SEX DISCRIMINATION PROBLEM

17. Brief Version of the Case Study

17.1 Problem Formulation

This discussion will be concerned with data on beginning salaries for all 32 male and 61 female skilled, entry-level clerical employees hired by the Harris Bank of Chicago between 1969 and 1971. These data come from a larger Harris Bank data file made public by the defense statisticians in a sex discrimination lawsuit against the bank.

The data for the problem is given in your textbook, pages 330-331 (see also pages 4-5). These data are also available in the SPSS file *discrim.sav* located on the FTP server. The instructions how to download the data files using FTP are available in the *Introduction to SPSS for Windows* manual (Appendix 1) or in the *Introduction to SPSS* module in STAT 252 Web site (Appendices).

The data give beginning salaries together with several valid measures of job qualification such as education level and previous experience. The following is a description of the variables in the study:

Column	Name of Variable	Description of Variable
1	BSAL	Beginning Annual Salary (dollars)
2	SAL77	Salary as of March 1977 (dollars)
3	FSEX	Sex (1 for females, 0 for males)
4	SENIOR	Seniority (months since first hired)
5	AGE	Age (months)
6	EDUC	Education (years)
7	EXPER	Experience prior to Employment with the bank (months)

We would like to use SPSS to answer the following two questions using the data:

1. Did the bank discriminatorily pay higher starting salaries to men than to women?
2. Did the females tend to receive smaller pay increases than similarly experienced (in terms of seniority) males during their employment with the bank?

The word *discriminatorily* in the above question makes our analysis especially difficult. Indeed, you will see that it will be relatively easy to demonstrate that the bank paid higher starting salaries to men than to women, but it will be hard to prove that this disparity was due to gender alone.

We will also discuss other related problems such as the changes in gender structure of the employees in the bank over the study period or the relationship between salary in 1977 and seniority. It will be also interesting to answer the question whether females tended to receive smaller pay increases than similarly qualified males during their employment with the bank.

17.2 Data Collection

There are seven variables considered in the study: starting salary, salary as of March 1977, gender, seniority, age, education, and experience prior to employment with the bank. Obviously, starting salary is affected by gender, seniority, age, education, and prior experience but it is not affected by salary in 1977.

Does seniority affect starting salary? It does not, but the time of hire certainly does. The data comes from a larger data file containing the data about entry-level clerical employees hired by the bank between 1965 and 1977. To account for the general effect of beginning salaries increasing over this period (inflation), seniority measured as the number of months since first hired, was also included in the data. Thus seniority expresses indirectly the information about time of hire and hence it is an important variable affecting starting salary. All of the employees in our case study were hired between 1969 and 1971. This is why seniority measured as the number of months since first hired is between 65 and 98 in our data file.

Summarizing, all the considered variables except for salary in 1977 are important variables affecting starting salary.

We have to be aware of the limitations of available data. The starting salary is affected by several factors such as job qualifications, experience, and job market situation. Only some of these factors could be measured somehow whereas starting salary could be measured precisely. In general, good measures of job qualifications and performance are seldom available.

All 93 persons considered in the study are described as entry-level clerical employees, but it is reasonable to assume that this group might be divided into several non-homogeneous subgroups with different job responsibilities. It is likely that initial placement in one of these groups affected salary determination. However, this information is not available in our case study.

17.3 Study Design

In order to support the discrimination claim we have to use the data to show that gender is the only *cause* of the observed disparity in starting salaries between males and females.

The case study is an example of an observational study because the sex of each of the 93 employees was not decided by the investigator. In other words, allocation of employees to the two gender groups (males, females) was not determined by any chance mechanism.

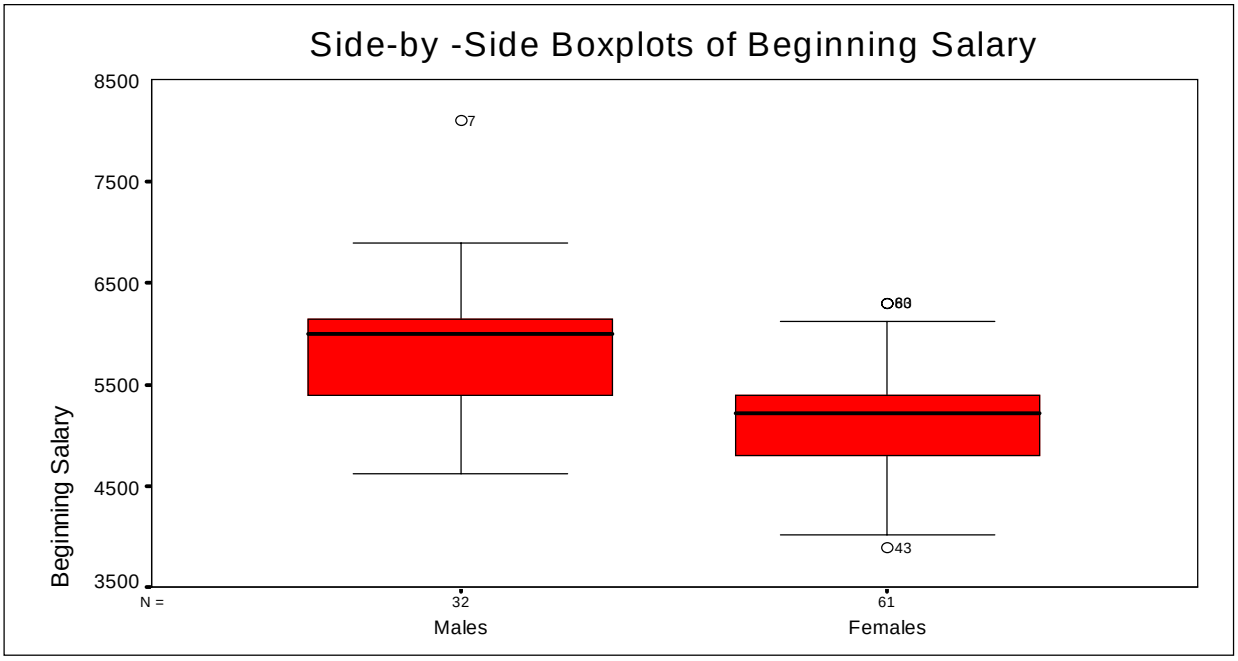
As the study is an observational study, we are not able to draw any causal conclusions from the statistical analysis alone. It is possible that some confounding variables are responsible for the disparity in the starting salaries. You will see later that the males generally did have more years of education than the females, and this, not gender, may have been responsible for the observed differences in the starting salaries for males and females. Thus, the effect of gender cannot be separated from the effect of education.

