Statistik för Biologer
F10: Repetition

Shaobo Jin

Matematiska institutionen

Från kursplanen: kursmål för statistik

Efter godkänd kurs ska studenten kunna:

1. redogöra för grunderna till statistiska undersökningar och ha kunskap om några metoder för beskrivande statistik;
2. uppvisa en grundläggande förtrogenhet med statistiska begrepp och metoder som kan förekomma i kvantitativ biologi, och en allmän förståelse för hur statistik kan tillämpas inom några områden av biologi;
3. använda enklare matematisk och statistisk programvara.

Innehåll: Statistik

1. Population, stickprov, naturlig variation. Beskrivande statistik.
2. Skattning av väntevärde, varians och standardavvikelse.
3. Replikat av försök. Allmänt om sampling.
4. Diskreta och kontinuerliga data.
    1. normalfördelningen,
    2. t-fördelningen,
    3. något om Poisson-, exponential- och chi2-fördelningarna.
5. Idéer bakom hypotesprövning. Statistiska test,
    1. binomialfördelning och teckentest.
    2. Test vid en eller två normalfördelningar. Stickprov i par.
    3. Chi2-test.
    4. Wilcoxons rangsummetest.
    5. Envägs och tvåvägs variansanalys,
    6. Multipla jämförelser.
6. Korrelation. Enkel linjär regression.
7. randomiserade block.
8. Matematisk programvara.

Tentan

Delar av materialet har redan examinerats via inlämningsuppgifterna och deltagande vid datorövningarna.

Vad ska examineras vid tentan är

1. Idéer bakom hypotesprövning. Statistiska test,
   1. binomialfördelning och teckentest.
   2. Test vid en eller två normalfördelningar. Stickprov i par.
   3. Chi2-test.
   4. Wilcoxons rangsummetest.
   5. Envägs och tvåvägs variansanalys,
   6. Multipla jämförelser.
2. Korrelation. Enkel linjär regression.
3. Matematisk programvara: R

Sammanfattning

Se till att ni kan använda R för följande saker

1. Idéer bakom hypotesprövning. Statistiska test,
   1. binomialfördelning och teckentest: dbinom() och pbinom()
   2. Test vid en eller två normalfördelningar. Stickprov i par: t.test() och boot.t.test()
   3. Chi2-test: chisq.test() och cor.test(). En variant är fisher.text().
   4. Wilcoxons rangsummetest: wilcox.test()
   5. Envägs och tvåvägs variansanalys: aov() och ANOVA.boot()
   6. Multipla jämförelser: TukeyHSD() och p.adjust()
2. Korrelation. Enkel linjär regression: cor(), cor.test(), lm(), boot_summary() och predict()

Ni ska kunna

- Du ska kunna tolka resultaten, samt veta när de olika metoderna används och för vilka sorters problem.
  - Uppgifterna kommer att ha formuleringar i stil med “finns det någon skillnad mellan gruppernas medelvärden?” och inte “gör ett t-test”. Du ska själv lista ut vilken metod som passar.
- Du ska kunna kontrollera förutsättningar.
  - Du ska själv lista ut vilken metod som passar, t.ex, är det rimligt att använda `t.test()`? Är det bättre att använda `wilcox.test()`?

Statistikproblem

1. Statistikdelen av tentan kommer att bestå av tre problem.
   1. Mattedel: 25 poäng.
   2. Statistikdel: 15 poäng.
2. För godkänt: minst 21 poäng totalt, varav minst 10 i matte och 5 i statistik.
3. Du får och bör använda R för att lösa problemen.
4. Du kommer att ha tillgång till GM.
5. Du lämnar in dina lösningar och din R-kod för statistikdelen **digitalt**!
   - Båda behövs!

Hur ska en lösning se ut?

1. Alltid: ange tydligt vilken metod du använder
   - "Jag använder ett t-test för att jämföra grupperna."
2. Vid hypotesprövning: ange vilka hypoteser som testas
   - "Jag testar H_0: ingen skillnad mellan gruppernas medelvärden mot H_1: gruppernas medelvärden skiljer sig åt."
3. Vid hypotesprövning: ange vilken signifikansnivå du använder
   - "p-värdet blir 0,03. Skillnaden är signifikant vid 5% signifikansnivå."
   - Det räcker inte med just "Skillnaden är signifikant".
4. Alltid: tolka resultatet
   - "Det finns en skillnad i vingslagsfrekvens mellan arterna."

Hur ska den inlämnade koden se ut?

