# 9. Distributions

| | *Group I* | *Group II* | *Group III* | *Group IV* |
|---|---|---|---|---|
| *# of Packets* | | | | |
| *Mean # of M&Ms* | | | | |
| *St. Dev of # of M&Ms* | | | | |
| *Overall # of M&Ms* | | | | |
| *Overall # Non-Red M&Ms* | | | | |
| *Proportion Non-Red* | | | | |

• What is the population from which these samples are drawn?

• Why is there a different answer for each group for the proportion of non-red M&M's?

• What do you suppose the results would look like if we did this experiment with 20,000 groups?

**Experiment.** Suppose that the true population proportion is $\boldsymbol{p} = 0.75$. Take a sample of size 100 from this population and compute $\hat{\boldsymbol{p}}$. Record the result.

Now take another sample of size 100 from the population; again compute $\hat{\boldsymbol{p}}$ and write down the result. You now have computed two (probably different) values for $\hat{\boldsymbol{p}}$.

Repeat the process a third time, getting a third computed value for $\hat{\boldsymbol{p}}$.

Continue until you have 20,000 computed values for $\hat{\boldsymbol{p}}$, each based on a randomly selected sample of size 100.

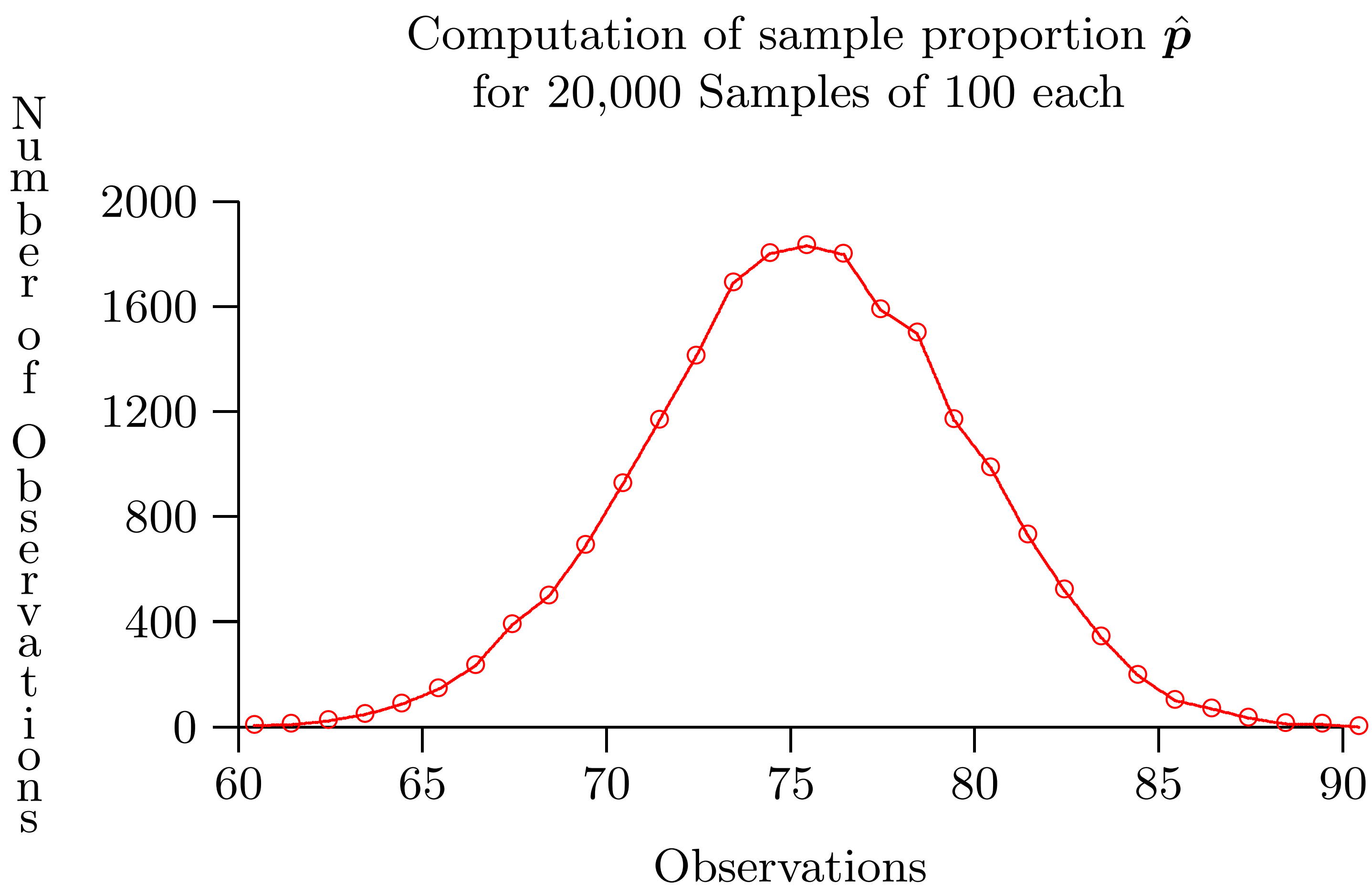
Make a frequency table of the results.

