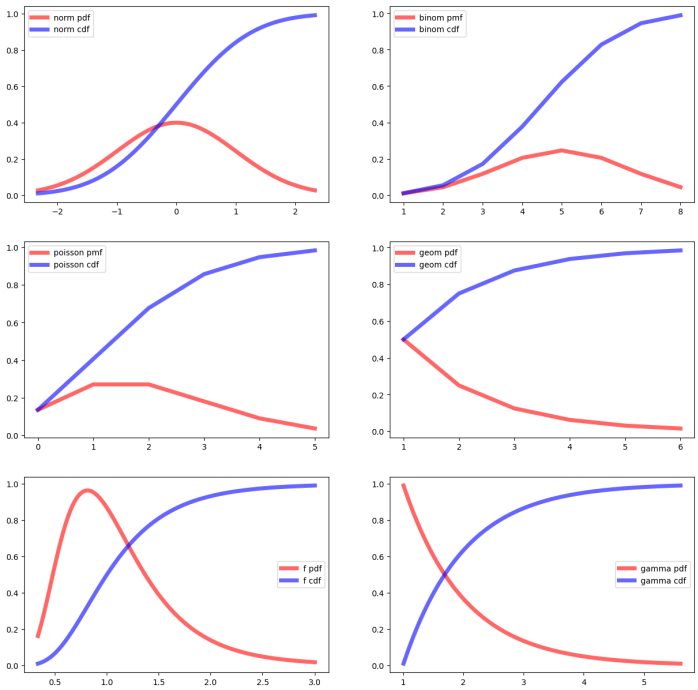

Figure 1: Sannolikhetsfunktion och fördelningsfunktion för några besläktade distributioner.

2.1 Slumpvariabler

En *slumpvariabel* (*random variable*) eller en *stokastisk* variabel, förkortat s.v. (r.v.), är en variabel som ger ett värde genom dragning (sampling) med någon viss sannolikhet för varje möjligt värde. En *sannolikhetsfunktion*, även kallad *täthetsfunktion* (*probability function*, *probability density function*) beskriver denna sannolikhetsfördelning och defineras som

$$f(x) = P[X = x] \tag{5}$$

under villkor att

$f(x) \geq 0$ *för alla* x

$$\sum_{alla\ x} f(x) = 1$$

Villkoren är nödvändiga (och tillräckliga) för att täthestfunktionen skall beskriva en sannolikhetfördelning. Figur 1 visar några täthetsfunktioner i rött/orange. I blått visas ett alternativt sätt att visa fördelningen. En *fördelningsfunktion* (*cumulative distribution function*) är bland annat mindre känslig för transformationer vid jämförselser mellan stickprov. Den defineras som:

$$F(x) = P[X \leq x] = \sum_{x_i \leq x} f(x_i) \tag{6}$$

Notera att lutningen hos fördelningsfunktionen är precis sannolikhetsfunktionen, det vill säga att derivatan av F är f: $\frac{d}{dx}F(x) = f(x)$. Fördelningsfunktionen är alltså arean under sannolikhetsfunktionen från vänster till höger.

2.2 Moment

2.3 Väntevärde (Expected Value)

Det första *momentet* hos en fördelning är dess *väntevärde*. Detta är i statistik det teoretiska medelvärdet av alla möjliga värden en slumpvariabel kan anta. Det anger en mittpunkt av alla värdena i fördelningen. I allmänhet räknar vi på stickprovsmedel och antar att detta är en rättvis skattning av det sanna medelvärdet, med hjälp av vissa korrektioner (se kapitel 3).

$$\mu = E[X] = \sum_{x \in S} x f(x) \tag{7}$$

2.4 Varians och standardavvikelse

Det andra momentet

$$E[X^2] = \sum_{x \in S} x^2 f(x)$$

är ett icke-centrerat mått på fördelningens *spridning*. Efter centering får vi ett mått på *varians*.

$$\texttt{Var}X = \sigma^2 = E[(X - \mu)^2] = E[X^2] - (E[X])^2 \tag{8}$$

Eftersom det andra momentet innehåller en kvadrat är variansen inte omedelbart jämförbar med värdena i fördelningen. Standardavvikelsen återgår till samma skala som slumpvariabeln har.

$$\sigma = \sqrt{\texttt{Var}X} = \sqrt{\sigma^2} \tag{9}$$

De två följande momenten benämns i ordning *skevförskjutning* (skew) $E[X^3]$ och *svanstyngd* (kurtosis) $E[X^4]$, men är utanför denna kurs. I Figur 1 kan dock anas vad dessa moment medföljer utifrån de svenska namnen.

