STAT021 Statistical Methods II

Lecture 3 Statistical Modeling

Lu Chen
Swarthmore College
9/11/2018

Review

- Data structure, variables and relationships
  - *Rows (observations) and columns (variables)*
  - *Response variable and explanatory variable*
  - *Association and causation*
- Distribution
- Normal distribution $X \sim N(\mu, \sigma)$
  - *68-95-99.7 rule* `dnorm()`, `pnorm()`, `qnorm()`
  - *Standard Normal distribution* $Z \sim N(0, 1)$
  - *Normal Q-Q plot* `qqnorm()`
- Population, sample, parameter, statistic
- Random sampling
- Sampling distribution
- Central Limit Theorem (CLT)

Outline

Statistical modeling

- Statistical model and its purposes
- Four-step process of statistical modeling
  - Choose
  - Fit
  - Assess
  - Use
- Data example
  - Two-sample t test
  - Four-step process

Statistical model and its purposes

> A **statistical model** is a class of mathematical model, which represents, often in considerably idealized form, the data-generating process.

Making predictions

- Predicting the probability of acceptance to medical school based on GPA.
- Predicting the price of a car based on its age, mileage, and model.

Understanding relationships

- Is there relationship between a certain gene and disease? Does this relationship differ between men and women?

Assessing differences

- How ages of first walking are different between an exercise group and a control group.

Statistical model

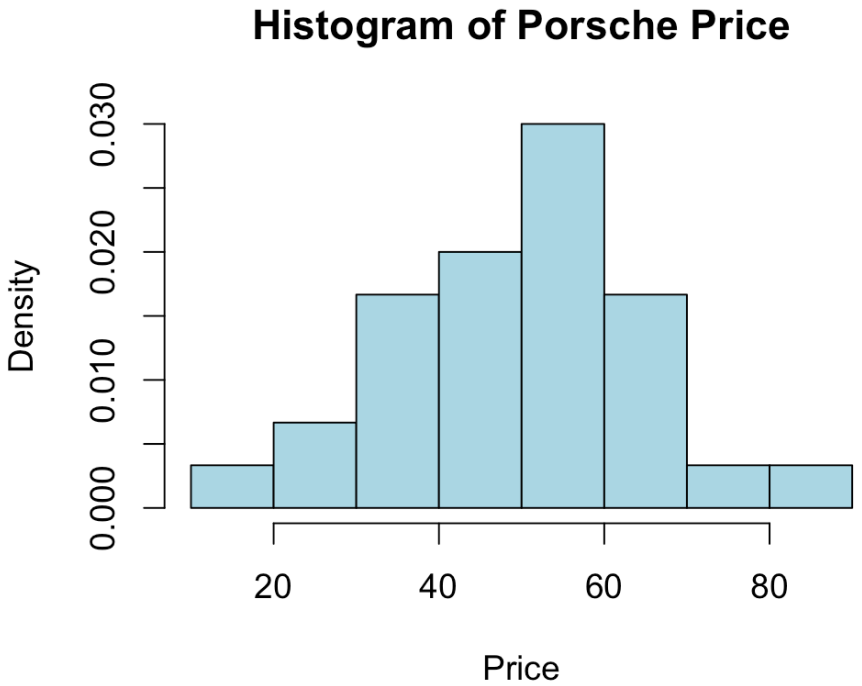
***Price*, *Age* and *Mileage* of used Porsche for sale** (in thousand dollars)

```
porsche <- read.table("../Datasets/PorschePrice.txt",sep="\t",header=T)
dim(porsche)
```

```
## [1] 30  3
```

```
head(porsche, 10)
```

