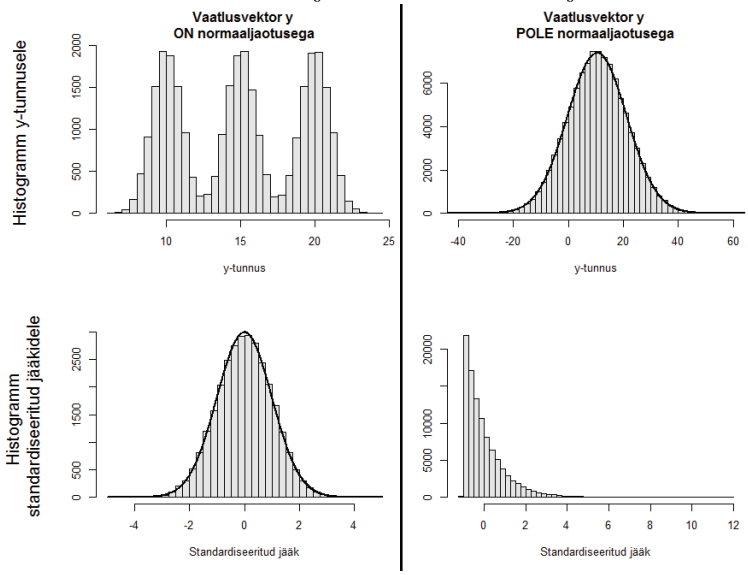
7.6 Normaaljaotuse eeldusest

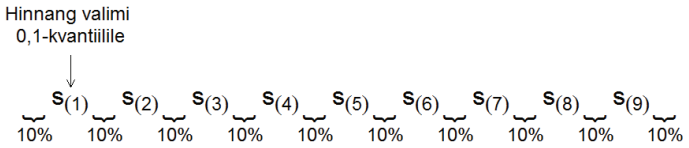
Mitmete tulemuste saamiseks (F-test, t-test, prognoosi- ja usaldusintervallid) oleme eeldanud, et vaatlusvektor $\mathbf{y}$ on normaaljaotusega juhuslik suurus. Selle eelduse kontrollimisel peame siiski meeles hoidma, et tegemist on mitmemõõtmelise normaaljatusega (igal vaatlusel on oma keskväärtus). Kuna kõik vaatlusvektori elemendid pole sama normaaljaotusega, siis ei saa me vaatlusvektori $\mathbf{y}$ elemente kasutades lihtsalt histogrammi joonistada ja selle põhjal otsustada normaaljaotuse eelduse kehtivuse üle — sest normaaljaotuste segu ei ole normaaljaotusega! Histogramm (või teisedki normaaljaotuse eelduse kontrollimiseks mõeldud tehnikad) eeldavad, et vaatlused oleksid sama normaaljaotusega. Kuidas teisendada erinevaid normaaljatuseid samaks normaaljaotuseks? Selleks lahutame esmalt igast vaatlusest maha tema keskväärtuse (leiame mudeli jäägid). Jääkide keskväärtus (kui mudeli kuju on õige) on 0. Seejärel peame jäägid jagama tema oodatava standardhälbega — sest (hinnatud) jääkide hajuvused on erinevad. Seega peame normaaljaotuse eelduse kontrollimiseks kasutama standardiseeritud jääke — standardiseeritud jäägid peaksid olema kõik sama jaotusega (normaaljaotusega keskväärtusega 0, dispersiooniga 1). Seda muidugi vaid siis, kui mudeli kuju ja jääkide konstantse dispersiooni eeldused kehtivad. Vaata ka joonist 7.1.

Joonis 7.1: Normaaljatuse eeldus. Kaks juhtumit.



Histogramm pole muidugi mitte kõige mugavam graafik jaotuse eelduse kontrollimiseks. Üheks levinuimaks graafikuks on tõenäosuspaber ehk kvantiil-kvantiil graafik (*Q-Q plot*). Kvantiil-kvantiil graafikutelt on märksa kergem märgata kõrvalekaldeid eeldatavast jaotusest, eriti kui kõrvalekalded leiavad pigem aset jaotuste sabaosas (tüüpiline häda praktikas). Sellel graafikul kujutatakse y-teljel standardiseeritud jääkide variatsioonrea elemente (mida võib vaadata kui mitteparameetrilisi hinnanguid kvantiilidele, vaata ka joonist 7.2), y-teljele aga kantakse variatsioonirea elementide oodatavad väärtused (ehk keskväärtused), mis on arvutatud eeldades, et standardiseeritud jäägid on standardse normaaljatusega.

