

STATS 500 HW1

AUTHOR
Minxuan Chen

PUBLISHED
September 9, 2023

Github repo: https://github.com/PKUUniiiice/STATS_500

Problem 1

The dataset `teengamb` concerns a study of teenage gambling in Britain.

Glimpse

▼ Code

```
1 data(teengamb)
2 head(teengamb)
```

	sex	status	income	verbal	gamble
1	1	51	2.00	8	0.0
2	1	28	2.50	8	0.0
3	1	37	2.00	6	0.0
4	1	28	7.00	4	7.3
5	1	65	2.00	8	19.6
6	1	61	3.47	6	0.1

Descriptive Statistics

Note that `sex` is a categorical variable so here we only provide descriptions for the other variables.

▼ Code

```
1 summary(teengamb[,2:5])
```

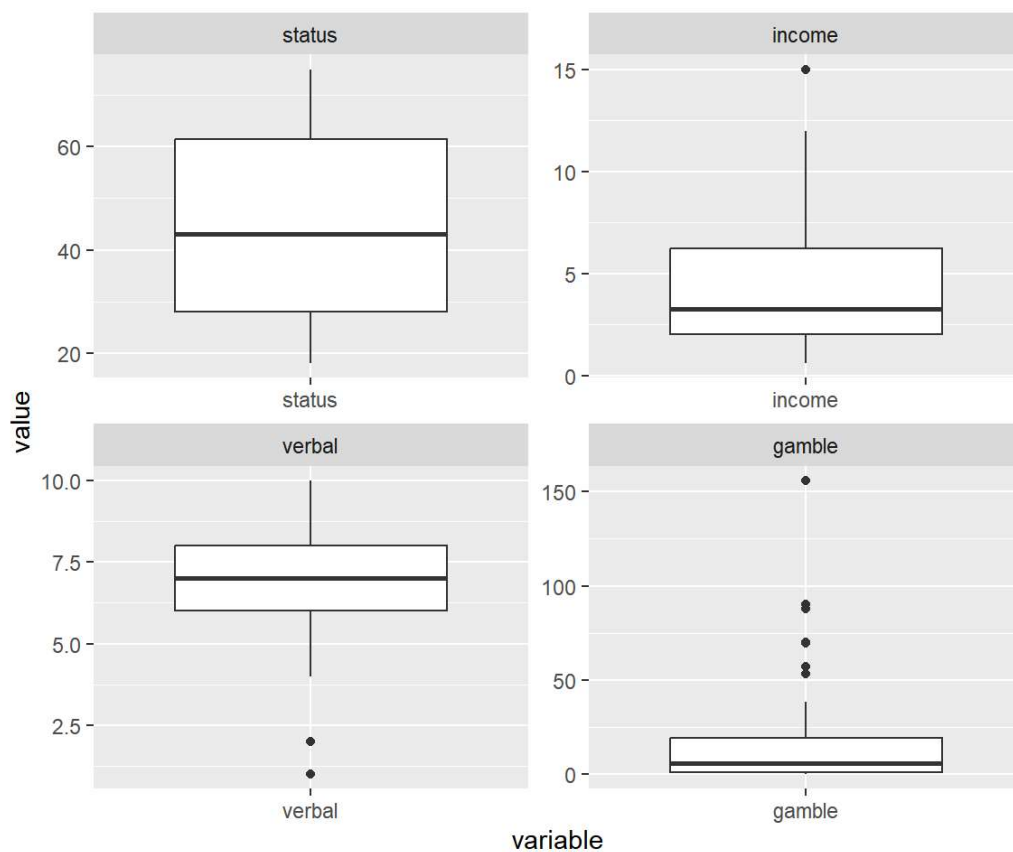
status		income		verbal		gamble	
Min.	:18.00	Min.	: 0.600	Min.	: 1.00	Min.	: 0.0
1st Qu.	:28.00	1st Qu.	: 2.000	1st Qu.	: 6.00	1st Qu.	: 1.1
Median	:43.00	Median	: 3.250	Median	: 7.00	Median	: 6.0
Mean	:45.23	Mean	: 4.642	Mean	: 6.66	Mean	: 19.3
3rd Qu.	:61.50	3rd Qu.	: 6.210	3rd Qu.	: 8.00	3rd Qu.	: 19.4
Max.	:75.00	Max.	:15.000	Max.	:10.00	Max.	:156.0

It seems that the variable `gamble` is heavily right-skewed.

