(A)

Denne opgave bygger på resultaterne fra 2 forsøg med epo-behandling af for tidligt fødte børn, idet gruppe 1 og 3 stammer fra første forsøg, mens gruppe 2 og 4 stammer fra det andet. Det må antages, at behandlingen af børnene i første forsøg må være udført under sammenlignelige forhold, og tilsvarende om forsøg 2, men det vides ikke, hvorvidt forholdene under forsøg 1 og 2 kan sammenlignes. Fra forsøg 1 fremkommer observationen $x^1 = (x_{rs})_{r=1,3,s=1,\ldots,n_r}$, hvor $n_1 = 15$ og $n_3 = 16$, idet $x_1 = (x_{1r})_{r=1,\ldots,15}$ betegner observationen fra epo-gruppe 1, og $x_3 = (x_{3r})_{r=1,\ldots,16}$ betegner observationen fra epo-gruppe 3. Der antages, at denne observation er en realisation af den stokatiske vektor $X^1 = (X_{rs})_{r=1,3,s=1,\ldots,n_r}$. Da ændringen i hæmoglobinprocent i de børn, der deltog i forsøget, ikke kan påvirke hinanden, og forsøgene må antages at være foregået særskilt, kan det antages, at X_{rs} for $r = 1, 3$ og $s = 1, \ldots, n_r$ er uafhængige. Da x_{1s} for $s = 1, \ldots, 15$ således repræsenterer uafhængige gentagelser af samme forsøg under sammenlignelige forsøg, må det antages, at X_{1s} for $s = 1, \ldots, 15$ har samme fordeling. Analogt antages det, at X_{3s} for $s = 1, \ldots, 16$ er uafhængige identisk fordelte stokatiske variable.

Fra forsøg 2 fremkom observationen $x^2 = (x_{rs})_{r=2,4,s=1,\ldots,n_r}$, hvor $n_2 = 15$ og $n_4 = 14$, idet $x_2 = (x_{2r})_{r=1,\ldots,15}$ betegner observationen fra epo-gruppe 2, og $x_4 = (x_{4r})_{r=1,\ldots,14}$ betegner observationen fra epo-gruppe 4. Det antages, at denne observation er en realisation af den stokatiske vektor $X^2 = (X_{rs})_{r=2,4,s=1,\ldots,n_r}$. Ved analoge argumenter antages det, at X_{rs} for $r = 2, 4$ og $s = 1, \ldots n_r$ er indbyrdes uafhængige, og at X_{2s} for $s = 1, \ldots, 15$, samt X_{4s} for $s = 1, \ldots, 14$ er identisk fordelte. Da de to forsøg er udført adskilt fra hinanden, kan det antages X_{rs} for $r = 1, 2, 3, 4$ og $s = 1, \ldots n_r$ alle er indbyrdes uafhængige.

Det skal nu undersøges, om X_{rs} for $r = 1, 2, 3, 4$ og $s = 1, \ldots, n_r$ kan antages at være normalfordelte. Først beregnes den empiriske middelværdi

$$\bar{x}_r = \frac{1}{n_r} \sum_{s=1}^{n_r} x_{rs}$$

og den empiriske varians

$$s_r^2 = \frac{1}{n_r - 1} \sum_{s=1}^{n_r} (x_{rs} - \bar{x}_r)^2,$$

og der tegnes et histogram for datasættet x_r sammen med tætheden for normalfordelingen med middelværdi $\bar{x}_r$ og varians s_r^2. På figur 1 betragtes test-1 datahistogrammet mod den estimerede normalfordeling, og det bemærkes, at histogrammet kun svagt har den ønskede klokkeform, idet der især mangler observationer tæt ved den emperiske middelværdi. På figur 3 er det histogrammet fra epogruppe 2, og der er en tydelig afvigelse fra normalfordelingskurven, idet der fremkommer mange observationer lavere end den empiriske middelværdi, men histogrammet har delvist den ønskede form. På figur 5 er det epogruppe 3, og der ses en vis overensstemmelse mellem histogrammet og normalfordelingens tæthed, hvilket også gør sig gældende for figur 7 med epogruppe 4.

Generelt er overensstemmelsen mellem histogrammerne og de givne tætheder for normalfordelingen ikke overbevisende, men dette kan skyldes, at de benyttede data er behæftet med en vis usikkerhed samt det meget lave antal observationer, hvorved man ikke bør forvente den store sammenhæng mellem normalfordelingens tætheder og histogrammet. Man kan således pga det lave antal observation ikke på baggrund af figur 1,3,5 og 7 bekræfte normalfordelingshypotesen. Man kan dog heller ikke

afkræfte denne, da selv et histogram bygget på få observationer, som vides at være normalfordelte, kan adskille sig væsentligt fra tætheden, da normalfordelingen altid har en vis varians.

