**Quantitative Variables: Summary Statistics Center and Spread**

1. **Five Number Summary and Boxplots**

For a set of numerical data:

Median

First Quartile (Q1)

Third Quartile (Q3)

Min

Max

Interquartile Range (IQR)

Five Number Summary: min Q1 med Q3 max

**By Hand Example** 3,5,2,3,7,5,8,6,7

Five Number Summary:

IQR:

Boxplot:

**Five Number Summary Using R** (for the variable VAR in the data frame FRAME)

quantile(~VAR, data=FRAME)

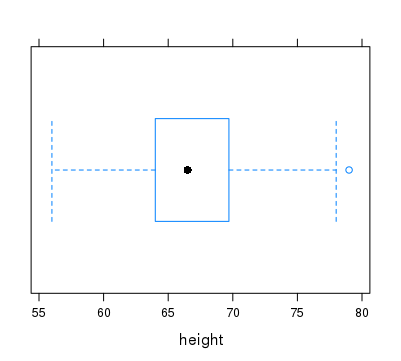
Find the Five Number Summary for the heights of the adult children in the data frame Galton.

> quantile(~height,data=Galton)

0% 25% 50% 75% 100%

56.0 64.0 66.5 69.7 79.0

> bwplot(~height, data= Galton)



Boxplot of a data set skewed to right

Boxplot of a data set skewed to the left

1. **Mean, Standard Deviation**

**Generic notation for quantitative data**

n = number of data values

x = variable

x1, x2, …, xn = data values

**Mean**



**Example:** x: 1,1,3,4,6,7,7,9

**NOTES** 1. If the data is symmetric:

2. If the data is skewed to the right:

3. If the data is skewed to the left:

4. Effect of outliers:

**Variance and Standard Deviation**

variance of x = 

standard deviation of x = =

**Example**: x: 1,2,3,4,5

**Computing summary statistics with R**

mean(~VAR, data = FRAME)

var(~VAR, data = FRAME)

sd(~VAR, data = FRAME)

median(~VAR, data = FRAME)

quantile(~VAR, data = FRAME)

IQR(~VAR, data = FRAME)

favstats(~VAR, data = FRAME)

> mean(~Assaults,data=USArrests)

[1] 7.788

> var(~Assaults,data=USArrests)

[1] 18.97047

> sd(~Assaults,data=USArrests)

[1] 4.35551

> median(~Assaults,data=USArrests)

[1] 7.25

> quantile(~Assaults,data=USArrests)

0% 25% 50% 75% 100%

0.800 4.075 7.250 11.250 17.400

> IQR(~Assaults,data=USArrests)

[1] 7.175

> favstats(~Assaults,data=USArrests)

min Q1 median Q3 max mean sd n missing

0.8 4.075 7.25 11.25 17.4 7.788 4.35551 50 0

**Example:** The data frame **HELPrct** contains data on an experiment testing a treatment of substance abusers. It is a “large” data frame. We can get the top “few” lines of the data frame using the command **head.**

age anysubstatus anysub cesd d1 daysanysub dayslink drugrisk e2b female sex g1b homeless

1 37 1 yes 49 3 177 225 0 NA 0 male yes housed

2 37 1 yes 30 22 2 NA 0 NA 0 male yes homeless

3 26 1 yes 39 0 3 365 20 NA 0 male no housed

4 39 1 yes 15 2 189 343 0 1 1 female no housed

5 32 1 yes 39 12 2 57 0 1 0 male no homeless

6 47 1 yes 6 1 31 365 0 NA 1 female no housed

i1 i2 id indtot linkstatus link mcs pcs pss\_fr racegrp satreat sexrisk substance

1 13 26 1 39 1 yes 25.111990 58.41369 0 black no 4 cocaine

2 56 62 2 43 NA <NA> 26.670307 36.03694 1 white no 7 alcohol

3 0 0 3 41 0 no 6.762923 74.80633 13 black no 2 heroin

4 5 5 4 28 0 no 43.967880 61.93168 11 white yes 4 heroin

5 10 13 5 38 1 yes 21.675755 37.34558 10 black no 6 cocaine

6 4 4 6 29 0 no 55.508991 46.47521 5 black no 5 cocaine

treat

1 yes

2 yes

3 no

4 no

5 no

6 yes

**Summary statistics by groups**

We can get summary statistics about age for each of the three abuse groups.