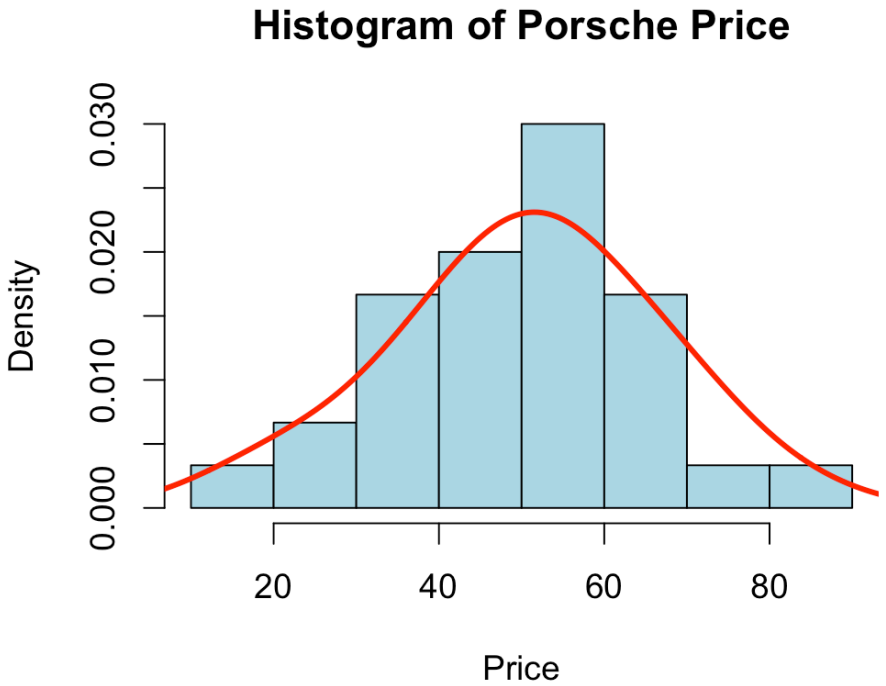
```
##    Price Age Mileage
## 1   69.4   3   21.50
## 2   56.9   3   43.00
## 3   49.9   2   19.90
## 4   47.4   4   36.00
## 5   42.9   4   44.00
## 6   36.9   6   49.80
## 7   83.0   0    1.30
## 8   72.9   0    0.67
## 9   69.9   2   13.40
## 10  67.9   0    9.70
```



Statistical model

> A **statistical model** is a class of mathematical model, which represents, often in **considerably idealized** form, the **data-generating process**.

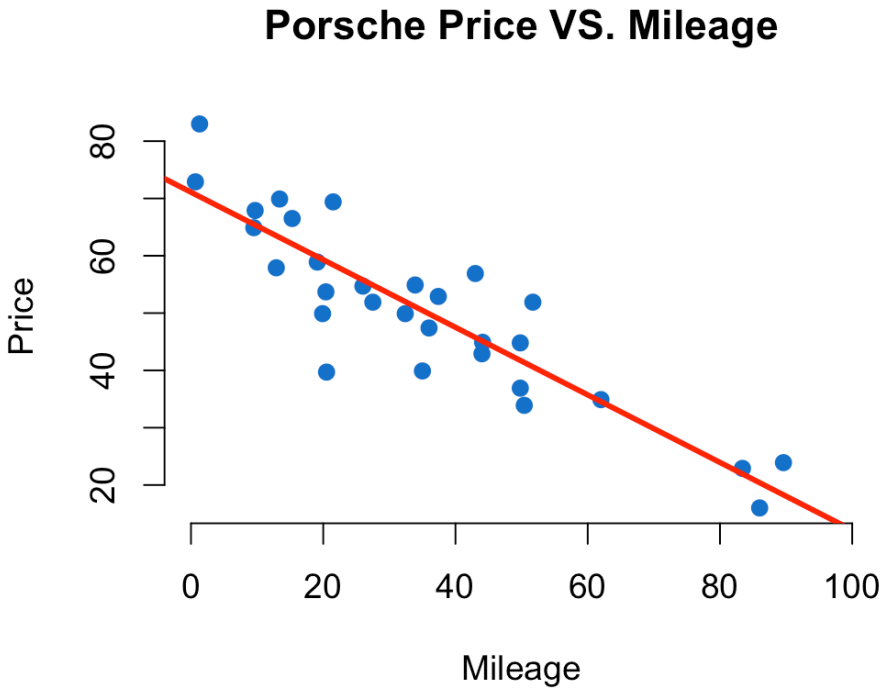
- "data-generating process"
  - $Price \overset{approx.}{\sim} N(50.5, 15.5)$
  - The values of *Price* follow an approximately Normal distribution with mean 50.5 thousand dollars and SD 15.5 thousand dollars.
- "considerably idealized"
  - $Price \overset{approx.}{\sim} N(50.5, 15.5)$is simplified and not exact.
  - How *Price* changes cannot be fully explained.