Af denne grund tegnes QQ-plot for det observerede data mod den estimede normalfordeling, idet punkterne $(x_r(s), F^{-1}(\frac{n_r}{s}))$ for $s = 1, \ldots, n_r$, hvor F betegner fordelingsfunktionen for den estimerede normalfordelingen, plottes. Hvis observationer kan antages at stamme fra den estimerede normalfordeling, skal punkterne ligge tæt ved diagonalen, som ligledes er indtegnet. Betragter man således figur 2,4,6 og 8, der er QQ-plot for de 4 grupper, er det overordnede indtryk, at der er god overensstemmelse mellem punkterne og diagonalen, især omkring 0, der repræsenterer observationer omkring den emperiske middelværdi. Der er dog enkelte værdi, som ligger langt fra den empiriske middelværdi, som medfører at QQ-plottet har enkelte punkter, der ligger langt fra diagonalen. Dette skyldes, at normalfordelingen er koncentreret omkring middelværdien, og der er således lille sandsynlighed for at få sådanne observationer, man da sandsynligheden ikke er nul, er enkelte observationer af denne type ikke nok til at forkaste hypotesen. På baggrund af den store tilnærmelse i QQ-plottene godkendes normalfordelingshypotesen, og i resten af opgaven antages det, at X_{rs} for $r = 1, 2, 3, 4$ og $s = 1, \ldots, n_r$ er normalfordelte med middel μ_r og varians σ_r^2, hvor μ_r kan estimeres af $\bar{x}_r$ og σ_r^2 af s_r^2.

Vi ønsker nu at undersøge hvorvidt de to forsøg er sammenlignelige. Da gruppe 3 og 4 har fået samme behandling, og de stammer fra hver sit forsøg, kan dette undersøges på baggrund af deres data. Betragt observationen $x = (x_{rs})_{r=3,4,s=1,\ldots,n_r}$, hvor $n_3 = 16$ og $n_4 = 14$, idet $x_3 = (x_{2r})_{r=1,\ldots,16}$ betegner observationen fra epo-gruppe 3, og $x_4 = (x_{4r})_{r=1,\ldots,14}$ betegner observationen fra epo-gruppe 4. Der antages, at denne observation er en realisation af den stokatiske vektor $X = (X_{rs})_{r=3,4,s=1,\ldots,n_r}$, hvor X_{rs} for $r = 3, 4$ og $s = 1, \ldots n_r$ er indbyrdes uafhængige og $X_{rs} \sim N(\mu_r, \sigma_r^2)$. Ved t-test i SAS beregnes testsandsynligheden for at $\sigma_3^2 = \sigma_4^2$ og da $\varepsilon(x) = 0.4755$, godkendes hypotesen på niveau 0.05, hvorfor der efterfølgende antages, at $\sigma_3^2 = \sigma_4^2 = \sigma^2$. Herved antages da, at $X_{rs} \sim N(\mu_r, \sigma^2)$ for alle $r = 3, 4$ og $s = 1, \ldots, n_r$. Undersøgelse om hvorvidt de to forsøg er sammenlignelige kan således formaliseres til, at undersøgelsen hvorvidt X_{rs} har samme fordeling for alle $r = 3, 4$ og $s = 1, \ldots, n_r$, dvs. der skal testes for om $\mu_3 = \mu_4$. Til dette anvendes sætning 11.4.1.

Da er x en observation fra den statistiske model

$$\left(\mathbb{R}^{30}, (N_{\mu_3,\mu_4,\sigma^2})_{(\mu_3,\mu_4,\sigma^2)\in\mathbb{R}^2\times]0,\infty[}\right),$$

hvor N_{μ_3,μ_4,σ^2} har tæthed

$$\phi_{\mu_3,\mu_4,\sigma^2}(x) = \frac{1}{(\sqrt{2\pi\sigma^2})^{30}} e^{-\frac{1}{2\sigma^2}\sum_{r=3}^{4}\sum_{s=1}^{n_r}(x_{rs}-\mu_r)^2}.$$

Da $\sum_{r=3}^{4}\sum_{s=1}^{n_r}(x_{rs} - \mu_r)^2 > 0$, grundet der i hver forsøgsgruppe er to forskellige observationer, kan sætningen benyttes. Vi betragter hypotesen

$$\mu_3 = \mu_4 = \mu$$

Under modellen estimeres (μ_3, μ_4, σ^2) i henhold til sætning 11.4.1 og bemærkning 11.4.3. ved

$$\hat{\mu}_3 = \bar{x}_3 = \frac{1}{16}\sum_{s=1}^{16} x_{3s}, \quad \hat{\mu}_4 = \bar{x}_4 = \frac{1}{14}\sum_{s=1}^{14} x_{4s}, \quad s^2 = \frac{1}{28}\sum_{r=3}^{4}\sum_{s=1}^{n_r}(x_{rs} - \bar{x}_r)^2$$

hvor $\hat{\mu}_3 \sim N(\mu_3, \frac{\sigma^2}{16})$,$\hat{\mu}_4 \sim N(\mu_4, \frac{\sigma^2}{14})$ og $s^2 \sim \frac{\sigma^2}{28}\chi^2_{28}$.

Under hypotesen estimeres (μ, σ^2) i henhold til sætning 11.4.1 og bemærkning 11.4.3. ved

$$\hat{\mu} = \bar{x} = \frac{1}{30}\sum_{r=3}^{4}\sum_{s=1}^{n_r} x_{rs}, \quad \tilde{s}^2 = \frac{1}{29}\sum_{r=3}^{4}\sum_{s=1}^{n_r}(x_{rs} - \bar{x})^2$$

hvor $\hat{\mu} \sim N(\mu, \frac{\sigma^2}{30})$ og $\tilde{s}^2 \sim \frac{\sigma^2}{29}\chi^2_{29}$.

