Lecture 1 Handout: Data and EDA

Statistics 139 Team

September 07, 2023

Topics

- Dataset manipulation in R
- Numerical summaries: mean, SD, median, IQR
- Graphical summaries: barplots, boxplots, histograms, scatterplots

The material in this lab corresponds to the Lecture 1 Notes.

The General Social Survey (GSS) is a biennial, nationally representative survey conducted by the National Opinion Research Center at the University of Chicago. The GSS is second only to the U.S. Census as the most cited social science dataset in the country,' Even though the data are collected via a complex sampling design, the data can reasonably be analyzed validly as if it were a simple random sample from the US population (we will get into handling this more carefully via survey weights later in the course), see this blog by Andrew Gelman, a world-class statistician.

The following questions will be explored in this lab with the GSS 2018 data:

1. At what age do Americans no longer further their education?
2. How is marital status linked to education and low income status?
3. How does income relate to being a government vs. private sector employee?

The full GSS 2018 data are available in the data file 'gss18.csv'. For convenience, descriptions of the variables used in this lab exercise are included below. To view the complete list of study variables and their descriptions, access the GSS documentation code book by clicking here.

- `age`: age of respondent, in years. Respondents 89 years or older were recorded as 89 years of age.
- `educ`: years of education of respondent (beyond kindergarten). Respondents with 20 or more years of education were recorded as 20 years.
- `rincom16`: the respondent's income, categorized into many groups. See gssdataexplorer.norc.org/variables/6168/vshow for full details.
- `hrs1`: number of hours respondent worked the previous week.
- `marital`: marital status, with categories 1 = married, 2 = widowed, 3 = divorced, 4 = separated, and 5 = never married.

- `wrkgovt`: does respondent work for the government? 1 = government job, 2 = private sector job.

Concept Checks (Stat 110/111 Review):

a) What is the distinction between $\bar{X}$, $\bar{x}$ and μ?

$\bar{X}$ is the sample mean as a random variable (the potential values it could take on), $\bar{x}$ is the observed sample mean from data (an actual, observed value), and μ is the population mean (an unknown value, typically).

b) What is the sampling distribution of $\bar{X}$? When is this exact? When is this an approximation?

$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ (note: this assumes approximate independence of observations). This is exact if the individual measurements in the population (or data generating process) is Normally distributed. This is an approximation when the sample size, n, is 'large enough' (typically, $n \geq 30$ works well enough unless there is extreme skewness or severe outliers).

c) What is the sampling distribution of the sample variance, S^2? When is this exact? When is this an approximation?

$S^2 \sim \left(\frac{\sigma^2}{n-1}\right)\chi^2_{n-1}$. This is exact if the observations are normally distributed and are independent. This is approximate if the sample size is large enough (as it will be approximately Normal based on CLT!).

Question 1.

a) Using numerical and graphical summaries, describe the distribution of ages of the respondents.

The survey clearly samples adults as ages vary from 18 to at least 89 years of age. The distribution may be bimodal, with more individuals in their thirties and late fifties, with a clear tailing off after about 70 years of age.

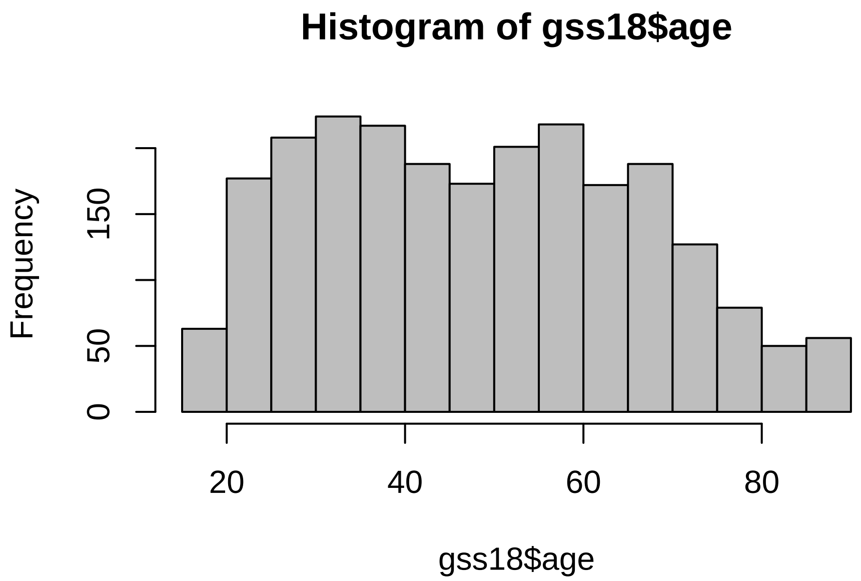
```r
gss18 = read.csv("data/gss18.csv")
summary(gss18$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##   18.00   34.00   48.00   48.97   63.00   89.00       7
```

```r
sd(gss18$age,na.rm=T)
```

```
## [1] 18.06088
```

```r
hist(gss18$age,col="gray",breaks=20)
```



b) Calculate the median and interquartile range of the distribution of the variable `hrs1`. Write a sentence explaining the median in the context of these data.

