Problem 2

Hindy Rossignol, Riya Parikh, Mrugank Pednekar, Ioannis Panagiotopoulos

2025-09-11

R setup

```
library(readr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(fitdistrplus)
```

```
## Loading required package: MASS
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
## Loading required package: survival
```

Problem 2

Part a

```
# loading df
df_insurance <- read_csv("/Users/riyaparikh_computeracct/Downloads/MIT/15.072_AdvancedAn
alyticsEdge/deliverable1-analyticsedge-mit/Data/insurance.csv", show_col_types = FALSE)
```

Part b

```
mean_charges = mean(df_insurance$charges)
sd_charges = sd(df_insurance$charges)
```

Mean = 13270.42, SD = 12110.01

Part c

```
probability_between_values = pnorm(14000, mean_charges, sd_charges) - pnorm(8000, mean_c
harges, sd_charges)
```

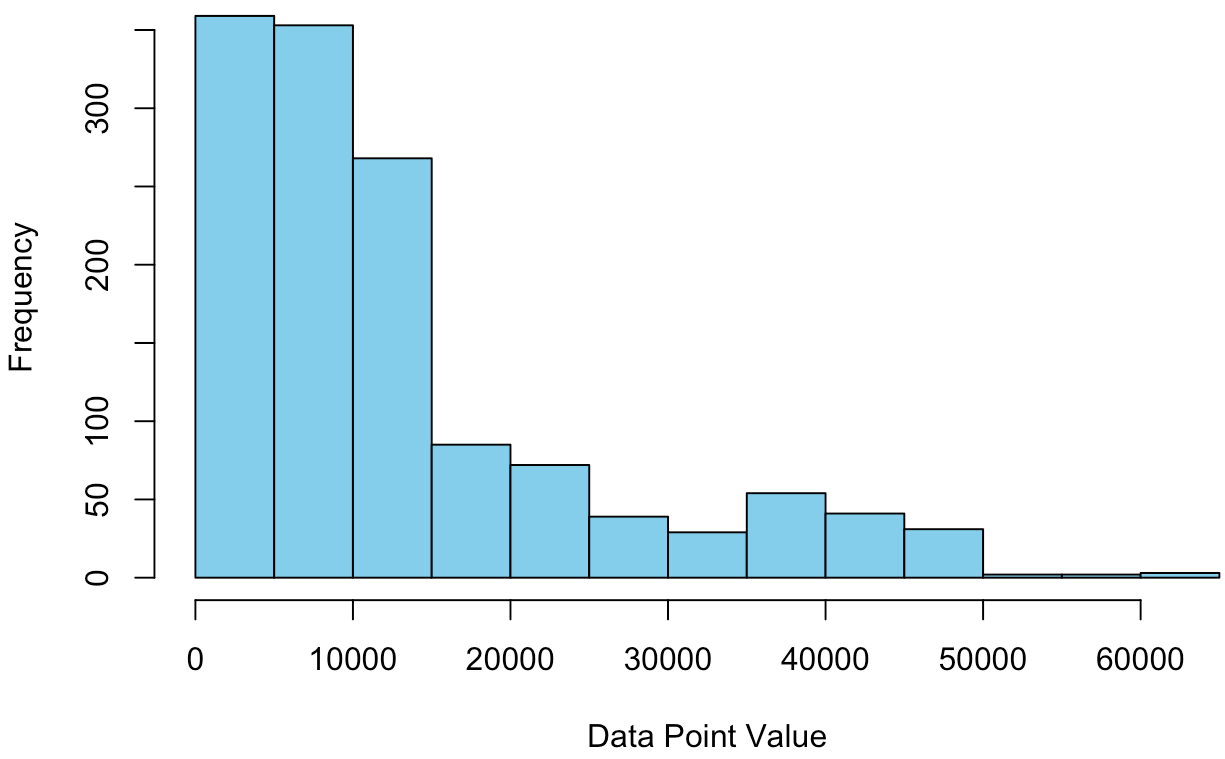
The probability of being between $8000 and $14000 based on the Normal distribution is 0.19.

