Københavns Universitet: Sandsynlighedsregning & Statistik - Obligatorisk opg. 2

Hold 8
Daniel Friis-Hasché - rcb933
Victor Vangkilde Jørgensen - kft410

December 19, 2024

Contents

1 Indlæs datasættet som beskrevet ovenfor. Lav en ny variabel, *LogPay*, i datasættet. Den nye variabel skal indeholde den naturlige logaritme til værdierne i *Pay*. Bestem derefter median, gennemsnit, stikprøvevarians og stikprøvespredning for variablene *Pay* og *LogPay*, dvs. udfyld følgende skema:

	Median	Gennemsnit	Stikprøvevarians	Stikprøvespredning
Pay	**	**	**	**
LogPay	**	**	**	**

Table 1: Tabel stillet i opgaven.

Vi bruger følgende kode i R til at importere data'en og beregne værdierne til tabellen:

```
data <- read.table("paydata2017.txt", header=TRUE)
data <- transfor m(data, logPay=log(Pay))

payMedian <- median(data$Pay)
payAverage <- mean(data$Pay)
paySampleVariance <- var(data$Pay)
paySampleDeviation <- sd(data$Pay)
logPayMedian <- median(data$logPay)
logPayAverage <- mean(data$logPay)
logPaySampleVariance <- var(data$logPay)
logPaySampleDeviation <- sd(data$logPay)
```

Vi får så vores værdier vi kan indsætte i tabellen givet i opgaven:

	Median	Gennemsnit	Stikprøvevarians	Stikprøvespredning
Pay	81459,685	85577,221	1076703934,22	32813,167
LogPay	11,308	11,293	0,125	0,354

Table 2: Tabel med data fra R

2 Tegn et histogram for Pay på "sandsynlighedsskala", dvs. således at det samlede areal under histogrammet er 1. Tegn tætheden for normalfordelingen med middelværdi og varians lig gennemsnit og stikprøvevarians for *Pay* i samme figur. Lav den tilsvarende figur for *LogPay*. Diskutér om det er mest fornuftigt at antage at lønnen er normalfordelt eller at logaritmen til lønnen er normalfordelt. Argumentér ud fra figurerne og (nogle af) tallene fra tabellen ovenfor.

Vi bruger kommandoen

```
hist_data <- hist(data$Pay, prob=TRUE), breaks=50}
```

til at tegne histogrammet med et samlet areal på 1. Her får vi så følgende:

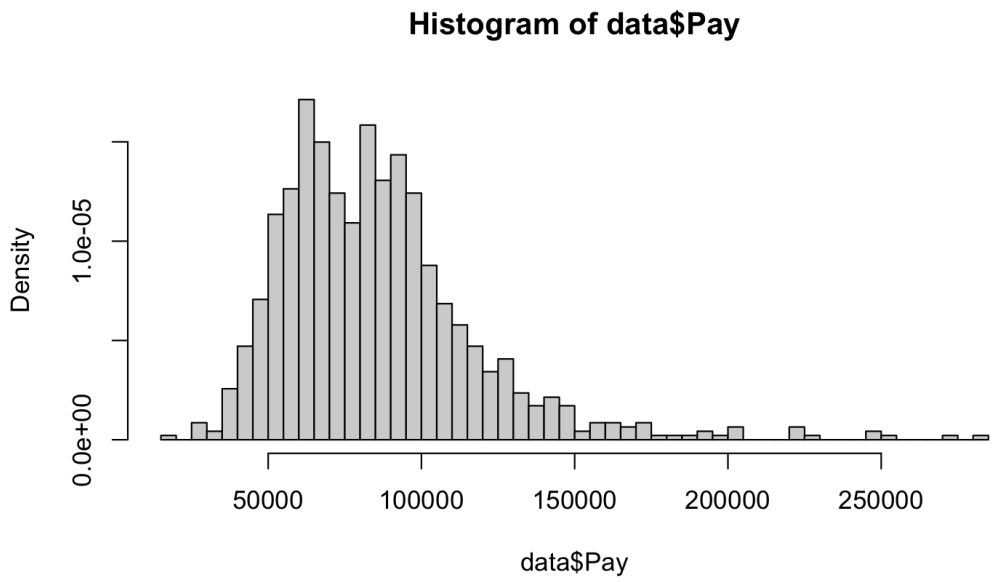

Figure 1: histData histogram plot

Vi kan derefter bekræfte at vi har gjort det rigtige, ved at gøre følgende:

```
sum(hist_data$density * diff(hist_data$breaks))
```

