Normal Probability Plot

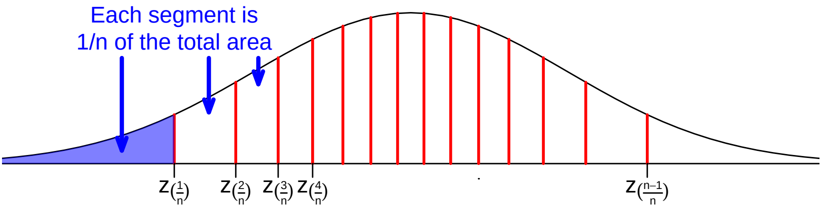
Yibi Huang
Department of Statistics
University of Chicago

How to Check the Normality of Errors

- Histogram of the residuals: if normal, should be bell-shaped
  - Pros: simple, easy to understand
  - Cons: for a small sample, histogram may not be bell-shaped even though the sample is from a normal distribution
- *Normal probability plot* of the residuals
  - aka. *normal QQ plot*,
    QQ stands for “quantile-quantile”
  - best tool to assess normality
  - See next slide for details

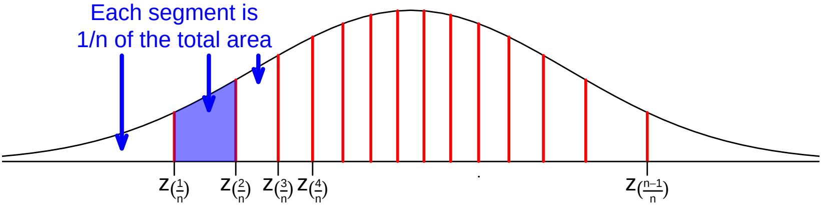
Ideas Behind the Normal Probability Plot (1)

- Data: $y_1, y_2, \ldots, y_n$
- Sorted Data: $y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(n)}$,
  call the **Sample Quantiles**
- **Theoretical Quantiles** of the N$(0, 1)$: $z_{(\frac{1}{n})}, z_{(\frac{2}{n})}, \ldots, z_{(\frac{n-1}{n})}$,
  where, $z_{(\frac{k}{n})}$ is a value such that $\mathrm{P}(Z \leq z_{(\frac{k}{n})}) = \dfrac{k}{n}$ for $Z \sim N(0, 1)$.



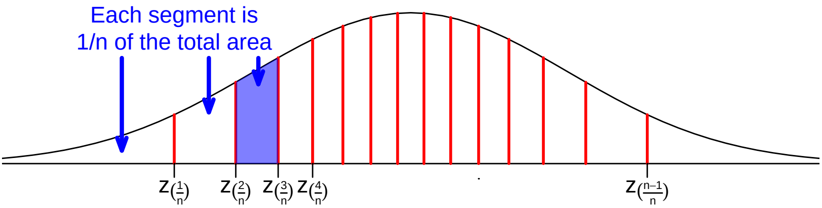
Ideas Behind the Normal Probability Plot (1)

- Data: $y_1, y_2, \ldots, y_n$
- Sorted Data: $y_{(1)} \le y_{(2)} \le \ldots \le y_{(n)}$,
  call the **Sample Quantiles**
- **Theoretical Quantiles** of the $\mathrm{N}(0, 1)$: $z_{(\frac{1}{n})}, z_{(\frac{2}{n})}, \ldots, z_{(\frac{n-1}{n})}$,
  where, $z_{(\frac{k}{n})}$ is a value such that $\mathrm{P}(Z \le z_{(\frac{k}{n})}) = \dfrac{k}{n}$ for $Z \sim N(0, 1)$.



Ideas Behind the Normal Probability Plot (1)

- Data: $y_1, y_2, \ldots, y_n$
- Sorted Data: $y_{(1)} \le y_{(2)} \le \ldots \le y_{(n)}$,
  call the **Sample Quantiles**
- **Theoretical Quantiles** of the N(0, 1): $z_{(\frac{1}{n})}, z_{(\frac{2}{n})}, \ldots, z_{(\frac{n-1}{n})}$,
  where, $z_{(\frac{k}{n})}$ is a value such that $P(Z \le z_{(\frac{k}{n})}) = \dfrac{k}{n}$ for $Z \sim N(0, 1)$.



