STOCKHOLMS UNIVERSITET
Statistiska institutionen
Jessica Franzén

TENTAMEN I INTRODUKTION TILL STATISTIK FÖR STATSVETARE

2017-04-27

Skrivtid: 10.00-15.00
Godkända hjälpmedel: Miniräknare.
Tentamen består av fem uppgifter. För full poäng på en uppgift krävs tydliga, utförliga och väl motiverade lösningar.

Uppgift 1 (20 poäng)

Det totala antalet anmälningar som inkommit till Skolinspektionen under åren 2012-2016 gällande brister i skolhälsan visas i tabellen nedan.

År	2012	2013	2014	2015	2016
Antal anmälningar	8	10	23	13	11

a) Vad har variabeln "Antal anmälningar" för variabeltyp och datanivå?

b) Räkna ut medelvärdet och medianen för antal anmälningar under den angivna tidsperioden. Förklara eventuella skillnader mellan de två lägesmåtten.

c) För att räkna ut variansen för antal anmälningar under den angivna tidsperioden använder vi formeln $\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}$ istället för formeln $S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$. Förklara varför.

d) Skapa en indexserie av variabeln "Antal anmälningar" med 2013 som bastidpunkt. Förklara hur indexserien tolkas.

Uppgift 2 (20 poäng)

Vattenverket i en kommun vill undersöka halten av järn i dricksvattnet. Man drar ett slumpmässigt urval om 40 hushåll och beräknar för dessa en genomsnittlig järnhalt på 0.16 mg/l (milligram per liter) och en standardavvikelse på 0.09 mg/l.

a) Beräkna ett 90%-igt konfidensintervall för den genomsnittliga järnhalten i kommunens dricksvatten.

b) Förklara hur ett 90%-igt konfidensintervall ska tolkas.

c) Om man vill att den totala bredden på det 90%-iga konfidensintervallet ska vara som mest 0.03, hur stort stickprov måste man då minst ta?

Uppgift 3 (20 poäng)
Vi går tillbaka till uppgift 2 och använder återigen insamlade data från de 40 hushållen. Livsmedelsverkets riktlinjer är att dricksvattnet är tjänligt utan anmärkning om järnhalten understiger 0.2 mg/l.

a) Utför ett hypotestest för att undersöka om kommunens dricksvatten är tjänligt utan anmärkning. Använd signifikansnivån $\alpha = 0.01$.

I ett hypotestest anges oftast p-värdet. Antag att p-värdet för ett test blev 0.02.

b) Förklara vad som menas med ett p-värde och hur vi använder det för att besluta om vi förkastar eller inte förkastar en nollhypotes för olika signifikansnivåer α.

Uppgift 4 (20 poäng)
En småföretagare gjorde följande sammanställning över ålder och sjukfrånvarodagar under ett år för sina 6 anställda.

Namn	A-son	B-son	C-son	D-son	E-son	F-holm
Ålder	20	31	35	44	52	60
Sjukfrånvaro (dagar)	16	7	18	23	20	28

a) Anpassa regressionslinjen $\widehat{y} = bx + a$ d.v.s. skatta a och b då y = antal sjukfrånvarodagar.

b) Tolka den skattade koefficienten b i termer av ålder och sjukfrånvaro.

c) Kan man göra en meningsfull tolkning av koefficienten a? Förklara varför eller varför inte.

d) Korrelationskoefficienten $r = 0.74$. Förklara begreppet korrelation generellt (med hjälp av en eller flera ritade bilder) samt ge en tolkning av det givna värdet på korrelationskoefficienten i detta fall.

Uppgift 5 (20 poäng)
a) Förklara begreppet determinationskoefficient

b) Förklara skillnaden mellan en diskret och en kontinuerlig fördelning. Rita ett exempel på vardera fördelning.

c) Förklara skillnaden mellan korrelation och kausalitet.

d) En tidsseriemodell kan tänkas bestå av 4 komponenter. Ange dessa fyra komponenter och förklara vad respektive komponent innebär.

e) Förklara begreppet "bias i en skattning" och vad det kan bero på.

Lycka till!

FORMELBLAD

DESKRIPTIV STATISTIK

Ett urval består av n stycken observationer.

