Statistics 2B (Practical)

Compiled by Mr. V van Appel, Department of Statistics - University of Johannesburg
Practical video done by Mr. T Hansragh, Department of Statistics - University of Johannesburg

Practical 5: Chapter 1 & 2 (The Statistical Sleuth)

Part 1: https://youtu.be/lLgzRTgSpoo
Part 2: https://youtu.be/60cT03o0g3A

5.1 Introduction

This document is intended to help describe how to undertake analyses introduced as examples in the Third Edition of the Statistical Sleuth (2013) by Fred Ramsey and Dan Schafer.

This work leverages initiatives undertaken by Project MOSAIC (http://www.mosaic-web.org), an NSF-funded effort to improve the teaching of statistics, calculus, science and computing in the undergraduate curriculum. In particular, we utilize the mosaic package, which was written to simplify the use of R for introductory statistics courses.

To use a package within R, it must be installed (one time), and loaded (each session). The package can be installed using the following command:

```
> #install.packages("mosaic") # note the quotation marks
>
> library(mosaic)
> #install.packages("Sleuth3") # note the quotation marks
> library(Sleuth3)
```

This needs to be done once per session.

The specific goal of this document is to demonstrate how to calculate the quantities described in the Statistical Sleuth, Chapter 1: Drawing Statistical Conclusions using R.

5.2 Case Study 1: Motivation and Creativity

Do grading systems promote creativity in students? Do ranking systems and incentive awards increase productivity among employees? Do rewards and praise stimulate children to learn? The data for Case Study 1 was collected by psychologist Teresa Amabile in an experiment concerning the effects of intrinsic and extrinsic motivation on creativity. Subjects with considerable experience in creative writing were randomly assigned to one of two treatment groups: 24 of the subjects were placed in the "intrinsic" treatment group, and 23 in the "extrinsic" treatment group.

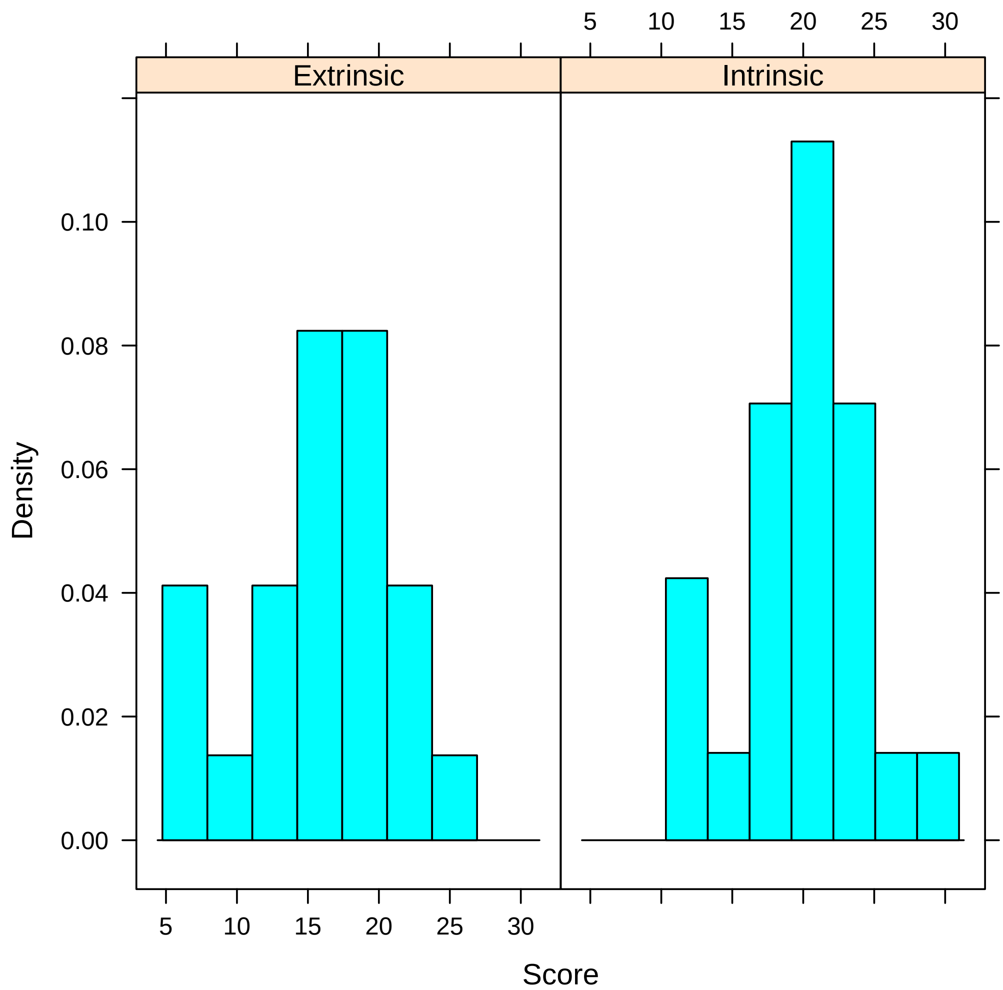
1 Summarize the essential features of the data using graphical display and summary statistics and interpret the results. You may compute the followings.

