# AMS 597: Statistical Computing

Pei-Fen Kuan (c)

Applied Math and Stats, Stony Brook University

# One sample t-test

- The t-tests are based on an assumption that data come from the normal distribution
- In the one-sample case we assume that data $x_1, ..., x_n$ are normal random variables with mean $\mu$ and variance $\sigma^2$. We wish to test the null hypothesis that $H_0 : \mu = \mu_0$
- One can check for normality using the Shapiro Wilk test, implemented in `shapiro.test()` in R.

# One sample t-test

- The test statistics is

$$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$$

  under $H_0$
- The p-value is
  - $P(T_0 \geq t_0 | H_0)$ for $H_a : \mu > \mu_0$
  - $P(T_0 \leq t_0 | H_0)$ for $H_a : \mu < \mu_0$
  - $2P(T_0 \geq |t_0| | H_0)$ for $H_a : \mu \neq \mu_0$
- Consider an example concerning daily energy intake in kJ for 11 women (Altman, 1991, p. 183). First, the values are placed in a data vector.

# One sample t-test

```
daily.intake <- c(5260, 5470, 5640, 6180, 6390, 6515,
    6805, 7515, 8230, 8770)
mean(daily.intake)
```

```
## [1] 6677.5
```

```
sd(daily.intake)
```

```
## [1] 1174.11
```

```
quantile(daily.intake)
```

```
##     0%    25%    50%    75%   100%
## 5260.0 5775.0 6452.5 7337.5 8770.0
```

# One sample t-test

```r
shapiro.test(daily.intake)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  daily.intake
## W = 0.93468, p-value = 0.4955
```

# One sample t-test

```
res <- t.test(daily.intake, mu = 7725)
names(res)
```

```
##  [1] "statistic"   "parameter"   "p.value"     "conf.int"
##  [6] "null.value"  "stderr"      "alternative" "method"
```

```
res$para
```

```
## df
##  9
```

```
res$conf.int
```

```
## [1] 5837.592 7517.408
## attr(,"conf.level")
## [1] 0.95
```

# One sample t-test

```r
res$statistic
```

```
##         t
## -2.821273
```

```r
res$p.value
```

```
## [1] 0.02000537
```

```r
res$method
```

```
## [1] "One Sample t-test"
```

# One sample t-test

- Exercise: Can you write your own one sample t-test function for a two-sided alternative hypothesis? Your function will return the test statistic and p-value.

# Wilcoxon signed-rank test

- The t tests are fairly robust against departures from the normal distribution especially in larger samples, but sometimes you wish to avoid making that assumption. To this end, the distribution-free methods are convenient.
- For the one-sample Wilcoxon test, the procedure is to subtract the theoretical $\mu_0$ and rank the differences according to their numerical value, ignoring the sign, and then calculate the sum of the positive or negative ranks.

# Wilcoxon signed-rank test

- The point is that, assuming only that the distribution is symmetric around $\mu_0$, the test statistic corresponds to selecting each number from 1 to n with probability 1/2 and calculating the sum.
- The distribution of the test statistic can be calculated exactly, at least in principle. It becomes computationally excessive in large samples, but the distribution is then very well approximated by a normal distribution.