Medelvärde:

$$\overline{x} = \frac{\sum x_i}{n}$$

Medelvärde från frekvenstabell:

$$\overline{x} = \frac{\sum f_i x_i}{n}$$

Vägt medelvärde:

$$\overline{x} = \sum_{i=1}^{L} w_i \overline{x}_i$$

Kvartiler:

$$q_1 = \frac{n+1}{4} \qquad\qquad q_3 = \frac{3(n+1)}{4}$$

Varians:

$$s^2 = \frac{\sum (x_i - \overline{x})^2}{n-1} = \frac{\sum x_i^2 - n\overline{x}^2}{n-1}$$

Varians från frekvenstabell:

$$s^2 = \frac{\sum f_i (x_i - \overline{x})^2}{n-1} = \frac{\sum f_i x_i^2 - n\overline{x}^2}{n-1}$$

VÄNTEVÄRDE OCH VARIANS

Väntevärde för X:

$$\mu = E(X) = \sum x f(x)$$

Varians för X:

$$\sigma^2 = V(X) = \sum [x - E(X)]^2 f(x) = \sum x^2 f(x) - [E(X)]^2$$

SAMPLINGFÖRDELNINGAR OCH CENTRALA GRÄNSVÄRDESSATSEN

Om populationen är normalfördelad med väntevärde μ och varians σ^2 dvs $x \sim N(\mu, \sigma^2)$

så är $\overline{x} \sim N(\mu, \frac{\sigma^2}{n})$

Om populationen har en annan fördelning (vilken som helst) med väntevärde μ och varians σ^2

så är $\overline{x} \sim N(\mu, \frac{\sigma^2}{n})$ om $n \geq 30$ enligt centrala gränsvärdessatsen

STATISTISK INFERENS (normalfördelade data eller $n \geq 30$)

Konfidensintervall:

$\overline{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ alternativt $\overline{x} \pm z_{\alpha/2}\frac{S}{\sqrt{n}}$

$P \pm z_{\alpha/2}\sqrt{\frac{P(1-P)}{n}}$

Antal observationer för en given bredd $2B$ på konfidensintervallet:

$n = \frac{Z_{\alpha/2}^2\sigma^2}{B^2}$

Hypotesprövning:

$Z = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}$ alternativt $Z = \frac{\overline{X}-\mu_0}{S/\sqrt{n}}$

$Z = \frac{P-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$

Tabellvärden för standardiserad normalfördelning:

α	Z_α
0.005	2.58
0.01	2.33
0.025	1.96
0.05	1.64
0.1	1.28

REGRESSION

Skattning av regressionslinjen $\widehat{y} = bx + a$

$$b = \frac{\sum(x_i-\overline{x})(y_i-\overline{y})}{\sum(x_i-\overline{x})^2} = \frac{n\sum x_iy_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}$$

$$a = \overline{y} - b\overline{x}$$

$$\widehat{y} = bx + a$$

Residualvarians:

$$s_e^2 = \frac{\sum e^2}{n-2} = \frac{\sum(y-\widehat{y})^2}{n-2}$$

Residualspridning:

$$s_e = \sqrt{\frac{\sum e^2}{n-2}} = \sqrt{\frac{\sum(y-\widehat{y})^2}{n-2}}$$

Korrelation:

$$r = \frac{\sum(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum(x_i-\overline{x})^2\sum(y_i-\overline{y})^2}} = \frac{n\sum x_iy_i - \sum x_i\sum y_i}{\sqrt{\left[n\sum x_i^2 - (\sum x_i)^2\right]\left[n\sum y_i^2 - (\sum y_i)^2\right]}}$$

Determinationskoefficient (enkel linjär regression):

$$R^2 = r^2$$

INDEX

$I_t = \frac{x_t}{x_0} 100$ $\qquad$ $I_t = \frac{I_t}{I_0} 100$

Laspeyres index: $\qquad$ Paasches index:

$P_t^L = \frac{\sum p_t q_0}{\sum p_0 q_0} \cdot 100$ $\qquad$ $P_t^P = \frac{\sum p_t q_t}{\sum p_0 q_t} \cdot 100$

Statistiska institutionen

Stockholms universitet

Rättningsblad

Datum: 27/4-2017

Sal: ~~Brunnsvikssalen~~ Ugglevikssalen

Tenta: Statistik för statsvetare

Kurs: Introduktion till statistik för statsvetare

ANONYMKOD: SFS-0013

☒ Jag godkänner att min tenta får läggas ut anonymt på hemsidan som studentsvar.