1. Du får använda vilka funktioner i R du vill för att lösa problemen.
2. Jag rekommenderar att du kommenterar din kod med förklaringar på vad du gör i de olika stegen.
   - Det underlättar din egen förståelse
   - Det gör det lättare för rättande lärare att förstå hur du har tänkt

Exempel:

```
# Beraekna medelvaerdet av vikt
mean(body_mass_g)
```

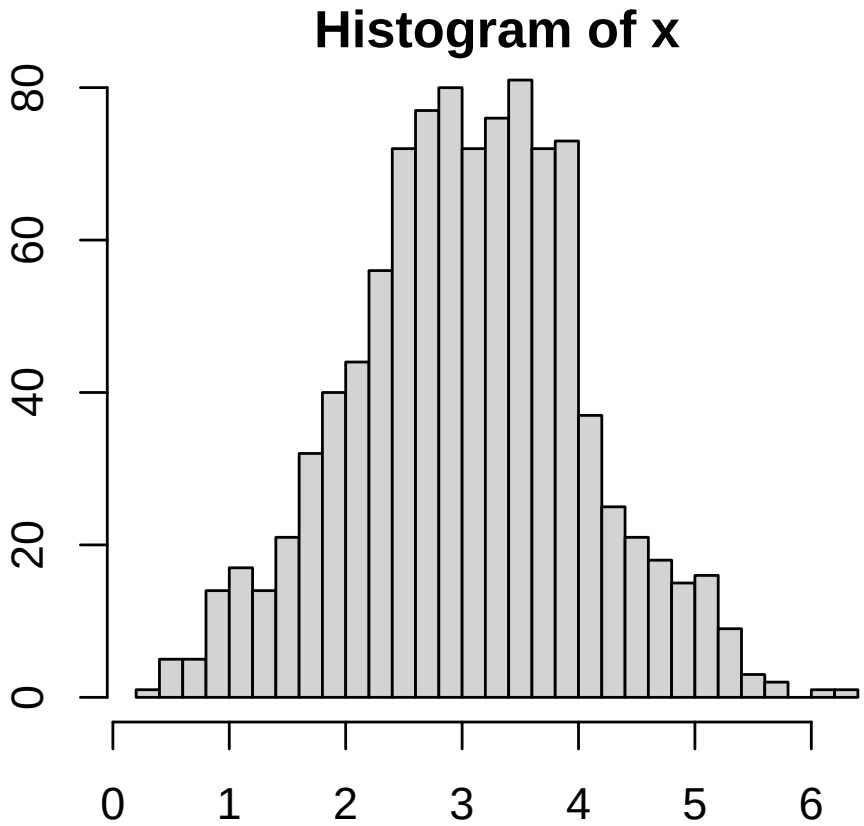
Gråzoner

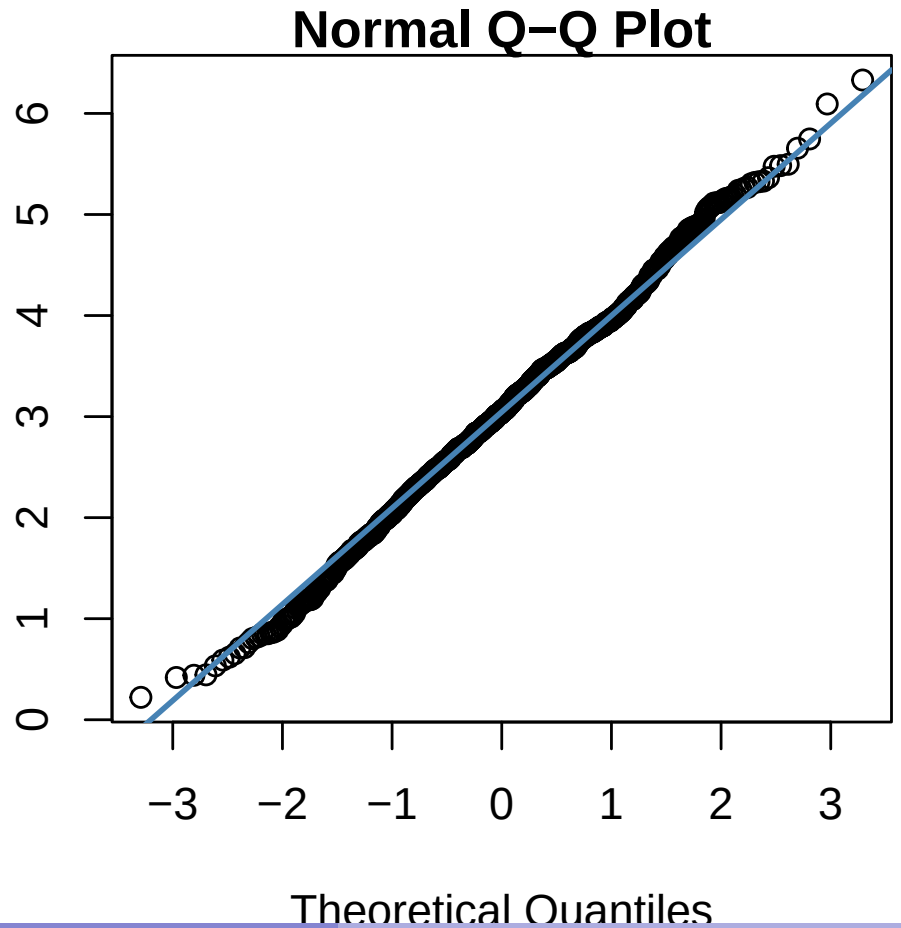
I många delar av statistiken behöver vi göra bedömningar:

- Är data normalfördelade?
- Är sambandet linjärt?
- Har vi några outliers?
- Osv.

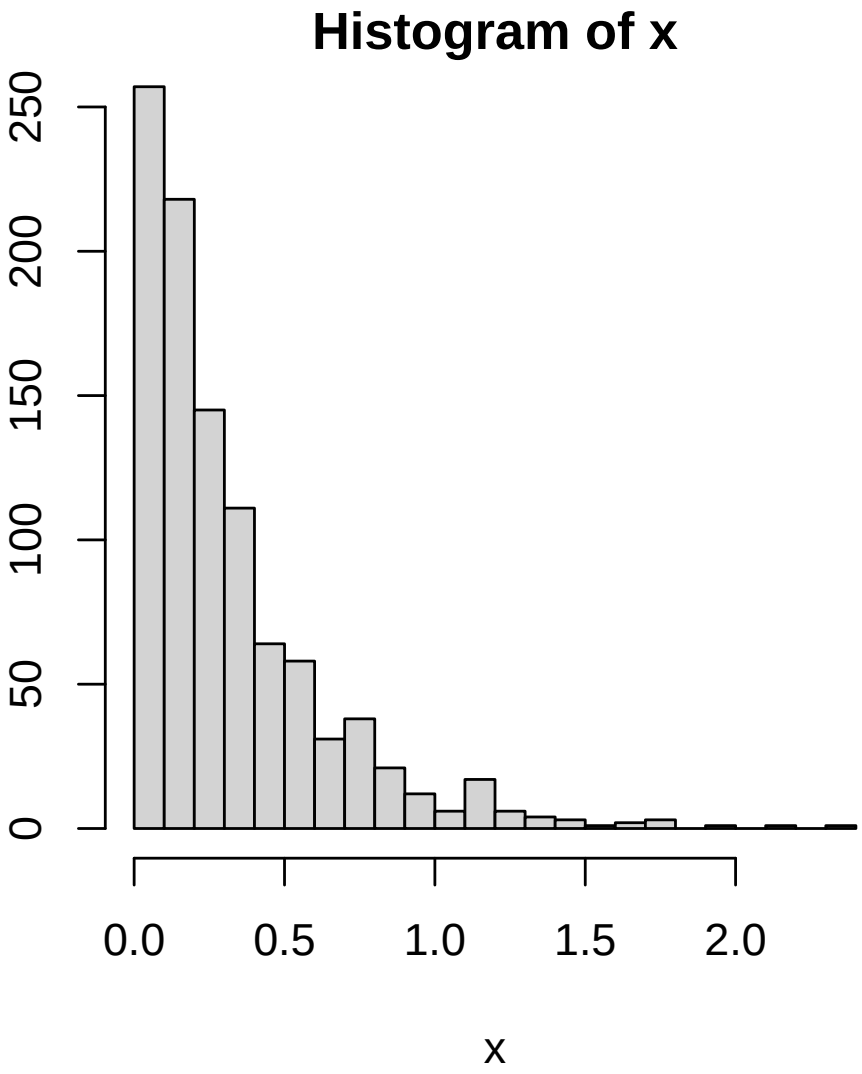
Exempel: normalfördelad?