- Summarize the data.

```
> summary(case0101)

     Score           Treatment
 Min.   : 5.00   Extrinsic:23
 1st Qu.:14.90   Intrinsic:24
 Median :18.70
 Mean   :17.86
 3rd Qu.:21.25
 Max.   :29.70

> favstats(Score~Treatment,data=case0101)

  Treatment min     Q1 median    Q3  max     mean       sd  n missing
1 Extrinsic   5 12.150   17.2 18.95 24.0 15.73913 5.252596 23       0
2 Intrinsic  12 17.425   20.4 22.30 29.7 19.88333 4.439513 24       0
```

- Draw the histogram for Score versus treatment.

```
> histogram(~Score|Treatment, data = case0101)
```



- Generate Stem and leaf plots for "Extrinsic" and "Intrinsic" motivations.

```
> Treatment=case0101$Treatment
> with(subset(case0101,Treatment =="Extrinsic"), stem(Score, scale = 5))

  The decimal point is at the |

   5 | 04
   6 | 1
   7 |
   8 |
   9 |
  10 | 9
  11 | 8
  12 | 03
  13 |
  14 | 8
  15 | 0
  16 | 8
  17 | 2245
  18 | 577
  19 | 25
  20 | 7
  21 | 2
  22 | 1
  23 |
  24 | 0

> with(subset(case0101,Treatment =="Intrinsic"), stem(Score, scale = 5))

  The decimal point is at the |

  12 | 009
  13 | 6
  14 |
  15 |
  16 | 6
  17 | 25
  18 | 2
  19 | 138
  20 | 356
  21 | 36
  22 | 126
  23 | 1
  24 | 03
  25 |
  26 | 7
  27 |
  28 |
  29 | 7
```

2 Inferential procedures: test the hypothesis that students receiving the "intrinsic" rather than the "extrinsic" questionnaire caused students in this study to score higher on poem creativity.

- Perform two sample t-test.

```
> t.test(Score ~ Treatment, alternative = "two.sided", data =case0101)

        Welch Two Sample t-test

data:  Score by Treatment
t = -2.9153, df = 43.108, p-value = 0.005618
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -7.010803 -1.277603
sample estimates:
mean in group Extrinsic mean in group Intrinsic
               15.73913                19.88333
```

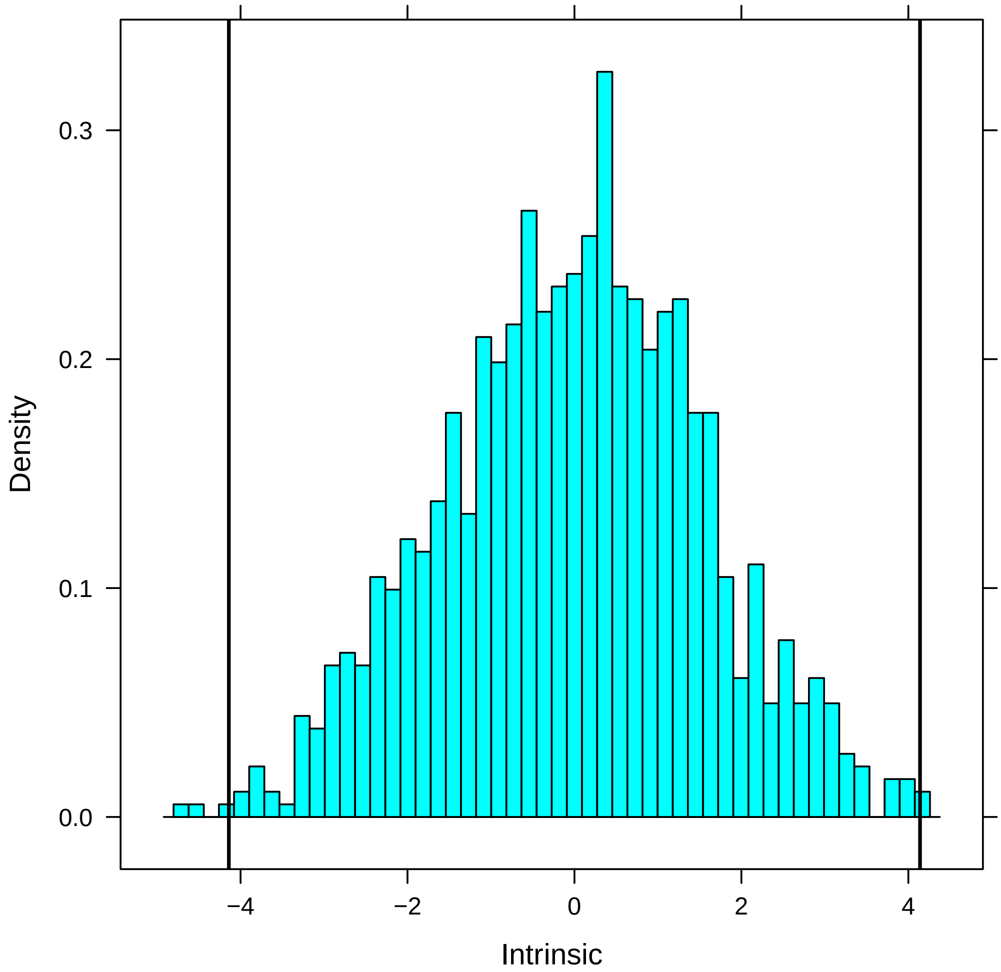
- Perform simple linear regression.

```
> summary(lm(Score ~ Treatment, data = case0101))

Call:
lm(formula = Score ~ Treatment, data = case0101)

Residuals:
    Min      1Q  Median      3Q     Max
-10.739  -2.983   1.061   2.961   9.817

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)          15.739      1.012  15.550  < 2e-16 ***
TreatmentIntrinsic    4.144      1.416   2.926  0.00537 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.854 on 45 degrees of freedom
Multiple R-squared:  0.1598,	Adjusted R-squared:  0.1412
F-statistic: 8.561 on 1 and 45 DF,  p-value: 0.005366
```