After handling the missingness with the argument `na.rm = T`, we find that the median is 40 hours per week and the interquartile range is 15 hours per week. The median is not surprising as the 'typical' American work week is 40 hours per week (in fact we find that 30.4% of respondents reported to have worked 40 hours per week).

```
median(gss18$hrs1, na.rm = T)
```

```
## [1] 40
```

```
IQR(gss18$hrs1, na.rm = T)
```

```
## [1] 15
```

```
mean(gss18$hrs1 == 40, na.rm = T)
```

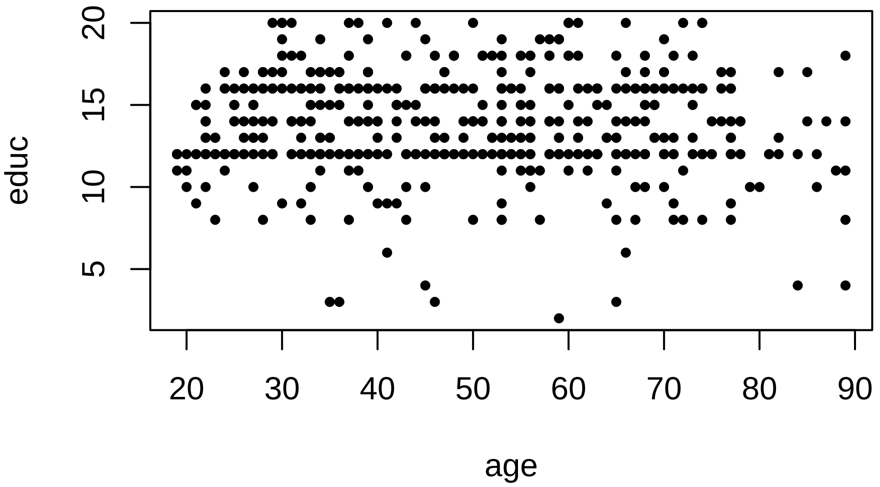
```
## [1] 0.3041274
```

c) Use the following code to draw a random sample of 500 participants from the entire dataset. Using the random sample, `gss18.samp`, visually investigate the relationship between age and education. Based on this visual, at what age do respondents appear to no longer further their education? Use this smaller sample only for this part of the problem.

The scatterplot below illustrates any possible relationship. The only emerging pattern is that younger individuals do not have the extremes (no points in the upper left in the area of 20 years of age and above 15 years of schooling). The full distribution of education (up to 20 years) begins in the late twenties or even 30 years of age, so that is reasonably when most Americans no longer pursue further education. Note: the variable `educ` may be misleading since it is right-censored at 20.

```
#draw a random sample
set.seed(139)
row.num = sample(1:nrow(gss18), 500, replace = FALSE)
gss18.samp = gss18[row.num, ]
```

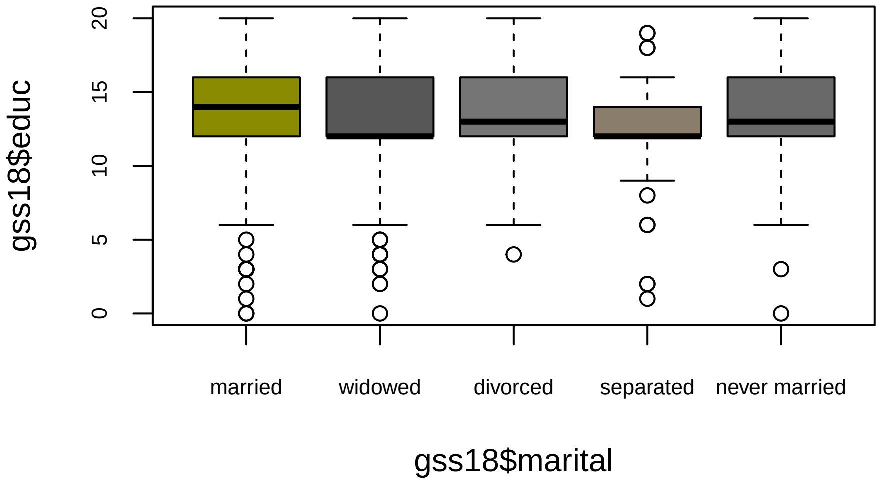
```
plot(educ~age, data = gss18.samp, pch = 16, cex = 0.7)
```



d) Compare the distribution of `educ` across each group in `marital` among adults (defined as individuals 25 years of age or older). Describe any trends or interesting observations.