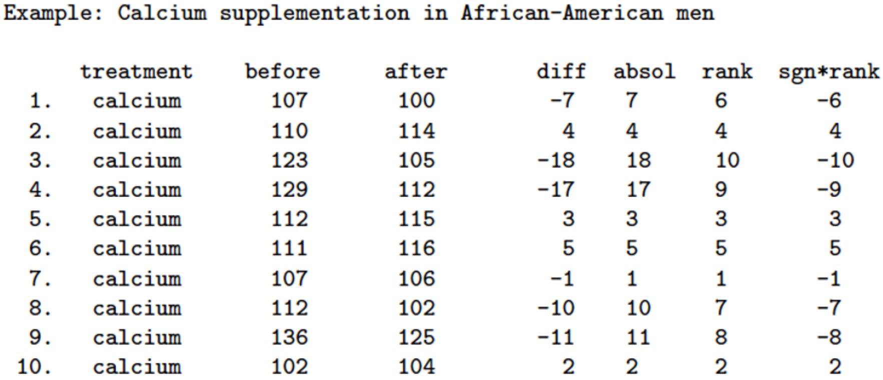
# Wilcoxon signed-rank test

- Suppose $Y_1, Y_2, \ldots, Y_n$ iid according to a symmetric distribution $F$ with median $\tilde{\mu}$
- Hypotheses $H_0 : \tilde{\mu} = \tilde{\mu}_0$ vs $H_a : \tilde{\mu} > \tilde{\mu}_0$

# Wilcoxon signed-rank test

- Delete $Y_i$'s equal $\tilde{\mu}_0$, adjust $n$
- Compute $Y_i' = Y_i - \tilde{\mu}_0$
- Rank $|Y_i'|$'s from smallest to largest
- The statistic $S^+$ is the sum of ranks from observation with $Y_i'$ positive
- $S^-$ defined similarly

# Wilcoxon signed-rank test



Example: Calcium supplementation in African-American men

| | treatment | before | after | diff | absol | rank | sgn*rank |
|---|---|---|---|---|---|---|---|
| 1. | calcium | 107 | 100 | -7 | 7 | 6 | -6 |
| 2. | calcium | 110 | 114 | 4 | 4 | 4 | 4 |
| 3. | calcium | 123 | 105 | -18 | 18 | 10 | -10 |
| 4. | calcium | 129 | 112 | -17 | 17 | 9 | -9 |
| 5. | calcium | 112 | 115 | 3 | 3 | 3 | 3 |
| 6. | calcium | 111 | 116 | 5 | 5 | 5 | 5 |
| 7. | calcium | 107 | 106 | -1 | 1 | 1 | -1 |
| 8. | calcium | 112 | 102 | -10 | 10 | 7 | -7 |
| 9. | calcium | 136 | 125 | -11 | 11 | 8 | -8 |
| 10. | calcium | 102 | 104 | 2 | 2 | 2 | 2 |

# Wilcoxon signed-rank test

- $S^+ = 4 + 3 + 5 + 2 = 14$; $S^- = 41$

```
x <- c(-7, 4, -18, -17, 3, 5, -1, -10, -11, 2)
wilcox.test(x)
```

```
##
##  Wilcoxon signed rank exact test
##
## data:  x
## V = 14, p-value = 0.1934
## alternative hypothesis: true location is not equal to 0
```

# Wilcoxon signed-rank test

- Calculating the null distribution for $n = 4$; an x in the column indicates that the sign of the rank is positive

| 1 | 2 | 3 | 4 | $S_0^+$ |
|---|---|---|---|---|
| | | | | 0 |
| x | | | | 1 |
| | x | | | 2 |
| | | x | | 3 |
| | | | x | 4 |
| x | x | | | 3 |
| x | | x | | 4 |
| x | | | x | 5 |
| | x | x | | 5 |
| | x | | x | 6 |
| | | x | x | 7 |
| x | x | x | | 6 |
| x | x | | x | 7 |
| x | | x | x | 8 |
| | x | x | x | 9 |
| x | x | x | x | 10 |

# Wilcoxon signed-rank test

- Exercise: Can you write your own exact Wilcoxon signed-rank test function for a two-sided alternative hypothesis? Your function will return the test statistic and p-value. You may define the two-sided p-value as $2\min(p_1, p_2)$ where $p_1 = \frac{\sum_{k=1}^{K} I(S_{0k}^{+} \geq S^{+})}{K}$ and $p_2 = \frac{\sum_{k=1}^{K} I(S_{0k}^{+} \leq S^{+})}{K}$

# Wilcoxon signed-rank test

- Large sample distribution
- Can show

$$E(S^+) = \frac{n(n+1)}{4} \text{ and } V(S^+) = \frac{n(n+1)(2n+1)}{24}$$

- If $n \geq 20$,

$$Z = \frac{S^+ - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}} \sim N(0,1)$$