3 Permutation test

```
> diffmeans = diff(mean(Score ~ Treatment, data = case0101))
> diffmeans # observed difference

Intrinsic
 4.144203

> numsim = 1000 # set to a sufficient number
> nulldist = do(numsim) * diff(mean(Score ~ shuffle(Treatment), data = case0101))
> confint(nulldist)

       name     lower    upper level     method estimate
1 Intrinsic -3.050756 3.003687  0.95 percentile 4.144203
```

```
> # Display 1.8 Sleuth
> histogram(~Intrinsic, nint = 50, data = nulldist, v = c(-4.14, 4.14))
```



As described in the Sleuth on page 12, if the group assignment changes, we will get different results. First, the test statistics will be just as likely to be negative as positive. Second, the majority of values fall in the range from -3.0 to +3.0. Third, only few of the 1,000 randomization produced test statistics as large as 4.14. This last point indicates that 4.14 is a value corresponding to an unusually uneven randomization outcome, if the null hypothesis is correct.

5.3 Case Study 2: Gender Discrimination

For Case Study 2: Gender discrimination the following questions are posed: Did a bank discriminatorily pay higher starting salaries to men than to women? The data in Display 1.3 are the beginning salaries for all 32 male and all 61 female skilled, entry-level clerical employees hired by the bank between 1969 and 1977.

1 Summarize the essential features of the data using graphical display and summary statistics and interpret the results. You may compute the followings.

- Summarize the data.

