Probability Distributions

Random Variable

- A random variable *X* takes on a defined set of values with different probabilities.
    - For example, if you roll a die, the outcome is random (not fixed) and there are 6 possible outcomes, each of which occur with probability one-sixth.
    - For example, if you poll people about their voting preferences, the percentage of the sample that responds "Yes on Proposition 100" is a also a random variable (the percentage will be slightly different every time you poll).

- Roughly, <u>probability</u> is how frequently we expect different outcomes to occur if we repeat the experiment over and over ("frequentist" view)

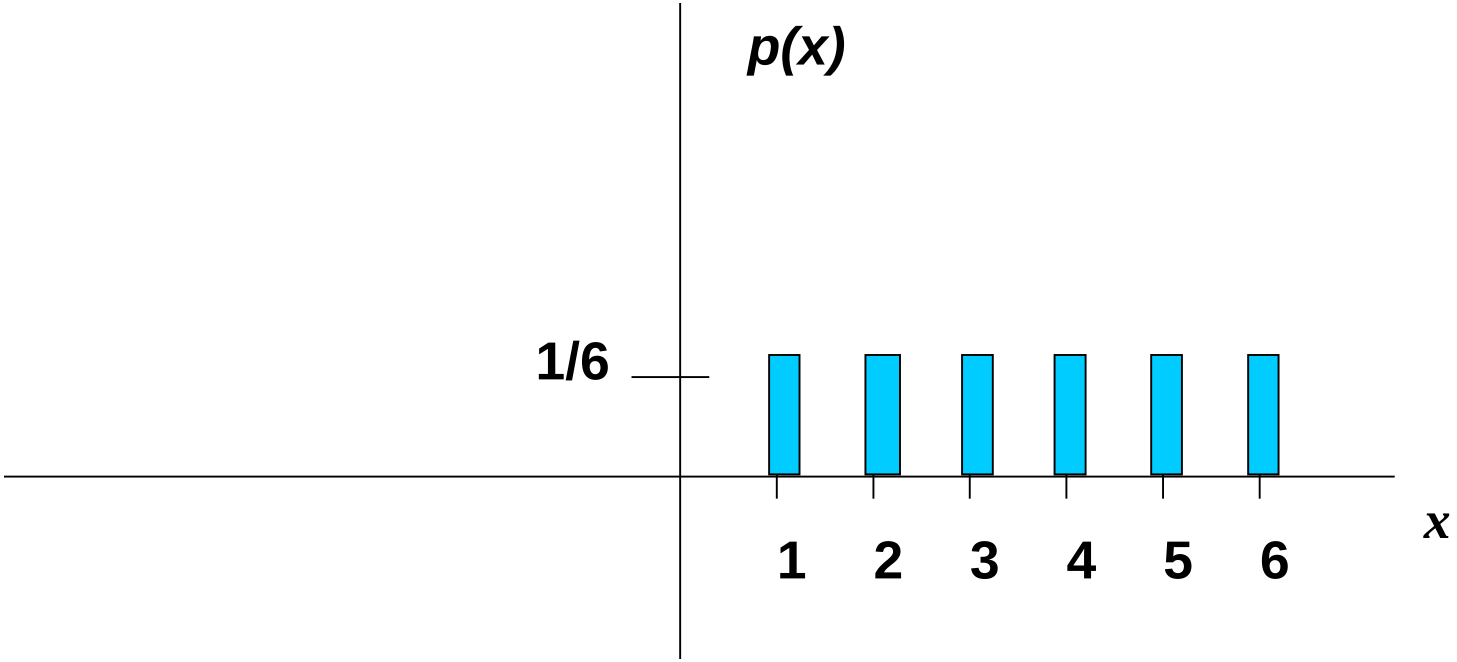
Random variables can be discrete or continuous

- **Discrete** random variables have a countable number of outcomes
  - Examples: Dead/alive, treatment/placebo, dice, counts, etc.
- **Continuous** random variables have an infinite continuum of possible values.
  - Examples: blood pressure, weight, the speed of a car, the real numbers from 1 to 6.

Probability functions

- A probability function maps the possible values of *x* against their respective probabilities of occurrence, *p(x)*
- *p(x)* is a number from 0 to 1.0.
- The area under a probability function is always 1.