Ideas Behind the Normal Probability Plot (2)

- If $Y \sim N(\mu, \sigma^2)$, then

$$P(Y \leq \mu + \sigma z_{(\frac{k}{n})}) = P\Bigg(\underbrace{\frac{Y - \mu}{\sigma}}_{\sim N(0,1)} \leq z_{(\frac{k}{n})}\Bigg) = \frac{k}{n}$$

We expected k/n of the observations to be $\leq \mu + \sigma z_{(\frac{k}{n})}$

- We observe k/n of the observations are $\leq y_{(k)}$.
- If the data are indeed $N(\mu, \sigma^2)$, we expect

$$y_{(k)} \approx \mu + \sigma z_{(\frac{k}{n})}$$

- If one plots the Sample Quantiles $y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(n)}$ against the Theoretical Quantiles $z_{(\frac{1}{n})}, z_{(\frac{2}{n})}, \ldots, z_{(\frac{n-1}{n})}$, the points would fall on the straight line

$$y = \mu + \sigma z.$$

if the data follow $N(\mu, \sigma^2)$

A Technical Remark

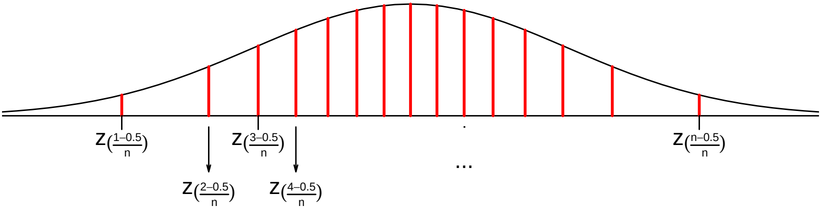
R actually uses the Theoretical Quantiles:

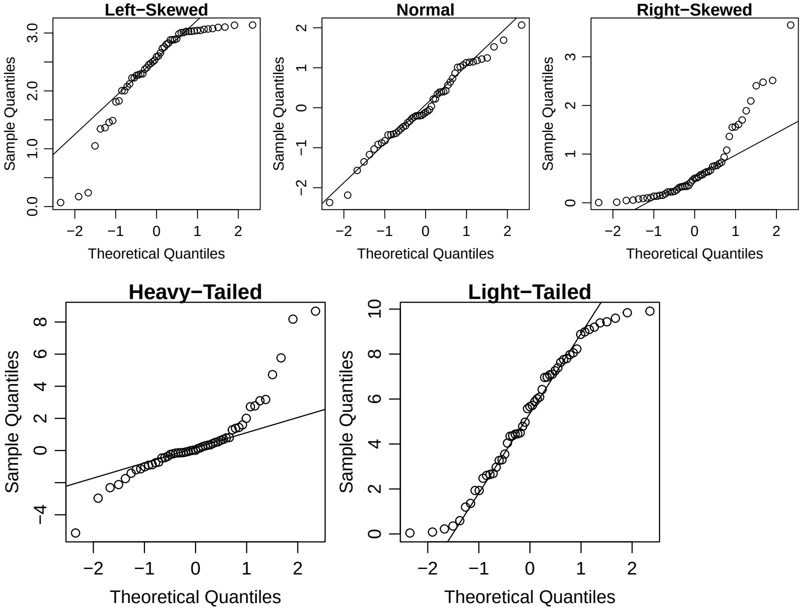
$$z_{(\frac{1-0.5}{n})}, \quad z_{(\frac{2-0.5}{n})}, \quad z_{(\frac{3-0.5}{n})}, \quad \ldots, \quad z_{(\frac{n-0.5}{n})}$$

instead of

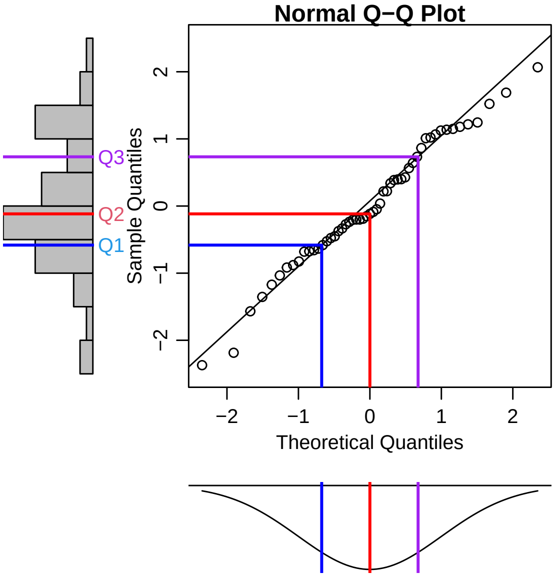
$$z_{(\frac{1}{n})}, \quad z_{(\frac{2}{n})}, \ldots, z_{(\frac{n-1}{n})}, z_{(\frac{n}{n})},$$

since $z_{(n/n)} = \infty$.