OBS! SKRIV ÄVEN PÅ BAKSIDAN AV SKRIVBLADEN

Markera besvarade uppgifter med kryss

1	2	3	4	5	6	7	8	9	Antal inl. blad
X	X	X	X	X					6
Lär.ant. 20	17	17	19	18,5					

SL

POÄNG	BETYG	Lärarens sign.
91,5	A	JF

SU, STATISTIK

Skrivsal: UG **Anonymkod:** SFS-0013 **Blad nr:** 1.

1. a) Kvantitativ, diskret variabel, då den är siffermässig och endast antar heltal.

Datanivån är en kvotnivå. Detta då avståndet mellan värdena på skalan går att mäta, samt att det finns en given nollpunkt. 4 anmälningar är dubbelt så mycket som 2 anmälningar osv.

5

b) Medelvärdet: Det genomsnittliga värdet i datamaterialet.

$$\bar{X} = \frac{\sum x_i}{n} = \frac{8 + 10 + 23 + 13 + 11}{5} = 13$$

Median: Det mittersta värdet i ett storleksordnat datamaterial.

8, 10, 11, 13, 23

Pågrund av värdet 23 vilket är något extremt i relation till övriga observationerna blir medelvärdet snedvridet och således högre än medianen som inte alls tar hänsyn till extremer åt något håll.

5

c) σ^2 = verkliga variansen i hela populationen (Parameter).

s^2 = variansen i ett urval. (variabel).

Då vi här utgår från data över det totala antalet anmälningar, räknar vi därför ut variansen för en totalundersökning σ^2, och inte för en urvalsundersökning, s^2

5

d)

t (År)	x (antal anmälningar)	Index med 2013 som basår
2012	8	80
2013	10	100
2014	23	230
2015	13	130
2016	11	110

$$\boxed{I_t = \frac{x_t}{x_b} \cdot 100}$$

$$I_{2012} = \frac{8}{10} \cdot 100 = 80 \qquad I_{2014} = \frac{23}{10} \cdot 100 = 230$$

$$I_{2015} = \frac{13}{10} \cdot 100 = 130 \qquad I_{2016} = \frac{11}{10} \cdot 100 = 110$$

Indexserien visar med hur mycket antalet anmälningarna skiljer sig för åren 2012, 2014, 2015 och 2016 i relation till anmälningarna vid basåret 2013. Man kan t.ex utläsa att det kom in 10% fler anmälningar år 2016 i relation till antalet 2013.

5

SU, STATISTIK

Skrivsal: UG Anonymkod: SFS-0013 Blad nr: 2

2.

$n = 40$ (40 hushåll)

$\bar{x} = 0{,}16$ (mg/liter)

$S = 0{,}09$ (mg/liter)

a) 90%-igt konfidensintervall.

Ger $\alpha = 0{,}1$

Vi tar $\frac{\alpha}{2}$ och tar ut motsvarande värde på Z_α ur tabellen över standardiserad normalfördelning.

$\frac{0{,}1}{2} = 0{,}05$.

$\alpha : 0{,}05$ ger $Z_\alpha : 1{,}64$

Sätter in värdena i formeln: $\bar{x} \pm Z_{\alpha/2} \frac{S}{\sqrt{n}}$

$0{,}16 \pm 1{,}64 \cdot \frac{0{,}09}{\sqrt{40}}$

$0{,}16 \pm 0{,}023$

Ger intervallet =

$(0{,}137 ; 0{,}183)$

7

b) Den verkliga populationsparametern, i detta fall verkliga medelvärdet, μ, befinner sig med 90% sannolikhet i intervallet (0,137 – 0,183.)

4

c) Den totala bredden på konfidensintervallet = 2B.

B är således halva bredden.

Då 2B är givet till 0,03, blir $B = \frac{0,03}{2} = 0,015$

Vi utgår från samma Z_α och S men använder oss nu av formeln för att räkna ut n vid den givna bredden.

$$n = \frac{Z_{\alpha/2}^2 \cdot (s)^2}{B^2}$$

Då vi ej har standardavvikelsen för hela populationen tar vi standardavvikelsen för vårt urval, s.

$$n = \frac{1,64^2 \cdot 0,09^2}{0,015^2} = 96,8256$$, avrundar uppåt ≈ 97.