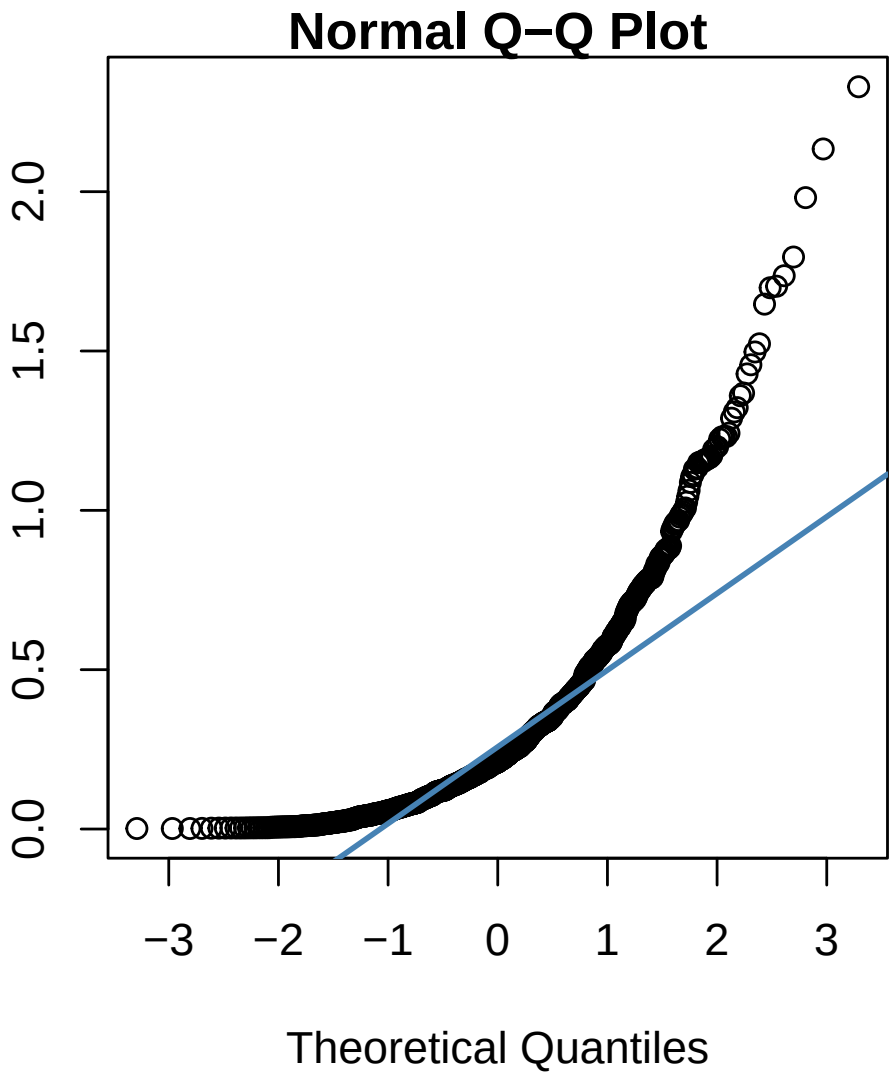
Till exempel, vi har data före och efter en behandling. Vi vill testa om det finns någon skillnad före och efter behandlingen. Vi tänker använda t.test(..., paired = TRUE) men vill kolla om data (efter minus före) är normalfördelade.