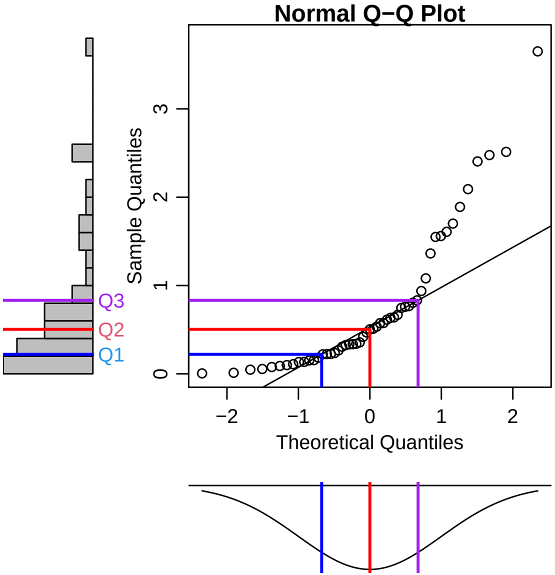



Left−Skewed
Normal
Right−Skewed
Heavy−Tailed
Light−Tailed
Sample Quantiles
Theoretical Quantiles

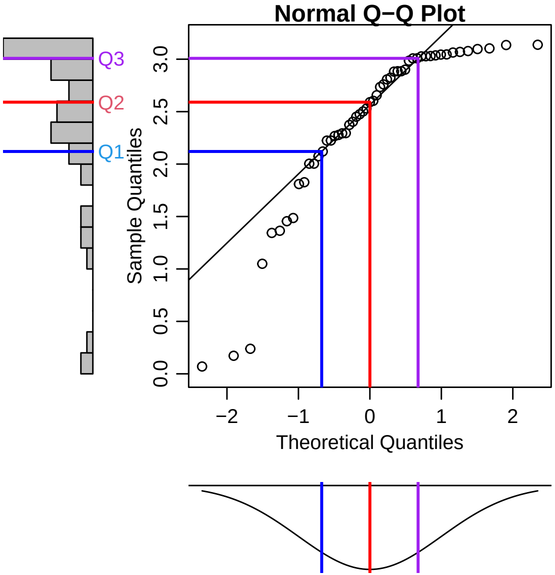
Normal QQ Plot — Normal Data



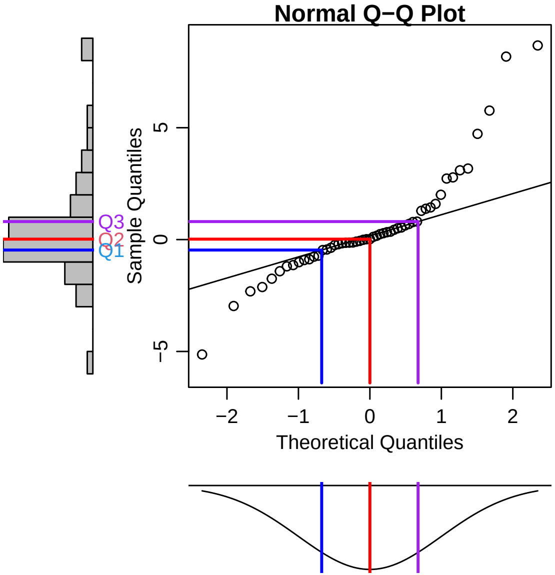
Normal QQ Plot — Right-Skewed Data



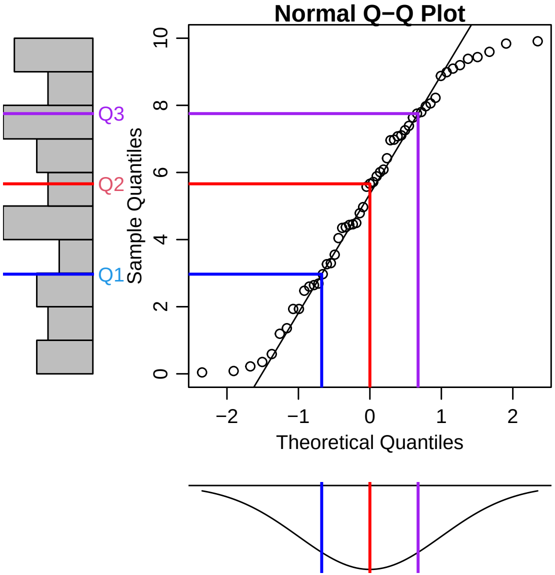
Normal QQ Plot — Left-Skewed Data



Normal QQ Plot — Heavy-Tailed Data



Normal QQ Plot — Light-Tailed Data

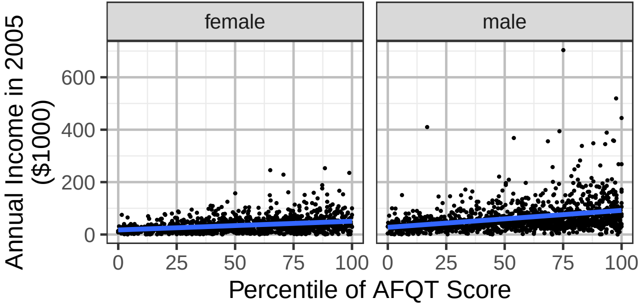


Example: NLSY Data in HW1

Subjects in the National Longitudinal Study of Youth (NLSY) data by U.S. Bureau of Labor Statistics `https://www.bls.gov/nls/` are 1306 American men and 1278 American women aged 14-22 in 1979. The variables include

- `Gender`
- `AFQT`: the percentile scores on the Armed Forces Qualifying Test, which is designed for evaluating the suitability of military recruits but which is also used by researchers as a general intelligence test
- `Income2005`: annual income in thousands of dollars in 2005

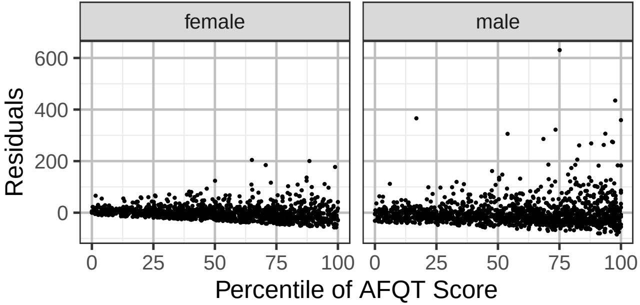
```r
NLSY = read.table(
  "http://www.stat.uchicago.edu/~yibi/s224/data/NLSY.txt",
  header=T)
library(ggplot2)
ggplot(NLSY, aes(x = AFQT, y = Income2005)) +
  geom_point(size = 0.2) +
  xlab("Percentile of AFQT Score") +
