Messy Data Lab Part 2

ben Goldstone

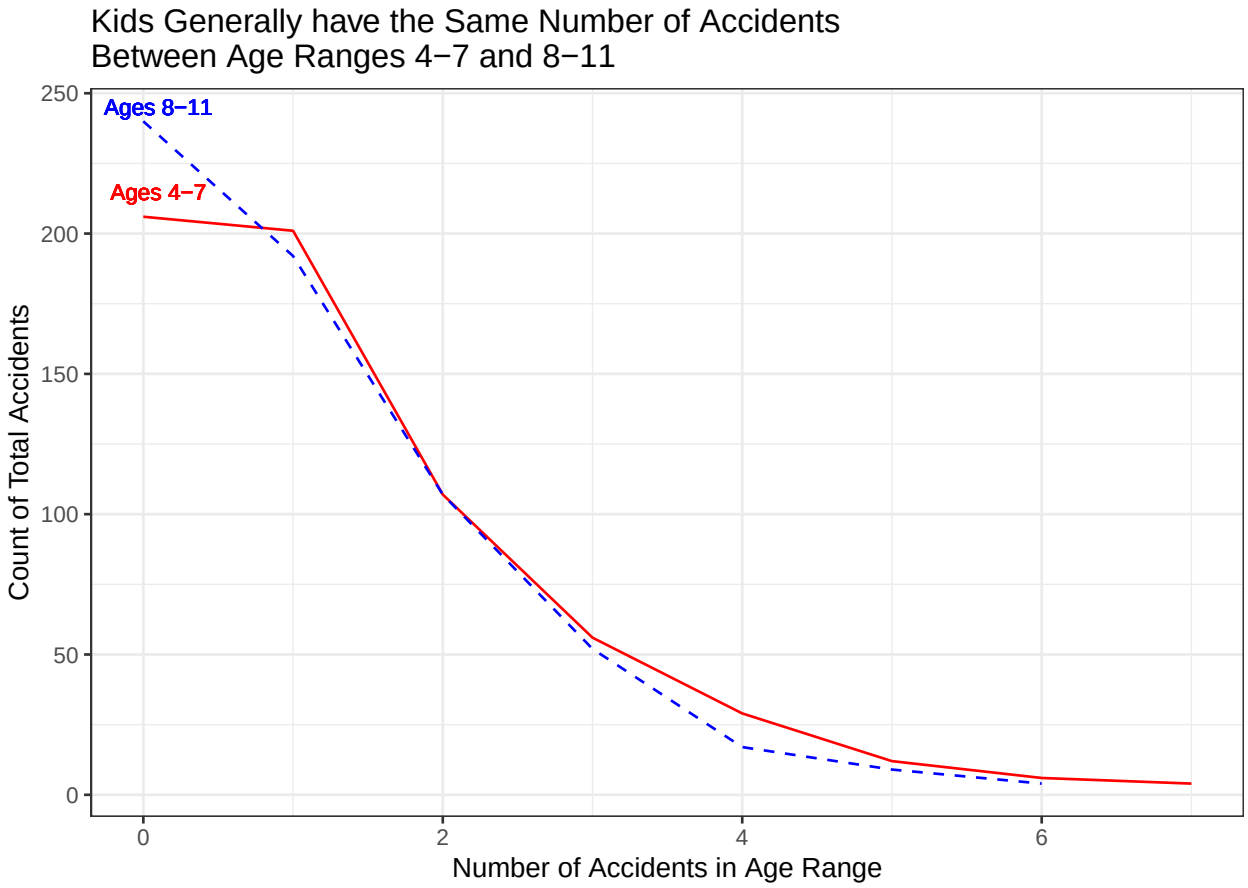
11/29/2023

Dataset # 209: Accident-Repeated Ness Among Children

```
# Random Numbers: 345 208 149 154 465 104  63 209  87 187
accident = read.table("https://raw.githubusercontent.com/JedStephens/Handbook-of-Small-Data-Sets/master/
head(accident)
```

```
##    V1 V2 V3 V4 V5 V6 V7 V8  V9
## 1 101 76 35 15  7  3  3  0 240
## 2  67 61 32 14 12  4  1  1 192
## 3  24 36 22 15  6  1  2  1 107
## 4  10 19 10  5  2  4  0  2  52
## 5   1  7  3  4  2  0  0  0  17
## 6   2  1  4  2  0  0  0  0   9
```

```
accident$T2 = c(0,1,2,3,4,5,6,7)
newAccident = gather(accident,"T1","accidents", 1:9)
newAccident$T1 = parse_number(newAccident$T1)
newAccident$T1 = newAccident$T1 - 1
gf_line(accidents~T1,data=subset(newAccident, T2 == 7 & T1!=8),color = "red") %>%
  gf_line(accidents~T2,data=subset(newAccident, T1 == 8 & T2!=7),color = "blue",linetype = "dashed") %>%
  gf_text(x=0.1,y=215,label="Ages 4-7", color="red",size = 3.2) %>%
  gf_text(x=0.1,y=245,label="Ages 8-11", color="blue",size=3.2) %>%
  gf_labs(y="Count of Total Accidents",x="Number of Accidents in Age Range",
          title="Kids Generally have the Same Number of Accidents\nBetween Age Ranges 4-7 and 8-11")
```



I chose this visualization as I think it accurately portrays what the statisticians were envisioning which is comparing the difference in amount of accidents between the two age groups. I used color and different line weights to differentiate between the age groups in this dataset. Given there is no clear indication of what these statisticians were trying to find, I think this is the best way to represent this data in this situation.

Dataset # 149: Silver content of Byzantine coins

```
silver = read.delim("https://raw.githubusercontent.com/JedStephens/Handbook-of-Small-Data-Sets/master/S
head(silver)
```

```
##    V1  V2  V3  V4
## 1 5.9 6.9 4.9 5.3
## 2 6.8 9.0 5.5 5.6
## 3 6.4 6.6 4.6 5.5
## 4 7.0 8.1 4.5 5.1
## 5 6.6 9.3  NA 6.2
## 6 7.7 9.2  NA 5.8
```