Part d

```
count_in_range <- sum(df_insurance$charges >= 8000 & df_insurance$charges <= 14000)
total_vals = length(df_insurance$charges)
true_prop_between_values = count_in_range/total_vals
median(df_insurance$charges)
```

```
## [1] 9382.033
```

```
hist(df_insurance$charges, main = "Histogram of Charges Data Points", xlab = "Data Point
Value", col = "skyblue")
```



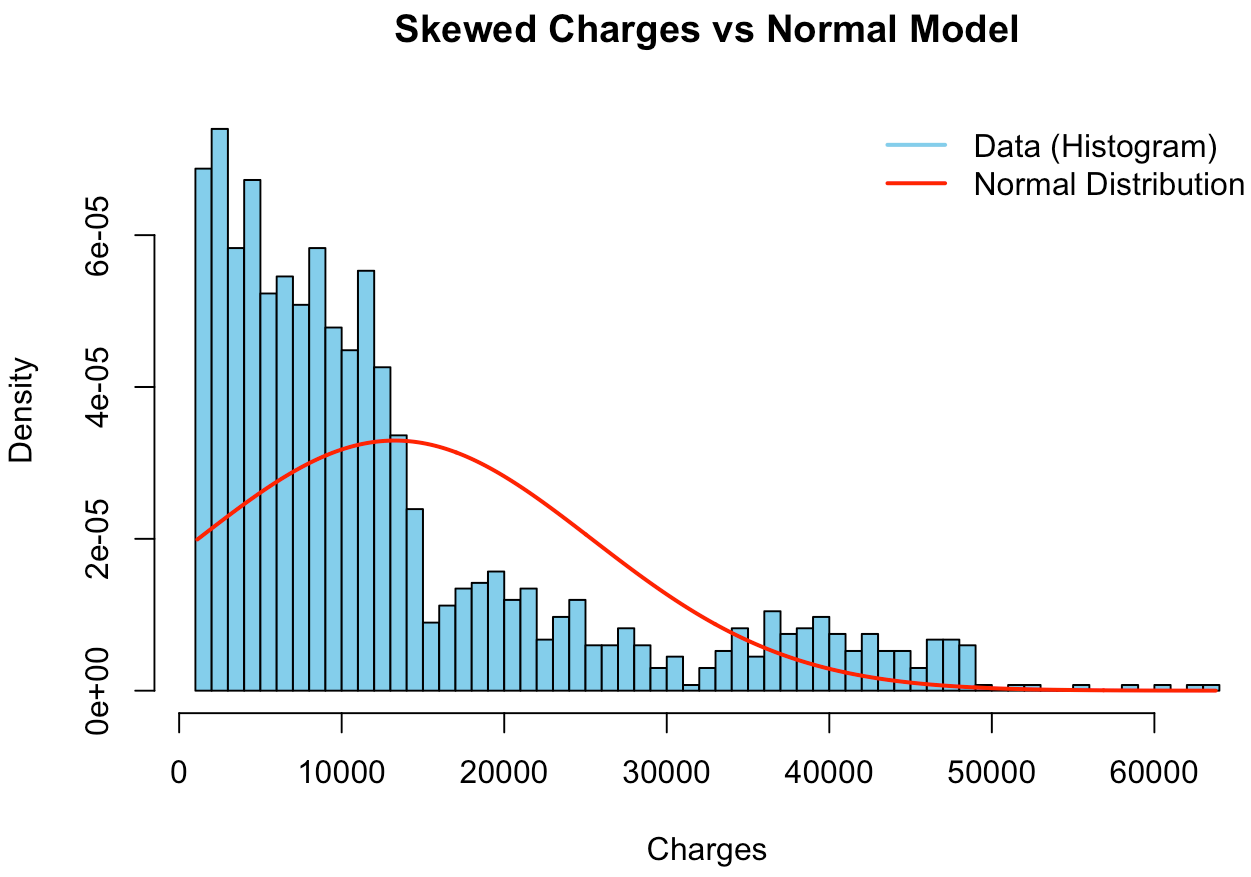

```
# to see overlapping stats plots - ours vs normal model
# calculate values for comparison
count_in_range <- sum(df_insurance$charges >= 8000 & df_insurance$charges <= 14000)
total_vals <- length(df_insurance$charges)
true_prop_between_values <- count_in_range / total_vals
print(true_prop_between_values)
```

```
## [1] 0.2825112
```

```
median(df_insurance$charges)
```

```
## [1] 9382.033
```

```r
# histogram of charges (empirical distribution)
hist(df_insurance$charges,
     breaks = 50,                # more detail
     freq = FALSE,               # scale histogram to density
     main = "Skewed Charges vs Normal Model",
     xlab = "Charges",
     col = "skyblue")

# overlay fitted normal distribution curve
x_vals <- seq(min(df_insurance$charges), max(df_insurance$charges), length.out = 1000)
y_norm <- dnorm(x_vals, mean = mean_charges, sd = sd_charges)

lines(x_vals, y_norm, col = "red", lwd = 2)

# add legend
legend("topright", legend = c("Data (Histogram)", "Normal Distribution"),
       col = c("skyblue", "red"), lwd = 2, bty = "n")
```