With the current study design we are not able to prove sex discrimination because we are not able to isolate all confounding variables to see the effects of gender alone on starting salaries. Although statistical analysis alone is not able to prove sex discrimination, it can be useful in a court of law to establish discrimination.

17.4 Displaying Relevent Variables

1. **The side-by-side boxplots for male and female starting salaries.**
2. **Did the females tend to be less educated than the males?**
3. **Did the females tend to have less experience prior to employment with the bank than the males?**

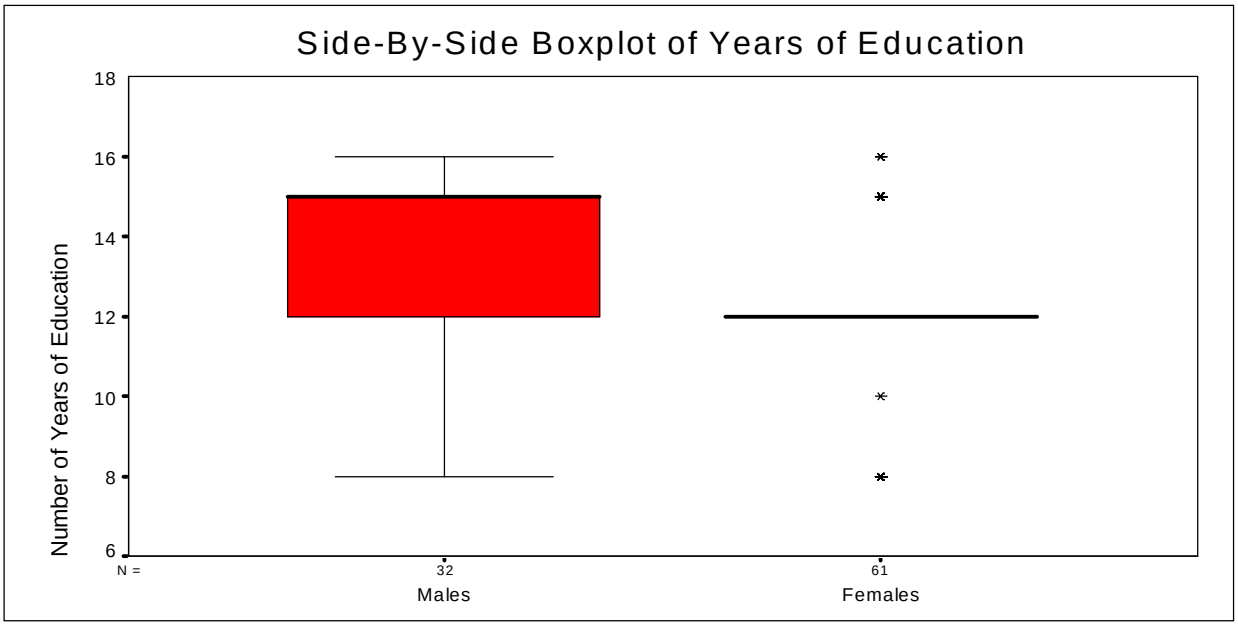
1. Now we will obtain the side-by-side boxplots for male and female starting salaries. We will identify outliers (if any) by the 1.5*IQR criterion in each distribution.



The distribution of male starting salaries is shifted up compared to the distribution of female starting salaries. The median male starting salary indicated by the position of the horizontal line within the red box for males is significantly higher than the corresponding value for females. The spread of data, represented by the width of the box (interquartile range) is larger for male than female observations.

The positions of outliers in the data sets are indicated above or below the whiskers in the boxplots. Thus the seventh observation (O7) in the file is an outlier in the male data. This corresponds to the salary of $8,100. The 43rd and 80th observations are outliers in the female data. This corresponds to $3,900 and $6,300, respectively.