In the plots below, we see that most marital groups are very similar, but there may be slight evidence that the separated group has slightly less education, on average. This could be a sample size issue though since it is the smallest group. Note: we first turned `marital` into a factor to make interpreting the plot easier.

```
#draw a random sample
gss18$marital = factor(gss18$marital, labels = c("married","widowed",
                       "divorced","separated","never married"))
boxplot(gss18$educ~gss18$marital, col = sample(colors(),5),cex.axis=0.65)
```



```
table(gss18$marital)
```

```
##
##       married       widowed      divorced     separated never married
##           998           200           403            75           670
```

Question 2.

a) Create a dummy/indicator/binary variable `lowincome` to indicate those individuals that make less than $15,000. Construct a two-way table, with `marital` as the row variable and `lowincome` as the column variable. Which group is at lowest risk of being low income? Highest risk?

Based on the second table below (with proportions within group), we see that separated respondents had the lowest risk at 14.0%, and never married respondents had the highest risk at 33.8%.

```
# 10 is the cut-off based on teh variable defintion.  Click on the link in the
# introduction to the data set above (the link for the `rincom16` variable)
gss18$lowincome = 1*(gss18$rincom16 <= 10)
table(gss18$marital, gss18$lowincome)
```

```
##
##                  0   1
##   married       521  93
##   widowed        39  16
##   divorced      200  43
##   separated      37   6
##   never married 270 138
```

```
prop.table(table(gss18$marital, gss18$lowincome),1)
```

```
##
##                         0         1
##   married       0.8485342 0.1514658
##   widowed       0.7090909 0.2909091
##   divorced      0.8230453 0.1769547
##   separated     0.8604651 0.1395349
##   never married 0.6617647 0.3382353
```

b) Relative risk can measure how two categorical variables are related (really , two binary variables). Here, we are interested in measuring the relative risk as the ratio of the proportion of respondents who are low income among those who are divorced to the proportion of respondents who are low income among those who are married. Calculate this relative risk for these respondents.

From these calculations, is it possible to conclude that getting divorced reduces or raises one's chance of being low income?

The relative risk is estiamted to be $RR = 0.177/0.151 = 1.168$, meaning divorced repsondents are have a roughly 17% higher chance of being low income than married respondents. There is no way to conclude that this is a causal link as the data are observational allowing for many possible confounding factors (could be the other way around: those with low income mauy have extra stressors leading to divorce).

```
usefulrows = prop.table(table(gss18$marital, gss18$lowincome),1)[c("divorced","married"),]
rr = usefulrows[1,2]/usefulrows[2,2]
rr
```

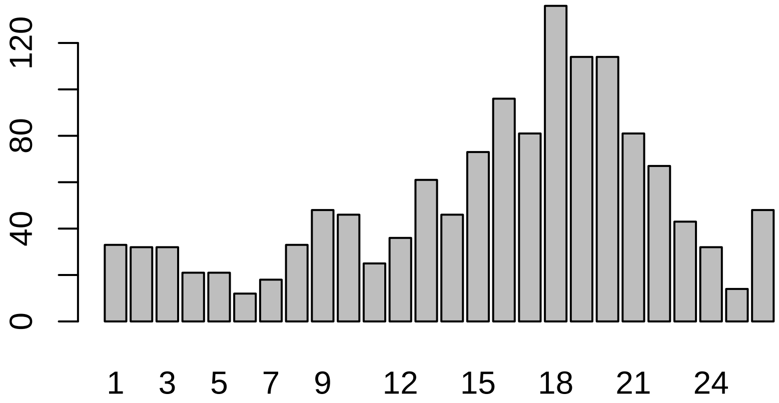
```
## [1] 1.168282
```

Question 3.

a) Describe the distribution of income of the respondents. Estimate the median income, and provide a rough estimate for both the mean income and standard deviation of incomes.