2.5 Fördelningsparametrar

Till varje fördelning hör ett antal parametrar, vilket något förenklat är fördelningens moment. Till exempel skrivs en *normalfördelning* med medel μ (mittpunkt, väntevärde) och varians σ^2:

$$\begin{aligned} X &\sim N(\mu, \sigma^2) && \text{Normalfördelning} \\ \mu &= E[X] = \mu \\ \sigma^2 &= E[(X - \mu)^2] = \sigma^2 \\ f(x) &= \frac{1}{\sigma\sqrt{2\pi}e^{[(x-\mu)/\sigma]^2}} \end{aligned}$$

En *standard-normal* fördelning är en normalfördelning centrerad kring 0, vilket alltså är dess väntevärde. Variansen är normaliserad till 1, vilket alltså innebär att även standardavvikelsen blir 1 ($\sqrt{1} = 1$). Denna är särksilt användbar eftersom värdet på x-axeln i den översta figuren i vänstra hörnet av Figur 1 helt enkelt anger hur många standardavvikelser bort från medlet detta är. Standard-normalfördelningen skrivs med en särskild symbol $\mathcal{Z}$:

$$\mathcal{Z} \sim N(0, 1)$$

För andra fördelningar gäller andra parametrar. Till exempel för en binomial fördelning:

$$X \sim Binom(n, p) \tag{10}$$

där n är antalet försök och p är sannolikheten att vardera försök lyckas. För en binomial fördelning gäller:

$$\begin{aligned} X &\sim Binom(n,p) && \text{Binomialfördelning} \\ \mu &= np \\ \sigma^2 &= np(1-p) \\ f(x) &= \binom{n}{x} p^x (1-p)^{n-x} \end{aligned}$$

Alltså kan momenten beräknas från parametrarna, vilket i allmänhet gäller när man skriver fördelningar på detta sätt. När antalet försök går mot oändligheten, dvs $n \to \infty$, övergår den diskreta binomialfördelningen i den kontinuerliga normalfördelningen. Ett till exempel på en kontinuerlig fördelning är Γ-fördelningen, en kontinuerlig exponentiell fördelning:

$$\begin{aligned} X &\sim \Gamma(\alpha, \beta) && \text{Gammafördelning} \\ \mu &= \alpha\beta \\ \sigma^2 &= \alpha\beta^2 \\ f(x) &= \frac{1}{\beta^\alpha \int_0^\infty z^{\alpha-1} e^{-z} dz} x^{\alpha-1} e^{-x/\beta} \\ x, \alpha, \beta &> 0 \end{aligned}$$

Som synes är täthetsfunktionen för denna familj av fördelningar väldigt komplex. När parametern $\alpha \to \infty$ så blir denna fördelning identiskt med normalfördelningen. F-fördelningen är en jämförelse mellan två χ^2-fördelningar, som i sin tur är Γ-fördelningar med vissa parametrar. Dessa likheter går djupt– villkoren för sannolikhetsfördelningar gör att de faller i några breda kategorier som beror på deras moment, grovt de som mer liknar normalfördelning och de som lutar åt exponentialfördelning. Skevmoment och svanstyngd beskriver skillnaderna i detaljerna.

2.6 Rimlighet

I maskininlärning är en annan sorts statistik väldigt viktig: rimlighetsstatistik (likelihood statistics). Till skillnad från sannolikheter så följer inte rimlighet en fördelning, men har ett förhållande till en fördelningsfunktion. Om vi tänker oss en generell fördelningsfunktion $f(X, \alpha_1, .., \alpha_n)$ där α_i är fördelningsens parameterar så kan vi bilda en gemensam fördelning (se 4.2)

$$f(X_1, .., X_m, \alpha_1, ..., \alpha_n)$$

genom att gör upprepade stickprov. Rimlighetsfunktionen bildas genom att ersätta slumpvariablerna X_i med sina observerade värden x_i och istället låta parametrarna α_j vara variabler:

$$\mathcal{L}(\mathbf{a}; \mathbf{x}) = f(x_1, ..., x_n, \mathbf{a}),$$

där $\mathbf{a}$ är en vektor av parameter variablerna. Rimlighet ”vänder” i ett visst avseende på sannolikheten. Givet utfallet (observationerna) och ett antagande om fördelningen så är rimlighet sannolikheten att värdena drogs ur den antagna fördelningen med parametrarna $\mathbf{a}$. Varje fördelning

har en rimlighetsfunktion, men vi nöjer oss i denna kurs med några av de diskreta. Se även *logistisk regression* i regressionskompendiumet för mer om hur rimlighet används i ML.

$$\text{Binomial} \equiv \mathcal{L}(p; n, x) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$\text{NegBinom} \equiv L(p; n, x) = \prod_{k=1}^{n} \binom{x-1}{k-1} p^k (1-p)^{x-k}$$

$$\text{Geometrisk} \equiv \mathcal{L}(\surd; \S) = (1-p)^x p$$

3 Beskrivande statistik och skattning

För att kunna härleda en numerisk represenation av grundsanningen, det vill säga bygga en modell, behöver vi klura ut vad för sorts slumpvariabler vi har att göra med. Är de mätvärden eller abstrakta kategorier? Är de diskreta eller kontinuerliga? Vad för sannolikhetsfördelning har de? Vilken maskininlärningsmetod är lämplig? Den sista av dessa frågor lämnar vi till Maskininlärningskursen, i denna kurs begränsar vi oss till linjär regression. För att besvara övriga frågor använder vi statistiska metoder.

3.1 Stickprov

Om vi inte på förhand kan klura ut och utföra ett experiment som har en känd fördelning (t.ex en binomialfördelning) och därmed resonera om hela populationen, så måste vi göra stickprov. Ett stickprov defineras formellt enligt

Stickprov *ett slumpmässigt urval av storlek n från en fördelning X som är en samling av n oberoende slumpvariabler, var och en med samma fördelning som X.*

Detta innebär att statistiska urval alltid skall göras med återläggning. En tumregel är att så länge stickprovet utgör som mest 5% av populationen kan ändå oberoende antas. I de flesta maskininlärningsproblem är populationen oändlig (till exempel alla teoretiskt möjliga bilder på katter), men paradoxalt nog kan tillgången på data vara så begränsad att stickproven är så små att de måste utvidgas genom så kallad dataförstärkning (data augmentation). I maskininlärningsproblem yttrar sig beroenden i datan som *överanpassning*, vilket innebär att systemet får väldigt bra resultat under träningen men presterar bedrövligt på okänd data.

3.2 Histogram, sannolikhets- och fördelningsfunktionen

Ett första sätt att visualisera fördelningen är ett *histogram*. Dessa finns direkt tillgänliga i `pandas`:

```
1 import pandas as pd
2 import numpy as np
3 df = pd.DataFrame(np.column_stack([np.random.gamma(2, 1, size=100),
4                                    np.random.standard_normal(size=100),
5                                    np.random.binomial(5, .3, size=100)]),
6                   columns=["Gamma/Chi^2", "Standard Normal", "Binomial"])
7 df.hist()
```

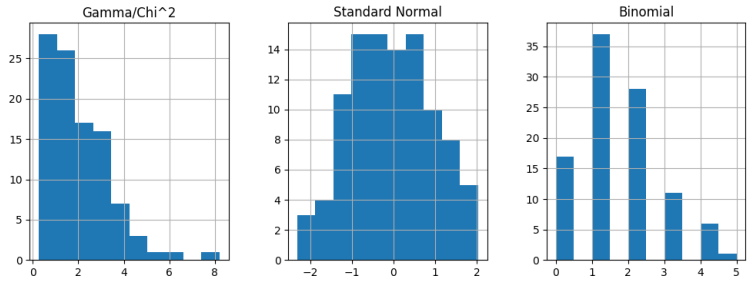

Figure 2: Histogram

Dessa visualiserar sannolikhetsfunktionen. Genom att ändra rad 7 i ovanstående listning till `df.hist(cumulative=True)` erhålls istället kumulativa histogram (*ogive*), vilka följer fördelningsfunktionen.

