Vad ska ni kunna till tentan?

1 Skilja på histogram för fördelningar.

diskret: Geometrisk, binomial

kontinuerlig: Γ (gamma), normal

2. Korrelation, varians/kovarians

pearsonr letar bara efter linjära förhållanden!

3. boxplots, probplots

Jämför mot normalfördelning. Hjälper att avgöra om lin. reg. är lämpligt.

4. Väntevärde, Varians, standardavvikelse

$E[X]$, $E[(X-\mu)^2]$, $\sqrt{\mathrm{Var}\,X}$

Dessa är deskriptiva mått; även kända som moment. $\sigma = \sqrt{\mathrm{Var}\,X}$, standardavvikelsen är en sorts förenkling av $\sigma^2 = \mathrm{Var}\,X$.

5. Sannolikheter, fördelningar

$f(x) = P[X = x]$ sannolikhets funktion

$F(x) = P[X \leq x]$ fördelnings funktion

6. Två definitioner av sannolikhet

relativ frekvens (stick prov): $P[A] = \frac{f}{n}$

klassisk sannolikhet (populationen): $P[A] = \frac{n(A)}{n(S)}$

$[1, 2, 3, 4, 5, 6, 6, 7, 8, 10]$

$P[A = 9] = 0$ enligt relativ frekvens.

7. Regression

Förklaringsgrad: $R^2 = \frac{SSR}{S_{yy}}$, $0 \leq R^2 \leq 1$

Total varians i Y kan delas upp enligt

$$S_{yy} = SSE + SSR$$

$$(TSS = RSS + SSR)$$

R^2 mäter "hur mycket" av datan som regressionen förklarar.

Framförallt ger R^2 en god indikation för vilken konfidensnivå vi skall välja.

8 Lin. reg. fortsättning

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_d X_d$$

Interaktionseffekter och linjärisering:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 \qquad (1)$$

$$f(\alpha_1, \alpha_2, \alpha_3) = \beta_0 + \beta_1 \alpha_1 + \beta_2 \alpha_2 + \beta_3 \alpha_3$$

$$f(x_1, x_2, x_1 x_2) = (1)$$

Enkla fördelningar

$$E[X] = \sum_x x \, f(x)$$

$A, B \sim X$

$$E[A+B] = \sum_x (x_1 - x_2) \, f(x)$$

$x_1 \in A$

$x_2 \in B$

Gemensamma fördelningar

$f_{XY}(x,y)$ sannolikhetsfunktion

Givet:

$$E[X+Y] = \sum_x \sum_y (x+y) f(x,y)$$

$$E[X] = \sum_x \sum_y x f(x,y)$$

Visa att $E[X+Y] = E[X] + E[Y]$ givet en gemensamfördelning (X, Y).

Notera att

$$E[X] = \sum_x \sum_y x f(x,y)$$

$$E[Y] = \sum_x \sum_y y f(x,y)$$

då gäller

$$E[X+Y] = \sum_x \sum_y (x+y) f_{XY}(x,y)$$

$$= \sum_x \sum_y x f(x,y) + \sum_x \sum_y y f(x,y) = E[X] + E[Y]$$

■

9. Signifikans (och hypotesprövning)

En sorts hypotesprövning.

Tex.

noll-hypotes:

$H_0: \beta = 0$

dvs alla $\beta_i = 0$

$$\frac{SSR/d}{s^2} \sim F_{d,n-d-1}$$

ensidigt: sf

$H_0: \beta_i = 0$

dvs en viss parameter är 0.

$$\frac{\hat{\beta}_i}{s\sqrt{c_{ii}}} \sim X$$ okänd fördelning

om H_0 är sann

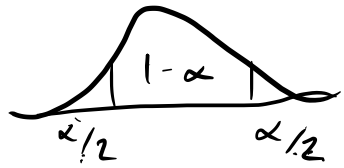
så $\sim T_{n-d-1}$

tvåsidigt test: $p = 2 \cdot \min(\text{cdf}, sf)$

10. Konfidensintervall

$$\bar{X} \pm z_{\alpha/2}\left(\sigma/\sqrt{n}\right)$$ (för medel)

Om σ och μ är kända



$z_{\alpha/2}$ hittar vi med stats.standard_normal.ppf($\alpha/2$)
eller normal.ppf($\alpha/2$, 0, 1)

$100(1-\alpha)\%$ konfidensintervall
t.ex 95% -> $\alpha = 0.05$, $\alpha/2 = 0.025$

Om σ och μ är obända

$\bar{x} \pm t_{\alpha/2}(s/\sqrt{n})$ (för medlet)

men! förfarand antagande om $X \sim N(\sigma, \mu)$

för multipel lin. reg. Välj först α!

$\hat{\beta}_i \pm t_{\alpha/2} s\sqrt{c_{ii}}$