Inference:

What can we infer about population ***parameters*** *from sample* ***statistics****?*

(*) Intuition: If the sample is 'good', then the sample statistic should be close to the corresponding population parameter.

⇒ The sample percentage of *brown tribbles* should be close to the population percentage of *brown tribbles*.

⇒ The sample average weight of a *tribble* should be close to the population average weight of a *tribble*.

Etc.

(*) A 'good' sample is one that matches the population in all/most meaningful ways. It is 'representative'.

(*) Simple random samples produce good samples *almost always*.

(*) A **point estimate** is a (single) number that estimates the population parameter in question.

Example 1. In a simple random sample of 1100 U.S. voters, 56% favored a constitutional amendment banning selfies from social media.

$\Rightarrow$ 56% is a point estimate for the population percentage of voters that favor this amendment.

(*) It is almost certain that the true population percentage is something other than the sample percentage.

(*) It is also almost certain that the true population percentage is *close to* the sample percentage. This leads to the idea of an *interval estimate*.

Definition: A 95%-confidence interval for the population parameter α is an interval of the form $\mathcal{I} = (A - \varepsilon, A + \varepsilon)$, where

- A is the sample statistic that corresponds to α.
- ε is a *margin of error*, a *give-or-take* number that accounts for *chance error*.
- There is a 95% chance that the interval $\mathcal{I}$ contains α.

Constructing a 95%-confidence interval for population percentage:

Suppose that $\hat{p} \times 100\%$ is the ***sample*** percentage of *blank*, in a simple random sample of size n, taken from the population.

$\Rightarrow$ $\hat{p} \times 100\%$ is a *point estimate* for $p \times 100\%$, the ***population*** percentage of *blank*.

$$\Rightarrow \quad SE_{\%} = \frac{\sqrt{p(1-p)}}{\sqrt{n}} \times 100\% \approx \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} \times 100\%$$

is the ***estimated*** *standard error for percentage*

$\Rightarrow$ If n is sufficiently large, then

$$P(\overbrace{(\hat{p} \times 100\%)}^{\text{sample }\%} - 2SE_{\%} < \overbrace{(p \times 100\%)}^{\text{population }\%} < \overbrace{(\hat{p} \times 100\%)}^{\text{sample }\%} + 2SE_{\%}) \approx 95\%,$$

according to the ***normal approximation***, so that the interval

$$\mathcal{I} = ((\hat{p} \times 100\%) - 2SE_{\%}, (\hat{p} \times 100\%) + 2SE_{\%}) = (\hat{p} \times 100\%) \pm 2SE_{\%}$$

is a 95%-confidence interval for the population percentage of *blank*.

Example 1. (cont.) The sample percentage of U.S. voters who favor the *no-selfie-on-social-media* amendment is 56%.

(*) The standard error for percentage is estimated to be

$$SE_{\%} \approx \frac{\sqrt{0.56 \cdot 0.44}}{\sqrt{1100}} \times 100\% \approx 1.5,$$

(*) A 95% confidence interval for the population percentage of voters who favor the no-selfie amendment is

$$56\% \pm 3\% = (53\%, 59\%).$$

Example 1. (cont. some more) In the same survey, 23% of the voters said they prefer Burger King to McDonalds. Find a 95%-confidence interval for the percentage of all US voters who prefer BK to McD.

(*) The $SE_{\%}$ is estimated by $SE_{\%} \approx \frac{\sqrt{0.23 \times 0.77}}{\sqrt{1100}} \times 100\% \approx 1.27\%$.

(*) A 95% confidence interval for the percentage of all US voters who prefer BK to McD. is given by

$$23\% \pm 2.54\%.$$

What does "95%-confidence" *mean*?

(*) The population percentage of *whatever* is unknown but ***fixed***.

(*) Different samples generally produce different sample data — e.g., different sample percentages. Therefore different samples will produce different 95%-confidence intervals — though most of them will be very similar to each other.

(*) The term "95%-confidence" means that if we were to construct all possible 95% confidence intervals (for a given sample size), then 95% of these intervals would contain the (unknown) population percentage.

(*) Taking one simple random sample and constructing just one confidence interval is like randomly choosing one of all possible such intervals, so it has a 95% chance of being one of the good ones — an interval that contain the true population percentage.

Observations:

1. To increase the *likelihood* that a sample interval contains the population parameter, we can increase the margin of error.

E.g., the interval (sample %) $\pm 3SE_{\%}$ is a 99.7%-confidence interval for the population percentage (assuming a simple random sample and large enough sample size).

2. To increase the *accuracy* of the estimate, we decrease the margin of error...

How can we both decrease the margin of error and increase the likelihood that the resulting interval contains the true value?

Increase the sample size!

Example 2. A second simple random sample of 13247 US voters was surveyed, and 57.1% of those surveyed supported the no-selfie amendment.

(*) The (new) $SE_{\%}$ is estimated to be

$$SE_{\%} = \frac{\sqrt{0.571 \times 0.429}}{\sqrt{13247}}\% \approx 0.43\%.$$

(*) A 95% confidence interval for the percentage of all US voters who favor the no-selfies amendment is

$$57.1\% \pm 0.86\% = (56.24\%, 57.96\%).$$

(*) A 99.7% confidence interval for the percentage of all US voters who favor the no-selfies amendment is

$$57.1\% \pm 1.29\% = (55.81\%, 58.39\%).$$

Next... Estimating averages.

(*) The expected value and standard error of the sum of n tickets drawn at random with replacement from a box of numbered tickets are

$$EV(\text{sum}) = (\text{Average of box}) \times n \quad \text{and} \quad SE(\text{sum}) = SD(\text{box}) \times \sqrt{n}.$$

The average of the draws is the sum of the draws divided by n, so...

(*) The expected value of the average of n tickets drawn at random with replacement from a box of numbered tickets is

$$EV(\text{Avg}) = \frac{(\text{Average of box}) \times n}{n} = \text{Average of box}.$$

Likewise

(*) The standard error for the average of the draws is

$$SE(\text{Avg}) = \frac{SE(\text{sum})}{n} = \frac{SD(\text{box}) \times \sqrt{n}}{n} = \frac{SD(\text{box})}{\sqrt{n}}.$$

Observations:

- The SE for the ***sum*** of n draws from a given box of numbered tickets is the standard deviation of the (hypothetical) box of ***all possible sums of n draws*** from the original box.
- As the number of draws (n) grows larger, there will be *more variation* in the observed sums. This means that the SD of the box-of-sums increases...

$\Rightarrow$ The SE for the sum of the draws ***increases*** with the number of draws:

$$\boxed{SE(\text{sum of } n \text{ draws from box}) = \sqrt{n} \times (\text{SD of the box})}$$

(more) Observations:

- The SE for the ***average*** of n draws from a given box of numbered tickets is the standard deviation of the (hypothetical) box of ***all possible averages of n draws*** from the original box.

- As the number of draws (n) grows larger, there will be *less variation* in the observed averages. This means that the SD of the box-of-averages decreases...