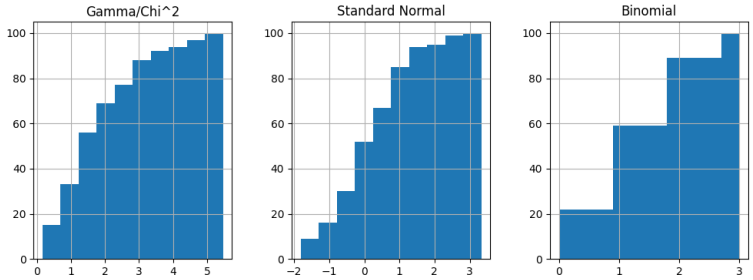

Figure 3: Kumulativa histogram (ogives)

Histogram konstrueras med relativ frekvens och är alltså ett stickprovsmått. För att konstruera histogram måste slumpvariabelns värden delas in i kategorier, vilket innebär slutna intervall för kontinuerliga värden. Dessa måste väljas på ett sätt som inte döljer eller förvrider den sanna fördelningen. Det finns många metoder för att göra detta men ingen är dock felfri. Problemet kallas *binning* och är ett svårlöst problem i allmänhet. Just för histogram kommer man undan med enklare algoritmer, men i mer komplicerade fall kan det hända att man manuellt får göra uppdelningen. Till exempel vid s.k. χ^2-analys, även känt som *goodness-of-fit*, så fungerar metoden inte om det förekommer tomma intervall, vilket är svårt att balansera mot kravet om rättvishet vad gäller storleken på intervallen man skapar.

Konsekvensen är att det kan vara svårt att se skillnad på fördelningar genom histogram. Det är en bra idé att titta på både det kumulativa och relativa frekvenshistogrammet, som illustreras i Figurer 2 och 3.

3.3 Lägesmått, Spridningsmått

Mittpunkten för stickprovet är *stickprovsmedlet* (*sample mean*):

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{11}$$

där n är stickprovets storlek. Medlet är obegränsat och kan både gå mot noll, en konstant eller mot oändligheten när stickprovets storlek går mot oändligheten. Stickprovsmedlet är en

rättvis uppskattning av populationens medelvärde bara om stickprovet följer samma sannolikhetsfördelning. Åtminstone för uppräkneligt oändliga utfallsrum är det ett rimligt antagande att detta måste vara sant när $n \to \infty$ men när det gäller oräkneligt oändliga utfallsrum (kontinuerliga) så kan vi inte göra samma antagande. För kontinuerliga slumpvariabler måste vi istället förlita oss på Centrala Gränsvärdessatsen och de Stora Talens Lag för samma resultat.

Ett ytterligare beskrivande lägesmått är *medianen*. Stickprovsmedianen i ett sorterat utfallsrum med storlek n är talet som finns på plats

$$\frac{n+1}{2} \qquad \text{Medianläge}$$

Om n är jämnt är det medlet av värdena ovanför och under medianläget (median location).

Ett sista lägesmått är *typvärdet* (*mode*) vilket helt enkelt är det värde som förekommer flest gånger (eller har högst täthet för kontinuerliga slumpvariabler). För diskreta slumpvariabler är det bara att räkna förekomsterna i datan, men för kontinuerliga behöver mer komplicerad analys tillämpas, vilket vi lämnar därhän.

Fördelningens spridning beräknas genom *stickprovsvarians*:

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \tag{12}$$

där *Bessels korrigering* $(n-1)$ har tillämpats för att stickprovsvariansen skall vara en rättvis skattning av populationsvariansen.

3.4 Boxplot

Ett annat sätt att visualisera en datamängd är en boxplot. En boxplot jämför datamängden med en normalfördelning och synliggör avvikelser från en sådan. Liksom för histogrammen är det en något inveklad, men i detta fall enkelt automatiserbar, procedur att konstruera en boxplot. De finns tillgängliga i `pandas`:

```
1    df.boxplot()
```