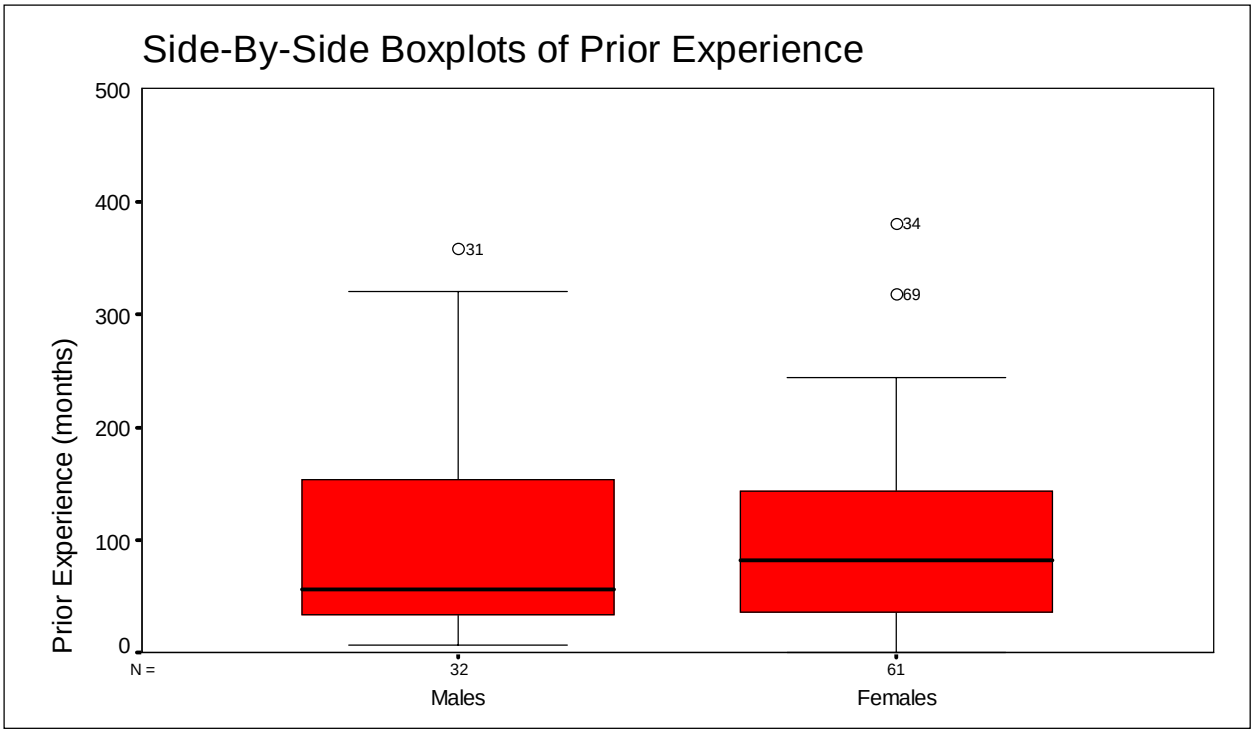
2. Did the females tend to be less educated than the males? We will answer the question by producing side-by-side boxplots of number of years of education for males and females.



The males have more years of education, on average, than females. The median number of years of education for males is 15 (upper side of the red box) and the median number of years of education for females is only 12. The boxplot for the females is flat indicating that the interquartile range is zero. The first, the second, and the third quartiles are equal to 12. Most females have 12 years of education. The extreme observations (more than 3*IQR from the end of the box) are marked with asterisks. As the interquartile range is zero, all observations different from 12 are considered extreme observations.

Most of the employees have either 8, 12, or 15 years of education and 10 of the 11 individuals with 8 years of education are females.

3. Did the females tend to have less experience prior to employment with the bank than the males? We will answer the question by obtaining side-by-side boxplots of prior experience for males and females.



As you can see, the median number of months of prior experience for females is slightly larger than the corresponding median for males. The spread is similar. The distribution is skewed to the right for the males (look at the position of the median inside the box), and it is approximately symmetric for the females.

17.5 Summarizing Relevant Variables

1. **The summary statistics for male and female starting salaries.**
2. **The summary statistics for starting salary, number of years of education, and number of years of prior experience for males and females.**

1. Obtain the summary statistics for male and female starting salaries. In particular, obtain the mean, median, quartiles, standard deviation, minimum, and maximum.

The following is a part of the output produced by *Explore*. See the details in *Computing Instructions*.

Beginning Salaries By Gender

Males

Valid cases:	32.0	Missing cases:	.0	Percent missing:	.0		
Mean	5956.875	Std Err	122.1056	Min	4620.000	Skewness	.7674
Median	6000.000	Variance	477112.5	Max	8100.000	S E Skew	.4145
5% Trim	5928.333	Std Dev	690.7333	Range	3480.000	Kurtosis	1.7728
95% CI for Mean (5707.839, 6205.911)				IQR	825.0000	S E Kurt	.8094

Females

Valid cases:	61.0	Missing cases:	.0	Percent missing:	.0		
Mean	5138.852	Std Err	69.1234	Min	3900.000	Skewness	-.0780
Median	5220.000	Variance	291460.3	Max	6300.000	S E Skew	.3063
5% Trim	5136.995	Std Dev	539.8707	Range	2400.000	Kurtosis	-.2863
95% CI for Mean (5000.585, 5277.120)				IQR	600.0000	S E Kurt	.6038

Now we will obtain similar outputs for other variables and summarize our results in the form of a table.

2. We will obtain the mean, median, standard deviation, minimum, maximum for starting salary, number of years of education, and number of years of prior experience for males and females.