⇒ The SE for the average of the draws ***decreases*** with the number of draws:

$$\boxed{SE(\text{avg. of } n \text{ draws from box}) = \frac{\text{SD of the box}}{\sqrt{n}}}$$

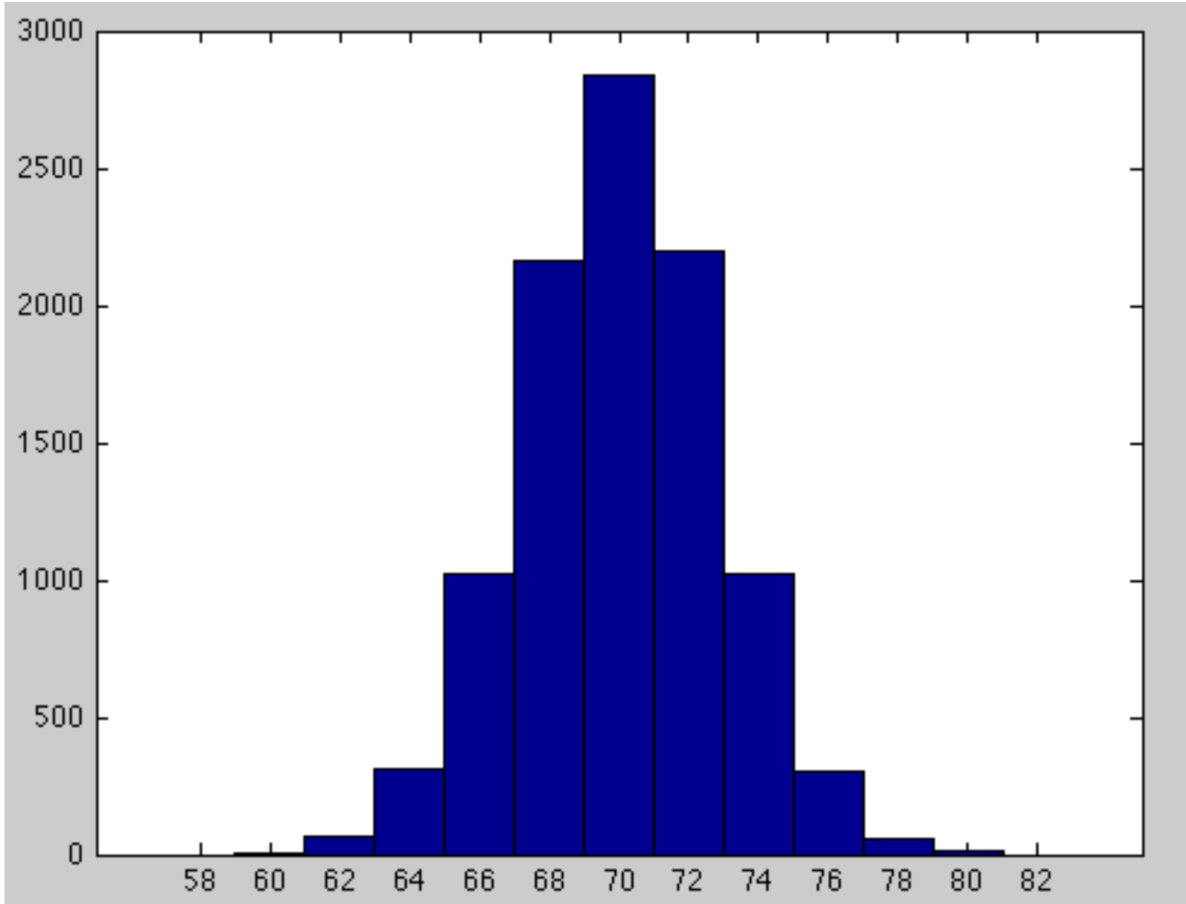
- ***Important:*** The SE for the sample average is ***not*** a measure of variation in the original box (population) — it is a measure of the variation in the sample averages across all samples.

Comments:

- The formulas given above for the SE(sum) and SE(avg) are exactly correct when the draws are done *with replacement*. If the draws are done *without replacement*, then the standard errors tend to be smaller, but if the number of draws is small compared to the size of the box, the difference is negligible.
- The ***Central Limit Theorem*** tells us that the distribution of sample averages of n draws from a box of numbered tickets is well-approximated by the normal distribution if...
    - The observed sample averages are converted to standard units:

$$\frac{\text{(sample average)} - \text{(expected average)}}{SE(\text{avg})}.$$

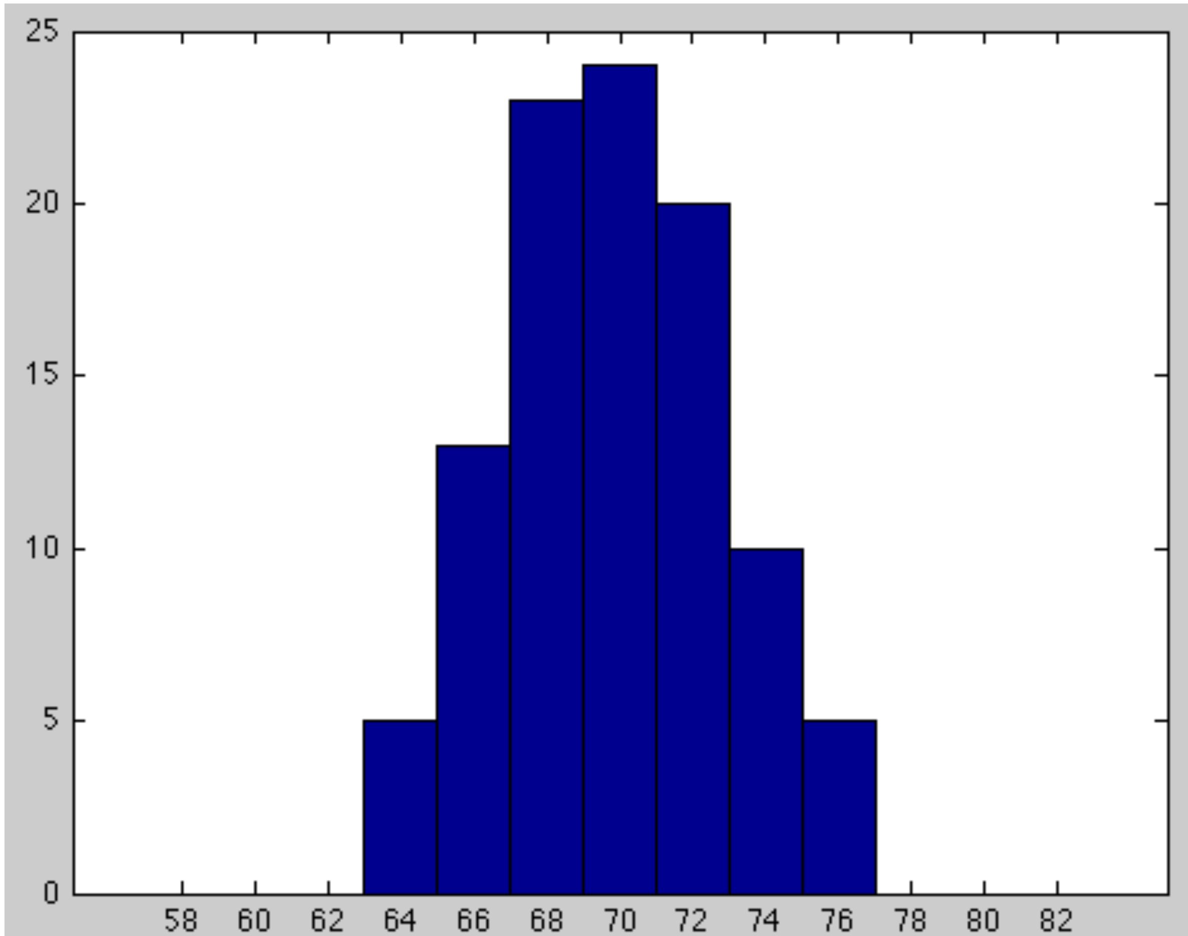
    - The number of draws (n) is large enough.
- Recall: the ***expected average*** is the same as the average of the box.