Statistical model

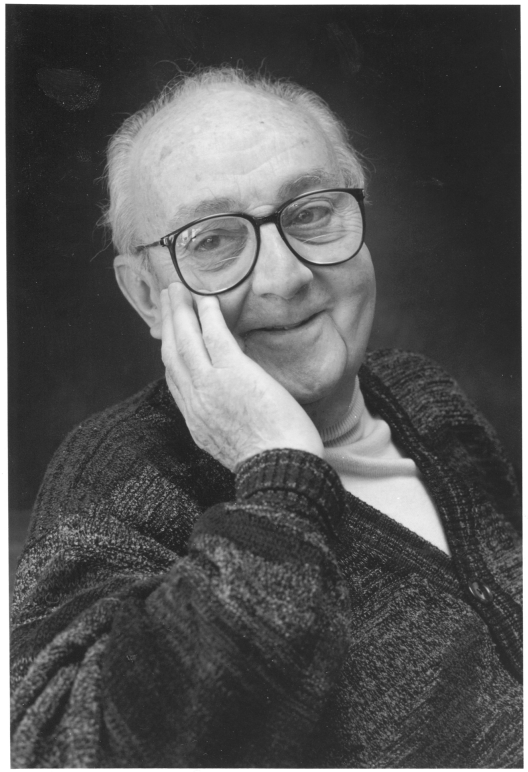
> A **statistical model** is a class of mathematical model, which represents, often in **considerably idealized** form, the **data-generating process**.

- "data-generating process"
  - $Price = 71.1 - 0.6 \times Mileage$
  - *Price* is 71.1 thousand dollars when *Mileage* is 0. As *Mileage* inceases 1000 miles, *Price* decreases 0.6 thousand dollars.
- "considerably idealized"
  - The regression model is still simplied and not exact.
  - How *Price* changes can only be explained partially.



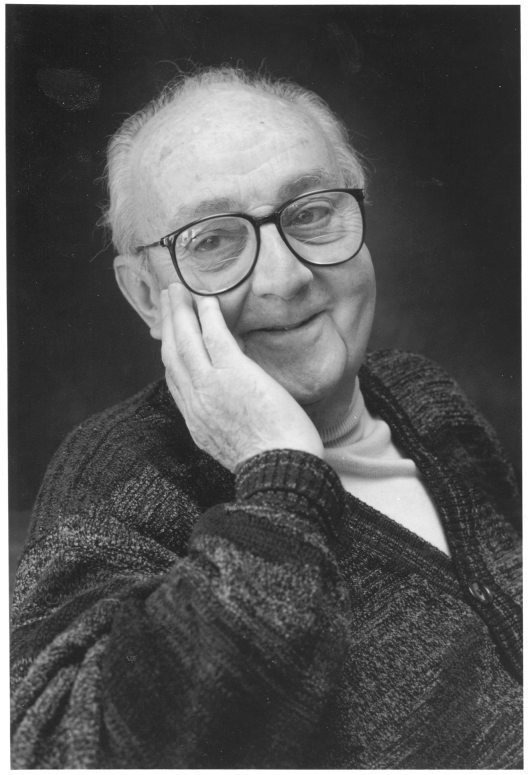
"All models are wrong"

> **"All models are wrong, but some are useful."**
>
> – George E. P. Box



- 10/18/1919 - 3/28/2013
- Director of the Statistical Research Group, Princeton University
- Established the Department of Statistics at University of Wisconsin-Madison (1960-1992)
- PhD in Statistics; supervised by Egon Pearson, son of Karl Pearson
- Married Joan Fisher, second daughter of Ronald Fisher

"All models are wrong"



George E. P. Box (1978)

- "Now it would be very remarkable if any system existing in the real world could be exactly represented by any simple model. However, cunningly chosen parsimonious models often do provide remarkably useful approximations."
- "For such a model there is no need to ask the question 'Is the model true?'. If 'truth' is to be the 'whole truth' the answer must be 'No'. The only question of interest is ' Is the model illuminating and useful'."

