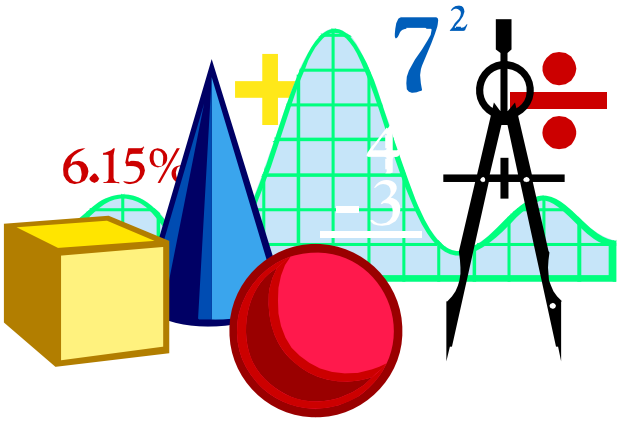

Formelsamling i statistik

Version 4.1
2004-01-19

Innehåll:

1 Beteckningar 3
2 Beskrivande statistik 4
2.1 CENTRAL- OCH SPRIDNINGSMÅTT 4
2.2 STANDARDVÄGNING 6
2.3 KORRELATION OCH REGRESSION 7
3 Sannolikhetsteori 8
3.1 RÄKNA MED SANNOLIKHETER 8
3.2 KOMBINATORIK 8
3.2 STOKASTISKA VARIABLER (SLUMPVARIABLER) 9
3.3 SANNOLIKHETSFÖRDELNINGAR 10
3.4 APPROXIMATIONSREGLER 11
3.5 STICKPROVSFÖRDELNINGAR 11
4 Statistisk slutledning 12
4.1 KONFIDENSINTERVALL 12
4.2 HYPOTESPRÖVNING 14
4.3 x^2 - TEST 16
4.4 TEST AV KORRELATION 16
4.5 ICKE PARAMETRISKA TEST 17

1 Beteckningar

Stora bokstäver, **X,Y** etc, betecknar slumpvariabler.
Små bokstäver, **x, y** etc, betecknar faktiska värden på observationer.

Populationsstorleken betecknas med **N**.
Stickprovsstorleken betecknas med **n**.

Grekiska bokstäver betecknar populationens parameter. Ex **σ**, som betecknar populationens standardavvikelse
Undantag: andelen, proportionen betecknas i vissa böcker med **p** resp $\hat{p}$

Latinska bokstäver betecknar skattningen av parametern. Den skattas utifrån stickprovet. Ex **s**, som beteckar stickprovets standardavvikelse.

	Parameter (i populationen/ sanno-likhetsfördelningen)	Parameterskattning (statistika) (i stickprovet)
Medelvärde	μ	$\bar{x}$
Varians	σ^2	s^2
Andel, proportion	p (alt π)	$\hat{p}$ (alt p)
Korrelation	ρ	r
Generellt	θ	$\hat{\theta}$

2 Beskrivande statistik

2.1 Central- och spridningsmått

Median: Värdet av den mittersta observationen vid udda antal observationer

Obs nr $\frac{N+1}{2}$ i storleksordning

alt medelvärdet av de båda mittersta observationerna vid jämnt antal observationer

Observationerna: $\frac{N}{2}$ och $\frac{N+2}{2}$

Typvärde: Det vanligaste, mest förekommande värdet i ett material

Stickprovsmedelvärde: $\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$

Populationsmedelvärde: $\mu_X = \frac{\sum_{i=1}^{N} x_i}{N}$

Variationsvidd: Differensen mellan det största och det minsta värdet, max-min

Kvartilavstånd: Avståndet mellan övre och undre kvartilen, $Q_3 - Q_1$

Kvartilavvikelse: Genomsnittligt avstånd från kvartilerna till medianen, $\frac{Q_3 - Q_1}{2}$

Stickprovsstandardavvikelse: $s_X = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{\sum x_i^2 - n \cdot \bar{x}^2}{n-1}} = \sqrt{\frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1}}$

Populationsstandardavvikelse: $\sigma_X = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu_X)^2}{N}} = \sqrt{\frac{\sum x_i^2}{N} - \mu_X^2}$

Om materialet anges i en frekvenstabell:

K = Antal olika observerade värden/olika klasser
f_i = absolut frekvens; $\sum_{i=1}^{K} f_i = n$ (stickprov)
f_i = absolut frekvens; $\sum_{i=1}^{K} f_i = N$ (population)

