F7, Samplingfördelningar

Christian Tallberg

Statistiska institutionen

Stockholms universitet

Stickprovsmedelvärdets fördelning

Exempel: Vi har en stor population med följande sannolikhetsfördelning

x	1	7	13
$p(x)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

Beräkna medelvärdet och standardavvikelsen i populationen?

$$\begin{aligned}\mu &= \sum_{alla\ x} xp(x) = 1\cdot\frac{1}{3}+7\cdot\frac{1}{3}+13\cdot\frac{1}{3}=7\\ \sigma^2 &= \sum_{alla\ x} xp(x) = (1-7)^2\cdot\frac{1}{3}+(7-7)^2\cdot\frac{1}{3}\\ &\quad +(13-7)^2\cdot\frac{1}{3}\\ &= 24\\ \sigma &= sqrt(24)\end{aligned}$$

Från denna population väljer man helt slumpmässigt och utan återläggning $n = 2$ element och beräknar $\bar{x}$. Observationerna kan ses som oberoende då populationen är stor.

Möjliga stickprov:

1,1	7,1	13,1
1,7	7,7	13,7
1,13	7,13	13,13

Möjliga stickprovsmedelvärden:

$\bar{x}$		
1	4	7
4	7	10
7	10	13

Sannolikheten för samtliga utfall är $1/3 \cdot 1/3 = 1/9$.

Samplingfördelningen (sannolikhetsfördelningen för $\bar{X}$) blir då

$\bar{x}$	1	4	7	10	13
$p(\bar{x})$	$\frac{1}{9}$	$\frac{2}{9}$	$\frac{3}{9}$	$\frac{2}{9}$	$\frac{1}{9}$

och medelvärdet, varians och standardavvikelse för $\bar{X}$

$$\begin{aligned}E\left(\bar{X}\right) &= 1\cdot\frac{1}{9}+4\cdot\frac{2}{9}+...+13\cdot\frac{1}{9}=7\\ V\left(\bar{X}\right) &= (1-7)^2\cdot\frac{1}{9}+(4-7)^2\cdot\frac{2}{9}+...\\ &\quad +(13-7)^2\cdot\frac{1}{9}\\ &= 12\\ SD\left(\bar{X}\right) &= sqrt(12)\end{aligned}$$

Vi ser att

$$\begin{aligned}E\left(\bar{X}\right) &= E(X)=\mu=7\\ V\left(\bar{X}\right) &= \frac{V(X)}{2}=\frac{\sigma^2}{2}=\frac{24}{2}\end{aligned}$$

Vi drar ett slumpmässigt urval av storlek n från en population, där $X_1, X_2, ..., X_n$ är observationer från slumpvariabeln X. Om X har medelvärdet μ och variansen σ^2, gäller för de enskilda observationerna (innan de observerats) att

$$\begin{aligned} E(X_1) &= E(X_2) = ... = E(X_n) = \mu \\ V(X_1) &= V(X_2) = ... = V(X_n) = \sigma^2. \end{aligned}$$

Låt

$$\begin{aligned} \sum X &= X_1 + X_2 + ... + X_n \text{ och} \\ \bar{X} &= \frac{\sum X}{n}. \end{aligned}$$

Vi har då att

$$\begin{aligned} E\left(\sum X\right) &= E(X_1 + X_2 + ... + X_n) \\ &= E(X_1) + E(X_2) + ... + E(X_n) \\ &= n\mu \text{ och} \\ E\left(\bar{X}\right) &= E\left(\frac{\sum X}{n}\right) = \frac{1}{n} E\left(\sum X\right) = \frac{n\mu}{n} = \mu. \end{aligned}$$

Dessutom har vi att

$$\begin{aligned} V\left(\sum X\right) &= V(X_1 + X_2 + ... + X_n) \\ &= V(X_1) + V(X_2) + ... + V(X_n) \\ &= n\sigma^2 \text{ och} \\ V\left(\bar{X}\right) &= V\left(\frac{\sum X}{n}\right) = \frac{1}{n^2} V\left(\sum X\right) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}. \end{aligned}$$

Egenskaper för stickprovsmedelvärdet ($\bar{X}$):