Nu vil vi så tegne tætheden for normalfordeling. Det gør vi også med det vink givet i opgaven, dog med en tilføjet kommando `lwd=**` for at ændre tykkelsen.

```
f1 <- function(x) dnorm(x, mean=payAverage, sd=paySampleDeviation)
curve(f1, add=TRUE, col="red", lwd=2.5)
```

Og her får vi:

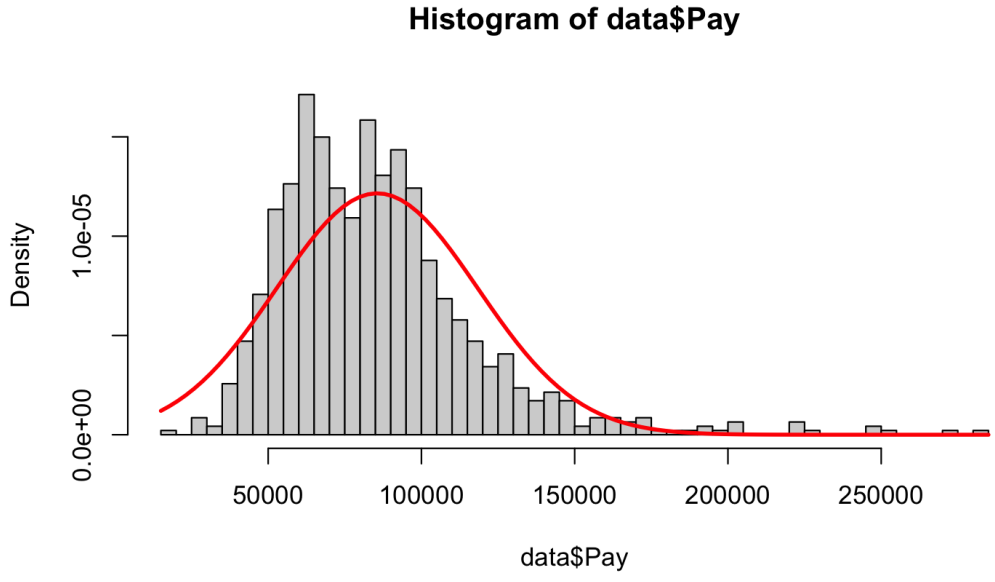

Figure 2: histData histogram plot

Vi har så nu for ***LogPay***:
For selve kodedelen har vi

```
hist(data$logPay, prob=TRUE, breaks=50)
f2 <- function(x) dnorm(x, mean=logPayAverage, sd=logPaySampleDeviation
    )
curve(f2, add=TRUE, col="red", lwd=2.5)
```

Og histogrammet ender med at se ud som:

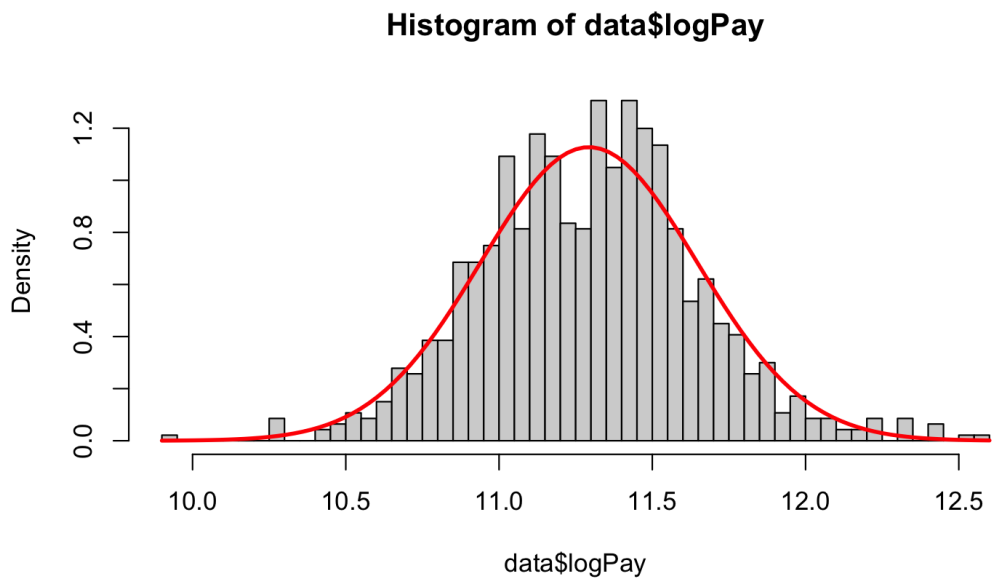

Figure 3: histData histogram plot

For at analysere om Pay eller LogPay er normalfordelt, kan vi tegne deres QQ-plots i RStudio:

```
# For Pay
qqPay <- qqnorm(data$Pay, main="QQ-Plot af Pay")
abline(payAverage, paySampleDeviation, col="red", lwd=2.2, lty=2)
# For LogPay
qqLogPay <- qqnorm(data$logPay, main="QQ-Plot af LogPay")
abline(logPayAverage, logPaySampleDeviation, col="red", lwd=2.2, lty=2)
```

Og vi får så følgende to plots der visualiserer, hvilken er normalfordelt:

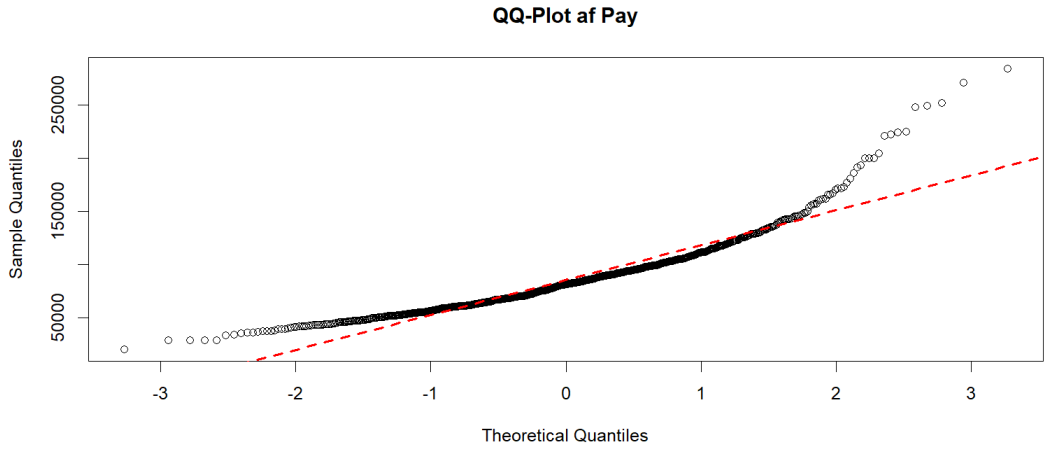

Figure 4: QQ-Plot af Pay

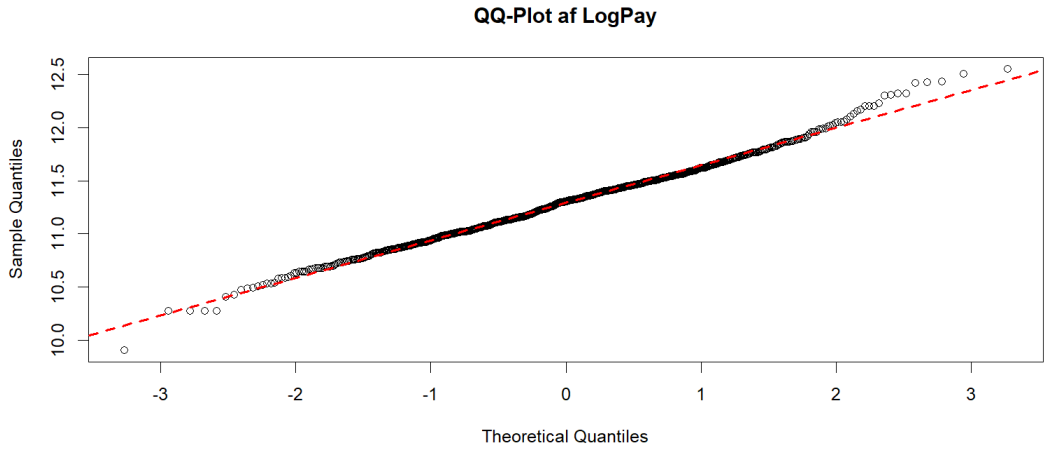

Figure 5: QQ-Plot af LogPay

Ud fra QQ-plotsne kan vi se, at fordeling for LogPay er mere normalfordel en Pay, da punkterne ligger mere som en linje.
Vi vurderer derfor, at LogPay er normalfordelt, mens Pay ikke er normalfordelt.

3 Brug en normalfordeling til at bestemme et fornuftigt estimat for sandsynligheden for at en tilfældig statsansat i Connecticut tjente mere end 100000 USD i 2017. Beregn et andet fornuftigt estimat for den samme sandsynlighed, men hvor du ikke laver nogle normalfordelingsantagelser. Ligger de to estimater tæt på eller langt fra hinanden?

For at bestemme sandsynligheden for, at en tilfældig ansat i Connecticut tjente mere end 100.000 USD, kan vi bestemme arealet under fordelingsfunktionen fra 100.000 til den højeste kendt lønnede statsansat:

```
# med normalfordeling
integrate(f1, 100000, max(data$Pay))
OUTPUT: 0.3301343 with absolute error < 6.1e-09
```

Vi estimerer dermed, at der er 0,33 sandsynlighed for, at en tilfældig statsansat har en løn, som er højere end 100.000 USD ud fra vores data, når vi antager, at lønnen er normalfordelt.
Der bruges her funktionen, `f1` fundet i tidligere opgave.

```
# uden normalfordelingsantagelse
length(data$Pay[data$Pay>=100000])/length(data$Pay)
OUTPUT: 0.245182
```

Vi kan se, at vi får et meget mere præcist resultat ved at bruge integration for en normalfordeling, end ved at dividere antallet af ansatte, der tjener over 100000 USD med det totale antal testdeltagere.

$$Forskel = 0.33 - 0.245 = 0.085$$

Der er stor forskel på, hvad vores sandsynligheder bliver, baseret på om vi antager, at lønningen er normalfordelt eller ej.

Lad i det følgende X ***være en stokastisk variabel med*** $X \sim N(\mu, \sigma^2)$***, og definer*** $Y = e^X$***.***

4 Vis at tætheden for Y er givet ved

$$f(y) = \frac{1}{y\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\log y - \mu)^2}{2\sigma^2}\right), y > 0.$$

Du kan enten bruge transformationssætningen (BH, Sætning 8.1) eller først bestemme fordelingsfunktionen (cdf'en) for Y og dernæst tætheden. Bemærk at transformationssætningen gennemgås ved forelæsningerne fredag i kursusuge 4 (eller måske først mandag i kursusuge 5), hvorimod I allerede nu har de nødvendige redskaber til rådighed til at løse opgaven ved at gå via cdf'en.

Fordelingen med tæthed f kaldes den logaritmiske normalfordeling, eller bare log-normalfordelingen — fordi logaritmen til en stokastisk variabel med denne tæthed er normalfordelt

Den generelle tæthedsfunktion for en normalfordeling givet som $X \sim N(\mu, \sigma^2)$ kan beskrives som:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

Vi ved ud fra opgavebeskrivelsen, at:

$$Y = e^X \Leftrightarrow \log(Y) = X$$

For at transformere tæthedsfunktionen for X til at bestemme tæthedsfunktionen for Y, kan vi bruge teorem 8.1.1 fra BH:

$$f(y) = f(x) \cdot \left|\frac{\partial x}{\partial y}\right|$$

Vi indsætter tæthedsfunktionen for X, og vi differentierer X, som er log(Y) med hensyn til Y:

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \cdot \left|\frac{\partial \log(y)}{\partial y}\right|$$

$$\Leftrightarrow \; f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \cdot \left|\frac{1}{y}\right|$$

$$\Leftrightarrow \; f(y) = \frac{1}{y\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Til sidst kan vi også erstatte x med log(y), da dette er det samme:

$$f(y) = \frac{1}{y\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\log(y)-\mu)^2}{2\sigma^2}\right)$$

5 Forklar hvorfor det er fornuftigt at bruge log-normal-fordelingen til at beskrive fordelingen af Pay. Gør herunder rede for hvilke af de værdier I beregnede i spørgsmål 1, som er fornuftige at bruge som μ og σ^2 i definitionen af tætheden f. Tegn histogrammet over Pay igen. Indtegn derefter grafen for tætheden f med de relevante værdier af μ og σ i samme graf, og kommentér grafen.

Hvis man visuelt kigger på de to QQ-plots vi har lavet tidligere, kan vi se at plottet over ***Pay*** viser en ikke normalfordeling i løn hos de forskellige medarbejdere, hvorimod QQ-plottet og fordelingen af ***LogPay*** er normalfordelt.
Vi fortrækker derfor, at kigge på fordelingen af logPay, da denne repræsenterer løsnen som en normalfordeling.

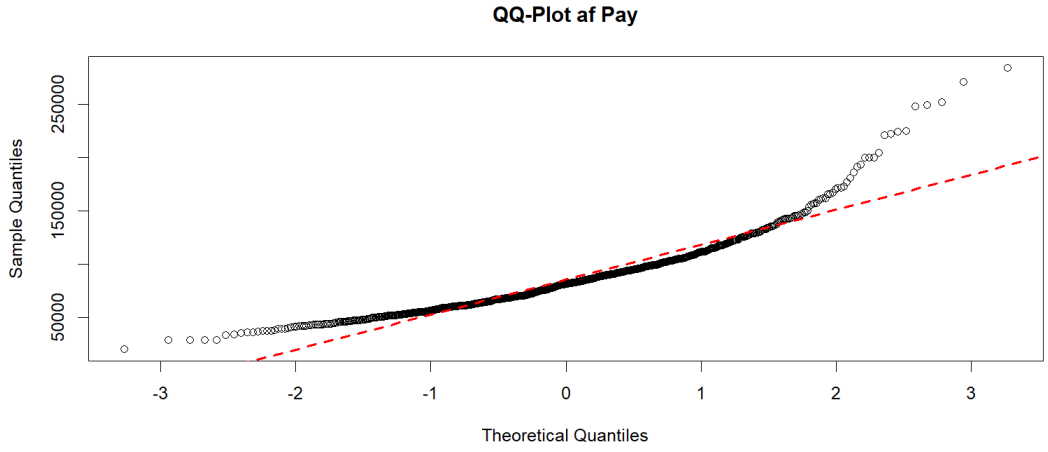

Figure 6: QQ-Plot af Pay

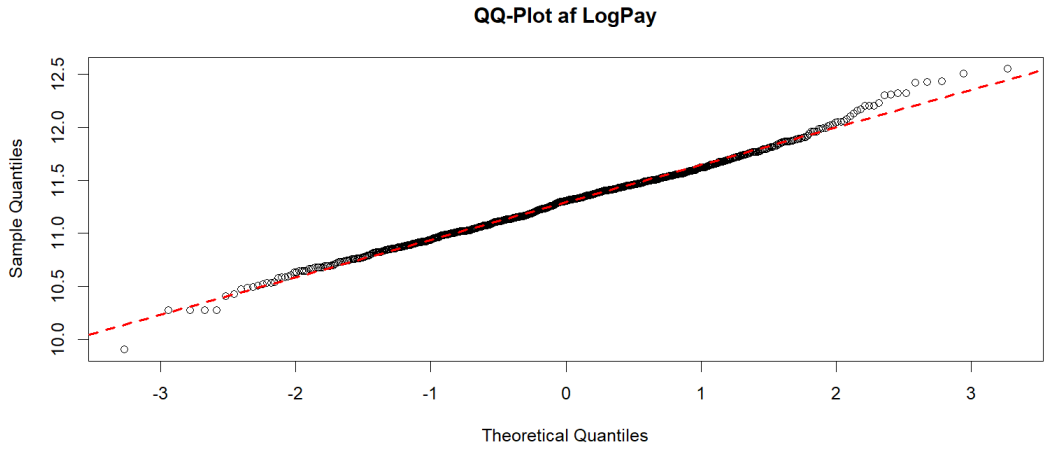

Figure 7: QQ-Plot af LogPay

Vi kan altså meget nemmere se hvordan fordelingen af løn forløber sig ved at bruge log-normalfordelingen.