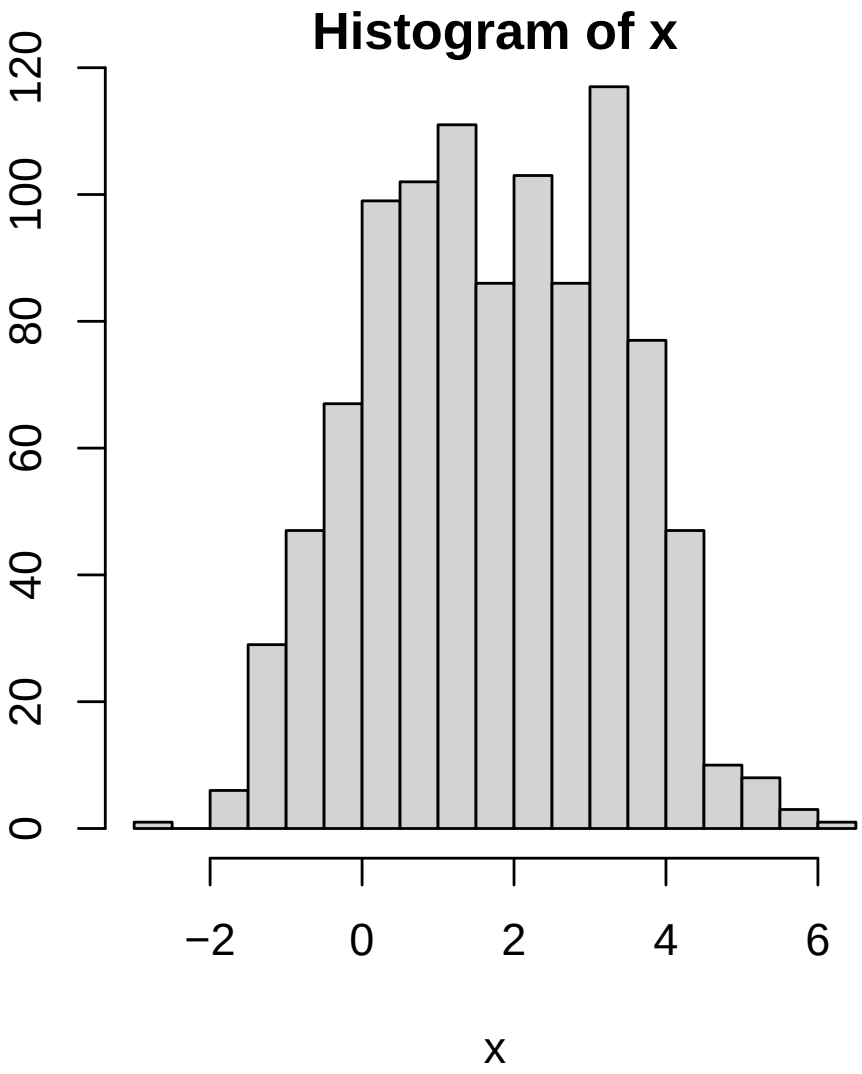


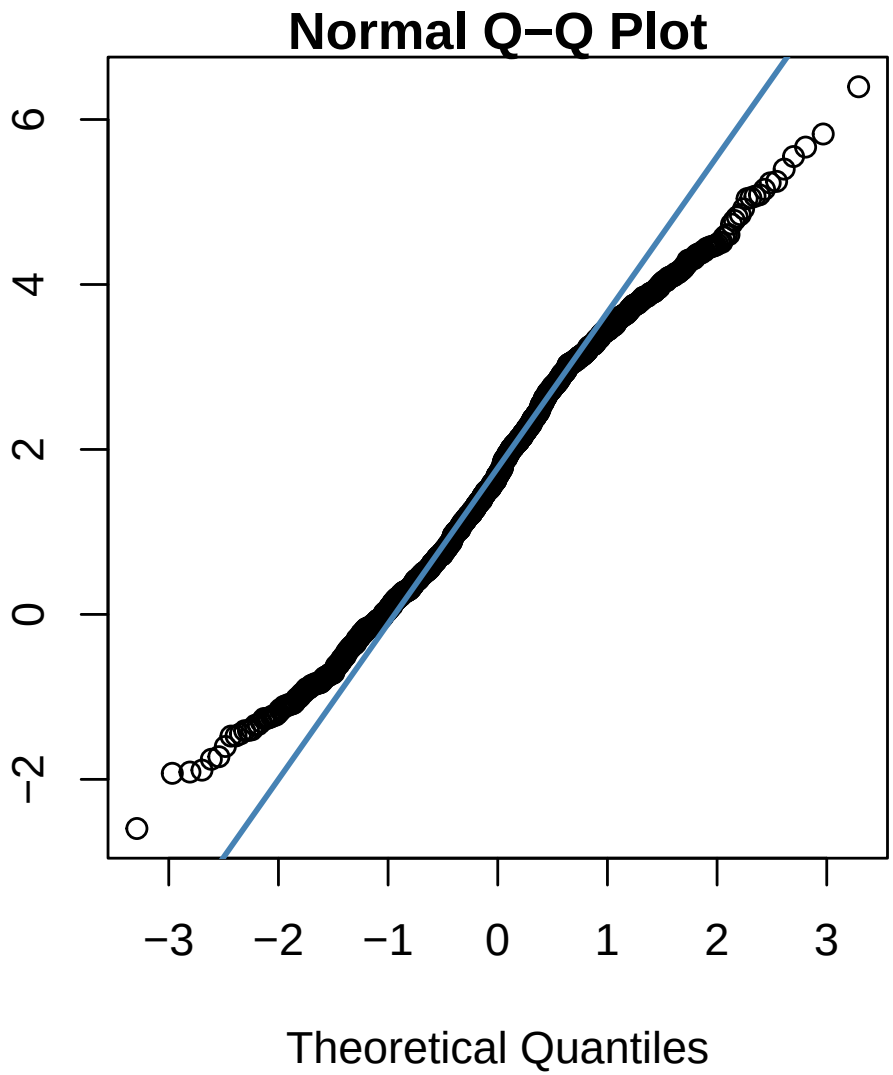
Exempel: normalfördelad?





