Tennis: Comebacks in Best-Of-5 format.

Certain tournaments in Association of Tennis Professionals (ATP) are held according to a Best-of-5 format: one needs to win three sets to obtain victory.

Tennis history is filled with examples of players completing epic comebacks from two sets down to win in 5 sets (0-2 => 3-2).

Assume that in a match of Player A vs Player B:
- Player A loses sets 1 & 2, while
- Player B loses sets 3 & 4,

**Question:** what is \( p = P(\text{Player A wins 5th set}) \)?

Below we compile data on all ATP tennis matches into one data frame.

```r
setwd("/home/usdandres/Documents/Study_stuff/Sports Research/Tennis/tennis_atp-master")
years <- c(1968:2017)
n.years <- length(years)
all.matches <- NULL
for (i in 1:n.years){
  all.matches <- rbind(all.matches,read.csv(paste("atp_matches_",years[i],".csv",sep="")))
}
nrow(all.matches)
head(all.matches,4)
```

<table>
<thead>
<tr>
<th>tourney_id</th>
<th>tourney_name</th>
<th>surface</th>
<th>draw_size</th>
<th>tourney_level</th>
<th>tourney_date</th>
<th>match_num</th>
<th>winner_id</th>
</tr>
</thead>
<tbody>
<tr>
<td>1968-580</td>
<td>Australian Chps.</td>
<td>Grass</td>
<td>64</td>
<td>G</td>
<td>19680119</td>
<td>1</td>
<td>1002:</td>
</tr>
<tr>
<td>1968-580</td>
<td>Australian Chps.</td>
<td>Grass</td>
<td>64</td>
<td>G</td>
<td>19680119</td>
<td>2</td>
<td>10980:</td>
</tr>
<tr>
<td>1968-580</td>
<td>Australian Chps.</td>
<td>Grass</td>
<td>64</td>
<td>G</td>
<td>19680119</td>
<td>3</td>
<td>10025:</td>
</tr>
<tr>
<td>1968-580</td>
<td>Australian Chps.</td>
<td>Grass</td>
<td>64</td>
<td>G</td>
<td>19680119</td>
<td>4</td>
<td>10010:</td>
</tr>
</tbody>
</table>

Too many features, selecting just a few interesting ones, and getting rid of some 'NA' and other irrelevant rows.
In [16]:
```r
data.5sets <- subset(all.matches, str_count(as.character(score), "\-\") == 5)
head(data.5sets, 4)
```

<table>
<thead>
<tr>
<th>winner_name</th>
<th>winner_rank</th>
<th>loser_name</th>
<th>loser_rank</th>
<th>score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vijay Amritraj</td>
<td>22</td>
<td>Raul Ramirez</td>
<td>32</td>
<td>5-7 6-4 2-6 6-3 8-6</td>
</tr>
<tr>
<td>Jiri Hrebec</td>
<td>43</td>
<td>Jan Kodes</td>
<td>8</td>
<td>4-6 6-1 3-6 6-0 7-5</td>
</tr>
<tr>
<td>Rod Laver</td>
<td>10</td>
<td>John Newcombe</td>
<td>3</td>
<td>3-6 7-5 6-3 3-6 6-4</td>
</tr>
<tr>
<td>Roger Taylor</td>
<td>13</td>
<td>Antonio Munoz</td>
<td>81</td>
<td>3-6 6-4 6-1 3-6 6-3</td>
</tr>
</tbody>
</table>

Data Wrangling, Task #2:

How to get the 5-set matches RELEVANT to our question?

Hint: Need matches where sets 1 & 2 are won by the SAME PLAYER.

--
In [18]:

```r
# Now we need to record amounts of matches with the following distinct score dynamics (Best-of-5 sets):

- 0-2 => 2-2 => 3-2 (the player COMPLETES the comeback from 0-2 down)
- 0-2 => 2-2 => 2-3 (the player TIES it at 2-2, but LOSES the 5th set)
```

```
In [19]:

### Extracting scores of sets 1 & 2

```r
data.5sets$FirstSet <- unlist(lapply(strsplit(as.character(data.5sets$score),"\s+"),unlist))
data.5sets$SecondSet <- unlist(lapply(strsplit(as.character(data.5sets$score),"\s+"),unlist))
```

```r
head(data.5sets,4)
```

<table>
<thead>
<tr>
<th>winner_name</th>
<th>winner_rank</th>
<th>loser_name</th>
<th>loser_rank</th>
<th>score</th>
<th>FirstSet</th>
<th>SecondSet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vijay Amritraj</td>
<td>22</td>
<td>Raul Ramirez</td>
<td>32</td>
<td>5-7 6-4 2-6 6-3 8-6</td>
<td>5-7</td>
<td>6-4</td>
</tr>
<tr>
<td>Jiri Hrebec</td>
<td>43</td>
<td>Jan Kodes</td>
<td>8</td>
<td>4-6 6-1 3-6 6-0 7-5</td>
<td>4-6</td>
<td>6-1</td>
</tr>
<tr>
<td>Rod Laver</td>
<td>10</td>
<td>John Newcombe</td>
<td>2</td>
<td>3-6 7-5 6-3 3-6 6-4</td>
<td>3-6</td>
<td>7-5</td>
</tr>
<tr>
<td>Roger Taylor</td>
<td>13</td>
<td>Antonio Munoz</td>
<td>81</td>
<td>3-6 6-4 6-1 3-6 6-3</td>
<td>3-6</td>
<td>6-4</td>
</tr>
</tbody>
</table>
```