The *Explore* procedure in SPSS produces the following summaries:

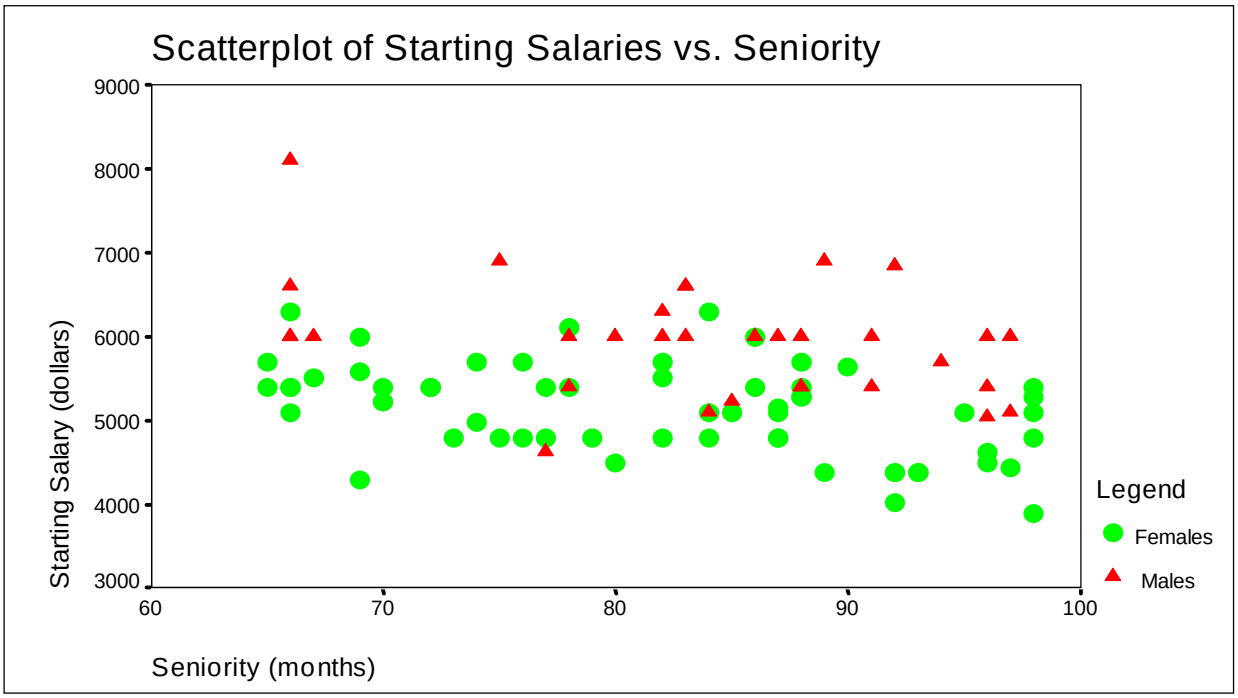
	Gender	Number	Mean	Median	St. Deviation	Min	Max
Starting Salary (dollars)	Males	32	5956.87	6000	690.73	4620	8100
	Females	61	5138.85	5220	539.87	3900	6300
Education (years)	Males	32	13.53	15	1.87	8	16
	Females	61	11.97	12	2.31	8	16
Experience (months)	Males	32	103.05	56	102.10	7	359
	Females	61	99.82	82	85.40	0	381

As you can see, the mean beginning salaries for women were lower than those for men, but the men had higher mean years of education and months of experience. This is consistent with the conclusions we obtained by analyzing the graphical displays for the variables. Notice that although the average number of months of prior experience is higher for men, the median is larger for women as we noticed while examining the side-by-side boxplots of starting salaries for males and females. The distribution of number of months of experience is highly skewed for men.

17.6 Displaying Relationships

1. **Did the bank pay higher starting salaries to men than to women hired at the same time?**
2. **Did the bank pay higher starting salaries to men than to women with approximately the same previous experience?**
3. **Did the women tend to receive lower starting salaries than similarly educated men?**
4. **Scatterplot matrix.**

1. Did the bank pay higher starting salaries to men than to women hired at the same time? In order to answer the question, we will obtain a scatterplot of starting salaries versus seniority for males and females. Plotting salaries against seniority ensures that we will be able to compare the salaries for both gender groups hired at the same time. We will use different marking symbol in the plot to denote male and female subjects.



As you can see the starting salaries of males tend to be higher than the salaries of females hired at the same time. No matter when the clerks have been employed, the highest paid employees are males. The situation has not improved for those hired at the end of the three-year period (low seniority), even it has worsened because almost all new male employees get higher salaries than the females. The plot indicates increasing disparity over the considered period.