I henhold til SAS er t-størrelsen 0.41 og testsandsynligheden $\varepsilon(x) = 0.6855$, hvorved testen er ikke signifikant. Hypotesen godkendes således på 0.05 niveau. Forsøgene i gruppe 3 og 4 kan således antages at have samme middelværdi, og idet deres varians blev antaget ens, må X_{rs} for $r = 3, 4$ og $s = 1, \ldots, n_r$ være identisk fordelt. Da de to grupper repræsenterer det samme forsøg gentaget under forskellige forhold, kan disse antages at være sammenlignelige, da forsøgsresultaterne kan antages at stamme fra samme fordeling. Under den afsluttende model kan det antages, at $X_{rs} \sim N(\mu, \sigma^2)$, hvor μ estimes ved $\hat{\mu} = -0,5967$ og σ^2 ved $\tilde{s}^2 = 1,6410$, hvor $\hat{\mu} \sim N(\mu, \frac{\sigma^2}{30})$ og $\tilde{s}^2 \sim \frac{\sigma^2}{29}\chi^2_{29}$.

(B)

Vi ønsker at undersøge, om der kan antages at være en lineær sammenhæng mellem epo-dosis og hæmoglobinprocent ved for tidligt fødte børn. Vi plotter derfor vores observationer mod dosis i et diagram (se Figur 9 i Bilag 3) og får SAS til at beregne og indtegne den bedste regressionslinie og ser, at der tilnærmelsesvist er tale om en lineær sammenhæng mellem dosis og hæmoglobinprocent. Dette bekræftes yderligere ved at plotte de standardiserede residualer mod dosis (se Figur 10 i Bilag 3), hvor vi ser, at residualerne tilnærmelsesvist spreder sig om x-aksen som en standard normalfordeling, jævnfør IH afsnit 12.6.1. Derudover plottes et histogram over residualerne med en indtegnet standard normalfordeling (se Figur 11 i Bilag 3), og det ses, at de to stemmer fint overens. Endvidere laves et QQ-plot af residualernes fraktiler mod standardnormalfordelingens fraktiler (se Figur 12 i Bilag 3), og det ses, at punkterne fordeler sig ganske pænt om en ret linie, hvilket igen antyder en lineær sammenhæng.

Vi antager derfor endelig (!), at der findes en lineær sammenhæng (dvs. at der gælder følgende om middelværdien: $EX_r = \alpha + \beta(t - \bar{t})$ for $r = 1, \ldots, 60$ og $(\alpha, \beta) \in \mathbb{R}^2$) mellem ændringen i hæmoglobinniveauet (for for tidligt fødte børn) og dosen af epo. Videre antager vi, at variansen for ændringen i hæmoglobinprocenterne er den samme ligegyldigt om børnene får en epo-dosis på 0 U/kg, 50 U/kg eller 100 U/kg. Vi bruger så IH sætning 12.3.1 og får modellen

$$(\mathbb{R}^{60}, (N_{\alpha,\beta,\sigma^2})_{\alpha,\beta,\sigma^2 \in \mathbb{R}^2 \times]0,\infty[}),$$

hvor $N_{\alpha,\beta,\sigma^2}$ har tæthed

$$\phi_{\alpha,\beta,\sigma^2}(x) = \frac{1}{(\sqrt{2\pi\sigma^2})^{60}} e^{-\frac{1}{2\sigma^2}\sum_{s=1}^{60}(x_s - \alpha - \beta(t_s - \bar{t}))^2}.$$

Her er maksimaliseringsestimatoren for $(\alpha, \beta, \sigma^2)$ bestemt ved

$$\begin{aligned}
\hat{\alpha} &= \bar{x} \\
\hat{\beta} &= \frac{\sum_{r=1}^{60}(x_r - \bar{x})(t_r - \bar{t})}{\mathrm{SSD}_t} \\
\hat{\sigma}_l^2 &= \frac{1}{60}\sum_{r=1}^{60}(x_r - \bar{x} - \hat{\beta}(t_r - \bar{t}))^2
\end{aligned}$$

hvor $x_1, \ldots, x_{60}$ er ændringerne i hæmoglobinprocenterne og $t_1, \ldots, t_{60}$ er de tilsvarende doser epo. Endvidere gælder der $\hat{\alpha} \sim N(\alpha, \frac{\sigma^2}{60})$, $\hat{\beta} \sim N(\beta, \frac{\sigma^2}{\mathrm{SSD}_t})$ og $\hat{\sigma}_l^2 \sim \frac{\sigma^2}{60}\chi^2_{58}$. Vi vil dog (jvf. IH bemærkning 12.3.3) bruge $s_l^2 = \frac{1}{58}\sum_{r=1}^{60}(x_r - \bar{x} - \hat{\beta}(t_r - \bar{t}))^2$ som estimator for spredningen, og da $s_l^2 = \frac{60}{58}\hat{\sigma}_l^2$ gælder der $s_l^2 \sim \frac{\sigma^2}{58}\chi^2_{58}$.