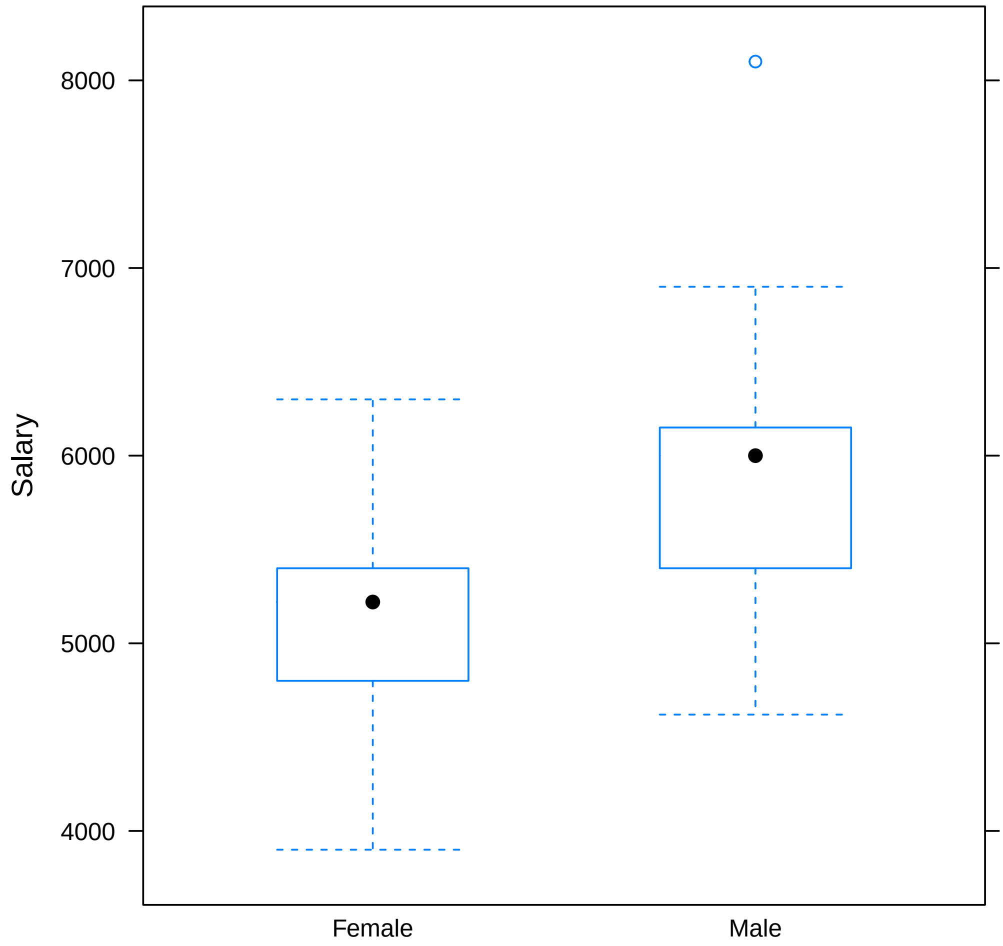
```
> (case0102) # Display 1.3 Sleuth p4

   Salary    Sex
1    3900 Female
2    4020 Female
3    4290 Female
4    4380 Female
5    4380 Female
6    4380 Female
7    4380 Female
8    4380 Female
9    4440 Female
10   4500 Female
11   4500 Female
12   4620 Female
13   4800 Female
14   4800 Female
15   4800 Female
16   4800 Female
17   4800 Female
18   4800 Female
19   4800 Female
20   4800 Female
21   4800 Female
22   4800 Female
23   4980 Female
24   5100 Female
25   5100 Female
26   5100 Female
27   5100 Female
28   5100 Female
29   5100 Female
30   5160 Female
31   5220 Female
32   5220 Female
33   5280 Female
34   5280 Female
35   5280 Female
36   5400 Female
37   5400 Female
38   5400 Female
39   5400 Female
40   5400 Female
41   5400 Female
42   5400 Female
43   5400 Female
44   5400 Female
45   5400 Female
```

```
46    5400 Female
47    5400 Female
48    5520 Female
49    5520 Female
50    5580 Female
51    5640 Female
52    5700 Female
53    5700 Female
54    5700 Female
55    5700 Female
56    5700 Female
57    6000 Female
58    6000 Female
59    6120 Female
60    6300 Female
61    6300 Female
62    4620   Male
63    5040   Male
64    5100   Male
65    5100   Male
66    5220   Male
67    5400   Male
68    5400   Male
69    5400   Male
70    5400   Male
71    5400   Male
72    5700   Male
73    6000   Male
74    6000   Male
75    6000   Male
76    6000   Male
77    6000   Male
78    6000   Male
79    6000   Male
80    6000   Male
81    6000   Male
82    6000   Male
83    6000   Male
84    6000   Male
85    6000   Male
86    6300   Male
87    6600   Male
88    6600   Male
89    6600   Male
90    6840   Male
91    6900   Male
92    6900   Male
93    8100   Male
> favstats(Salary ~ Sex, data = case0102)
```

```
     Sex  min   Q1 median   Q3  max     mean       sd  n missing
1 Female 3900 4800   5220 5400 6300 5138.852 539.8707 61       0
2   Male 4620 5400   6000 6075 8100 5956.875 690.7333 32       0
```

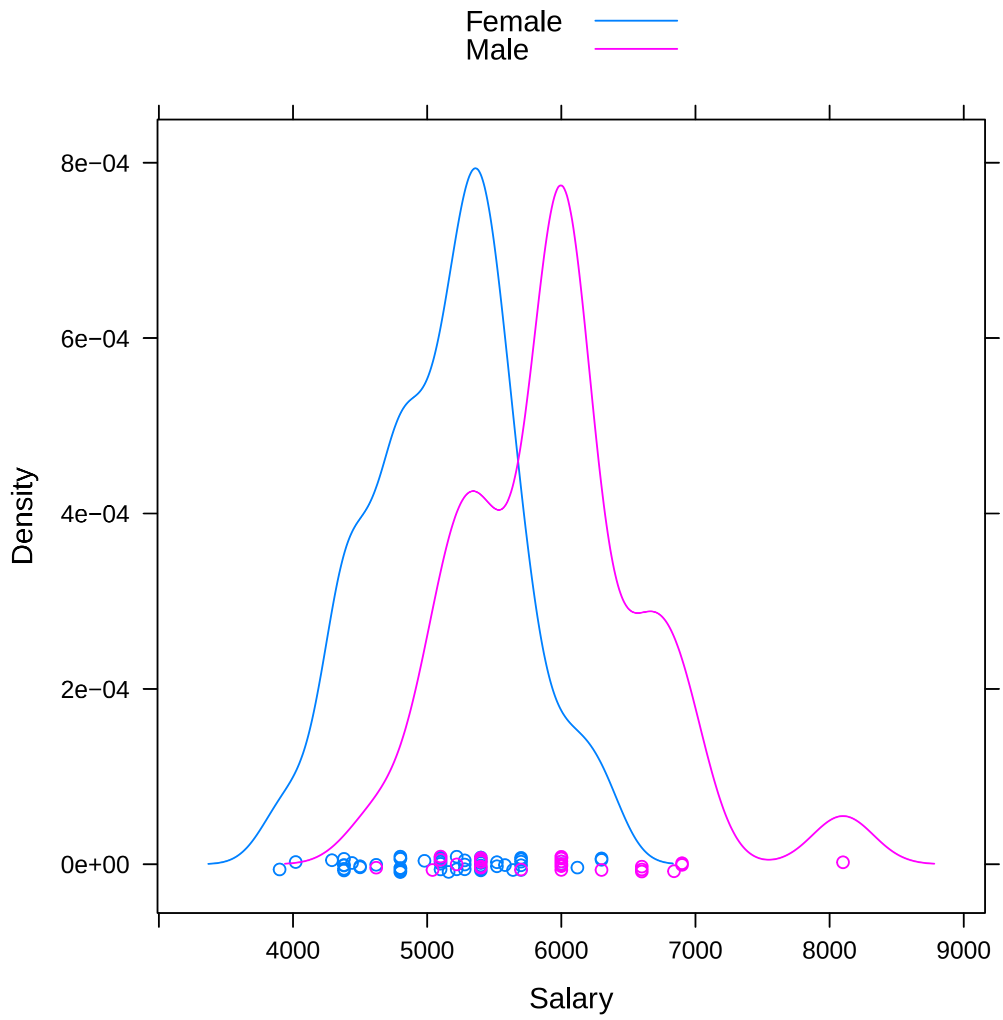
- Draw the boxplot.

```
> bwplot(Salary ~ Sex, data = case0102)
```