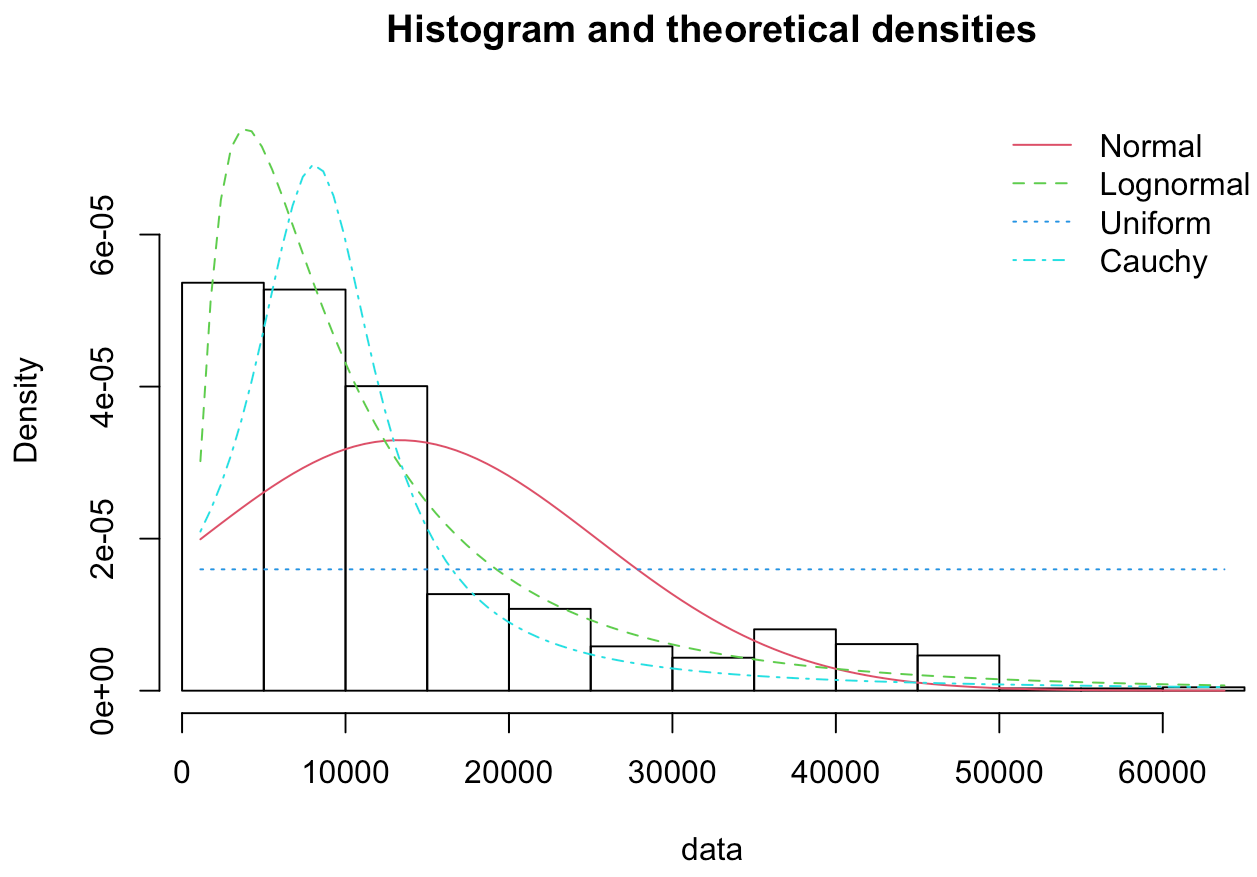
Actual proportion of charges between $8000 and $14000 is 0.28. This is not the same as the probability we got when assuming the charges follow a Normal distribution in part c. Our initial assumption is that the charges do not follow a Normal distribution. The histogram we plotted of charges data points shows that the distribution is right skewed towards higher charges. We also see that the median charge is lower than the mean charge ($9382.033 vs $13270.42).