Stickprovsmedelvärdet: $\bar{x} = \frac{\sum_{i=1}^{K} f_i x_i}{n}$

Populationsmedelvärdet: $\mu_X = \frac{\sum_{i=1}^{K} f_i x_i}{N}$

Stickprovsstandardavvikelse: $s_X = \sqrt{\frac{\sum_{i=1}^{K} f_i (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{\sum f_i x_i^2 - n\bar{x}^2}{n-1}} = \sqrt{\frac{\sum f_i x_i^2 - \frac{\left(\sum f_i x_i\right)^2}{n}}{n-1}}$

Populationsstandardavvikelse: $\sigma_X = \sqrt{\frac{\sum_{i=1}^{K} f_i (x_i - \mu_X)^2}{N}} = \sqrt{\frac{\sum f_i x_i^2}{N} - \mu_X^2}$

Fraktilmått vid klassindelade material:

$$P_i = x_0 + w\frac{(i/100)\cdot n - kf_0}{f}$$

i = den eftersökta percentilen
P_i = variabelvärdet för den sökta percentilen
x_0 = den nedre klassgränsen där den sökta percentilen finns
w = klassbredd
kf_0 = kumulerad frekvens för x_0
(i/100)n = kumulerad frekvens för P_i
f = frekvens i klassen

2.2 Standardvägning

Problem: En variabel, y, har olika medelvärden eller olika andelar mellan kategorierna A och B. Detta datasamband ska kontrolleras för en uppdelning i klasser/ kategorier efter en tredje variabel.

Standardpopulationsmetoden:
Välj en standardpopulation och använd dess vikter genomgående. På så sätt kan vi direkt jämföra standardvägda medelvärden.

w_i = standardpopulationens vikter

$\bar{y}_i^A, \bar{y}_i^B$ = stratamedelvärden alt p_i^A, p_i^B = strataandelar

$$\bar{y}_{sv}^A = \frac{\sum_{i=1}^{K} w_i \bar{y}_i^A}{\sum_{i=1}^{K} w_i} \qquad \text{alt } p^A{}_{SV} = \frac{\sum_{i=1}^{K} w_i p_i^A}{\sum_{i=1}^{K} w_i}$$

$$\bar{y}_{sv}^B = \frac{\sum_{i=1}^{K} w_i \bar{y}_i^B}{\sum_{i=1}^{K} w_i} \qquad \text{alt } p^B{}_{SV} = \frac{\sum_{i=1}^{K} w_i p_i^B}{\sum_{i=1}^{K} w_i}$$

Kapacitetsmetoden:
Genom att beräkna hypotetiska medelvärden för de olika populationerna, utifrån en större populations medelvärden, och därefter beräkna indextal mellan de sanna och de hypotetiska medelvärdena kan populationerna på ett mer rättvist sätt jämföras.

$\bar{z}_i$ = stratamedelvärden för standardpopulationen (en större populations stratamedelvärden)
w_i = frekvenser, antal, vikter

$$\bar{y}_{hyp}^A = \frac{\sum_{i=1}^{K} w_i^A \cdot \bar{z}_i}{\sum_{i=1}^{K} w_i^A} \qquad I_A = 100 \cdot \frac{\bar{y}_{sant}^A}{\bar{y}_{hyp}^A}$$

$$\bar{y}_{hyp}^B = \frac{\sum_{i=1}^{K} w_i^B \bar{z}_i}{\sum_{i=1}^{K} w_i^B} \qquad I_B = 100 \cdot \frac{\bar{y}_{sant}^B}{\bar{y}_{hyp}^B}$$

2.3 Korrelation och regression

Kovariansen i stickprovet: $\text{Cov}[X,Y] = \frac{1}{n-1}\sum(x_i - \bar{x})(y_i - \bar{y})$

Korrelationskoefficienten i stickprovet:

$$r = \frac{Cov[X,Y]}{s_X s_Y} = \frac{1}{n-1}\sum_{i=1}^{n}\frac{(x_i - \bar{x})(y_i - \bar{y})}{s_X \cdot s_Y} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}} =$$

$$\frac{n\sum xy - \sum x\sum y}{\sqrt{(n\sum x^2 - (\sum x)^2)\cdot(n\sum y^2 - (\sum y)^2)}} = \frac{\sum xy - n\bar{x}\bar{y}}{\sqrt{\left(\sum x^2 - n\bar{x}^2\right)\left(\sum y^2 - n\bar{y}^2\right)}}$$