Da vi tegnede histogrammet for henholdsvis ***Pay*** & ***LogPay***, valgte vi også at tegne en linje der bestod af μ & σ fordi det gjorde det mere visuelt intuitivt at forstå. Vi tegnede da:

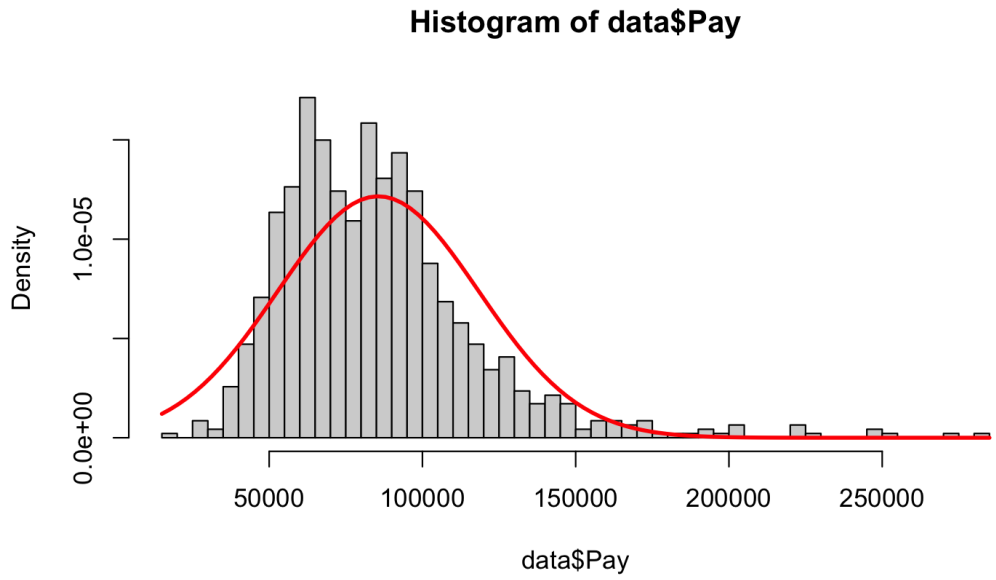

Figure 8: histData histogram plot

Histogrammet set er tegnet ved:

```
hist(data$Pay, prob=TRUE)
f1 <- function(x) dnorm(x, mean=payAverage, sd=paySampleDeviation)
curve(f1, add=TRUE, col="red", lwd=2.5)
```

Medianen for en kontinuert stokastisk variabel Z med fordelingsfunktion F er den værdi c hvor $F(c) = 0.5$, altså $P(Z \leq c) = 0.5$.
Antag som i spørgsmål 4 at $X \sim N(\mu, \sigma^2)$ og at $Y = e^X$

6 Gør rede for at X har median μ, og vis derefter at Y har median e^μ.

Ud fra vinket givet i opgaven, ved vi at $P(Z \leq c) = 0.5$. Vi kan omformulere dette til $P(X \leq \mu) = 0.5$, hvilket betyder, at sandsynligheden for at $X \leq \mu$ er 0.5, og altså må $P(X \geq \mu)$ også være lig 0.5, da vi får sandsynligheden for $X \geq \mu$ ved $1 - P(X \leq \mu) = 0.5$. Vi kan derfor sige at normalfordelingen er symmetrisk, og derfor er det gældende at X har medianen μ.
For vores normalfordeling $X \sim (0, 1)$ kan vi beskrive dette som:

$$P\left(X \leq \mu\right) = \int_{-\infty}^{0} f_X\left(x\right) \, \mathrm{d}x = 0,5$$

og

$$P\left(X \geq \mu\right) = \int_{0}^{\infty} f_X\left(x\right) \, \mathrm{d}x = 0,5$$

For Y's median ved vi fra forrige opgave at:

$$\log(Y) = X \Leftrightarrow Y = e^X$$

Og da X har medianen μ, må $\log(Y)$ altså også have samme median. Derfor må medianen til Y altså være givet ved e^μ.