# Wilcoxon signed-rank test

- If there are 2 or more observations with the same value of $Y'$, the observations are said to be tied
- For tied observations we assign the average rank or midrank
- Example: $\mathbf{Y} = \{23, 25, 45, 13, 23, 46\}$
- MidRanks: $\{2.5, 4, 5, 1, 2.5, 6\}$

# Wilcoxon signed-rank test

- Can show

$$E(S^+) = \frac{n(n+1)}{4}$$

- To accommodate ties, var is adjusted

$$V(S^+) = \frac{n(n+1)(2n+1) - \frac{1}{2}\sum_{i=1}^{q} t_i(t_i-1)(t_i+1)}{24}$$

where $q$ equals the number of sets of ties and $t_i$ is the number of observations in the $i$th set

- For example on previous slide, $q = 1$ and $t_1 = 2$ such that

$$V(S^+) = \frac{6(6+1)(2\cdot 6+1) - \frac{1}{2}\cdot 2\cdot 1\cdot 3}{24}$$

# Wilcoxon signed-rank test

```
res2 <- wilcox.test(daily.intake, mu = 7725)
names(res2)
```

```
## [1] "statistic"   "parameter"   "p.value"     "null.value"
## [6] "method"      "data.name"
```

```
res2
```

```
##
##  Wilcoxon signed rank exact test
##
## data:  daily.intake
## V = 6, p-value = 0.02734
## alternative hypothesis: true location is not equal to 7725
```

# Wilcoxon signed-rank test

```r
res3 <- wilcox.test(daily.intake, mu = 7725, exact = FALSE)
names(res3)
```

```
## [1] "statistic"   "parameter"   "p.value"     "null.value"
## [6] "method"      "data.name"
```

```r
res3
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  daily.intake
## V = 6, p-value = 0.03231
## alternative hypothesis: true location is not equal to 7725
```

# Two sample t-test

- The two-sample t test is used to test the hypothesis that two samples may be assumed to come from distributions with the same mean.
- The theory for the two-sample t test is not very different in principle from that of the one-sample test.
- Data are now from two groups, $x_1, x_2, ..., x_{n_1}$ and $y_1, y_2, ..., y_{n_2}$, which we assume are sampled from $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$

# Two sample t-test

- It is desired to test the null hypothesis

$$H_0 : \mu_1 - \mu_2 = c_0$$

- Equal variance assumption:

$$t_0 = \frac{(\bar{x} - \bar{y}) - c_0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2} \text{ under } H_0$$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

# Two sample t-test

- Unequal variance assumption:

$$t_0 = \frac{(\bar{x} - \bar{y}) - c_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t_{df} \text{ under } H_0$$

$$df = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

# Comparison of variances

- Suppose $H_0 : \sigma_1^2 = \sigma_2^2 \Leftrightarrow H_0 : \sigma_1^2/\sigma_2^2 = 1$
- Test statistic:

$$F_0 = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \sim F_{n_1-1,n_2-1} \text{ under } H_0$$

# Comparison of variances

- Data url: “http://www.ams.sunysb.edu/~pfkuan/Teaching/AMS597/Data/d_logret_6stocks.txt” , also on Brightspace.

```
ret <- read.table(paste0(dataPath, "d_logret_6stocks.txt"),
    header = T)
var.test(ret$Pfizer, ret$Intel)
```

```
##
##  F test to compare two variances
##
## data:  ret$Pfizer and ret$Intel
## F = 0.11577, num df = 63, denom df = 63, p-value = 3.703e-1
## alternative hypothesis: true ratio of variances is not equa
## 95 percent confidence interval:
##  0.07033263 0.19055829
## sample estimates:
## ratio of variances
##           0.115769
```

# Two sample t-test

```
t.test(x, y = NULL, alternative = c("two.sided", "less",
"greater"), mu = 0, paired = FALSE, var.equal = FALSE,
conf.level = 0.95, ...)
```

