

The goal of this project is to perform an exploratory analysis of a dataset using SPSS.

Data description

The data set gss2004.xls contains 575 observations on 5 variables:

SEX = Respondent's sex – {1 for Male, 2 for Female}

AGE = Age of respondent

WWWHR = Hours on the WWW per week for Internet users

NEWS30 = Respondent has used news site in the past 30 days (1= "never", 2 = "1-2 times", 3="3-5times" 4="more than 5 times")

EMAILHR = Hours of e-mail per week for Internet users

The data were collected from the 2004 General Social Survey for adult respondents (18 years of age or older), living in the United States. The GSS is one of the largest and longest projects that have been conducted to monitor social change and the growing complexity of American society (see <http://www.norc.org> for more information).

The analysis described below will study the number of hours spent by Internet users using email. The study will also explore whether men and women use email differently.

Case Processing Summary

	sex	Valid		Missing		Total	
		N	Percent	N	Percent	N	Percent
emailhr	Male	239	100.0%	0	0.0%	239	100.0%
	Female	336	100.0%	0	0.0%	336	100.0%

Descriptive

sex	Statistic	Std. Error
Male	Mean	6.33
	95% Confidence Interval for Mean	Lower Bound: 5.12, Upper Bound: 7.54
	5% Trimmed Mean	4.85
	Median	2.00
	Variance	90.364
	Std. Deviation	9.506
	Minimum	0
	Maximum	50
	Range	50
	Interquartile Range	7
	Skewness	2.658
Kurtosis	7.585	.314
Female	Mean	5.93
	95% Confidence Interval for Mean	Lower Bound: 4.98, Upper Bound: 6.88
	5% Trimmed Mean	4.57
	Median	2.00
	Variance	78.924
	Std. Deviation	8.884
	Minimum	0
	Maximum	50
	Range	50
	Interquartile Range	6
	Skewness	2.675
Kurtosis	7.834	.265