- $\bar{X}$ är en slumpvariabel
- väntevärdet för $\bar{X}$ sammanfaller med populationsmedelvärdet, dvs $E\left(\bar{X}\right) = E(X) = \mu$
- variansen för $\bar{X}$ är omvänt proportionell mot stickprovets storlek, dvs $V\left(\bar{X}\right) = V(X)/n = \sigma^2/n$

För ett stickprov omfattande n observationer $X_1, X_2, ..., X_n$ på en normalfördelad variabel med medelvärdet μ och standardavvikelsen σ gäller (enligt satsen om linjära kombinationer) att summan

$$\sum X = X_1 + X_2 + ... + X_n$$

är normalfördelad med

$$\begin{aligned} E\left(\sum X\right) &= n\mu \\ V\left(\sum X\right) &= n\sigma^2 \end{aligned}$$

och medelvärdet $\bar{X}$ är normalfördelat med

$$\begin{aligned} E\left(\bar{X}\right) &= \mu \\ V\left(\bar{X}\right) &= \frac{\sigma^2}{n}. \end{aligned}$$

Dvs

$$\begin{aligned} \sum X &\sim N\left(n\mu; n\sigma^2\right) \\ \bar{X} &\sim N\left(\mu; \frac{\sigma^2}{n}\right). \end{aligned}$$

Dessutom är de standardiserade variablerna

$$\begin{aligned} \frac{\sum X - n\mu}{sqrt\left(n\sigma\right)} &\sim N\left(0;1\right) \\ \frac{\bar{X}-\mu}{\frac{\sigma}{sqrt(n)}} &\sim N\left(0;1\right) \end{aligned}$$

Centrala gränsvärdessatsen (CGS)
Summan av n oberoende slumpvariabler med samma fördelning är ungefär normalfördelade om n är tillräckligt stort. (Tumregel $n \geq 30$).

Dvs

$$\begin{aligned} \sum X &\sim approx\ N\left(n\mu; n\sigma^2\right) \\ \bar{X} &\sim approx\ N\left(\mu; \frac{\sigma^2}{n}\right). \end{aligned}$$

Dessutom är de standardiserade variablerna

$$\begin{aligned} \frac{\sum X - n\mu}{sqrt\left(n\sigma\right)} &\sim approx\ N\left(0;1\right) \\ \frac{\bar{X}-\mu}{\frac{\sigma}{sqrt(n)}} &\sim approx\ N\left(0;1\right). \end{aligned}$$

Exempel:

I en stor population är medelvärdet $\mu = 65$ och standardavvikelsen $\sigma = 8$. Man väljer helt slumpmässigt ut 100 element. Vad är sannolikheten att man får ett stickprovsmedelvärde som understiger 64?

Fördelningen för X är okänd, men stickprovet är stort ($n = 100$). Alltså $\bar{X} \sim approx\ N\left(65; \frac{8}{10}\right)$ enligt CGS.

$$\begin{aligned} \Pr\left(\bar{X} < 64\right) &= \Phi\left(\frac{64-\mu}{\frac{\sigma}{sqrt(n)}}\right) = \Phi\left(\frac{64-65}{\frac{8}{10}}\right) \\ &= \Phi\left(-1.25\right) \\ &= 1 - \Phi\left(1.25\right) = 1 - 0.8944 = 0.1056 \end{aligned}$$

Binomialfördelningen kan approximeras med normalfördelning om $npq > 5$. Dvs om

$$X \sim Bin\left(n;p\right)$$

är

$$\begin{aligned} X &\sim approx\ N\left(\mu;\sigma\right) \\ \text{där } \mu &= np \\ \text{och } \sigma &= sqrt\left(npq\right) \end{aligned}$$

Utan halvkorrektion (kontinuitetskorrektion)

$$\Pr\left(a \leq X \leq b\right) \approx \Pr\left(\frac{a-np}{sqrt\left(npq\right)} \leq Z \leq \frac{b-np}{sqrt\left(npq\right)}\right)$$

Med halvkorrektion

$$\Pr\left(a \leq X \leq b\right) \approx \Pr\left(\frac{a-0.5-np}{sqrt\left(npq\right)} \leq Z \leq \frac{b+0.5-np}{sqrt\left(npq\right)}\right)$$

Exempel:

Variabeln X är $Bin\,(44; 0.45)$. Beräkna $\Pr(X \leq 26)$.