# Two sample t-test

```r
res3 <- t.test(ret$Pfizer, ret$Intel)
names(res3)
```

```
##  [1] "statistic"   "parameter"   "p.value"     "conf.int"
##  [6] "null.value"  "stderr"      "alternative" "method"
```

```r
res3$stat
```

```
##         t
## 0.2170671
```

# Two sample t-test

```r
t.test(ret$Pfizer, ret$Intel)
```

```
##
##  Welch Two Sample t-test
##
## data:  ret$Pfizer and ret$Intel
## t = 0.21707, df = 77.394, p-value = 0.8287
## alternative hypothesis: true difference in means is not equ
## 95 percent confidence interval:
##  -0.01588991  0.01977844
## sample estimates:
##    mean of x    mean of y
## -0.004041315 -0.005985579
```

# Exercise

- Perform for ‘Citigroup’ one sample test with the null hypothesis that the mean is zero
- Perform the Wilcoxon signed-rank test for ‘Citigroup’
- Perform the two-sample test for ‘Pfizer’ and ‘Citigroup’

# Wilcoxon Rank Sum test

- Also known as Mann-Whitney test
- Assume $Y_{1j}, \ldots, Y_{n_j j}$ iid $F_j(y)$; $j = 1, 2$

$$H_0 : F_1(y) = F_2(y)$$

$$H_a : F_1(y) = F_2(y + \Delta)$$

  where $\Delta$ is a constant
- Pool the two samples
- Rank them from smallest to largest
- Compute the sum of the ranks, $W_1$, in group 1

# Wilcoxon Rank Sum test

- There are $N = n_1 + n_2$ subjects in our study
- Thus there are $\binom{N}{n_1}$ possible outcomes
- Under $H_0$, each is equally likely
- We compute the distribution of $W_1$ by enumeration

# Wilcoxon Rank Sum test

- A new drug is being test in humans for the first time to assess effect on CD4+ T cells in patients with HIV
- 7 individuals are randomized to 2 groups: control ($n_1 = 3$) or drug ($n_2 = 4$)
- Endpoint is percent change in CD4+ count from baseline
- Null hypothesis is the drug has no effect

$$H_0 : \Delta = 0; H_a : \Delta \neq 0$$

- Data: control (65, 73, 69); drug (89, 70, 92, 88)
- There are $\binom{7}{3} = 35$ possible outcomes of the study, i.e. there are 35 possible rankings for group 1

# Wilcoxon Rank Sum test

| Ranks | $W_1$ | Ranks | $W_1$ | Ranks | $W_1$ |
|---|---|---|---|---|---|
| 1,2,3 | 6 | 1,5,6 | 12 | 2,6,7 | 15 |
| 1,2,4 | 7 | 1,5,7 | 13 | 3,4,5 | 12 |
| 1,2,5 | 8 | 1,6,7 | 14 | 3,4,6 | 13 |
| 1,2,6 | 9 | 2,3,4 | 9 | 3,4,7 | 14 |
| 1,2,7 | 10 | 2,3,5 | 10 | 3,5,6 | 14 |
| 1,3,4 | 8 | 2,3,6 | 11 | 3,5,7 | 15 |
| 1,3,5 | 9 | 2,3,7 | 12 | 3,6,7 | 16 |
| 1,3,6 | 10 | 2,4,5 | 11 | 4,5,6 | 15 |
| 1,3,7 | 11 | 2,4,6 | 12 | 4,5,7 | 16 |
| 1,4,5 | 10 | 2,4,7 | 13 | 4,6,7 | 17 |
| 1,4,6 | 11 | 2,5,6 | 13 | 5,6,7 | 18 |
| 1,4,7 | 12 | 2,5,7 | 14 | | |