Korrelationskoefficienten i populationen: $\rho = \frac{Cov[X,Y]}{\sigma_X \sigma_Y}$

Regressionslinjen $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ alternativt $\hat{y} = a + b \cdot x$

$$\hat{\beta}_1 = r\frac{s_Y}{s_X} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{n\sum xy - \sum x\sum y}{n\sum x^2 - \left(\sum x\right)^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = \frac{\sum y - \hat{\beta}_1\sum x}{n}$$

3 Sannolikhetsteori

3.1 Räkna med sannolikheter

n_A, N_A= antal element med egenskapen A

$$pr(A) = \frac{n_A}{n} \qquad \text{alt} \qquad pr(A) = \frac{N_A}{N}$$

$\overline{A}$ = Komplementet till A (icke-A)

$$pr(\overline{A}) = 1 - pr(A)$$
$$pr(\overline{A}|B) = 1 - pr(A|B)$$

$$pr(A \text{ eller } B) = pr(A) + pr(B) - pr(A \text{ och } B) \qquad \text{Additionssatsen}$$

$$pr(A \text{ och } B) = pr(A|B) \cdot pr(B) \text{ eller } pr(B|A) \cdot pr(A) \qquad \text{Multiplikationssatsen}$$

$$pr(A) = \sum_{i=1}^{K} pr(E_i \text{ och } A) = \sum_{i=1}^{K} pr(E_i) \cdot pr(A|E_i) \qquad \text{Satsen om total sannolikhet}$$

$$pr(E_i|A) = \frac{pr(E_i \text{ och } A)}{pr(A)} = \frac{pr(E_i) \cdot pr(A|E_i)}{\sum_{i=1}^{K} pr(E_i) \cdot pr(A|E_i)} \qquad \text{Bayes sats}$$

Oberoende händelser
Händelserna A och B kan betraktas som oberoende händelser om:

$$pr(A|B) = P(A)$$

Detta leder till att snittet mellan två oberoende händelser kan beräknas enligt:

$$pr(A \text{ och } B) = pr(A) \cdot P(B)$$

Om A och B är oberoende händelser, så är även A och icke-B, icke-A och B samt icke-A och icke-B oberoende händelser.

Disjunkta händelser
Två händelser sägs vara disjunkta om de är varandra uteslutande, dvs de kan inte inträffa samtidigt.

$$pr(A \text{ and } B) = 0$$

3.2 Kombinatorik

Permutationer (ordningen är av betydelse)

En ordnad följd av *x* objekt valda från *n* objekt

Antal permutationer: $${}_{n}P_{x} = \frac{n!}{(n-x)!}$$

Kombinationer (ordningen saknar betydelse)

Ett urval av *x* objekt valda från *n* objekt (utan hänsyn till ordningen)

Antal kombinationer: $\binom{n}{x} = \frac{n!}{x!(n-x)!}$ *Skrivs också* ${}_nC_x$

3.2 Stokastiska variabler (slumpvariabler)

$pr(x) = pr(X = x)$

Väntevärdet $E[X] = \mu_X = \sum x \cdot pr(x)$

Variansen

$Var[X] = \sigma_X{}^2 = E\left[\left((x-\mu)^2\right)\right] = \sum (x-\mu)^2 pr(x) = \sum x^2 \cdot pr(x) - \mu^2 = E\left(X^2\right) - E\left(X\right)^2$

Standardavvikelse: $sd(X) = \sigma = \sqrt{E\left[(X-\mu)^2\right]}$

Linjära kombinationer av en slumpvariabel

Om Y = aX + b, där a och b är konstanter så är :

E(aX+b) = aE(X) + b dvs $\mu_Y = a\mu_X + b$

Var[Y] = a^2Var(X) dvs $\sigma^2_Y = a^2\sigma^2_X$

$\sigma_Y = |a|\sigma_X$

Standardisering av slumpvariabel

Om X är en slumpvariabel med medelvärdet μ och standardavvikelsen σ,
så har $\frac{X-\mu}{\sigma}$ medelvärdet = 1 och standardavvikelsen = 0.