"All models are wrong"

When we do statistical modeling, we pay close attention to simplifications and imperfections, seeking to quantify how much the model explains and how much it does not.

We do DOT expect to

- predict the response variable exactly
- determine the exact relationship between two variables
- assess exactly how much two groups may differ

We want to

- predict and state how confident we are about our predictions
- examine the relationship and quantify how far off our model is likely to be
- evaluate how likely the two groups are to differ and by what magnitude

Lies, damn lies and statistics

Statistics lies?

Yes and **No**.

- **Yes**. All models are wrong. Statistical models never lead to truth.
- **No**, Statistics does not lie. **We fool ourselves** if we expect (exact) truth from statistics.

Statistical model

$$\begin{array}{ccccc} \text{Data} & = & \text{Model} & + & \text{Error} \\ Y & = & f(X) & + & \epsilon \end{array}$$

- Data, Y: the variable of interest - the response variable.
- Model, $f(X)$: how the explanatory variable(s) can be used to explain the response variable - function of the explanatory variable.
- Error, ϵ: part of the data that cannot be explained by the explanatory variable

In a statistical model, we specify

- the form of $f(X)$
- the distribution of ϵ

Four-step process of statistical modeling

- **CHOOSE**. Identify response and explanatory variables and their types, and do exploratory analysis.
- **FIT**. Estimate model parameters.
- **ASSESS**. Evaluate how well the model describes the data. Two components:
  - Assess model: compare the proposed model to a simpler model. Significance test; confidence interval.
  - Assess error: check the conditions/assumptions of the error term.
- **USE**.
  - Make predictions
  - Understand relationships
  - Assess differences
  - Discuss the limitations of the analysis

Four-step process of statistical modeling

Usually statistical modeling is carried out as:

- CHOOSE
- FIT
- ASSESS
- re-CHOOSE
- re-FIT
- re-ASSESS
- ...
- USE

Example: financial incentives for weight loss

Researchers investigated whether financial incentives would help people lose weight. Subjects in the treatment group were offered financial incentives for achieving weight loss goals, while the control group subjects did not use financial incentives. After four months the weight change ($Before - After$ in pounds) was recorded for each individual.

```
wl <- read.table("../Datasets/WeightLossIncentive4.txt",sep="\t",header=T)
dim(wl)
```

```
## [1] 36  2
```

```
head(wl)
```

```
##   WeightLoss   Group
## 1       12.5 Control
## 2       12.0 Control
## 3        1.0 Control
## 4       -5.0 Control
## 5        3.0 Control
## 6       -5.0 Control
```

Example: financial incentives for weight loss

```
wl$WeightLoss[wl$Group == "Control"]
```

```
##  [1]  12.5  12.0   1.0  -5.0   3.0  -5.0   7.5  -2.5  20.0  -1.0   2.0
## [12]   4.5  -2.0 -17.0  19.0  -2.0  12.0  10.5   5.0
```

```
wl$WeightLoss[wl$Group == "Incentive"]
```

```
##  [1] 25.5 24.0  8.0 15.5 21.0  4.5 30.0  7.5 10.0 18.0  5.0 -0.5 27.0
## [14]  6.0 25.5 21.0 18.5
```

- Significance test: two-sample t test
  1. Null and alternative hypothesis
  2. Test statistic
  3. P-value
  4. Conclusion
- Confidence interval: based on the two-sample t distribution

Example: financial incentives for weight loss

Population parameters and sample statistics

Group	Size	Population Mean	Population SD	Sample Mean	Sample SD
Control	$n_1 = 19$	μ_1	σ_1	$\bar{y}_1 = 3.9$	$s_1 = 9.1$
Incentive	$n_2 = 17$	μ_2	σ_2	$\bar{y}_2 = 15.7$	$s_2 = 9.4$

Two-sample t test: null and alternative hypotheses

- Null hypothesis $H_0 : \mu_1 = \mu_2$. The two groups have the same amount of weight loss.
- Alternative hypothesis $H_a : \mu_1 \neq \mu_2$. The *Control* and the *Incentive* group lost significantly different amount of weight.

