Biostatistics

Chapter 3 Data Distribution and Sampling

Jing Li

jing.li@sjtu.edu.cn

http://cbb.sjtu.edu.cn/~jingli/courses/2017fall/bi372/

Dept of Bioinformatics & Biostatistics, SJTU

Review Questions (5 min)

- Describe briefly measures of data dispersion.

Review lecture2

Displaying the data

- Frequency table
- Bar chart
- Histogram, box-plot, violin plot

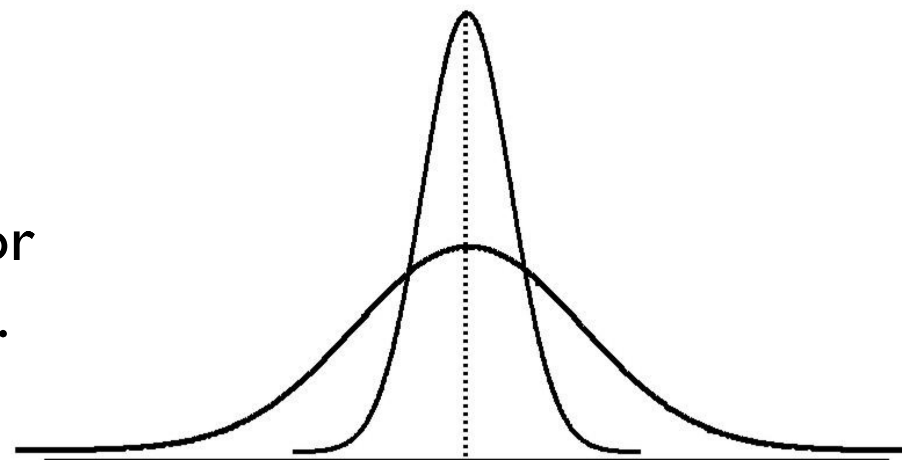
Descriptive statistics

- Mean, median, mode, range, IQR
- Quantiles
- Var, sd, cv

Measures of Dispersion

- Dispersion
  - Range
  - Interquartile Range
  - Variance
  - Standard Deviation
  - Coefficient of Variation

- Measures of variation give information on the spread or variability of the data values.



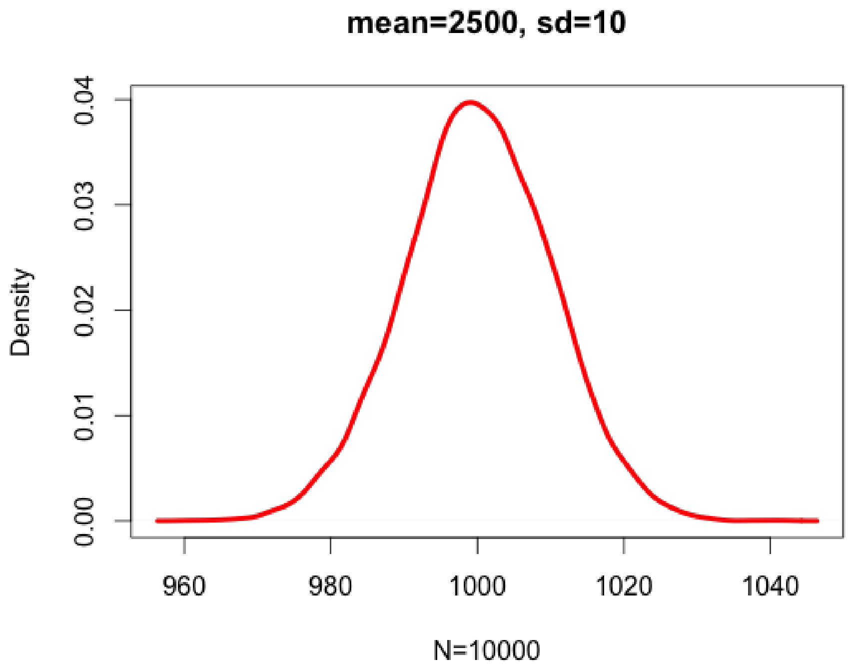
Same center, different variation

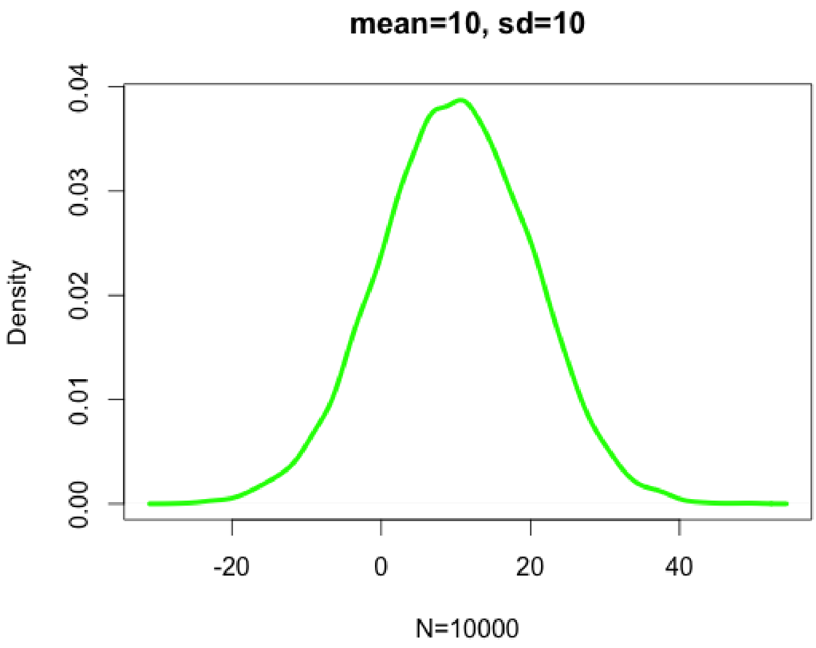
• Sample Standard Deviation (SD, 标准差)

- ✓ Most commonly used measure of variation
- ✓ Shows variation about the mean
- ✓ Has the same units as the original data

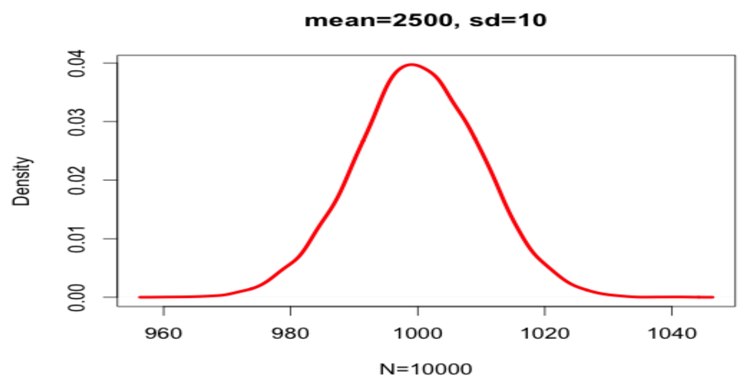
$$S = \sqrt{\frac{\sum_{i}^{n}(x_i - \overline{X})^2}{n-1}}$$

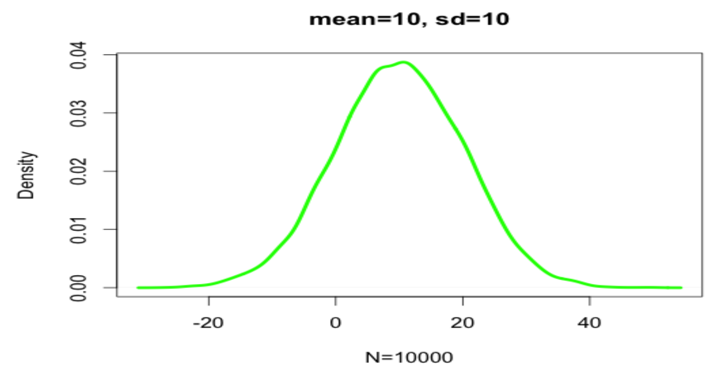
Mean-SD


mean=2500, sd=10
Density
0.00
0.01
0.02
0.03
0.04
960
980
1000
1020
1040
N=10000


mean=10, sd=10
Density
0.00
0.01
0.02
0.03
0.04
-20
0
20
40
N=10000

Mean-SD





Mean=2500kg, 10kg

Bienaymé-Chebyshev Rule

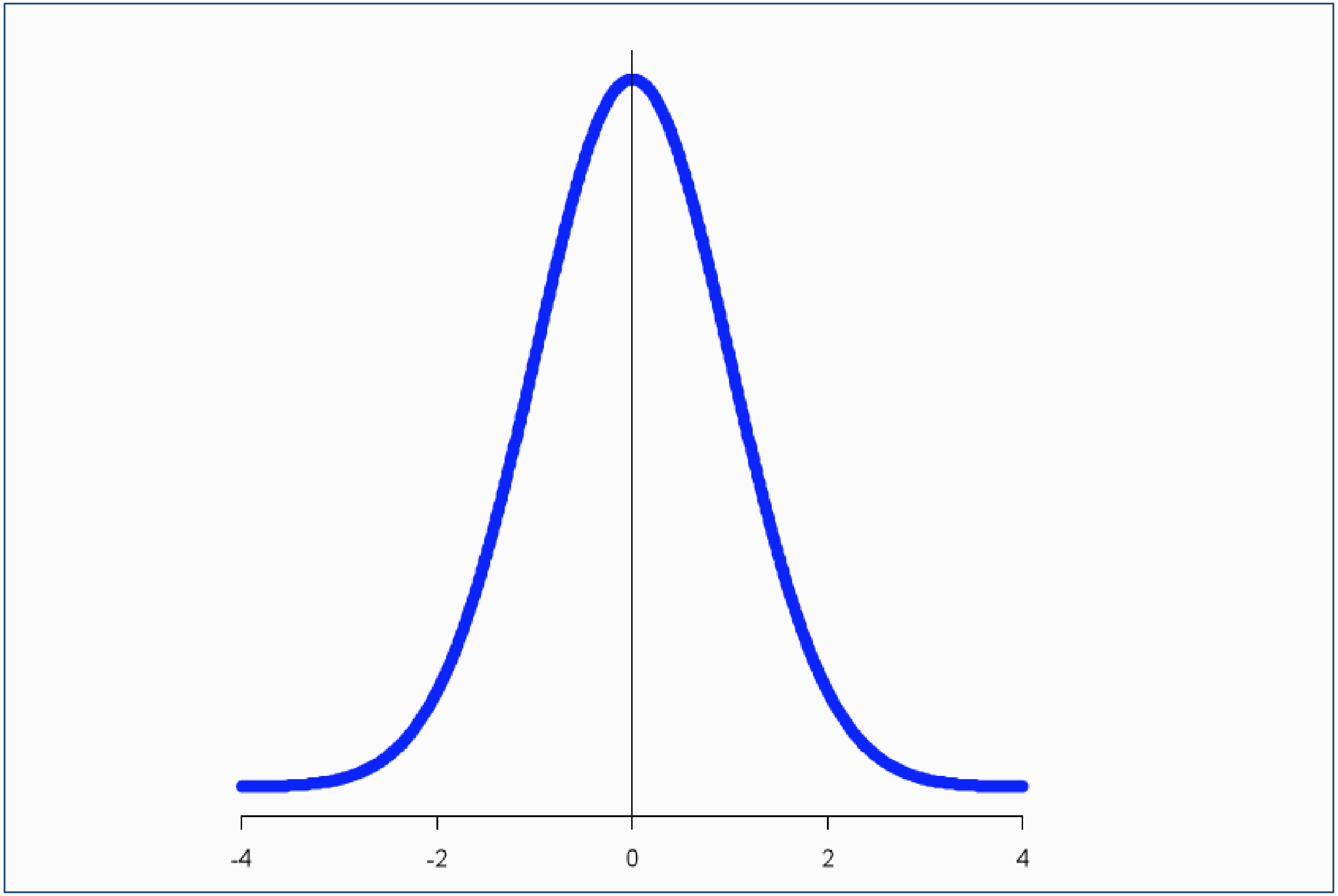
- Regardless of how the data are distributed, at least $(1 - 1/k^2)$ of the values will fall within k standard deviations of the mean (for $k > 1$)

Note use of μ (mu) to represent "mean".

Note use of σ (sigma) to represent "standard deviation."

At least	within
$(1 - 1/1^2) = 0\%$	$k=1$ $(\mu \pm 1\sigma)$
$(1 - 1/2^2) = 75\%$	$k=2$ $(\mu \pm 2\sigma)$
$(1 - 1/3^2) = 89\%$	$k=3$ $(\mu \pm 3\sigma)$

The Normal Distribution



The Normal Distribution

- The normal distribution is also called the "Gaussian distribution" in honor of its inventor Carl Friedrich Gauss

(1777–1855 Germany)

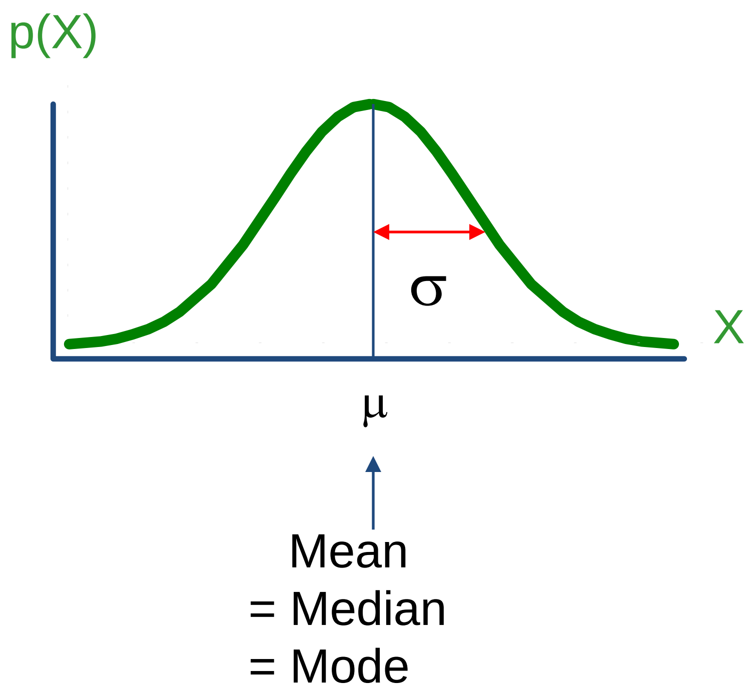
The Normal Distribution

- **Bell shaped**
- **Symmetrical**
- **Mean, median and mode are equal**

μ=mean

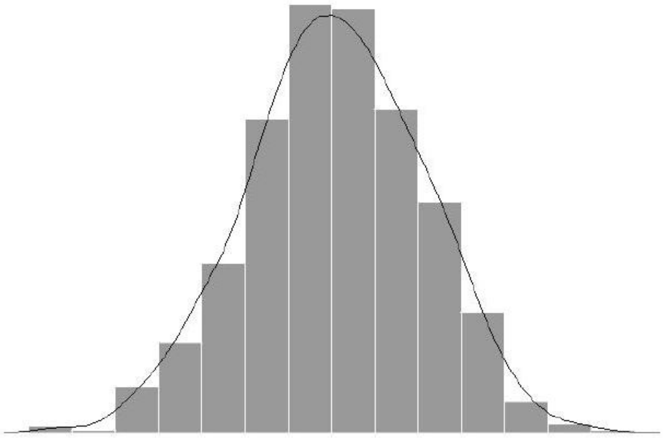
σ = standard deviation

The random variable has an infinite theoretical range:
$+\infty$ to $-\infty$



Examples:

- height
- weight
- age
- bone density
- IQ (mean=100; SD=15)
- SAT(Scholastic Assessment Test) scores
- blood pressure



The Normal PDF

It's a probability function, so no matter what the values of μ and σ, must integrate to 1!

$$\int_{-\infty}^{+\infty} \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx = 1$$

<u>Note constants:</u>

π=3.14159

e=2.71828

The Normal Distribution

- Normal distribution is defined by its mean and standard dev.

$$E(X)=\mu = \int_{-\infty}^{+\infty} x \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx$$

$$Var(X)=\sigma^2 = \int_{-\infty}^{+\infty} x^2 \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx) - \mu^2$$

Standard Deviation(X)=σ

The Normal PDF

It's a probability function, so no matter what the values of μ and σ, must integrate to 1!

$$\int_{-\infty}^{+\infty} \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = 1$$

This is a bell shaped curve with different centers and spreads depending on μ and σ

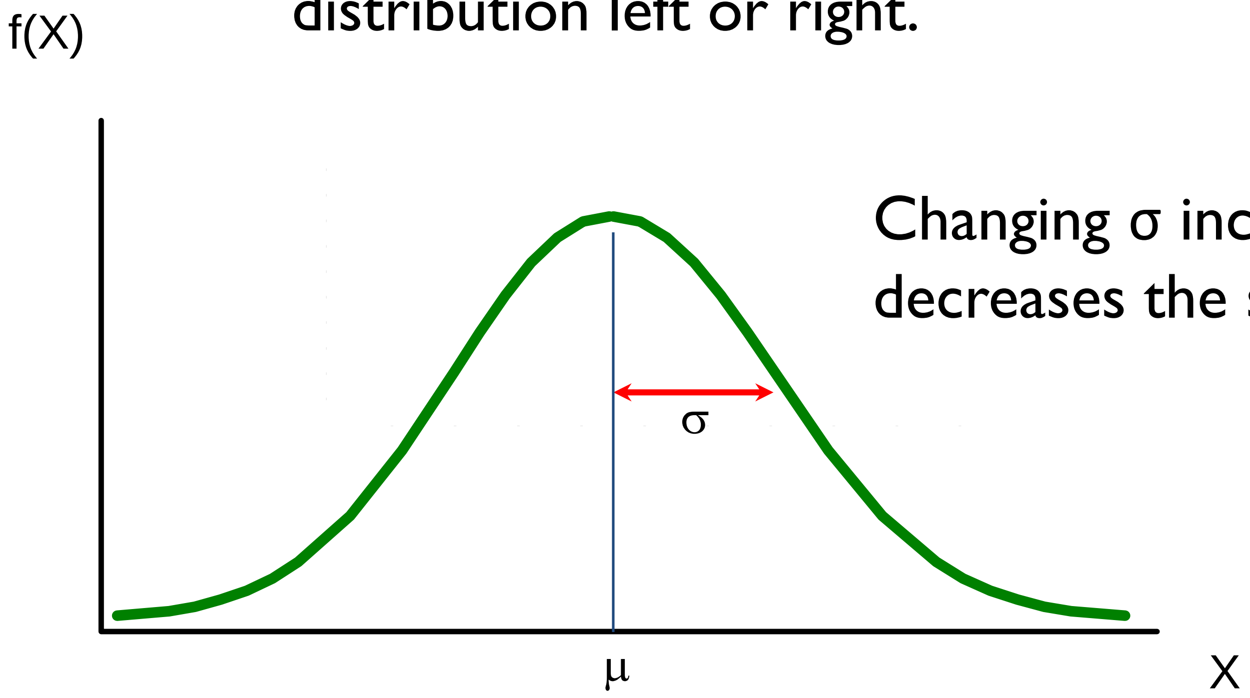
Note constants:

π=3.14159

e=2.71828

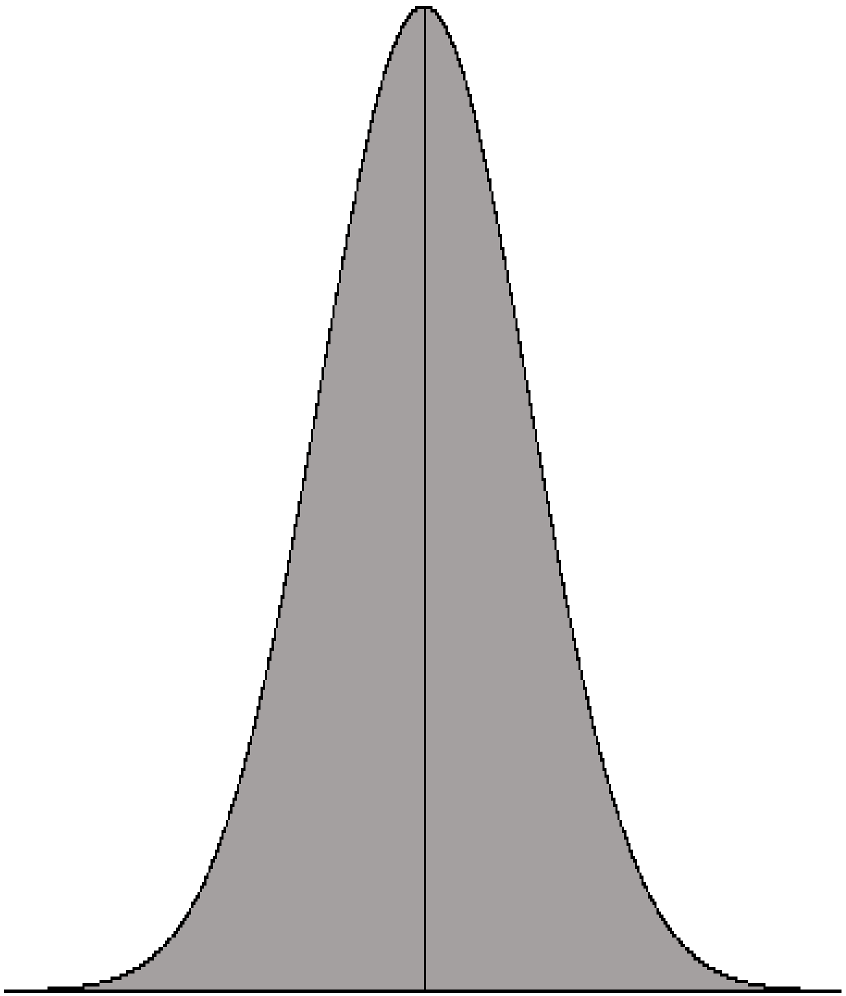
The Normal Distribution

Changing μ shifts the distribution left or right.

