Föreläsning 8, Matematisk statistik 7.5 hp för E
Punktskattningar

Stas Volkov

Matematisk statistik – slumpens matematik

Sannolikhetsteori: Hur beskriver man slumpen och slumpmässiga händelser?

- Slh. för 3 st 1:or på 10 tärningsslag?
- Givet fördelningen för vågor, hur höga/stora kan de 5 % ”värsta” vågorna vara?
- Vi observerar ett radioaktivt material med känd halveringstid under 10 mintuer; vilken fördelning kommer det observerade antalet sönderfall att följa?

Statistikteori: Vilka slutsatser kan man dra av ett datamaterial?

- Givet 3 st 1:or på 10 tärningslag, är tärningen ”rättvis”?
- Givet 10 års mätningar av vågor, vad kan vi säga om fördelningen?
- Under 10 minuter observerar vi 5 sönderfall, vad är halveringstiden?

Statistik

Från mätningar (insamlad data) dra slutsatser om verkligheten.

Vi behöver då en modell för våra mätingar!

Ofta innehåller vår modell okända parametrar samt ett antagande om fördelning för observationerna.

Exempel: Kvalitetskontroll

Vi kontrollerar n st slumpmässigt utvalda komponenter från ett stort parti och ser om de fungerar.

Modell: X =antalet trasiga komponenter
$X \underset{\sim}{\in} \mathsf{Bin}(n, p)$, där p är andelen trasiga kommponenter. Parametern p är okänd.

Möjliga frågeställlningar:

1. Vad är en bra uppskattning av p?
2. Hur stor är osäkerheten i uppskattningen?
3. Vilket intervall tror vi p ligger inom?
4. Hur stort måste n vara för att uppnå en ”tillräckligt liten” osäkerhet?

Statistikteori — översikt

Punktskattning

Hur gör man en bra gissning av en okänd storhet? Hur vet man att den är bra?

Intervallskattning

Hitta istället ett intervall som täcker den okända storheten med en given (stor) sannolikhet.

Hypotestest

Om gissningen blev 0.013, kan rätt värde på den okända storheten ändå vara 0.01?

Regression

Sambandsanalys:
hur vet vi om två variabler påverkar varandra?

Statistikteori, grundläggande begrepp

Stickprov

Ett **stickprov**, $x_1, x_2, \ldots, x_n$, är **observationer** av s.v. $X_1, \ldots, X_n$ från någon fördelning $X_i \in F(\vartheta)$ där ϑ är en okänd **parameter**.

Skattning

En **skattning** av ϑ, $\vartheta^*(x_1, \ldots, x_n)$ är en observation av den s.v. $\vartheta^*(X_1, \ldots, X_n)$. Båda betecknas oftast bara med ϑ^*.