7 Et af følgende udtryk for $E(Y)$ er korrekt:

$$E(Y) = e^\mu, \quad E(Y) = e^{\mu - \sigma^2}, \quad E(Y) = e^{\mu + \sigma^2},$$
$$E(Y) = e^{\mu + \sigma^2/2}, \quad E(Y) = e^{\mu + \sigma}$$

Undersøg ved hjælp af simulation med $\mu = 0$ og $\sigma = 1.5$ hvilket af udtrykkene der er det korrekte. Brug mindst 10000 simulerede værdier. (Man kan ikke vise ved simulation at et sådant udtryk gælder generelt, men fire af udtrykkene kan udelukkes, og et af dem er faktisk korrekt!)

Først bruger vi `rnorm(n, mean, sd` til at finde alle X's værdier. Herefter tager vi alle værdierne i `exp()` for at få deres eksponentielle værdi (der følger beregningerne tidligere, $Y = e^X$). Vi tager herefter middelværdien af de værdier, vi har beregnet, for at få et udgangspunkt, vi kan bruge. Til sidst trækker vi den fundne middelværdi fra de forskellige formler stillet i opgaven, for at se hvilken formel der kommer tættest på middelværdien. I R skriver vi:

```r
simX <- rnorm(100000, 0, 1.5)
simY <- exp(simX)
# Vi beregner middelv rdien for simulationerne for Y
mean(simY)
# Vi beregner middelv rdien for alle andre givne mulige beskrivelser
abs(exp(0)-mean(simY))
abs(exp(0 - (1.5^2))-mean(simY))
abs(exp(0 + (1.5^2))-mean(simY))
abs(exp(0 + (1.5^2) / 2)-mean(simY))
abs(exp(0 + 1.5)-mean(simY))
# OUTPUT:
# Vi beregner middelv rdien for simulationerne for Y
# mean(simY)
# [1] 2.985198
# Vi beregner middelv rdien for alle andre givne mulige udtryk, og
    tager differencen
# abs(exp(0)-mean(simY))
# [1] 1.985198
# abs(exp(0 - (1.5^2))-mean(simY))
# [1] 2.879798
# abs(exp(0 + (1.5^2))-mean(simY))
# [1] 6.502538
# abs(exp(0 + (1.5^2) / 2)-mean(simY))
# [1] 0.09501914
# abs(exp(0 + 1.5)-mean(simY))
# [1] 1.496491
```

Vi kan ud fra vores simulationer se, at den praktiske middelværdi for Y er 2,986.
Vi har derefter valgt at beregne differencen mellem middelværdierne for alle de givne udtryk og den simulerede middelværdi for Y.
Ud fra vores resultater, kan vi konkludere, at $E(Y) = e^{\mu+\sigma^2/2}$, er det korrekte udtryk for middelværdien af Y.

8 Brug det relevante udtryk og de relevante værdier for μ og σ^2 til at bestemme et fornuftigt estimat for den gennemsnitlige løn for statsansatte i Connecticut i 2017. Sammenlign med det relevante tal fra tabellen fra Spørgsmål 1 (hvilket?).

Hvis vi sammenligner med tabellen vi lavede ved Spørgsmål 1, kan vi se, at vi der fik gennemsnittet til 85577,221 USD:

	Median	Gennemsnit	Stikprøvevarians	Stikprøvespredning
Pay	81459,685	85577,221	1076703934,22	32813,167
LogPay	11,308	11,293	0,125	0,354

Table 3: Tabel med data fra R

Vi kan bruge vores fundne udtryk for middelværdien for Y, som vi fandt i opgaven før:

$$E(Y) = e^{\mu+\sigma^2/2}$$

Vi indsætter nu vores gennemsnit og spredning:

$$E(Y) = e^{11,293+0,354^2/2} = 85447,55$$

Dette er vores estimat for den gennemsnitlige løn for en statsansat in Connecticut
Man bruger normalt såkaldte QQ-plots til at vurdere om data kan antages at være normalfordelte. Dette er beskrevet DS på side 68-71 og side 76 (om R).