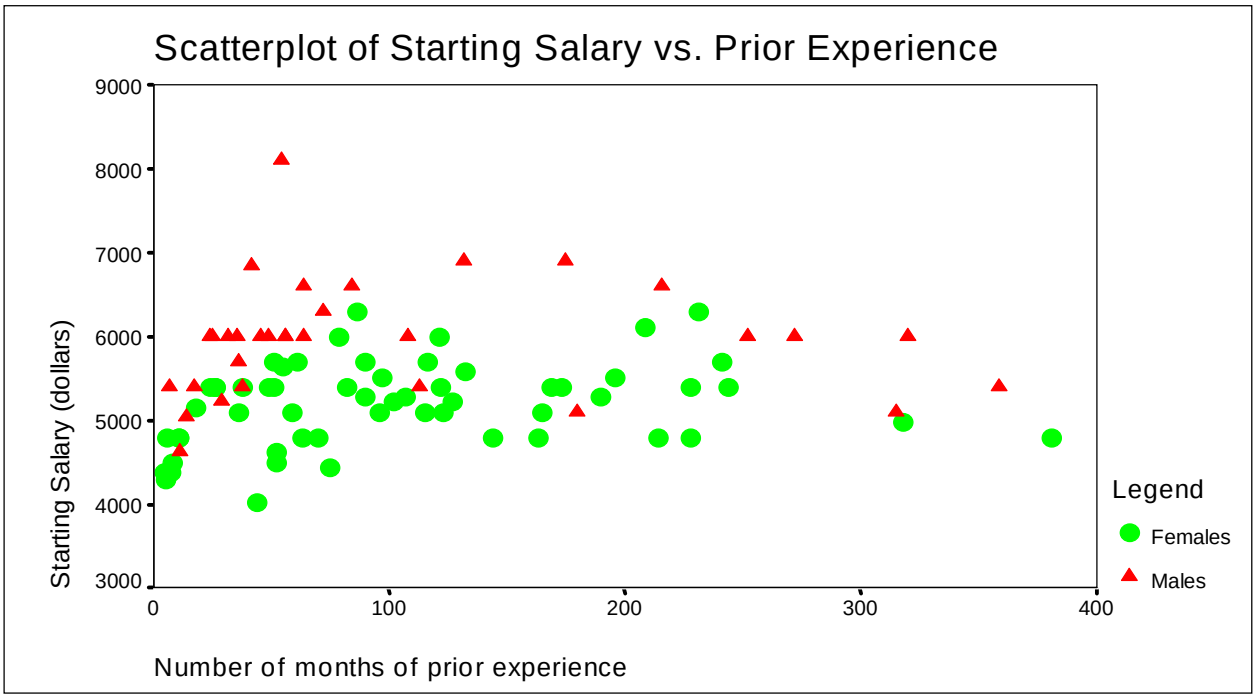
A slow upward drift of salaries over the study period is discernible in the plot. However, the rate of increase is smaller for females. The female starting salaries seem to be rather flat. The spread increases over time for both male and female salaries. On the plot, several males stand out as having much higher salaries than other employees hired at approximately the same time.

Does the scatterplot prove sex discrimination (that males receive higher starting salaries because they are males)? Not necessarily. Although the scatterplot clearly indicates that the males, as a group, received larger starting salaries than the females, we cannot claim that this disparity is attributable to sex discrimination. The scatterplot is consistent with discrimination, but other possible explanations cannot be ruled out; for example, the males may have had more years of education or previous experience.

Notice also that the above plot shows also the change in the gender structure over the time period. Most new clerks hired at the end of the period are females.

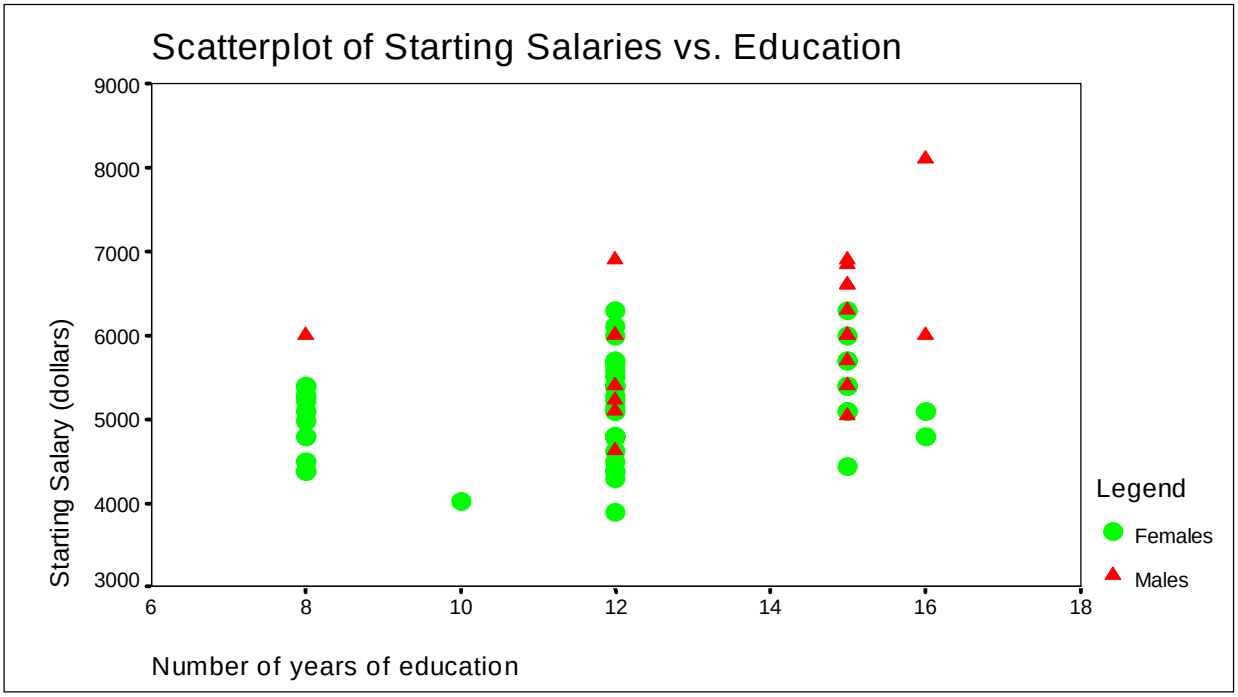
2. Did the bank pay higher starting salaries to men than to women with approximately the same previous experience? In order to answer the question, obtain a scatterplot of starting salaries versus the number of months of prior experience for males and females. Use different marking symbol on the plot to denote male and female subjects.