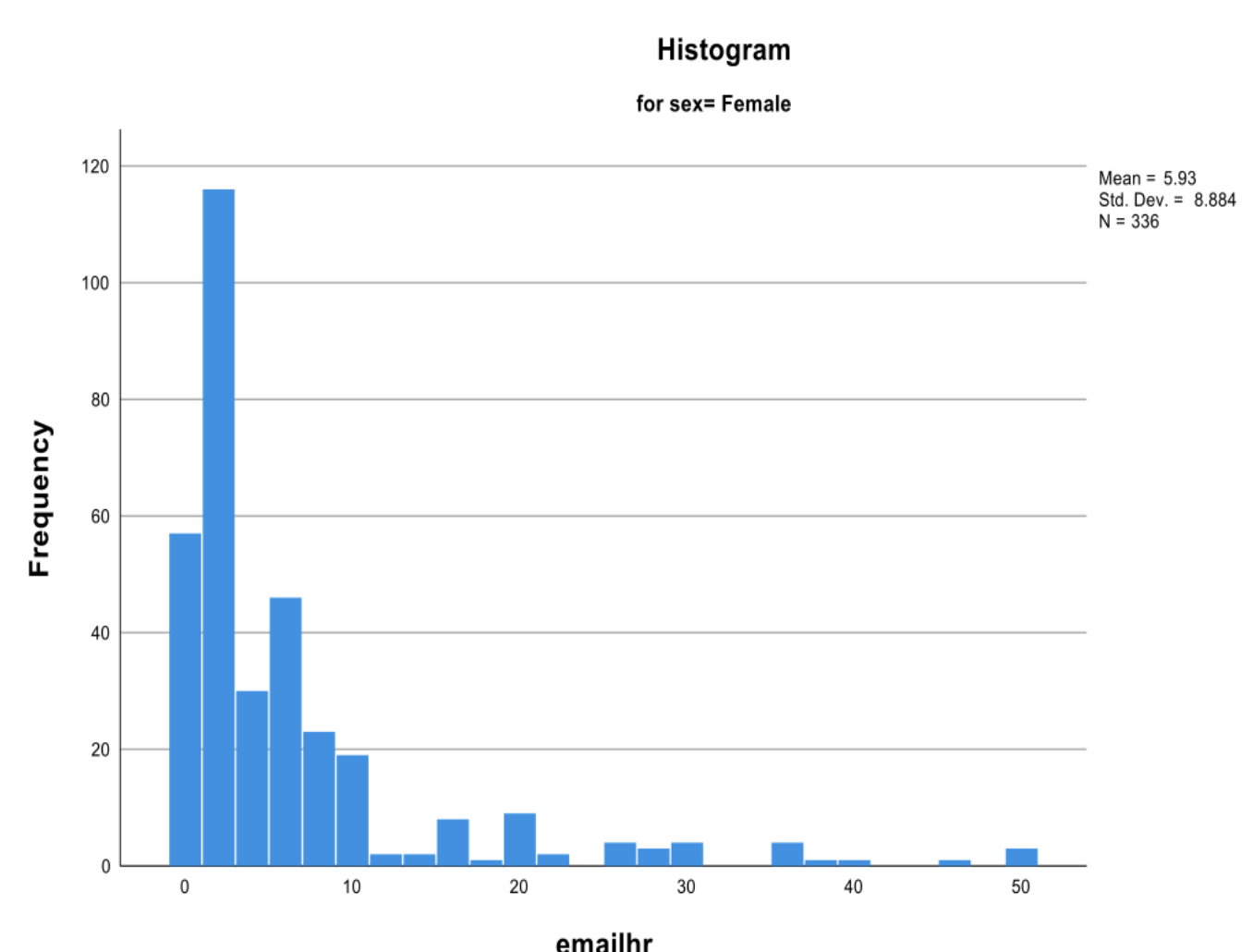