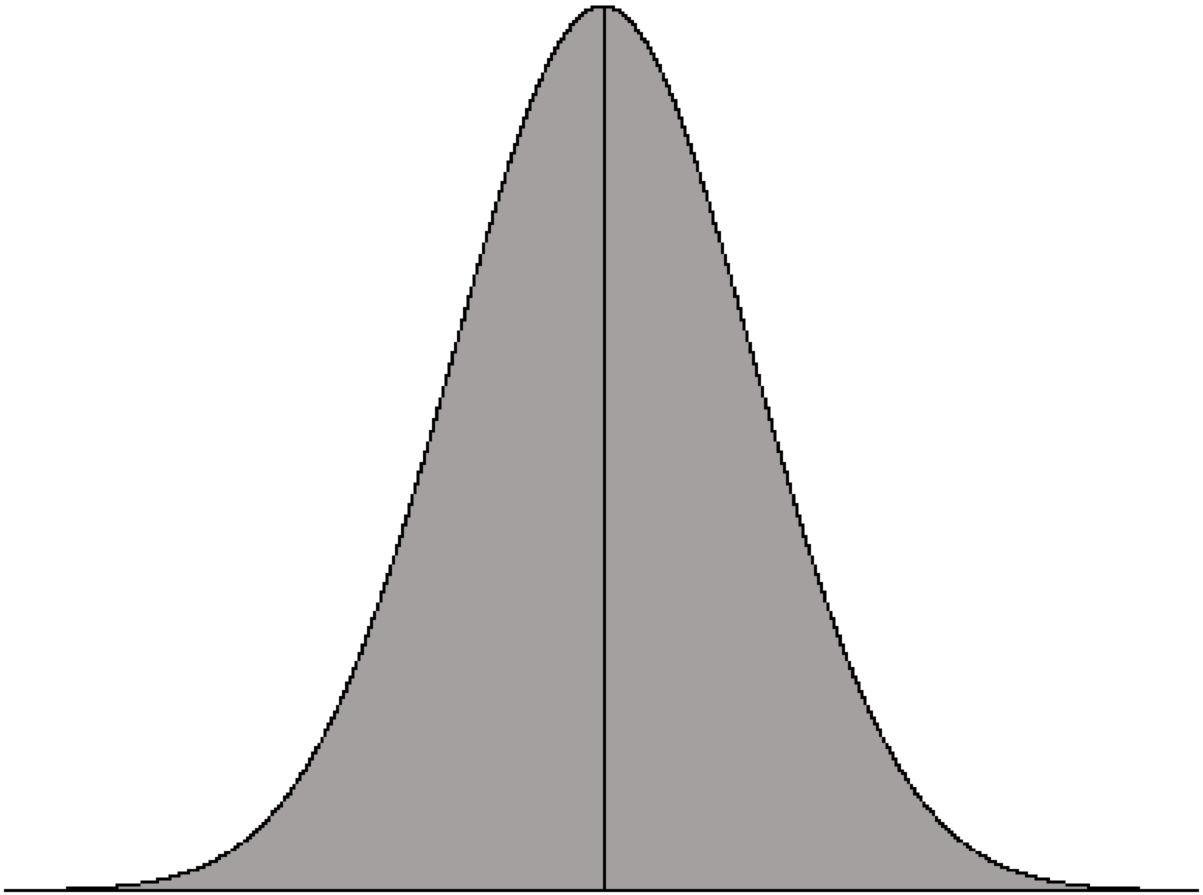


Changing σ increases or decreases the spread.

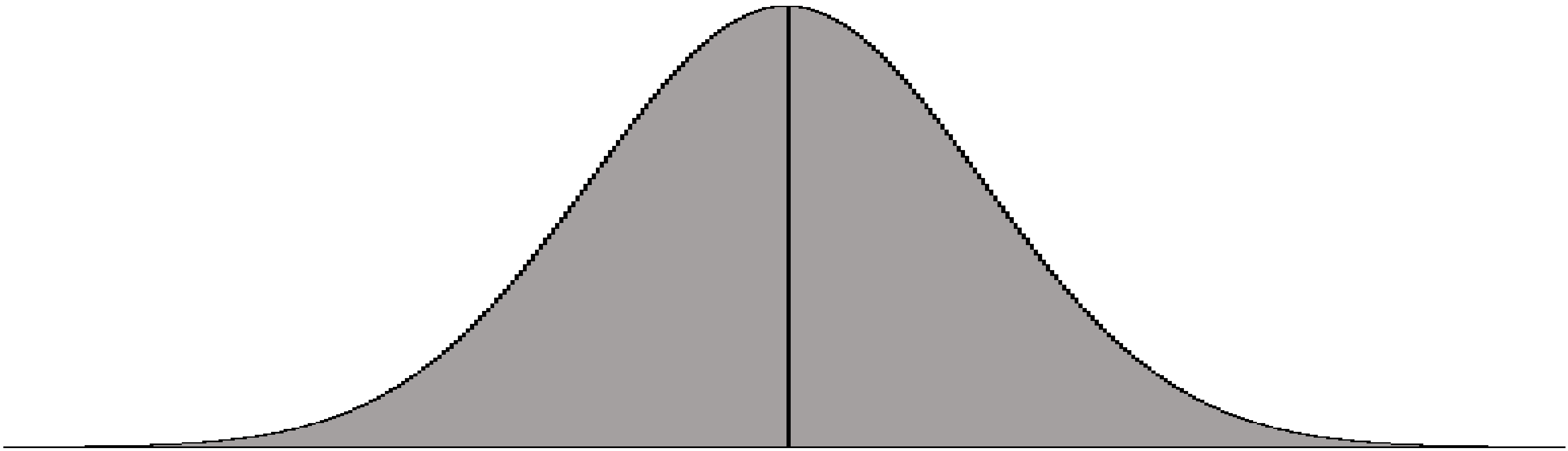
Small standard deviation



Larger standard deviation



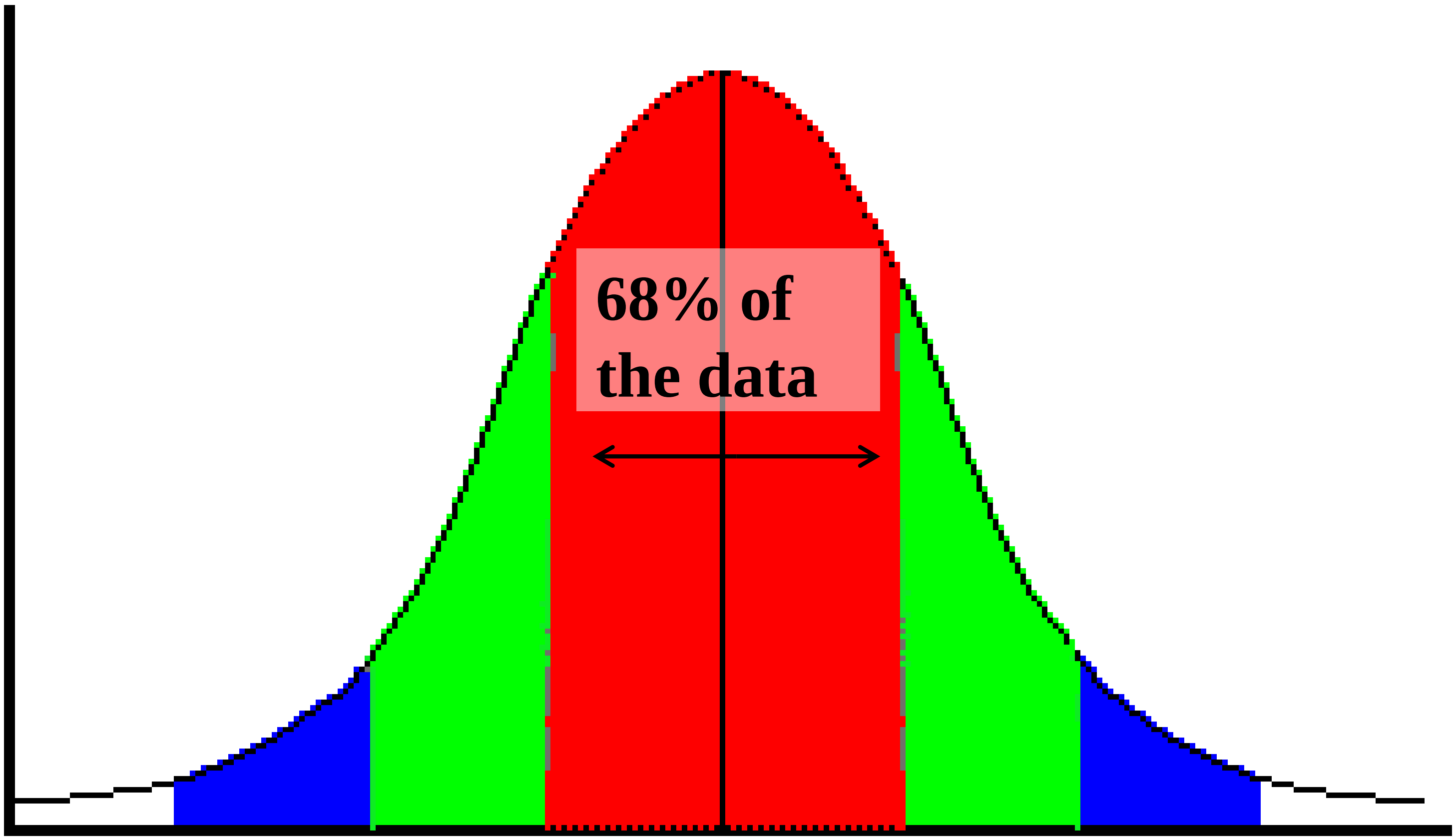
Even larger standard deviation



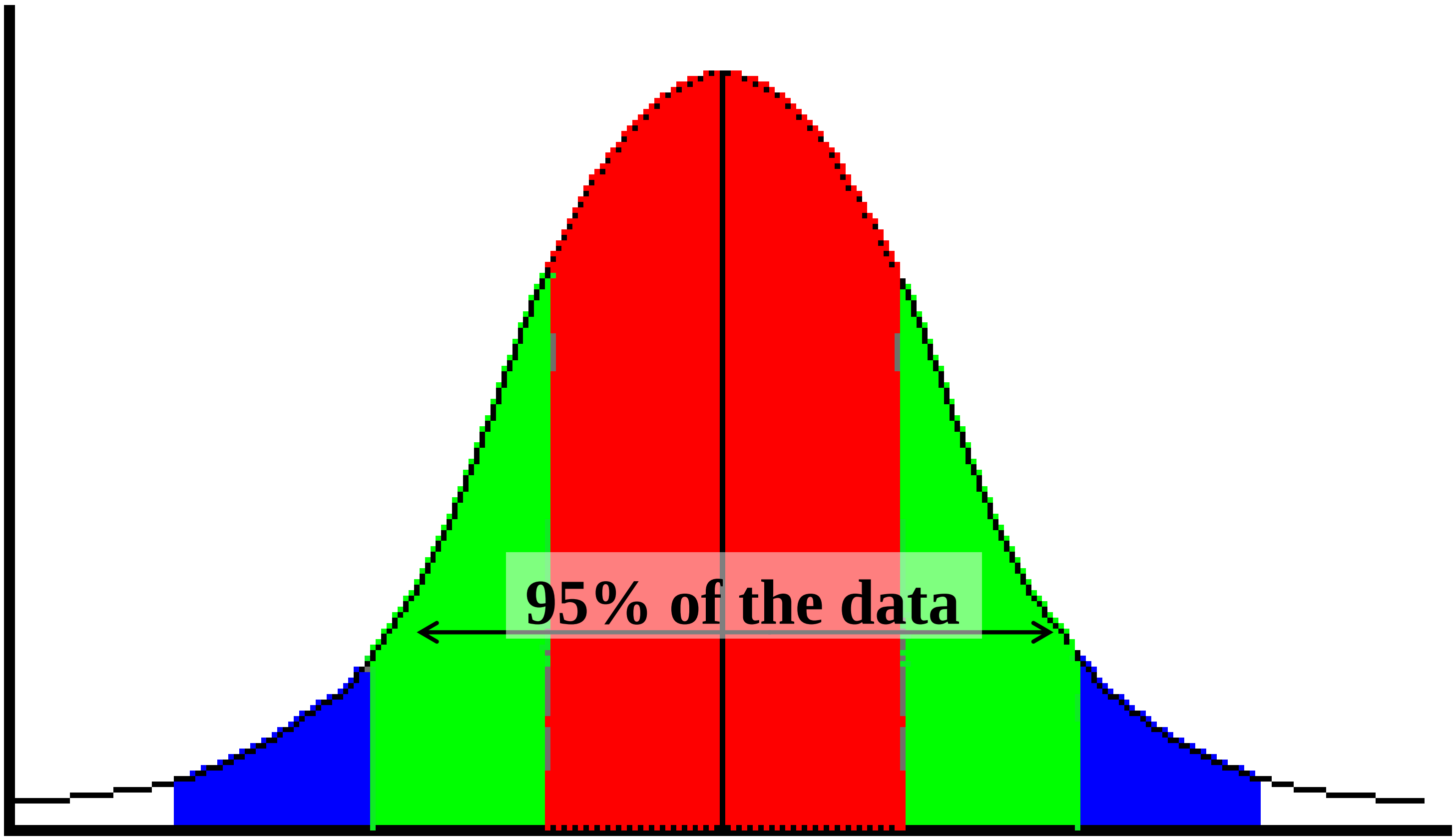
**The beauty of the normal curve

No matter what μ and σ are, the area between μ-σ and μ+σ is about 68%; the area between μ-2σ and μ+2σ is about 95%; and the area between μ-3σ and μ+3σ is about 99.7%. Almost all values fall within 3 standard deviations.