Distribution and Correlation

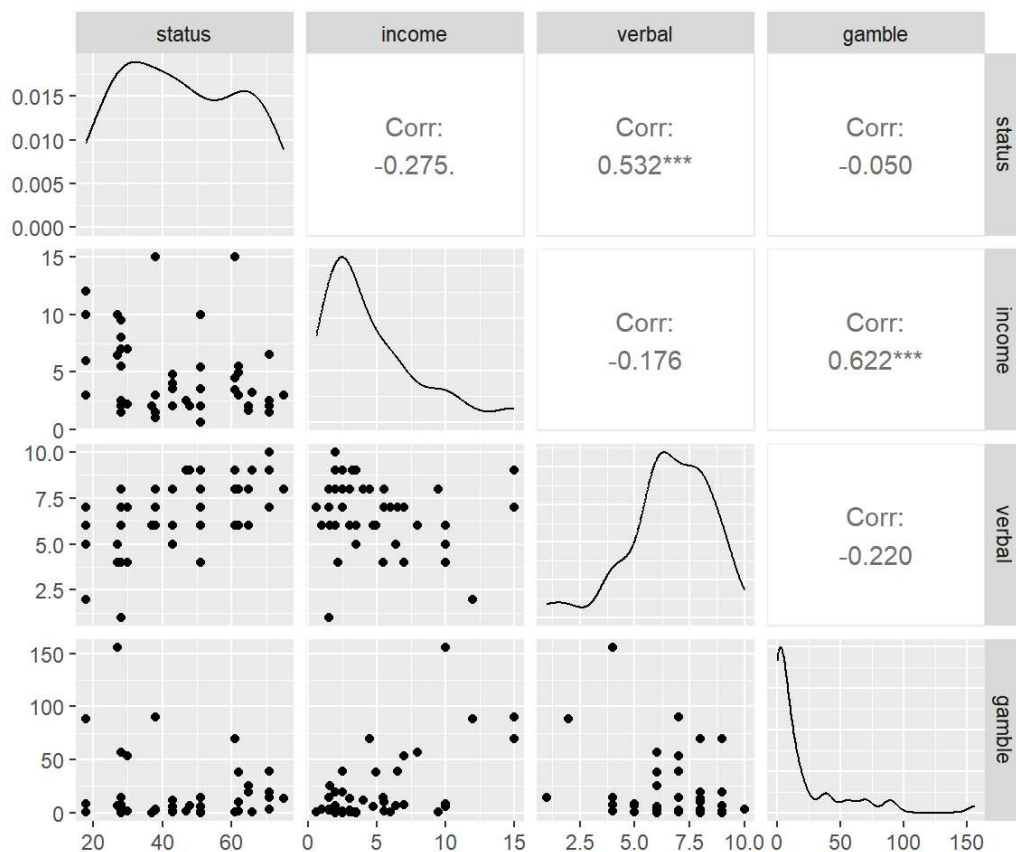
▼ Code

```
1 md <- melt(teengamb[2:5], id.vars=NULL)
2 ggplot(md, aes(variable, value)) +
3   geom_boxplot() +
4   facet_wrap(~variable, scales="free")
```



▼ Code

```
1 ggpairs(teengamb[2:5])
```



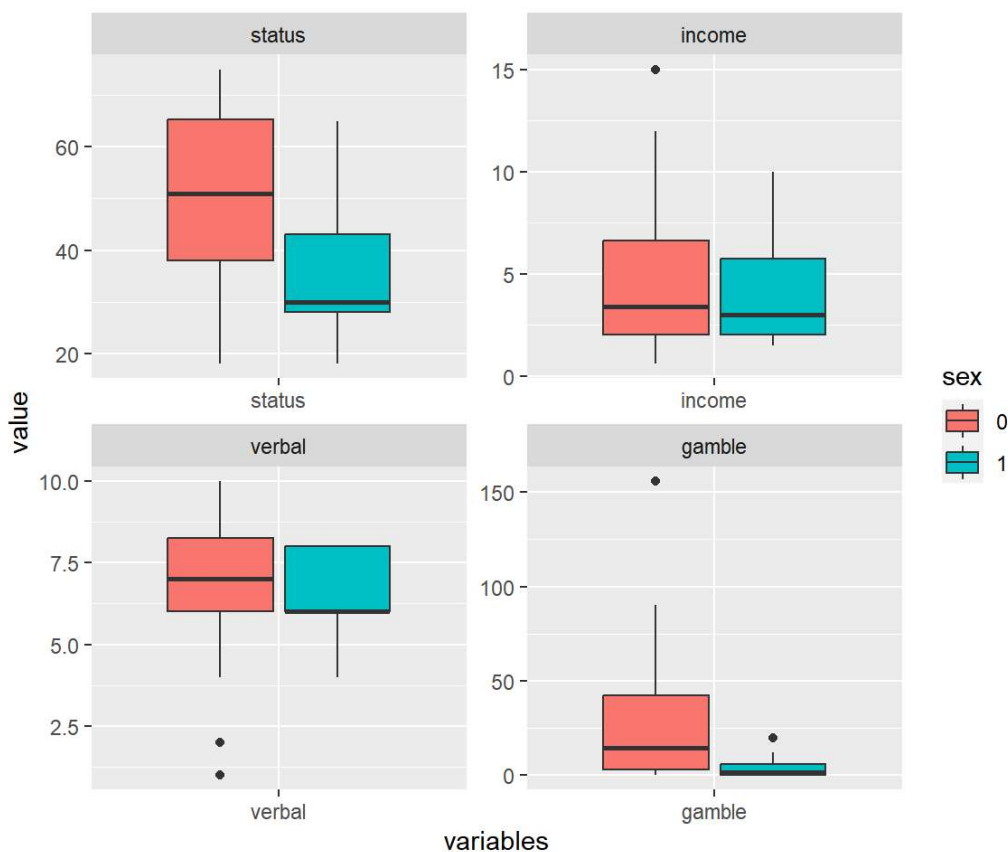
In the boxplot, we can observe outliers in all variables except for `status`, with `gamble` having the highest number of outliers.