# Wilcoxon Rank Sum test

- It can be shown that

$$E(W_1) = \frac{n_1}{N}\frac{N(N+1)}{2} = \frac{n_1(N+1)}{2}$$

$$V(W_1) = \frac{n_1 n_2(N+1)}{12}$$

# Wilcoxon Rank Sum test

- If $n_1$ and $n_2$ are large

$$Z = \frac{W_1 - E(W_1)}{\sqrt{V(W_1)}}$$

  will be approx $N(0,1)$

- Approximation is good for $n_1, n_2 \geq 12$
- If there are ties

$$V(W_1) = \frac{n_1 n_2 (N+1)}{12} - \frac{n_1 n_2}{12N(N-1)} \sum_{i=1}^{q} t_i(t_i - 1)(t_i + 1)$$

# Wilcoxon Rank Sum test

```r
wilcox.test(ret$Pfizer, ret$Intel, exact = TRUE)
```

```
##
##  Wilcoxon rank sum exact test
##
## data:  ret$Pfizer and ret$Intel
## W = 2019, p-value = 0.8923
## alternative hypothesis: true location shift is not equal to
```

# Wilcoxon Rank Sum test

```
wilcox.test(ret$Pfizer, ret$Intel)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  ret$Pfizer and ret$Intel
## W = 2019, p-value = 0.892
## alternative hypothesis: true location shift is not equal to
```

# Simple linear regression

- The linear regression model is given by $y_i = \alpha + \beta x_i + \epsilon_i$ in which $\epsilon_i$ are assumed independent and $N(0, \sigma^2)$
- The parameters $\alpha$, $\beta$ and $\sigma^2$ can be estimated using the method of least squares.
- In particular, the values of $\alpha$ and $\beta$ can be obtained by minimizing the sum of squared residuals, and $\sigma^2$ can be estimated via the sum of squared residuals. This will be studied in details later in this course.

# Simple linear regression

- It is usually of prime interest to test the null hypothesis $\beta = 0$ for which we can use a t-test

```
fit1 <- lm(ret$Pfizer ~ ret$Intel)
```

# Simple linear regression

```
summary(fit1)
```

```
##
## Call:
## lm(formula = ret$Pfizer ~ ret$Intel)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.055920 -0.013845  0.000851  0.017246  0.045693
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.003903   0.002913  -1.340    0.185
## ret$Intel    0.023078   0.043112   0.535    0.594
##
## Residual standard error: 0.02321 on 62 degrees of freedom
## Multiple R-squared:  0.0046, Adjusted R-squared:  -0.01145
## F-statistic: 0.2865 on 1 and 62 DF,  p-value: 0.5944
```

# Simple linear regression

```r
names(fit1)
```

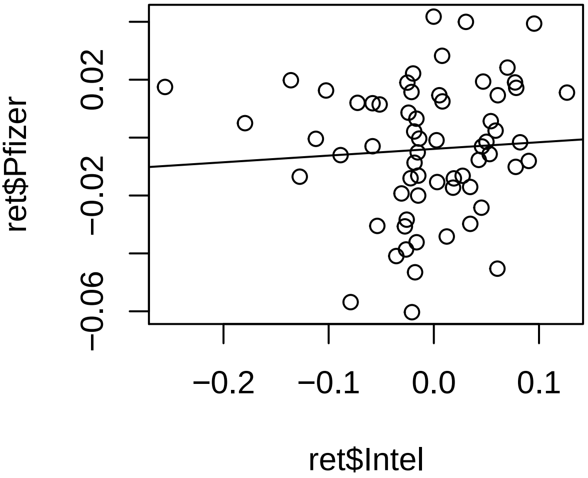
```
##  [1] "coefficients"  "residuals"     "effects"       "rank'
##  [5] "fitted.values" "assign"        "qr"            "df.re
##  [9] "xlevels"       "call"          "terms"         "model
```

```r
fit1$coeff
```

```
## (Intercept)   ret$Intel
## -0.00390318  0.02307791
```

```r
plot(ret$Intel, ret$Pfizer)
abline(lm(ret$Pfizer ~ ret$Intel))
```

# Simple linear regression



# Simple linear regression

```r
# regression without intercept
fit2 <- lm(ret$Pfizer ~ -1 + ret$Intel)
summary(fit2)
```