Tolkning: Desto större stickprov, alltså desto större antal n, desto högre precision i intervallet. Tar vi 97 st i vårt urval istället för 40 st, minskar längden på intervallet. Det blir lättare att se mer precist var det verkliga värdet befinner sig. För en total bredd på 0,03 måste vi ta med minst 97 observationer i urvalet.

SU, STATISTIK

Skrivsal: UG Anonymkod: SFS-0013 Blad nr: 3

3. a) Nollhypotes: $H_0: \mu_0 = 0,2$ R

Alternativhypotes: $H_A: \mu < 0,2$ – vänstersidigt test

Sign $\alpha = 0,01$ ger enligt tabellen $z_\alpha = 2,33$

Använder följande formel för att räkna ut vår teststatistika:

$$z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Använd tidigare angivna värden för $\bar{x}$, n och s.

$\bar{x} = 0,16$

$n = 40$

$s = 0,09$

$$z = \frac{(0,16 - 0,2)}{(0,09/\sqrt{40})} = \frac{-0,04}{0,01423024 95} = -2,81$$ K

Då −2,81 är mindre än −2,33 i detta vänstersidiga test så kan vi förkasta nollhypotesen. Dricksvattnet är tjänligt utan anmärkningar på signifikansnivån $\alpha = 0,01$.

K 10

b) P-värdet säger oss vad sannolikheten är att få vårt värde på z_{obs} eller något ännu mer extremt. I detta fall är sannolikheten 2% då P-värdet = 0,02.

Är P-värdet större än signifikansnivån bör nollhypotesen förkastas, är värdet mindre bör nollhypotesen inte förkastas utan accepteras.

7

P-värdet används som komplement då olika hypotestest av samma datamaterial kan få olika resultat beroende på vald signifikansnivå.

4. a)

Ålder x_i	x_i^2	Sjukfrånvaro dagar y_i	y_i^2	$x_i \cdot y_i$
20	400	16	256	320
31	961	7	49	.
35	1225	18	.	.
44	1936	23	.	.
52	2704	20	.	.
60	3600	28	.	

$\Sigma x_i = 242$ $\quad \Sigma x_i^2 = 10826$ $\quad \Sigma y_i = 112$ $\quad \Sigma y_i^2 = 2342$ $\quad \Sigma x_i \cdot y_i = 4899$

$n = 6$

$$b = \frac{n \cdot (\Sigma x_i \cdot y_i) - (\Sigma x_i) \cdot (\Sigma y_i)}{n \cdot (\Sigma x_i^2) - (\Sigma x_i)^2}$$

$$b = \frac{6 \cdot 4899 - 242 \cdot 112}{6 \cdot 10826 - 242^2} = \frac{2290}{6392} = 0,358 \approx \underline{0,36}$$ R

SU, STATISTIK

Skrivsal: UG Anonymkod: SFS-0013 Blad nr: 4

Fortsättning 4 a) →

$a = \bar{y} - b \cdot \bar{x}$

$a = \frac{112}{6} - 0{,}36 \cdot \frac{292}{6} = \underline{4{,}15}$ R

$\hat{y} = bx + a$ — Regressionslinje

$\boxed{\hat{y} = 0{,}36x + 4{,}15}$ R

5

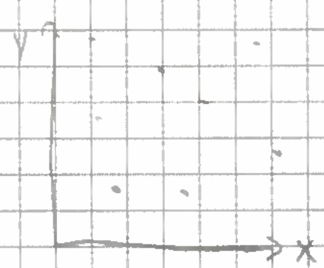
b) b = riktningskoefficient. Bestämmer om ett samband är positivt eller negativt, samt hur stor lutningen är.
b säger i detta fall att desto äldre en person är desto högre blir sjukfrånvaron, då sambandet är positivt. Lutningen är dock relativt liten vilket innebär att det inte är någon drastisk skillnad i sjukfrånvaro mellan åldrarna. Hade b varit = 1, hade detta inneburit att +1 år i ålder leder till +1 sjukdag mer/år.
Är man 20 år hade sjukfrånvaron/år blivit 24,15 dagar, 21 år inneburit 25,15 dagar osv.