Treating it like an ordered categorical variable (which it truly is), we look at the barplot and see that the majority of individuals are between categories 12 and 24 (between $17,500 and $150,000), with a peak around category 18 ($40,000 to $50,000). The median is category 17, which is ($35,000 to $40,000). Be careful not to call this left-skewed as the categories are not equal in width on the income scale. The mean is a tough one, but income is likely right-skewed, so a good chance it would be above $40,000. Standard devation is nearly impossible to estimate from this variable, but IQR is around $48,750 ($67500 - 18750$), based on the midpoints of categories 12 and 20. Standard deviation is typically less than IQR, so a guess of around $40,000 is reasonable. Note: the transformation of actual income to these income brackets is non-linear (but is monotonic), and the median will hold up since the transformation is monotonic. The mean and sd will certainly be affected (remember: for a non-linear transformation: $E(g(X)) \neq g(E(X))$).

```r
barplot(table(gss18$rincom16))
```



```r
median(gss18$rincom16,na.rm=T)
```

```
## [1] 17
```

```r
quantile(gss18$rincom16,c(0.25,0.75),na.rm=T)
```

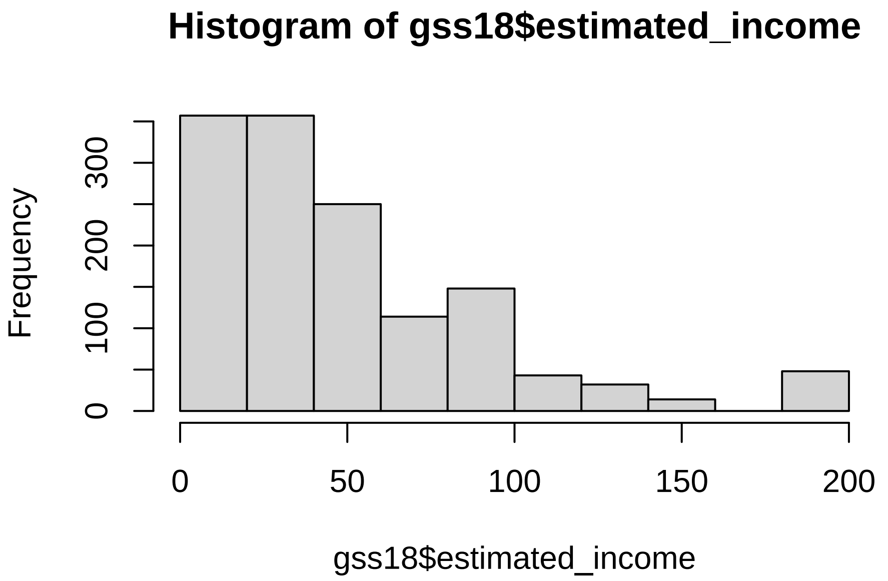
```
## 25% 75%
##  12  20
```

b) The following code creates a new variable, `estimated_income`, within `gss18` that records the rough median of each income group from the variable `rincom16` for each respondent (in thousands of dollars). Use this variable to substantiate your 3 estimates in the previous part. Which of the 3 estimates will be biased?

The estimates of mean, median, and standard deviation are 37.5, 50.4, and 45.3, respectively (in thousand of dollars). These are reasonably close to the guesses in the previous part (note the distribution is VERY right-skewed, and thus the mean and sd are higher than expected).

```r
medians = c(0.5, 2, 3.5, 4.5, 5.5, 6.5, 7.5, 9, 11.25, 13.75, 16.25, 18.75, 21.25,
            23.75, 27.5, 32.5, 37.5, 45, 55, 67.5, 82.5, 100, 120, 140, 160, 200)
gss18$estimated_income = medians[gss18$rincom16]
```

```
hist(gss18$estimated_income)
```



```
median(gss18$estimated_income, na.rm = T)
```

```
## [1] 37.5
```

```
mean(gss18$estimated_income, na.rm = T)
```

```
## [1] 50.37968
```

```
sd(gss18$estimated_income, na.rm = T)
```

```
## [1] 45.26101
```

c) Which of the 3 estimates will be biased in the previous part? Will it be over or under estimated? Why?