Example: financial incentives for weight loss

Population parameters and sample statistics

Group	Size	Population Mean	Population SD	Sample Mean	Sample SD
Control	$n_1 = 19$	μ_1	σ_1	$\bar{y}_1 = 3.9$	$s_1 = 9.1$
Incentive	$n_2 = 17$	μ_2	σ_2	$\bar{y}_2 = 15.7$	$s_2 = 9.4$

Two-sample t test: test statistic, P-value and conclusion

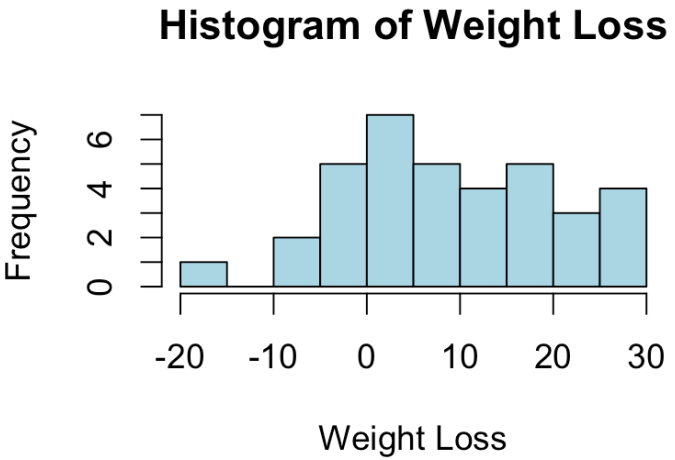
- $$\text{Test statistic } t = \frac{(\bar{y}_1 - \bar{y}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = -3.8 \sim t(33.3)$$

- P-value $= 2P(T \geq 3.8) = 0.0006 < 0.05$ for $T \sim t(33.3)$.
- Conclusion: The *Control* and the *Incentive* group lost significantly different amount of weight after 4 months at level $\alpha = 0.05$.

Four-step process of the example: CHOOSE

Exploratory data analysis: each variable

- Response variable: *WeightLoss*
  - Quantitative. Mean: 9.5; SD: 10.9.
- Explanatory variable: *Group*
  - Binary: *Control* ($n_1 = 19$) or *Incentive* ($n_2 = 17$)

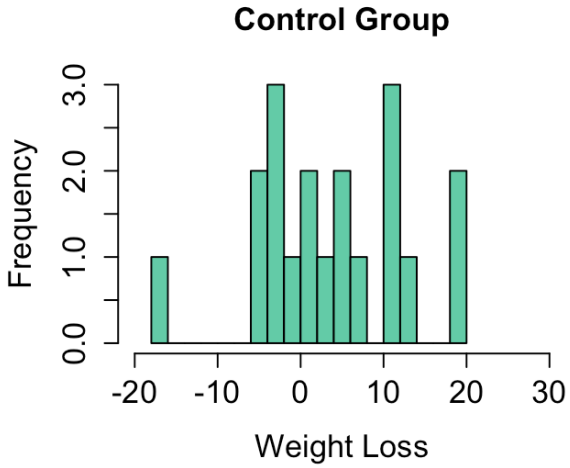


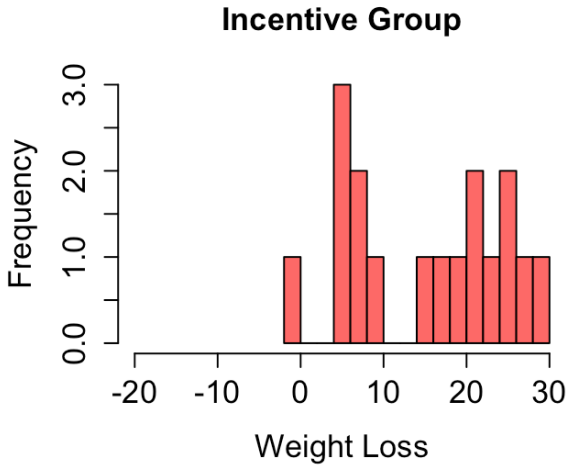
Exploratory data analysis: relationship between variables