The following plot shows the scatterplot of salaries versus prior experience:



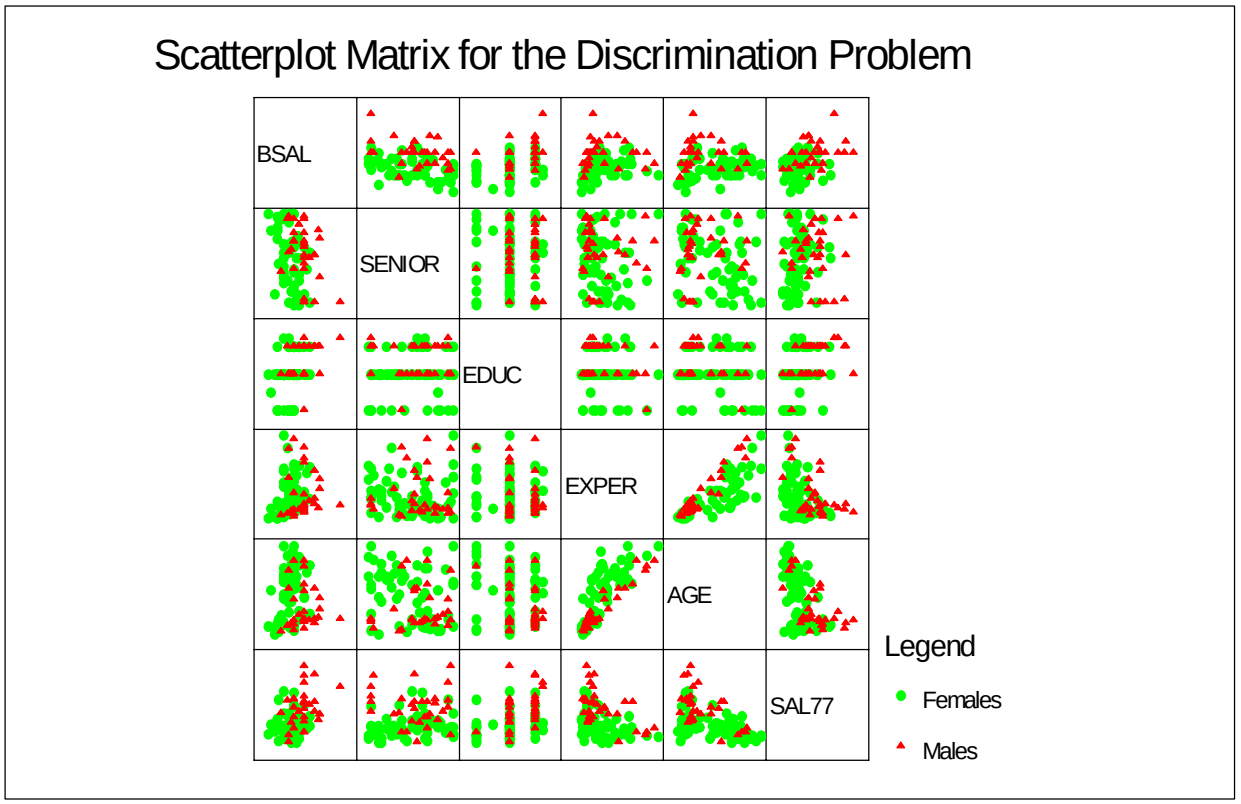
It is clear from the above plot that the males tend to receive higher salaries than females with the same number of months of prior experience. The plot also shows that male employees tend to have less previous experience than females. Since only entry-level jobs are being considered, there is an effect of diminishing returns in the relationship of experience on beginning salary. There is an evident increase of beginning salaries up to about 80 month of prior experience. But then relationship seems to level off. For an entry-level position, very large amounts of experience do not correspond to large beginning salaries.

3. Did the females tend to receive lower starting salaries than similarly educated males? In order to answer the question, obtain a scatterplot of starting salaries versus the number of years of education for males and females.



The starting salaries increase with the number of years of education. The rate of increase is faster for males.

4. Use SPSS to create a scatterplot matrix with the following six variables: Age, EDUC, SAL77, SENIOR, BSAL, EXPER. The variable FSEX (gender) should be treated as a grouping variable.



The scatterplot matrix displays scatterplots for pairs of variables in an array of rows and columns. The variable labels are given on the diagonal. For example, the bottom row shows EXPER versus each of the remaining variables in the various columns, with EXPER on the vertical axis.

We have seen above that females usually have less years of education than males. Is the extent of the disparity between males and females starting salaries justified by this factor? In other words: After accounting for the differences in education background and prior experience, did females tend to receive smaller starting salaries than males?

It is impossible to answer the question using scatterplots. We need statistical tools that make it possible to measure the effects of gender alone on starting salary. These tools are based on multiple regression and will be discussed in one of the future labs.

17.7 Checking the Assumptions Underlying Inferential Methods

Before we use SPSS to obtain confidence intervals and carry out appropriate tests, it is important to explain what the population of interest is and why we use hypothesis testing although the 61 females and 32 males were not selected from any well-defined population.