Joonis 7.2: Variatsioonrea element kui kvantiili hinnang. Näide $n = 9$ jaoks.



Kuna variatsioonrea elementide keskväärtust on täpselt leida üsna tülikas (seda on tehtud mõnede valimi suuruste jaoks), siis üritatakse arvutust lihtsustada kasutades ligikaudseid arvutusvalemeid. Nimelt kui vaatame valimit suurusega n, siis on suuruselt i. standardiseeritud jäägi $s_{(i)}$ keskväärtus (juhul kui jääkide jaotuseks on normaaljaotus) ligikaudu leitav kui (kasutusel on erinevaid lähendeid):

$$\begin{aligned} \mathrm{E}(s_{(i)}) &\approx z_{i/(n+1)} \\ &\approx z_{(i-3/8)/(n+1-2\cdot 3/8)} \\ &\approx z_{(i-1/2)/(n+1-2\cdot 1/2)}, \end{aligned}$$

Kus z_α tähistab standardse normaaljaotuse α-kvantiili.

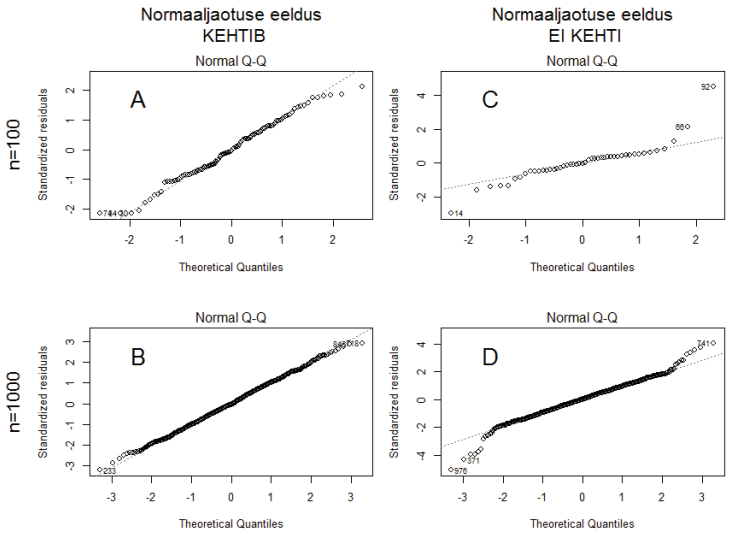
Variatsioonrea viie esimese elemendi keskväärtused ja nende lähendid standardse normaaljatuse kvantiilide abil juhu $n = 9$ jaoks:

i	$Es_{(i)}$	$z_{i/(n+1)}$	$z_{(i-3/8)/(n+1-2\cdot 3/8)}$
1	-1.485	-1.282	-1.494
2	-0.932	-0.842	-0.932
3	-0.572	-0.524	-0.572
4	-0.275	-0.253	-0.274
5	0	0	0

Märkus: need keskväärtused on leitud eeldades vaatluste sõltumatust. Standardiseeritud jäägid pole sõltumatud, kuid enamasti on sõltuvused piisavalt väikesed, et neid ignoreerida.

Kui standardiseeritud jääkide jaotuseks on tegelikult ka normaaljaotus, siis (järjestatud) standardiseeritud jäägid ja nende oodatavad väärtused (leitud eeldades normaaljaotust) peaksid tulema sarnased. Seega kvantiil-kvantiil graafik võiks normaaljaotuse eelduse kehtides olla ligilähedaselt sirge (vaata joonist 7.3). Kui kvantiil-kvantiil graafik pole küll täiesti sirge, aga püsib referentsjoonel näiteks vahemikus $-1{,}96 \ldots 1{,}96$ (joonis 7.3, D) siis käituvad jääkide kvantiilid vahemikus 0,025-kvantiilist kuni 0,975-kvantiilini nii nagu nad normaaljaotuse puhul peaksid käituma ja näiteks normaaljaotuse eeldust kasutades leitud 0,95-prognoosiintervalli võib siiski usaldada.

Joonis 7.3: Kvantiil-kvantiil graafikud. Neli andmestikku.