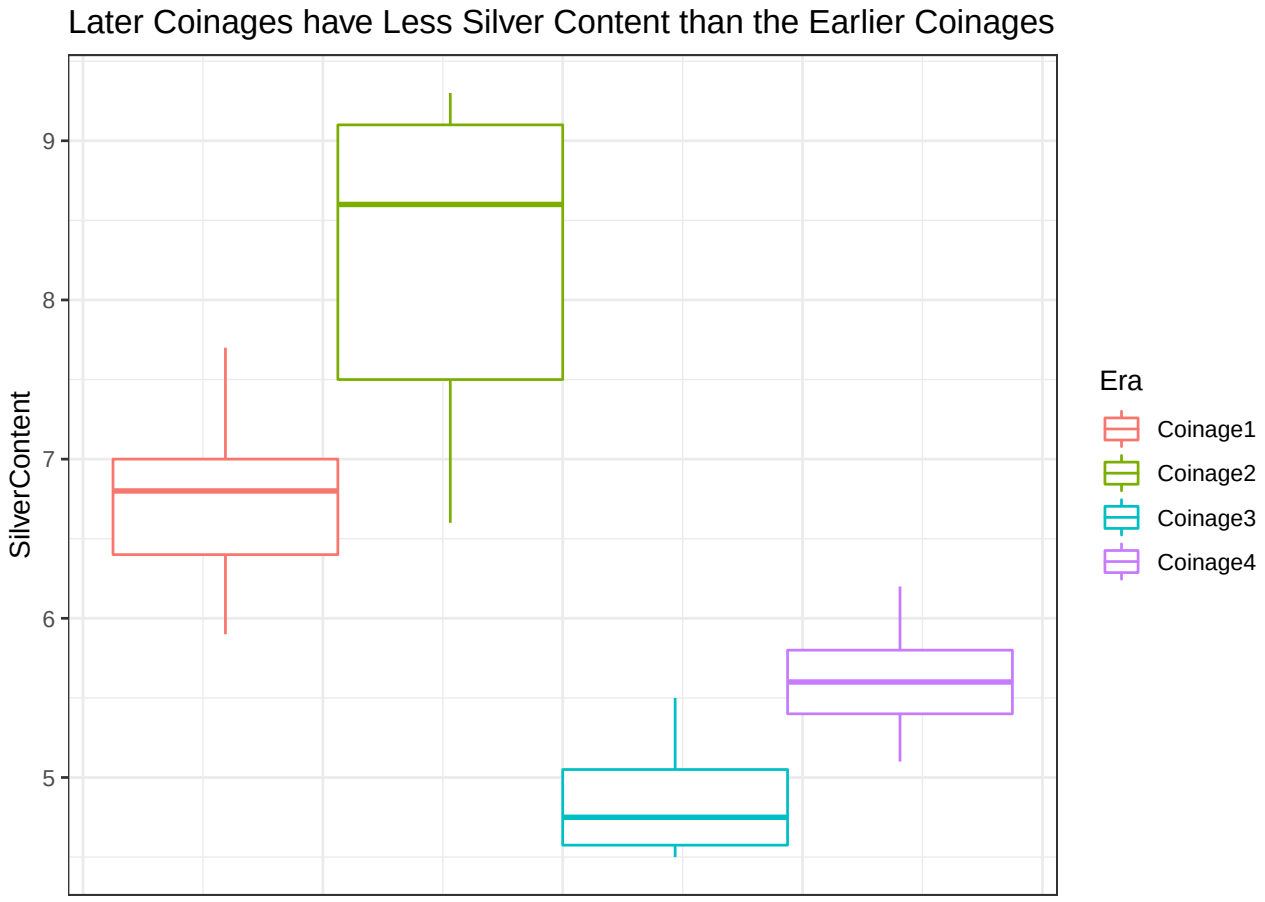
```
silver = rename(silver, Coinage1=V1,Coinage2=V2,Coinage3=V3,Coinage4=V4)
newSilver = gather(silver,"Era","SilverContent",1:4)
head(newSilver)
```

```
##        Era SilverContent
## 1 Coinage1           5.9
## 2 Coinage1           6.8
```

```
## 3 Coinage1           6.4
## 4 Coinage1           7.0
## 5 Coinage1           6.6
## 6 Coinage1           7.7
```

```
gf_boxplot(~SilverContent, data=newSilver,color=~Era) %>%
  gf_labs(title="Later Coinages have Less Silver Content than the Earlier Coinages") %>%
  gf_theme(axis.text.x = element_blank(), axis.ticks.x = element_blank()) %>%
  gf_refine(coord_flip())
```

```
## Warning: Removed 9 rows containing non-finite values (stat_boxplot).
```




```
silver_model=aov(SilverContent~Era, data=newSilver)
summary(silver_model)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## Era           3  37.75  12.583   26.27 1.31e-07 ***
## Residuals    23  11.02   0.479
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 9 observations deleted due to missingness
```

```
TukeyHSD(silver_model)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
```

```