\textcolor{blue}{ Standard deviation is essentially guaranteed to be an underestimate of truth since we are ignoring spread within each income category (assume everyone to be at the midpoints). The mean is likely biased (likely underestimated) too since we do not know how high the incomes in the top category could go (the outliers are truncated to $200,000). The median may not be biased at all!}

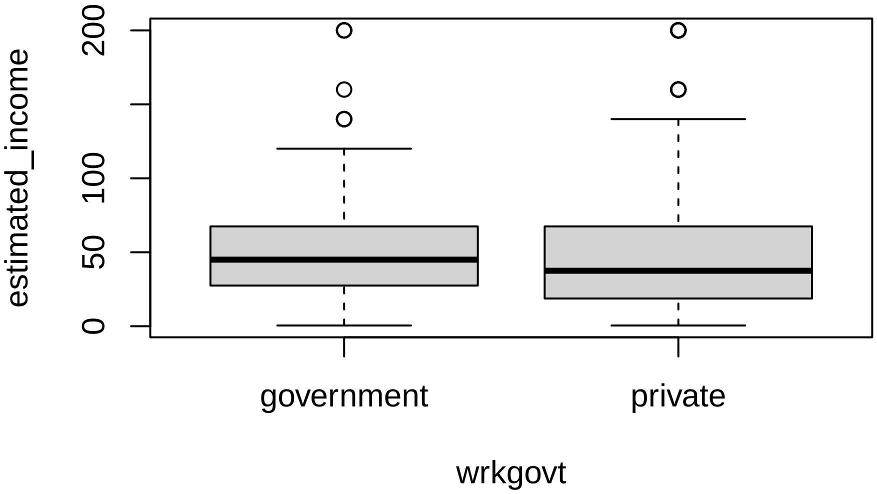
d) Propose a better way to create the estimated income for each respondent that will end up with a less biased estimate than in part (b)? You do not need to implement this.

A better way to implement this would be to sample from within each income category. The simplest way would be to sample from uniform distributions between the bounds of each category, but most likely the distributions within each category are also right-skewed. We could assume a general log-normal distribution of incomes, and apply that across the income categories (would take some tweeking to get right).

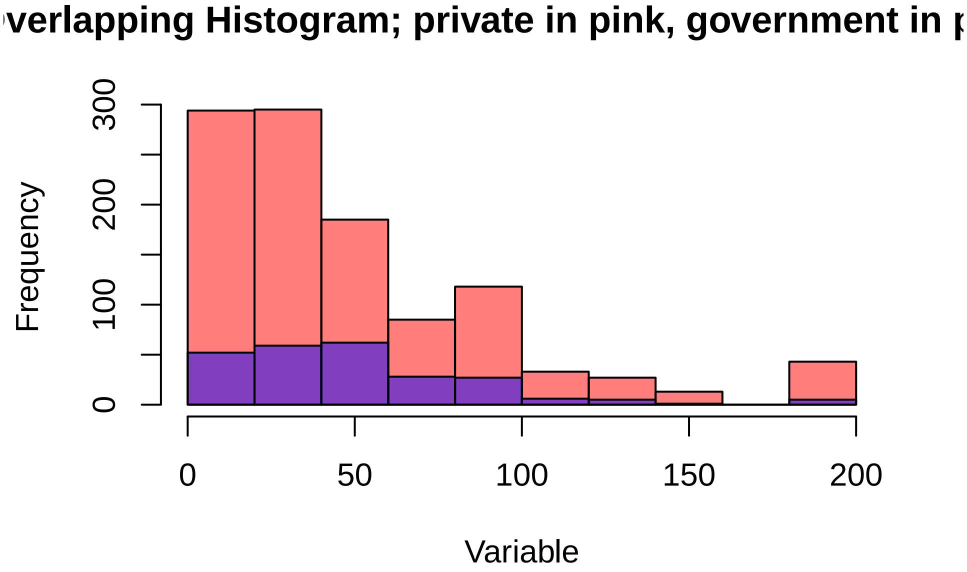
e) Compare the distribution of `estimated_income` for government employees vs. private industry employees. Describe what you see in a few sentences.

Both distributions are right-skewed, the private sector more so. The estimated mean is slightly higher for the private sector ($50.58K vs $49.99K), but the median is higher in the government group ($45K vs $37.5K). Not surprisingly: there is much greater opportunity to make a lot of money (or very little money) in the private sector.

```
gss18$wrkgovt = factor(gss18$wrkgovt, labels = c("government","private"))
boxplot(estimated_income~wrkgovt, data=gss18)
```



```
hist(gss18$estimated_income[gss18$wrkgovt=="private"], col=rgb(1,0,0,0.5),
     main="Overlapping Histogram; private in pink, government in purple",
     xlab="Variable")
hist(gss18$estimated_income[gss18$wrkgovt=="government"],data=gss18,
     col=rgb(0,0,1,0.5), add=T)
```



```
summary(gss18$estimated_income[gss18$wrkgovt=="government"])
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's 
##     0.50    27.50    45.00    49.99    67.50   200.00      338
```

```
summary(gss18$estimated_income[gss18$wrkgovt=="private"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##    0.50   18.75   37.50   50.58   67.50  200.00     806
```