

Stat 505 Assignment 1 Solutions

1. Sequences

- (a) Integers from 128 to 143. There are 16 of them

```
> options(width = 90)
> 128:143
 [1] 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143
> length(128:143)
[1] 16
```

- (b) integers from 100 to 80. There are 21 of them

```
> 100:80
 [1] 100 99 98 97 96 95 94 93 92 91 90 89 88 87 86 85 84 83 82 81 80
> length(100:80)
[1] 21
```

- (c) nine equally spaced real numbers from 0 to π .

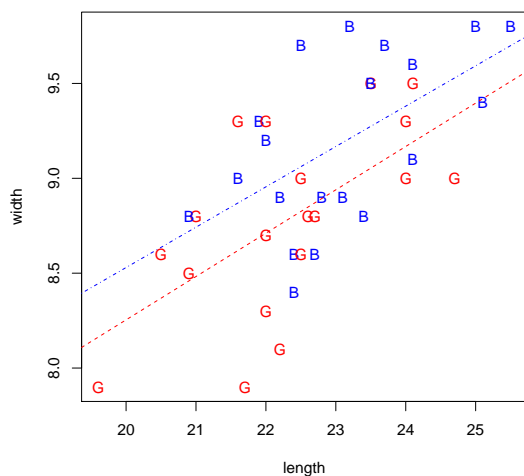
```
> pi.seq <- seq(0, pi, length = 9)
> round(rbind(pi.seq, zapsmall(sin(pi.seq)), zapsmall(cos(pi.seq))), 4)
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]
pi.seq  0 0.3927 0.7854 1.1781 1.5708 1.9635 2.3562 2.7489 3.1416
      0 0.3827 0.7071 0.9239 1.0000 0.9239 0.7071 0.3827 0.0000
      1 0.9239 0.7071 0.3827 0.0000 -0.3827 -0.7071 -0.9239 -1.0000
```

I used `rbind` to bind them together as rows of a matrix. The first and last (0 and π) have sin of 0, the middle one, $\pi/2$, has cos 0.

2. Kids feet data set was collected to see if the length-width relationship of feet varies between 4th grade boys and girls.

- (a) Data input and plot:

```
> kidsfeet <- read.table("http://www.amstat.org/publications/jse/datasets/kidsfeet.dat")
> names(kidsfeet) <- c("month", "year", "length", "width", "sex", "foot",
+ "hand")
> plot(width ~ length, data = kidsfeet, pch = as.character(sex), col = 6 -
+ unclass(sex) * 2)
> feet.fit <- lm(width ~ 0 + sex + length:sex, data = kidsfeet)
> abline(coef(feet.fit)[c(1, 3)], col = "blue", lty = 4)
> abline(coef(feet.fit)[c(2, 4)], col = "red", lty = 2)
```



- (b) Relationships: *Boys feet are generally longer and wider, but the relationship between length and width seems to be the same for both sexes.*
- (c) A summary of the model gives

```
> (foot.fit.summary <- summary(lm(width ~ sex * length, data = kidsfeet)))$coef)
              Estimate Std. Error    t value    Pr(>|t|)
(Intercept)  4.27732807  1.70215249   2.5128936  0.016732500
sexG         -0.59223769  2.29937443  -0.2575647  0.798251230
length       0.21262376  0.07357345   2.8899521  0.006574429
sexG:length  0.01582067  0.10096313   0.1566975  0.876383655
```

0.0158 with standard error 0.101. The variance estimate has 35 degrees of freedom, so a 95% CI for the difference in slope is:

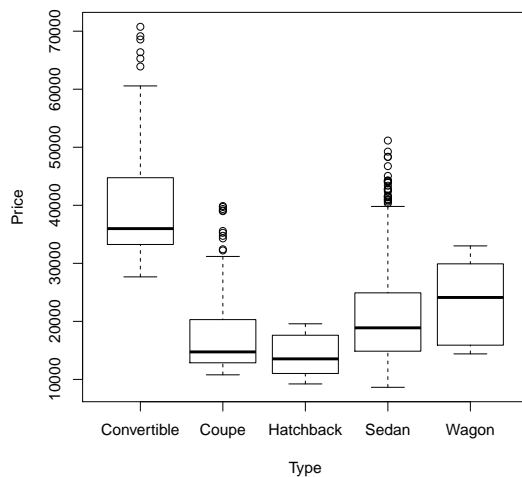
$$0.01582 + c(-1,1) * qt(.975,35) * 0.10096 = (-0.189, 0.221).$$

The interval contains 0, so we cannot reject $H_0 : \Delta_{\text{slope}} = 0$ at the $\alpha = .05$ significance level. A model with parallel lines fits well, so I conclude that the relationship between length and width is the same for boys and girls.