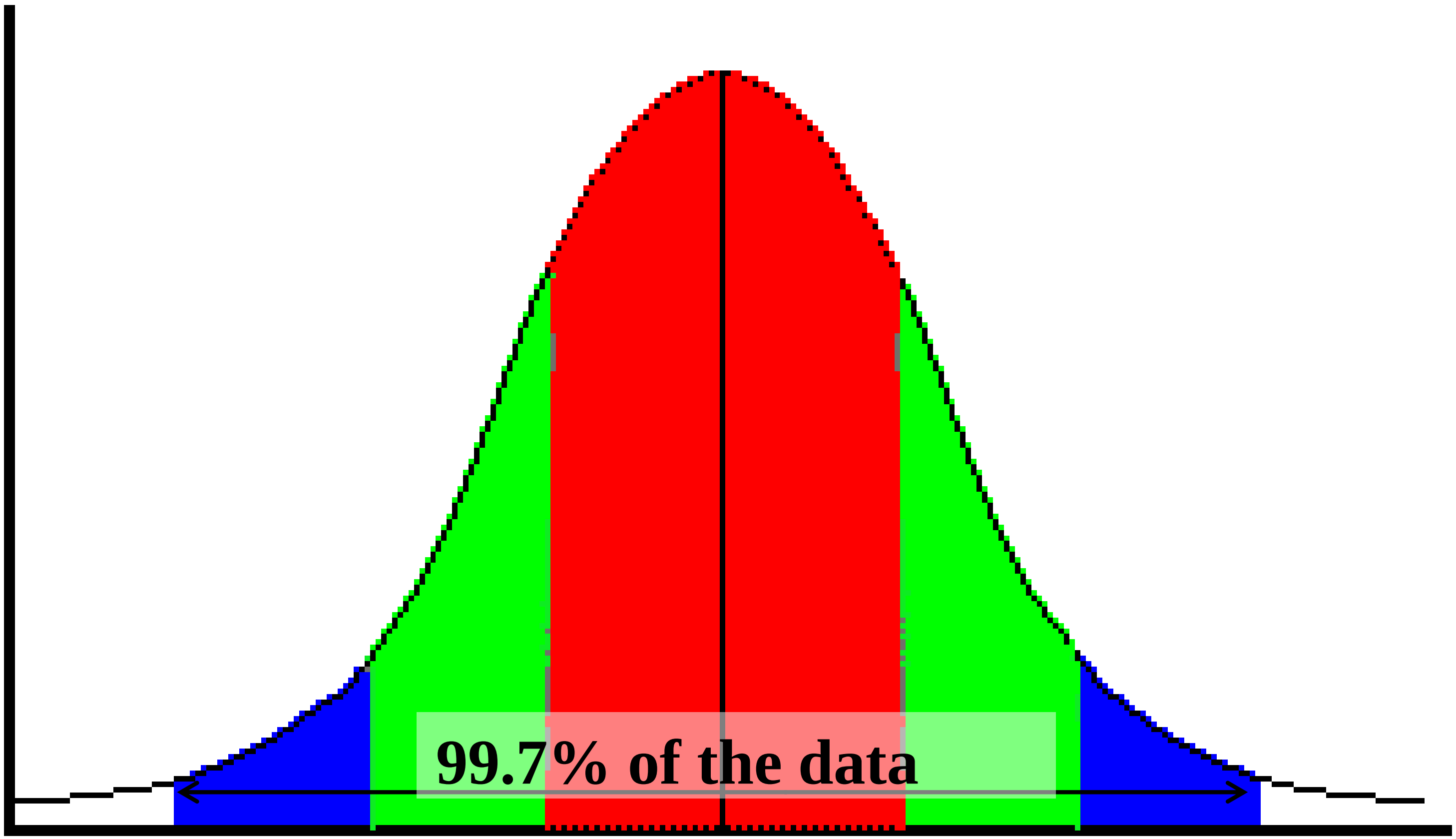
68-95-99.7 Rule



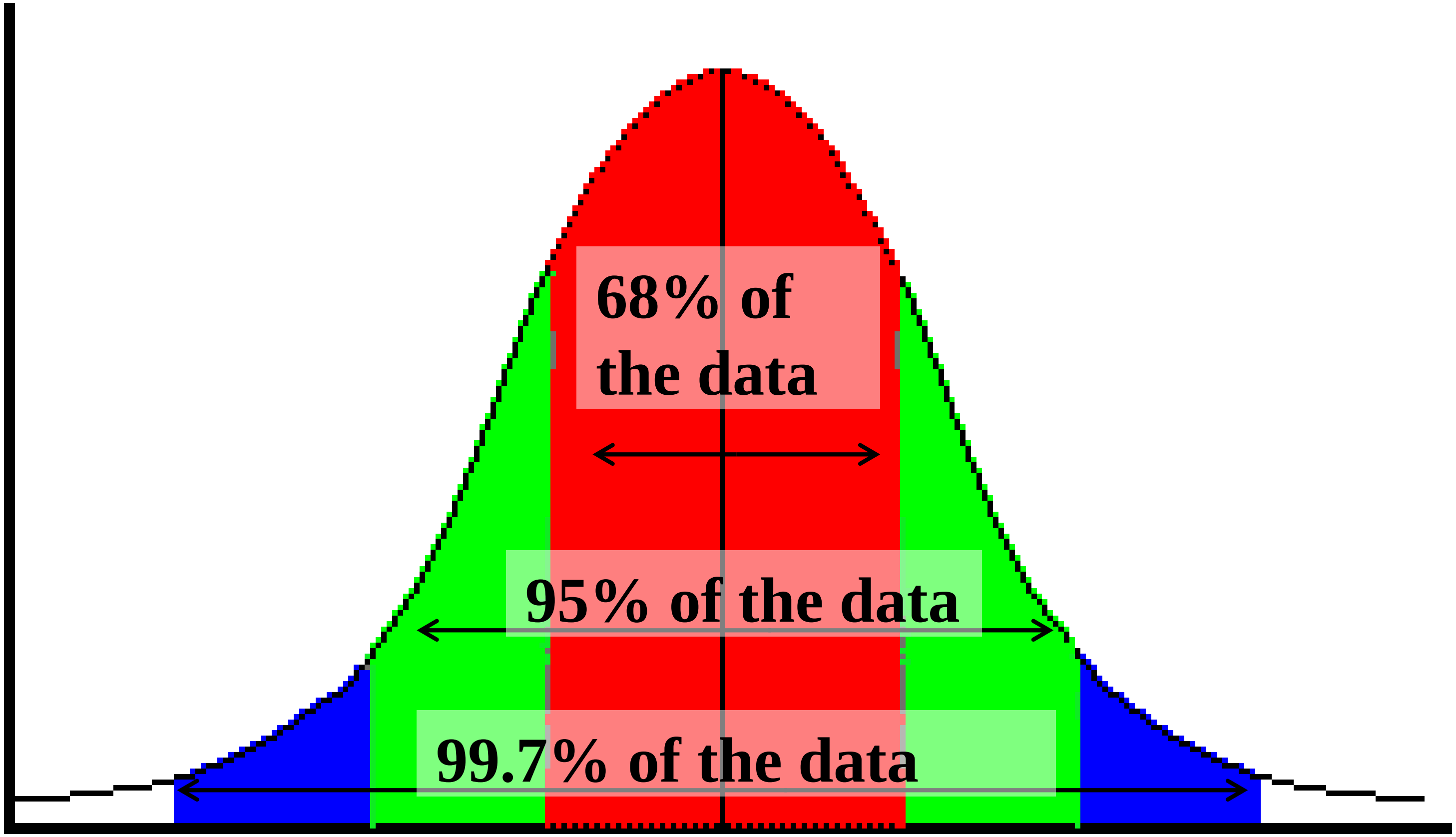
68-95-99.7 Rule



68-95-99.7 Rule



68-95-99.7 Rule

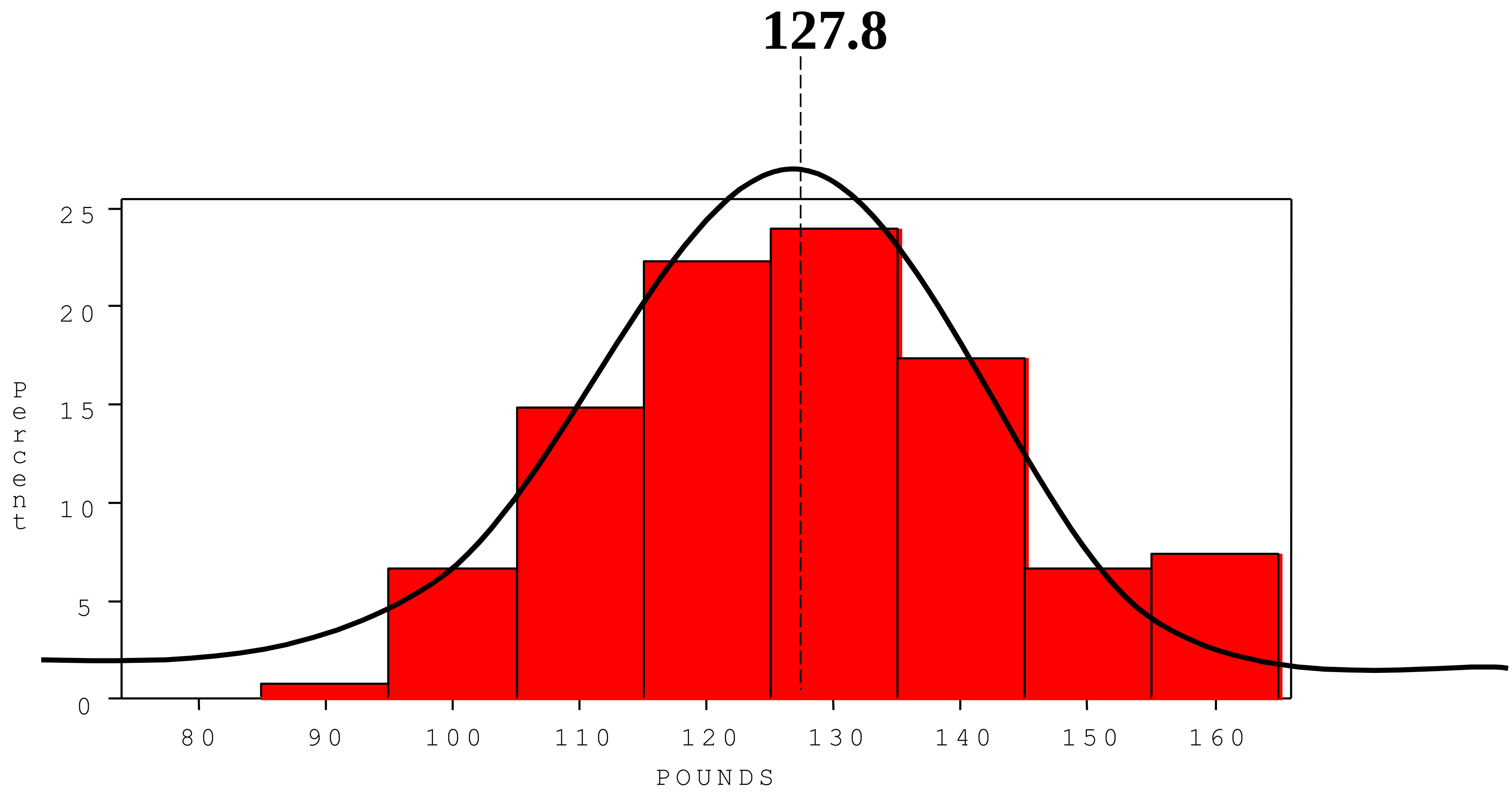


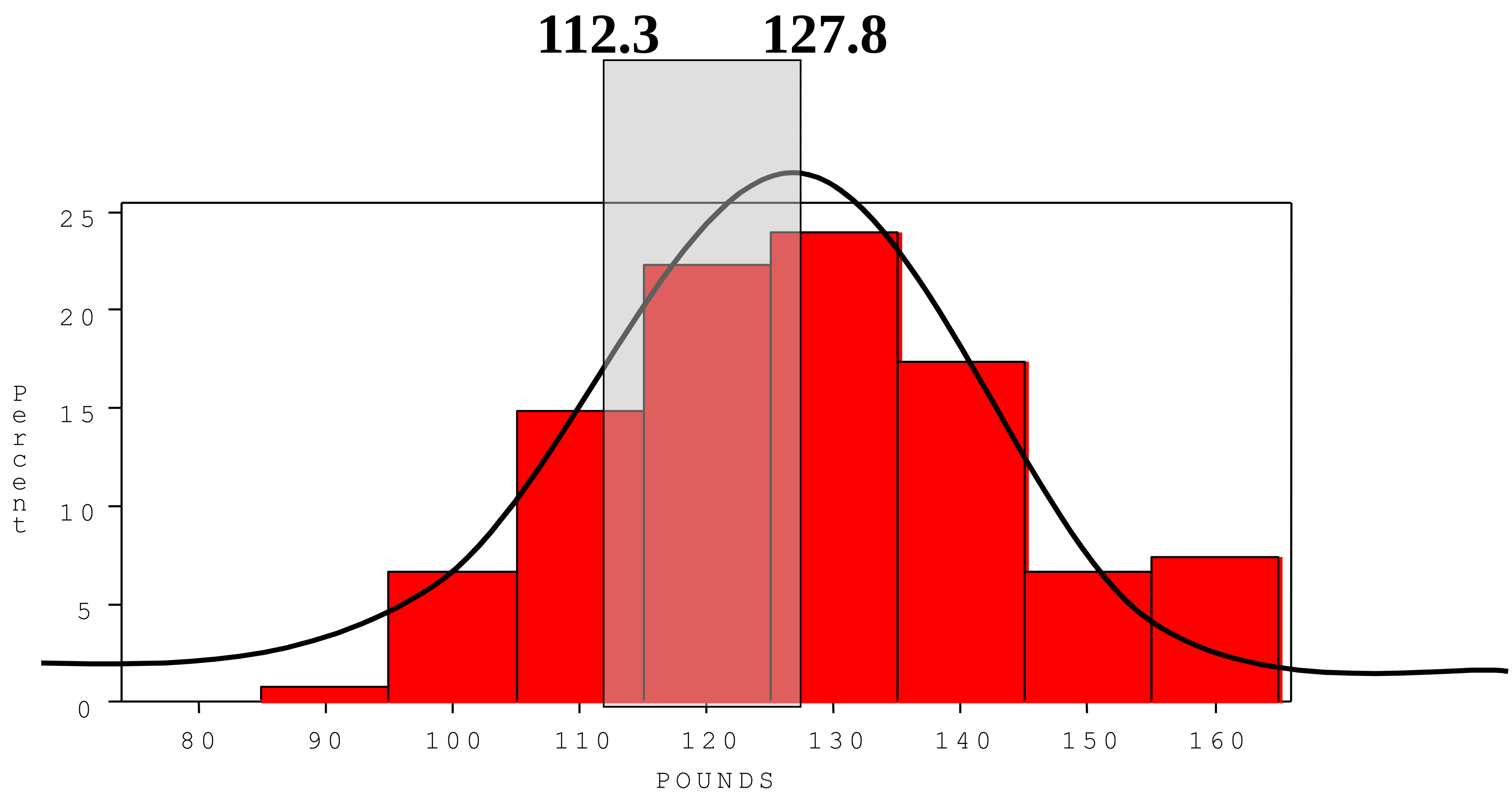
How good is rule for real data?

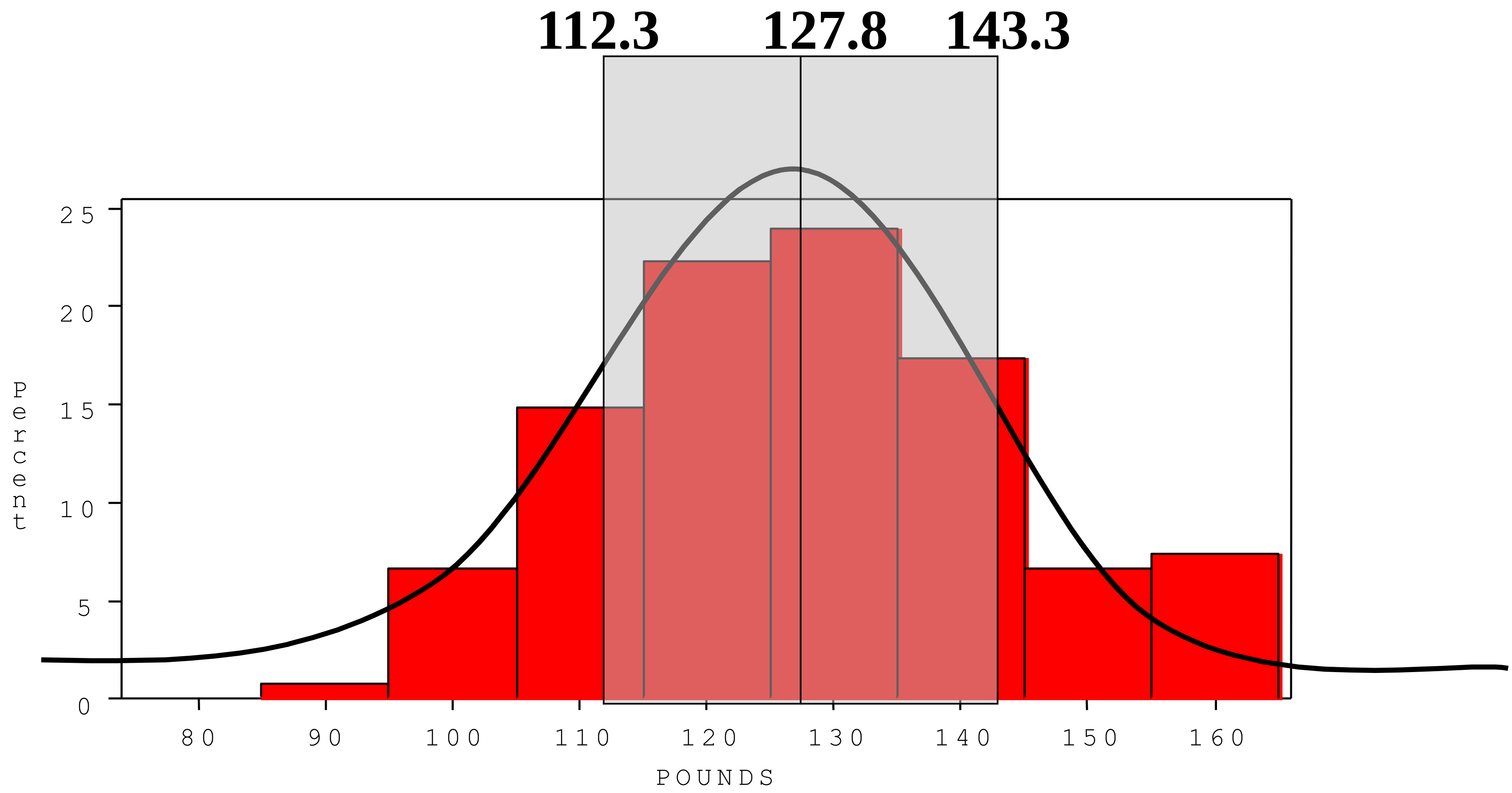
Check some example data:

The mean of the weight of the women = 127.8

The standard deviation (SD) = 15.5

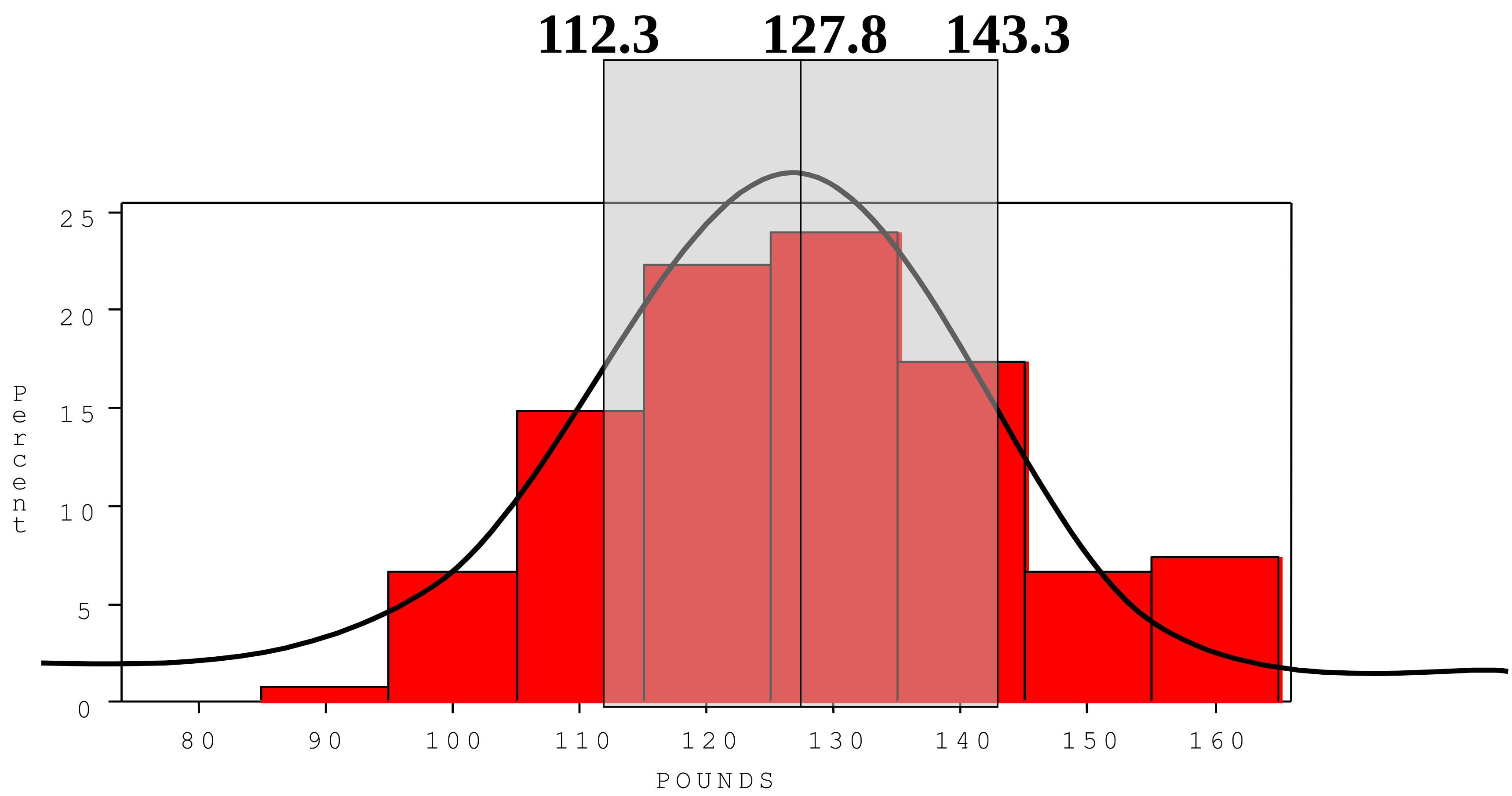

127.8
Percent
0
5
10
15
20
25
80
90
100
110
120
130
140
150
160
POUNDS

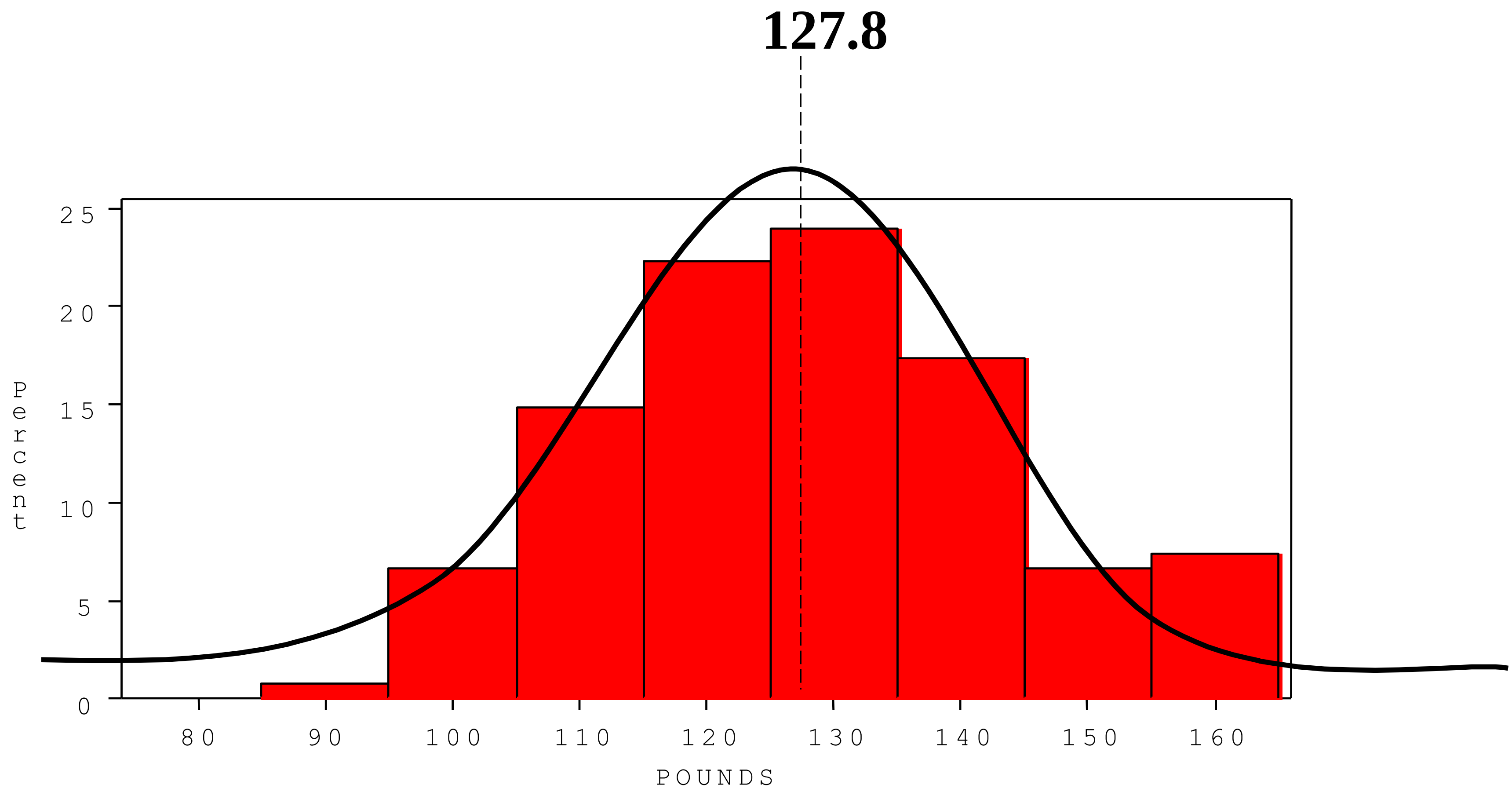

112.3
127.8
25
20
15
10
5
0
Percent
80
90
100
110
120
130
140
150
160
POUNDS

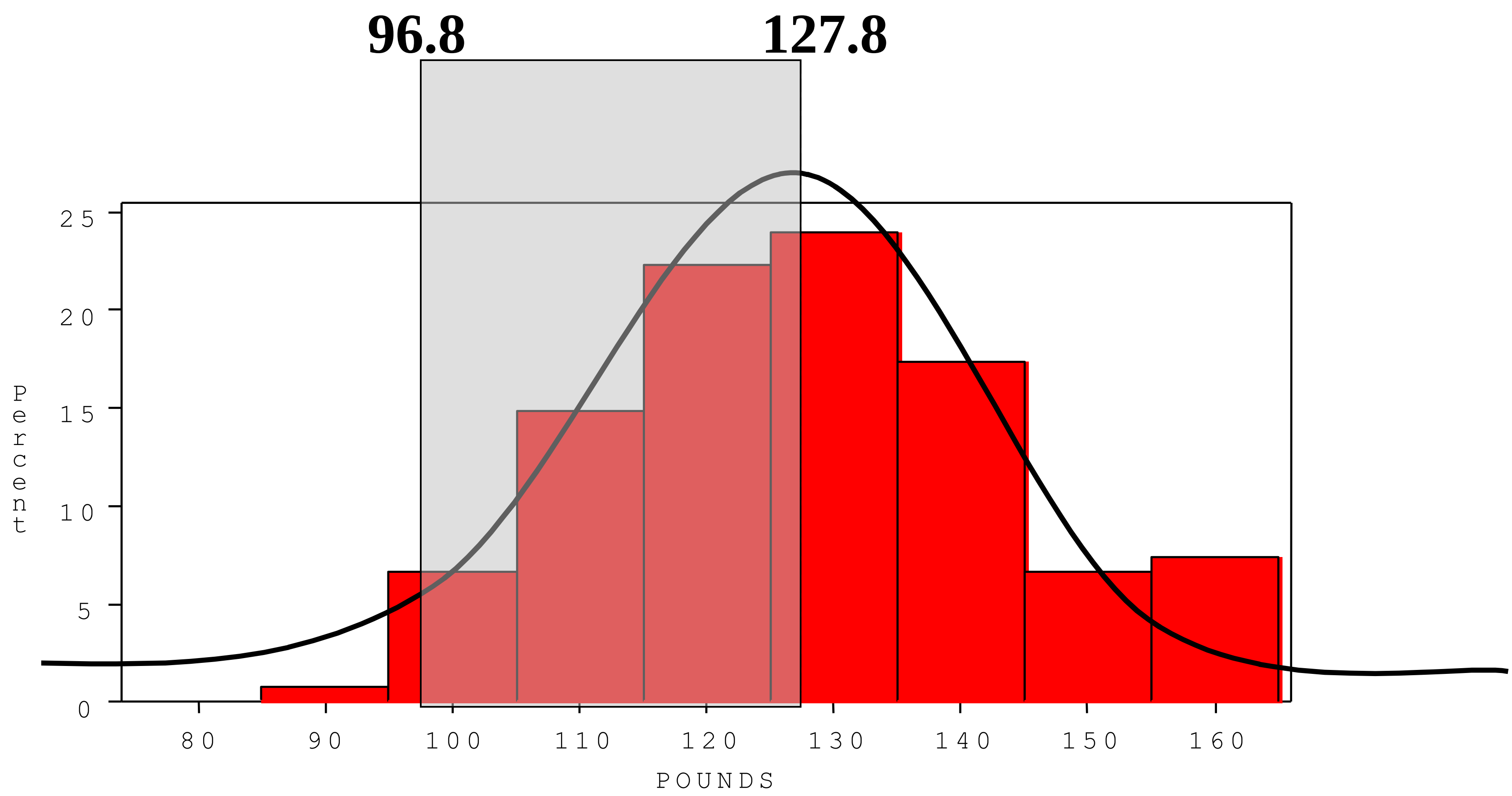

112.3
127.8
143.3
Percent
0
5
10
15
20
25
80
90
100
110
120
130
140
150
160
POUNDS

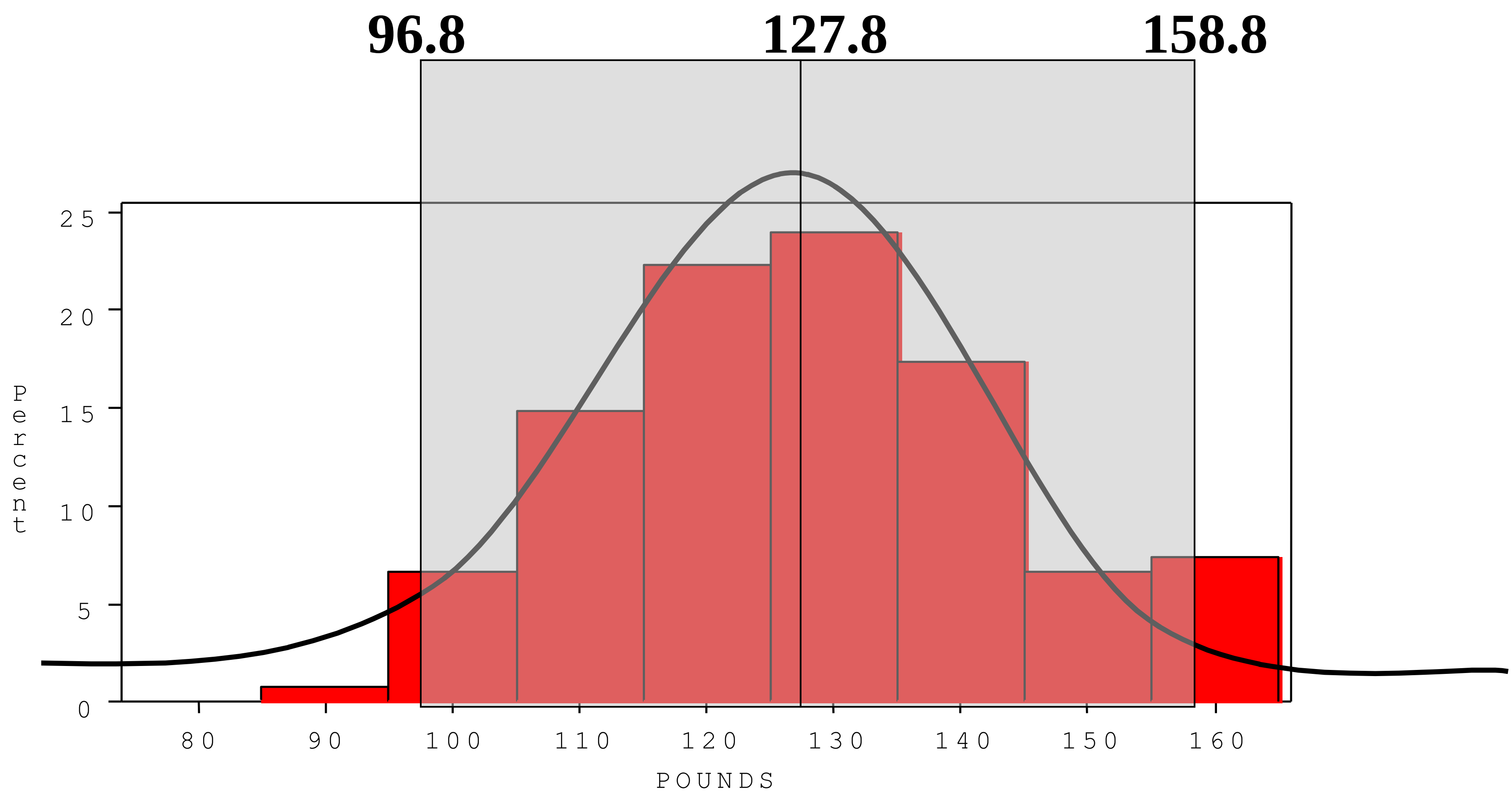
68% of 120 = .68x120 = ~ 82 runners

In fact, 79 runners fall within 1-SD (15.5 lbs) of the mean.



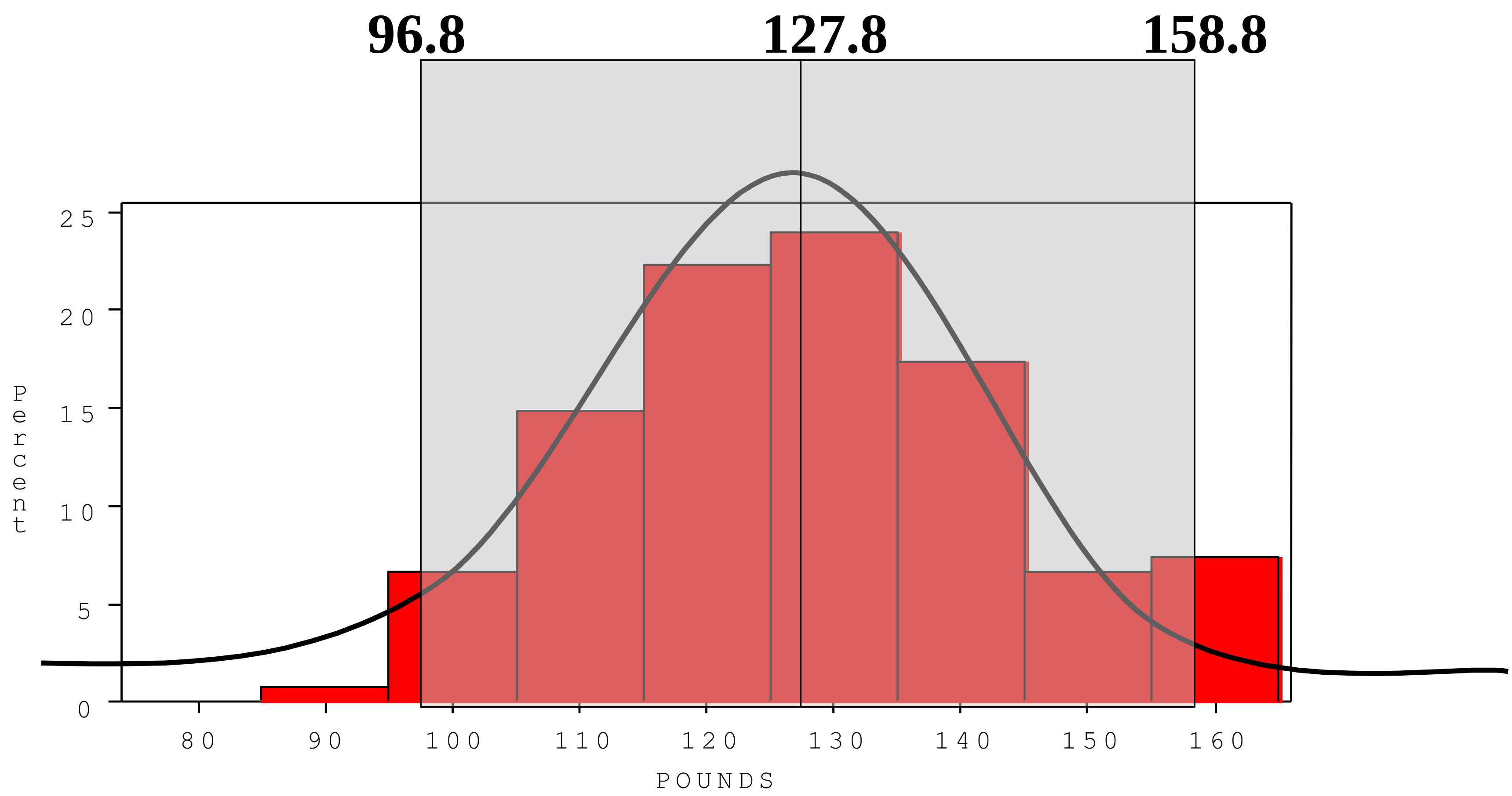

127.8
Percent
0
5
10
15
20
25
80
90
100
110
120
130
140
150
160
POUNDS

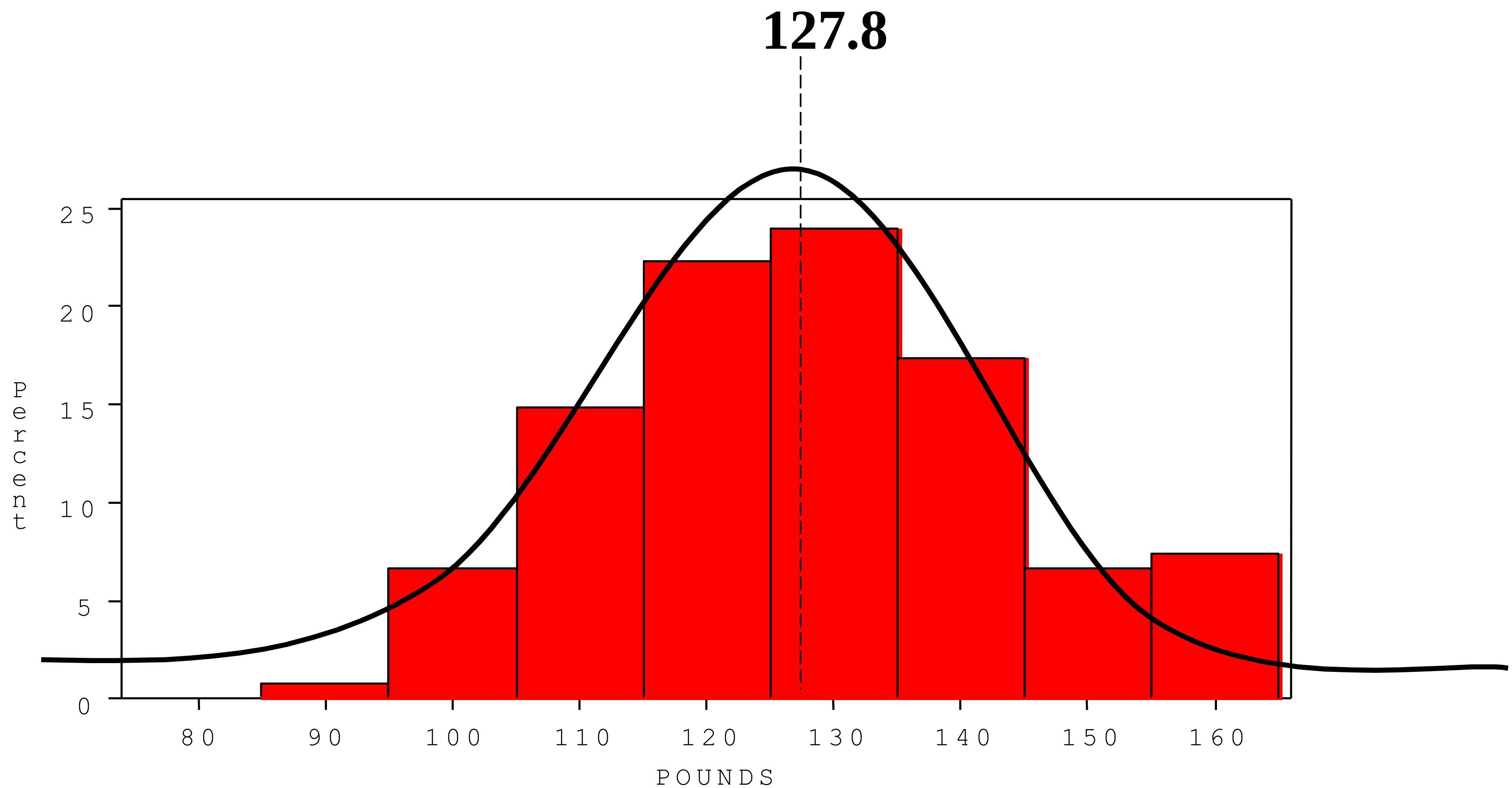

96.8
127.8
Percent
0
5
10
15
20
25
80
90
100
110
120
130
140
150
160
POUNDS

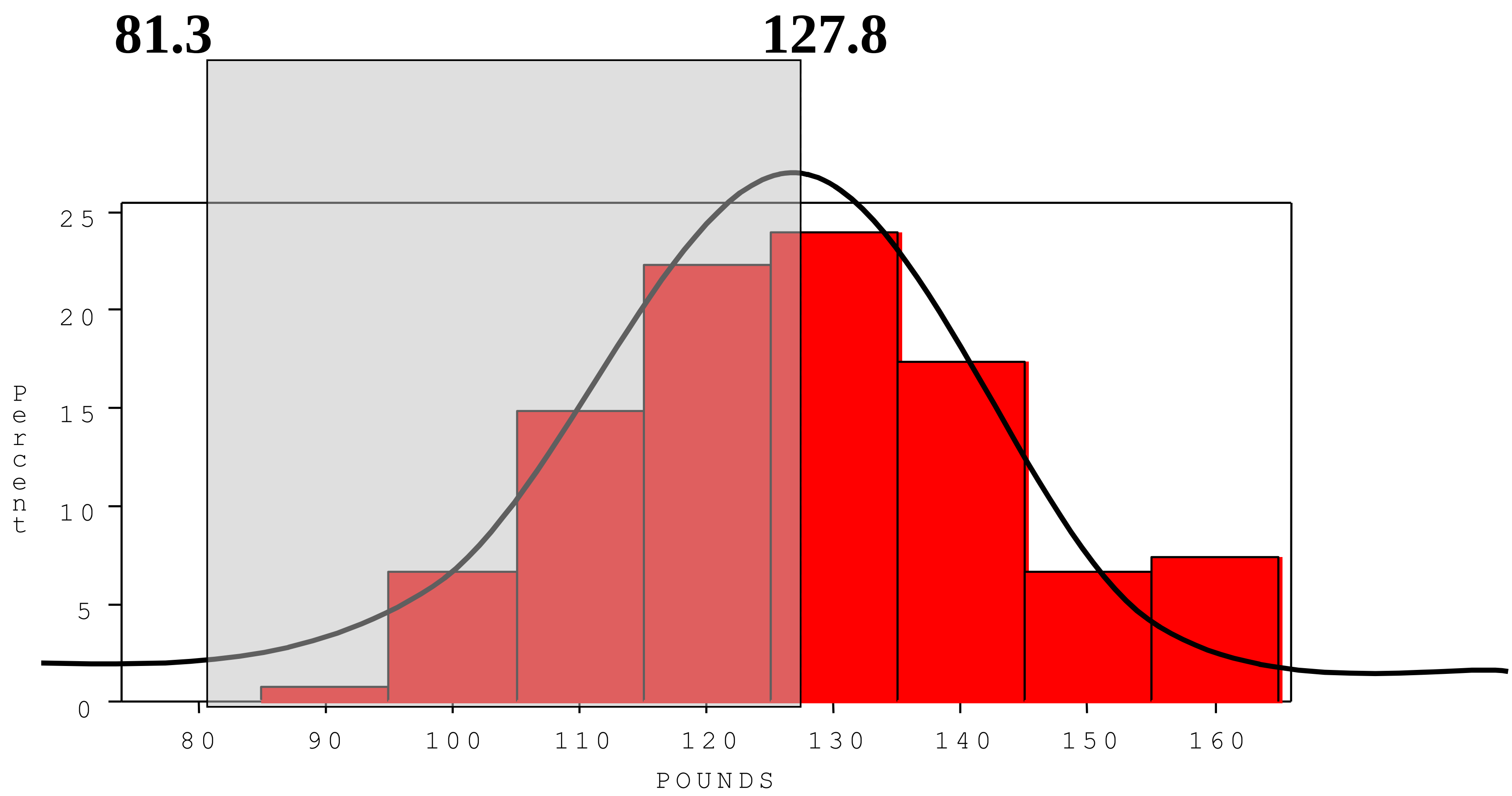

96.8
127.8
158.8
Percent
25
20
15
10
5
0
80
90
100
110
120
130
140
150
160
POUNDS