Discrete example: roll of a die



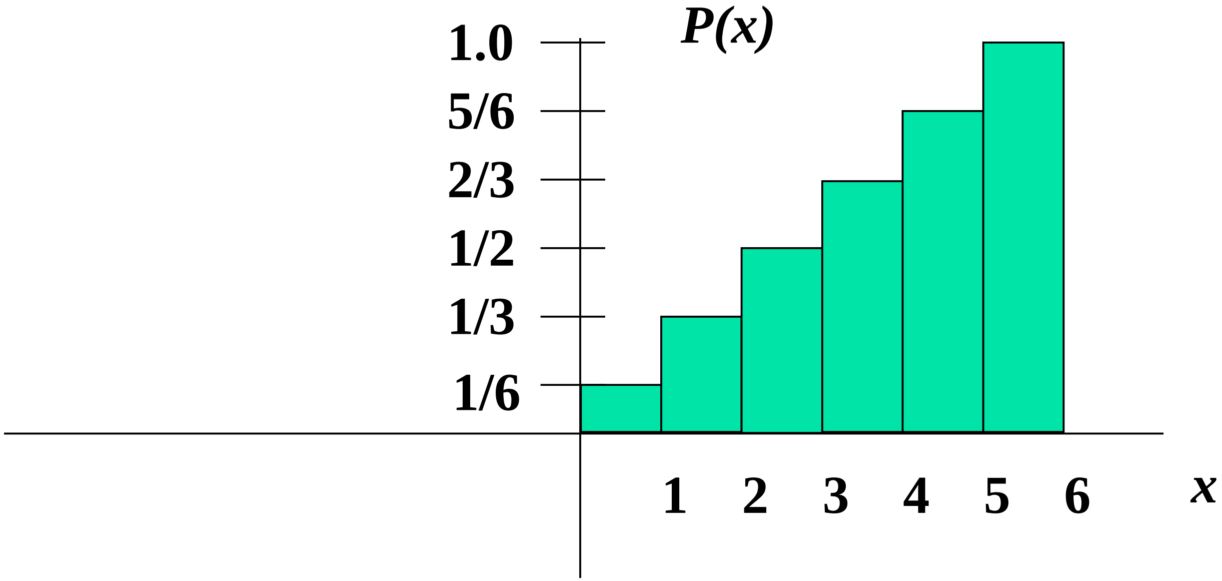
$$\sum_{\text{all } x} P(x) = 1$$

Probability mass function (pmf)

x	$p(x)$
1	$p(x=1)=1/6$
2	$p(x=2)=1/6$
3	$p(x=3)=1/6$
4	$p(x=4)=1/6$
5	$p(x=5)=1/6$
6	$\underline{p(x=6)=1/6}$

1.0

Cumulative distribution function (CDF)



Cumulative distribution function

x	$P(x \le A)$
1	$P(x \le 1)=1/6$
2	$P(x \le 2)=2/6$
3	$P(x \le 3)=3/6$
4	$P(x \le 4)=4/6$
5	$P(x \le 5)=5/6$
6	$P(x \le 6)=6/6$

Examples

1. What’s the probability that you roll a 3 or less?

$P(x \leq 3)$=1/2

2. What’s the probability that you roll a 5 or higher?

$P(x \geq 5)$ = 1 – $P(x \leq 4)$ = 1-2/3 = 1/3

Practice Problem

Which of the following are probability functions?

a. $f(x)=.25$ for x=9,10,11,12

b. $f(x)= (3-x)/2$ for x=1,2,3,4

c. $f(x)= (x^2+x+1)/25$ for x=0,1,2,3

Answer (a)

a. $f(x)=.25$ for $x=9,10,11,12$

x	***f(x)***
9	**.25**
10	**.25**
11	**.25**
12	**.25**

1.0

Yes, probability function!

Answer (b)

b. *f(x)*= *(3-x)*/2 for x=1,2,3,4

x	***f(x)***
1	**(3-1)/2=1.0**
2	**(3-2)/2=.5**
3	**(3-3)/2=0**
4	**(3-4)/2=-.5**

Though this sums to 1, you can't have a negative probability; therefore, it's not a probability function.

Answer (c)

c. $f(x)= (x^2+x+1)/25$ for x=0,1,2,3

x	f(x)
0	1/25
1	3/25
2	7/25
3	13/25

24/25

Doesn’t sum to 1. Thus, it’s not a probability function.

Practice Problem:

- The number of times that Rohan wakes up in the night is a random variable represented by *x*. The probability distribution for *x* is:

x	**1**	**2**	**3**	**4**	**5**
P(x)	**.1**	**.1**	**.4**	**.3**	**.1**

Find the probability that on a given night:

a. He wakes exactly 3 times **$p(x=3)= .4$**

b. He wakes at least 3 times **$p(x\geq3)= (.4 + .3 +.1) = .8$**

c. He wakes less than 3 times **$p(x<3)= (.1 +.1) = .2$**

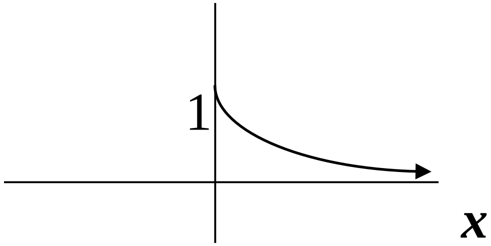
Important discrete distributions in epidemiology…