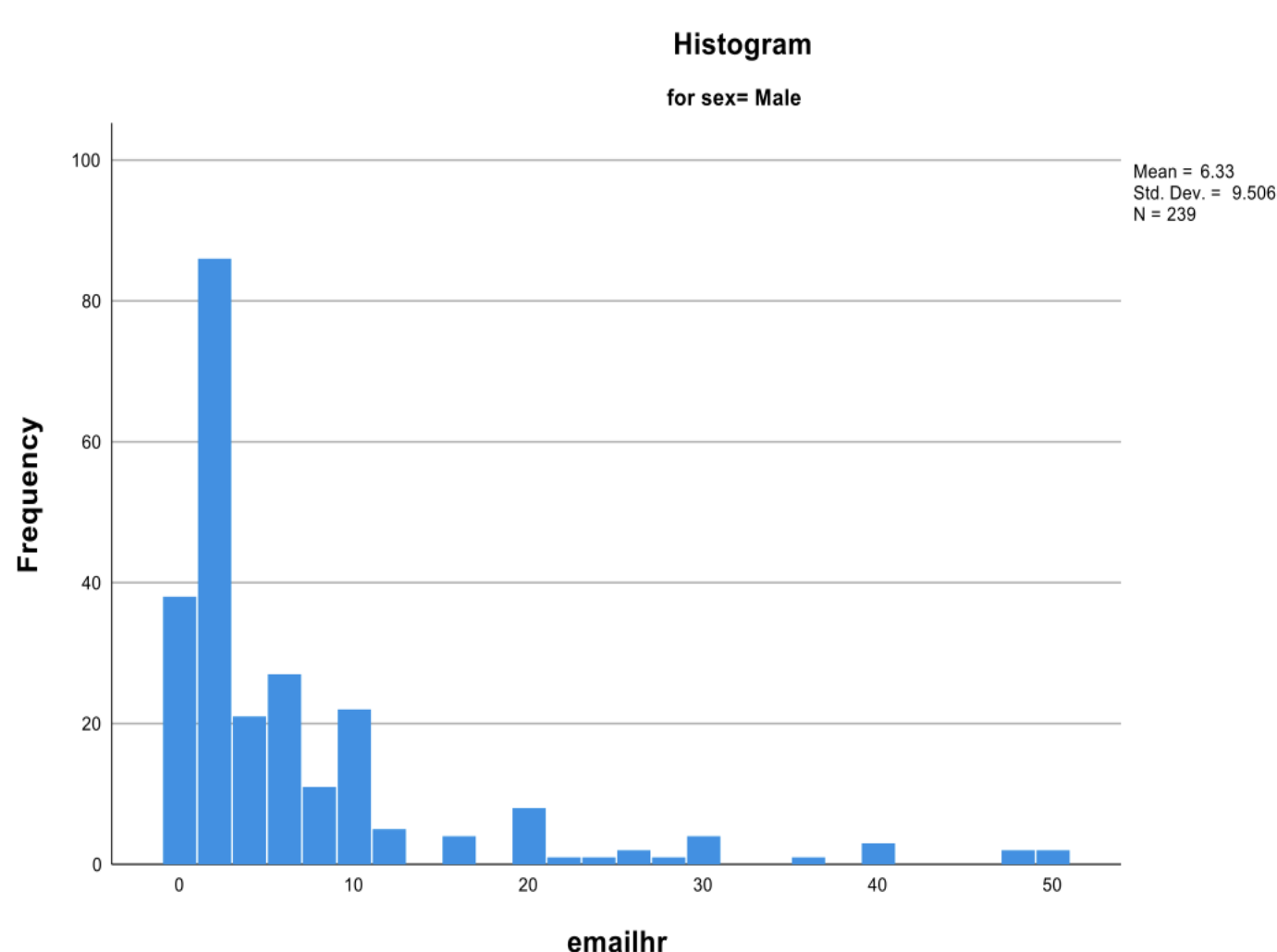
Percentiles