Summor och differenser

För sinsemellan <u>oberoende</u> slumpvariabler X_1, X_2 etc gäller:

$$E(X_1 + X_2) = E(X_1) + E(X_2)$$
$$sd(X_1 + X_2) = \sqrt{(sd(X_1))^2 + (sd(X_2))^2}$$

$$E(X_1 - X_2) = E(X_1) - E(X_2)$$
$$sd(X_1 - X_2) = \sqrt{(sd(X_1))^2 + (sd(X_2))^2}$$

Medelvärde och standardavvikelse för summan av *n* observationer dragna slumpmässigt ur en fördelning med medelvärde μ och standardavvikelse σ:

$\mu_{Sum} = n\mu$ $\sigma_{Sum} = \sqrt{n}\sigma$

3.3 Sannolikhetsfördelningar

Diskreta sannolikhetsfördelningar

Binomialfördelningen, Bi(n, p).

Antal försök = n, sannolikheten för att ”lyckas” = p, variabeln X = antal lyckade av n försök.
Sannolikheterna (massfunktionen):

$$pr(X = x) = \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x} = \frac{n!}{x!(n-x)!} \cdot p^x \cdot (1-p)^{n-x} \qquad x = 0, 1, 2, ..., n$$

Väntevärde: $E(X) = \mu = n \cdot p$

Varians: $Var(X) = \sigma^2 = n \cdot p \cdot (1-p)$

Standardavvikelse: $\sigma = \sqrt{np(1-p)}$

Hypergeometriska fördelningen, Hyp(N,n,S)

N = populationens storlek, n = stickprovets storlek, S = antalet i populationen av en viss kategori, variabeln X = antal i stickprovet påträffade av kategorin. $p = S/N$

Sannolikheterna (massfunktionen): $pr(X = x) = \frac{\binom{S}{X}\binom{N-S}{n-X}}{\binom{N}{n}} = \frac{{}_sC_x \cdot {}_{N-s}C_{n-x}}{{}_NC_n}$ x = 0, 1, 2, ..., n

Väntevärde: $E(X) = \mu = n \cdot p$

Varians: $Var(X) = \sigma^2 = \left(\frac{N-n}{N-1}\right) \cdot n \cdot p \cdot (1-p)$

Standardavvikelse: $\sigma = \sqrt{\sigma^2}$

Kontinuerliga sannolikhetsfördelningar

Likformiga fördelningen, Uni(a,b)

$a \le x \le b$

Sannolikhetstäthet: $f(x) = \frac{1}{b-a}$

Väntevärde: $E(X) = \mu = \frac{a+b}{2}$ Varians: $Var(X) = \sigma^2 = \frac{(b-a)^2}{12}$

Normalfördelningen, Nf(μ, σ)

$-\infty \le x \le \infty$

Täthetsfunktion: $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Väntevärde: $E(X) = \mu$

Varians: $Var(X) = \sigma^2$

3.4 Approximationsregler

Approximera den hypergeometriska fördelningen med binomialfördelningen när urvalsfraktionen understiger 0,1. ($n/N \leq 0,1$)

Binomialfördelningen får approximeras med Normalfördelningen när $np(1-p) \geq 9$. (n stort och p litet)

Kontinuitetskorrektion förbättrar approximationen eftersom Binomialfördelningen är diskret och Normalfördelningen är kontinuerlig. X=1 i en diskret fördelning motsvarar intervallet 0,5<X<1,5 i en kontinuerlig fördelning. (Ex $X > 10 \Rightarrow X \geq 10,5$ och $X \leq 20 \Rightarrow X \leq 20,5$)

3.5 Stickprovsfördelningar

n = stickprovets storlek

Sannolikhetsfördelningen för stickprovsandelen $\hat{p}$

Medelvärdet (väntevärdet) för $\hat{p}$ är p och standardavvikelsen är $\sqrt{\frac{p(1-p)}{n}}$.

Fördelningen för $\hat{p}$ är binomial men kan då $np(1-p) > 9\ (ca.)$ approximeras med normalfördelningen.

Sannolikhetsfördelningen för stickprovsmedelvärdet $\overline{X}$

Antag att vi har en kvantitativ variabel med medelvärdet μ och standardavvikelsen σ i populationen.
Om mätvariabeln X är normalfördelad så är $\overline{X}$ normalfördelad. Om n > 30 (ca.) så är $\overline{X}$ approximativt normalfördelad även om X inte är det (enligt centrala gränsvärdessatsen).
$\overline{X}$ har väntevärdet μ och standardavvikelsen $\sigma/\sqrt{n}$

4 Statistisk slutledning

Den statistiska inferensen bygger ofta på att vi kan använda oss av normalfördelningen. Antingen kan vi förutsätta normalfördelade variabler, vilket garanterar att stickprovsmedelvärdet är normalfördelat, eller också kan vi utnyttja CGS som säger att stickprovsmedelvärdet är normalfördelat när vi har ”stora” stickprov.