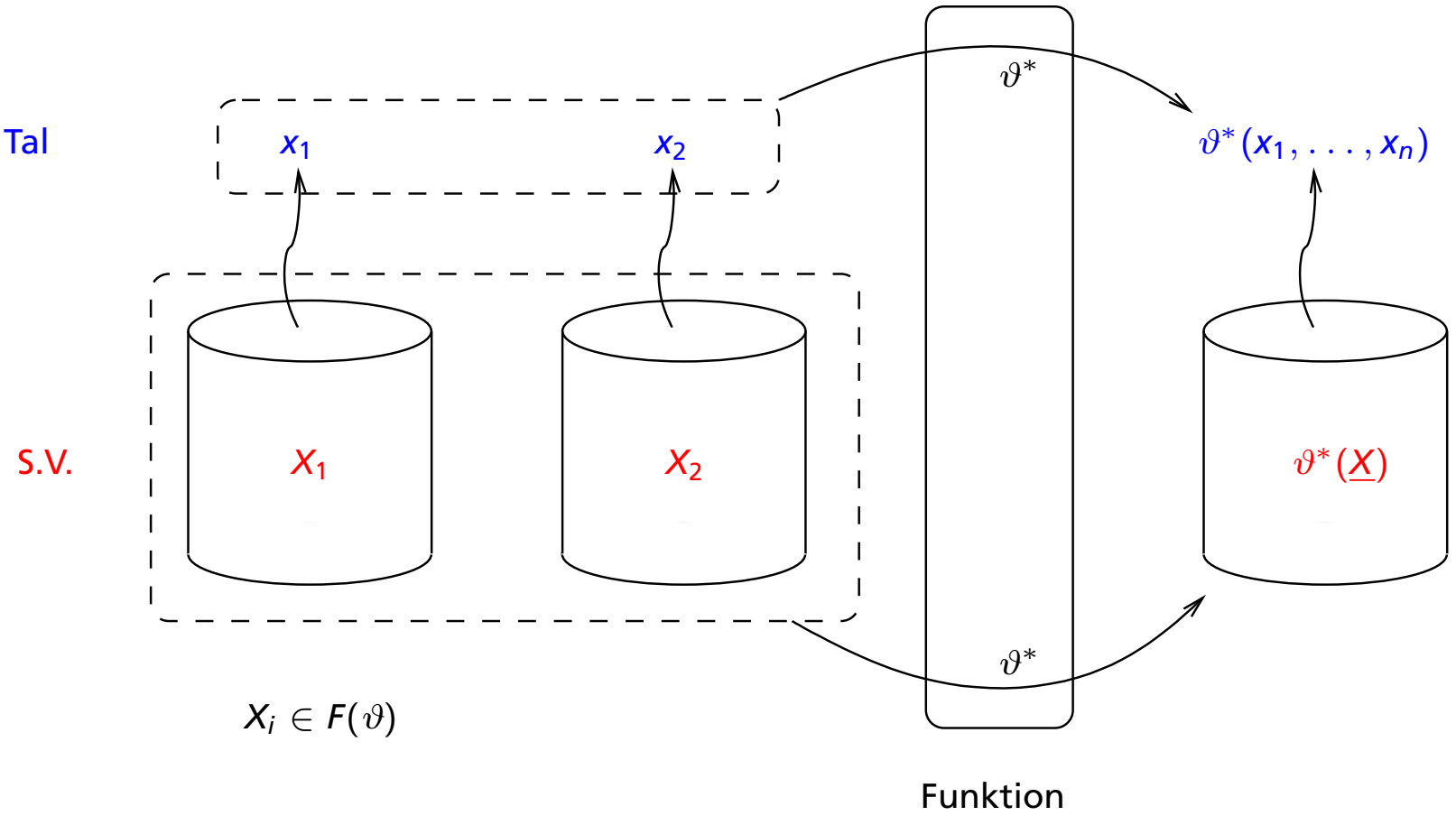
Bra egenskaper för en skattning är

Väntevärdesriktig : $E(\vartheta^*) = \vartheta$, inget systematiskt fel.

Effektiv: liten varians (osäkerhet) $V(\vartheta^*)$.

Konsistent: $P(|\vartheta_n^* - \vartheta| > \varepsilon) \to 0, \quad n \to \infty$, dvs ”Blir bättre när vi får fler observationer”,

En skattning ϑ^* är både ett tal, en s.v. och en funktion



Modell för mätning med slumpmässigt mätfel

Antag att vi vill mäta en storhet μ. Om man gör n st mätvärden, $x_1, \ldots, x_n$ är dessa observationer av

$$X_i = \mu + \varepsilon_i = \text{”Rätt värde”} + \text{”Mätfel”}$$

där ε_i är ett slumpmässigt mätfel.
Ofta antas att ε_i är oberoende och

$$\varepsilon_i \in N(0, \sigma)$$

Detta ger att våra observationer blir

$$X_i \in N(\mu, \sigma)$$

Därför ser vi att väntevärdet är den storhet vi försöker mäta upp.