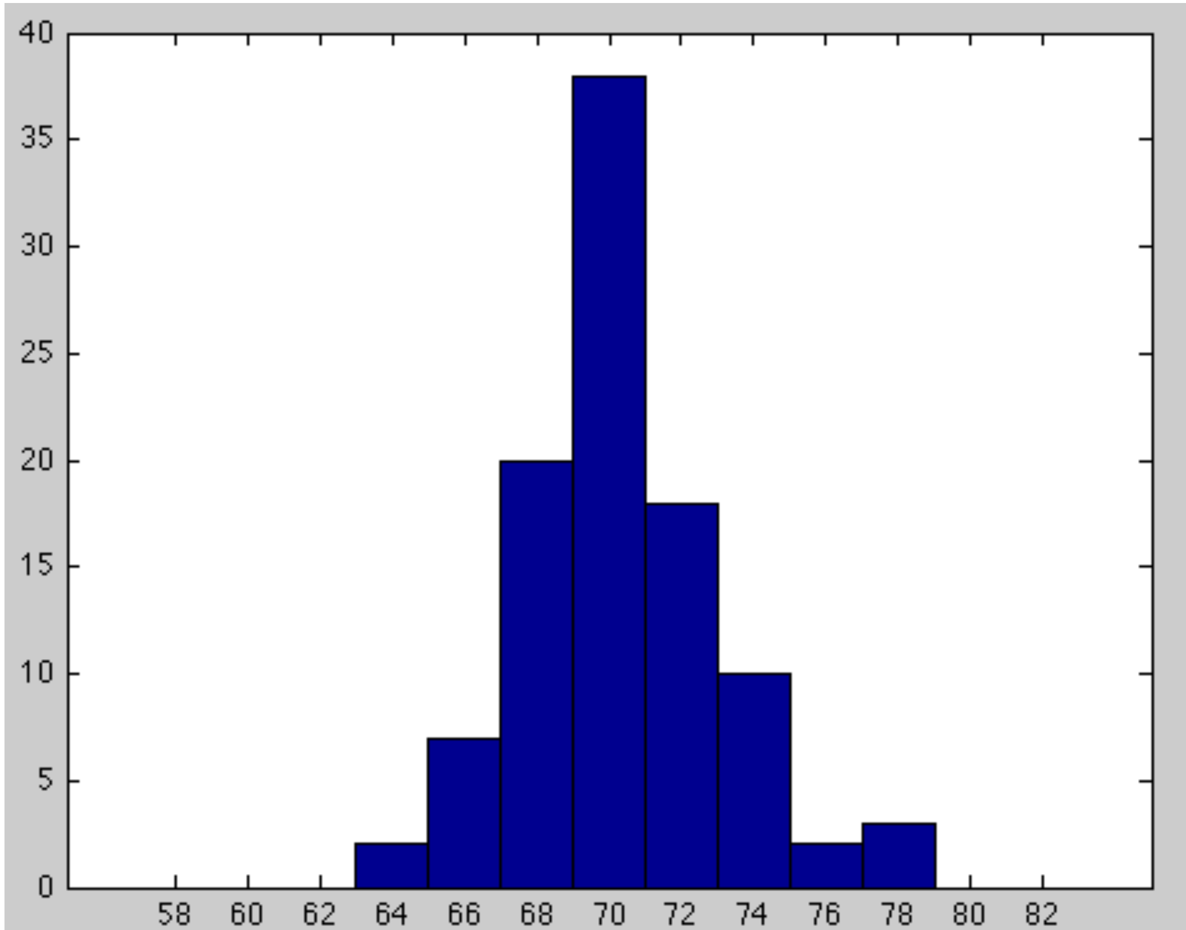
Data histogram for a population.

$$\boldsymbol{mean} = 70.003, \qquad \boldsymbol{SD} = 2.8032.$$

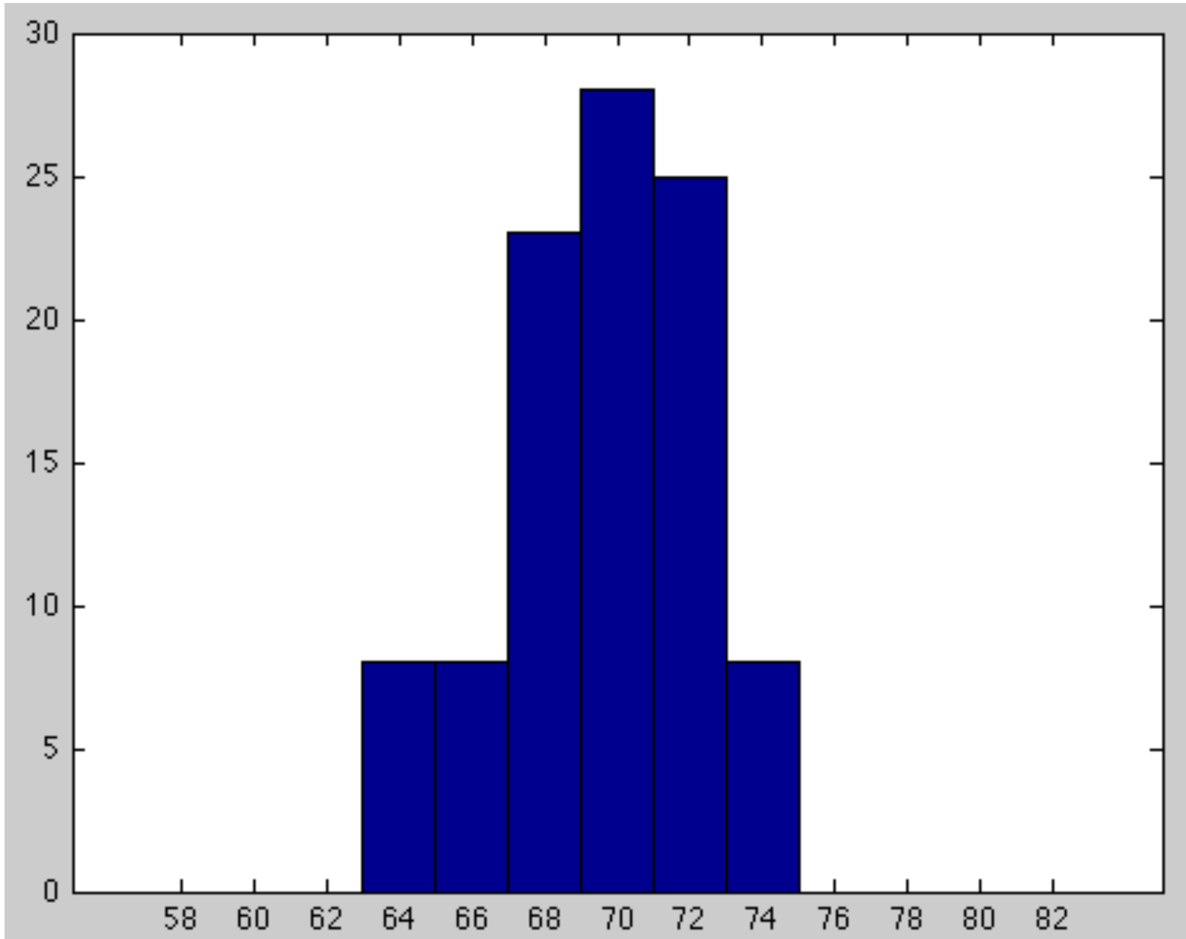
Data histograms for four simple random samples of size 100:



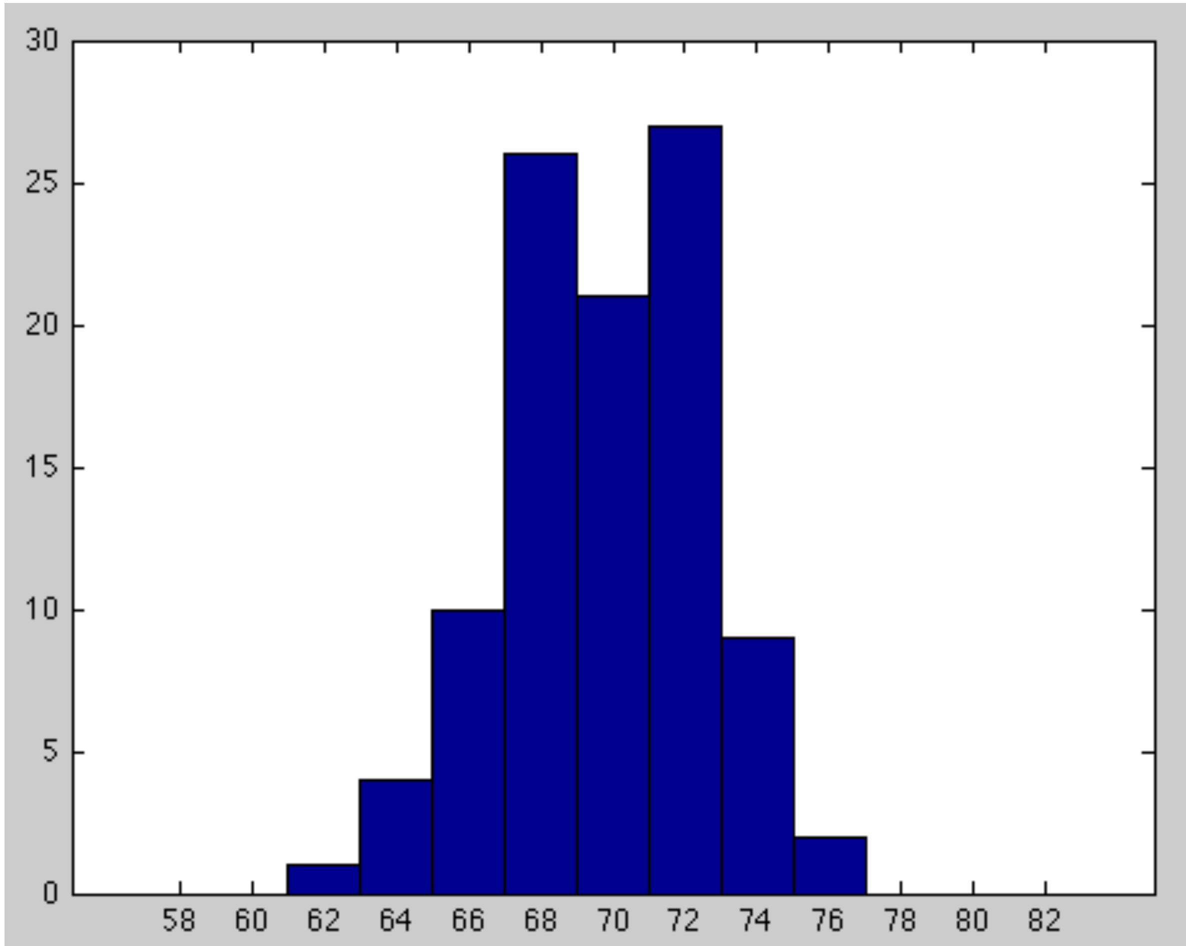
Sample 1: Mean = 69.77, SD $\approx$ 3.



Sample 2: Mean = 70.22, SD $\approx$ 2.75.