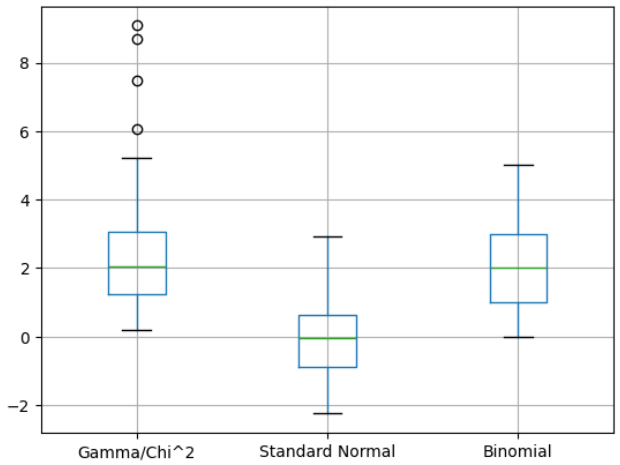

Figure 4: Boxplot för datan genererad i tidigare exempel. Den gröna linjen anger medelvärdet. Boxen anger den mittersta interkvartilen. Strecken avgränsar området som anses normalt och ringarna anger extremvärden. Notera förskjutningarna av medlen och kvartilerna, slumpmässiga för normalfördelningen men parametriskt för binomialfördelningen ($p = .3$) och Γ-fördelningen.

Kvartiler räknas enligt frekvens och följer fördelningsfunktionen. Q_1 är det minsta värde som täcker de lägre 25% av värdena i datan, Q_2 är medlet, dvs täcker 50% och Q_3 täcker 75% av datan.

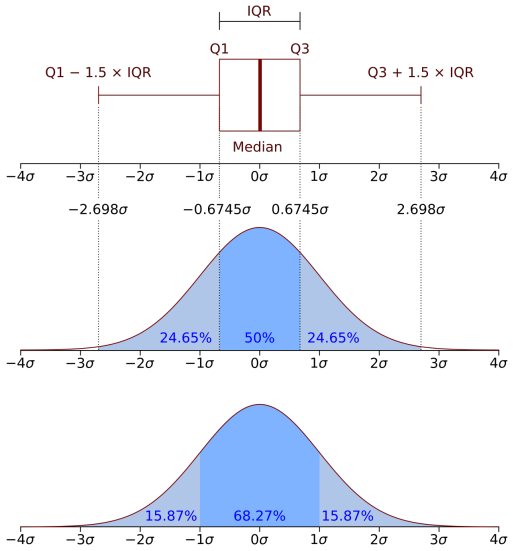

Figure 5: Boxplot och kvartiler jämfört med en standard-normalfördelning (mitten) samt standardavvikelser (nederst). Skapad av Jhguch at en.wikipedia, CC BY-SA 2.5

3.5 Probplot

Ett annat sätt att jämföra med normaldistribution mer direkt är en `probplot`:

```
1 import scipy.stats as stats
2 import matplotlib.pyplot as plt
3 fig, ax = plt.subplots(3,1)
4 stats.probplot(df["Gamma/Chi^2"].to_numpy(), plot=ax[0])
5 stats.probplot(df["Standard Normal"].to_numpy(), plot=ax[1])
6 stats.probplot(df["Binomial"].to_numpy(), plot=ax[2])
7 plt.show()
```

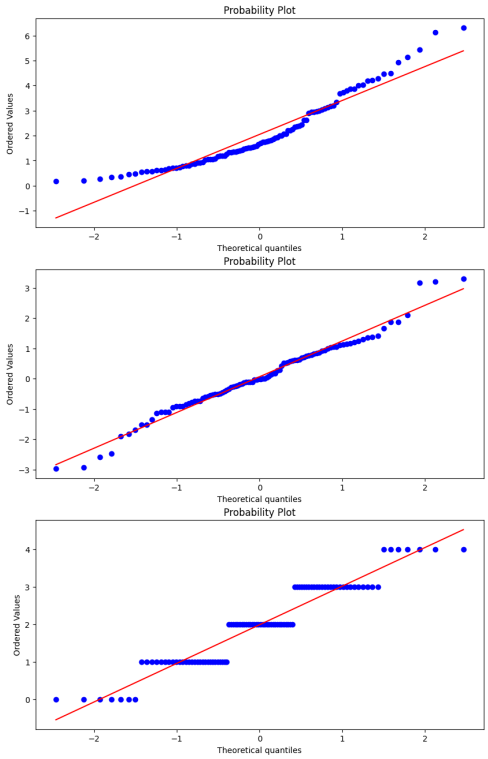

Figure 6: Probplot för samma data som tidigare exempel. Överst syns en typisk kurva för hur en exponentiell fördelning skiljer sig från normalfördelningen. Mitterst är ett positivt fall, datan är normalfördelad. Nederst syns hur en diskret slumpvariabel som är ungefär normal ser ut. Då $n \to \infty$ blir värdena mer och mer linjärt placerade inom kvartilerna och närmar sig mittengrafen. När $\alpha \to \infty$ rätar den översta grafen ut sig och börjar likna mittengrafen.