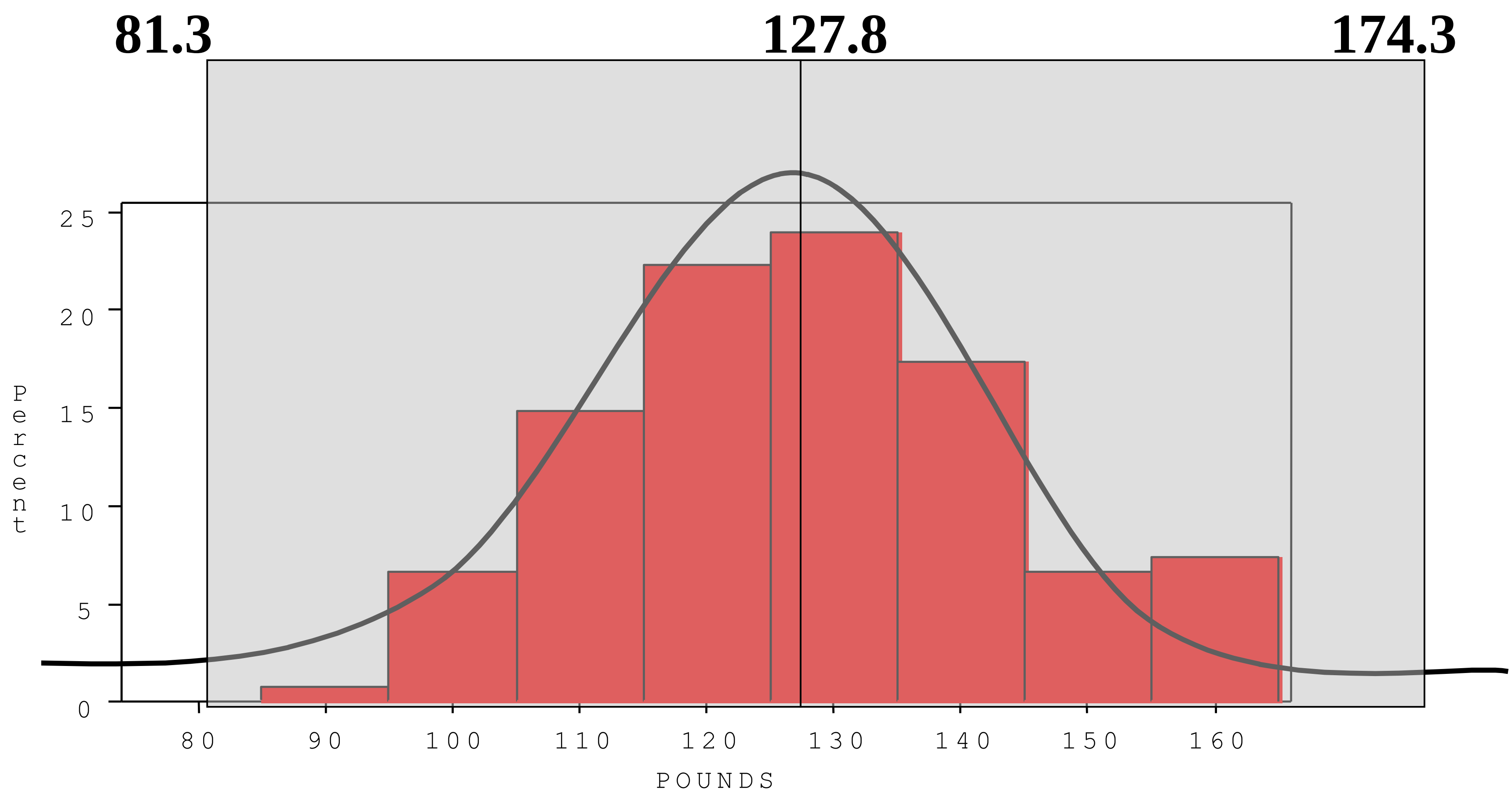
95% of 120 = .95 x 120 = ~ 114 runners

In fact, 115 runners fall within 2-SD's of the mean.



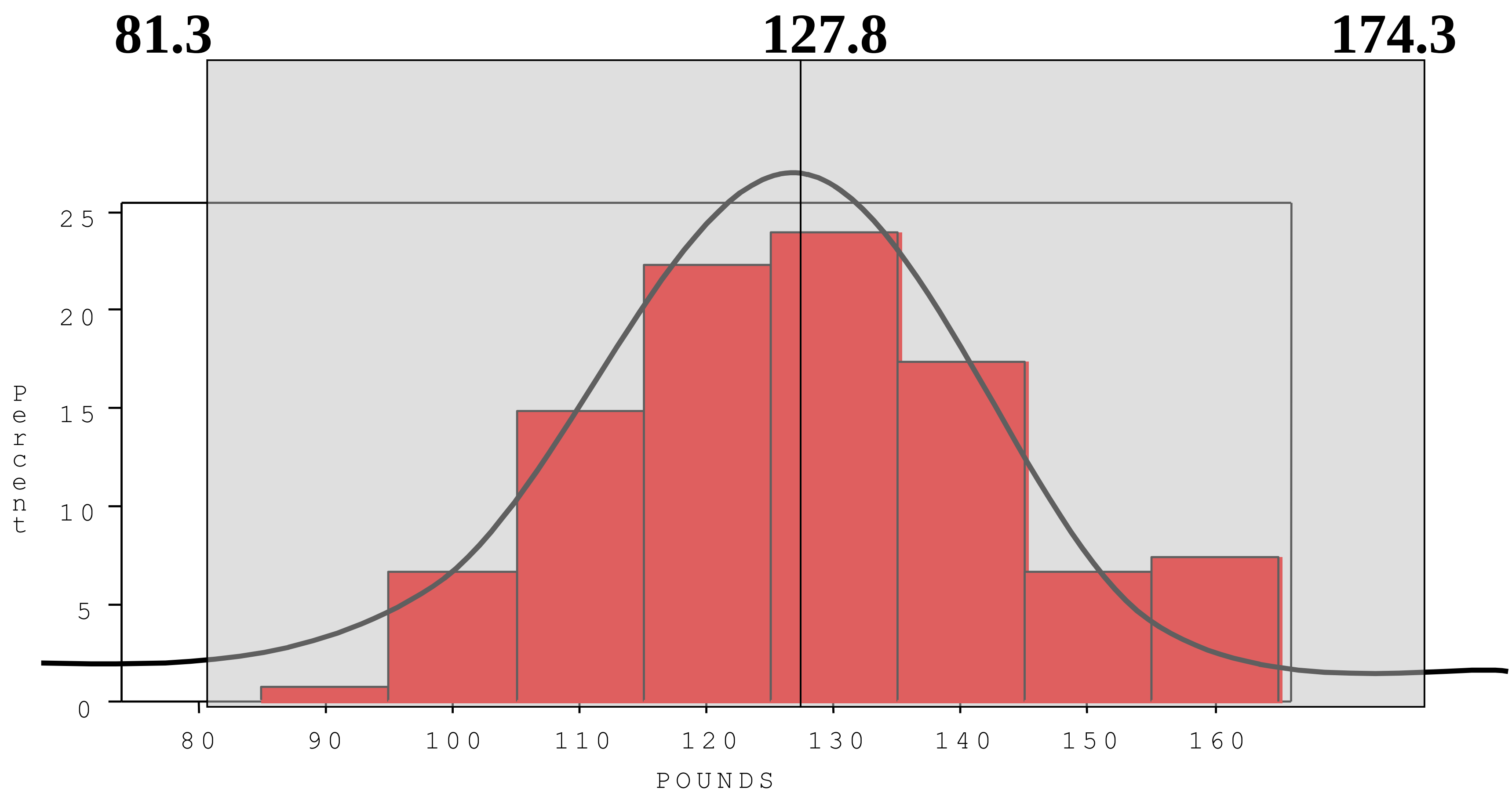

127.8
25
20
15
10
5
0
Percent
80
90
100
110
120
130
140
150
160
POUNDS


81.3
127.8
25
20
15
10
5
0
Percent
80
90
100
110
120
130
140
150
160
POUNDS


81.3
127.8
174.3
Percent
25
20
15
10
5
0
80
90
100
110
120
130
140
150
160
POUNDS

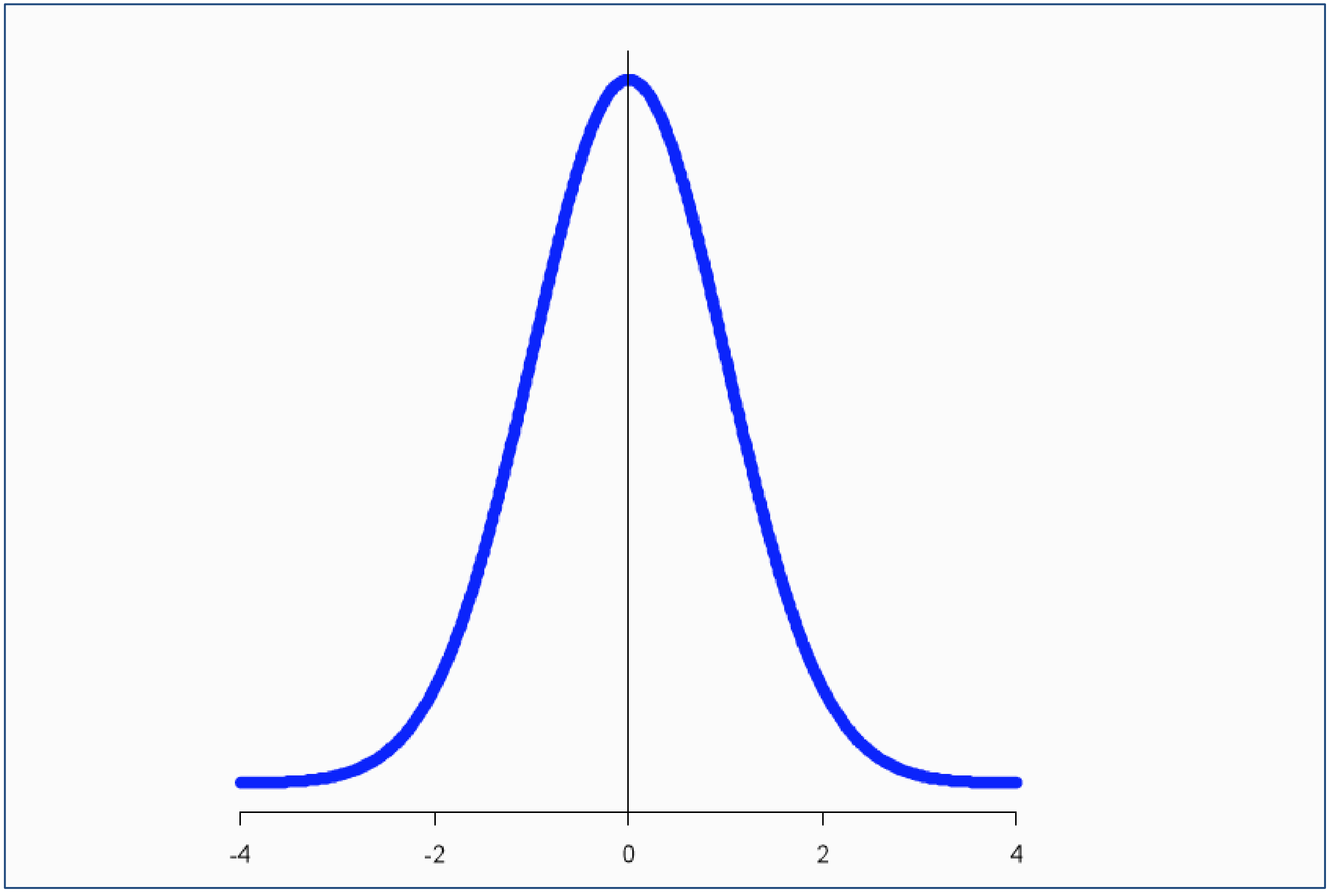
99.7% of 120 = .997 x 120 = 119.6 runners

In fact, all 120 runners fall within 3-SD's of the mean.



The Standard Normal Distribution

- The standard normal distribution has a mean of 0, and standard deviation of 1



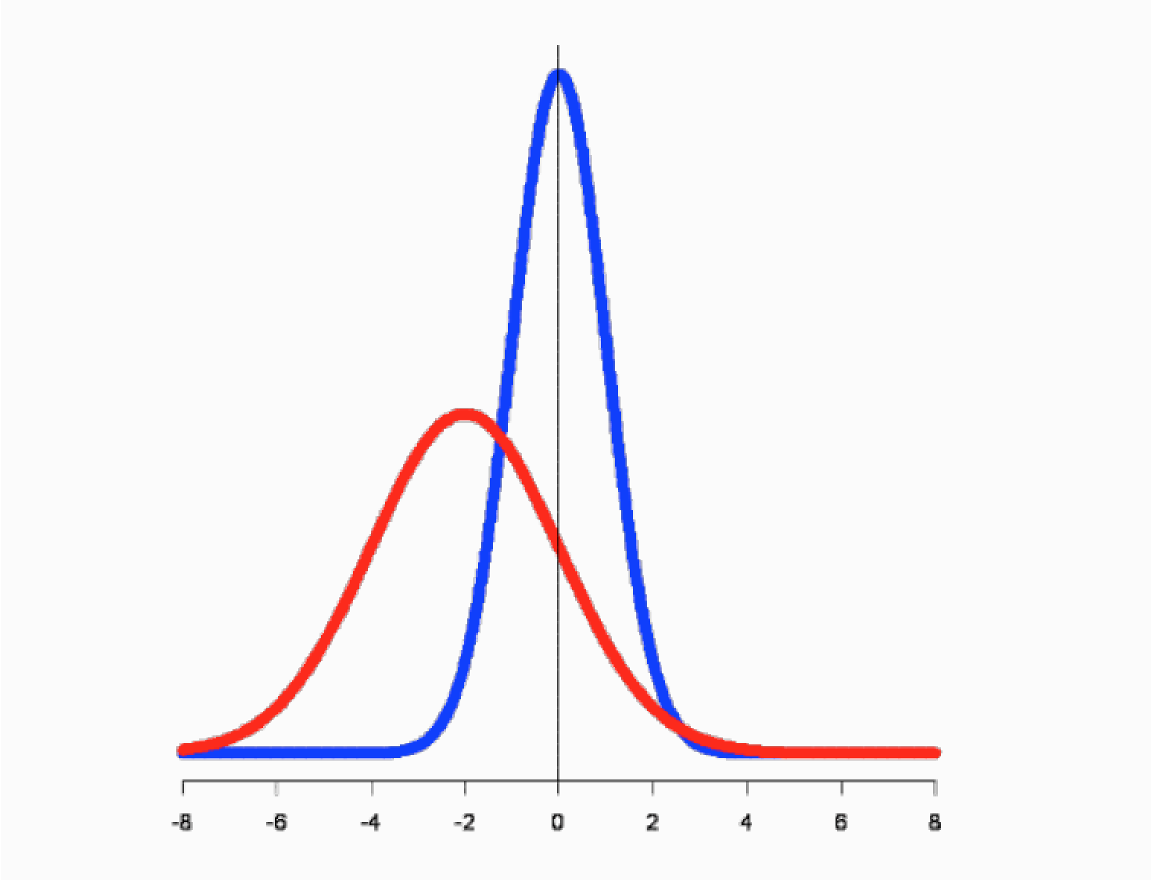
The Standard Normal Distribution

All normal distributions can be converted into the standard normal curve by subtracting the mean and dividing by the standard deviation:

$$Z = \frac{X - \mu}{\sigma}$$

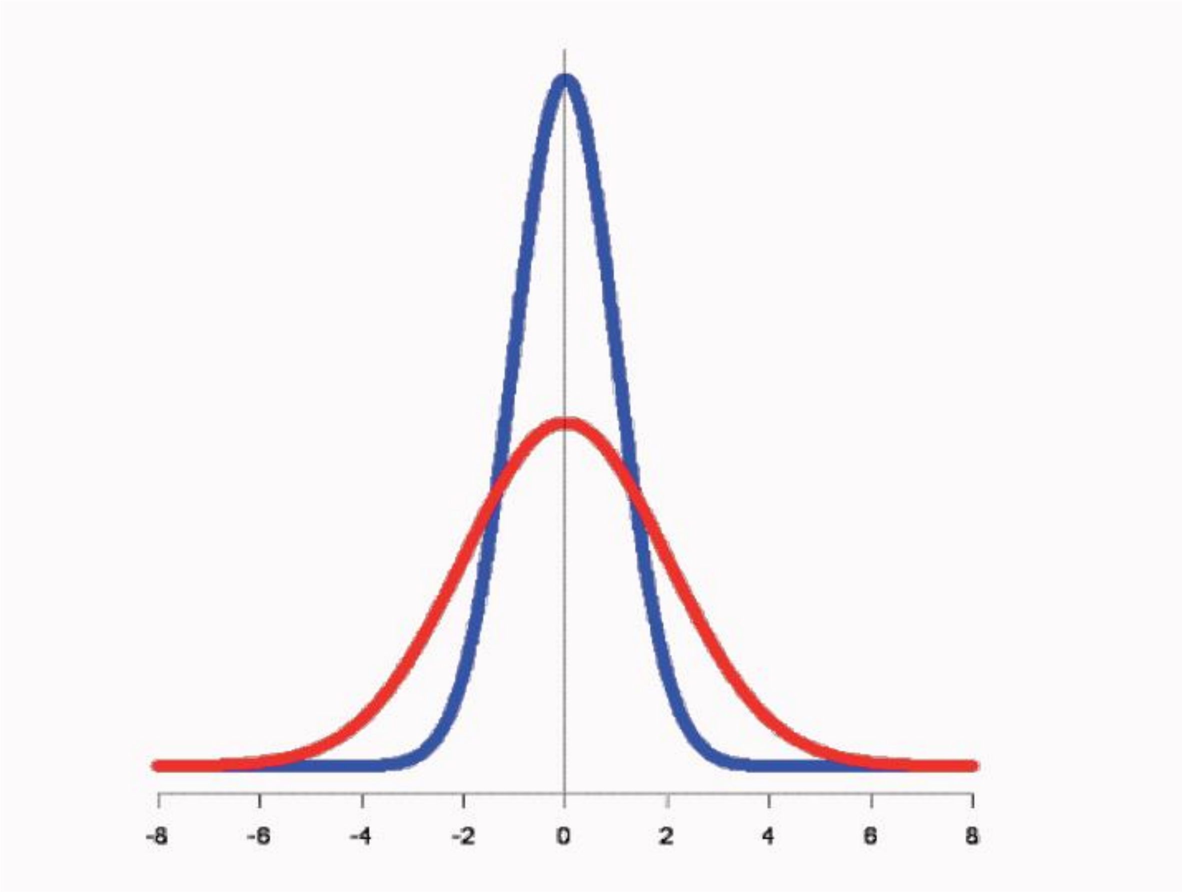
Transforming to Standard Normal

- The standard normal curve (blue) and another normal with mean -2, and standard deviation 2



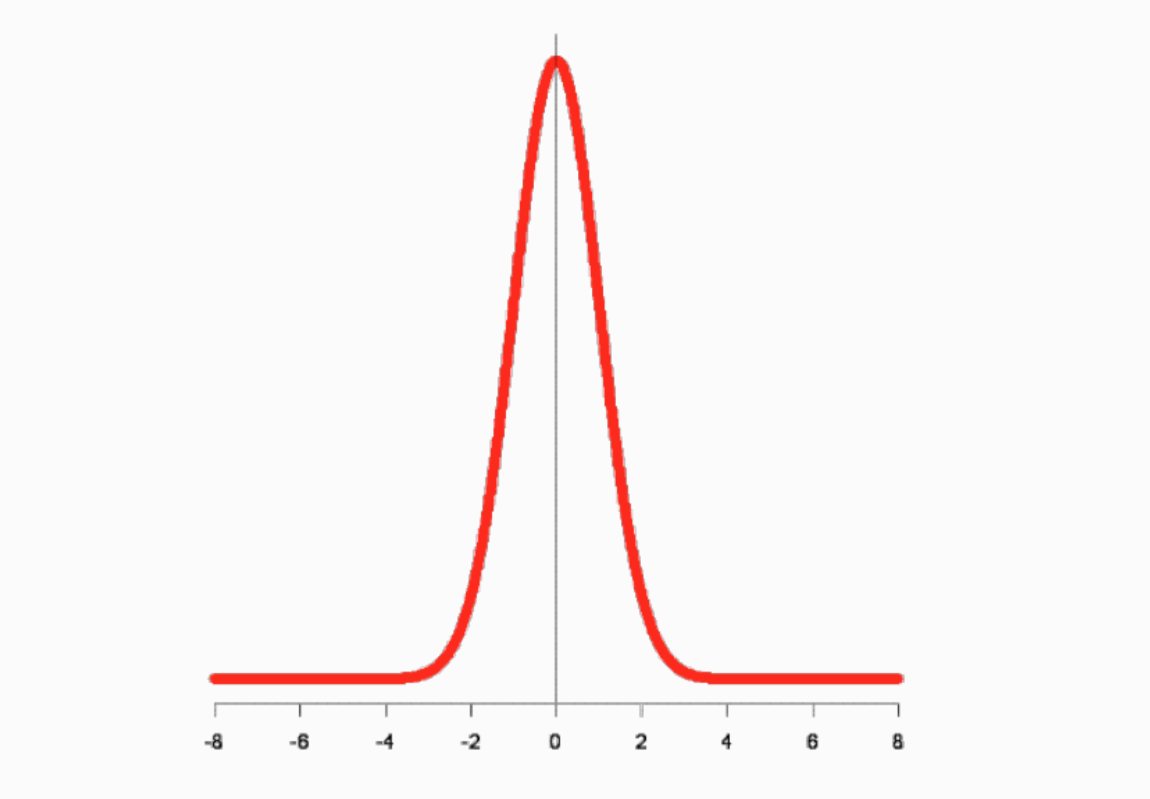
Transforming to Standard Normal

- To center at zero, subtract of mean of -2 from each observation under the red curve



Transforming to Standard Normal

- To "change shape" (i.e., change spread; i.e., standard deviation) divide each "new observation" by standard deviation of 2



The Standard Normal Curve

Z~Normal(μ=0, σ=1)

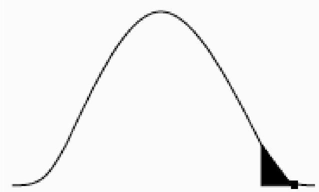
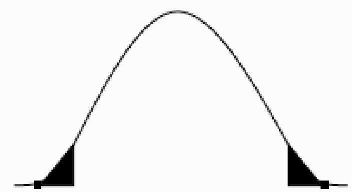
$$f(Z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z)^2}$$

The Standard Normal Distribution (Z)

Somebody calculated all the integrals for the standard normal and put them in a table! So we never have to integrate!