Af SAS-udskriften (Lineær regressionsdelen) ses det at

$$\begin{aligned} \bar{x} &= -1,12667 \\ \hat{\beta} &= 0,01537 \\ 58s_l^2 &= 160,79636 \\ \hat{\beta}^2\mathrm{SSD}_t &= 24,36097 \end{aligned}$$

og dermed gælder der

$$\begin{aligned} s_l^2 &= \frac{160,79636}{58} = 2,77235 \\ \mathrm{SSD}_t &= \frac{24,36097}{0,01537^2} = 103120,93496. \end{aligned}$$

Så vi estimerer α til $-1,12667$ β til $0,01537$ og σ^2 til $2,77235$ og $\hat{\beta}$ er normalfordelt med middelværdi β og varians $\frac{\sigma^2}{103120,93496}$.

(c)

Vi anvender nu modellen fra opgave (b) og ønsker at undersøge hypotesen

$$H_\beta : \ EX_r = \alpha + \beta_0(t_r - \bar{t}),$$

hvor $\beta_0 = 0$. Forkastes denne hypotese, vil det betyde, at vi ikke kan se bort fra vores baggrundsvariabel (dosis), eller med andre ord, der er en ikke-triviel lineær sammenhæng mellem dosis og hæmoglobinprocent.

Under hypotesen bliver maksimaliseringsestimatorerne for (α, σ^2) (jævnfør IH 12.5.2):

$$\begin{aligned} \hat{\alpha} &= \bar{x} \\ \hat{\sigma}_\beta^2 &= \frac{1}{60}\sum_{r=1}^{60}(x_r - \bar{x})^2, \end{aligned}$$

hvor $\hat{\alpha} \sim N\left(\alpha, \frac{1}{60}\sigma^2\right)$ og $\hat{\sigma}_\beta^2 \sim \frac{\sigma^2}{60}\chi^2_{59}$.

Fra SAS får vi vores t-værdi til 2.96, hvilket giver en testsandsynlighed på 0.0044. Derfor forkaster vi hypotesen om, at $\beta_0 = 0$ og kan konkludere, at der altså *er* en ikke-triviel lineær sammenhæng mellem dosis og hæmoglobinprocent. Hvilket så vil sige, at epo givet til for tidligt fødte børn *har* en indflydelse på deres hæmoglobinniveau. Den ikke-trivielle statistiske slutmodel bliver så modellen fra opgave (b).

Bilag 1: SAS-kode

Her er alle observationer selvfølgelig indtastet i de anvendte datasæt, som af pladsgrunde dog er udeladt.

```
/* Normalfordelinghypoteser for alle fire forsøg: */

PROC SORT DATA=epoalle;
BY test;
PROC UNIVARIATE DATA=epoalle NOPRINT;
  VAR hgp;
    HISTOGRAM/NORMAL(MU=est SIGMA=est) MIDPOINTS=-4 to 3 by 1;
    QQPLOT/NORMAL(MU=est SIGMA=est);
  BY test;
RUN;
QUIT;

/* Opgave (a): T-test for sammenligneligheden af de to forsøg: */

PROC TTEST DATA=epo;
  CLASS TEST;
  VAR RESULT;
RUN;

/* Opgave (b): */

/* Plot af observationer med indtegnet regressionslinie
som udgangspunkt for regressionsanalyse og estimater:*/

SYMBOL1 v=none i=rl c=red;
SYMBOL2 v=none I=STDPMT c=blue;
PROC GPLOT DATA=opgb;
  PLOT hgp*dosis=1 hgp*dosis=2/OVERLAY;
RUN;
QUIT;

/* Plot af residualer, for at illustrere normalfordelte residualer:*/

SYMBOL1 v=star i=rl c=red;
SYMBOL2 v=none I=STDPMT c=blue;
PROC REG DATA=opgb;
  MODEL hgp=dosis/R CLM;
  PLOT STUDENT.*(P. dosis);
  OUTPUT OUT=res STUDENT=residual;
RUN;
QUIT;

/* Plots af residualhistogram mod normalfordeling
og residualernes fraktiler mod normalfordelingens fraktiler: */

PROC UNIVARIATE NOPRINT;
  VAR residual;
  HISTOGRAM residual/NORMAL MIDPOINTS=-3 to 3 BY 0.5;
  QQPLOT residual/NORMAL(MU=est SIGMA=est);
RUN;
QUIT;

/* Opgave (c): Test for hypotesen beta=0: */

PROC REG DATA=opgb;
  MODEL hgp=dosis/R CLM;
  TEST dosis=0/PRINT;
RUN;
QUIT;
```

Bilag 2: SAS-udskrifter

T-test for sammenhæng mellem de to grupper (3 og 4):

The TTEST Procedure

T-Tests

Variable	Method	Variances	DF	t Value	Pr > |t|
RESULT	Pooled	Equal	28	-0.41	0.6855
RESULT	Satterthwaite	Unequal	27.9	-0.41	0.6815

Equality of Variances

Variable	Method	Num DF	Den DF	F Value	Pr > F
RESULT	Folded F	15	13	1.49	0.4755

Regressionstest for hypotesen $\beta = 0$:

The REG Procedure
Model: MODEL1
Dependent Variable: hgp

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	24.36097	24.36097	8.79	0.0044
Error	58	160.79636	2.77235		
Corrected Total	59	185.15733			

Root MSE	1.66504	R-Square	0.1316
Dependent Mean	-1.12667	Adj R-Sq	0.1166
Coeff Var	-147.78443		

