Sampling distributions part 1 (proportions)

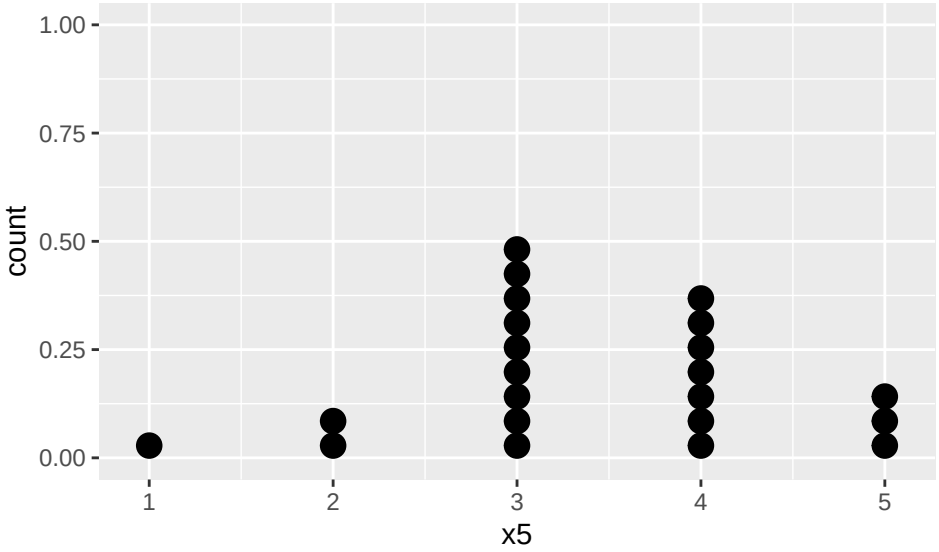
Thomas Scofield

Feb. 20, 2020

In class you were given paper bags containing bottle caps, an unknown proportion of which were green. You drew from the bag, $n = 5$ times with replacement, and counted the number of green caps. The reported numbers are placed in a vector called `x5`, with the accompanying dotplot showing the distribution for the counts of greens:

```
x5 <- c(3,3,5,4,3,4,3,4,1,2,4,4,3,4,3,5,3,5,3,4,3,2)
gf_dotplot(~x5)
```

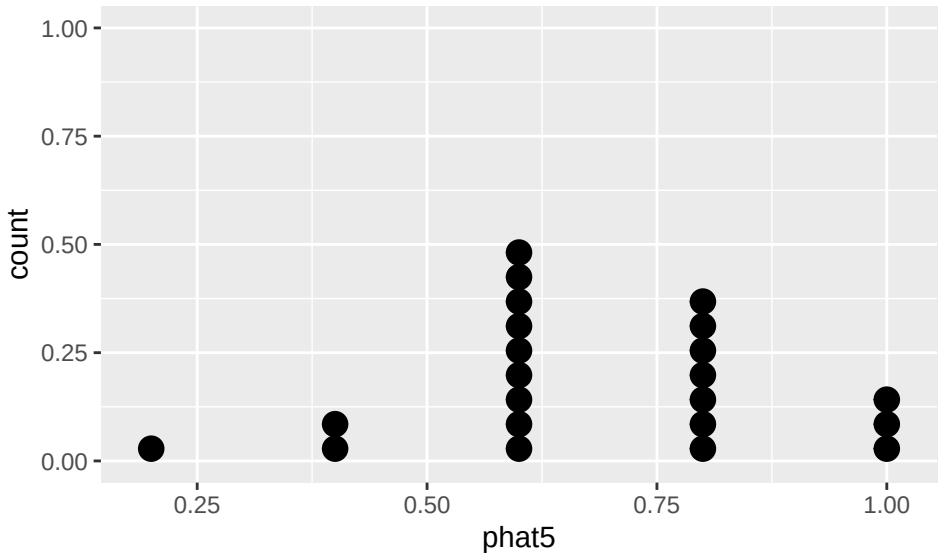
```
## `stat_bindot()` using `bins = 30`. Pick better value with `binwidth`.
```



While I did not ask you to report the *proportion* of greens in each sample, it is easy to convert from counts to proportions. You just divide by 5.

```
phat5 <- x5 / 5
gf_dotplot(~phat5)
```

```
## `stat_bindot()` using `bins = 30`. Pick better value with `binwidth`.
```



Look carefully at the two dotplots, and see if you understand both

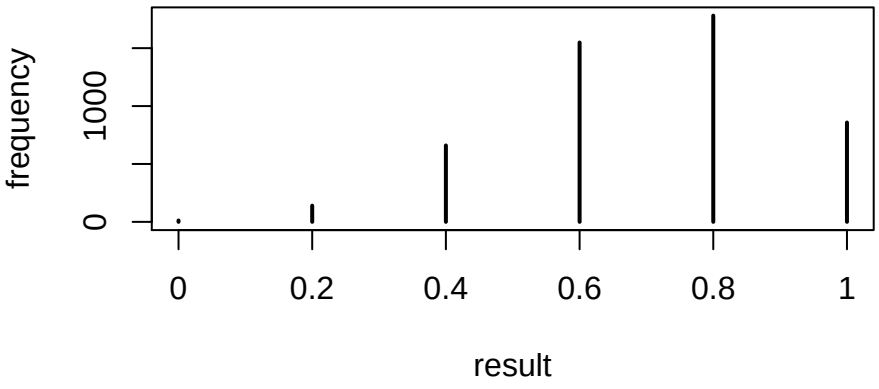
- why they are so similar, and
- what makes them different.