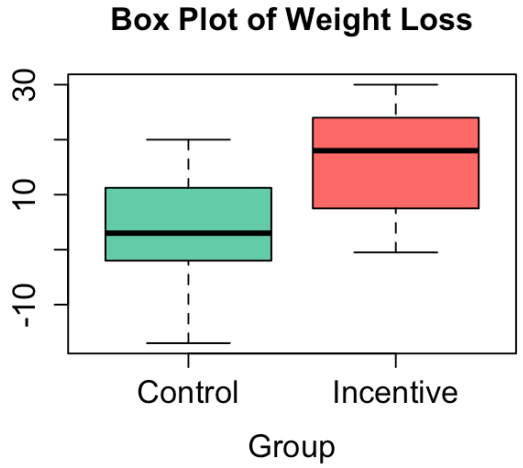
Group	Size	Sample Mean	Sample SD
Control	$n_1 = 19$	$\bar{y}_1 = 3.9$	$s_1 = 9.1$
Incentive	$n_2 = 17$	$\bar{y}_2 = 15.7$	$s_2 = 9.4$

Four-step process of the example: CHOOSE

Exploratory data analysis: relationship between variables







Assume weight losses of each group follow a Normal distribution

- Data = Model + Error
  - *Control* group: $Y = \mu_1 + \epsilon_1$, where $\epsilon_1 \sim N(0, \sigma_1) \Rightarrow Y \sim N(\mu_1, \sigma_1)$
  - *Incentive* group: $Y = \mu_2 + \epsilon_2$, where $\epsilon_2 \sim N(0, \sigma_2) \Rightarrow Y \sim N(\mu_2, \sigma_2)$

Four-step process of the example: FIT

- Population parameters assumed: μ_1 , μ_2 , σ_1 , σ_2
- In the **FIT** step, we need to estimate the parameters assumed in the model using the data sample.
- The sample statistics for population mean μ and population SD σ are sample mean $\bar{y}$ and sample SD s.
- Based on the table of summary statistics, we have

  $\bar{y}_1 = 3.9$ as the estimate to μ_1

  $\bar{y}_2 = 15.7$ as the estimate to μ_2

  $s_1 = 9.1$ as the estimate to σ_1

  $s_2 = 9.4$ as the estimate to σ_2

Four-step process of the example: ASSESS

ASSESS Model

Compare the **current model**

$$Y = \mu_1 + \epsilon_1, \text{ where } \epsilon_1 \sim N(0, \sigma_1) \text{ for the control group}$$
$$Y = \mu_2 + \epsilon_2, \text{ where } \epsilon_2 \sim N(0, \sigma_2) \text{ for the incentive group}$$

to a **simpler model**

$$Y = \mu + \epsilon, \text{ where } \epsilon \sim N(0, \sigma) \text{ for both groups}$$

to see whether it is necessary to assume different means for the two groups.

- This assessment is equivalent to a two-sample t test for whether $\mu_1 = \mu_2 = \mu$.

Four-step process of the example: ASSESS

```r
t.test(WeightLoss ~ Group, data=wl) # two-sample t test
```

```
##
##  Welch Two Sample t-test
##
## data:  WeightLoss by Group
## t = -3.7982, df = 33.276, p-value = 0.0005889
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -18.05026  -5.46058
## sample estimates:
##   mean in group Control mean in group Incentive
##                3.921053                15.676471
```