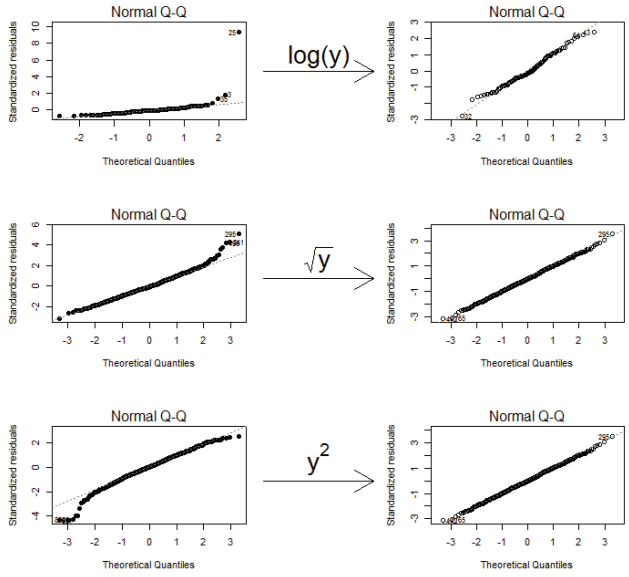
Kui oluline on normaaljaotuse eeldus F-testi ja t-testi jaoks? Ühe kasuliku vihje võib leida järgmisest artiklist:

Arnold, S. F. (1980). Asymptotic validity of F tests for the ordinary linear model and the multiple correlation model. *Journal of the American Statistical Association*, 75(372), 890-894.

Selles artiklis näidatakse, et kui maksimaalne mõjukus koondub nulliks (ja kui $\text{rank}(X)/n \to 0$ ning mudeli jääkide dispersioon on lõpmatusest väiksem) siis F-test ja t-test jäävad asümptootiliselt korrektseteks testideks (ja ka usaldusintervallid keskväärtusele on asümptootiliselt korrektsed) ka siis, kui vaatlused pole normaaljaotusega. Samas näiteks usaldusintervallid jääkide dispersioonile ei ole isegi asümptootiliselt korrektsed kui jääkide jaotuseks pole normaaljaotus.

Väiksemate valimite (ja näiteks prognoosiintervallidega) jääb normaaljaotuse eeldus ikkagi oluliseks. Mida siis teha, kui uuritava tunnuse jaotuseks pole normaaljaotus? Üheks võimaluseks on uuritavat tunnust transformeerida, kasutada näiteks lineaarses mudelis sõltuva tunnusena $\log(y)$-t y-i asemel, vaata ka joonist 7.4.

Joonis 7.4: Enamkasutatavad transformatsioonid



Aga vahel võib ka kogenud praktikul esineda raskuseid sobiva transformatsiooni leidmisel (rääkimata siis oma lennuka karjääri alguses olevast tudengist). Sestap võib vahel abi otsida (pool)automaatsetest meetoditest, näiteks kasutada Box-Cox'i transformatsiooni.

7.6.1 Box-Cox'i transformatsioon

Box-Cox'i transformatsioon otsib sobivaimat astmefunktsiooni (kuid kaalutakse ka logaritmi), mis transformeeriks y-tunnuse nii, et transformeeritud tunnust kasutava mudeli jääkide jaotus oleks võimalikult lähedane normaaljaotusele. Täpsemalt, otsitakse transformatsiooni kujul:

$$Y^* = \begin{cases} (Y^\lambda - 1)/\lambda & \lambda \neq 0 \\ \ln(Y) & \lambda = 0 \end{cases}$$

Box-Cox'i transformatsiooni saab kasutada siis, kui y-tunnuse väärtused on positiivsed.

Miks on otsitav transformatsiooni just sellisel kujul? Antud transformatsioon sõltub ühestainsast parameetrist λ. Teades λ väärtust teame ka otsitavat transformatsiooni. Aga soovitakse, et transformatsioon oleks otsitava parameetri λ suhtes pidev funktsioon ja ühe parameetri väärtuse ($\lambda = 0$) korral saaksime tulemuseks ka sageli kasutatava log-transformatsiooni. Valitud kuju puhul on transformatsiooni piirväärtus juhul kui $\lambda \to 0$ võrdne logaritmiga (kasutame L'Hospitali reeglit ja reeglit $\frac{\partial a^x}{\partial x} = a^x \ln(a)$):

$$\begin{aligned} \lim_{\lambda \to 0} (Y^\lambda - 1)/\lambda &= \lim_{\lambda \to 0} \frac{\partial (Y^\lambda - 1)/\partial \lambda}{\partial \lambda / \partial \lambda} \\ &= \lim_{\lambda \to 0} \frac{Y^\lambda \cdot \ln(Y)}{1} \\ &= \ln(Y). \end{aligned}$$