Part e

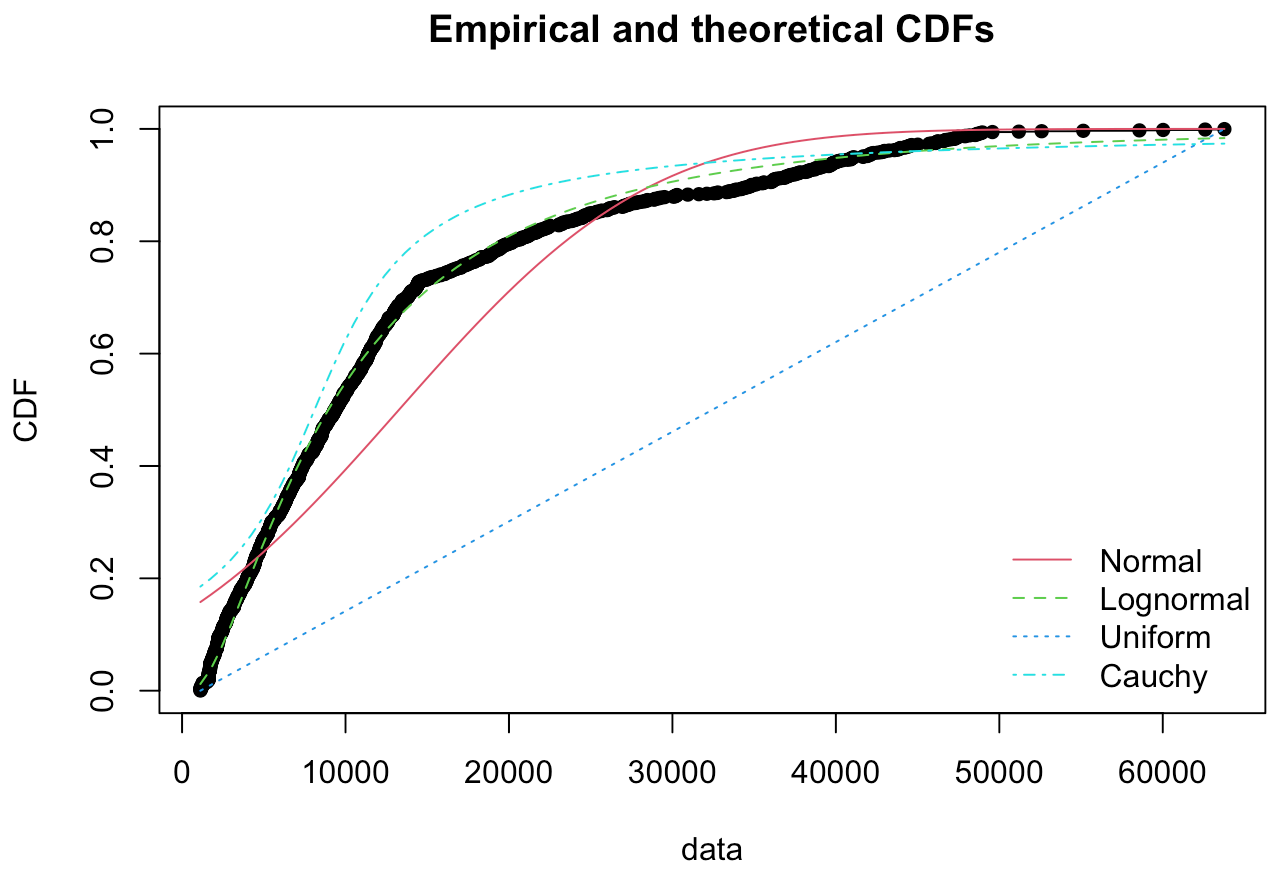
```
fit_norm <- fitdist(df_insurance$charges, "norm")
fit_lognorm <- fitdist(df_insurance$charges, "lnorm")
fit_unif <- fitdist(df_insurance$charges, "unif")
fit_cauchy <- fitdist(df_insurance$charges, "cauchy")

gofstat(list(fit_norm, fit_lognorm, fit_unif, fit_cauchy))
```

```
## Goodness-of-fit statistics
##                                1-mle-norm 2-mle-lnorm  3-mle-unif 4-mle-cauchy
## Kolmogorov-Smirnov statistic   0.188462   0.0365844   0.5147556     0.185312
## Cramer-von Mises statistic    14.829729   0.3973136 155.5524079     9.501239
## Anderson-Darling statistic    85.138872   3.9424972         Inf    65.856208
##
## Goodness-of-fit criteria
##                                  1-mle-norm 2-mle-lnorm 3-mle-unif 4-mle-cauchy
## Akaike's Information Criterion   28959.26    27923.58   29561.21     28716.76
## Bayesian Information Criterion   28969.66    27933.98   29571.61     28727.16
```