*What do you suppose that a graph of the results would look like?*

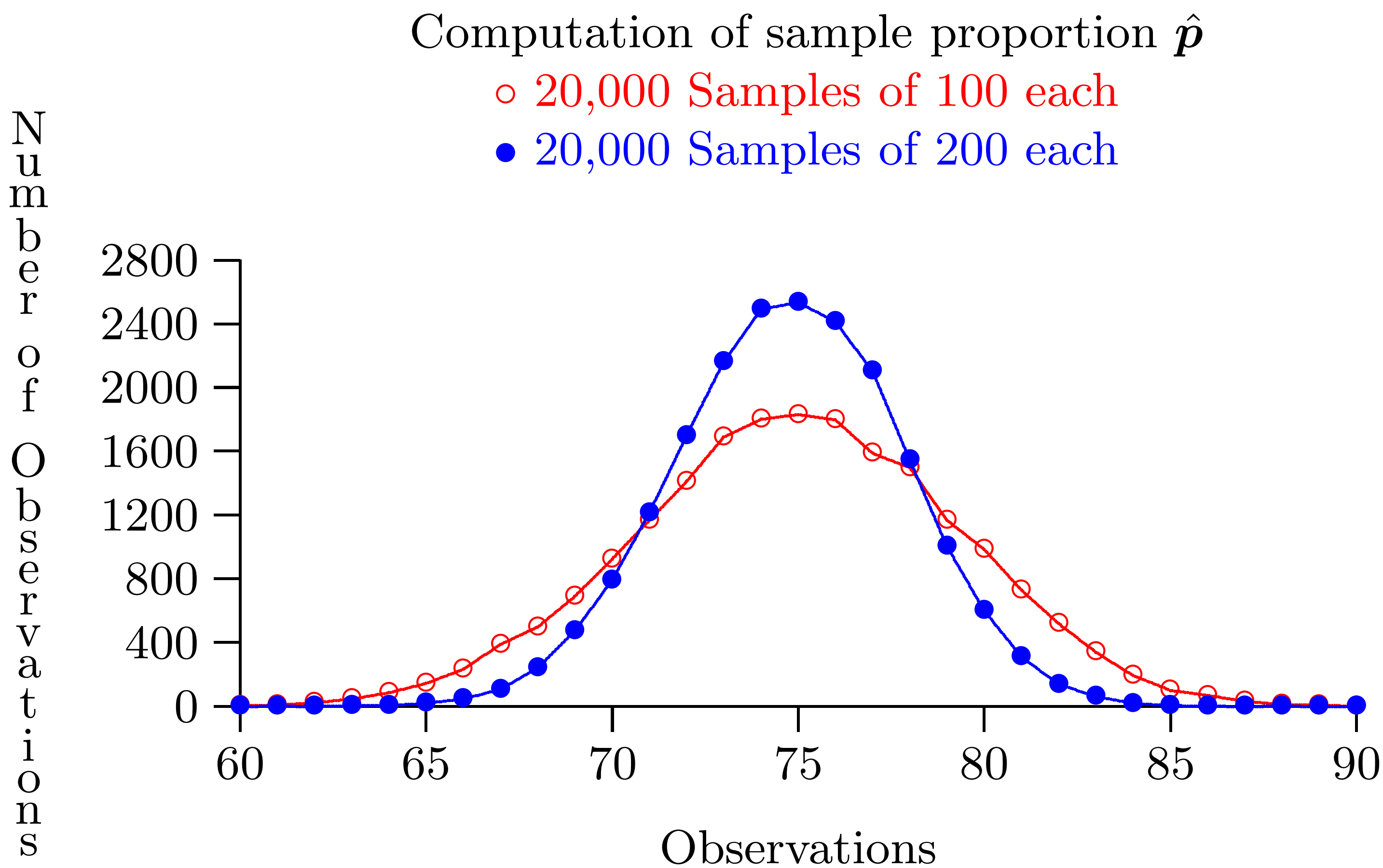
I performed this experiment (using a random number generator on a computer to simulate taking the samples and computing $\hat{p}$). I got the following frequency table:

| observed % | frequency |
|---|---|
| 60 | 6 |
| 61 | 9 |
| 62 | 23 |
| 63 | 47 |
| 64 | 86 |
| 65 | 144 |
| 66 | 233 |
| 67 | 389 |
| 68 | 498 |
| 69 | 690 |
| 70 | 924 |
| 71 | 1166 |
| 72 | 1410 |
| 73 | 1689 |
| 74 | 1801 |
| 75 | 1831 |
| 76 | 1798 |
| 77 | 1588 |
| 78 | 1498 |
| 79 | 1168 |
| 80 | 986 |
| 81 | 730 |
| 82 | 520 |
| 83 | 341 |
| 84 | 195 |
| 85 | 101 |
| 86 | 69 |
| 87 | 34 |
| 88 | 12 |
| 89 | 10 |

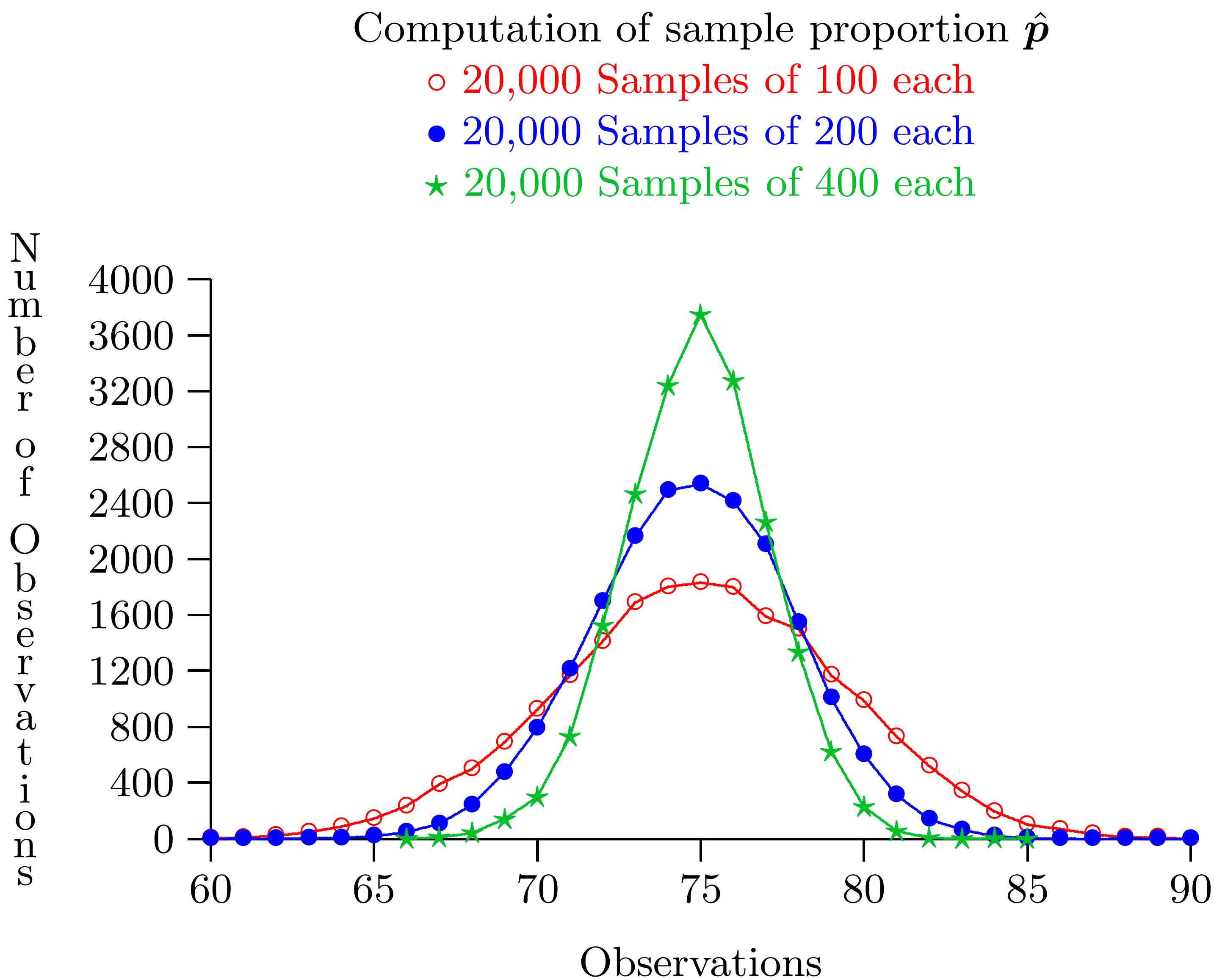
A plot of this data reveals




Suppose that you double sample size from 100 to 200, repeating again with 20,000 sample?




Note that the larger samples are correct more often (taller in the middle) and incorrect less often (lower in the tails). However doubling the sample size does not increase the peak by two.


Computation of sample proportion $\hat{p}$
20,000 Samples of 100 each
20,000 Samples of 200 each
20,000 Samples of 400 each
Number of Observations
4000
3600
3200
2800
2400
2000
1600
1200
800
400
0
60
65
70
75
80
85
90
Observations

Observations:

- The distribution of repeated samples is approximately normal.
- Larger sample sizes are “more accurate” than smaller sample sizes.
- Larger samples have less variability than smaller samples.
- In order to double the accuracy you must quadruple the sample size.

Note that we have found a sample of 20,000 sample proportions. In principle it is possible to keep doing this forever. We have discovered that the population of all possible sample proportions is approximately normally distributed. Since it is normally distributed it must have an associated mean and standard deviation. We’ve also discovered that there is less variability in larger samples and quadrupling the sample size doubles the accuracy.

The Central Limit Theorem summarizes these results.

## 9.1. Central Limit Theorem

For large sample sizes, the distribution of the sample mean and the sample proportion are approximately normal. In particular, for large sample sizes the sample mean $\bar{x}$ is approximately normal with mean $\mu$ and standard deviation

$$\text{sampling standard deviation of } \bar{x} = \frac{\sigma}{\sqrt{n}}$$

while $\hat{p}$ is approximately normal with mean $p$ and

$$\text{sampling standard deviation of } \hat{p} = \sqrt{\frac{p(1-p)}{n}}$$