Week 6 Challenge Problem Solutions

Let's clear the global computing environment:

```
rm( list = ls() )
```

Now let's load in the Problem Set 6 R object file:

```
load( "Problem Set 6 R Objects.Rdata" )
```

Week 6 Challenge Problem: Pie Chart vs. Barplot

I've mentioned that many people disapprove of pie charts. Instead, they advocate using barplots to display this information. In this problem, we'll explore this issue.

Other than the graph in part (a), the answers for this problem are not really "right" or "wrong", and we'll give you credit as long as you write something reasonable. However, I think you'll get more out of the problem if you engage with it and really try to think about this issues here.

This challenge problem is a continuation of Problem 4 from Problem Set 6, so you should solve that first before working on this.

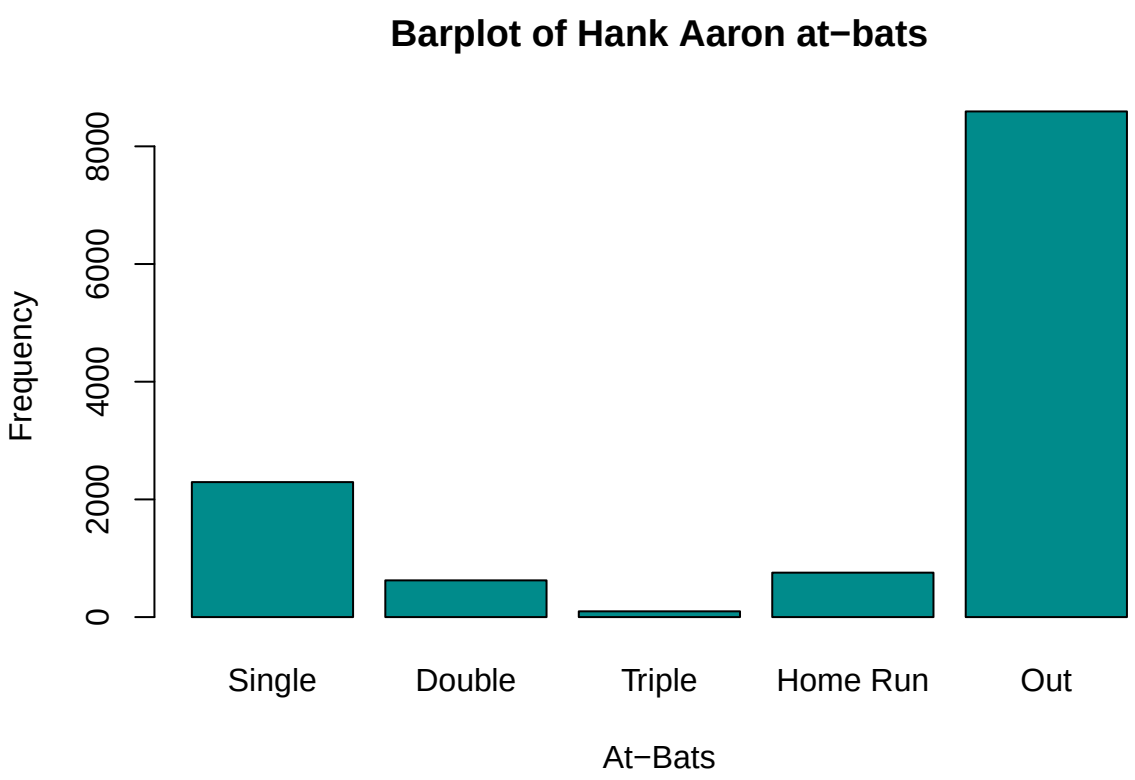
Part (a)

Construct a barplot of the data from problem 4.

Solution

```
hank.aaron.at.bats.factor <-
    factor(
        problem.4.data,
        levels =
            c( "Single", "Double", "Triple",
               "Home Run", "Out" )
    )

barplot(
    table( hank.aaron.at.bats.factor ),
    main = "Barplot of Hank Aaron at-bats",
    xlab = "At-Bats",
    ylab = "Frequency",
    col = "cyan4"
)
```



Part (b)

Using the barplot from part (a), can you easily determine Hank Aaron's batting average? Don't do any calculations in your head – just by looking at the shapes, can you make a reasonable estimate of Hank Aaron's batting average? Explain your answer with a few sentences.

Solution

Personally, I find it difficult to estimate Hank Aaron's batting average just by looking at the bar plot. It's clear that most of his at-bats were outs, but it's hard for me to form any sort of sense of the actual numerical value. The problem here is that it's difficult to intuitively compare areas of different shape, and there are many bars for the hits that I somehow have to add all together in my mind and then compare with the bar for the outs.