> favstats(~age,groups=substance,data=HELPrct)

substance min Q1 median Q3 max mean sd n missing

1 alcohol 20 33 38.0 43.00 58 38.19774 7.652272 177 0

2 cocaine 23 30 33.5 37.25 60 34.49342 6.692881 152 0

3 heroin 19 27 33.0 39.00 55 33.44355 7.986068 124 0

> favstats(~age |substance,data=HELPrct)

substance min Q1 median Q3 max mean sd n missing

1 alcohol 20 33 38.0 43.00 58 38.19774 7.652272 177 0

2 cocaine 23 30 33.5 37.25 60 34.49342 6.692881 152 0

3 heroin 19 27 33.0 39.00 55 33.44355 7.986068 124 0

> favstats(age~substance,data=HELPrct)

substance min Q1 median Q3 max mean sd n missing

1 alcohol 20 33 38.0 43.00 58 38.19774 7.652272 177 0

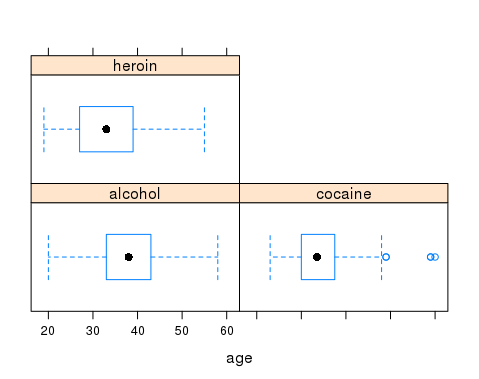
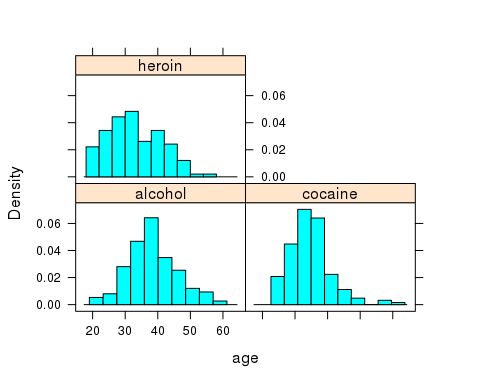
2 cocaine 23 30 33.5 37.25 60 34.49342 6.692881 152 0

3 heroin 19 27 33.0 39.00 55 33.44355 7.986068 124 0

**Histograms and boxplots by groups**

> histogram(~age|substance,data = HELPrct)

> bwplot(~age|substance,data = HELPrct)



**Quantitative vs Categorical Variables**

A quantitative variable is a variable that takes on numerical values in such a way that it makes sense to perform arithmetic operations on its values. A categorical variable is a variable whose values are categories. Length (in inches) is a quantitative variable; gender (male/female) is a categorical variable. Sometimes numbers are used to represent categories. E.g., 1 = male and 2 = female. Using numbers to represent categories does not turn a categorical variable into a quantitative variable.

The data frame KidsFeet contains data on thirty-nine 4th grade children

> head(KidsFeet)

name birthmonth birthyear length width sex biggerfoot domhand

1 David 5 88 24.4 8.4 B L R

2 Lars 10 87 25.4 8.8 B L L

3 Zach 12 87 24.5 9.7 B R R

4 Josh 1 88 25.2 9.8 B L R

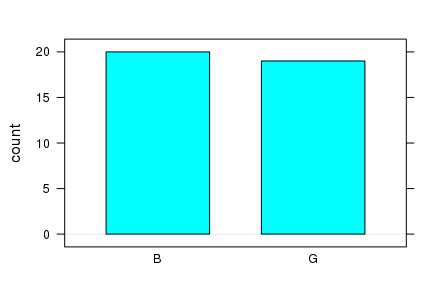
5 Lang 2 88 25.1 8.9 B L R

6 Scotty 3 88 25.7 9.7 B R R

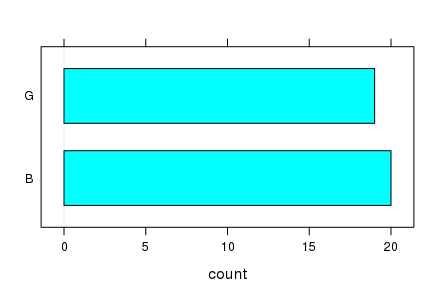
**Categorical Variables**

Distributions of categorical variables are given in terms of the number (frequency) of individuals in each category or by the percentages (relative frequencies) of individuals that fall into each of the categories. The standard picture is a bargraph.