3. 2005 used GM cars

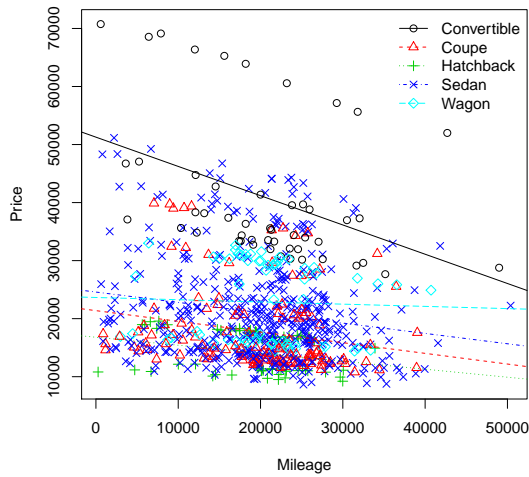
- (a) Price versus type.

```
> cars <- read.csv("../data/kuiper.csv", head = T)
> plot(Price ~ Type, data = cars)
```



- (b) Price versus mileage.

```
> plot(Price ~ Mileage, data = cars, pch = unclass(Type), col = unclass(Type))
> for (typ in 1:5) abline(lm(Price ~ Mileage, data = cars, subset = unclass(Type) ==
+   typ), lty = typ, col = typ)
> legend("topright", levels(cars$Type), pch = 1:5, col = 1:5, lty = 1:5,
+   bty = "n")
```

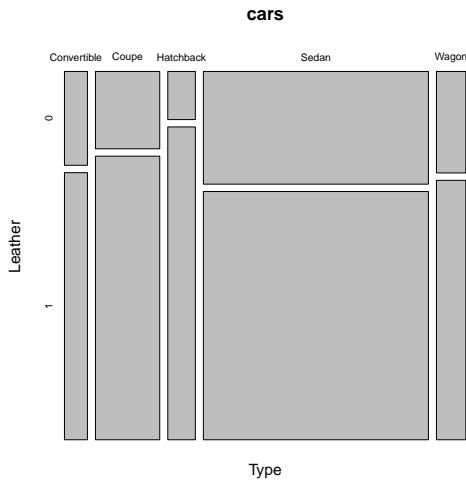


In general, price declines with the mileage of the used car. However, there are huge anomalies in these data, like the line of convertibles at the top are all Cadillac XLR-V8 Hardtops. This is clearly not a random sample of used cars.

(c) Type versus leather.

```
> mosaicplot(Type ~ Leather, data = cars)
> prop.table(with(cars, table(Type, Leather)), 1)
```

Type	Leather	
	0	1
Convertible	0.2600000	0.7400000
Coupe	0.2142857	0.7857143
Hatchback	0.1333333	0.8666667
Sedan	0.3122449	0.6877551
Wagon	0.2812500	0.7187500



It seems quite odd that over half the cars in each category have leather seats – even hatchbacks and wagons.

4. Superbowl data.

(a) The file provided has several problems. Here's how I addressed them:

- The first three dates had quotes around them. I removed those.