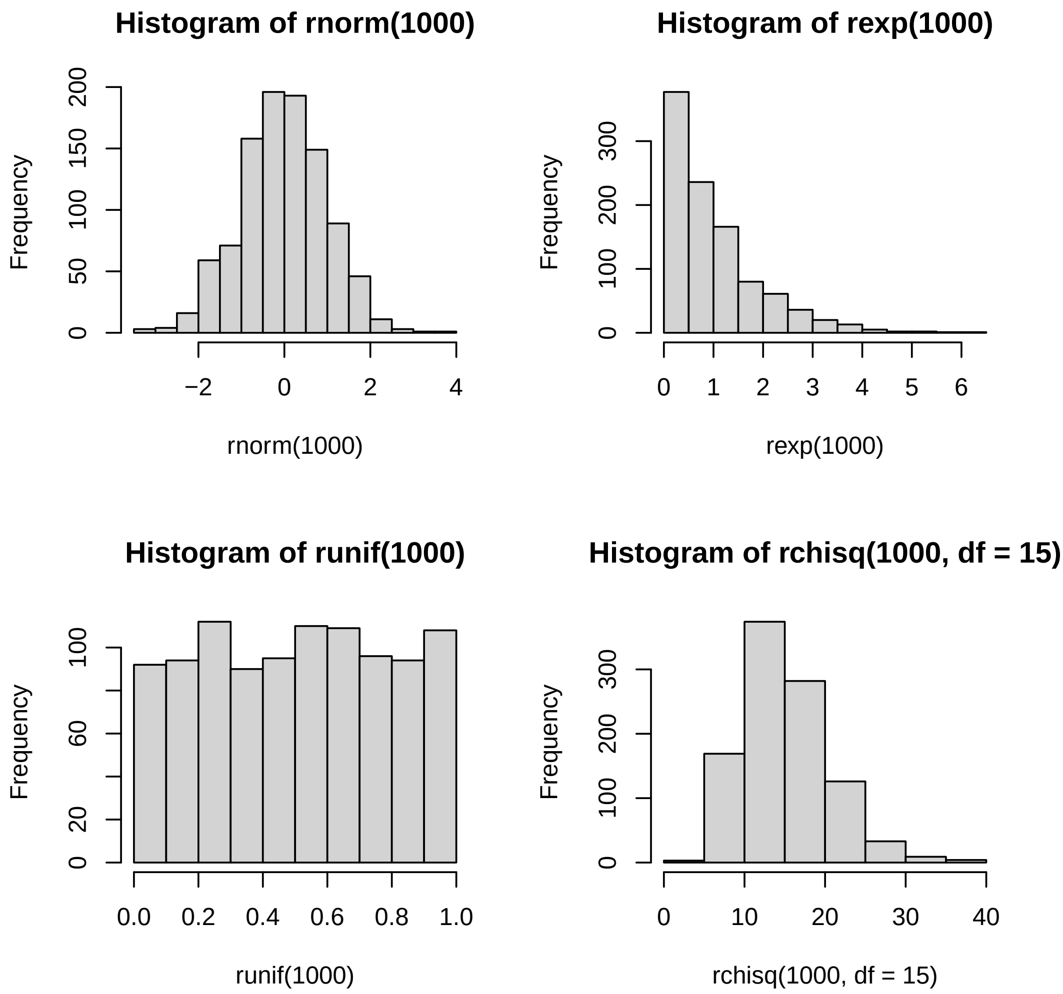
- Compute the density plot

```
> densityplot(~Salary, groups = Sex, auto.key = TRUE, data = case0102)
```