- Binomial (coming soon…)
  - Yes/no outcomes (dead/alive, treated/untreated, smoker/non-smoker, sick/well, etc.)
- Poisson
  - Counts (e.g., how many cases of disease in a given area)

Continuous case

- The probability function that accompanies a continuous random variable is a continuous mathematical function that integrates to 1.
  - For example, recall the negative exponential function (in probability, this is called an "exponential distribution"): $f(x) = e^{-x}$
  - This function integrates to 1:

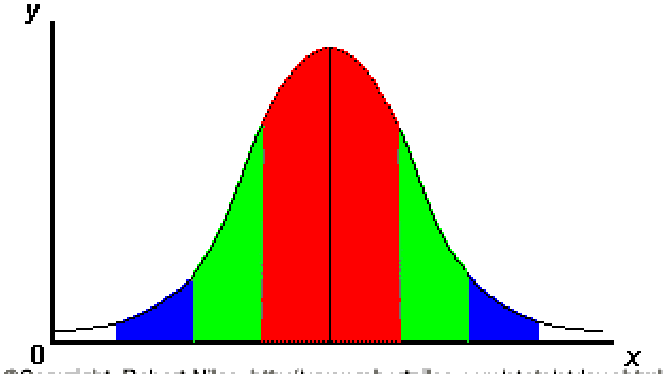
$$\int_0^{+\infty} e^{-x} = -e^{-x} \Big|_0^{+\infty} = 0 + 1 = 1$$



Review: Continuous case

- The normal distribution function also integrates to 1 (i.e., the area under a bell curve is always 1):

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



©Copyright, Robert Niles, http://www.robertniles.com/stats/stdev.shtml

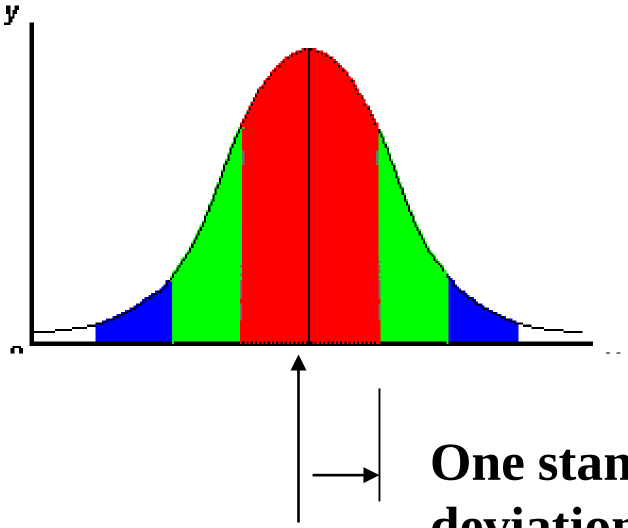
Review: Continuous case

- The probabilities associated with continuous functions are just areas under the curve (integrals!).
- Probabilities are given for a range of values, rather than a particular value (e.g., the probability of getting a math SAT score between 700 and 800 is 2%).

Expected Value and Variance

- All probability distributions are characterized by an expected value (=mean!) and a variance (standard deviation squared).

For example, bell-curve (normal) distribution:



Mean (μ)

One standard deviation from the mean (σ)

Expected value, or mean

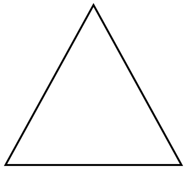
- If we understand the underlying probability function of a certain phenomenon, then we can make informed decisions based on how we expect *x* to behave on-average over the long-run…(so called "frequentist" theory of probability).

- Expected value is just the weighted average or mean (μ) of random variable *x*. Imagine placing the masses *p*(*x*) at the points *X* on a beam; the balance point of the beam is the expected value of *x*.

Example: expected value

- Recall the following probability distribution of Rohan’s waking pattern:

x	**1**	**2**	**3**	**4**	**5**
P(x)	**.1**	**.1**	**.4**	**.3**	**.1**



$$\boxed{\sum_{i=1}^{5} x_i\, p(x) = 1(.1) + 2(.1) + 3(.4) + 4(.3) + 5(.1) = 3.2}$$

Expected value, formally

Discrete case:

$$\boxed{E(X) = \mu = \sum_{\text{all x}} x_i p(x_i)}$$

Continuous case:

$$\boxed{E(X) = \mu = \int_{\text{all x}} x_i p(x_i) dx}$$

Sample Mean is a special case of Expected Value…

Sample mean, for a sample of n subjects: =

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \sum_{i=1}^{n} x_i \left(\frac{1}{n}\right)$$

The probability (frequency) of each person in the sample is 1/n.