* bargraph(~sex,data=KidsFeet)



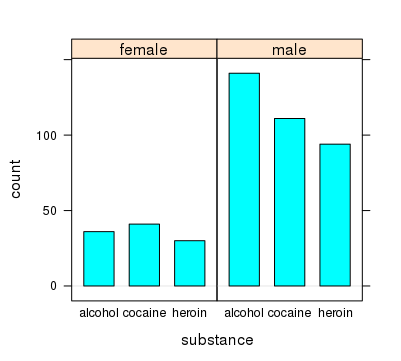
* bargraph(~sex,data=KidsFeet,horizontal=TRUE)



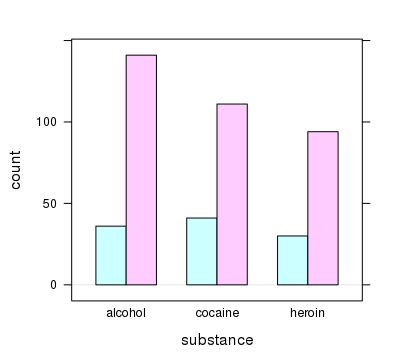
**Grouping with categorical variables**

**Compare how the type of abuse is distributed with respect to gender (sex)**

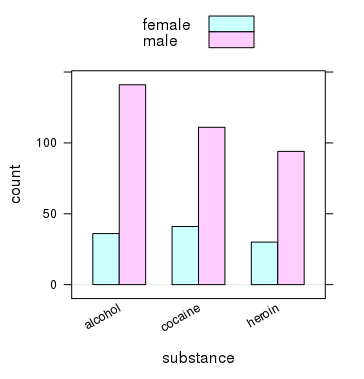
> bargraph(~substance | sex, data = HELPrct)

****

> bargraph(~substance, groups =sex, data = HELPrct)

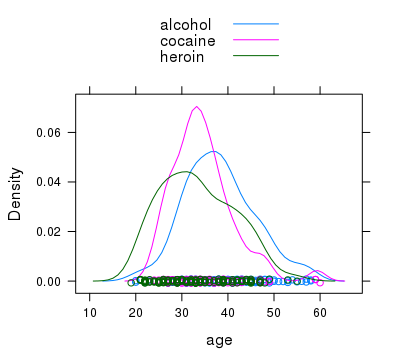
****

> bargraph(~substance, groups =sex, data = HELPrct, auto.key=TRUE)

****

**Compare the distribution of age for the three types of abuse.**

> densityplot(~age,groups=substance, data = HELPrct, auto.key = TRUE)

****

EXERCISES 2

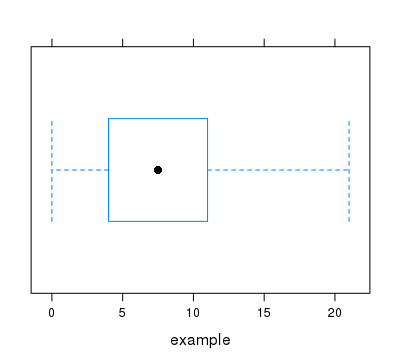
2.1 In the data frame HELPrct, the variable **i1** gives the average number of drinks per day, **substance** gives the kind of substance abuse, and **sex** gives the gender.

1. Create and print out histograms comparing the average number of drinks per day by gender,
2. Repeat (a) with boxplots.
3. Compute the means and standard deviations for the daily number of drinks for male and for females.
4. Based on the information in (c) would you conclude that, on average, the males in the study consume more alcohol than the females? For which gender is there more variability in alcohol consumption?
5. Create and print out **frequency** histograms comparing the alcohol consumption by type of substance abuse. Is the distribution of alcohol consumption symmetric for those whose are alcohol abusers? If it is skewed, in which direction is it skewed?
6. Compute the median alcohol consumption for each of the three types of abuse.

2.2 Compute the mean, variance and standard deviation for the data set x: 1,5,3,7,9 **by HAND.**

2.3 Create a set of 6 numbers in the range 0 – 10 (inclusive) that will have the largest possible standard deviation.

2.4 Below is the boxplot of a data set.



1. What is the median of the data set?
2. Is the mean of the data set equal to, smaller than, or larger than the median?

2.5 From the data frame HELPrct, create and print out a bar graph picture that shows how homelessness (VAR = homeless) is distributed with respect to gender (VAR = sex). Produce the version where the bars are adjacent and the color codes for the gender categories are above the bar graph. Based on the picture, should you conclude that males and females in the study are homeless at the same rate?