Om vi har en känd populationvarians kan vi använda z-fördelningen, men då vi inte känner populationsvariansen ska vi ofta använda t-fördelningen. Denna fördelning beskrivs av ett antal frihetsgrader, *df*, beroende på hur stora stickproven är. När antalet frihetsgrader ökar tenderar t-fördelningen att bli allt mera lik z-fördelningen. Man kan använda z-fördelningen när antalet frihetsgrader överstiger 100 utan att felet blir av större betydelse.

Har vi urval från ändliga populationer och urvalsfraktionen är större än 10%, (n/N > 0,1) använder vi ändlighetskorrektion, se sidan 16.

4.1 Konfidensintervall

Konfidensintervall är uppbyggda på följande sätt:

Punktskattning ± konfidensgrad·standardfel; Där standardfelet (se) är punktskattningens standardavvikelse

Konfidensintervall för μ:

Fall 1:X är NF, σ är känd $\bar{x} \pm z_{\alpha/2} \cdot se(\bar{x}) = \bar{x} \pm z_{\alpha/2} \frac{\sigma_X}{\sqrt{n}}$

Fall 2:X är NF, σ är okänd $\bar{x} \pm t \cdot se\left(\bar{x}\right) = \bar{x} \pm t \cdot \frac{s_X}{\sqrt{n}} \qquad df = n-1$

Fall 3: Fördelningen är okänd men n ≥ 30-50 $\bar{x} \pm t \cdot se\left(\bar{x}\right) = \bar{x} \pm t \cdot \frac{s_X}{\sqrt{n}} \qquad df = n-1$

Fall 4: Fördelningen är okänd och n är stort $\bar{x} \pm z_{\alpha/2}.se\left(\bar{x}\right) = \bar{x} \pm z_{\alpha/2} \frac{s_x}{\sqrt{n}}$

Konfidensintervall för parvisa observationer, matchning:

Matchande par, X och Y är NF $\bar{d} \pm t \cdot se\left(\bar{d}\right) = \bar{d} \pm t \cdot \frac{s_d}{\sqrt{n}} \qquad df = n-1$

Matchande par, n ≥ 30 $\bar{d} \pm t \cdot se\left(\bar{d}\right) = \bar{d} \pm t \cdot \frac{s_d}{\sqrt{n}} \qquad df = n-1$

Konfidensintervall för μ_X - μ_Y, två oberoende stickprov:

Fall 1 : X och Y är NF, σ_X och σ_Y är kända $(\bar{x} - \bar{y}) \pm z_{\alpha/2} \cdot \sqrt{\frac{\sigma_X^{\,2}}{n_X} + \frac{\sigma_Y^{\,2}}{n_Y}}$

Fall 2 : X och Y är NF, $\sigma_X = \sigma_Y$ men okända, $\left(\bar{x}-\bar{y}\right) \pm t \cdot se\left(\bar{x}-\bar{y}\right) = \left(\bar{x}-\bar{y}\right) \pm t \cdot s\sqrt{\frac{n_X + n_Y}{n_X n_Y}}$

n_X eller $n_Y < 30$, $s^2 = \frac{(n_X - 1)s_X^{\,2} + (n_Y - 1)s_Y^{\,2}}{n_X + n_Y - 2}$ $df = n_X + n_Y - 2$

Fall 3 : X och Y är NF, $\sigma_X \neq \sigma_Y$ men okända, $\left(\bar{x}-\bar{y}\right) \pm t \cdot se\left(\bar{x}-\bar{y}\right) = \left(\bar{x}-\bar{y}\right) \pm t \cdot \sqrt{\frac{s_X^{\,2}}{n_X} + \frac{s_Y^{\,2}}{n_Y}}$

n_X eller $n_Y \leq 30$

$$df = \frac{\left((s_X^{\,2}/n_X) + (s_Y^{\,2}/n_Y)\right)^2}{\frac{(s^2{}_X/n_X)^2}{n_X - 1} + \frac{(s^2{}_Y/n_Y)^2}{n_Y - 1}} \quad \text{alternativt } Min\left(\left[n_X - 1\right], \left[n_Y - 1\right]\right)$$