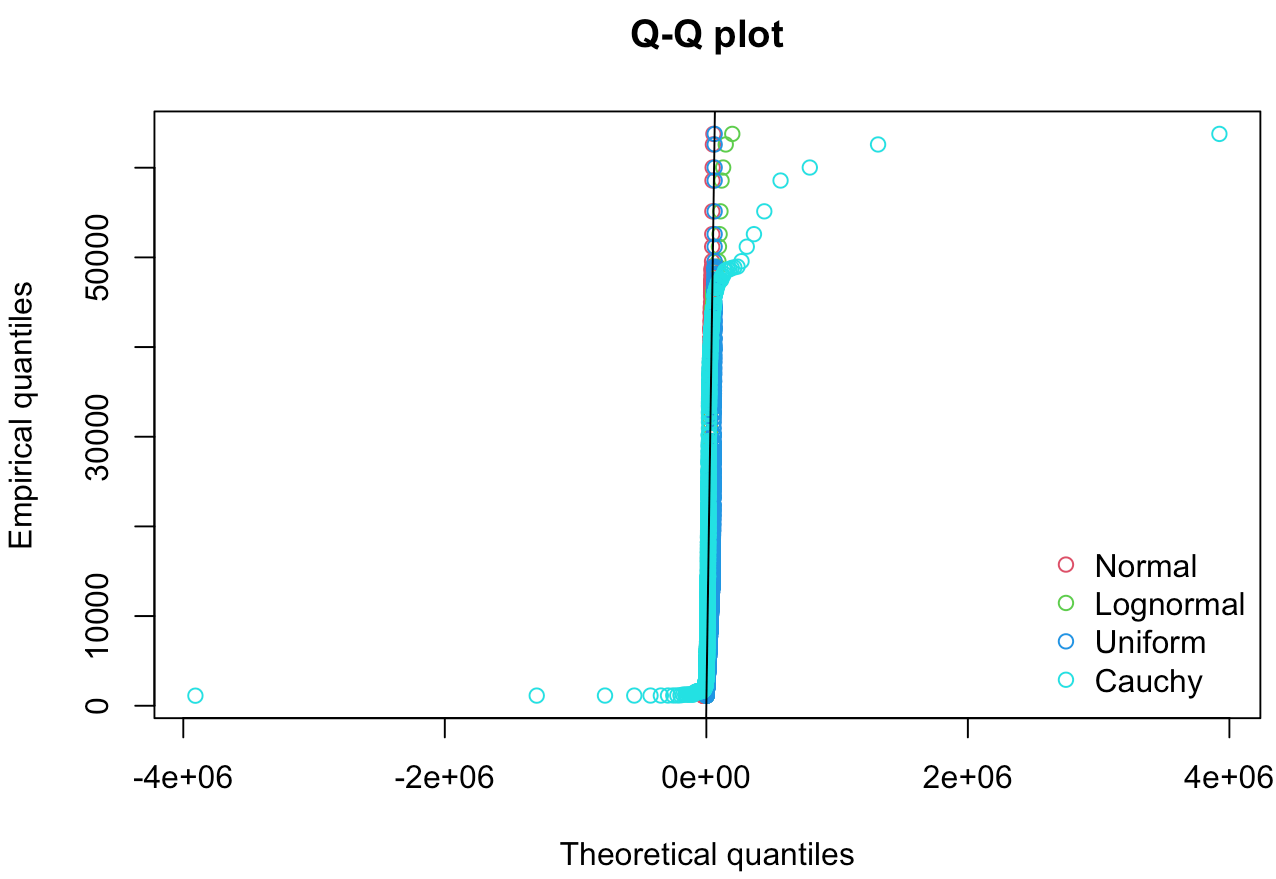
```
plot.legend <- c("Normal", "Lognormal", "Uniform", "Cauchy")

denscomp(list(fit_norm, fit_lognorm, fit_unif, fit_cauchy), legendtext = plot.legend)
```