Exempel: normalfördelad?





Exempel: normalfördelad?

Om du inte är säker på om något är **normalfördelat** eller inte så är det upp till dig att avgöra vad som ser mest rimligt ut.

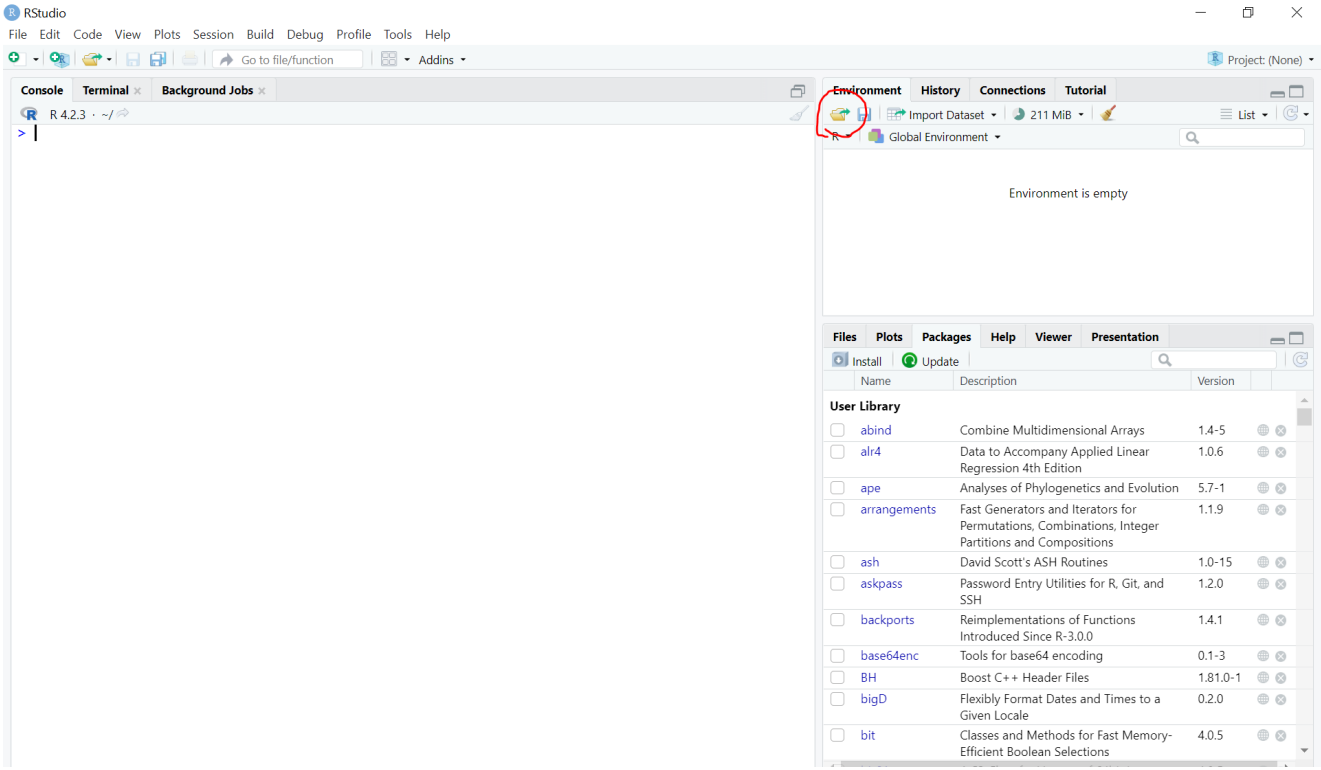
- Men du ska använda metod som passar.
- Det går inte med, t.ex. "Differensen är inte normalfördelad så jag använder ett t-test med funktionen t.test(..., paired = TRUE)".

Om du inte är säker på om något är **linjärt** eller inte så är det upp till dig att avgöra vad som ser mest rimligt ut.