  ylab("Annual Income in 2005\n($1000)") +
  geom_smooth(method = 'lm') + facet_wrap(~Gender)
```



Residual plot of the MLR model

`lm1 = lm(Income2005 ~ Gender + AFQT, data=NLSY),`

```
lm1 = lm(Income2005 ~ Gender + AFQT, data=NLSY)
ggplot(NLSY, aes(x = AFQT, y = lm1$res)) +
  geom_point(size = 0.2) +
  xlab("Percentile of AFQT Score") +
  ylab("Residuals") + facet_wrap(~Gender)
```



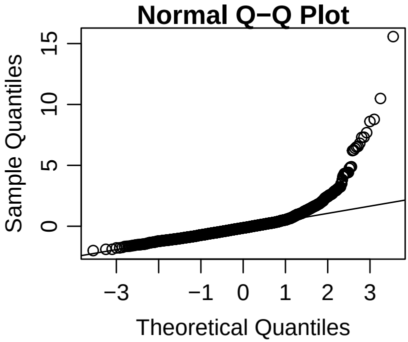
Normal QQ Plots in R

The R command `qqnorm()` can make normal QQ plots.

The `qqline()` command will add a straight line to the normal QQ plot to help gauging normality.

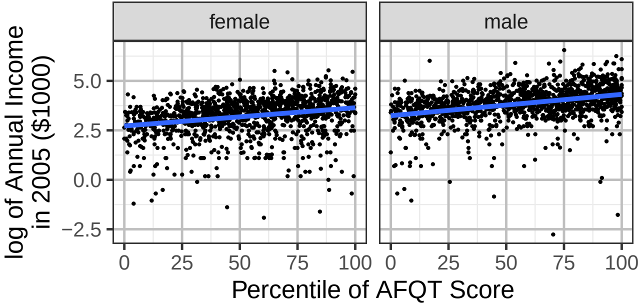
```
qqnorm(rstudent(lm1))
qqline(rstudent(lm1))
```

Are the residuals normal, right-skewed, or left-skewed?



After Taking Log of Income2005 ...

```
ggplot(NLSY, aes(x = AFQT, y = log(Income2005))) +
  geom_point(size = 0.2) +
  xlab("Percentile of AFQT Score") +
  ylab("log of Annual Income\nin 2005 ($1000)") +
  geom_smooth(method = 'lm') + facet_wrap(~Gender)
```



Normal QQ plot of the residuals after taking log of `Income2005`

```
lm2 = lm(log(Income2005) ~ Gender + AFQT, data=NLSY)
qqnorm(rstudent(lm2))
qqline(rstudent(lm2))
```

Are the residuals normal, right-skewed, or left-skewed?