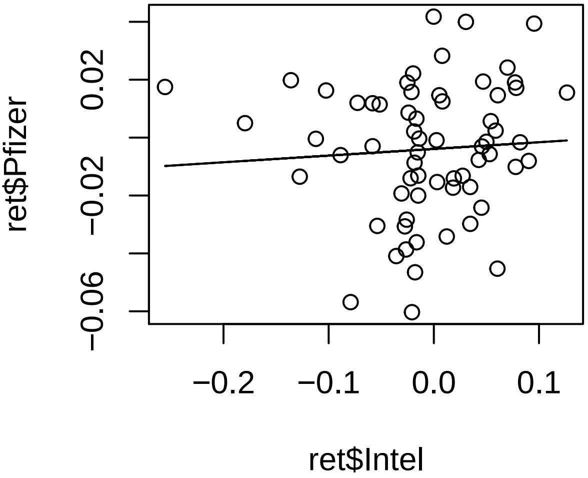
```
##
## Call:
## lm(formula = ret$Pfizer ~ -1 + ret$Intel)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.059716 -0.017863 -0.003199  0.013654  0.041790
##
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## ret$Intel  0.02819    0.04321   0.652    0.516
##
## Residual standard error: 0.02336 on 63 degrees of freedom
## Multiple R-squared:  0.006712,   Adjusted R-squared:  -0.00
```

# Residuals and fitted values

- We have seen how summary can be used to extract information about the results of a regression analysis. Two further extraction functions are `fitted` and `resid`

```
fitted(fit1)[1:10]
resid(fit1)[1:10]
plot(ret$Intel, ret$Pfizer)
lines(ret$Intel, fitted(fit1))
```

# Residuals and fitted values



# Residuals and fitted values

- To visualize the residual plots, you may type `plot(fit1)`
- The plot of fitted values vs residuals is usually used for checking constant variance and linearity assumptions
- QQplot on residuals can be used to check for normality assumption

# Prediction and confidence bands

- Fitted lines are often presented with uncertainty bands around them. There are two kinds of bands, often referred to as the “narrow” and “wide” limits.
- The narrow bands, confidence bands, reflect the uncertainty about the line itself. The wide bands, prediction bands, include the uncertainty about future observations.

# Prediction and confidence bands

- Predicted values, with or without prediction and confidence bands, may be extracted with the function predict. With no arguments, it just gives the fitted values:

```
predict(fit1)
```

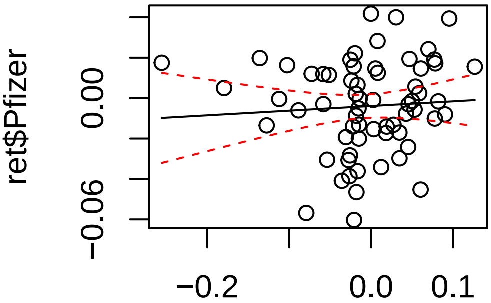
- If you add `interval="confidence"` or `interval="prediction"`, then you get the vector of predicted values augmented with limits.

```
predict(fit1, interval = "confidence", level = 0.95)
predict(fit1, interval = "prediction", level = 0.95)
```

# Prediction and confidence bands

- The best way to add prediction and confidence intervals to a scatterplot is to use the `matlines` function, which plots the columns of a matrix against a vector.

```
plot(ret$Intel, ret$Pfizer)
pp <- predict(fit1, int = "c")
matlines(sort(ret$Intel), pp[order(ret$Intel), ], lty = c(1,
    2, 2), col = c("black", "red", "red"))
```



# Correlation

- The function cor() can be used to compute the correlation between two or more vectors.
- Pearson correlation coefficient

$$r = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i (Y_i - \bar{Y})^2 \sum_i (X_i - \bar{X})^2}}$$

- cor.test() can be used to perform hypothesis test on correlation

```
cor(ret$Intel, ret$Pfizer)
```

```
## [1] 0.06782663
```

```
cor.test(ret$Intel, ret$Pfizer)
```

```
##
##  Pearson's product-moment correlation
##
## data:  ret$Intel and ret$Pfizer
## t = 0.5353, df = 62, p-value = 0.5944
```