Kom ihåg: väntevärde och varians

Väntevärdet anger **tyngdpunkten** för fördelningen

$$E(X) = \begin{cases} \int_{-\infty}^{\infty} x \cdot f_X(x)\,dx & \text{kontinuerlig} \\ \sum_k k \cdot p_X(k) & \text{diskret} \end{cases}$$

Variansen anger hur utspridd X är kring sitt väntevärde.

$$V(X) = E\left([X - E(X)]^2\right) = E(X^2) - E(X)^2 \geq 0.$$

Nuttiga formler:

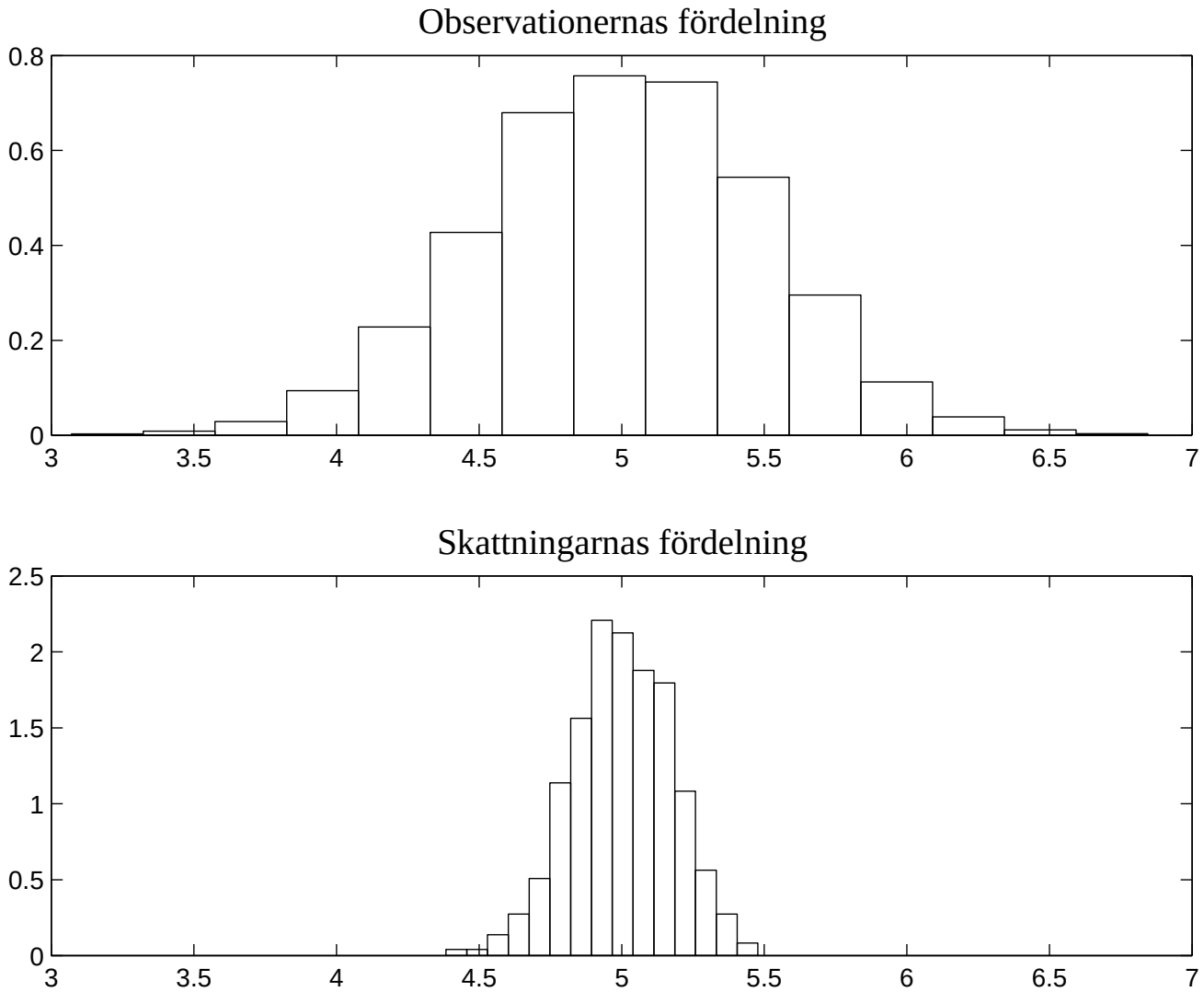
$$E\left(\sum a_i X_i + b\right) = \sum a_i E(X_i) + b$$

$$V\left(\sum_i a_i X_i + b\right) = \sum_i a_i^2 V(X_i) + 2 \underbrace{\sum_{i<j} a_i a_j C(X_i, X_j)}_{=0 \text{ om okorrelerade}}$$