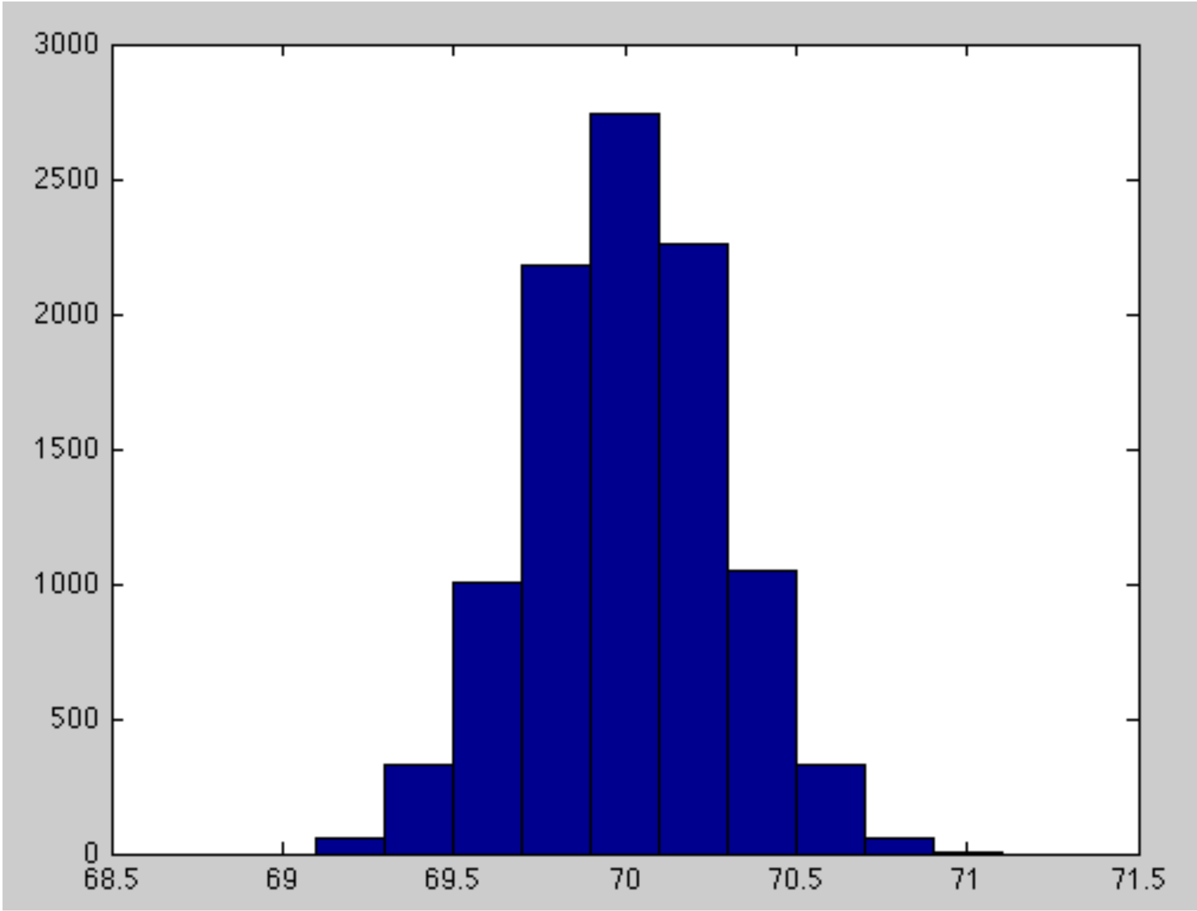


Sample 3: Mean = 69.52, SD $\approx$ 2.69.



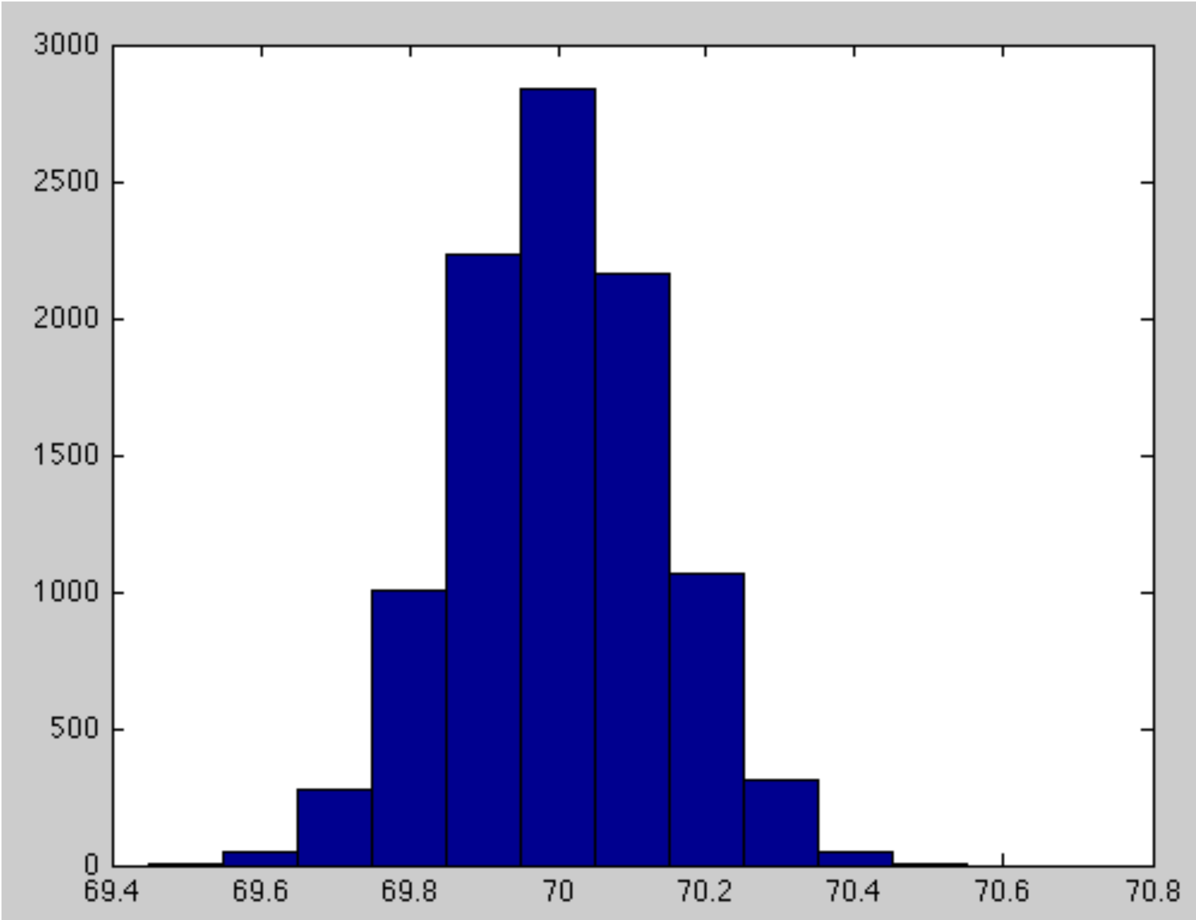
Sample 4: Mean = 69.89, SD $\approx$ 2.83.

Histogram for distribution of averages of 10,000 samples of size 100:



Mean= 70.0037**, SD**= 0.2797

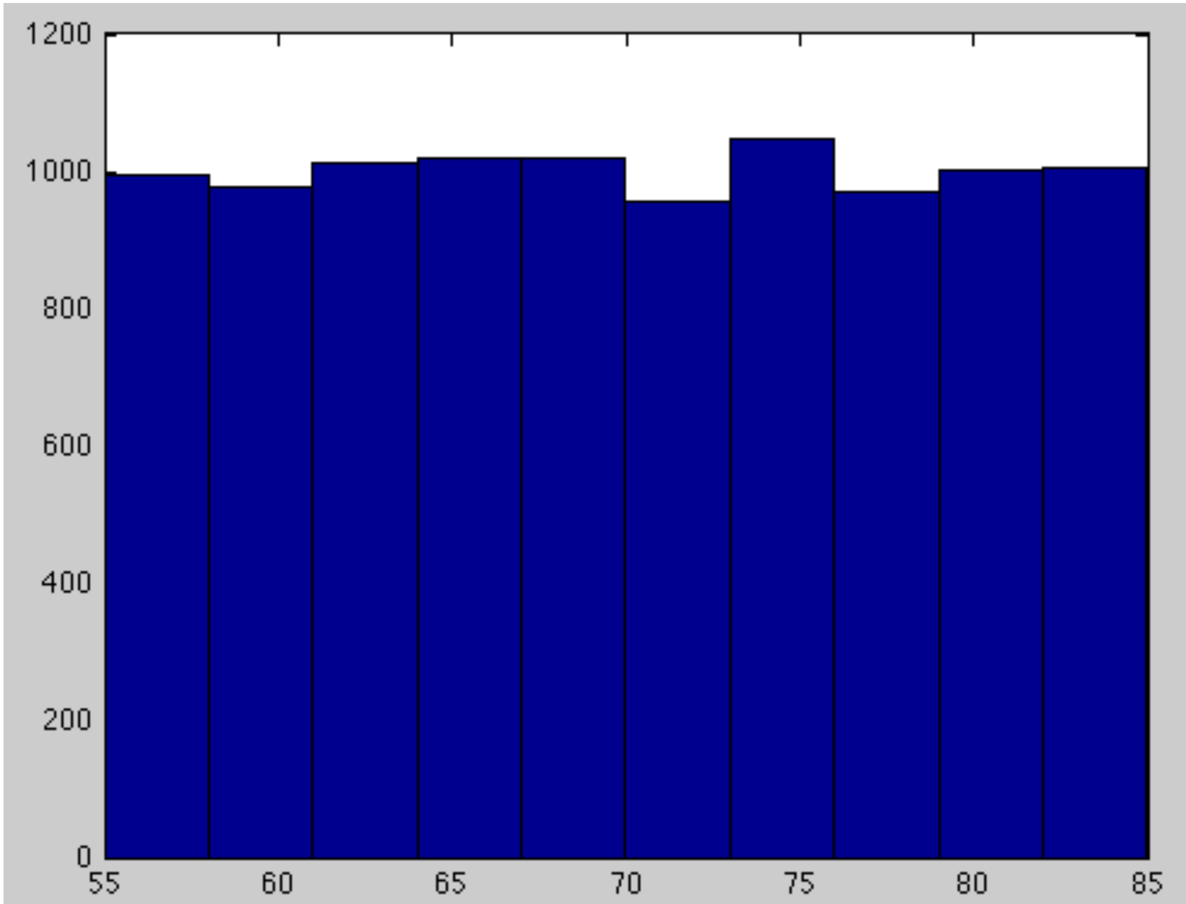
Histogram for distribution of averages of 10,000 samples of size 400:



Mean$= 70.0017$**, SD** $= 0.1376$

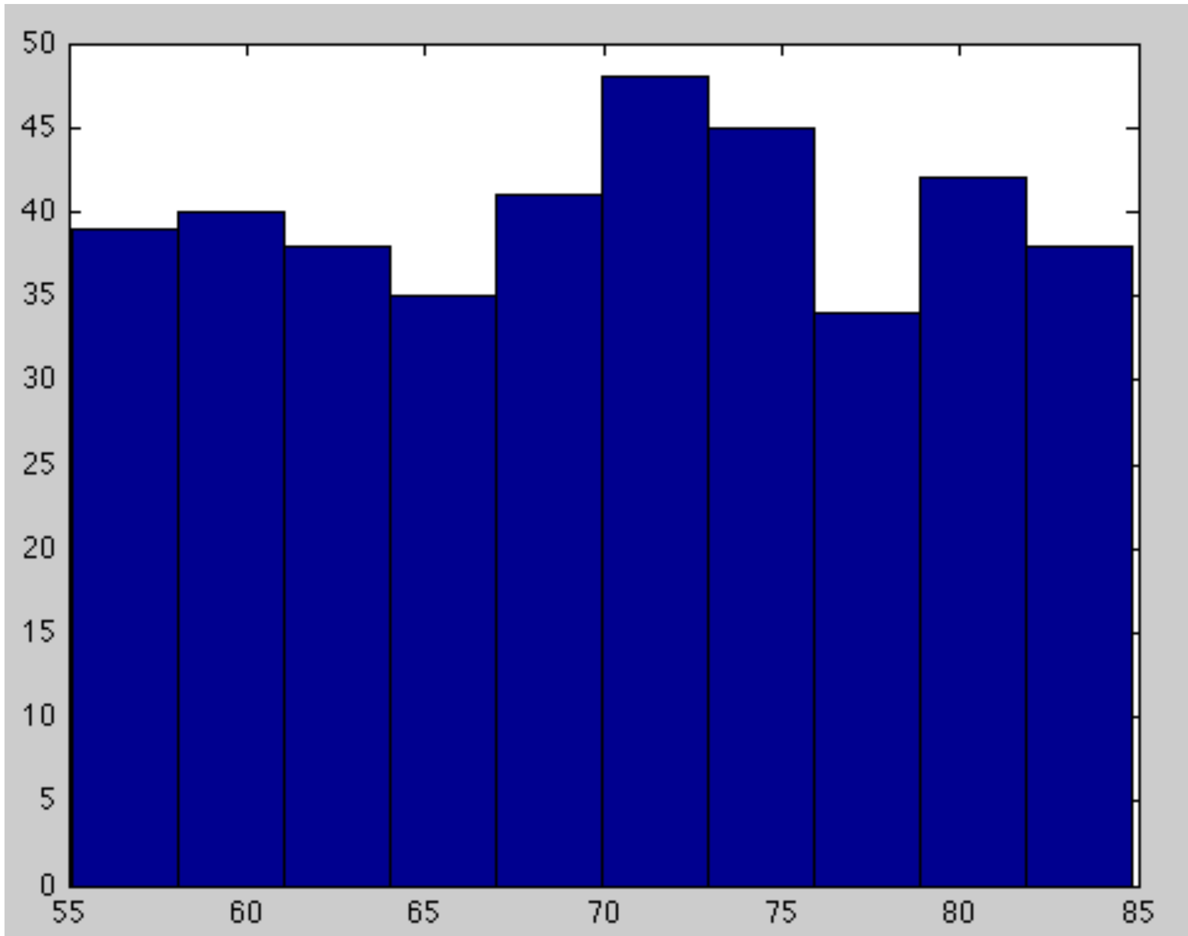
Example 2:

Data histogram for another population

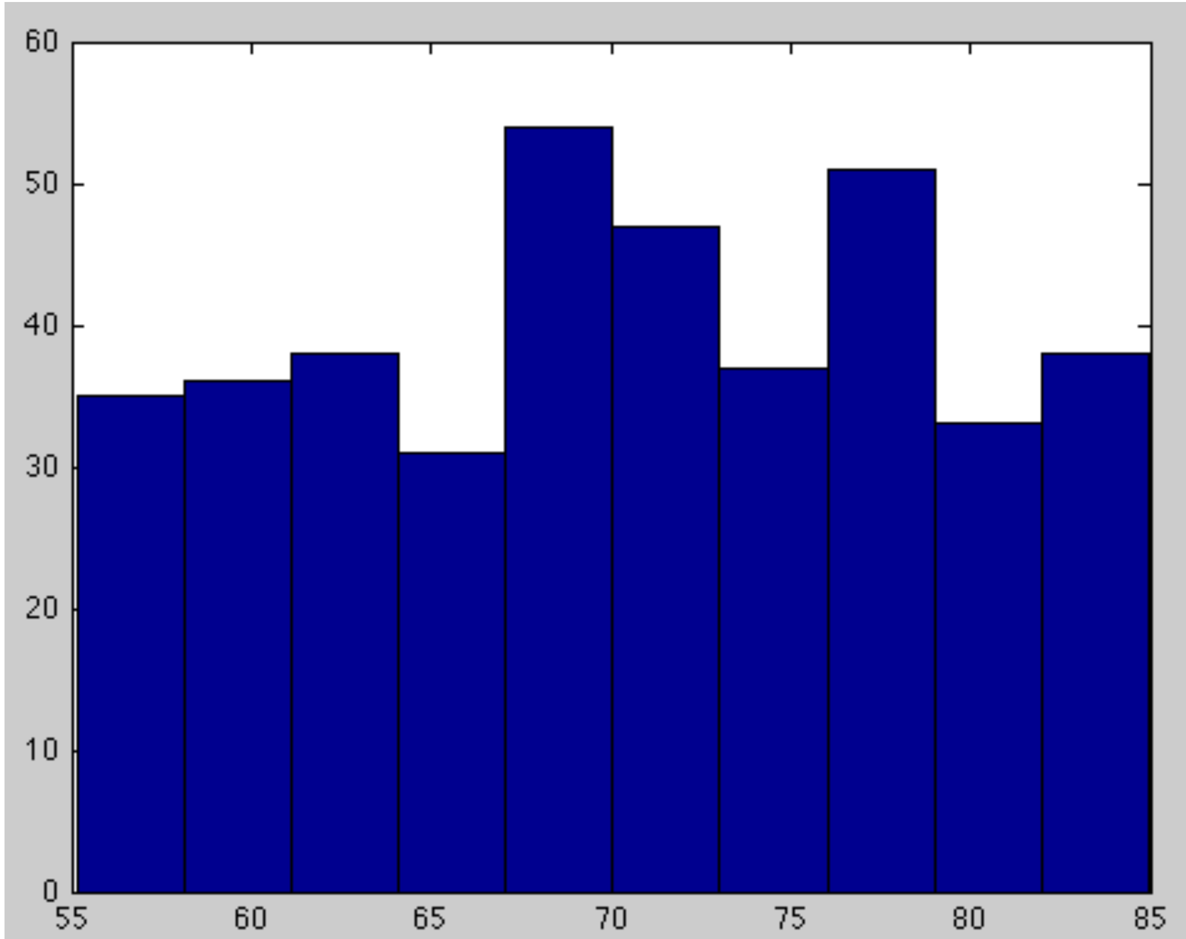


mean $= 70.046$***,*** ***SD*** $= 8.657$ ***(why is the SD bigger?).***

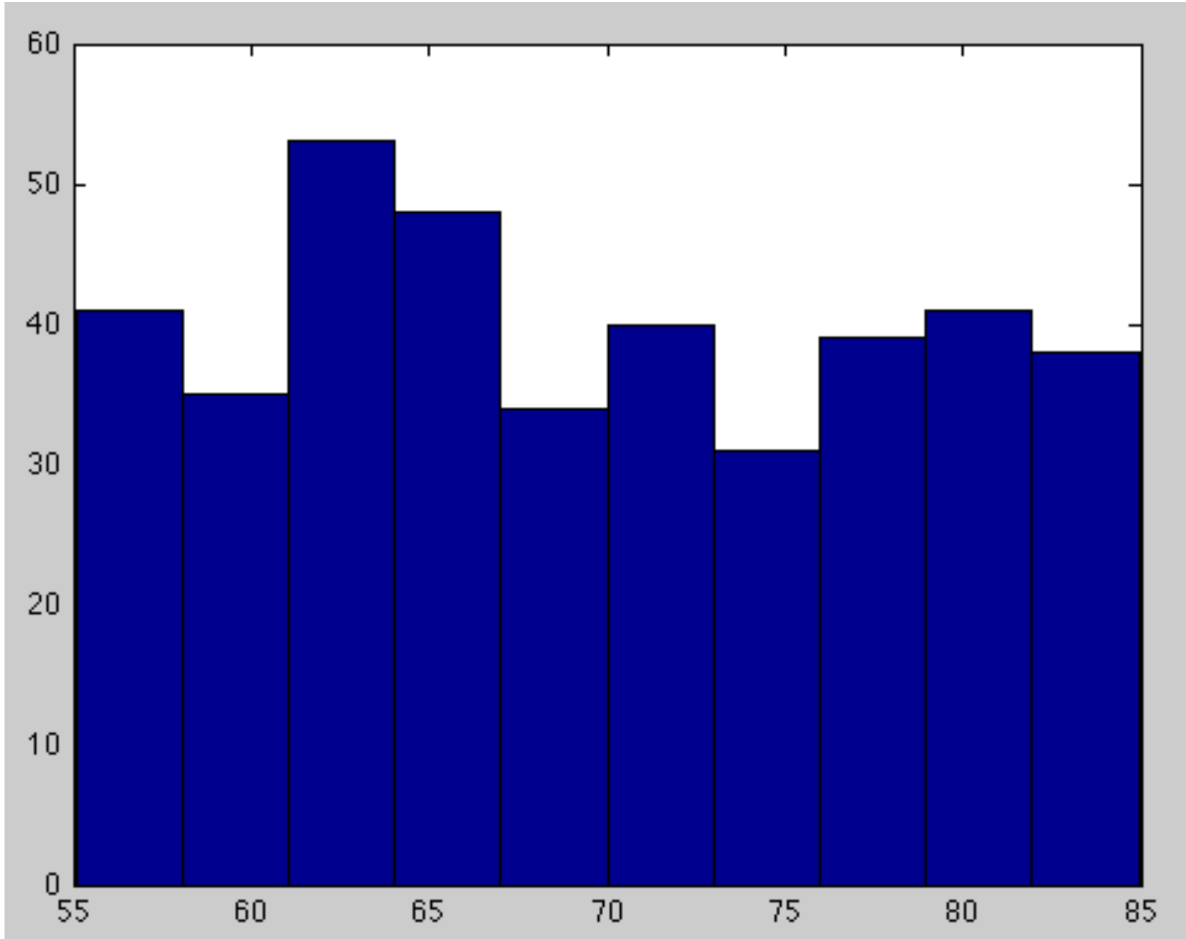
Data histograms for four simple random samples of size 400 taken from the uniform distribution above...