- Scores were written together separated by a strange dash. I did a global replace to change the dash to a space.
- Attendance had a comma. I removed those again with global replace.
- Now the header row did not have the right number of entries. I changed it to
number month date year winner w.score l.score loser venue city attendance

(b) Reading the data into an R dataframe.

```
> super <- read.table("../data/mySuperBowl.data", head = T)
```

(c) Two more columns are needed to show (1) the total score and (2) the winning margin (also called point spread).

```
> super$total <- super$w.score + super$l.score
> super$spread <- super$w.score - super$l.score
```

(d) A summary of the data frame.

```
> summary(super)
```

number	month	date	year	winner
Min. : 1	February: 9	Min. : 1.00	Min. :1967	Pittsburgh Steelers : 6
1st Qu.:12	January :36	1st Qu.:12.00	1st Qu.:1978	Dallas Cowboys : 5
Median :23		Median :20.00	Median :1989	San Francisco 49ers : 5
Mean :23		Mean :18.51	Mean :1989	Green Bay Packers : 4
3rd Qu.:34		3rd Qu.:26.00	3rd Qu.:2000	New England Patriots: 3
Max. :45		Max. :31.00	Max. :2011	New York Giants : 3
				(Other) :19

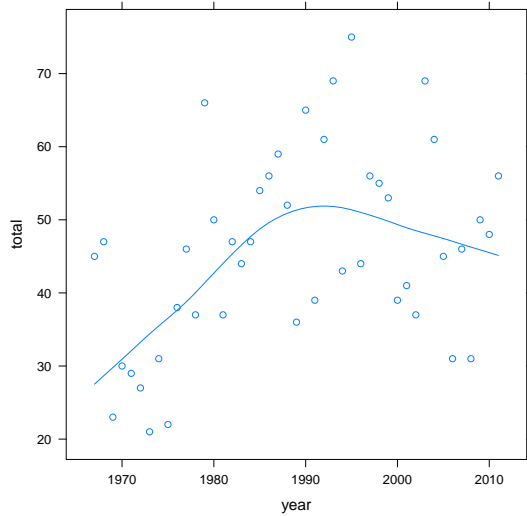
w.score	l.score	loser	venue
Min. :14.00	Min. : 3.00	Buffalo Bills : 4	Louisiana Superdome: 6
1st Qu.:23.00	1st Qu.:10.00	Denver Broncos : 4	Rose Bowl (stadium): 5
Median :30.00	Median :16.00	Minnesota Vikings : 4	Miami Orange Bowl : 4
Mean :30.16	Mean :15.58	Dallas Cowboys : 3	Tulane Stadium : 3
3rd Qu.:35.00	3rd Qu.:20.00	Miami Dolphins : 3	Georgia Dome : 2
Max. :55.00	Max. :31.00	New England Patriots: 3	Joe Robbie Stadium : 2
		(Other) :24	(Other) :23

city	attendance	total	spread
Miami Florida :10	Min. : 61946	Min. :21.00	Min. : 1.00
New Orleans Louisiana: 9	1st Qu.: 72544	1st Qu.:37.00	1st Qu.: 6.00
Pasadena California : 5	Median : 74658	Median :46.00	Median :13.00
Tampa Florida : 4	Mean : 77784	Mean :45.73	Mean :14.58
San Diego California : 3	3rd Qu.: 80281	3rd Qu.:55.00	3rd Qu.:19.00
Atlanta Georgia : 2	Max. :103985	Max. :75.00	Max. :45.00
(Other) :12	NA's : 1		

The Steelers have won 6 times, Cowboys and 49ers have each won 5. Bills, Broncos and Vikings have each lost 4.

(e) Looking for a trend to the total score over time.

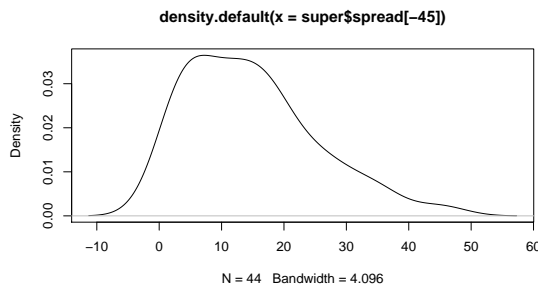
```
> require(lattice)
> print(xyplot(total ~ year, super, type = c("p", "smooth")))
```



There is not a continuous trend, but total score did rise over the period 1967 to 1990. Since then it has stayed flat or dropped a bit.

(f) Point spread is of interest to odds makers. Plot and summarize its distribution.