Even better, computers now do all the integration.

The Standard Normal Distribution (Z)

	Within Z SDs of the mean	More than Z SDs above the mean	More than Z SDs above or below the mean
Z			
1.0	68.27%	15.87%	31.73%
2.0	95.45%	2.28%	4.55%
2.5	98.76%	0.62%	1.24%
3.0	99.73%	0.13%	0.27%

Example

- For example, what's the probability of getting a math SAT score below 575 if SAT scores are normally distributed with a mean of 500 and a standard deviation of 50??

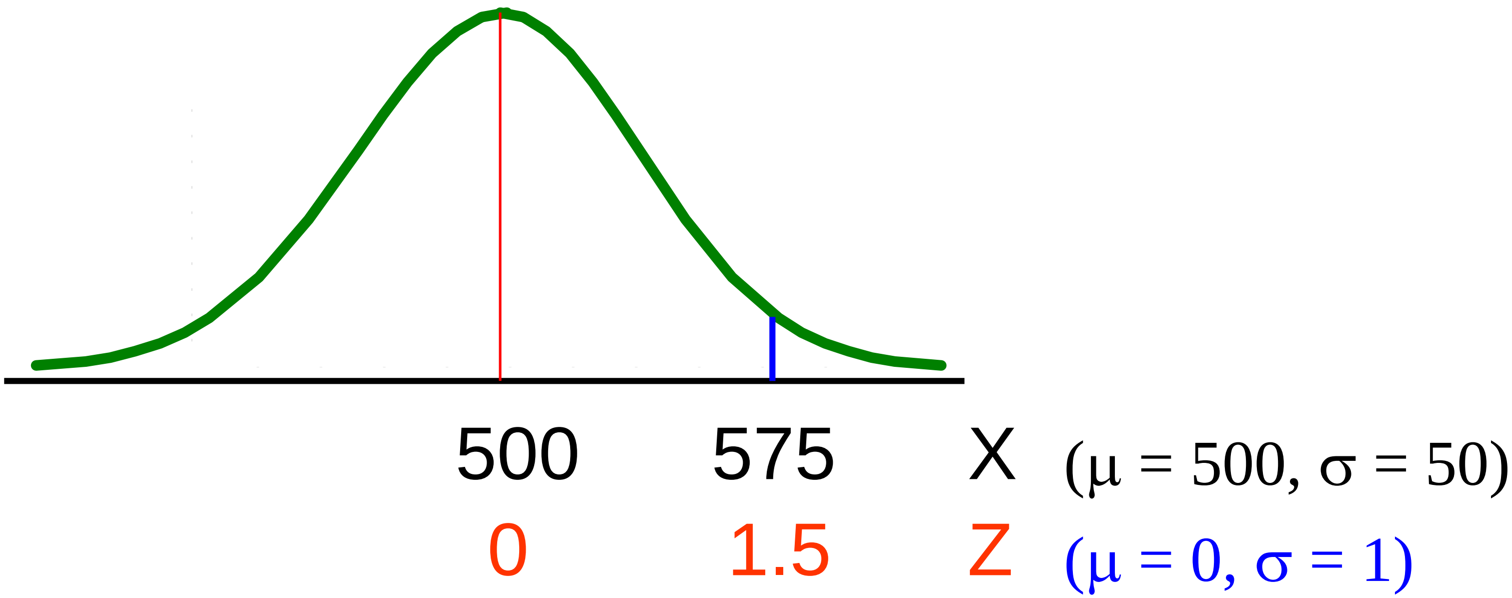
Example

- For example, what's the probability of getting a math SAT score below 575 if SAT scores are normally distributed with a mean of 500 and a standard deviation of 50??

$$\therefore P(X \leq 575) = \int_{200}^{575} \frac{1}{(50)\sqrt{2\pi}} \cdot e^{-\frac{1}{2}(\frac{x-500}{50})^2} dx$$

Solve this?.... (5min)

Comparing X and Z units



Example

So, What' s the probability of getting a math SAT score of 575 or less, μ=500 and σ=50?

Example

So, What's the probability of getting a math SAT score of 575 or less, μ=500 and σ=50?

$$\boxed{Z = \frac{575-500}{50} = 1.5}$$

Example

So, What's the probability of getting a math SAT score of 575 or less, μ=500 and σ=50?

$$Z = \frac{575-500}{50} = 1.5$$

$$\therefore P(X \leq 575) = \int_{200}^{575} \frac{1}{(50)\sqrt{2\pi}} \cdot e^{-\frac{1}{2}(\frac{x-500}{50})^2} dx \longrightarrow \int_{-\infty}^{1.5} \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}Z^2} dz$$

No need to do the integration!

Just look up Z= 1.5 in standard normal chart → no problem! = .9332

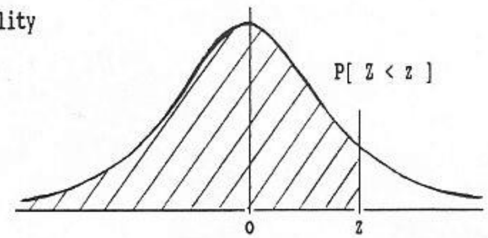
Looking up probabilities in the standard normal table

STANDARD STATISTICAL TABLES

1. Areas under the Normal Distribution

The table gives the cumulative probability up to the standardised normal value z
i.e.

$$P[\, Z < z \,] = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp(-\tfrac{1}{2}Z^2)\, dZ$$



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5159	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7854
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8804	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767

What is the area to the left of Z=1.50 in a standard normal curve?

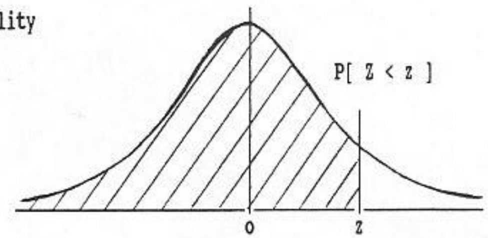
Looking up probabilities in the standard normal table

STANDARD STATISTICAL TABLES

1. Areas under the Normal Distribution

The table gives the cumulative probability up to the standardised normal value z i.e.

$$P[\,Z < z\,] = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp(-\tfrac{1}{2}Z^2)\, dZ$$



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5159	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7854
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8804	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767

Z=1.50

Z=1.50

What is the area to the left of Z=1.50 in a standard normal curve?

Area is 93.32%

Is my data "normal"?

- Not all continuous random variables are normally distributed!!
- It is important to evaluate how well the data are approximated by a normal distribution, because many statistics (ttest, ANOVA, linear regression) assume that the outcome variable is normally distributed.

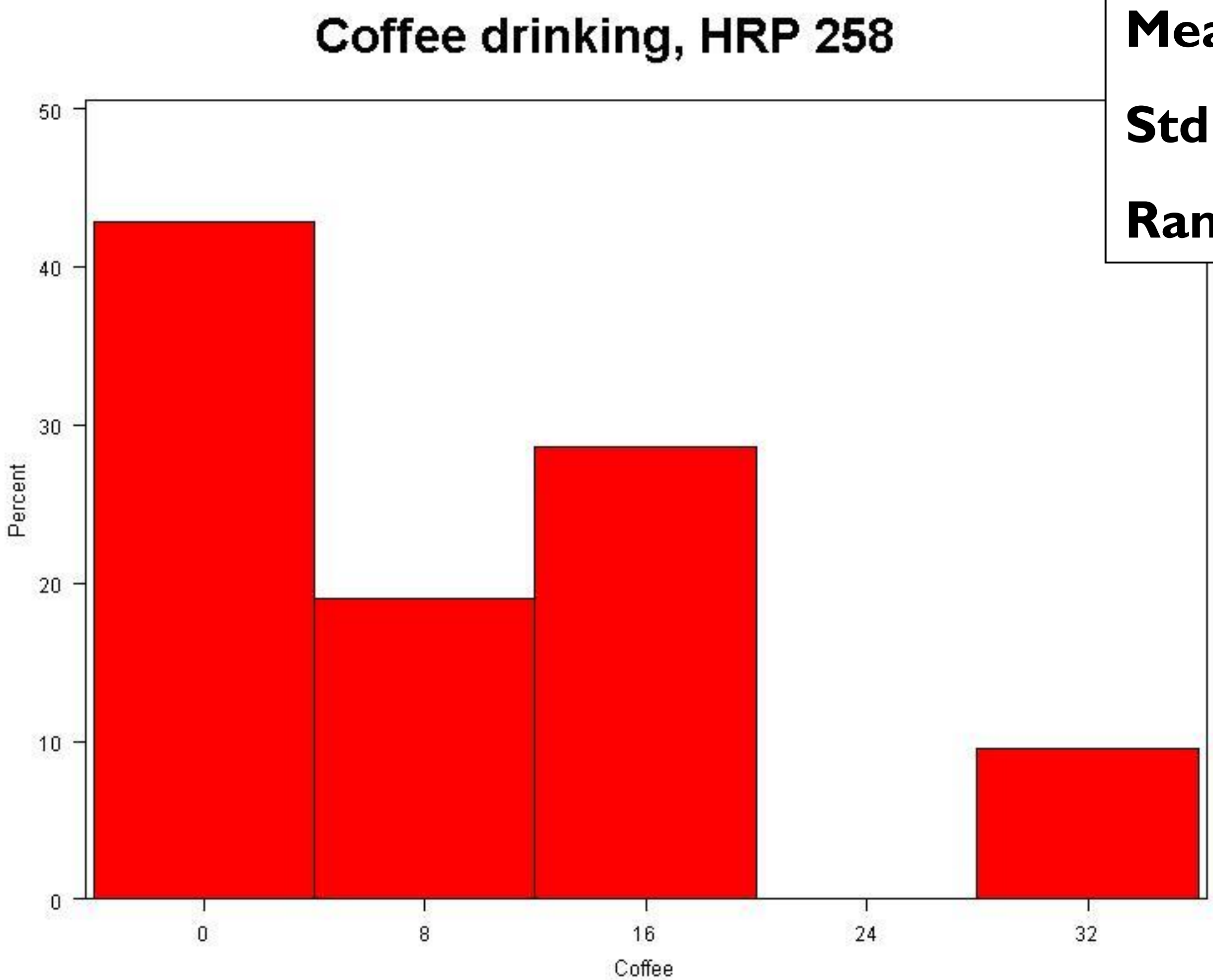
Is my data normally distributed?

Group Discussion

Are my data normally distributed?

1. Look at the histogram! Does it appear bell shaped?
2. Compute descriptive summary measures—are mean, median, and mode similar?
3. Do 2/3 of observations lie within 1 std dev of the mean? Do 95% of observations lie within 2 std dev of the mean?
4. Look at a normal probability plot（正态概率图, QQ plot）—is it approximately linear?
5. Run tests of normality (such as Kolmogorov-Smirnov). But, be cautious, highly influenced by sample size!

Example: coffee drinking (n=21)

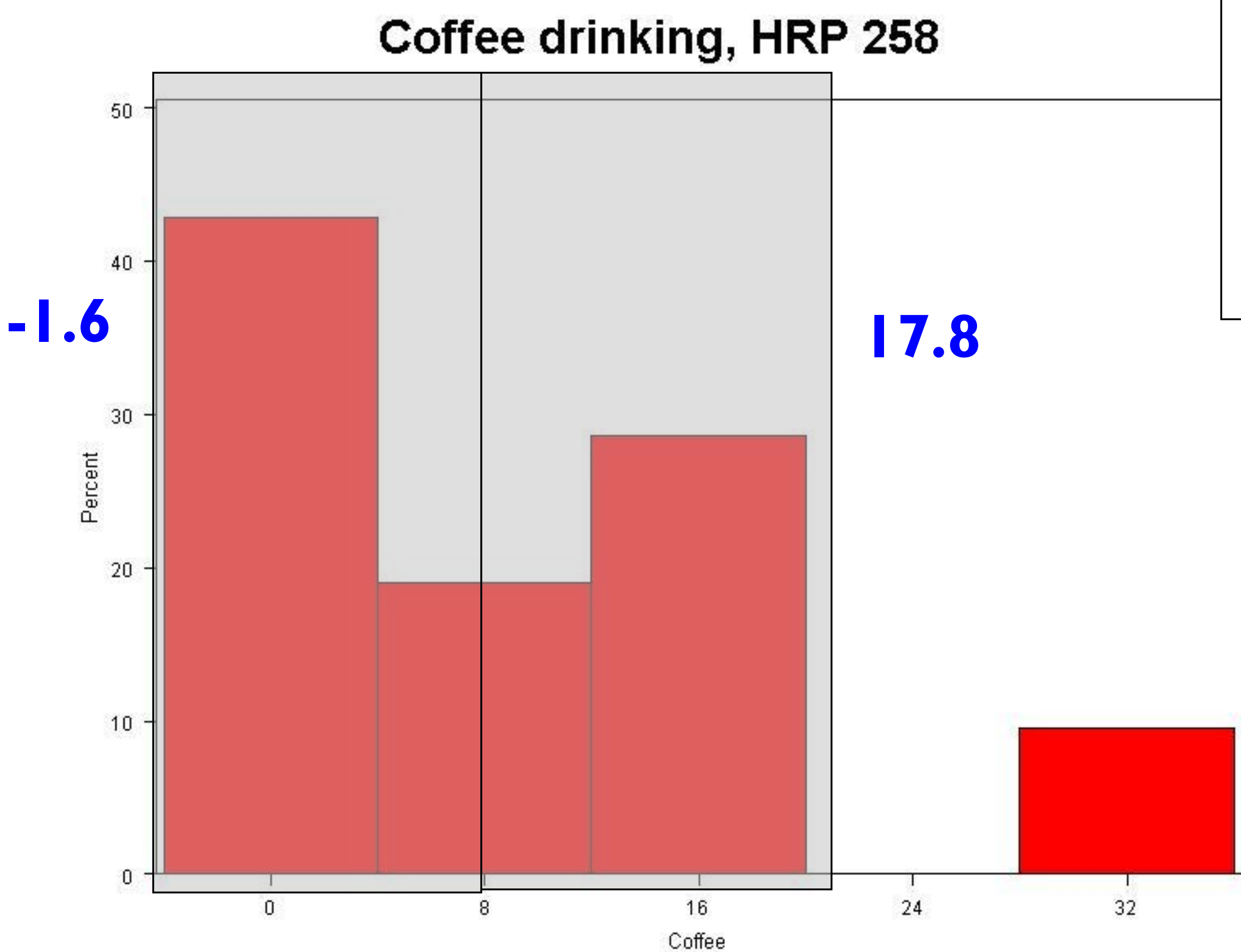


Mean=8.1 ounces/day

Std Dev=9.7 ounces/day

Range: 0 to 32

Example: coffee drinking (n=21)

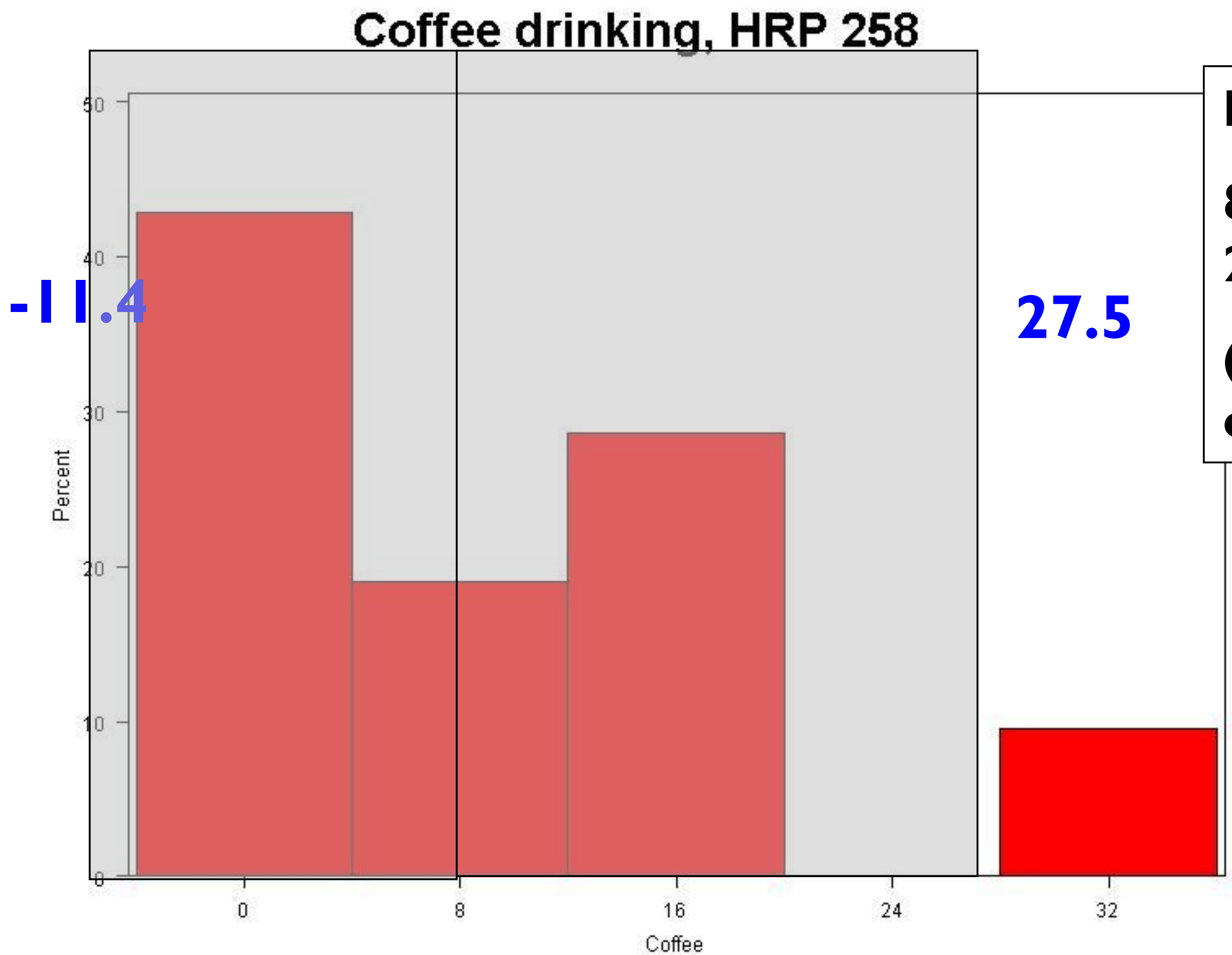


Mean±1 Std Dev=

8.1 ± 9.7 = -1.6 to 17.8

(covers 90% of observations)

Example: Class coffee drinking (n=21)

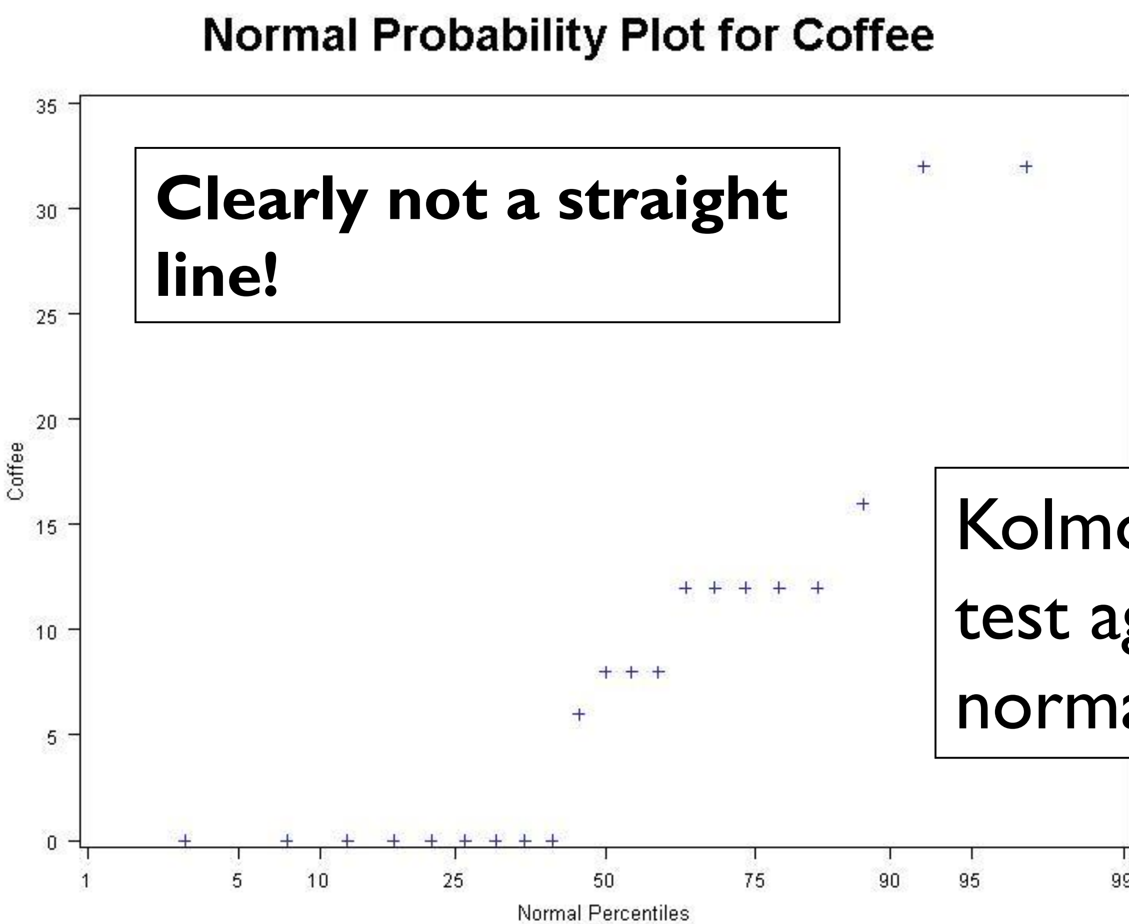


Mean±2 Std Dev=

8.1 ± 19.4 = -11.4 to 27.5

(covers 90% of observations)

Normal Probability Plot



Clearly not a straight line!