Vi har att

$$npq = 44 \cdot 0.45 \cdot 0.55 = 10.89 > 5$$

Alltså kan vi approximera med normalförd.

$$\begin{aligned} E(X) &= \mu = np = 44 \cdot 0.45 = 19.8 \\ V(X) &= \sigma^2 = npq = 44 \cdot 0.45 \cdot 0.55 = 10.89, \end{aligned}$$

dvs $X \sim approx\ N\,(19.8; 10.89)$

1. Utan halvkorrektion

$$\begin{aligned} \Pr(X \leq 26) &= \Phi\left(\frac{26 - 19.8}{sqrt\,(10.89)}\right) = \Phi\,(1.88) \\ &= 0.9699 \end{aligned}$$

2. Med halvkorrektion

$$\begin{aligned} \Pr(X \leq 26) &= \Phi\left(\frac{26 + 0.5 - 19.8}{sqrt\,(10.89)}\right) = \Phi\,(2.03) \\ &= 0.9788 \end{aligned}$$

Exakt (utan approximation)

$$\Pr(X \leq 26) = 0.9786$$

Proportionen

Vi har $0 - 1$ kodade variabler. T ex

kvinna $= 1$	man $= 0$
ja $= 1$	nej $= 0$
pos $= 1$	neg $= 0$

Låt $X \sim Bernoulli\,(p)$. X antar värdet 1 vid lyckat försök och 0 vid misslyckat försök. I en population är

$$p = \mu = \frac{\sum x}{N}$$

proportionen (andelen) ettor i populationen. Variansen för X blir

$$\begin{aligned} \sigma^2 &= V(X) = \frac{\sum (x - \mu)^2}{N} = \frac{\sum x^2}{N} - \mu^2 \\ &= \frac{\sum x}{N} - p^2 = p - p^2 = p\,(1 - p) = pq \end{aligned}$$

Låt den stokastiska variabeln $Y = \sum X$ vara antal lyckade försök (antal ettor) vid ett experiment. Då är $Y \sim Bin\,(n; p)$ (om förutsättningarna för binomialfördelningen är uppfyllda) och

$$\begin{aligned} E(Y) &= np \\ V(Y) &= npq \end{aligned}$$

Den stokastiska variabeln "proportionen ettor i stickprovet" som ges av

$$\hat{P} = \bar{X} = \frac{\sum X}{n} = \frac{Y}{n},$$

får då väntevärde och varians

$$\begin{aligned} E\left(\hat{P}\right) &= E\left(\frac{\sum X}{n}\right) = E\left(\frac{Y}{n}\right) = \frac{1}{n} E(Y) \\ &= \frac{np}{n} = p \\ V\left(\hat{P}\right) &= V\left(\frac{\sum X}{n}\right) = V\left(\frac{Y}{n}\right) = \frac{1}{n^2} V(Y) \\ &= \frac{npq}{n^2} = \frac{pq}{n}. \end{aligned}$$

Om $npq > 5$ är $\hat{P}$ approximativt $N\left(p; \frac{pq}{n}\right)$.

Exempel:

I ett adressregister är 10 % av adresserna felaktiga. Man gör ett slumpmässigt urval (OSU) av 100 adresser. Vad är sannolikheten att högst 16 % av de utvalda adresserna är felaktiga?

$\hat{P} =$ Andelen felaktiga adresser i stickprovet

$p = 0.1$

Då $npq = 100 \cdot 0.1 \cdot 0.9 = 9 > 5$, är $\hat{P} \sim approx N\left(p; \frac{pq}{n}\right)$

$$\begin{aligned} \Pr\left(\hat{P} \leq 0.16\right) &= \Phi\left(\frac{0.16 - 0.1}{sqrt\left(\frac{0.1 \cdot 0.9}{100}\right)}\right) = \Phi(2) \\ &= 0.97725 \end{aligned}$$