Fall 4 : X och Y är ej NF, men n_X och $n_Y \geq 30$ $\left(\bar{x}-\bar{y}\right) \pm z \cdot se\left(\bar{x}-\bar{y}\right) = \left(\bar{x}-\bar{y}\right) \pm z \cdot \sqrt{\frac{s_X^{\,2}}{n_X} + \frac{s_Y^{\,2}}{n_Y}}$

Konfidensintervall för proportionstal:

$np(1-p) \geq 9$ $\hat{p} \pm z \cdot se\left(\hat{p}\right) = \hat{p} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Konfidensintervall för skillnaden mellan två proportionstal:

$n_1p_1(1-p_1)$ och $n_2p_2(1-p_2) \geq 9$ $\left(\hat{p}_1 - \hat{p}_2\right) \pm z \cdot se\left(\hat{p}_1 - \hat{p}_2\right) = \left(\hat{p}_1 - \hat{p}_2\right) \pm z\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

Hur man beräknar erforderlig stickprovsstorlek:

m = tillåten felmarginal (halva konfidensintervallets längd)
z är beroende av konfidensnivån
*p** ”sämsta” värde på *p* i den givna situationen

$n > \left(\frac{z \cdot \sigma}{m}\right)^2$ För proportioner gäller: $n > \left(\frac{z}{m}\right)^2 p^* \cdot \left(1 - p^*\right)$

Används ändlighetskorrektion får formlerna följande utseende:

$n > \frac{z^2\sigma^2}{m^2 + \frac{z^2 \cdot \sigma^2}{N}}$ För proportioner gäller: $n > \frac{z^2 \cdot p^* \cdot \left(1 - p^*\right)}{m^2 + \frac{z^2 p^* \cdot (1 - p^*)}{N}}$

4.2 Hypotesprövning

α = Sannolikheten att förkasta en sann nollhypotes ; Sannolikheten för typ 1 fel
β = Sannolikheten att acceptera en falsk nollhypotes ; Sannolikheten för typ 2 fel

$$\text{testvariabel} = \frac{\text{aktuell punktskattning (estimator) - hypotetisk värde}}{\text{standardfel (se)}} = \frac{\hat{\theta} - \theta}{se\left(\hat{\theta}\right)}$$

Populationsmedelvärdet μ:

Fall 1: X är NF, σ är känd Testvariabel $Z = \dfrac{\overline{X} - \mu}{\sigma / \sqrt{n}}$

Fall 2 : X är NF, σ är okänd Testvariabel: $t = \dfrac{\overline{X} - \mu}{s / \sqrt{n}}$ *df = n-1*

Fall 3 : X är ej NF, men $n \geq 30$ Testvariabel: $t = \dfrac{\overline{X} - \mu}{s / \sqrt{n}}$ *df = n-1*

Fall 4 : X är ej NF, och n är mycket stort Testvariabel: $Z = \dfrac{\overline{X} - \mu}{s/\sqrt{n}}$

Skillnad mellan två medelvärden vid parvisa observationer, matchning:

Matchande par, X och Y är NF Testvariabel: $t = \dfrac{\overline{d} - D_0}{s_d / \sqrt{n}}$ *df = n-1*

Matchande par, X och Y är ej NF men $n > 30$ Testvariabel: $t = \dfrac{\overline{d} - D_0}{s_d / \sqrt{n}}$ *df = n-1*

Skillnad mellan två populationsmedelvärden, μ_X - μ_Y, två oberoende stickprov:

Fall 1: X och Y är NF, σ_X och σ_Y är kända: $Z = \dfrac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sqrt{\dfrac{{\sigma_X}^2}{n_X} + \dfrac{{\sigma_Y}^2}{n_Y}}}$

Fall 2: X och Y är NF, $\sigma_X = \sigma_Y$ men okända :

Testvariabel: $t = \dfrac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{s\sqrt{\dfrac{n_X + n_Y}{n_X n_Y}}}$ där $s^2 = \dfrac{(n_X - 1){s_X}^2 + (n_Y - 1){s_Y}^2}{n_X + n_Y - 2}$ $df = n_X + n_Y - 2$