Data

Om det finns en datafil till ett problem kommer det att vara en RData-fil. Den kan ofta läsas in genom att dubbelklicka på den. Om det inte går kan du

1. File -> Open File
2. eller



Vid Problem

```
Problem1 <- data.frame(frekvens = c(71.1,69.3,72.1,69.3,
66.7,69.3,67.6,69.2,69.6,70.9,70.6,69.8,71.7,73.4,71.2,
70.4,70.6,72.6,72.3,72.2,72.9,71.4,70.2,68.9,71.4,71.8,
71.8,73.4,72.5,71.2), underart = rep(c("calolaemus",
"homogenes", "pectroalis"), each = 10))
```

Exempelproblem 1

Den purpurstrupiga bergsjuvelen (Lampornis calolaemus) är en kolibriart som lever i Centralamerika. Man mätte vingfrekvensen hos två underarter:

1. calolaemus som lever i bergsområden i Costa Rica
2. homogenes som lever i lägre delar av landet.

Data finns i Problem1.RData

1. Beräkna ett 95% konfidensintervall för den genomsnittliga vingfrekvensen hos calolaemus.
2. Skiljer sig den genomsnittliga vingfrekvensen åt mellan de två underarterna?

Lösningar

1. Jag beräknar ett 95% konfidensintervall (genom t-test/t-fördelning). Konfidensintervallet för den genomsnittliga vingfrekvensen blir (68.4, 70.7) Hz.
2. Det är svårt att bedöma om något är normalfördelat när man bara har 10 mätpunkter. Jag ritade en QQ-plot och såg inga stora avvikelser från normalfördelning, och bedömer därför att jag kan använda ett t-test. T-testet används för att testa H0: ingen skillnad i medelvärden mellan grupperna mot H1: det finns en skillnad i medelvärden mellan grupperna. Jag fick $p = 0.006$. Skillnaden är signifikant vid 5% signifikansnivå. Det finns statistiska belägg för att det finns en skillnad mellan de två underarternas vingslagsfrekvens.

Varianter på exempelproblem 1

Om det är tydligt att data inte är normalfördelade använd du Wilcoxon-Mann-Whitney test, eller bootstrap.

```
## Wilcoxon-Mann-Whitney
wilcox.test(calolaemus, homogenes)
## Bootstrap
library(MKinfer)
boot.t.test(calolaemus, homogenes)
```

Stickprov i par (använd t-test för stickprov i par)

```
t.test(calolaemus, homogenes, paired = TRUE)
```

Lösning

Det är svårt att bedöma om något är normalfördelat när man bara har 10 mätpunkter. Jag använder därför ett ickeparametriskt test för att vara på den säkra sidan.

Jag använder ett Wilcoxon-Mann-Whitney-test för att testa H0: ingen skillnad i fördelning i vingfrekvens mellan underarterna mot H1: det finns en skillnad i vingfrekvensfördelningen mellan underarterna.

Jag fick p = 0.007. Skillnaden är signifikant vid 5% signifikansnivå. Det finns statistiska belägg för att det finns en skillnad mellan de två underarternas vingslagsfrekvens.

Exempelproblem 2

En forskare vill ta reda på om det är bättre att ge ett vaccin i armen eller i låret. Därför samlades data över en sorts biverkning in. Data finns i Problem2.RData.

Beror förekomsten av den här sortens biverkning på var patienten får vaccinet?

Lösningar

Jag använder ett chi2-test för att testa H0: ingen skillnad mellan behandlingarna mot H1: det finns en skillnad mellan behandlingarna.

Jag får p = 0.18. Skillnaden är inte signifikant vid 5% signifikansnivå. Vi har inga belägg för att det finns någon skillnad i förekomst av biverkningar mellan behandlingarna.

Varianter på exempelproblem 2

Om villkoren för χ^2-test är inte uppfyllda

- finns flera celler där den förväntade frekvens är mindre än 5

använd Fishers exakta test.

```
fisher.test(Problem2)
```

Exempelproblem 3

År 2001 tog företaget Agri-Tech i Virginia fram ett datorsystem som mätte storleken på ostron med en 3D-scanner (enhet: pixlar). Utifrån det måttet ville de sedan kunna uppskatta ostronets vikt (enhet: gram). Data från försöket finns i filen Problem3.RData.

1. Anpassa en lämplig regressionsmodell till data och ange hur den anpassade modellen ser ut.
2. Ett nytt ostron mättes med scannern till 5525945 pixlar. Använd din modell för att ge en uppskattning av ostronets vikt.
3. Kontrollera modellförutsättningar.