Denna funktion jämför direkt med en normalfördelning. Om datan passar en normalfördelning exakt skall alla datapunkter hamna på linjen.

3.6 Konfidensintervall

Ett konfidensintervall ger konkret information om variationen kring en statistisk storhet, till exempel ett uträknat medel eller en modellparameter. Ett konfidensintervall defineras enligt

Ett $100(1-\alpha)\%$ *konfidensinterval för en parameter* θ *är ett slumpinterval* $[L_1, L_2]$ *sådant att*

$$P[L_1 \leq \theta \leq L_2] = 1 - \alpha \tag{13}$$

Där (1-α) är intervallets *konfidensnivå*. Om $\alpha = 0.05$ så är alltså $100(1-\alpha) = 100(0.95) = 95\%$ konfidensnivån. Detta skall tolkas som att i 95% av fallen så ligger parametern inom det angivna intervallet. Beroende på vilken fördelning som en slumpvariabel följer används olika formler för att räkna ut ett konfidensintervall.

Låt $X_1, X_2, ..., X_n$ *vara ett stickprov av storlek* n *från en normalfördelning med känt medel* μ *och känd varians* σ^2. *Ett* $100(1-\alpha)\%$ *konfidensintervall på medlet* μ ges av

$$\bar{X} \pm z_{\alpha/2}(\sigma/\sqrt{n}), \qquad \mu, \sigma^2 \quad \textit{kända}$$

där $\bar{X}$ är stickprovsmedlet och $z_{\alpha/2}$ är punkten där normalfördelningens fördelningsfunktion har värdet $\alpha/2$. För att räkna ut detta $z_{\alpha/2}$ behövs alltså inversen till fördelningsfunktionen. Inversen av fördelningsfunktionen kallas på engelska *percentile point function* men på svenska har den inget så väldefinerat namn. "Alpha-funktionen" har jag hört den kallas, men det är inte ett så precist namn. I `scipy.stats` heter funktionen `ppf`.

Det är särskilt fall när både variansen och medlet är känt. I maskininlärning är det då vi centrerat och normaliserat datan till att vara standardnormal. Då är medlet 0 och variansen 1, vilket förenklar matematiken. Ett mer typiskt fall är att medlet är okänt och variansen skattad. Då används

$$\bar{X} \pm t_{\alpha/2}(S/\sqrt{n}) \qquad \mu, \sigma^2 \quad \textit{okända}$$

där $t_{\alpha/2}$ nu istället är motsvarande punkt för t-fördelningen med $n-1$ frihetsgrader. T-fördelningen kan sägas ta hänsyn till att stickprovet inte nödvändigtvis är exakt normalfördelat. Notera att $S/\sqrt{n}$ är en skattning av *standardavvikelsen i medlet* $\sigma/\sqrt{n}$. När stickprovets storlek går mot oändligheten går alltså konfidensintervallet mot exakt t-fördelningen. Motsvarande gäller förstås även när variansen och medlet är kända.

För linjär regression rör det sig inte längre om en enskild slumpvariabel med en fördelning, utan många slumpvariabler (modellens parametrar) med en gemensam fördelning. För linjär regressionräknar vi ut konfidensintervall på skattningar av koefficienterna $\hat{\beta}_i$ med formeln

$$\hat{\beta}_i \pm t_{\alpha/2}(S\sqrt{c_{ii}}) \qquad \text{modellparametrar}$$

där c_{ii} är variansen för skattningen av parametern β_i, vilken finns på rad i, kolumn i i varians/kovariansmatrisen för modellen (se kapitel 4). T-fördelningen har nu istället $n-d-1$ frihetsgrader, där d är antalet parametrar i den linjära modellen. Notera att det nu inte längre gäller att intervallet krymper med ökande stickprovsstorlek, då nämnaren i uttrycket försvunnit.