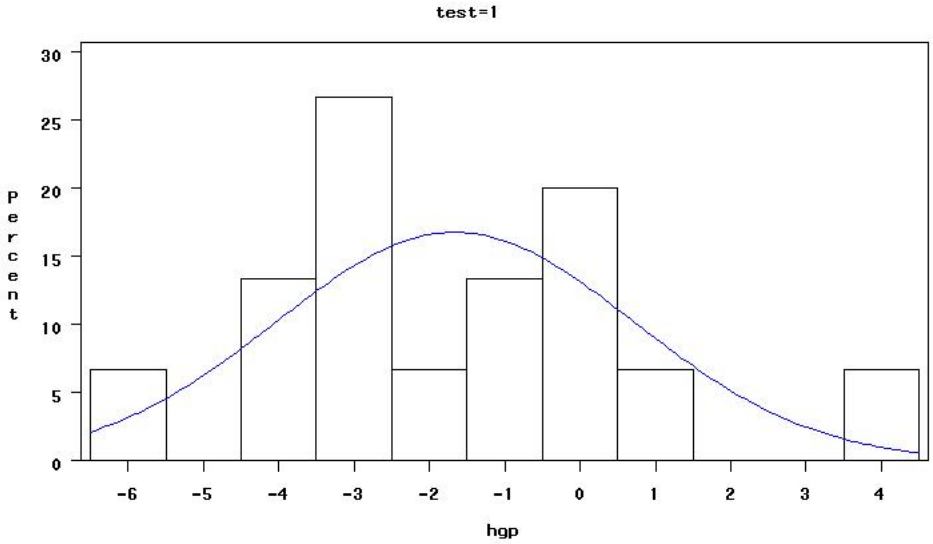
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-2.08727	0.38887	-5.37	<.0001
dosis	1	0.01537	0.00518	2.96	0.0044

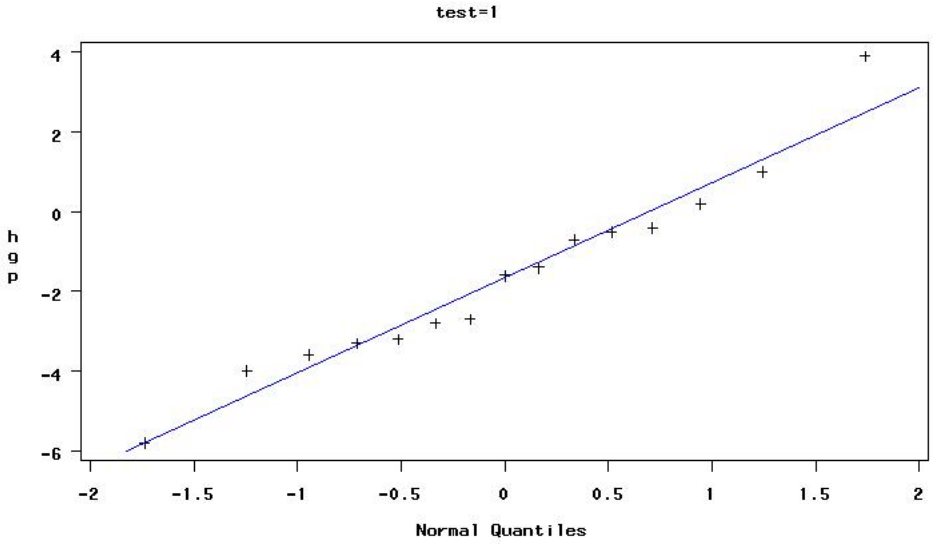
Test 1 Results for Dependent Variable hgp

Source	DF	Mean Square	F Value	Pr > F
Numerator	1	24.36097	8.79	0.0044
Denominator	58	2.77235		