- Compare the above density plot with the following plots

```
> par(mfrow = c(2,2))
> hist(rnorm(1000)) # Normal
> hist(rexp(1000)) # Long-tailed
> hist(runif(1000)) # Short-tailed
> hist(rchisq(1000, df = 15)) # Skewed
```



2 Inferential procedures: test the hypothesis that the mean salary of males is greater than the mean salary of females.

- Perform two sample t-test.

```
> t.test(Salary ~ Sex, var.equal = TRUE, data = case0102)

        Two Sample t-test

data:  Salary by Sex
t = -6.2926, df = 91, p-value = 1.076e-08
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1076.2465  -559.7985
sample estimates:
mean in group Female   mean in group Male
            5138.852             5956.875
```

3 Permutation test

We undertake a permutation test to assess whether the differences in the center of these samples that we are observing are due to chance, if the distributions are actually equivalent back in the populations of male and female possible clerical hires. We start by calculating our test statistic (the difference in means) for the observed data, then simulate from the null distribution (where the labels can be interchanged) and calculate our p-value.

```
> obsdiff = diff(mean(Salary ~ Sex, data = case0102))
> obsdiff
```

```
    Male
818.0225
```

```
> numsim = 1000
> res = do(numsim) * diff(mean(Salary ~ shuffle(Sex), data = case0102))
> densityplot(~Male, data = res)
> confint(res)
```

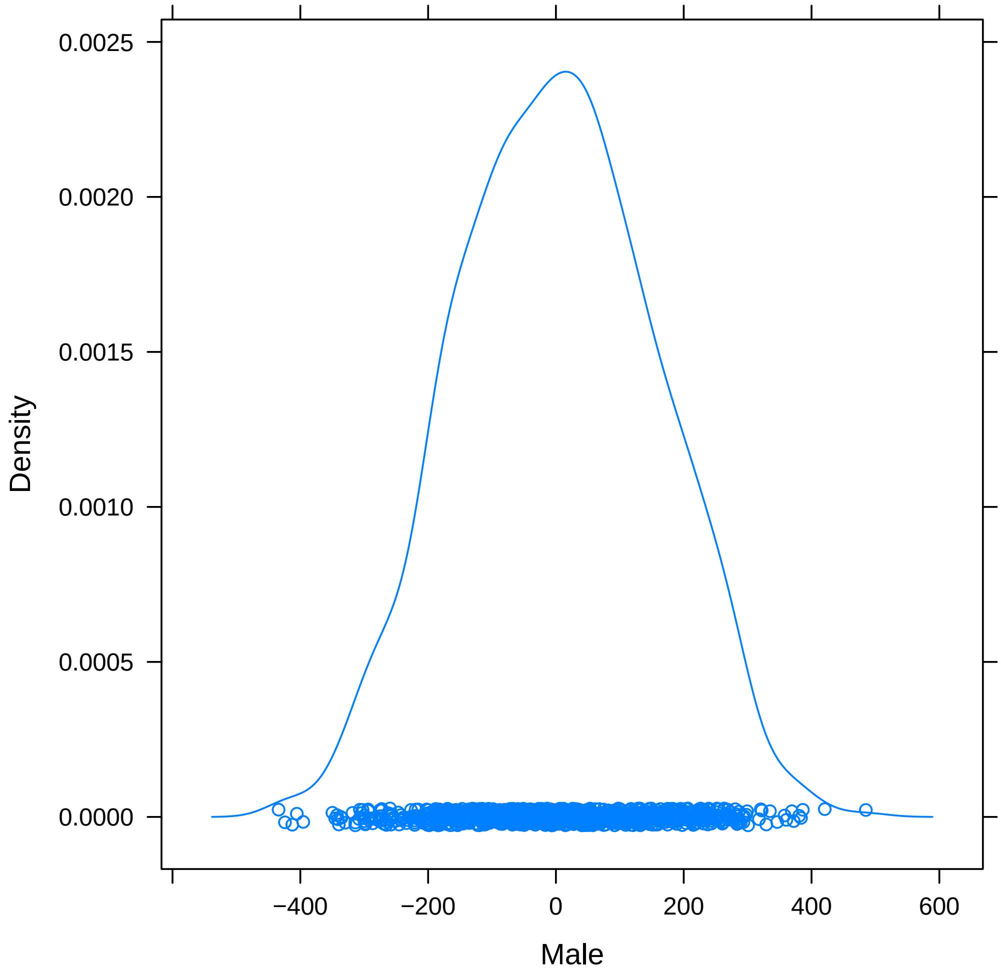
```
  name     lower    upper level     method estimate
1 Male -294.0826 283.4631  0.95 percentile 818.0225
```

```
> larger = sum(with(res, abs(Male) >= abs(obsdiff)))
> larger
```

```
[1] 0
```

```
> pval = larger/numsim
> pval
```

```
[1] 0
```



Through the permutation test, we observe that the mean starting salary for males is significantly larger than the mean starting salary for females, as we never see a permuted difference in means close to our observed value. Therefore, we reject the null hypothesis ($p < 0.001$) and conclude that the salaries of the men are higher than that of the women back in the population.