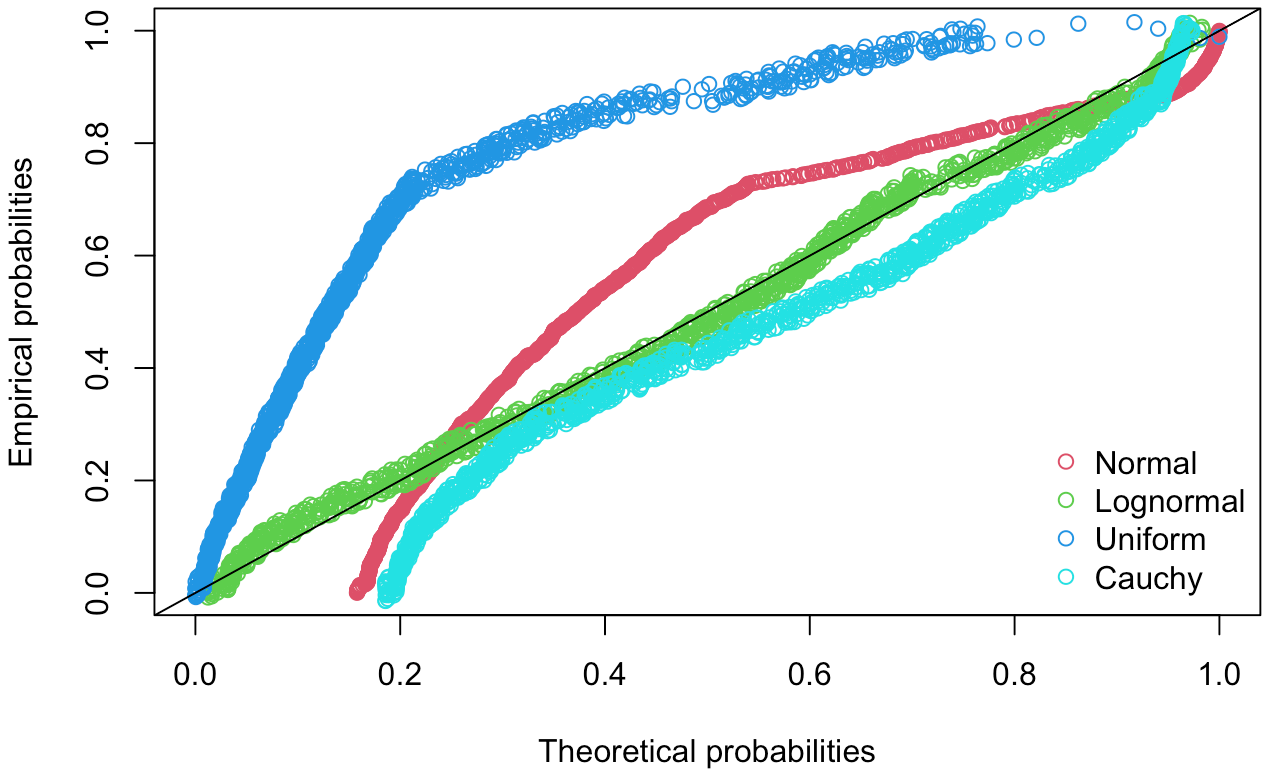
```
cdfcomp(list(fit_norm, fit_lognorm, fit_unif, fit_cauchy), legendtext = plot.legend)
```




```
qqcomp(list(fit_norm, fit_lognorm, fit_unif, fit_cauchy), legendtext = plot.legend)
```




```
ppcomp(list(fit_norm, fit_lognorm, fit_unif, fit_cauchy), legendtext = plot.legend)
```




```
quantile(df_insurance$charges, probs = c(0.25, 0.5, 0.75))
```

```
##       25%       50%       75%
##  4740.287  9382.033 16639.913
```

```
qlnorm(c(0.25, 0.5, 0.75),
       meanlog = fit_lognorm$estimate["meanlog"],
       sdlog = fit_lognorm$estimate["sdlog"])
```

```
## [1]  4811.090  8943.289 16624.595
```