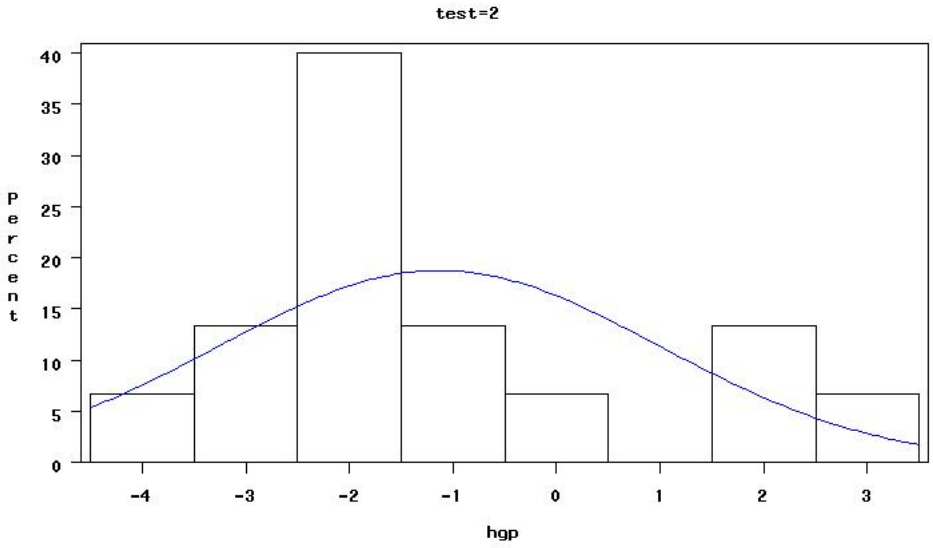
Bilag 3: SAS-plots



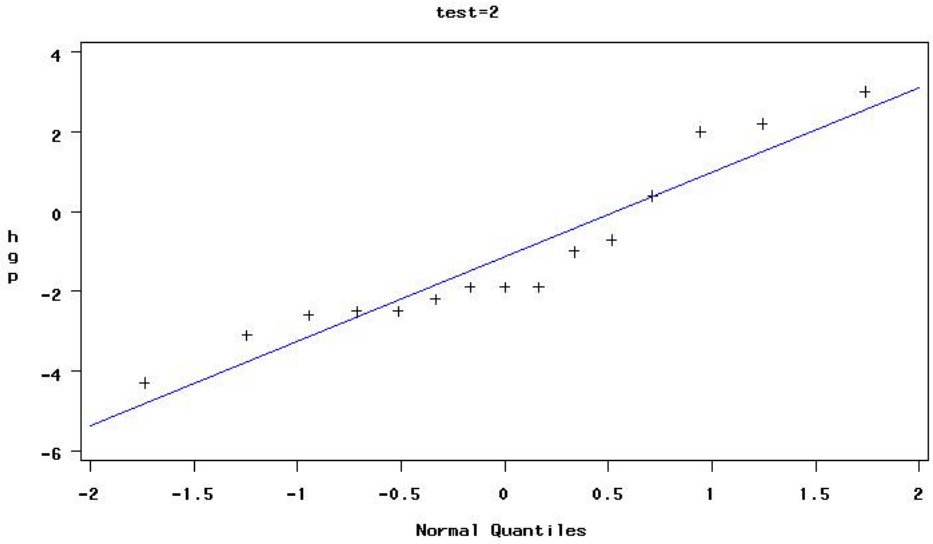
Figur 1. Plot af Test 1-datahistogram mod normalfordeling



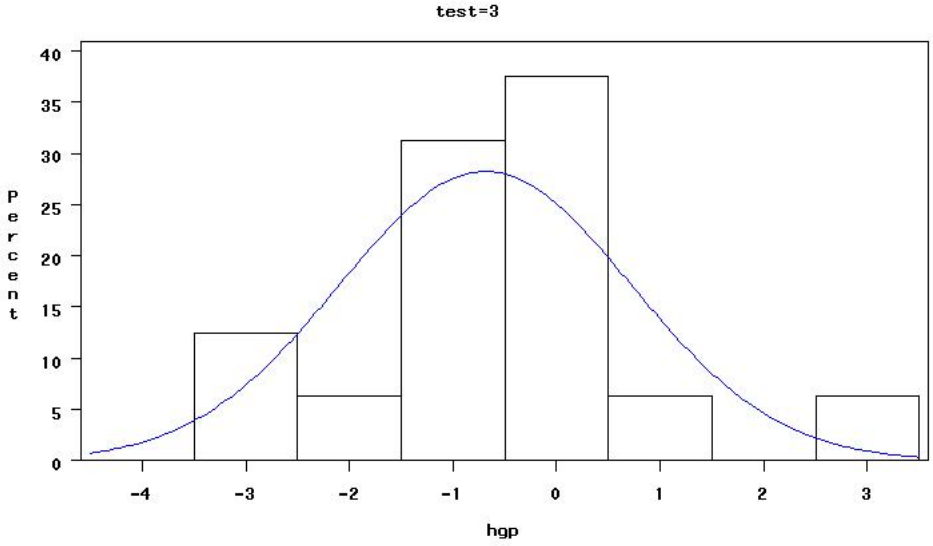
Figur 2. QQ-plot af Test 1



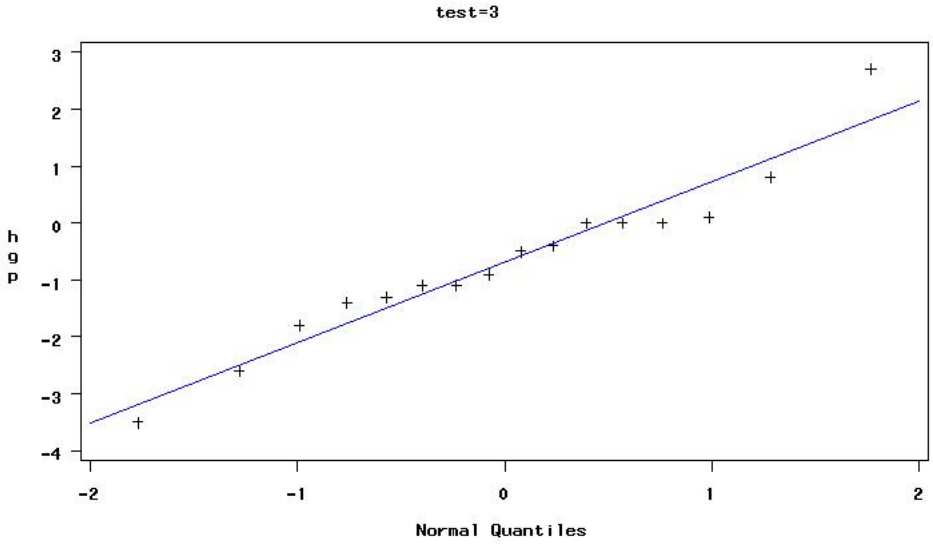
FIGUR 3. Plot af Test 2-datahistogram mod normalfordeling