Now we need to record amounts of matches with the following distinct score dynamics (Best-of-5 sets):

- 0-2 => 2-2 => 3-2 (the player COMPLETES the comeback from 0-2 down)
- 0-2 => 2-2 => 2-3 (the player TIES it at 2-2, but LOSES the 5th set)

```
In [19]:

### First case: COMEBACK KID wins.

```r
result.02.32 <- subset(data.5sets,
    (substr(FirstSet,1,1)<substr(FirstSet,3,3)) &
    substr(SecondSet,1,1)<substr(SecondSet,3,3))
paste("From 0-2 to 3-2: ",nrow(result.02.32)," matches")

### Second case: LEAD-BLOWER wins.

```r
result.02.23 <- subset(data.5sets,
    (substr(FirstSet,1,1)>substr(FirstSet,3,3)) &
    substr(SecondSet,1,1)>substr(SecondSet,3,3))
paste("From 0-2 to 2-3: ",nrow(result.02.23)," matches")
```

'From 0-2 to 3-2: 1027 matches'

'From 0-2 to 2-3: 783 matches'

We have the following statistical setup:

- **Sample:** All matches with 0-2 => 2-2 dynamic,
- **Binary Outcome:** Each match finishes either 3-2 or 2-3.
- **Parameter of Interest:** $p = P('3-2')$
Presume we'd like to account for difference in player level (maybe experience + skill helps them in the 5th set) =>

- Create a variable $X$ characterizing difference in level
- Perform logistic regression for binary outcome $Y$ ($= 1$ if "0-2 => 3-2", $= 0$ if "0-2 => 2-3")

$$\text{logit}(p_i) = \alpha + \beta X_i, \quad p_i = P(Y_i = 1)$$

```
In [20]:
n.02.32 <- nrow(result.02.32)  # number of 0-2 => 3-2
n.02.23 <- nrow(result.02.23)  # number of 0-2 => 2-3
n.total <- n.02.32+n.02.23
prop.test(n.02.32,n.total,0.5)
```

1-sample proportions test with continuity correction

data:  n.02.32 out of n.total, null probability 0.5
X-squared = 32.624, df = 1, p-value = 1.118e-08
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
0.5441827 0.5903350
sample estimates:
p
0.5674033

Presume we'd like to account for difference in player level (maybe experience + skill helps them in the 5th set) =>

1-sample proportions test with continuity correction

data:  n.02.32 out of n.total, null probability 0.5
X-squared = 32.624, df = 1, p-value = 1.118e-08
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
0.5441827 0.5903350
sample estimates:
p
0.5674033
### Creating an OUTCOME variable, =1 if "0-2 => 3-2", =0 if "0-2 => 2-3".

```r
my.table <- rbind(cbind(result.02.32, Won=1), cbind(result.02.23, Won=0))
```

### Creating a 'difference in level' variable.

```r
my.table$LogRankRatio <- ifelse(my.table$Won==1, log(my.table$winner_rank / log(my.table$loser_rank))
```

### Logistic Regression

```r
my.glm <- glm(Won ~ LogRankRatio, family=quasibinomial, data=my.table)
```

Call:
```
glm(formula = Won ~ LogRankRatio, family = quasibinomial, data = my.table)
```

Deviance Residuals:
```
    Min       1Q   Median       3Q      Max
-1.9321  -1.2192   0.8401   1.0448   1.6647
```

Coefficients:
```
             Estimate Std. Error t value Pr(>|t|)  
(Intercept)  0.25424    0.04856   5.235  1.84e-07 ***
LogRankRatio -0.30940    0.03572  -8.661  < 2e-16 ***
```

---

Signif. codes:  **'***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasibinomial family taken to be 1.002121)

Null deviance: 2476.2 on 1809 degrees of freedom
Residual deviance: 2395.9 on 1808 degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4

Result interpretations:

- \( \alpha > 0 \) and \( \alpha \) is **statistically significant** \( \implies \)
  
  "**Controlling** for difference in level, dynamic 0-2 \( \rightarrow \) 3-2 is MORE LIKELY than 0-2 \( \rightarrow \) 2-3."

- \( \beta < 0 \) and \( \beta \) is **statistically significant** \( \implies \)
  
  "If player ranks **lower** than opponent (\( X > 0 \) \( \implies \) lower ranking), he has **lower** chance to win."