```
dat = data.frame(
female = c(rep(0,6), rep(1,6)),
dept = rep(LETTERS[1:6],2),
apps = c(825,560,325,417,191,373,108,25,593,375,393,341),
admits = c(512,353,120,138,53,22,89,17,202,131,94,24))

head(dat)
```

```
##   female dept apps admits
## 1      0    A  825    512
## 2      0    B  560    353
## 3      0    C  325    120
## 4      0    D  417    138
## 5      0    E  191     53
## 6      0    F  373     22
```

Explain why Simpson's paradox occurs for these data.

When a relationship observed between two variables in a larger population,
either vanishes or reverses when the population is divided into subpopulations,
it is called Simpson's paradox.

In this particular case, it occurs due to the difference in the population distributions at a
n over-all level,
and at a department level.

The Simpson's paradox is observed due to the following reasons:

Uneven distribution of applicants:
- There were large differences in the number of male vs. female applicants to different de
partments.
- Some departments with high admission rates had very few female applicants, skewing the o
verall percentages.

In essence, the paradox arose because the relationship between gender and admissions was reve
rsed when the data was disaggregated by
department. This highlights the importance of considering potential confounding variables and
examining data at different levels of
granularity to avoid drawing incorrect conclusions about discrimination or bias.

```
admissions_df <- as.data.frame(UCBAdmissions)

head(admissions_df)
```

```
##      Admit Gender Dept Freq
## 1 Admitted   Male    A  512
## 2 Rejected   Male    A  313
## 3 Admitted Female    A   89
## 4 Rejected Female    A   19
## 5 Admitted   Male    B  353
## 6 Rejected   Male    B  207
```

```
# Creating binary columns for admission and gender

admissions_df$AdmitBinary <- ifelse(admissions_df$Admit == "Admitted", 1, 0)
admissions_df$GenderBinary <- ifelse(admissions_df$Gender == "Male", 1, 0)



head(admissions_df)
```

```
##      Admit Gender Dept Freq AdmitBinary GenderBinary
## 1 Admitted   Male    A  512           1            1
## 2 Rejected   Male    A  313           0            1
## 3 Admitted Female    A   89           1            0
## 4 Rejected Female    A   19           0            0
## 5 Admitted   Male    B  353           1            1
## 6 Rejected   Male    B  207           0            1
```

```
This is aggregated Admissions data for men and women.

The first column is dept.
The next 3 are applied, admitted and admitted% for men
The next 3 are applied, admitted and admitted% for women
The next 3 are applied, admitted and admitted% for the entire group

A       825     512     62.1    108  89     82.4    993   620       64.4
B       560     353     63.0    25   17     68.0    585     577     63.2
C       325     120     36.9    593  202      34.1    918     322       35.1
D       417     138     33.1    375  131      34.9    792     269       34.0
E       191     53      27.7    393  94     23.9    584   147       25.2
F       373     22      5.9     341  24     7.0     714   46        6.4
Total       2691    1198    44.5    1835    557   30.4    4526  1755      38.8



men --- odds --- admitted / (applied - admitted)
>>> 1198/(2691-1198)
0.8024112525117214



women --- odds --- admitted / (applied - admitted)
>>> 557/(1835-557)
0.43583724569640064


men vs women odds ratio
>>> 0.8024/0.4358
1.841211564938045

>>> math.log(1.8412)
0.6104175329609492
```

```
model <- glm(AdmitBinary ~ GenderBinary, data = admissions_df, family = binomial, weights = F
req)
summary(model)
```

```
##
## Call:
## glm(formula = AdmitBinary ~ GenderBinary, family = binomial,
##     data = admissions_df, weights = Freq)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.83049    0.05077 -16.358   <2e-16 ***
## GenderBinary  0.61035    0.06389   9.553   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6044.3  on 23  degrees of freedom
## Residual deviance: 5950.9  on 22  degrees of freedom
## AIC: 5954.9
##
## Number of Fisher Scoring iterations: 4
```

The coefficient for GenderBinary matches with the value calculated for ->

sample log odds ratio for Men versus Women given their admission rates at the bottom = 0.61035 ~ 0.6104

```
model <- glm(AdmitBinary ~ GenderBinary + as.factor(Dept), data = admissions_df, family = binomial, weights = Freq)
summary(model)
```

```
##
## Call:
## glm(formula = AdmitBinary ~ GenderBinary + as.factor(Dept), family = binomial,
##     data = admissions_df, weights = Freq)
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       0.68192    0.09911   6.880 5.97e-12 ***
## GenderBinary     -0.09987    0.08085  -1.235    0.217
## as.factor(Dept)B -0.04340    0.10984  -0.395    0.693
## as.factor(Dept)C -1.26260    0.10663 -11.841  < 2e-16 ***
## as.factor(Dept)D -1.29461    0.10582 -12.234  < 2e-16 ***
## as.factor(Dept)E -1.73931    0.12611 -13.792  < 2e-16 ***
## as.factor(Dept)F -3.30648    0.16998 -19.452  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6044.3  on 23  degrees of freedom
## Residual deviance: 5187.5  on 17  degrees of freedom
## AIC: 5201.5
##
## Number of Fisher Scoring iterations: 6
```

We could already see, that at a department level, the proportion of men who apply and get admitted is lower than that of women, primarily
due to the differences in the number of women applying to the departments.

On a closer look, it would become apparent, that the departments with lower-admission-rates in general (for both men and women), i.e. the
more competitive departments, receive larger number of applications from women, while the ones with higher admission rates receive lower
number of applications from women.

Hence, the gender being male, would lead to lowering of the probability of admission, according to the equation.

Thus, in this manner the Simpson's paradox can be seen in this data.