So, yes, it's clear that most of Hank Aaron's at-bats were outs, but it's hard to conclude anything else about the relative proportions.

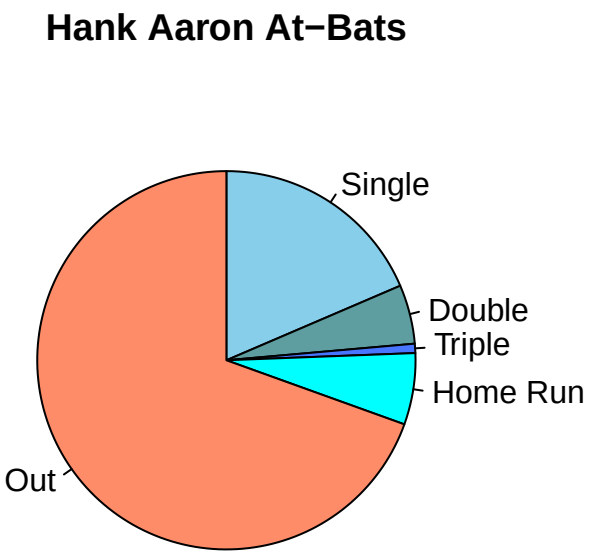
Part (c)

Construct a pie chart for the data from Problem 4.

Solution

```
pie(
  x = table( hank.aaron.at.bats.factor ),
  main = "Hank Aaron At-Bats",
```

```
  clockwise = TRUE,
  col =
    c(
      "skyblue",
      "cadetblue",
      "royalblue1",
      "cyan",
      "salmon1"
    )
)
```



Part (d)

Using the pie chart from part (c), can you easily determine Hank Aaron's batting average? Don't do any calculations in your head – just by looking at the shapes, can you make a reasonable estimate of Hank Aaron's batting average? Explain your answer with one or two sentences.

Solution

Now, with the pie chart, I *do* feel that it's possible to get a pretty reasonable sense of Hank Aaron's batting average just by looking at the graph, because it's immediately clear that the hits make up almost exactly 1/3 of all at-bats, although it seems to be a little less that.

By the way, let's check this very quickly – after all, we have the data in `problem.5.data`. Can we calculate the exact value of Hank Aaron's batting average? This wasn't part of the problem statement, and you're not required to do this, but I think it's an interesting quick problem to do.

Try to do this by yourself if you can: using the data in `problem.5.data`, calculate Hank Aaron's batting average.

Hank Aaron's at-bats are:

```
hank.aaron.at.bats <-
    length( problem.4.data )

hank.aaron.at.bats
```

```
## [1] 12364
```

We can use logical indexing to count the number of hits.

At first glance, this might seem complicated, because there are multiple different categories of hits.

However, there's a sneaky coding trick: since the alternative to a hit is an out, we can count all the items that are hits by counting all the things that are *not* outs:

```
hank.aaron.hits <-
    sum( problem.4.data != "Out" )
```

Now we can calculate Hank Aaron's batting average:

```
hank.aaron.batting.average <-
    hank.aaron.hits /
    hank.aaron.at.bats

cat(
    "Hank Aaron batting average:",
    formatC(
        hank.aaron.batting.average,
        format = "f",
        digits = 3
    )
)
```

```
## Hank Aaron batting average: 0.305
```

So Hank Aaron's career batting average was 0.305 (you can check this on the Internet). My original estimate based on the pie chart was about 1/3, or a batting average of 0.333. So, just based on the visual inspection of the pie chart, it was possible to estimate Hank Aaron's batting average to within 3 percentage points. Pretty good!

The point here is that a batting average is a *relative proportion*, and pie charts are good at displaying such quantities.

Part (e)

Comparing your answers from parts (b) and (d), do you agree with people who think we should always use barplots insted of pie charts? Explain your answer with one or two sentences.

Solution

Personally, I don't agree with this approach. I think barplots are very valuable, and can do many things that pie charts can't, but the reverse is true as well. The important thing to understand is that you have to think carefully about what information you're trying to convey, and then select the visualization method that best displays that information.

Again, there's no "right" or "wrong" answer here, in the sense that the TAs are going to grade you based on your response. As long as you write something vaguely sensible and use complete sentences we'll give you full credit. However, I think you'll get a lot more out of the problem if you engage with the ideas. Look at the graphs, think about Hank Aaron's batting average, and then decide for yourself which method you prefer.