Kolmogorov-Smirnov test agrees, not normal!

Standard Normal (Z) Table

http://www.sjsu.edu/faculty/gerstman/EpiInfo/z-table.htm

Non-normal Data

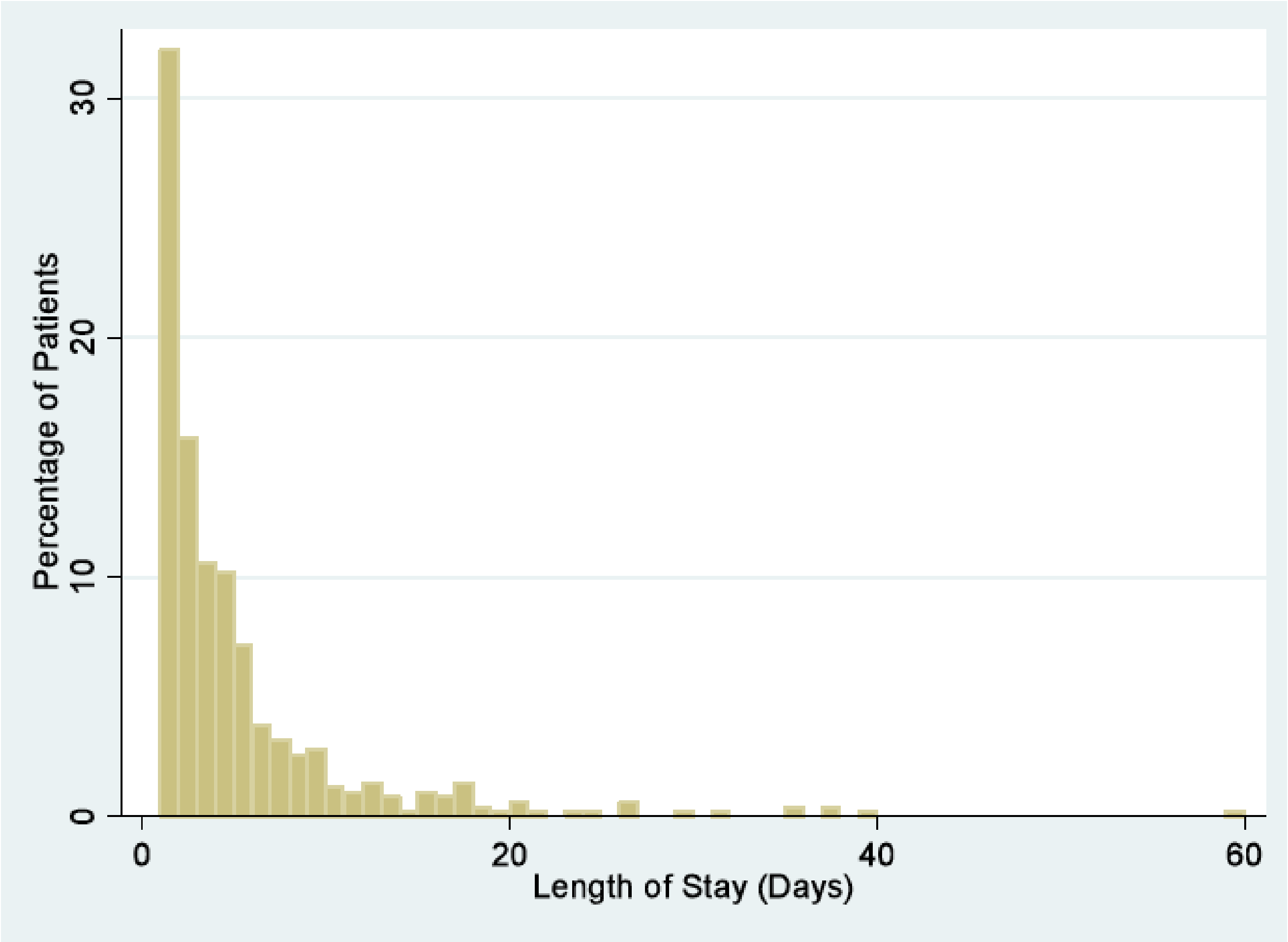
- Not all data is normal !
- Unless population/sample has a well known, "well behaved" (like a normal) distribution, we may not be able to use mean and standard deviation to create interpretable intervals, or measure "unusuality" of individual observations

Hospital Length of Stay Example

- Random sample of 500 patients
  - Mean length of stay: 4.8 days
  - Median length of stay: 3 days
  - Standard deviation: 6.3 days

Hospital Length of Stay Example

- Histogram of sample data



Hospital Length of Stay Example

- What percentage of patients had length of stay greater than five days?

(Wrong approach) z-score $z = \dfrac{5 - 4.8}{6.4} = 0.03$

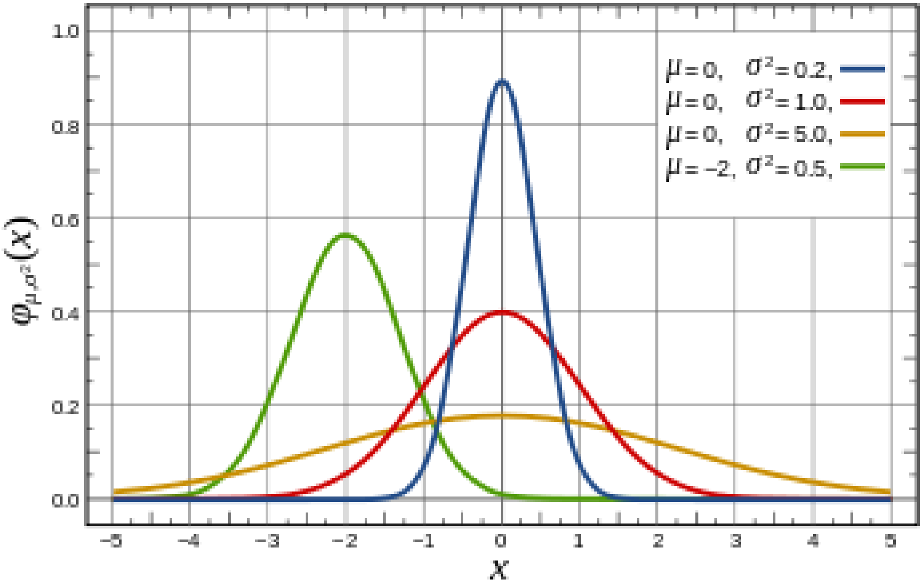
- Assuming normality, this would suggest that nearly 50% of the patients had length of stay greater than five days

Hospital Length of Stay Example

Percentiles	
1%	1
5%	1
10%	1
25%	1
50%	3
75%	5
90%	11
95%	17
99%	35

- According to percentiles, five days is the 75th percentile: so only 25% of the sample have length of stay over 5 days

Why normal distribution show up so much?



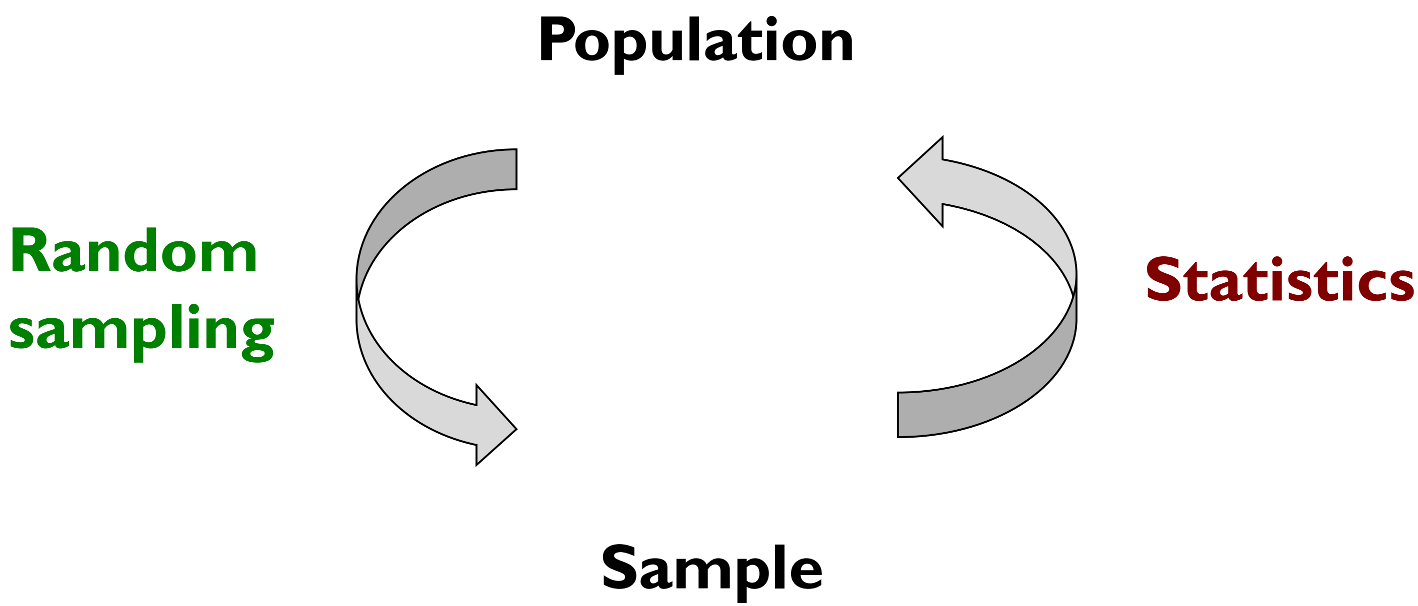
Central Limit Theorem

- Let $X_1, X_2, \ldots, X_n$ be a sequence of independently and identically distributed random variables with finite mean μ, and finite variance σ^2. Then:

$$\frac{\sqrt{n}\left(\overline{X}-\mu\right)}{\sigma} \xrightarrow{\text{Dist}} N(0,1) \quad \text{where } \overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

- Thus the limiting distribution of the sample mean is a normal distribution, regardless of the distribution of the individual measurements

Recall: Population and Samples



What is a statistic (统计量)?

- A **statistic** is any value that can be calculated from the sample data.
- Sample statistics are calculated to give us an idea about the larger population.

Examples of statistic:

- Mean
  - The average cost of a litre gas in Shanghai is ¥ 7.75
- Difference in means
  - The difference in the average gas price in 2013 (¥ 7.75) compared with 2003 (¥ 3.32) is ¥ 4.43.
- Proportion
  - 60% of students in SJTU eat breakfast regularly
- Difference in proportions
  - The difference in the proportion of female students who eat breakfast (70%) versus male students who do (50%) is 20%

Random Sample

- When a sample is randomly selected from a population, it is called a random sample
  - Technically speaking values in a random sample are representative of the distribution of the values in the population sample, regardless of size
- In a simple random sample, each individual in the population has an equal chance of being chosen for the sample
- Random sampling helps control systematic bias
- But even with random sampling, there is still sampling variability or error

Sampling Variability of a Sample Statistic

- If we repeatedly choose samples from the same population, a statistic will take different values in different samples

- If the statistic does not change much if you repeated the study (you get similar answers each time), then it is fairly reliable (not a lot of variability)

Sample statistics estimate population parameters

Sample statistic: mean IQ of 5 subjects

$$\frac{110+105+96+124+115}{5}=110$$

Truth (not observable)

Mean IQ of some population of 100,000 people =100

Sample (observation)

Make guesses about the whole population

Statistics vs. Parameters

- **Sample Statistic** – any summary measure calculated from data; e.g., could be a mean, a difference in means or proportions, an odds ratio, or a correlation coefficient
  - E.g., the mean vitamin D level in a sample of 100 men is 63 nmol/L
  - E.g., the correlation coefficient between vitamin D and cognitive function in the sample of 100 men is 0.15

- **Population parameter** – the true value/true effect in the entire population of interest

What is sampling variation?

- Statistics vary from sample to sample due to random chance.

- Example:
  - A population of 100,000 people has an average IQ of 100 (If you actually could measure them all!)
  - If you sample 5 random people from this population, what will you get?

Sampling Variation

<u>**Truth**</u> **(not observable)**

Mean IQ=100

Sampling Variation

$$\frac{120 + 160 + 180 + 95 + 95}{5} = 130$$

$$\frac{90 + 85 + 95 + 92 + 88}{5} = 90$$

$$\frac{110 + 105 + 96 + 124 + 115}{5} = 110$$

$$\frac{100 + 105 + 86 + 104 + 95}{5} = 98$$

<u>Truth</u> (not observable)

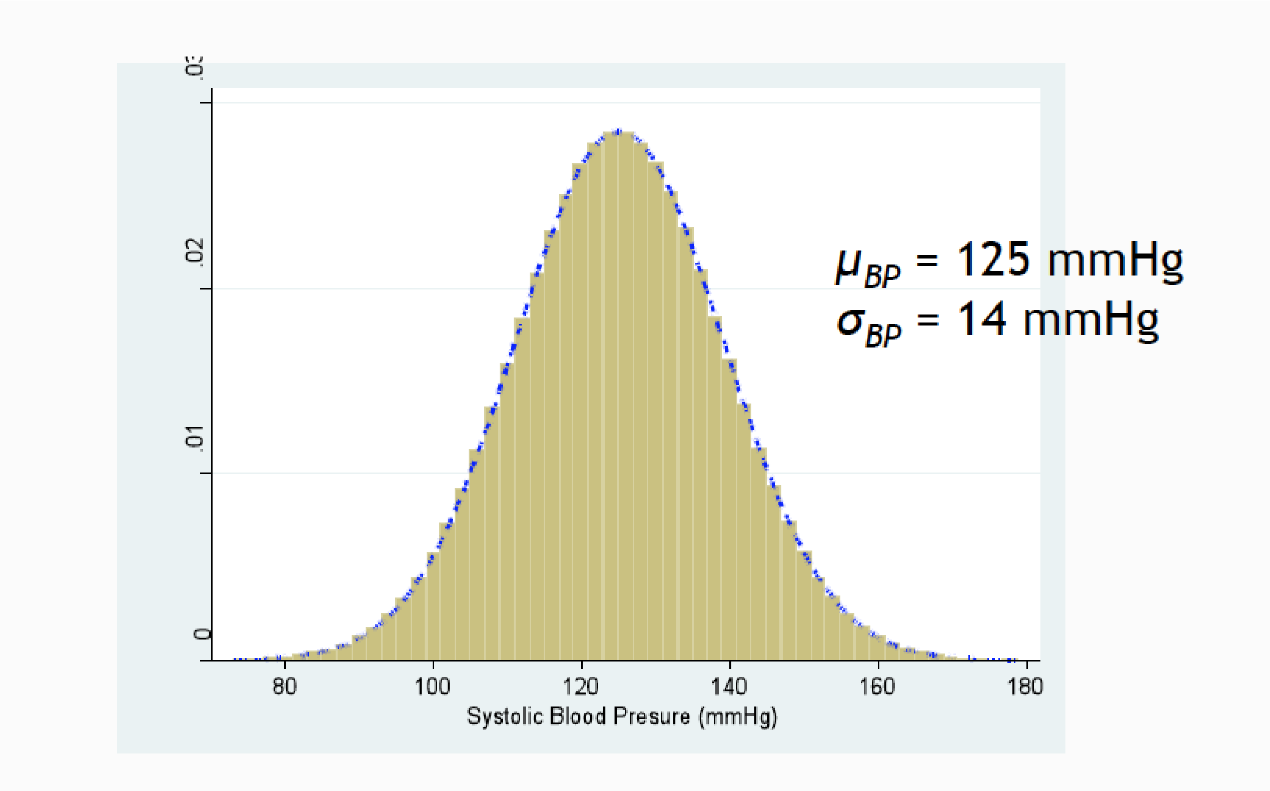
Mean IQ=100

Sampling Variation and Sample Size

- Do you expect more or less sampling variability in samples of 10 people?
- Of 50 people?
- Of 1000 people?
- Of 100,000 people?