9 Læs teksten om QQ-plots i DS, og lav QQ-plots for *Pay* og *LogPay*. Kommentér figurerne. Er konklusionen fra figurerne i overensstemmelse med jeres undersøgelser tidligere i opgaven?

Vi havde lavet QQ-plots i de tidligere opgaver, da vi mente, at QQ-plots'ne ville være nyttige for at analysere vores data.

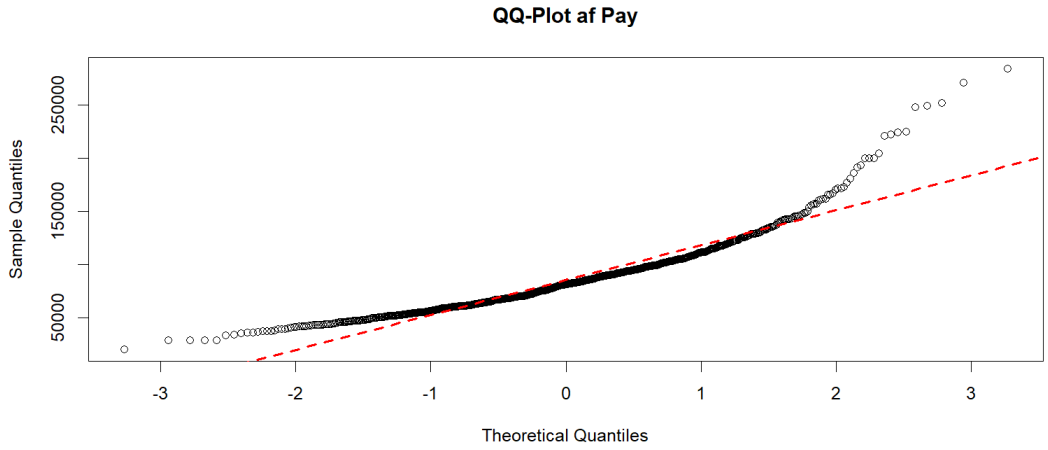

Figure 9: QQ-Plot af Pay

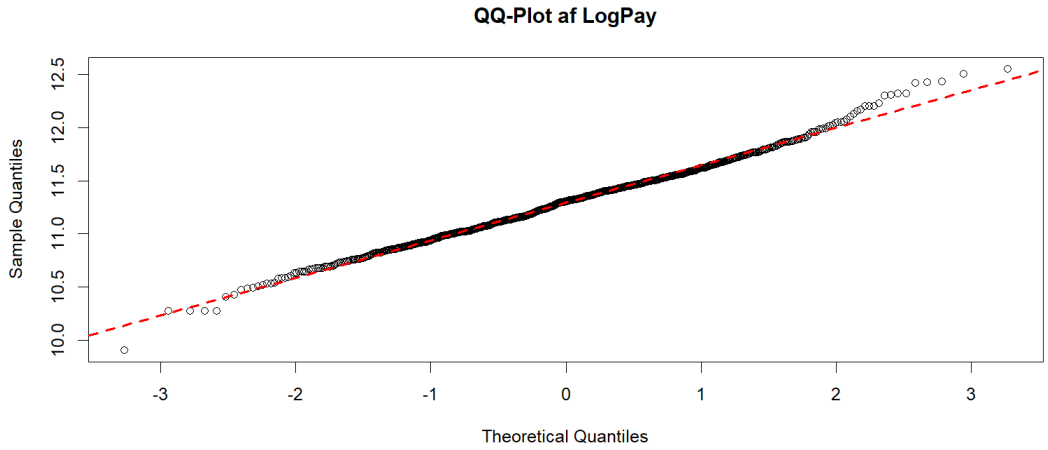

Figure 10: QQ-Plot af LogPay

De plots vi lavede stemmer fint overens med vores undersøgelser fra de tidligere opgaver, da vi kom frem til samme konklusion før.