4 Hypotesprövning och regression

4.1 Linjär regression

En *statistisk linjär regression* är en ungefärlig lösning av ett överbestämt ekvationssystem där vi antar att en slumpvariabel, Y, kan delas upp i en linjär kombination av andra slumpvariabler, X_i, vars gemensamma fördelning är den samma som Ys. Vi kallar Y en responsvariabel och X:n för parametrar eller prediktorer.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_d X_d \qquad \text{linjär modell}$$

Detta motsvarar ett linjärt ekvationssystem på matrisform

$$\mathbf{Y} = \mathbf{AX}$$

där vi istället söker $\mathbf{A}$, koefficientmatrisen som gör att Y är en affin transformation av X:en.

Parametrarna i X kan vara vad som helst som vi misstänker kan ha en relation till responsen i Y– så länge vi kan samla ett värde för parametern för varje mätning av en viss respons i Y. Vi behöver alltså kunna mäta X:ns värden för varje mätning av Y och forma en tabell med motsvarande y och x värden på varje rad. Detta kallas en *realisation* av slumpvariablerna Y respektive X. Vi söker sedan bestämma koefficienterna β så att varje rad i det överbestämda ekvationssystemet har ett höger- och vänsterled som är så nära varann som möjligt. Mer precist är kvadraterna av avståndet mellan höger- och vänsterledet minimerat enligt minsta kvadratmetoden, det vill säga

$$min(\sum_{k=1}^{n}(Y_k - \beta_{0,k} - \beta_{1,k}X_{1,k} - ... - \beta_{d,k}X_{d,k})^2)$$
$$\iff$$
$$min(\sum^{n}(Y_k - \beta_k X_k)^2)$$

Minsta-kvadratlösningen för β-koefficienterna är en kolumnvektor $\mathbf{b} = (b_0, b_1, ..., b_d)^T$. En skattning för Y är nu precis den linjära ekvationen

$$\hat{Y} = X\mathbf{b} \tag{14}$$

där alltså b:na multipliceras in på varje rad i X för att ge en uppskattning av värdet Y borde ha. Kvadraten av skillnaden mellan Y:na vi har och de vi förutsäger kallas Sum of Square Errors (eller Residual Sum of Squares).

$$\sum_{k=1}^{n}(Y_k - \hat{Y}_k)^2 \qquad \text{SSE, RSS}$$

Medlet av dessa kallas MSE (mean squared error) och kvadratroten av det är RMSE (root mean square error). RMSE är en skattning av standardavvikelsen i medelvärdet, men är inte rättvis och kommer alltså inte nödvändigtvis gå mot något bestämt värde när stickprovsstorleken ökar.

$$\frac{1}{n}\sum_{k=1}^{n}(Y_k - \hat{Y}_k)^2 \qquad \text{MSE}$$

$$\sqrt{\frac{1}{n}\sum_{k=1}^{n}(Y_k - \hat{Y}_k)^2} \qquad \text{RMSE}$$

För mer om regressioner, se regressionskompendiumet.

4.2 Gemensam fördelning

En gemensam fördelning (joint distribution) innefattar flera slumpvariabler. Vi begränsar oss här till fallet med två variabler, men alla uttryck kan förstås utvidgas till godtyckligt antal variabler, till exempel ett Y och alla X_i i en linjär modell.

$$f_{XY}(x, y) = P[X = x \wedge Y = y] \tag{15}$$

under villkor att

$$f_{XY}(x,y) \geq 0 \text{ } \textit{för alla par} \text{ } (x,y)$$

$$\sum_{alla\ x} f_{XY}(x,y) = 1$$

Marginella sannolikheter, det vill säga sannolikhetsfunktionen i en variabel givet en gemensam fördelning är