## Fit: aov(formula = SilverContent ~ Era, data = newSilver)
##
## $Era
##                          diff        lwr        upr     p adj
## Coinage2-Coinage1  1.4984127  0.5332876  2.4635378 0.0014326
## Coinage3-Coinage1 -1.8694444 -3.0202828 -0.7186061 0.0008811
## Coinage4-Coinage1 -1.1301587 -2.0952838 -0.1650336 0.0176461
## Coinage3-Coinage2 -3.3678571 -4.5682163 -2.1674979 0.0000004
## Coinage4-Coinage2 -2.6285714 -3.6522412 -1.6049017 0.0000018
## Coinage4-Coinage3  0.7392857 -0.4610735  1.9396449 0.3443290
```

I chose this visualization as it shows a the differences in silver content over many eras and provides a median for each era. I chouse to color the boxplots by era so we can see the differences in silver content for each era. Since there was no need for an x-axis I removed it to lessen confusion and flipped the boxplots so they were in a vertical orientation which I felt was more pleasent to look at. Given the boxplots and the ANOVA tests here, the silver content is not very significant between each of the different coinages. The coinanges that have the most significant difference is between Coinage 4 and Coingage 3 which is not what the statisticians were looking for which was a difference between the beginning (Coinage 1) and the end (Coinage 4).