Variation i observationer ger variation i skattningen

$$\mu_n^* = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \mathsf{E}(\mu_n^*) = \mu \qquad \mathsf{V}(\mu_n^*) = \frac{\sigma^2}{n}$$

Stickprov	**Observationer** x_{jk}								$\mu^* = \bar{x}_j$
1	4.83	4.93	5.24	5.12	5.10	4.69	5.62	4.73	5.03
2	5.09	5.13	4.53	4.59	4.70	4.10	4.96	5.26	4.79
3	5.53	5.10	4.34	5.05	5.21	4.43	4.30	4.56	4.82
4	4.48	5.10	4.75	5.17	4.98	5.01	5.82	5.12	5.05
5	5.14	5.10	4.79	5.48	4.70	5.89	5.22	5.91	5.28
6	4.80	5.33	5.22	5.26	4.45	4.12	5.29	5.09	4.95
7	5.20	5.26	5.49	5.60	4.83	5.28	4.38	5.18	5.15
8	4.48	4.81	4.62	4.61	5.04	4.81	4.32	4.41	4.64
⋮									


Observationernas fördelning
0.8
0.6
0.4
0.2
0
3
3.5
4
4.5
5
5.5
6
6.5
7
Skattningarnas fördelning
2.5
2
1.5
1
0.5
0
3
3.5
4
4.5
5
5.5
6
6.5
7

Minsta kvadrat-metoden, MK

Om $E(X_i) = \mu_i(\vartheta)$ så fås **MK-skattningen** av ϑ genom att **minimera förlustfunktionen**

$$Q(\vartheta) = \sum_{i=1}^{n} \Big(x_i - \mu_i(\vartheta) \Big)^2$$

med avseende på ϑ.

- Bestäm **hur** väntevärdet beror av ϑ, $E(X_i) = \mu_i(\vartheta)$.
- Sätt upp $Q(\vartheta)$
- Derivera $Q(\vartheta)$, sätt lika med noll och lös m.a.p. ϑ.
- Det ϑ som minimerar $Q(\vartheta)$ är MK-skattningen, ϑ^*_{MK}.

Maximum likelihood-metoden, ML

ML-skattningen av ϑ fås genom att **maximera likelihood-funktionen** $L(\vartheta; x_1, \ldots, x_n)$ m.a.p. ϑ.

$$L(\vartheta) = p_X(x_1) \cdot \ldots \cdot p_X(x_n) \qquad \text{(diskret)}$$
$$L(\vartheta) = f_X(x_1) \cdot \ldots \cdot f_X(x_n) \qquad \text{(kontinuerlig)}$$

I det diskreta fallet anger L-funktionen:
”Sannolikheten att få det stickprov som vi fått”.

- Sätt upp $L(\vartheta)$
- Logaritmera — $\ln L(\vartheta)$ maximeras av samma ϑ som $L(\vartheta)$.
- Derivera $\ln L(\vartheta)$ med avs. på ϑ, sätt lika med noll och lös för ϑ.
- Det ϑ som maximerar $L(\vartheta)$ är ML-skattningen ϑ^*_{ML}.

Exempel: Radon

Radonkoncentrationen i inomhusluft kan mätas genom att hänga upp en α-känslig film. Antalet hål i filmen beskrivs av en Poisson-process med

$$X_i \in \text{Po}(\mu K_i)$$

där μ är den okända radonkoncentrationen och K_i är kända konstanter som beror på bl.a. filmens känslighet, storlek och exponeringstiden.

Radon-data återkommer i lab 4.

Ex: Normalfördelning

Om $x_1, \ldots, x_n$ är observationer av $X_i \in \mathsf{N}(\mu, \sigma)$ blir ML- och MK-skattningen av μ och en korrigerad ML-skattning av σ^2

$$\mu^* = \bar{x}$$

$$(\sigma^2)^* = s^2(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Dessa används även för att skatta väntevärde och varians vid okänd fördelning.

Värför korrigerad?

$$\mathsf{E}(s^2(X)) = \sigma^2$$

Medelfel

Om $D(\vartheta^*(X)) = \sqrt{V(\vartheta^*(X))}$, dvs. standardavvikelsen av en skattning för ϑ^*, , innehåller okända parametrar ϑ då kan man inte räkna ut ett numeriskt värde på den. Istället stoppar man in skattningar på de okända parametrarna och får **medelfelet** $d(\vartheta^*)$.