sex	5	10	25	50	75
Weighted Average(Definition 1)	Male: .00	Male: .00	Male: 1.00	Male: 2.00	Male: 8.00
	Female: .00	Female: .00	Female: 1.00	Female: 2.00	Female: 7.00
Tukey's Hinges	Male: .00	Male: .00	Male: 1.00	Male: 2.00	Male: 7.50
	Female: .00	Female: .00	Female: 1.00	Female: 2.00	Female: 7.00

Tests of Normality

sex	Statistic	df	Sig.	Shapiro-Wilk	
				Statistic	df
emailhr Male	.253	239	<.001	.649	239
Female	.262	336	<.001	.651	336

Gender	Variables	Mean	S.D.	Max	Min	Median	First quartile	Third quartile
Male	EMAILHR	6.33	9.506	50	0	2	1	8
Female	EMAILHR	5.93	8.884	50	0	2	1	7



Shape:

The amount of emailing time for males and women is right-skewed (Positive right-skewed). This means that the majority of people spend about 10 hrs on reading/writing emails per week, however, we have very few people who spend up to 50 hrs per week.

Center:

The median shows that 50%(Q2) of the people spend 2 hrs on emails.

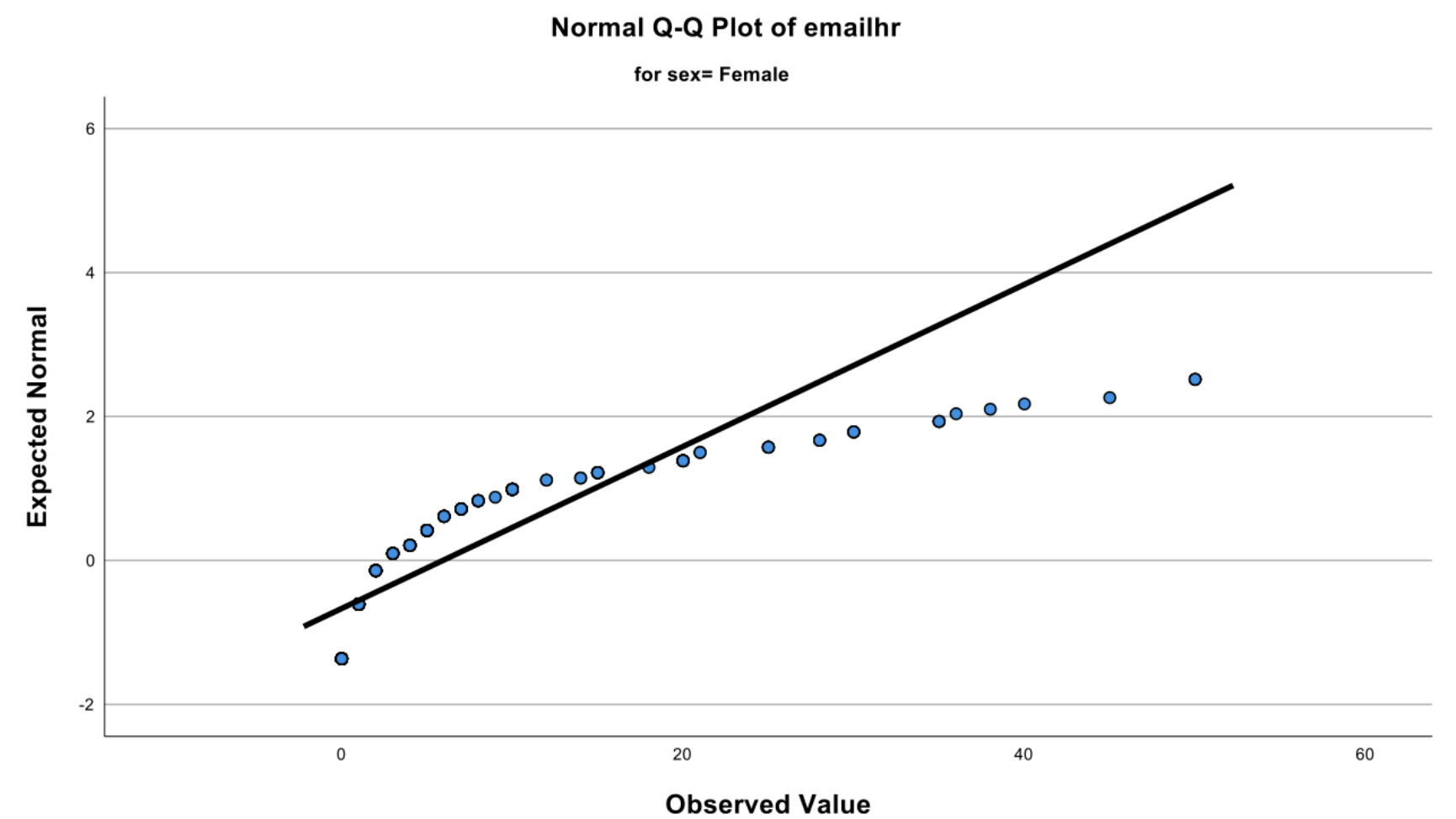
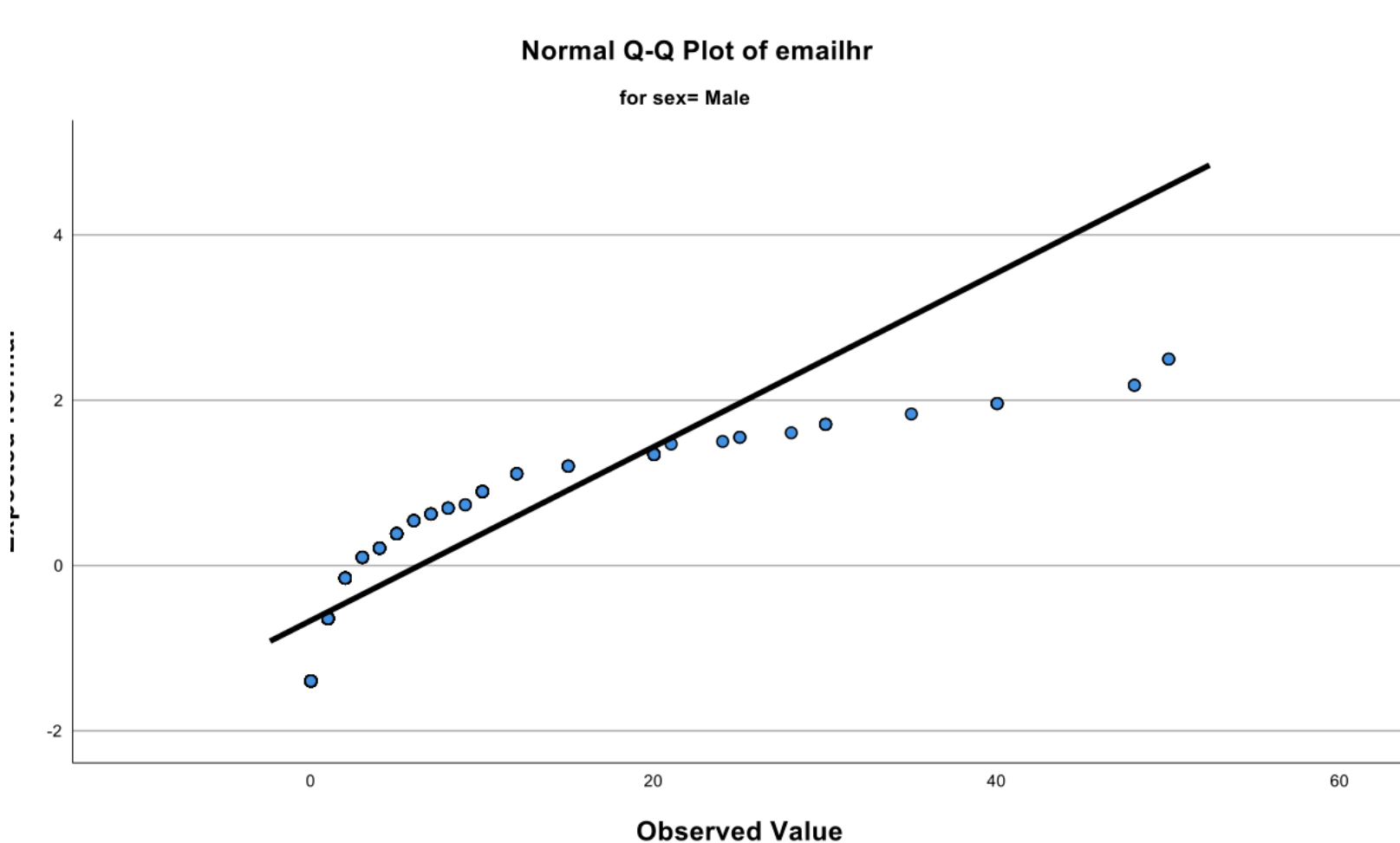
Distribution:

The Inter-Quartile Range (IQR = Q3-Q1), how spread the data is distributed. The IQR is 7 for men when compared to the median we use which is very large, meaning that there is a huge discrepancy among the email hours. Comparing the min=0 and max=50 email times it became evident that the emailing time varies widely.

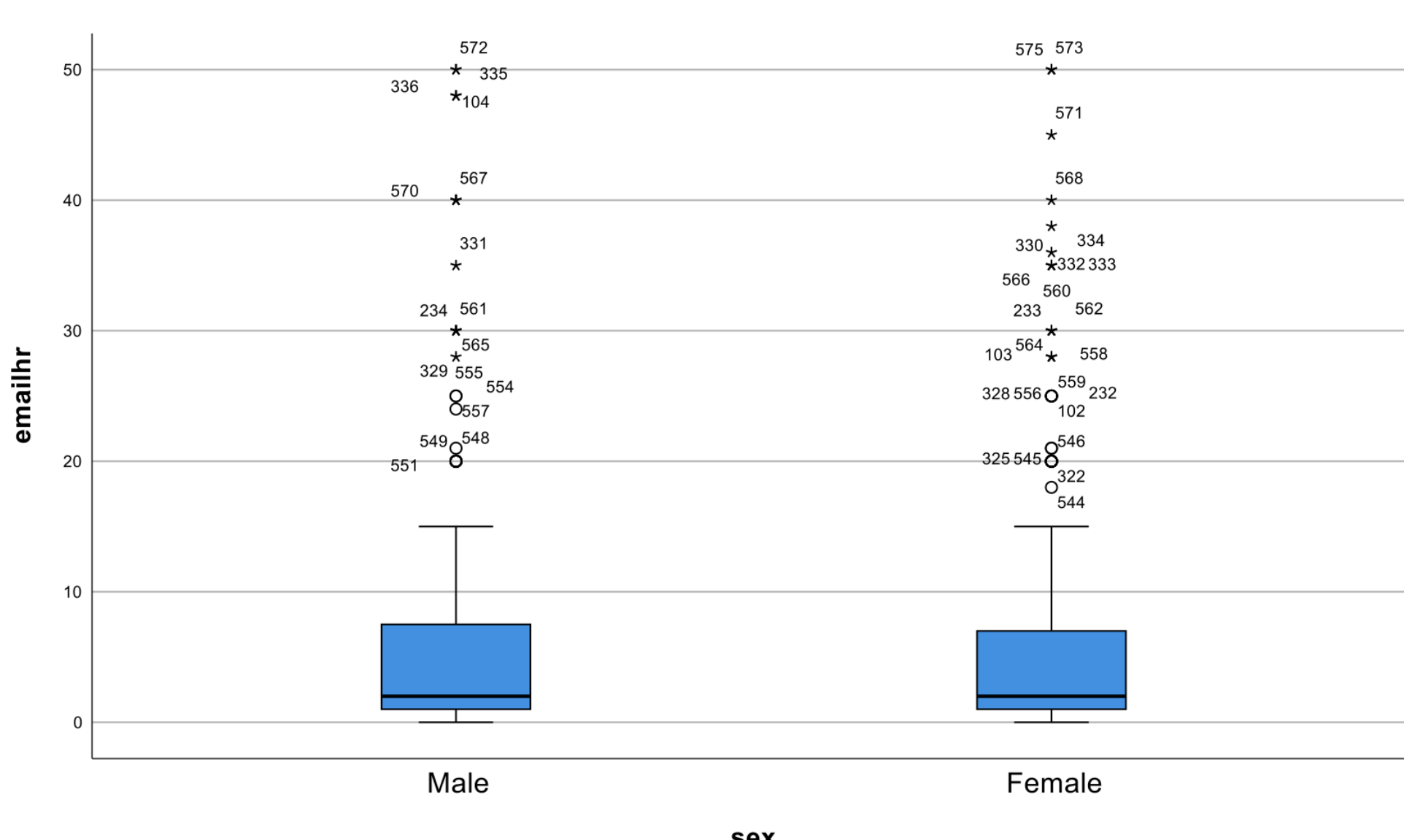
Since the distribution is right-skewed, I will use the median to describe the center.

The five-number summary shows 25% of the people will spend 1hr or less on emails per week while 50% will spend 2hrs or less and, 75% will spend 8hrs or less. The minimum amount of time spent by both males and females is 0hr, i.e., less than 60mins per week and the max time spent is 50hrs per week. According to the five-number summary, 25% of the people spend one hour or less on emails per week, while 50% spend two hours or less and 75% spend eight hours or less. The minimum amount of time spent on emails by both males and females is zero hours, i.e., less than 60 minutes per week, and the maximum time spent is 50 hours per week.

Normal Q-Q Plots



The graph indicates that the amount of time spent on writing and reading emails by both men and women does not follow a normal distribution. This is evidenced by the fact that the data points are not aligned along the diagonal line and there are numerous outliers present. The box plot further confirms this observation, as it reveals the presence of outliers beyond 20 hours of email usage per week for both genders.



For Male –

Upper outlier = Q3 + (1.5 * IQR) = 8 + (1.5 * 7) = 18.5 (18.5 hr and above are outliers)

Lower outlier = Q1 - (1.5 * IQR) = 1 - (1.5 * 7) = -9.5, because this is negative and time is not negative.

Based on the calculation above, there is no lower outlier as the value is negative and negative time is not possible. However, it is highly unusual for a person to spend more than 18 hours and 30 minutes reading or writing emails. If a man spends more time than this, it is considered atypical.

For Female-

Upper outlier = Q3 + (1.5 * IQR) = 7 + (1.5 * 6) = 16

Lower outlier = Q1 - (1.5 * IQR) = 1 - (1.5 * 6) = -8, because this is negative and time is not negative.

According to the calculation above, there are no lower outliers as negative time is not possible. However, spending more than 16 hours on email is unusual and is considered atypical for anyone.

The median time spent by both males and females on email per week is 2 hours, with a maximum usage of 50 hours. While there is no significant difference between the time spent by males and females, the extreme cut-off for males is one hour more than that of females.