Dataset # 154: Heights of Elderly Females With Osteoporosis

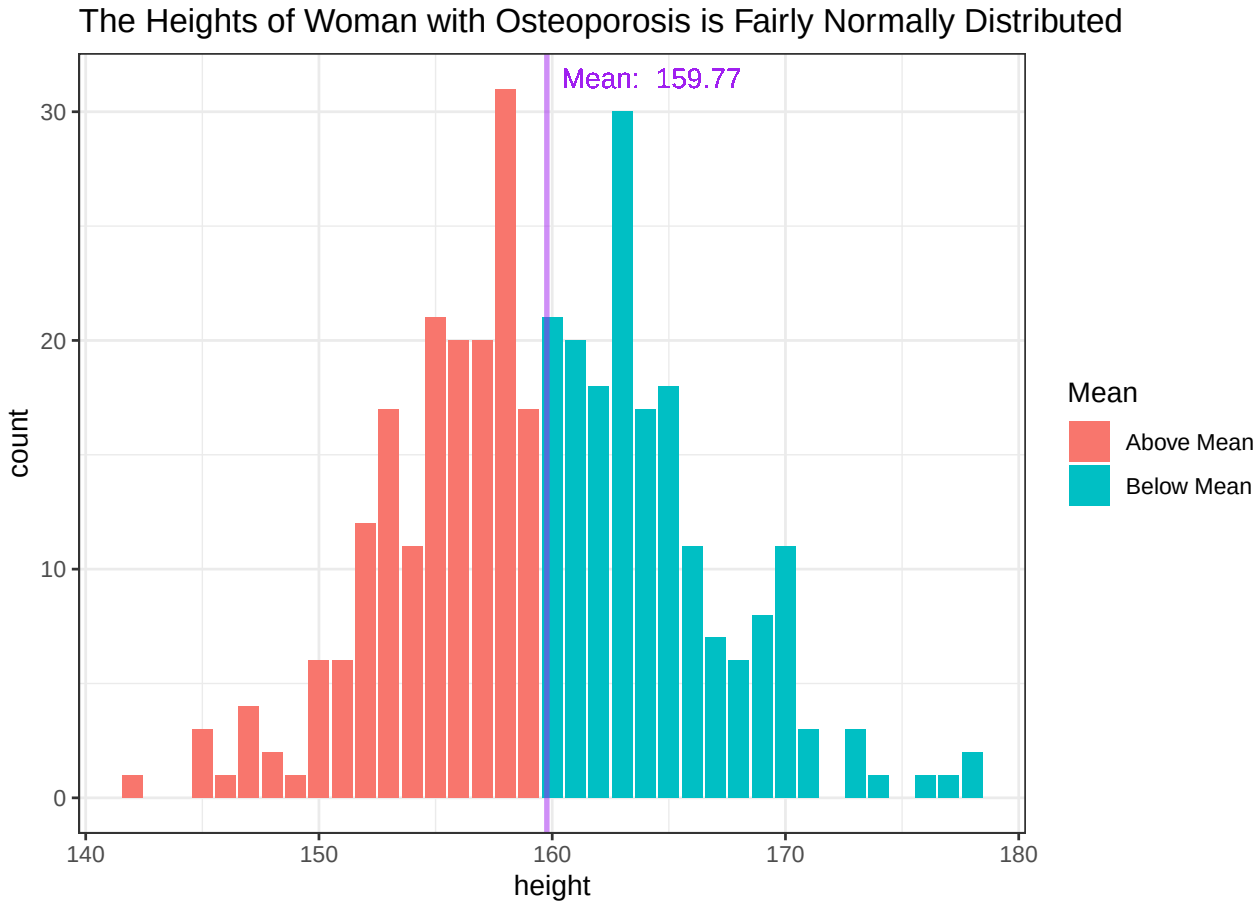
```
heights = read.delim("https://raw.githubusercontent.com/JedStephens/Handbook-of-Small-Data-Sets/master/
  header = FALSE)

newHeights = gather(heights, "ColName", "height", 1:ncol(heights)) %>% select(height)
heightMean = round(favstats(newHeights$height)[6]$mean,2)
newHeights = transform(newHeights, color=ifelse(heightMean - height < 0, "Below Mean", "Above Mean"))
head(newHeights)
```

```
##   height      color
## 1    156 Above Mean
## 2    150 Above Mean
## 3    156 Above Mean
## 4    155 Above Mean
## 5    164 Below Mean
## 6    160 Below Mean
```

```
meanText = paste("Mean: ",heightMean)
gf_bar( ~height, data=newHeights,fill=~color) %>%
  gf_labs(title="The Heights of Woman with Osteoporosis is Fairly Normally Distributed",fill="Mean") %>%
  gf_vline(xintercept = heightMean,color="purple",size=1,alpha=0.5) %>%
  gf_text(x=heightMean+4.5, y=31.5, label=meanText,size=4,color="purple", alpha=0.01)
```

```
## Warning: geom_vline(): Ignoring `mapping` because `xintercept` was provided.
```

```
## Warning: Removed 9 rows containing non-finite values (stat_count).
```



I chose this visualization as it shows the distribution of heights among women with Osteoporosis. I chouse to color the bars to show a below and above a mean to show how normally distributed the data is. Given the bargraph here, the heights are pretty normally distributed among women with Osteoporosis.