If we want to establish any sex discrimination pattern in the bank, it is possible to treat the 93 employees of the bank as a random sample from all *potential* employees with the bank. Indeed, the discrimination claim refers not only to the

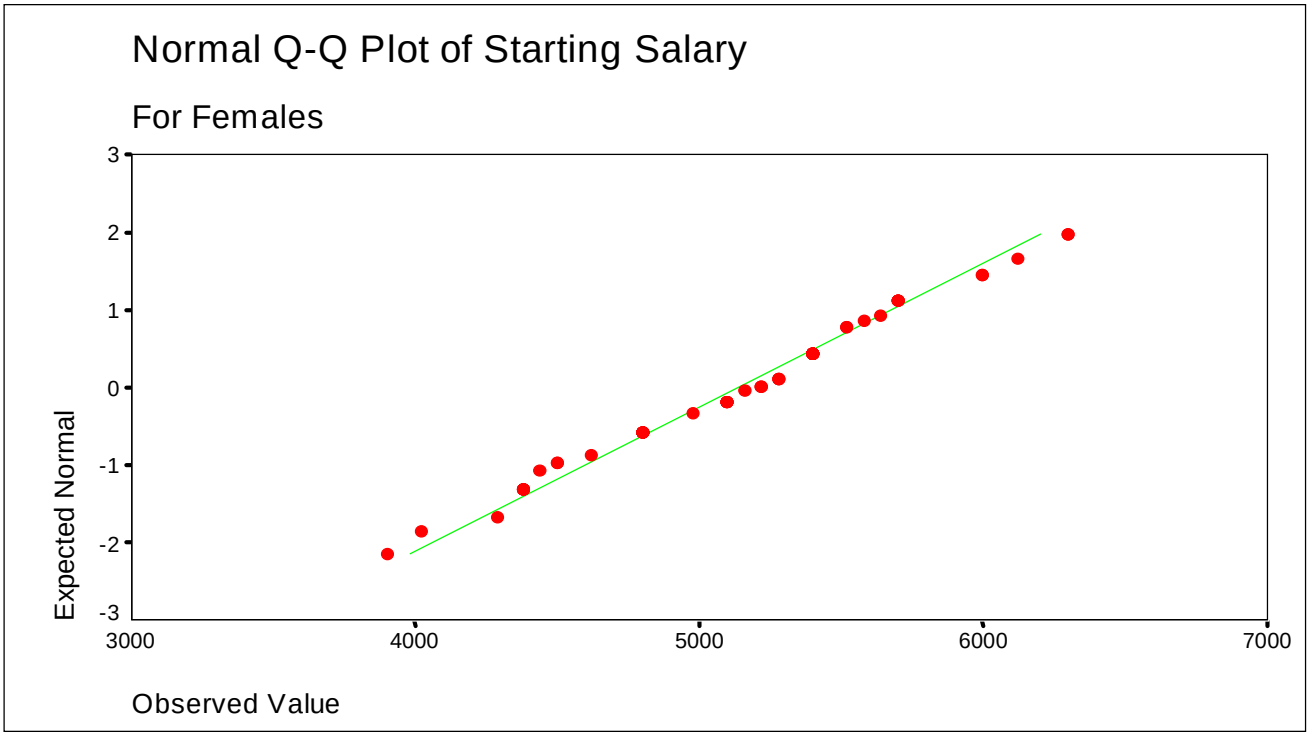
starting salaries of the current employees (with their unique individual characteristics), but it is more general, it refers to the bank's long-term practice for any potential pool of employees. With the meaning of the population, we can make some inferences using SPSS.

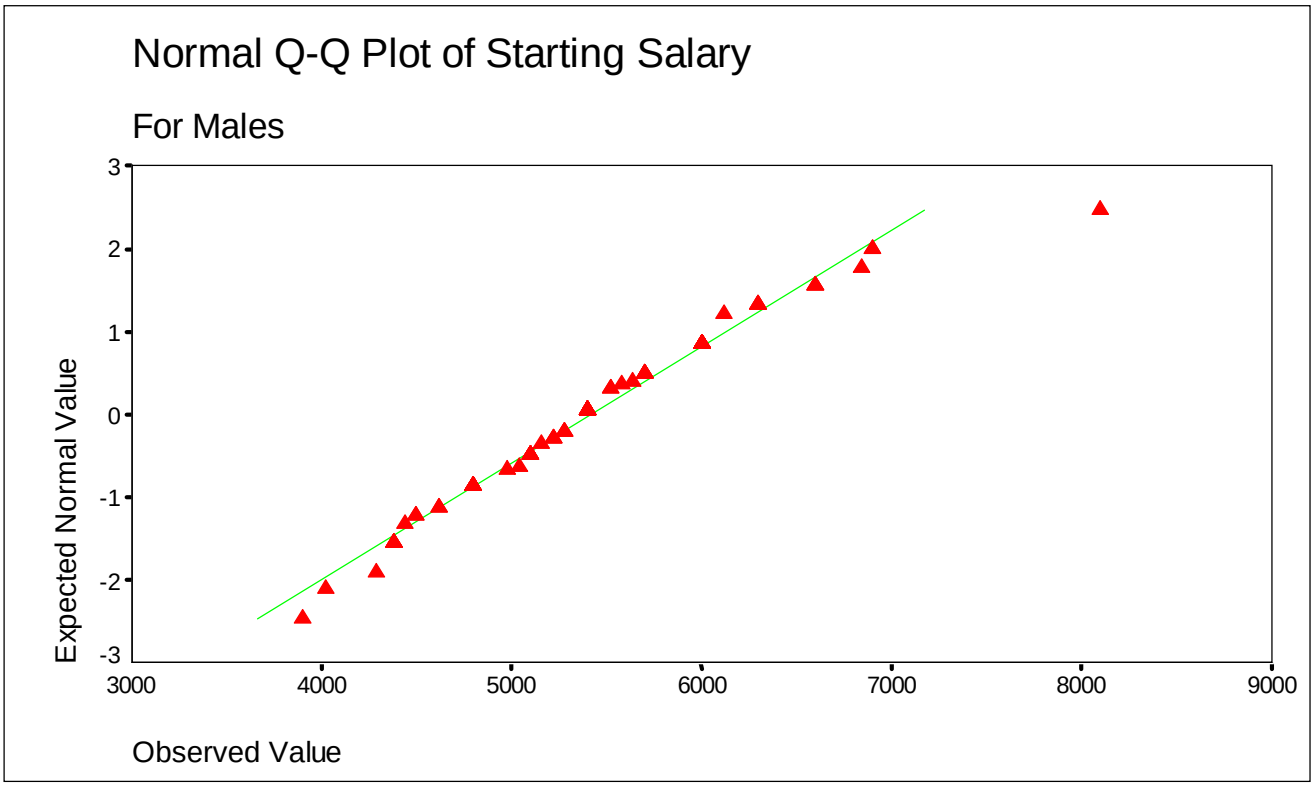
As the population standard deviations are unknown, two-sample t-test should be used to determine whether there are significant differences between the population means. The Independent-Samples T Test procedure available in SPSS compares the means of one variable (beginning salary) for two groups of cases (males, females).