5

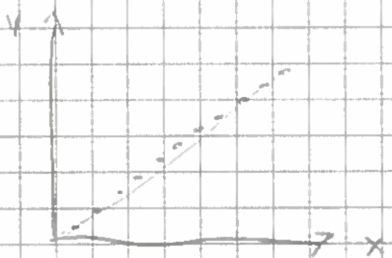
c) a = intercept. Bestämmer var linjen skär y-axeln.
Negativt värde = under origo, positivt värde = över origo.
a används för att veta var linjen ska ritas in. Är
a = 4,15 ska linjen skära y-axeln vid y = 4,15.
Man måste dock tänka kritiskt kring denna koefficient.
Om vi hypotetiskt har åldern x = 0 hade vi utifrån vår
formel [$y = 0{,}36 \cdot 0 + 4{,}15$] fått y = 4,15. Men någon
som är 0 år kan omöjligt ha 4 sjukfrånvarodagar.
Formeln måste sättas in i rätt sammanhang för att
kunna göra en meningfull tolkning.

5

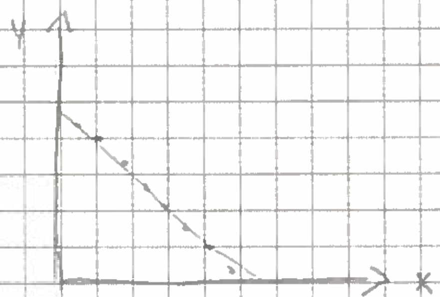
d) korrelation = hur starkt det ~~[illegible]~~ linjära sambandet är.



r = 0
Inget samband



r = 1
Perfekt positivt samband



r = -1
Perfekt negativt samband.

Då r = 0,79 innebär detta
ett ganska starkt positivt
samband mellan x och y.

4

SU, STATISTIK

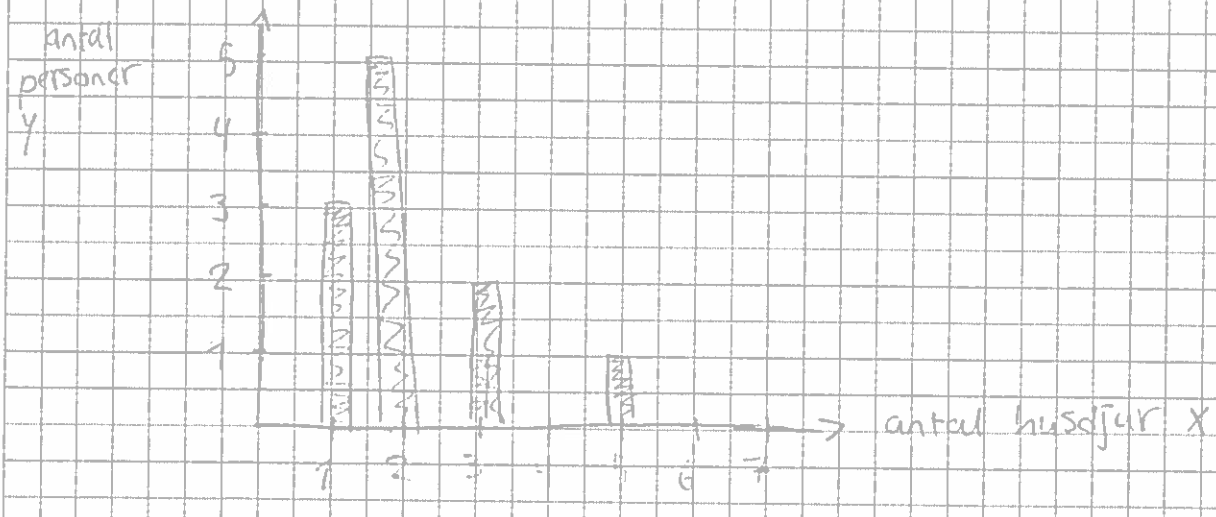
Skrivsal: UG Anonymkod: SFS-0013 Blad nr: 5

5. a) Determinationskoefficient = Förklaringsgrad. I hur stor grad förändringen i y beror på förändringen i x i den skattade regressionslijen. Skrivs som R^2, alltså korrelationen r^2. Kan anta värden mellan 0-1. Är $R^2 = 0,9$ innebär detta förändringen i y till 90% beror på förändringen i x. Övriga 10% kan bero på naturlig variation och andra förklaringsvariabler.

