Konfidensintervall för skillnad i väntevärden

Exempel: Opinionsundersökning

	Januari		December				
Parti	Antal	Andel	Antal	Andel	Förändring	Medelfel	Konf. int
(s)	668	35.0	669	35.5	0.4	1.6	$(-2.6,\ 3.5)$
(v)	139	7.3	132	7.0	-0.3	0.8	$(-1.9,\ 1.3)$
(mp)	109	5.7	107	5.7	0	0.8	$(-1.5,\ 1.4)$
(m)	463	24.3	473	25.1	0.8	1.4	$(-1.9,\ 3.5)$
(c)	126	6.6	123	6.5	-0.1	0.8	$(-1.7,\ 1.5)$
(fp)	244	12.8	240	12.7	-0.1	1.1	$(-2.2,\ 2.1)$
(kd)	101	5.3	87	4.6	-0.7	0.7	$(-2.1,\ 0.7)$
Övr	57	3.0	55	2.9	-0.1	0.5	$(-1.2,\ 1.0)$
	1907		1886				

Låt $x = 668$ vara antalet sympatisörer av $n = 1907$ intervjuade i januari och $y = 669$ motsvarande antal av $m = 1886$ i december. Vi modellerar x som ett utfall av X, X är $\text{Bin}(n, p_1)$, och y som ett utfall av Y, Y är $\text{Bin}(m, p_2)$. Andelarna sympatisörer p_1 och p_2 skattas med $p^*_{\text{obs}} = x/n = 0.3503$ respektive $\hat{p}_{\text{obs}} = y/m = 0.3547$. Förändringen $p_1 - p_2$ skattas med $p^*_{\text{obs}} - \hat{p}_{\text{obs}} = -0.0044$. Denna skattning är ett utfall av

$$p^* - \hat{p} = \frac{X}{n} - \frac{Y}{m} \quad \text{som är approximativt} \quad \text{N}\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{m}}\right).$$

Detta ger konfidensintervallet

$$p_1 - p_2 \in p^*_{\text{obs}} \pm -\hat{p}_{\text{obs}}\lambda_{\alpha/2}\sqrt{\frac{p^*_{\text{obs}}(1-p^*_{\text{obs}})}{n} + \frac{\hat{p}_{\text{obs}}(1-\hat{p}_{\text{obs}})}{m}} = -0.044 \pm 1.96 \cdot 1.6 = (-2.0,\ 1.1)$$

med konfidensgrad $\approx 95\%$.

Låt $x_1, \ldots, x_{n_x}$ och $y_1, \ldots, y_{n_y}$ vara två observationsserier. Vi ansätter modellen att $x_1, \ldots, x_{n_x}$ är utfall av $\text{N}(\mu_x, \sigma_x)$ fördelade stokastiska variabler och $y_1, \ldots, y_{n_y}$ är utfall av $\text{N}(\mu_y, \sigma_y)$ fördelade stokastiska variabler. Alla stokastiska variabler antas vara oberoende.

Skillnaden $\mu_y - \mu_x$ skattas med $\bar{y} - \bar{x}$ där

$$\overline{Y} - \overline{X} \quad \text{är} \quad \text{N}\left(\mu_y - \mu_x, \sqrt{\frac{\sigma_y^2}{n_y} + \frac{\sigma_x^2}{n_x}}\right)$$

Alltså ett konfidensintervall med konfidensgrad $1 - \alpha$ för $\mu_y - \mu_x$ ges av

$$\mu_y - \mu_x \in \bar{y} - \bar{x} \pm \lambda_{\alpha/2}\sqrt{\frac{\sigma_y^2}{n_y} + \frac{\sigma_x^2}{n_x}} \qquad (1-\alpha)$$

Om σ_x och σ_y inte är kända kan de skattas med s_x respektive s_y. Då erhålles det (approximativa) konfidensintervallet

$$\mu_y - \mu_x \in \bar{y} - \bar{x} \pm \lambda_{\alpha/2}\sqrt{\frac{s_y^2}{n_y} + \frac{s_x^2}{n_x}} \qquad (\approx 1-\alpha)$$