Exempel: $p^* = \frac{X}{n}$ där $X \in Bin(n, p)$ så $V(X) = np(1-p)$.

$$V(p^*) = V\left(\frac{X}{n}\right) = \frac{1}{n^2}V(X) = \frac{1}{n^2}np(1-p) = \frac{p(1-p)}{n},$$

$$\text{så } D(p^*) = \sqrt{\frac{p(1-p)}{n}} \quad \Longrightarrow \quad d(p^*) = \sqrt{\frac{p^*(1-p^*)}{n}}$$

Exempel. $\mu^* = \bar{X}$, där $X \in N(\mu, \sigma)$, σ okänd

$$V(\mu^*) = \frac{\sigma^2}{n}, \quad d(\mu^*) = \frac{s}{\sqrt{n}}, \text{ där } s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Exempel: två binomialfördelningar

Vi har två oberoende observationer: x_1 från $X_1 \in Bin(n_1, p)$ och x_2 från $X_2 \in Bin(n_2, 2p)$ där n_1 och n_2 är kända medan p är en okänd parameter (OBS! $0 < p < 1/2$).

- Bestäm MK-skattningen av p.
- Bestäm ML-skattningen av p.
- Beräkna skattningarnas värde när $n_1 = 5$ och $x_1 = 2$, samt $n_2 = 6$ och $x_2 = 3$.
- Är skattningarna väntesvärderiktiga?
- Vilken av skattningarna har lägst varians?

Exempel: två binomialfördelningar

- Enkelt: skatta p och $2p$ var för sig:

$$p_1^*(x_1) = p^* = \frac{x_1}{n_1} = \frac{2}{5} = 0.40$$

$$2p_2^*(x_2) = (2p)^* = \frac{x_2}{n_2} = \frac{3}{6} = 0.50 \quad \Rightarrow p^* = \frac{x_2}{2n_2} = 0.25$$

- Vilken är rätt? (Ingen av dem...)
- Vilken är bäst? (p_1^* kan bli $> 1/2$ som är omöjligt ☹☹☹)

$$V(p_1^*) = \frac{p(1-p)}{n_1} = \frac{p(1-p)}{5};$$

$$V(p_2^*) = \frac{2p(1-2p)}{2^2 n_1} = \frac{p(1-2p)}{12}.$$

Vikta ihop skattningarna, på olika sätt...

$$p_a^* = p_a^*(x_1, x_2) = \frac{p_1^* + p_2^*}{2} = \frac{\frac{x_1}{n_1} + \frac{x_2}{2n_2}}{2} = \frac{2n_2x_1 + n_1x_2}{4n_1n_2} = 0.325$$

$$p_b^* = p_b^*(x_1, x_2) = \frac{n_1p_1^* + n_2p_2^*}{n_1 + n_2} = \frac{x_1 + \frac{x_2}{2}}{n_1 + n_2} = \frac{2x_1 + x_2}{2(n_1 + n_2)} = 0.318$$

$$p_c^* = p_c^*(x_1, x_2) = \frac{p_1^* + 2p_2^*}{1 + 2} = \frac{\frac{x_1}{n_1} + \frac{x_2}{n_2}}{3} = \frac{n_2x_1 + n_1x_2}{3n_1n_2} = 0.300$$

$$p_d^* = p_d^*(x_1, x_2) = \frac{n_1p_1^* + 2n_2p_2^*}{n_1 + 2n_2} = \frac{x_1 + x_2}{n_1 + 2n_2} = 0.294$$

MK- och ML-skattningar

$$p^*_{MK} = \frac{n_1 x_2 + 2n_2 x_1}{n_1^2 + 4n_2^2} = 0.272$$

$$p^*_{ML} = \frac{n_1 + 2n_2 + 2x_1 + x_2}{4(n_1 + n_2)} - \frac{\sqrt{(n_1 + 2n_2)^2 + (2x_1 + x_2)^2 - (4n_1 x_1 + 6n_1 x_2 + 4n_2 x_2)}}{4(n_1 + n_2)} = 0.280$$