# Correlation

```r
Intel1 <- ret$Intel
Intel1[1] <- NA
cor(Intel1, ret$Pfizer)
```

```
## [1] NA
```

```r
cor(Intel1, ret$Pfizer, use = "complete.obs")
```

```
## [1] 0.06670035
```

# Correlation

- A non-parametric measure of correlation is Spearman rank correlation
- Let $R_{1i}$ and $R_{2i}$ be the ranks of the $Y_i$ and $X_i$, respectively
- Spearman correlation coefficient

$$r_s = \frac{\sum(R_{1i} - \bar{R}_1)(R_{2i} - \bar{R}_2)}{\sqrt{\sum_i(R_{1i} - \bar{R}_1)^2 \sum_i(R_{2i} - \bar{R}_2)^2}}$$

$$= 1 - \frac{6\sum d_i^2}{N^3 - N}$$

where $d_i = R_{1i} - R_{2i}$

# Correlation

```r
cor(ret$Intel, ret$Pfizer, method = "spearman")
```

```
## [1] 0.1315476
```

```r
cor.test(ret$Intel, ret$Pfizer, method = "spearman")
```

```
##
##  Spearman's rank correlation rho
##
## data:  ret$Intel and ret$Pfizer
## S = 37934, p-value = 0.2993
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.1315476
```

# Correlation

- The Spearman correlation coefficient can be used to test the null hypothesis of independence $H_0 : X \perp Y$ vs. $H_a : X \not\perp Y$ that is, $H_a : X$ and $Y$ not independent
- Distribution of $r_s$ under $H_0$ is derived using a permutation-based argument
- We can list the $R_{1i}$ in ascending order
- There are $N!$ possible orderings of the $R_{2i}$
- Under $H_0$, each of these orderings is equally likely

# Correlation

- Example: $N = 3$

| $R_{1i}$ | 1 | 2 | 3 | $\sum d_i^2$ | $r_s$ |
|---|---|---|---|---|---|
| $R_{2i}$ | 1 | 2 | 3 | 0 | 1.0 |
| $R_{2i}$ | 1 | 3 | 2 | 2 | 0.5 |
| $R_{2i}$ | 2 | 1 | 3 | 2 | 0.5 |
| $R_{2i}$ | 2 | 3 | 1 | 6 | −0.5 |
| $R_{2i}$ | 3 | 1 | 2 | 6 | −0.5 |
| $R_{2i}$ | 3 | 2 | 1 | 8 | −1.0 |

# Correlation

- Exercise: Write your own function to compute the Spearman rank correlation between two variables

# Correlation

- Kendall's $\tau$: Another rank correlation statistic
- Data: $(X_i, Y_i)$ for $i = 1, 2, \ldots, N$
- Definitions: Two pairs of observations are
  - concordant if $(X_i - X_j)(Y_i - Y_j) > 0$
  - discordant if $(X_i - X_j)(Y_i - Y_j) < 0$
- Let $p_c$ be the probability that a randomly chosen pair of observations is concordant; and $p_d$ the probability that they are discordant; then

$$\tau = p_c - p_d$$

# Correlation

- There are $\binom{N}{2}$ pairs of observations
- Let $P$ be the number of concordant pairs
- Let $Q$ be the number of discordant pairs
- The estimate of $\tau$ is
$$r_k = \frac{P - Q}{\binom{N}{2}}$$
- The distribution of $r_k$ under $H_0$ is computed using permutation principles
- As with $r_s$, there are $N!$ equally likely outcomes

# Correlation

```r
cor(ret$Intel, ret$Pfizer, method = "kendall")
```

```
## [1] 0.0922619
```

```r
cor.test(ret$Intel, ret$Pfizer, method = "kendall")
```

```
##
##  Kendall's rank correlation tau
##
## data:  ret$Intel and ret$Pfizer
## z = 1.0776, p-value = 0.2812
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##       tau
## 0.0922619
```