5.4 Case Study 2.1: Evidence Supporting Darwin's Theory of Natural Selection

Do birds evolve to adapt to their environments? That's the question being addressed by Case Study 2.1 in the Sleuth.

1 Summarize the essential features of the data using graphical display and summary statistics and interpret the results. You may compute the followings.

- Summarize the data.

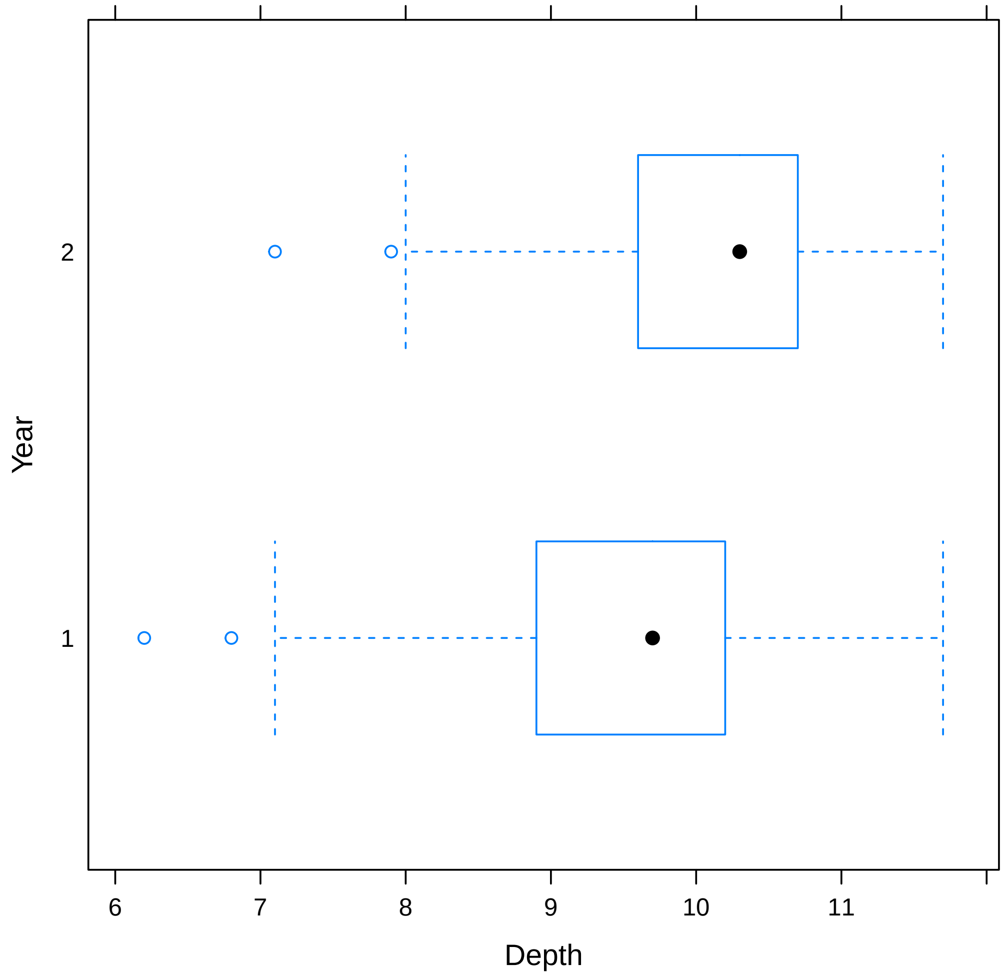
```
> summary(case0201)
      Year           Depth
 Min.   :1976   Min.   : 6.200
```

```
 1st Qu.:1976   1st Qu.: 9.100
 Median :1977   Median : 9.900
 Mean   :1977   Mean   : 9.804
 3rd Qu.:1978   3rd Qu.:10.500
 Max.   :1978   Max.   :11.700
```

```
> fav=favstats(Depth ~ Year, data=case0201)
```