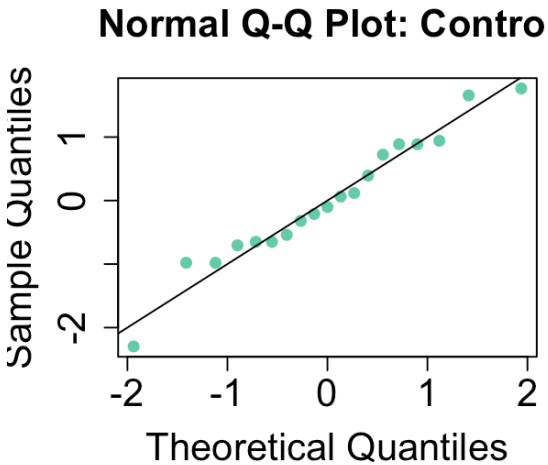
- Results are the same as previous calculations by hand.
- $t = -3.8, P = 0.00006 < 0.05$ The two groups are significantly different.
- Therefore, the current model that assumes different means of weight loss for the two groups is a better model.

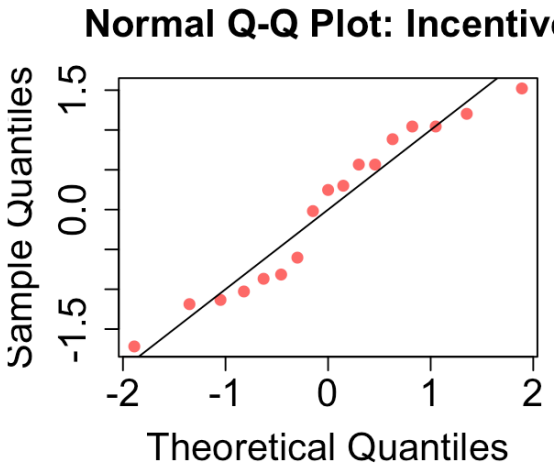
Four-step process of the example: ASSESS

ASSESS Error

$$Y = \mu_1 + \epsilon_1, \text{ where } \epsilon_1 \sim N(0, \sigma_1) \text{ for the control group}$$
$$Y = \mu_2 + \epsilon_2, \text{ where } \epsilon_2 \sim N(0, \sigma_2) \text{ for the incentive group}$$

- Denote e_1 and e_2 as the observed error for the two groups. Then $e_1 = y - \bar{y}_1$ and $e_2 = y - \bar{y}_2$. We need to check whether the error is Normally distributed.





- The Normal Q-Q plots show that *WeightLoss* of both groups is approximately Normally distributed.

Four-step process of the example: USE

- The financial incentives did produce a difference (11.8 pounds more) in the average weight loss over the four-month period.
- Since this was a designed experiment, we can infer a causal relationship between the treatment variable (control or incentive) and the weight change variable.

Be aware:

- 34 subjects were adult men, 2 subjects were adult women; the conclusion should be about adult men only.
- If not random, the conclusion cannot be extended to other adult men.
- Weight loss was measured at 4 months after the start of the study.
  Weight loss data measured at 7 months after the start of the study are available for analysis in Homework 2.

Summary: statistical modeling

- Statistical model

$$\begin{array}{ccccc} \text{Data} & = & \text{Model} & + & \text{Error} \\ Y & = & f(X) & + & \epsilon \end{array}$$

- Purposes of statistical modeling: *making predictions, understanding relationships, assessing differences.*
- Four-step process of statistical modeling
  - CHOOSE: *exploratory data analysis*
  - FIT: *estimating parameters*
  - ASSESS: *assessing model fitting and checking assumptions*
  - USE: *making predictions, understanding relationships, assessing differences, discussing limitations*

R codes

```r
# Scatterplot and regression line (slide 7)
m <- lm(Price ~ Mileage, data=porsche)
plot(x = porsche$Mileage, y = porsche$Price,
     main="Porsche Price VS. Mileage",
     xlab="Mileage", ylab="Price",
     col="dodgerblue3", pch=19, xlim=c(0,100))
abline(reg=m, lwd=3, col="red")

# Boxplot (slide 20)
boxplot(wl$WeightLoss~wl$Group, col=c("aquamarine3","indianred1"),
        xlab="Group", main="Box Plot of Weight Loss")
```