Vad skall ni kunna till tentan (chitills)?

1. Skilja på fördelningar via histogram (även kumulativa)

$$\text{binomial} \xrightarrow{n \to \infty} \text{normal}$$

Geometrisk	binomial	normal
(gamma) överkurs	neg-binomial	standardnormal

fler:

2. korrelation, kovarians

pearson r – letar bara efter linjära förhållanden och antar att responsvariabeln är normalfördelad.

3. boxplots, probplot

jämför mot normalfördelning - hjälper i avgörandet om lin. reg. metoder är lämpliga.

4. Väntevärde, varians, standard avvikelse

$\mathbb{E}[X]$ $\quad \mathbb{E}[(X-\mu)^2]$ $\quad \sqrt{\operatorname{Var} X}$

Dessa är deskriptiva mått, även kallade momenten.
$\sigma = \sqrt{\operatorname{Var} X}$ är en sorts ytterligare förenkling av andra momentet (som strikt ~~sett~~ bara är $\mathbb{E}[X^2]$).

5. Sannolikheter och fördelningar

$f(x) = P[X=x]$ sannolikhetsfunktion

$F(a) = P[X \leq x]$ fördelningsfunktion

6. Två definitioner av sannolikhet:

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

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

not: $[1, 2, 3, 4, 5, 6, 6, 7, 8, 10]$

relativt: $P[A=9] = 0$, $P[A=6] = \frac{2}{10} = 0.2$

7. Regressioner

Förklaringsgrad: $R^2 = \frac{SSR}{S_{yy}}$ $\quad (TSS = RSS + SSR)$

R^2 mäter "hur mycket" av datan som regressionen förklarar. Ger en god indikation på vilken konfidensnivå vi borde välja.

8. Enkla fördelningar:

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

9. Gemensamma fördelningar

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

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

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

$$E[\lambda X + \mu Y] = \lambda E[X] + \mu E[Y]$$

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

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

10. Signifikans (och hypotesprövning)

$H_0: \beta = 0$

dvs Alla $\beta_i = 0$

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

ensidigt test: stats.f.sf

(survival function)

$H_0: \beta_i = 0$

$$\frac{\hat{\beta}_i}{S\sqrt{c_{ii}}} \sim X \quad \text{okänd fördelning}$$

$$\sim T_{n-d-1}$$

två-sidigt test

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

11. Konfidensintervall

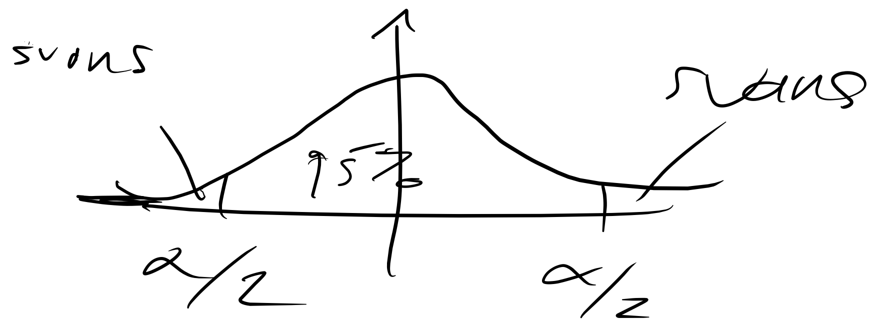
Om μ och σ är kända så är

$$\bar{X} \pm Z_{\alpha/2}(\sigma/\sqrt{n})$$ ett $100(1-\alpha)\%$ KI för medlet

$Z_{\alpha/2}$ hittar vi med stats.standard_normal.ppf($\alpha/2$)

= stats.normal($\alpha/2$, 0, 1)

tex: 95% KI $\rightarrow$ $\alpha = 0.05$ $\alpha/2 = 0.025$



Om σ, μ är okända

$\bar{X} \pm t_{\alpha/2} (S/\sqrt{n})$ är ett $100(1-\alpha)\%$ KI för medlet.

men! fortfarande antagande om att $Y \sim N(\mu, \sigma)$

där S är en rättvis skattning av σ.

För prediktorer/parametrar:

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

Välj först α!

12. $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$

Om interaktionseffekten är signifikant (dvs $H_0: \beta_3 = 0$ avvisas genom test) så skall β_1 och β_2 inkluderas i modellen oavsett deras signifikans.

Hierarkiprincipen