Based on both statistical metrics and visual diagnostics, the lognormal distribution is the best fit for the insurance charges data. It consistently outperforms alternatives across goodness-of-fit statistics, with a notably lower KS statistic indicating a closer match between empirical and theoretical distributions. The quantile comparison confirms that lognormal quantiles align tightly with the actual data, especially in the right tail.

To support this conclusion, we examine four key plots:

- Histogram with Density Curves: The histogram reveals a pronounced right skew in the data. Among the overlaid theoretical curves, the lognormal density (green dashed line) hugs the histogram most closely, especially in the peak and tail regions. The Normal and uniform curves fail to capture the asymmetry, while the Cauchy curve overestimates the tail.

- Empirical vs. Theoretical CDFs: The lognormal CDF tracks the empirical CDF almost perfectly across the entire range. The Normal CDF diverges in the upper tail, underestimating high charges. Uniform and Cauchy distributions show poor alignment, with noticeable deviations throughout.
- Q-Q Plot: The lognormal quantiles (green triangles) lie closest to the diagonal, indicating strong agreement with the empirical quantiles. The Normal distribution shows curvature, especially in the tails, while the Cauchy and uniform quantiles deviate substantially, suggesting poor fit.
- P-P Plot: Lognormal points cluster tightly around the diagonal, confirming that the predicted probabilities match the observed proportions well. Normal distribution points begin to drift in the upper range, and both uniform and Cauchy show systematic deviations.

Taken together, these plots reinforce the statistical conclusion: the lognormal distribution best captures the shape, spread, and tail behavior of the insurance charges data. It accommodates the skewness and heavy upper tail, which the normal distribution fails to model adequately.