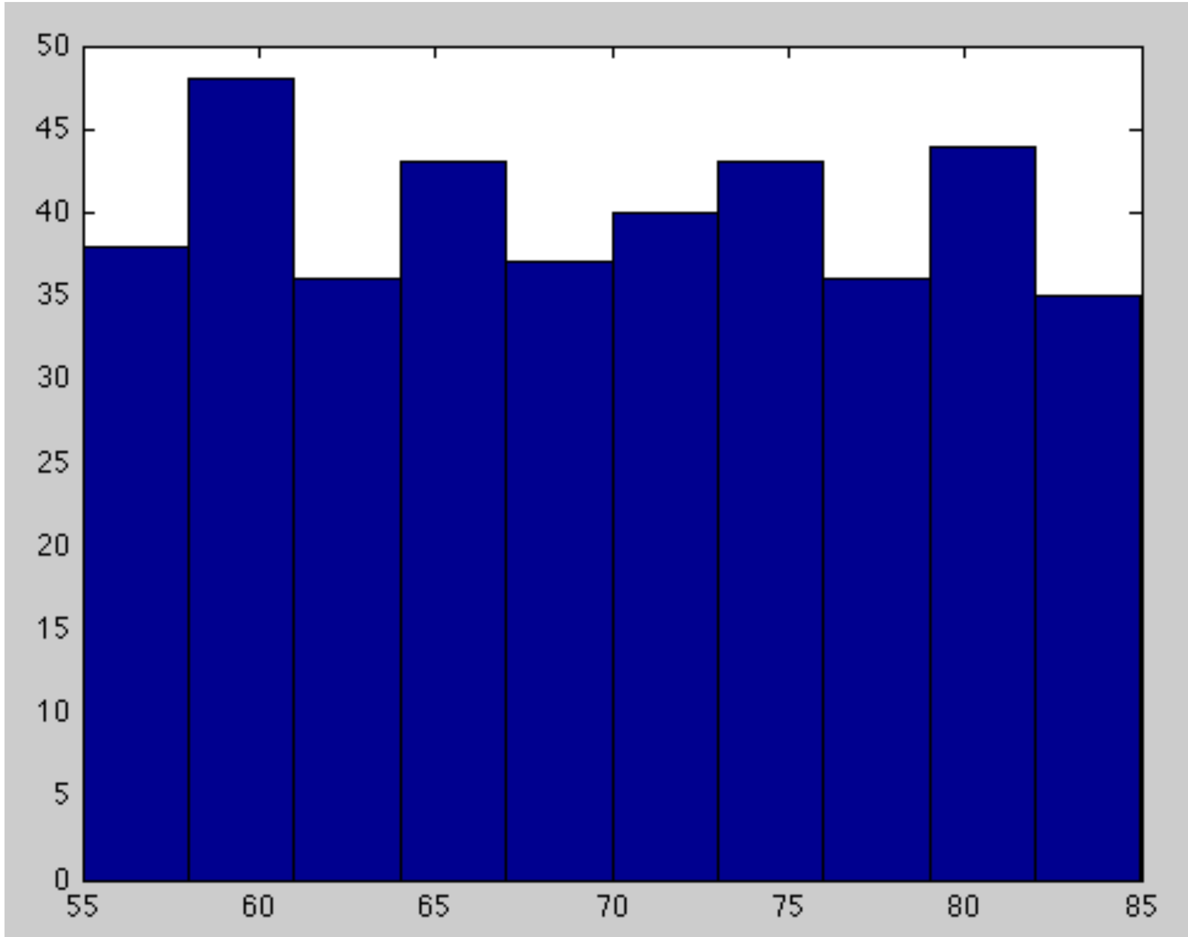
Sample 1: Mean = 70.72, SD $\approx$ 8.45



Sample 2: Mean = 70.67, SD $\approx$ 9.03

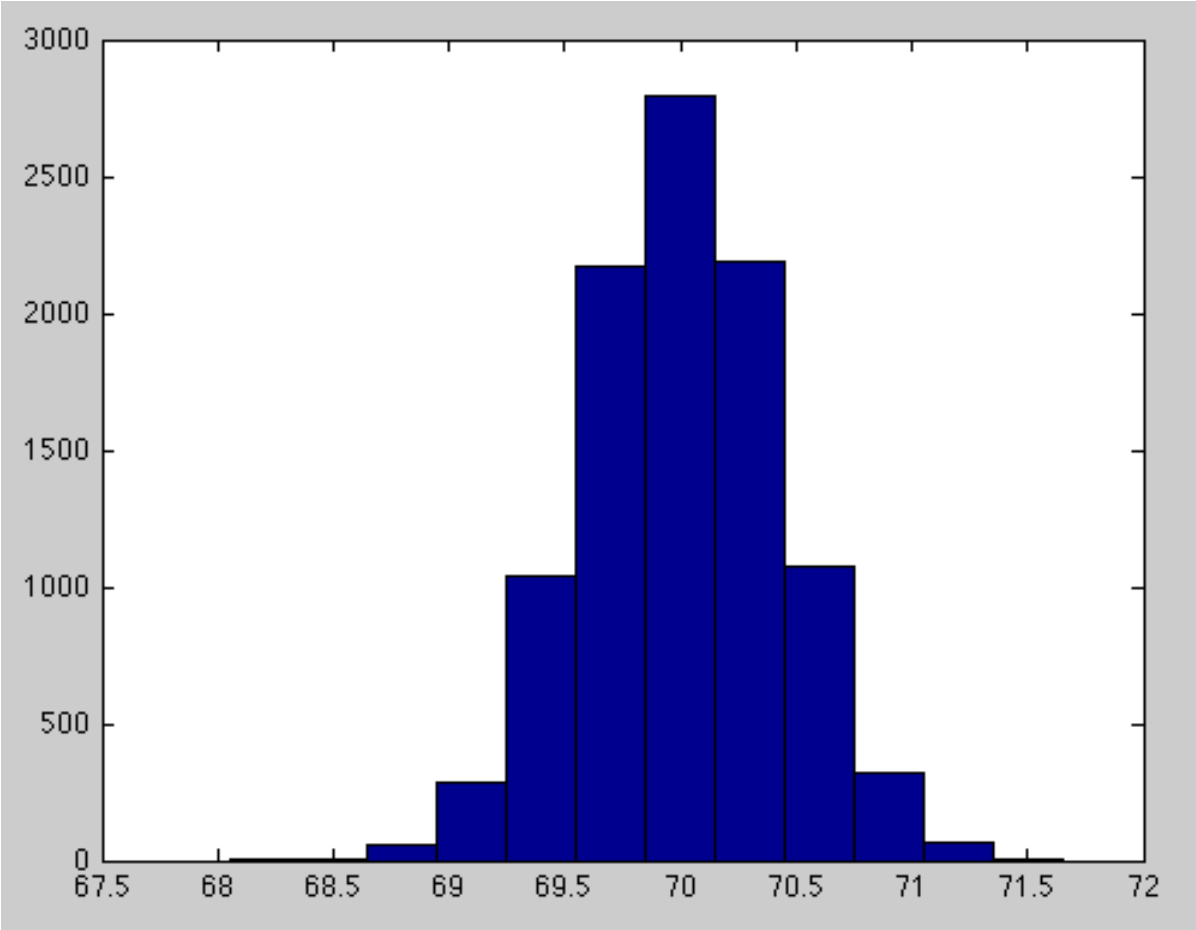


Sample 3: Mean = 69.32, SD ≈ 8.57



Sample 4: Mean = 70.11, SD $\approx$ 8.76

Histogram for distribution of averages of 10,000 samples of size 400 (from the uniform distribution above):



Mean = 70.0056, **SD**= 0.4186