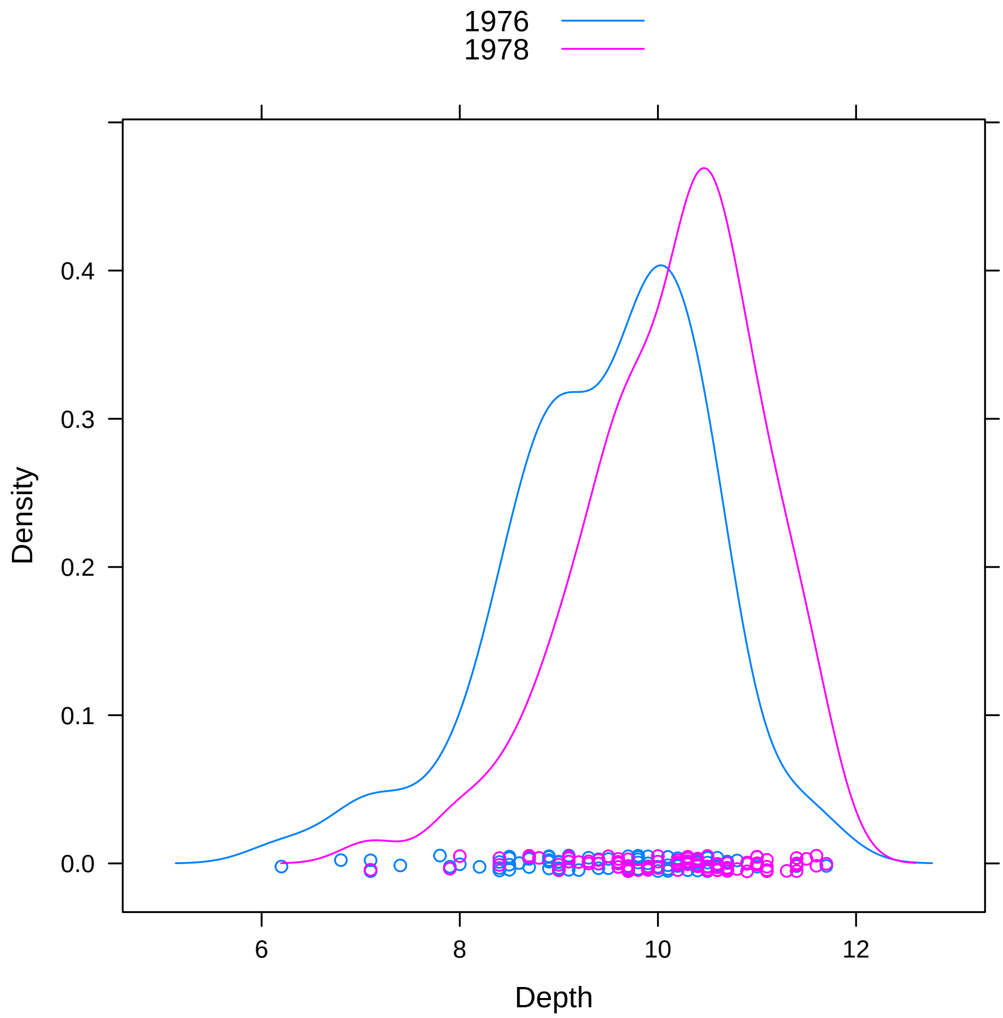
- Draw the boxplot.

```
> bwplot(Year~Depth, data=case0201)
```



- Compute the density plot

```
> densityplot(~Depth, groups = Year, auto.key = TRUE, data = case0201)
```



2 Inferential procedures:

- Perform two sample t-test.

```
> t.test(Depth ~ Year, var.equal = TRUE, data = case0201)

        Two Sample t-test

data:  Depth by Year
t = -4.5833, df = 176, p-value = 8.65e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.9564088 -0.3806698
sample estimates:
mean in group 1976 mean in group 1978
          9.469663          10.138202
```

3 Compute the 95% CI

```
> confint(lm(Depth ~ Year, data = case0201))
```

```
                  2.5 %       97.5 %
(Intercept) -935.6062223 -366.4881597
Year           0.1903349    0.4782044
```

5.5 Case Study 2.2: Anatomical Abnormalities Associated with Schizophrenia

Is the area of brain related to the development of schizophrenia? That's the question being addressed by case study 2.2 in the Sleuth.

1 Summarize the essential features of the data using graphical display and summary statistics and interpret the results. You may compute the followings.

- Summarize the data.

```
> summary(case0202)
```

```
  Unaffected       Affected
 Min.   :1.250   Min.   :1.02
 1st Qu.:1.600   1st Qu.:1.31
 Median :1.770   Median :1.59
 Mean   :1.759   Mean   :1.56
 3rd Qu.:1.935   3rd Qu.:1.78
 Max.   :2.080   Max.   :2.02
```

```
> case0202 = transform(case0202, DIFF = Unaffected - Affected)
> favstats(~DIFF, data = case0202)
```

```
   min    Q1 median    Q3  max      mean        sd  n missing
 -0.19 0.055   0.11 0.315 0.67 0.1986667 0.2382935 15       0
```

2 Inferential procedures:

- Calculate t-statistics.

```
> difmean = mean(~DIFF, data = case0202)
> difsd = sd(~DIFF, data = case0202)
> difn = nrow(case0202)
> difSE = difsd/sqrt(difn)
> tscore = (difmean - 0)/difSE # hypothesis difference=0
> tscore
```

```
[1] 3.228928
```

```
> twopvalue = 2*(1- pt(tscore, difn - 1))
> twopvalue
```

```
[1] 0.006061544
```

3 Compute the 95% CI

```
> tstar = qt(0.975, difn - 1)
> schizolower = difmean - tstar * difSE
> schizoupper = difmean + tstar * difSE
> with(case0202, t.test(Unaffected, Affected, paired = TRUE))

        Paired t-test

data:  Unaffected and Affected
t = 3.2289, df = 14, p-value = 0.006062
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.0667041 0.3306292
sample estimates:
mean of the differences
              0.1986667
```