Regarding correlation, significant linear relationships are apparent between `status` and `verbal`, as well as between `income` and `gamble`. These relationships are consistent with the meanings of these variables.

Gender Differences

▼ Code

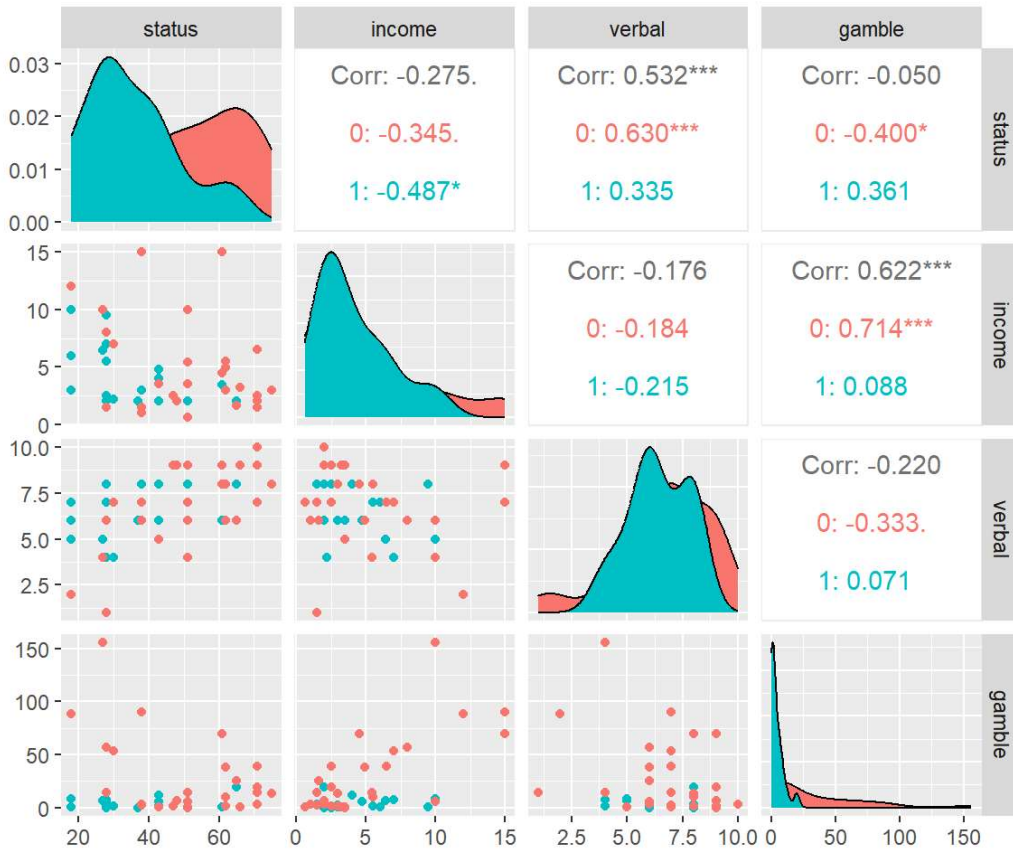
```
1 teengamb$sex <- as.factor(teengamb$sex)
2 md <- melt(teengamb, id.var=c("sex"), var="variables")
3 p <- ggplot(data=md, aes(x=variables, y=value, fill=sex))+
4   geom_boxplot()+
5   facet_wrap(~variables, scale="free")
6 show(p)
```



In the variables `status` and `gamble`, there appears to be noticeable differences between males and females. Males tend to have higher socioeconomic status scores as well as higher expenditures on gambling.

▼ Code

```
1 ggpairs(teengamb, columns=2:5, ggplot2::aes(color=sex))
```



Note that in the relationships `status - verbal` and `income - gamble`, there are noticeable differences between males and females (i.e. different corr coef). For instance, in the case of `income - gamble`, in males, teenagers with higher income tend to have higher expenditures on gambling. However, in females, gambling expenditure remains low, regardless of their income level.