Ett exakt konfidensintervall kan åstadkommas i specialfallet $\sigma_x = \sigma_y = \sigma$. Då skattas σ^2 med

$$s^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{(n_x - 1) + (n_y - 1)} = \frac{1}{n_x + n_y - 2}\left(\sum_{i=1}^{n_x}(x_i - \bar{x})^2 + \sum_{i=1}^{n_y}(y_i - \bar{y})^2\right).$$

Detta är en skattning sådan att

$$\frac{(n_x+n_y-2)S^2}{\sigma^2} = \underbrace{\frac{(n_x-1)S_x^2}{\sigma^2}}_{\in\chi^2(n_x-1)} + \underbrace{\frac{(n_y-1)S_y^2}{\sigma^2}}_{\in\chi^2(n_y-1)} \in \chi^2(n_x+n_y-2),$$

det vill säga en summa av två oberoende $\chi^2(n_x-1)$- resp. $\chi^2(n_y-1)$-fördelade stokastiska variabler och således är $\chi^2(n_x+n_y-2)$-fördelad. Detta kan utnyttjas för att sätta upp ett konfidensintervall för σ^2 (och σ) baserat på flera stickprovsvarianser.

Vidare så är

$$\frac{(\overline{Y}-\overline{X})-(\mu_y-\mu_x)}{S\sqrt{\frac{1}{n_y}+\frac{1}{n_x}}} \quad \text{är} \quad t(n_x+n_y-2)\text{-fördelad}$$

Ett exakt konfidensintervall fås som

$$\mu_y-\mu_x \in \bar{y}-\bar{x} \pm t_{\alpha/2}s\sqrt{\frac{1}{n_y}+\frac{1}{n_x}} \qquad (1-\alpha)$$

där kvantilen fås från $t(n_x+n_y-2)$-fördelningen.

Konfidensintervall för skillnad i väntevärden vid parade försök Här är $X_1,\ldots,X_n$ och $Y_1,\ldots,Y_n$ inte oberoende eller ens nödvändigtvis likafördelade. Detta uppstår typiskt vid mätningar på enheter före (X_i) och efter (Y_i) efter en insats/behandling. Modellen är att de parvisa skillnaderna

$$w_i = y_i - x_i, \quad i=1,2,\ldots,n$$

är utfall av oberoende $\mathrm{N}(\mu,\sigma)$-fördelade stokastiska variabler $W_1,\ldots,W_n$. Notera att här är

$$Y_i = X_i + W_i$$

där inga modelleringsantaganden ligger på fördelningarna för X_i och Y_i.

Hypotesprövning

Hypoteser är utsagor om parametervärden θ. Vi ställer vanligtvis upp två hypoteser:

$$H_0 : \theta \in \Theta_0,$$

en *nollhypotes*, och

$$H_1 : \theta \notin \Theta_0,$$

en *mothypotes/alternativhypotes*. Vi vill med observationer $x_1,\ldots,x_n$ försöka motbevisa utsagan H_0. Till vår hjälp har vi en teststorhet

$$t = t(x_1,\ldots,x_n)$$

som, då H_0 är sann, är ett utfall på en stokastisk variabel $T = t(X_1,\ldots,X_n)$ med känd fördelning. (Den ges oftast av en skattning av θ). Då kan vi bestämma en mängd rimliga värden för t, dvs. värden på t som är förenliga med H_0, och en mängd orimliga värden, betecknad C. Mängden C kallas *förkastelseområde* eller *kritiskt område*.

Beslutsregel: Om vi observerar $t \in C$ så förkastar vi H_0, annars inte. Vi kan tänka oss fyra scenarion:

	H_0 sann	H_0 falsk
$t \notin C$	OK.	Fel av 2:a slaget
$t \in C$	Fel av 1:a slaget	OK.

Vi vill begränsa sannolikheten för fel av första slaget, dvs sannolikheten att förkasta en korrekt nollhypotes. Vi väljer C så att

$$\mathsf{P}\left(T \in C | H_0 \text{ sann}\right) \leq \alpha$$

för ett litet α, kallat *signifikansnivån*.