Lösningar

1. Jag plottar sambandet och bedömer att det ser väldigt linjärt ut. Jag anpassar en linjär regresisonsmodell med vikt som responsvariabel och storlek i pixlar som förklarande variabel. Modellen får R2 = 0.95. Den anpassade modellen blir $y = -0.289413340803 + 0.000002586817x$.
2. Jag använder modellen för att göra en prediktion. Den predikterade vikten är 14.0 g.
3. Förutsättning 1: linjärt samband. Jag plottar sambandet och bedömer att det ser väldigt linjärt ut. Förutsättning 2: Lika varians. Jag utvärderar antagandet genom att undersöka om residulaerna beror på anpassat y. Jag tycker att antagandet är rimligt eftersom det inte finns någon tydlig tendens. Förutsättning 3: normalfördelning. Jag ritar en QQ-plot för residualerna och se inga stora avvikelser från normalfördelning.

Varianter på exempelproblem 3

- Frågor om p-värden
- Tolkning av modell ("hur mycket förändras vikten om storleken i pixlar ökar med 10000?")
- Prediktionsintervall för en ny observation

```
predict(LR, newdata = data.frame(Storlekpixlar = 5525945),
        interval = "prediction")
```

- Kondensintervall för $\alpha + \beta x$ (det förväntade värdet)

```
predict(LR, newdata = data.frame(Storlekpixlar = 5525945),
        interval = "confidence")
```

- Krävs en transformation av y-variabeln för att den linjära modellen ska bli rimlig

```
LR <- lm(log(Vikt) ~ Storlekpixlar, data = Problem3)
```

Exempelproblem 4

Den purpurstrupiga bergsjuvelen (Lampornis calolaemus) är en kolibriart som lever i Centralamerika. Man mätte vingfrekvensen hos två underarter:

1. calolaemus som lever i bergsområden i Costa Rica
2. homogenes som lever i lägre delar av landet.

Data finns i Problem1.RData.

1. Har alla underarterna samma medelvingfrekvens? Om inte, hur skiljer de sig åt? Kontrollera också förutsättningar.

Lösningar

Jag använder ANOVA för att testa H0: alla underarter har samma genomsnittliga vingfrekvens mot H1: minst en underart har en avvikande vingfrekvens. Vi får p=0.003. Vi förkastar nollhypotesen vid 5% signifikansnivå. Vi har statistiska belägg för att minst en av underarterna avviker.

För att se vilken art som avviker använder jag Tukeys HSD för multipla parvisa jämförelser. Resultat: skillnaden mellan homogenes och calolaemus är 1.97 Hz (p=0.009, signifikant), mellan pectoralis och calolaems är 2.04 Hz (p=0.007, signifikant) och mellan pectoralis och homogenes är 0.07 (p=0.99, ej signifikant). Vi ser att calolaemus har lägre vingfrekvens än de andra två underarterna.

Modelldiagnostikplottar visar att antagandena om normalfördelning och lika varians verkar vara uppfyllda.

Varianter på exempelproblem 4

1. Två faktorer i modellen, undersök samspelseffekt
2. Om vi inte har normalfördelade data kan vi använda bootstrap

```
library(lmboot)
ANOVA <- ANOVA.boot(frekvens ~ underart, data = Problem1)
ANOVA[["p-values"]]
```

Exempelproblem 5

År 2001 tog företaget Agri-Tech i Virginia fram ett datorsystem som mätte storleken på ostron med en 3D-scanner (enhet: pixlar). Utifrån det måttet ville de sedan kunna uppskatta ostronets vikt (enhet: gram). Data från försöket finns i filen Problem3.RData.

1. Anpassa två olika regressionsmodeller där vikt är responsvariabel. En som använder mätvärdet från 3D-scannern som förklarande variabel och en som använder mätvärdet från 2D-scannern som förklarande variabel. Vilken av modellera förklarar variationen i vikt bäst?
2. Ett nytt ostron mättes med scannern till 5525945 pixlar. Använd den modell du tycker verkar bäst för att ge ett prediktionsintervall för ostronets vikt.

Lösningar

Jag anpassar först en linjär regressionsmodell med vikt som responsvariabel och storlek i pixlar enligt 3D-scannern som förklarande variabel. R2 blir 0.95. Sambandet ser linjärt ut och modelldiagnostikplottarna ser bra ut.

Jag anpassar sedan en en linjär regressionsmodell med vikt som responsvariabel och storlek i pixlar enligt 2D-scannern som förklarande variabel. R2 blir 0.845. Sambandet ser linjärt ut och modelldiagnostikplottarna ser ganska bra ut.

Eftersom den första modellen har högre R2 (förklarar mer av variationen i vikt) och eftersom punkterna ligger närmare linjen för den första modellen så föredrar jag den.

Jag använder modellen för att beräkna ett 95 % konfidensintervall och får resultatet 12.6-15.5 g.