Example: Blood Pressure of Males

- We have data on blood pressures using a random sample of 113 men taken from the population of all men
- Assume the population distribution is given by the following:

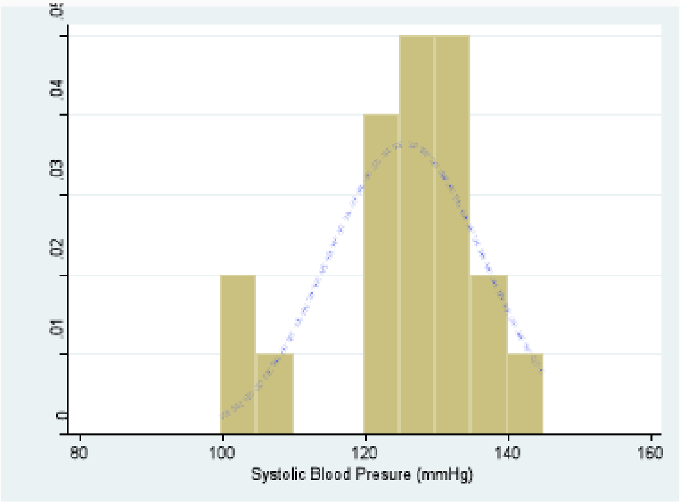


Example: Blood Pressure of Males

- Suppose we had all the time in the world
  - We decide to do an experiment
  - We are going to take 500 separate random samples from this
- population of men, each with 20 subjects
  - For each of the 500 samples, we will plot a histogram of the sample
- BP values, and record the sample mean and sample standard deviation

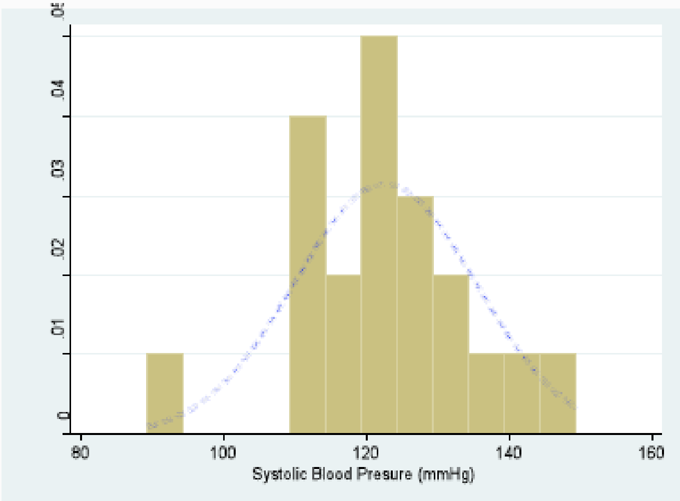
Random Samples

- Sample 1: n = 20



$$\bar{x}_{BP} = 125.7 \text{ mmHg}$$
$$s_{BP} = 10.9 \text{ mmHg}$$

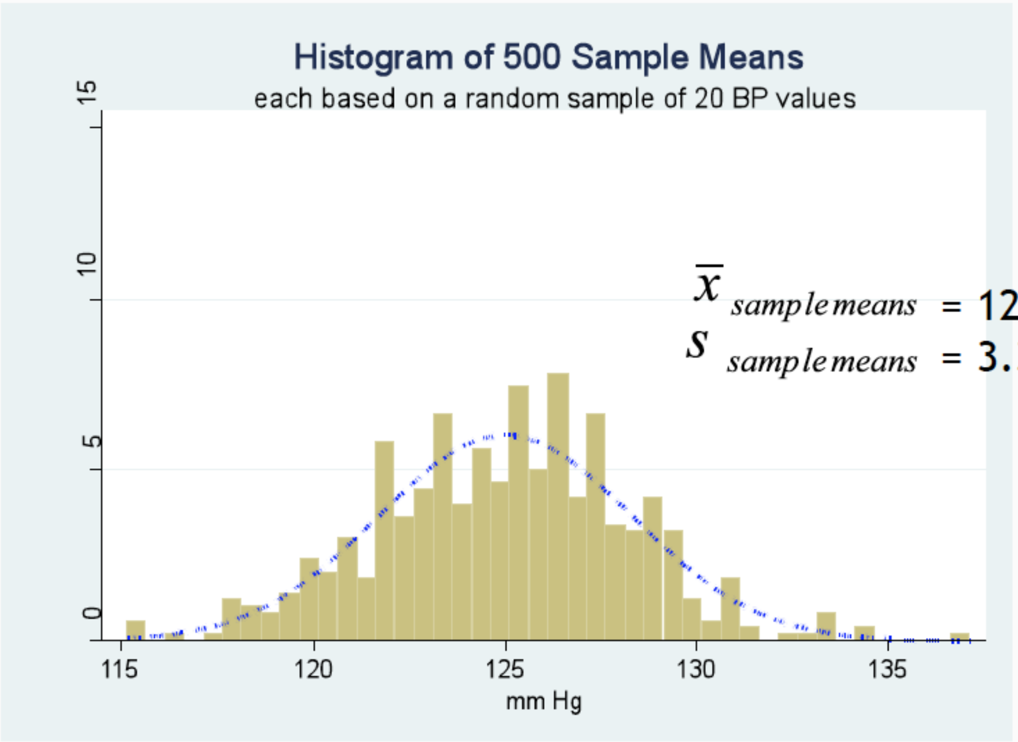
- Sample 2: n = 20



$$\bar{x}_{BP} = 122.6 \text{ mmHg}$$
$$s_{BP} = 12.7 \text{ mmHg}$$

Example: Blood Pressure of Males

- So we did this 500 times: now let's look at a histogram of the 500 sample means

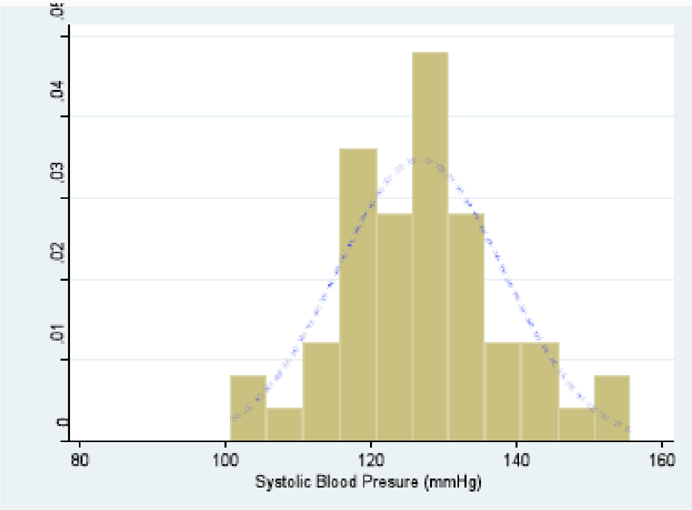


$\bar{x}_{sample\,means} = 125$ mmHg

$s_{sample\,means} = 3.3$ mmHg

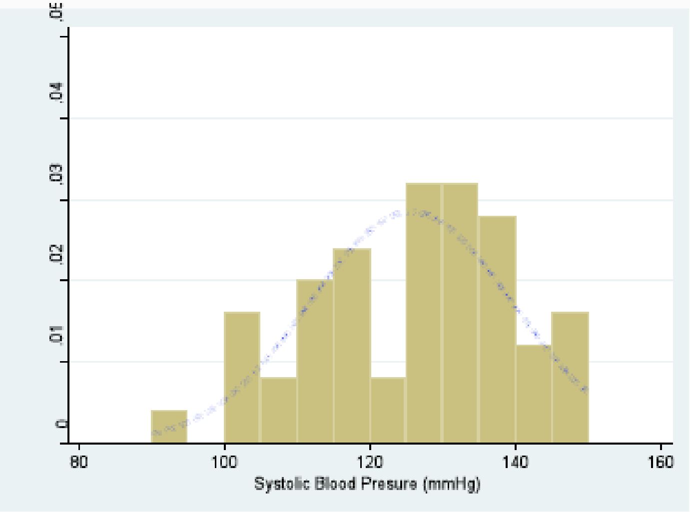
Random Samples

- Sample 1: n = 50



$$\bar{x}_{BP} = 126.7 \text{ mmHg}$$
$$s_{BP} = 11.5 \text{ mmHg}$$

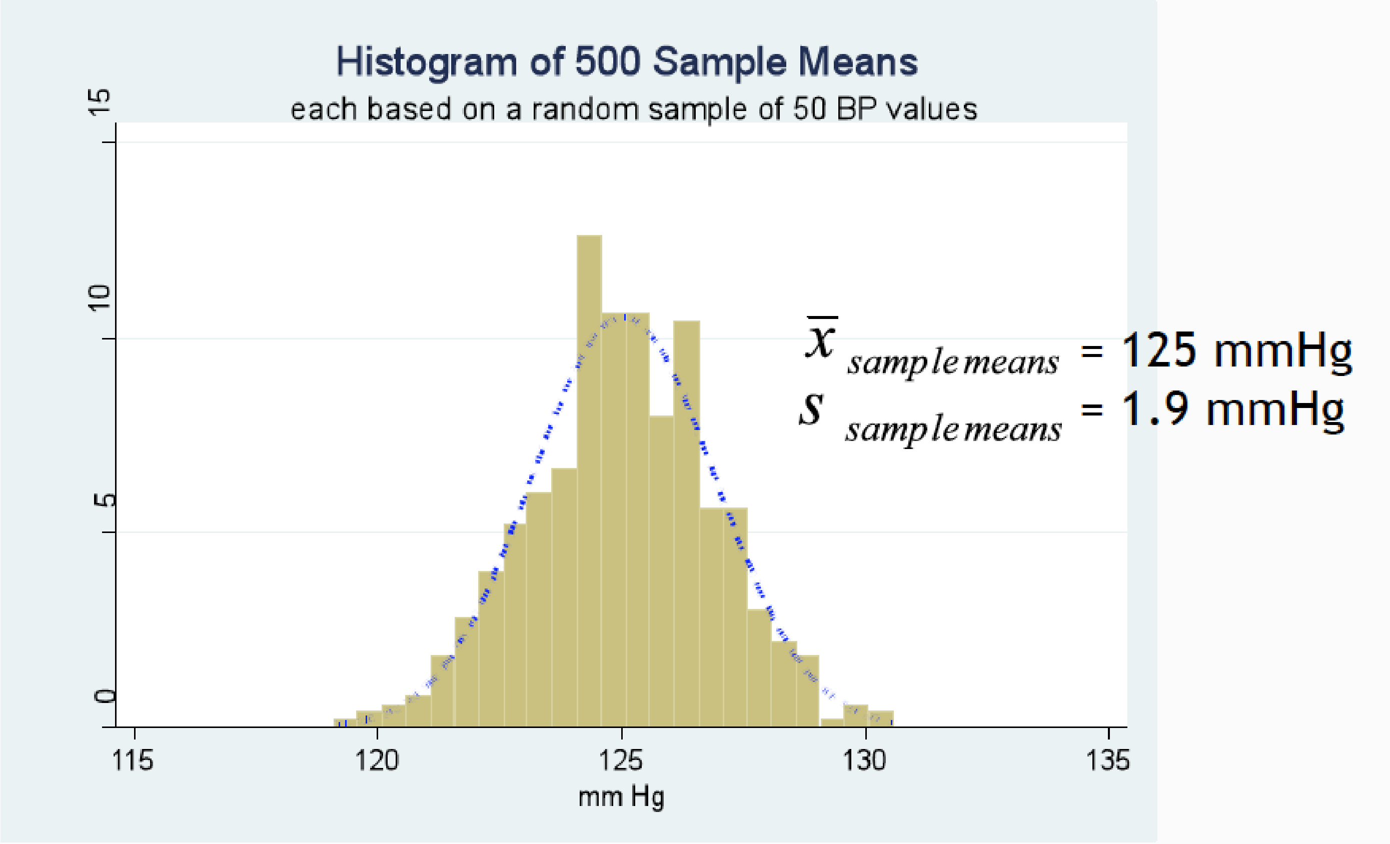
- Sample 2: n = 50



$$\bar{x}_{BP} = 125.5 \text{ mmHg}$$
$$s_{BP} = 14.0 \text{ mmHg}$$

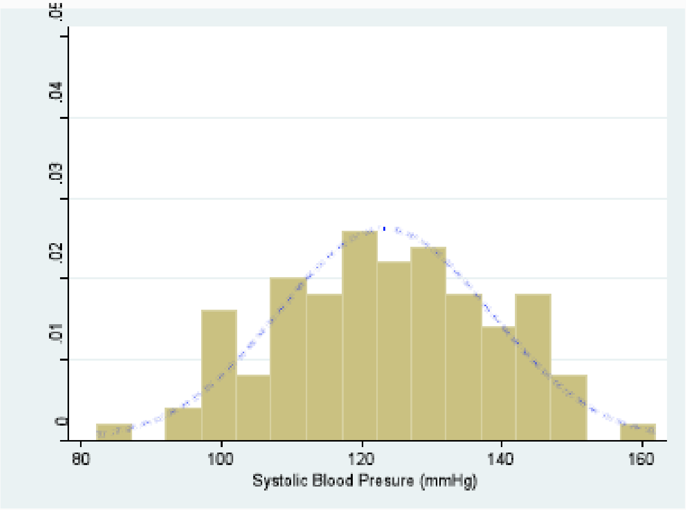
Example: Blood Pressure of Males

- So we did this 500 times: now let's look at a histogram of the 500 sample means



Random Samples

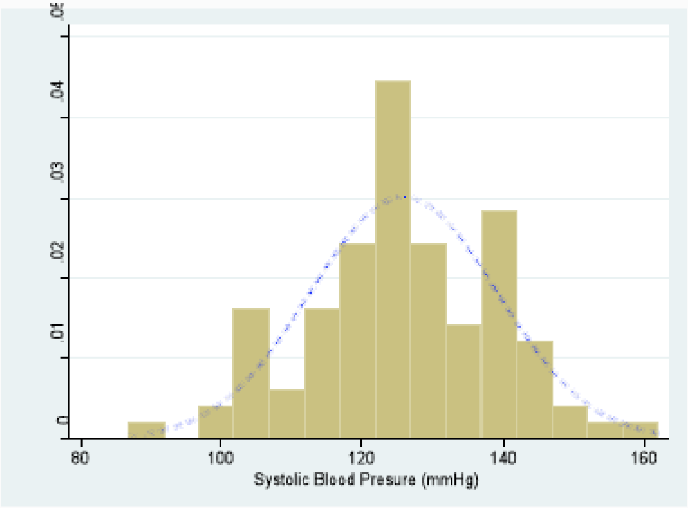
- Sample 1: n = 100



$\bar{x}_{BP}$ = 123.3 mmHg

s_{BP} = 15.2 mmHg

- Sample 2: n = 100

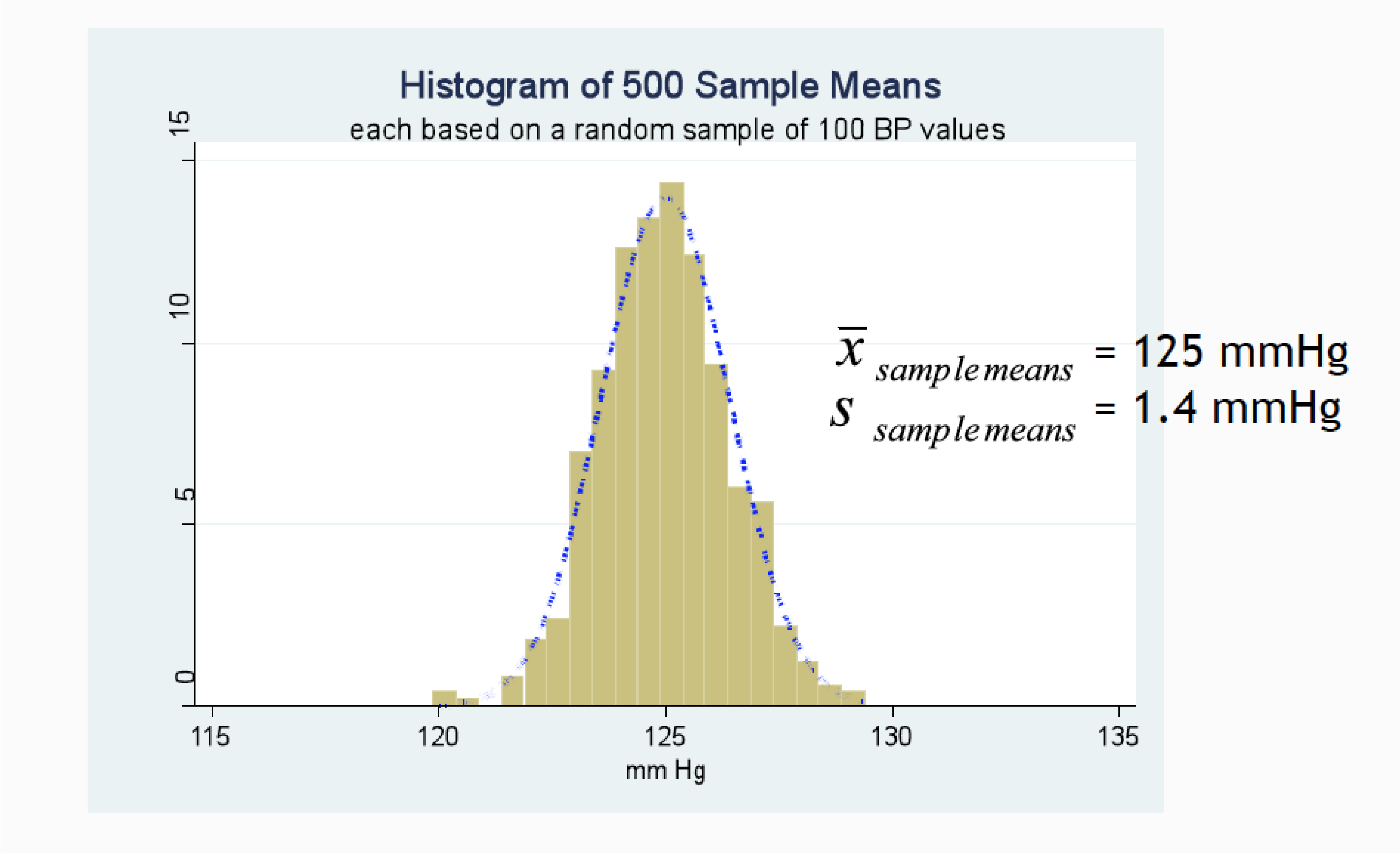


$\bar{x}_{BP}$ = 125.7 mmHg

s_{BP} = 13.2 mmHg

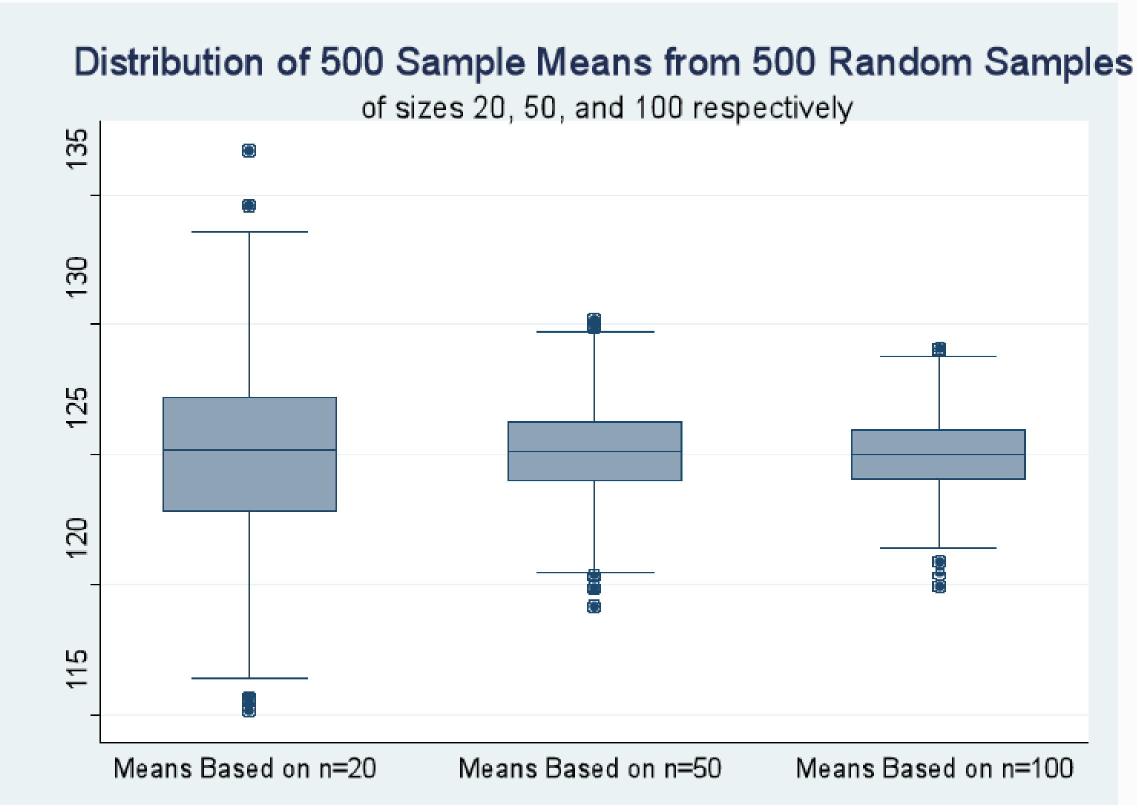
Example: Blood Pressure of Males

- So we did this 500 times: now let's look at a histogram of the 500 sample means



Example: Blood Pressure of Males

- Review the results



Example: Blood Pressure of Males

- Review the results

Sample Sizes	Means of 500 Sample Means	SD of 500 Sample Means	Shape of Distribution of 500 sample means
n = 20	125 mmHg	3.3 mm Hg	Approx normal
n = 50	125 mmHg	1.9 mm Hg	Approx normal
n = 100	125 mmHg	1.4 mm Hg	Approx normal

The standard error of a mean

$$\text{s.e.} = \frac{s}{\sqrt{n}}$$

- Standard deviation in the sample means of size n, often called the standard error of the sample mean
- The standard error measures the amount of variability in the sample mean; it indicates how closely the population mean is likely to be estimated by the sample mean.
- The standard deviation measures the amount of variability in the population
- Because standard deviations and standard errors are often confused it is very important that they are clearly labelled when presented in tables of results.

Standard error

- Standard Error is a measure of sampling variability.
- Standard error is the standard deviation of a sample statistic.
- Standard error decreases with increasing sample size and increases with increasing variability of the outcome (e.g., IQ).
- Standard errors can be predicted by computer simulation or mathematical theory (formulas).
  - **The formula for standard error is different for every type of statistic (e.g., mean, difference in means, odds ratio).**