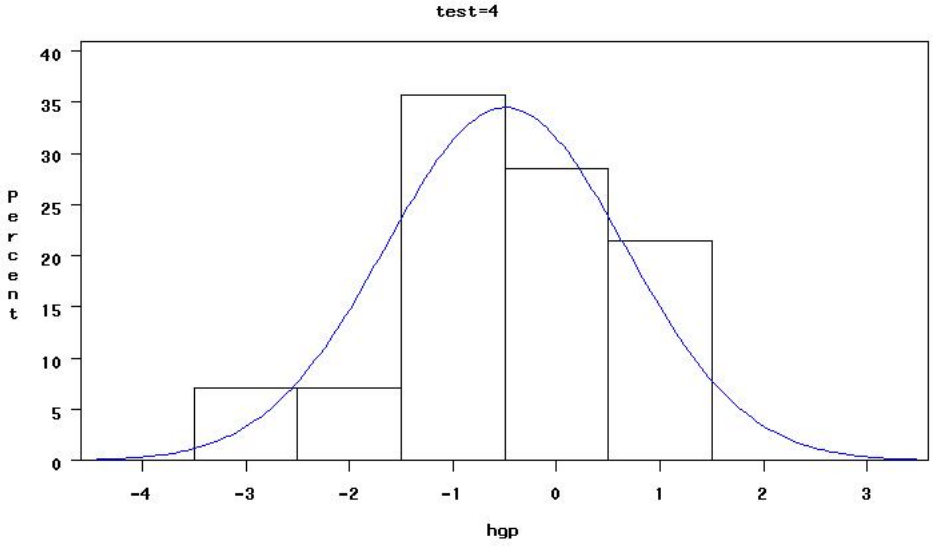
FIGUR 4. QQ-plot af Test 2



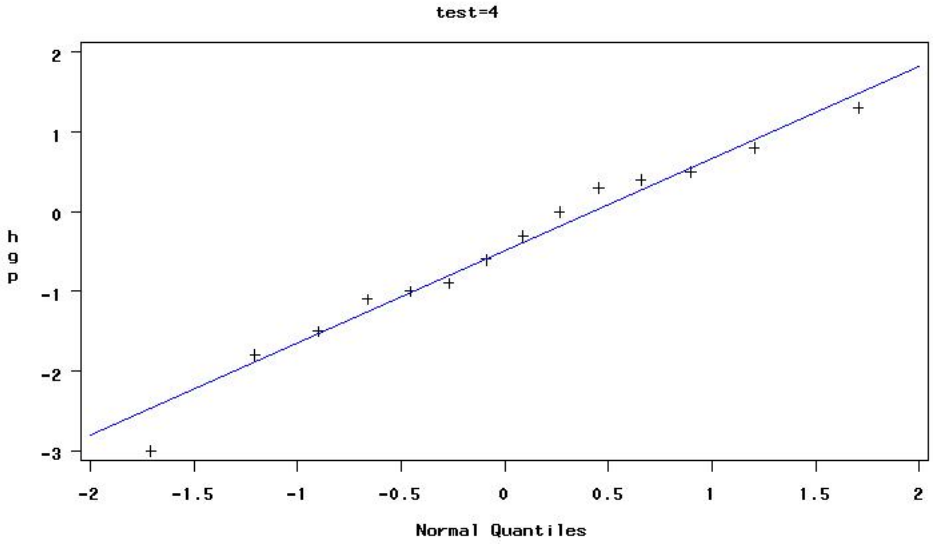
FIGUR 5. Plot af Test 3-datahistogram mod normalfordeling