The assumptions of the procedure are that both samples are random samples from their respective populations, the two samples are independent of one another, and both populations are normal.

In order to determine whether or not a variable is normally distributed, you can use one of the two available procedures in SPSS: *Normal Q-Q* Plot or *Normal P-P* Plot. The *Normal Q-Q* plot plots the quantiles of a variable's distribution against the quantiles of the normal distribution. If the data come from a normal distribution, the plot should resemble a straight line. The *Normal P-P* plot plots the cumulative proportions of a variable's distribution against the cumulative proportions of the normal distribution. Similarly, if the sample is from a normal distribution, points will cluster around a straight line.

The normal probability plot (Normal Q-Q plot) for each gender is displayed below.





Both plots do not indicate any significant departure from a straight line. Thus, there is no reason to suspect that the assumption of normality is violated.

17.8 Tests of Significance and Confidence Intervals

We will apply the independent samples t-test for the male and female observations in the three-year period. SPSS produces the following output:

t-tests for Independent Samples of FSEX

Variable	Number of Cases	Mean	SD	SE of Mean
BSAL				
Females	61	5138.8525	539.871	69.123
Males	32	5956.8750	690.733	122.106

Mean Difference = -818.0225

Levene's Test for Equality of Variances: F= .344 P= .559

t-test for Equality of Means

Variances	t-value	df	2-Tail Sig	SE of Diff	95% CI for Diff
Equal	-6.29	91	.000	129.997	(-1076.25, -559.799)
Unequal	-5.83	51.33	.000	140.313	(-1099.67, -536.376)

The output starts with statistics of the two groups, followed by the value of the difference between means. The Levene test for equality of variances is also included. Provided the F value is not significant (P>0.05), the variances can be assumed to be equal and the Equal Variances line of values for the t-test can be used. If P<0.05, then the equality of variances assumption has been violated and the t-test based on unequal variances should be used.

In our case, the high P-value of 0.559 in the Levene's Test for equality of variances strongly indicates that the sample data are consistent with the equality variances assumption. The P-value of the two-sided t-test for the equality of means is obtained by SPSS as zero. Hence, one-sided p-value is zero as well. That means that there is a very strong evidence that the population mean starting salaries are higher for males. The mean starting salary for males is estimated to be $559.80 to $1076.25 larger than the mean starting salary for females.

We found in Section 5 that in general starting salaries increased over the three-year period considered in the study. Hence, it makes sense to carry tests of significance to compare the average starting salaries of males and females for each of the three one-year periods. This approach is reflecting changes in mean starting salary over the three-year period. The results of the analysis are presented in **Section 8**.

17.9 Summary

Descriptive and inferential methods discussed above showed that in general the males received higher starting salaries than females hired at the same time. On the other hand, the males generally did have more years of education than the females, and this, not gender, may have been responsible for the observed differences in the starting salaries for males and females.

Is the extent of the disparity between males and females starting salaries justified by this factor? In general: After accounting for the differences in education background and other given measures of qualification, did females tend to receive smaller starting salaries than males?

It is impossible to answer the question using the statistical techniques discussed above. Scatterplots certainly prove that the males tend to receive higher salaries than females but they are not able to show how much of the disparity can be accounted by the differences in available measures of qualification.

We need statistical tools that make it possible to measure the effects of gender alone on starting salary. These tools are based on multiple regression and will be discussed in one of the future labs. You will see that the multiple regression techniques enable us to isolate the effects of gender on starting salaries from other available measures of qualification.

In general, some confounding variables may not be recognized or measured and these variables consequently cannot be accounted for the observed disparity. Obviously, it would be difficult or even impossible to consider all variables affecting starting salary. Therefore, it may be possible to conclude that males tend to receive larger starting salaries than females, even after accounting for all

available factors, and still not be possible to conclude, from the statistics alone, that this happens because they are males.

With the study design we are not able to prove sex discrimination because we are not able to isolate all confounding variables to see the effects of gender alone on starting salaries. Although statistical analysis alone is not able to prove sex discrimination, it can be useful in a court of law to establish discrimination.