4

b) Diskreta fördelningar:

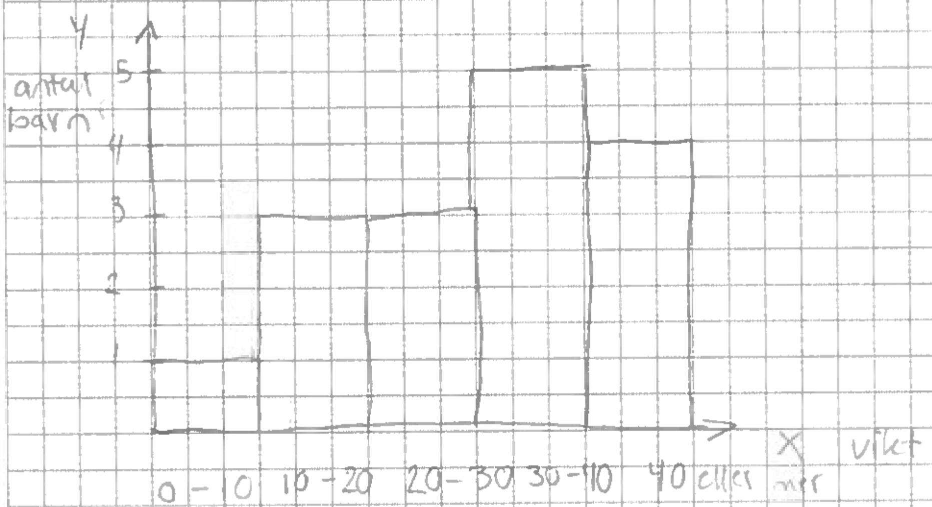
Variabeln kan endast anta vissa värden, ofta heltal. Kan t.ex illustreras genom ett stapeldiagram där y står för frekvensen (antal personer) och x står för antal husdjur/per person.



5. b) - fortsättning ->

Kontinuerliga fördelningar:

Variabeln kan anta alla tänkbara värden i ett givet intervall. Kan illustreras genom ett histogram där y står för frekvensen (antal barn) och x för vikten hos barnen.



4

c) korrelation står för huruvida ett samband finns eller inte och hur starkt detta samband i så fall är. kausalitet är ett orsaksamband där förändringen i en variabel direkt leder till förändring i en annan variabel, alltså orsak-verkan. kausalitet sker i en riktning: A leder till B men B leder inte till A. T.ex kan bra väder leda till ökad glassförsäljning, med detta innebär inte att ökad glassförsäljning leder till bra väder.

Det kan finnas en korrelation mellan två variabler utan att den ena orsakas av den andra. ->

SU, STATISTIK

Skrivsal: VG Anonymkod: SFS-0013 Blad nr: 6

5 c) - fortsättning →

T.ex kan det finnas en korrelation mellan drunkningsolyckor och glassförsäljning. Men ökat antal drunkningsolyckor orsakar inte ökad glassförsäljning. Istället orsakas dessa båda mer troligt av bra väder, vilket blir de kausala sambanden.

4

d) Tidsseriemodell med 4 komponenter:

1. Trend: tidsserien följer en trend, alltså de nya värdena baseras på hur utvecklingen hittills sett ut.
2. Slump: värdena i tidsserien beror på slumpen, kan ej kontrolleras, naturlig variation.
3. ~~Omvärldsvariabler: värdena påverkas av yttre förhållanden. Konjukturer, ekonomiska, sociala förhållanden etc.~~

2,5

e) I totalfelet ingår både urvalsfelet och systematiska fel. Bias är ett annat ord för de systematiska felen. Alltså fel som inte beror på skillnader i skattade värden för urvalet och verkliga värden i populationen orsakat av naturlig variation (urvalsfel).

Bias kan bero på en rad feltyper.

Täckningsfel: Fel målpopulation täcks in i ramen. Övertäckning då för många irrelevanta icke-relevanta täcks in. Undertäckning då individer/objekt från målpopen inte alls täcks in i rampopen.

Bortfallsfel: Fel då undersökningen saknar data/svar från individer/objekt som ingått i urvalet.

Mätfel: Fel på t.ex mätinstrumentet. En trasig våg ger fel resultat.

Bearbetningsfel: Fel under bearbetningsprocessen. Då t.ex data ska kodas och göras om till diagram, eller räknas om till olika mått.

4