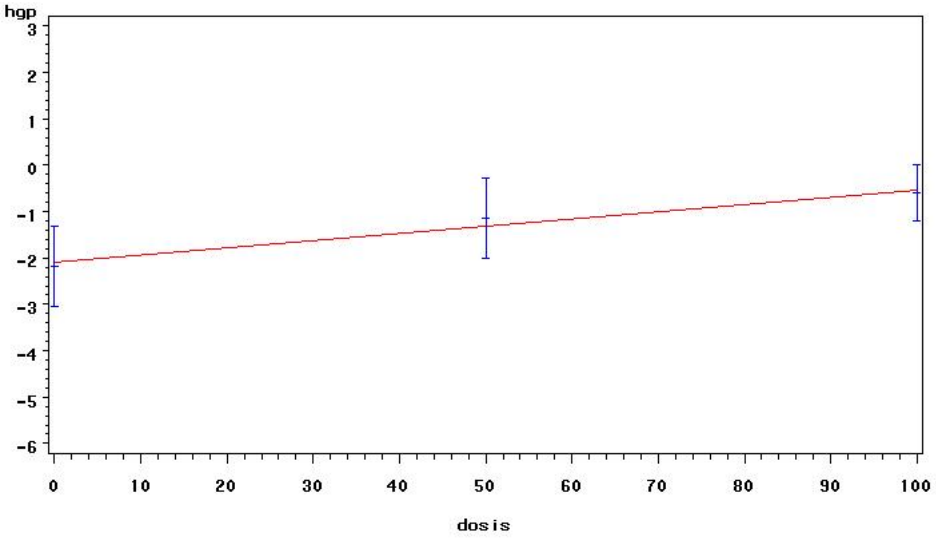
FIGUR 6. QQ-plot af Test 3



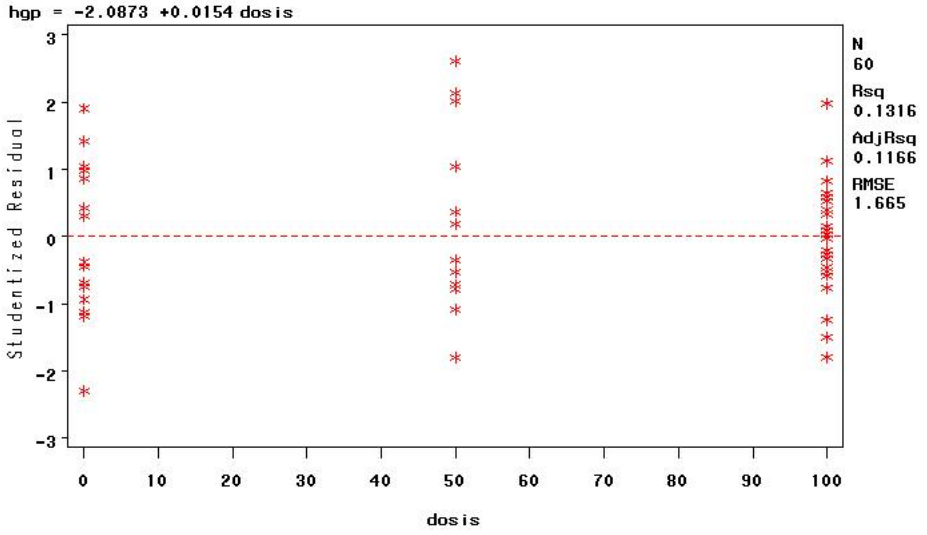
FIGUR 7. Plot af Test 4-datahistogram mod normalfordeling



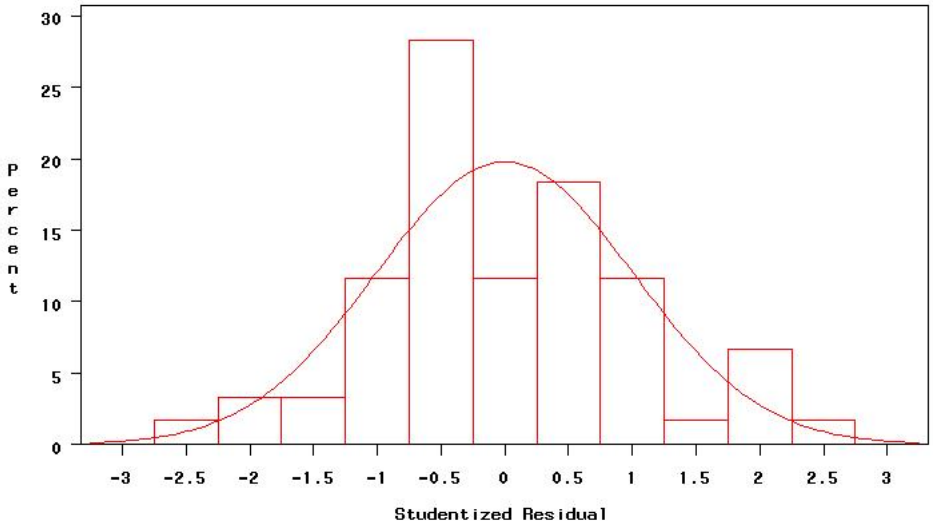
FIGUR 8. QQ-plot af Test 4



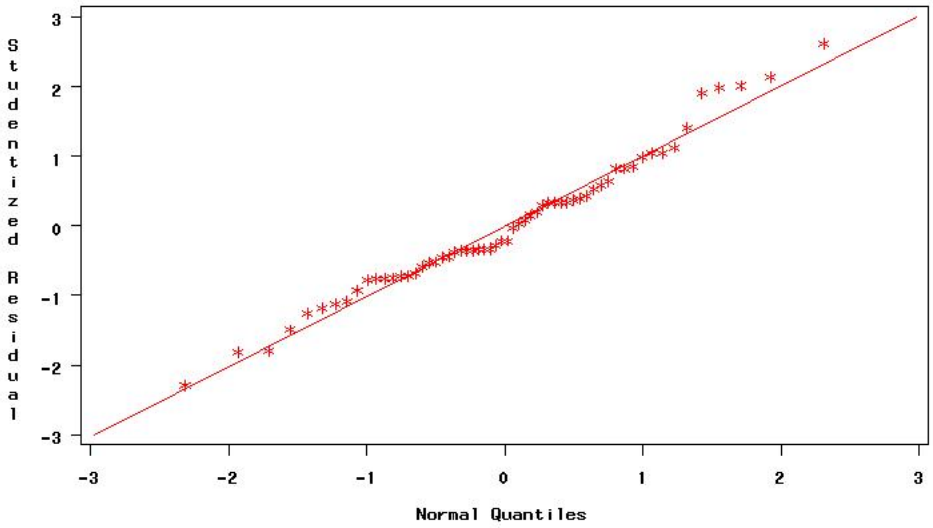
FIGUR 9. Plot af hæmoglubinændring mod epo-dosering



FIGUR 10. Plot af Studentized Residuals mod Dosis



FIGUR 11. Histogram af residualer med indtegnet stand. normalfordeling



FIGUR 12. QQ-plot af residualers fraktiler mod normalford. fraktiler.