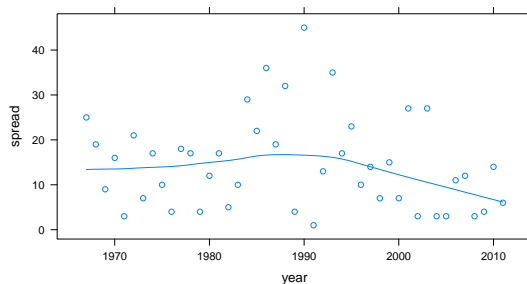
```
> summary(super$spread)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.00  6.00  13.00  14.58  19.00  45.00
> plot(density(super$spread[-45]))
```



Point spread is right skewed. The median is 13.5, quartiles 6.5 and 19.5, and extremes are 1 and 45.

(g) Looking for a trend to the point spread over time:

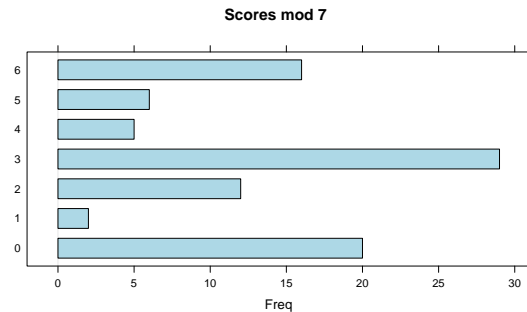
```
> print(xyplot(spread ~ year, super, type = c("p", "smooth")))
```



There's a tiny bit of decrease over time. It was flat until 1990, and since then has decreased slightly.

(h) Distribution of scores.

```
> scores <- c(super$w.score, super$l.score)
> table(scores%%7)
 0  1  2  3  4  5  6
20  2 12 29  5  6 16
> print(barchart(table(scores%%7), col = "lightblue", main = "Scores mod 7"))
```



We see that scores are most often 3 more than a multiple of seven. Next most common are multiples of seven and one less than a seven (two field goals or a missed point-after). There are numerous ways that one can get each different remainder.

R Code

```
#####
options(width=90)
128:143
length(128:143)

#####
100:80
length(100:80)

#####
pi.seq <- seq(0,pi, length=9)
round(rbind(pi.seq, zapsmall(sin(pi.seq)),zapsmall(cos(pi.seq))),4)

#####
kidsfeet <- read.table("http://www.amstat.org/publications/jse/datasets/kidsfeet.dat")
names(kidsfeet) <- c("month","year","length","width","sex","foot","hand")
plot(width ~ length, data= kidsfeet, pch = as.character(sex), col = 6-unclass(sex)*2)
feet.fit <- lm(width ~ 0 + sex + length:sex, data= kidsfeet)
abline(coef(feet.fit)[c(1,3)],col="blue", lty=4)
abline(coef(feet.fit)[c(2,4)],col="red", lty=2)
##or
#par(mfrow=c(1,2))
# plot(width ~ length, data=kidsfeet, subset = sex=="B", main = "Boys")
# plot(width ~ length, data=kidsfeet, subset = sex=="G", main = "Boys")

(foot.fit.summary <- summary( lm(width ~ sex * length, data= kidsfeet))$coef)

#####
cars <- read.csv("../data/kuiper.csv",head=T)
plot(Price ~ Type, data = cars)

#####
plot(Price ~ Mileage, data = cars, pch = unclass(Type), col = unclass(Type))
for(typ in 1:5)
  abline(lm(Price ~ Mileage, data = cars, subset = unclass(Type)==typ), lty=typ, col=typ)
legend("topright", levels(cars$Type), pch=1:5,col=1:5, lty=1:5, bty="n")

#####
```

```

mosaicplot(Type ~Leather, data = cars)
prop.table(with(cars, table(Type, Leather)),1)

#####
super <- read.table("../data/mySuperBowl.data",head=T)

#####
super$total <- super$w.score + super$l.score
super$spread <- super$w.score - super$l.score

#####
summary(super)

#####
require(lattice)
print(xyplot(total ~ year, super, type = c("p","smooth")))

#####
summary(super$spread)
plot(density(super$spread[-45]) )

#####
print(xyplot(spread ~ year, super, type = c("p","smooth")))

#####
scores <- c(super$w.score,super$l.score)
table(scores %% 7)
print(barchart(table(scores %% 7), col="lightblue", main = "Scores mod 7"))

```