Part f

```
mod <-  lm(charges ~ age + factor(sex) + bmi + children + factor(smoker) + factor(regio
n) , data = df_insurance)
summary_mod = summary(mod)
summary_mod
```

```
##
## Call:
## lm(formula = charges ~ age + factor(sex) + bmi + children + factor(smoker) +
##     factor(region), data = df_insurance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11304.9  -2848.1   -982.1   1393.9  29992.8
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -11938.5      987.8 -12.086  < 2e-16 ***
## age                        256.9       11.9  21.587  < 2e-16 ***
## factor(sex)male           -131.3      332.9  -0.394 0.693348
## bmi                        339.2       28.6  11.860  < 2e-16 ***
## children                   475.5      137.8   3.451 0.000577 ***
## factor(smoker)yes        23848.5      413.1  57.723  < 2e-16 ***
## factor(region)northwest   -353.0      476.3  -0.741 0.458769
## factor(region)southeast  -1035.0      478.7  -2.162 0.030782 *
## factor(region)southwest   -960.0      477.9  -2.009 0.044765 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6062 on 1329 degrees of freedom
## Multiple R-squared:  0.7509, Adjusted R-squared:  0.7494
## F-statistic: 500.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```

Part g

It is important to note that for the following interpretations, the change per unit that is being described only holds true if all other factors are held constant.

Age: for every increase of 1 yr in age, the charges are predicted to increase by $256.9, holding other variables constant. this variable is statistically significant as the pvalue is <0.05 at 2e-16.

Factor(sex)male: being a male, the charges are predicted to be lower by $131.3 compared to their equivalent female counterparts, holding other variables constant. this variable is not statistically significant as the pvalue is >0.05 at 0.693348.

BMI: for every increase of 1 unit in bmi, the charges are predicted to increase by $339.2, holding other variables constant. this variable is statistically significant as the pvalue is <0.05 at 2e-16.

Children: for every additional child the customer has, the charges are predicted to increase by $475.5, holding other variables constant. this variable is statistically significant as the pvalue is <0.05 at 0.000577.

Factor(smoker)yes: being a smoker, the charges are predicted to be higher by $23848.5 compared to their nonsmoking counterparts, holding other variables constant. this variable is statistically significant as the pvalue is <0.05 at 2e-16.

Factor(region)northwest: if a customer lives in the northwest, their charges are predicted to be lower by $353.0 holding all other variables constant. this variable is not statistically significant as the pvalue is <0.05 at 0.458769.

Factor(region)southeast: if a customer lives in the southeast, their charges are predicted to be lower by $1035.0 holding all other variables constant. this variable is statistically significant as the pvalue is <0.05 at 0.030782.

Factor(region)southwest: if a customer lives in the southwest, their charges are predicted to be lower by $960.0 holding all other variables constant. this variable is statistically significant as the pvalue is <0.05 at 0.044765.

Part h

```
mod2 <-  lm(charges ~ factor(sex) + bmi + children + factor(smoker) + factor(region) , d
ata = df_insurance)
summary_mod2 = summary(mod2)
summary_mod2
```

```
##
## Call:
## lm(formula = charges ~ factor(sex) + bmi + children + factor(smoker) +
##     factor(region), data = df_insurance)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -15013  -4646   -945   3652  32122
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -3953.0     1064.1  -3.715 0.000212 ***
## factor(sex)male           -310.9      386.7  -0.804 0.421590
## bmi                        411.3       33.0  12.464  < 2e-16 ***
## children                   597.5      160.0   3.735 0.000196 ***
## factor(smoker)yes        23658.9      479.9  49.303  < 2e-16 ***
## factor(region)northwest   -392.4      553.3  -0.709 0.478359
## factor(region)southeast  -1410.3      555.7  -2.538 0.011274 *
## factor(region)southwest  -1031.9      555.2  -1.859 0.063312 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7043 on 1330 degrees of freedom
## Multiple R-squared:  0.6636, Adjusted R-squared:  0.6618
## F-statistic: 374.8 on 7 and 1330 DF,  p-value: < 2.2e-16
```

The first model had an r^2 of .751 meaning that 75.1% of the variation in charges is accounted for by the model built on all the factors. the adj r^2 is .749. The second model (built without age) has an r^2 of .664 meaning that 66.4% of the variation in charges is accounted for by the model built on all the factors without age. the adj r^2 is .662. When age is not considered as an independent variable, the model is forced to “spread” the influence of age across other variables like BMI, smoking status, and region. Younger individuals are likely to be overcharged: without age, the model can’t distinguish between a healthy 25-year-old and a 55-year-old with similar BMI and smoking status. Older individuals are likely to be undercharged: older individuals typically incur higher medical costs, which the model without age can’t fully capture.

Part i

I would recommend applying a log transformation to the charges variable, and then redoing the analysis using the transformed variable to assess linear relationships and model assumptions more effectively.