Problem 2

(a)

By definition, for $E(AZ) = AE(Z)$, we need to verify

$$[E(AZ)]_i = [AE(Z)]_i$$

Note that

$$\begin{aligned}
 [E(AZ)]_i &= E([AZ]_i) \quad (\text{By def.}) \\
 &= E\left(\sum_{j=1}^m A_{ij}Z_j\right) \\
 &= \sum_{j=1}^m A_{ij}E(Z_j) \quad (\text{Linearity}) \\
 &= \sum_{j=1}^m A_{ij}[E(Z)]_j \quad (\text{By def.}) \\
 &= [AE(Z)]_i
 \end{aligned}$$

Similarly, for $\text{Cov}(AZ) = A \text{Cov}(Z) A^\top$, we need to verify

$$\text{Cov}(AZ)_{ij} = [A \text{Cov}(Z) A^\top]_{ij}$$

Note that

$$\begin{aligned}
 LHS &= \text{Cov}([AZ]_i, [AZ]_j) \\
 &= \text{Cov}\left(\sum_{s=1}^m A_{is} Z_s, \sum_{t=1}^m A_{jt} Z_t\right) \\
 &= \sum_{t=1}^m \sum_{s=1}^m A_{is} A_{jt} \text{Cov}(Z_s, Z_t) \\
 &= \sum_{t=1}^m \sum_{s=1}^m A_{is} A_{jt} \text{Cov}(Z)_{st}
 \end{aligned}$$

and

$$\begin{aligned}
 RHS &= [A \text{Cov}(Z) A^\top]_{ij} \\
 &= \sum_{t=1}^m [A \text{Cov}(Z)]_{it} [A^\top]_{tj} \\
 &= \sum_{t=1}^m \left(\sum_{s=1}^m A_{is} \text{Cov}(Z)_{st} \right) [A^\top]_{tj} \\
 &= \sum_{t=1}^m \left(\sum_{s=1}^m A_{is} \text{Cov}(Z)_{st} \right) A_{jt} \quad (A_{jt} = [A^\top]_{tj}) \\
 &= \sum_{t=1}^m \sum_{s=1}^m A_{is} A_{jt} \text{Cov}(Z)_{st} \\
 &= LHS
 \end{aligned}$$

(b)

We have

$$Y = \begin{bmatrix} Z_1 + 2Z_2 \\ Z_1 - 2Z_2 \\ -Z_1 + Z_2 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 1 & -2 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} = AZ$$

and

$$\text{Cov}(Z) = \begin{bmatrix} 3 & -1 \\ -1 & 3 \end{bmatrix}$$

Therefore,

$$\text{Cov}(Y) = A \text{Cov}(Z) A^\top$$

▼ Code

```

1 A <- matrix(c(1,1,-1,2,-2,2),ncol=2)
2 cov.z <- matrix(c(3,-1,-1,3),ncol=2)
3 cov.Y <- A%*%cov.z%*%t(A)
4 cov.Y

```

```

      [,1] [,2] [,3]
[1,]   11   -9    9
[2,]   -9   19  -19
[3,]    9  -19   19

```

and

$$\text{Corr}(Y) = [\text{diag}(\text{Cov}(Y))]^{-\frac{1}{2}} \text{Cov}(Y) [\text{diag}(\text{Cov}(Y))]^{-\frac{1}{2}}$$

▼ Code

```
1 diag(diag(cov.Y)^(-1/2))%*%cov.Y%*%diag(diag(cov.Y)^(-1/2))
```

```
      [,1]      [,2]      [,3]  
[1,]  1.000000 -0.622543  0.622543  
[2,] -0.622543  1.000000 -1.000000  
[3,]  0.622543 -1.000000  1.000000
```

Problem 3

It's known that

$$s_y = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2}, s_x = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$$
$$r = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{s_x \cdot s_y}$$

We go from $y = \hat{\beta}_0 + \hat{\beta}_1 x$

$$\begin{aligned} y &= \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x \\ &= \bar{y} + \hat{\beta}_1 (x - \bar{x}) \\ &= \bar{y} + \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} (x - \bar{x}) \\ &= \bar{y} + \frac{n \cdot r \cdot s_x \cdot s_y}{\sum_{i=1}^n (x_i - \bar{x})^2} (x - \bar{x}) \\ &= \bar{y} + \frac{r \cdot s_x \cdot s_y}{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} (x - \bar{x}) \\ &= \bar{y} + \frac{r \cdot s_x \cdot s_y}{s_x^2} (x - \bar{x}) \end{aligned}$$

Therefore, we have

$$\frac{y - \bar{y}}{s_y} = r \cdot \frac{x - \bar{x}}{s_x}$$