Fall 3: X och Y är NF, $\sigma_X \neq \sigma_Y$ men okända, n_X eller $n_Y \leq 30$: $$t = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}}$$

$$df = \frac{\left((s_X^2 / n_X) + (s_Y^2 / n_Y)\right)^2}{\frac{(s^2_X / n_X)^2}{n_X - 1} + \frac{(s^2_Y / n_Y)^2}{n_Y - 1}} \quad \text{alternativt } df = Min\left(\left[n_X - 1\right], \left[n_Y - 1\right]\right)$$

Fall 4: X och Y är ej NF, men n_X och $n_Y \geq 30$ $$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}}$$

Populationproportionen, p:

$np(1-p) \geq 9$ Testvariabel $$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

Skillnad mellan två populationsproportioner, $p_1 - p_2$:

$n_1p_1(1-p_1) \geq 9$ och $n_2p_2(1-p_2) \geq 9$ Testvariabel $$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_0(1-\hat{p}_0)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

där $$\hat{p}_0 = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

Ändlighetskorrektion:

Om vi har en ändlig population och tar ett stickprov som är mer än ca. 10 % av populationen, dvs. urvalskvoten $\frac{n}{N} > 0{,}1$, så används ändlighetskorrektion. Vi får en säkrare skattning och korrigerar därför medelfelet så att det minskar med faktorn $\sqrt{\frac{N-n}{N-1}} \approx \sqrt{1 - \frac{n}{N}}$

Ex på ändlighetskorrektion:

KI för μ: $$\bar{x} \pm z \cdot \frac{s_X}{\sqrt{n}} \cdot \sqrt{1 - \frac{n}{N}} = \bar{x} \pm z \cdot \sqrt{\frac{s_x^2}{n}\left(1 - \frac{n}{N}\right)}$$

Hypotesprövning för p: Testvariabel $$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n} \cdot \left(1 - \frac{n}{N}\right)}}$$

4.3 x^2 - test

4.3.1 **Goodness of Fit** (hur väl ansluter sig observerade värden till en given fördelning?)

Jämförelser mellan observerade (O_i) och förväntade (E_i) frekvenser.
Inga förväntade frekvenser får understiga 5

Testvariabel $\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$ Antal frihetsgrader = (r-1)

Test av oberoende

Jämförelser mellan observerade (O_{ii}) och förväntade (E_{ii}) frekvenser.
Inga förväntade frekvenser får understiga 5

	Var 2				
Var 1	1	2	. . .	c	Totalt
1	O_{11}	O_{12}	. . .	O_{1c}	R_1
2	O_{21}	O_{22}	. . .	O_{2c}	R_2
. . .	. . .	. . .	. . .	. . .	. . .
r	O_{r1}	O_{r2}	. . .	O_{rc}	R_r
Totalt	C_1	C_2	. . .	C_c	n

n = stickprovsstorlek
O_{ij} = observerad frekvens i cell i,j
R_i = radsumma för rad i
C_j = kolumnsumma för kolumn j
r = antal rader
c = antal kolumner

E_{ij} = förväntad frekvens i cell i,j = $\frac{R_i \cdot C_j}{n}$

Testvariabel $\chi^2 = \sum \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ Antal frihetsgrader = (r-1)(c-1)

4.4 Test av korrelation

r är korrelationskoefficienten mellan två normalfördelade variabler.

Testvariabel $t = \frac{r}{\sqrt{(1 - r^2)/(n-2)}}$ Antal frihetsgrader = n-2

4.5 Icke parametriska test

Icke parametriska test används när vi har små stickprov och inte kan anta att materialet är normalfördelat. Vi använder då sk fördelningsfria test.

Teckentest

Används när vi har parvisa observationer, sk matchning.
Ange för varje par om differensen är positiv eller negativ. Bortse från de fall där differensen = 0.
Räkna antalet plus- och minus-tecken.
Testfunktion: Det antal tecken som förekommer minst.
Vi testar hypotesen att andelen plustecken = 0,5
Utnyttja binomialfördelningen för att finna signifikansnivån.

Wilcoxons test

Används när vi har parvisa observationer, sk matchning.
Bilda differenser för varje par. Rangordna differenserna efter storleken på deras absolutbelopp.
Minsta differensen får rangtalet 1 osv. Vid flera lika differenser, beräkna medelvärdet av rangtalen .
Bortse från de fall där differensen = 0.
Beräkna rangsumman för de positiva och negativa differenserna.
Testfunktion: Den lägre rangsumman av de båda.
Vi testar hypotesen att medianen för de bildade differenserna = 0.
Titta i särskild tabell efter kritiska värden.