The variable you took note of was `color`, in this case a **binary** categorical variable (because there were only two colors of bottle caps in the bag). If we sampled 5 times and drew 3 greens, we would say our sample proportion is $\hat{p} = 0.6$. This is a value that can vary from sample to sample, and such things are called **sample statistics**. (Note: The *count* of green caps in 5 draws is an alternate sample statistic.)

You kept yourself in the dark about the true proportion of green caps in the bag. It was, in fact, 7/10. That is a value that is inherent to the population, not varying from sample to sample. We call it a **population parameter**. When our population parameter is a proportion, we denote it as p. So, in our population, $p = 0.7$.

The previous dotplot of sample proportions begins to give us a sense of what sorts of values occur often for $\hat{p}$, and what values are rarer. That is, they give us a sense of the **sampling distribution of the sample statistic** $\hat{p}$ when $n = 5$ and $p = 0.7$. Simulation is our best method, albeit an imperfect one, to get at this sampling distribution, but our approximation to it (the dotplot above) is pretty rough. We can greatly improve the approximation by adding another 5000 or so members to our class and including their contributed numbers, as well. That is what the code below attempts to mimick.

```
bag = c( rep("G", 7), rep("B", 3) )
phat5sim <- do(5000) * (sum(sample(bag, size=5, replace=T) == "G") / 5)
plot(tally(~result, data=phat5sim), ylab="frequency")
```



This, too, is an approximation, based on simulated results that use a random number generator. Nevertheless, from one simulation involving 5000 runs to the next, the plot doesn't change much; it's a fairly accurate picture of the sampling distribution for $\hat{p}$. Such a plot conveys to us that instances of $\hat{p} = 0$ or $\hat{p} = 0.2$ can arise, but are quite rare in comparison to the other possibilities. How rare? We can use `filter()` and `nrow()` commands to quantify it.

```
nrow(filter(phat5sim, result <= 0.2))
```

```
## [1] 152
```

which tells us there were just 152 instances out of 5000 samples for which $\hat{p}$ was this low.

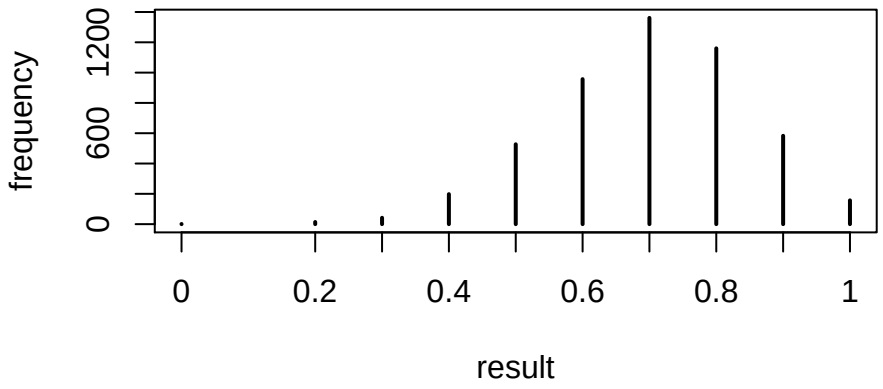
In class I asked you to repeat the exercise, now drawing $n = 10$ times. Your reported results, converted to proportions, were

```
phat10 <- c(9,4,6,7,7,8,4,10,8,8,7) / 10
phat10
```

```
##  [1] 0.9 0.4 0.6 0.7 0.7 0.8 0.4 1.0 0.8 0.8 0.7
```

In what follows, I build a larger (simulated) collection of $\hat{p}$, as if there had been 5000 students in our class, all drawing out bottle caps (with replacement) $n = 10$ times. The result is an approximate sampling distribution of $\hat{p}$ when $n = 10$ and $p = 0.7$:

```
phat10sim <- do(5000) * (sum(sample(bag, size=10, replace=T) == "G") / 10)
plot(tally(~result, data=phat10sim), ylab="frequency")
```



In comparison with the previous simulation (samples of size $n = 5$), there is a greater variety of $\hat{p}$-values when $n = 10$, but they also do not extend as far down on the low end.

Observations

1. The sampling distributions for $\hat{p}$ are different when different sample sizes are used. In particular, a larger sample size leads to a distribution which is less spread out. We can quantify this by computing the standard deviation for the sampling distribution. To get it for $\hat{p}$ when $n = 5$:

```
sd(~result, data=phat5sim)
```

```
## [1] 0.2053478
```

The corresponding standard deviation when $n = 10$ is

```
sd(~result, data=phat10sim)
```

```
## [1] 0.1471367
```

The standard deviation of a sampling distribution is usually referred to as **standard error**. We see the standard error for $\hat{p}$ is larger, near 0.205, for samples of size $n = 5$ than when $n = 10$, where the standard error is near 0.147.

2. We can calculate the mean of the sampling distribution. For samples of size $n = 10$, it is roughly

```
mean(~result, data=phat10sim)
```

```
## [1] 0.69894
```

which is the same (or very close) as the population proportion $p = 0.7$. When the mean of a sample statistic is the same as the population parameter it serves to estimate, we say that sample statistic is an **unbiased** estimator of the parameter.