$$f_X(x) = \sum_y f_{XY}(x,y) \tag{16}$$

$$f_Y(y) = \sum_x f_{XY}(x,y) \tag{17}$$

det vill säga, för $f_X(x)$ fixeras x och alla y:n summeras kring den x-koordinaten. Särskilt gäller att om $f_{XY}(x,y) = f_X(x)f_Y(y)$ så är slumpvariablerna X och Y *obereonde*. Om modellen är oberoende av responsen är den så dålig den kan vara – förutsägelserna har ingen koppling till det verkliga resultatet! Vi söker uttryckligen ett linjärt förhållande mellan Y och designmatrisen X. För att det skall vara möjligt så måste medlet av summan av X, dvs högerledet, vara detsamma som medlet hos Y. Uttryckt matematiskt måste alltså

$$E[Y] = E[X\boldsymbol{\beta}] = \sum (E[\beta_i X_i]) \tag{18}$$

Från linjäralgebran känner vi igen att detta innebär att $E[Y]$ måste vara en linjärkombination av $E[X]$. Om vi kan bevisa att väntevärdet av gemensamma fördelningar är linjärt så kan vi vara säkra på att denna uppdelning alltid går att göra, och att medlet blir rättvist så länge grundsanningen är ungefär linjär (inom vår valda konfidensnivå). Vi behöver först definera väntevärdet för gemensamma fördelningar:

$$E[X] = \sum_x \sum_y x f_{XY}(x,y)$$

$$E[Y] = \sum_x \sum_y y f_{XY}(x,y)$$

För väntevärden gäller att $E[cX] = cE[X]$. Därmed gäller att

$$\begin{aligned} E[\mu X + \nu Y] &= \sum_x \sum_y (\mu x + \nu y) f_{XY}(x,y) \\ &= \sum_x \sum_y \mu x f_{XY}(x,y) + \sum_x \sum_y \nu y f_{XY}(x,y) \\ &= E[\mu X] + E[\nu Y] \\ &= \mu E[X] + \nu E[Y] \end{aligned}$$

Alltså är väntevärdet E en linjär avbildning och uppdelningen av fördelningen för Y över X_i linjära komponenter kan väntas ge samma medelvärde.

4.3 Signifikans

Till vårt förfågande har vi några statistiskt mått på om en regression är användbar eller inte. Det första är ett statistiskt test för en signifikant regression. För att göra detta räknar vi ut

sannolikheten, givet vårt stickprov, att alla koefficienterna skulle vara 0, det vill säga att inga parametrar är korrelerade mot responsen. Vi uttrcker detta som att *nollhypotesen H_0 är att alla* $\beta_i = 0$. En test-statistika för detta är

$$\frac{SSR/k}{S^2} \sim F(d, n-d-1) \tag{19}$$

som följer en F-fördelning med d och $n-d-1$ frihetsgrader. För att testa detta, stoppar vi in värdet hos test-statistikan i överlevnadsfunktionen för F-fördelningen med angivna parametrar och utläser sannolikheten att alla parametrar är 0. Vi vill alltså att denna skall vara låg, så vi kan avfärda noll-hypotesen och konstatera att regressionen *är* statistiskt signifikant. För detta ensidiga F-test gäller att testet avfärdas för stora värden på statistikan. Det gäller att $SSR = S_{yy} - SSE$ då

$$S_{yy} = SSE + SSR \qquad \text{TSS}$$

Den totala variansen i responsvektorn Y kan alltså delas upp i två delar; SSE som är den residuala eller oförklarade delen runt regressionslinjen och SSR, variationen i Y som tillhör det linjära beroendet mellan parametrarna och medlet för Y. För att testa om enskilda parametrar är signifikanta, dvs om parametern är relevant, så används

$$\frac{\hat{\beta}_i}{S\sqrt{c_{ii}}} \sim T(n-d-1) \tag{20}$$

som är ett två-sidigt test och alltså avfärdar noll-hypotesen för både stora och små värden på statistikan. Sannolikheten räknas fram genom $2 * min(cdf, sf)$, dvs två gånger det minsta av fördelnings- och överlevnadsfunktionens värde för statistikan.

Ett sista men högst relevant mått är *förklaringsgraden* (coefficient of multiple determination) R^2. Denna räknas för multipel linjär regression ut som

$$R^2 = \frac{SSR}{S_{yy}}$$

och anger en proportion för hur mycket av variansen i Y som förklaras av regressionen. Detta värde ger en god indikation om vilken konfidensnivå som är lämplig. Om $R^2 = 0.96$ är till exempel 95% en lämplig konfidensnivå ($\alpha = 0.05$).