Aga kuidas leida parameetri λ väärtust? Suurima tõepära meetodil muidugi. Eeldame, et peale edukat transformatsiooni on transformeeritud tunnus normaaljaotusega:

$$\mathbf{y}^* \sim N(\mathbf{X}\boldsymbol{\beta}; \sigma^2 \mathbf{I}).$$

ehk transformeeritud valimivektori tihedus avaldub kujul:

$$f_{\mathbf{y}^*} = |2\pi\sigma\mathbf{I}|^{-0,5} \exp\left(-0,5(\mathbf{y}^* - \mathbf{X}\boldsymbol{\beta})^T (\sigma^2\mathbf{I})^{-1} (\mathbf{y}^* - \mathbf{X}\boldsymbol{\beta})\right).$$

Lähme siit üle esialgse juhusliku suuruse tiheduse peale:

$$\begin{aligned} f_{\mathbf{y}} &= |2\pi\sigma\mathbf{I}|^{-0,5} \exp\left(-0,5(\mathbf{y}^* - \mathbf{X}\boldsymbol{\beta})^T (\sigma^2\mathbf{I})^{-1} (\mathbf{y}^* - \mathbf{X}\boldsymbol{\beta})\right) \cdot |J(\mathbf{y}, \boldsymbol{\lambda})| \\ &= |2\pi\sigma\mathbf{I}|^{-0,5} \exp\left(-0,5(\mathbf{y}^* - \mathbf{X}\boldsymbol{\beta})^T (\sigma^2\mathbf{I})^{-1} (\mathbf{y}^* - \mathbf{X}\boldsymbol{\beta})\right) \cdot \prod_{i=1}^{n} y_i^{\lambda - 1}. \end{aligned}$$

Saadud tihedust võime vaadelda ka kui tõepärafunktsiooni (vaatleme seda $\mathbf{y}$ funktsiooni asemel tundmatute parameetrite funktsioonina) ja me võime leida, milline λ väärtus maksimiseerib selle tõepära (näiteks kasutades numbrilisi meetodeid).

R-is saab parimat transformatsiooni otsida käsu boxcox abil:

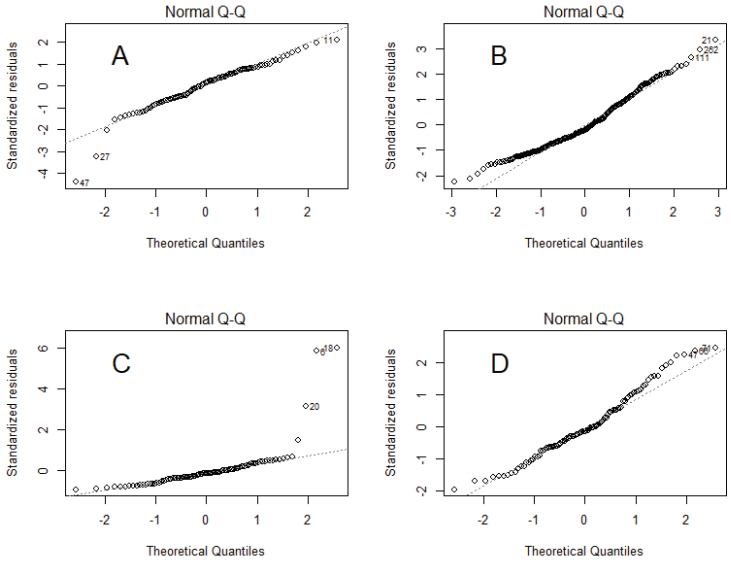
```
library(MASS)
boxcox(y~x, lambda=seq(-1,2,0.01))
```

Märksõna lambda= abil saab määrata, milliseid lambda-väärtuseid tõepära maksimiseerimisel proovitakse. Juhul kui parimat transformatsiooni näitava *lambda*-parameetri väärtus tuleb sarnane mõnele tuntud transformatsioonile (näiteks $\lambda = 0{,}092$ või $\lambda = 0{,}492$) siis soovitatakse pigem kasutada lähedast tuntud transformatsiooni (vastavalt siis logaritmi või ruutjuurt).

Ülesanded

Joonisel 7.5 on kujutatud standardiseeritud jääkide jaoks joonistatud tõenäosuspabereid. Milline transformatsioon (logaritmi vaatluseid, tõsta vaatlused ruutu, kasuta ruutjuurt või ära tee midagi) võiks tagada, et peale transformatsiooni normaaljaotuse eeldus oleks rahuldatud? Leia ka iga tõenäosuspaberi jaoks vastav histogramm jooniselt 7.6.

Joonis 7.5: Tõenäosuspaberid



Joonis 7.6: Standardiseeritud jääkide histogrammid