Variance/standard deviation

“The average (expected) squared distance (or deviation) from the mean”

$$\sigma^2 = Var(x) = E[(x-\mu)^2] = \sum_{\text{all x}} (x_i - \mu)^2 p(x_i)$$

***We square because squaring has better properties than absolute value. Take square root to get back linear average distance from the mean (=”standard deviation”).*

Variance, formally

Discrete case:

$$Var(X) = \sigma^2 = \sum_{\text{all x}} (x_i - \mu)^2 p(x_i)$$

Continuous case:

$$Var(X) = \sigma^2 = \int_{-\infty}^{\infty} (x_i - \mu)^2 p(x_i) dx$$

Sample variance is a special case…

The variance of a sample: s^2 =

$$\frac{\sum_{i=1}^{N}(x_i-\bar{x})^2}{n-1}=\sum_{i=1}^{N}(x_i-\bar{x})^2\left(\frac{1}{n-1}\right)$$

Division by n-1 reflects the fact that we have lost a “degree of freedom” (piece of information) because we had to estimate the sample mean before we could estimate the sample variance.

Practice Problem

A roulette wheel has the numbers 1 through 36, as well as 0 and 00. If you bet $1.00 that an odd number comes up, you win or lose $1.00 according to whether or not that event occurs. If *X* denotes your net gain, *X*=1 with probability 18/38 and *X*= -1 with probability 20/38.

We already calculated the mean to be = -$.053. What's the variance of *X*?

Answer

$$\sigma^2 = \sum_{\text{all x}} (x_i - \mu)^2 p(x_i)$$

$$= (+1 - -.053)^2 (18/38) + (-1 - -.053)^2 (20/38)$$

$$= (1.053)^2 (18/38) + (-1 + .053)^2 (20/38)$$

$$= (1.053)^2 (18/38) + (-.947)^2 (20/38)$$

$$= .997$$

$$\sigma = \sqrt{.997} = .99$$

Standard deviation is $.99. Interpretation: On average, you're either 1 dollar above or 1 dollar below the mean, which is just under zero. Makes sense!

calculation formula!

$$\boxed{Var(X) = \sum_{\text{all x}} (x_i - \mu)^2 p(x_i) = \sum_{\text{all x}} x_i^{\ 2} p(x_i) - (\mu)^2}$$

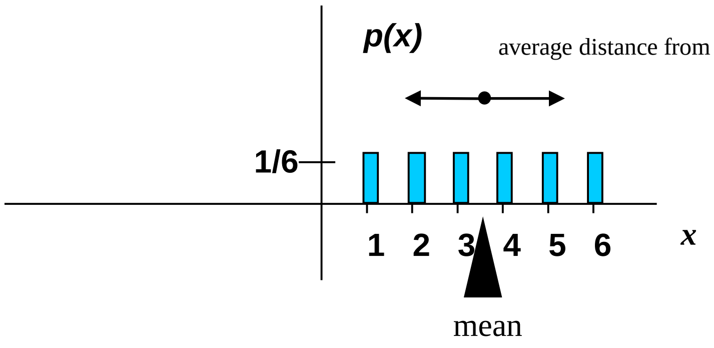
Intervening algebra!

$$\boxed{= E(x^2) - [E(x)]^2}$$

For example, what are the mean and standard deviation of the roll of a die?

x	*p(x)*
1	*p(x=1)*=1/6
2	*p(x=2)*=1/6
3	*p(x=3)*=1/6
4	*p(x=4)*=1/6
5	*p(x=5)*=1/6
6	*p(x=6)*=1/6

1.0



$$E(x)=\sum_{\text{all } x} x_i\, p(x_i)=(1)(\frac{1}{6})+2(\frac{1}{6})+3(\frac{1}{6})+4(\frac{1}{6})+5(\frac{1}{6})+6(\frac{1}{6})=\frac{21}{6}=3.5$$

$$E(x^2)=\sum_{\text{all } x} x_i^{\ 2}\, p(x_i)=(1)(\frac{1}{6})+4(\frac{1}{6})+9(\frac{1}{6})+16(\frac{1}{6})+25(\frac{1}{6})+36(\frac{1}{6})=15.17$$

$$\sigma_x^2 = Var(x)=E(x^2)-[E(x)]^2=15.17-3.5^2=2.92$$

$$\sigma_x=\sqrt{2.92}=1.71$$

Practice Problem

Find the variance and standard deviation for Rohan’s night wakings (recall that we already calculated the mean to be 3.2):

x	**1**	**2**	**3**	**4**	**5**
P(x)	**.1**	**.1**	**.4**	**.3**	**.1**

Answer:

x^2	1	4	9	16	25
P(x)	**.1**	**.1**	**.4**	**.3**	**.1**

$$E(x^2) = \sum_{i=1}^{5} x_i{}^2 p(x_i) = (1)(.1) + (4)(.1) + 9(.4) + 16(.3) + 25(.1) = 11.4$$

$$Var(x) = E(x^2) - [E(x)]^2 = 11.4 - 3.2^2 = 1.16$$

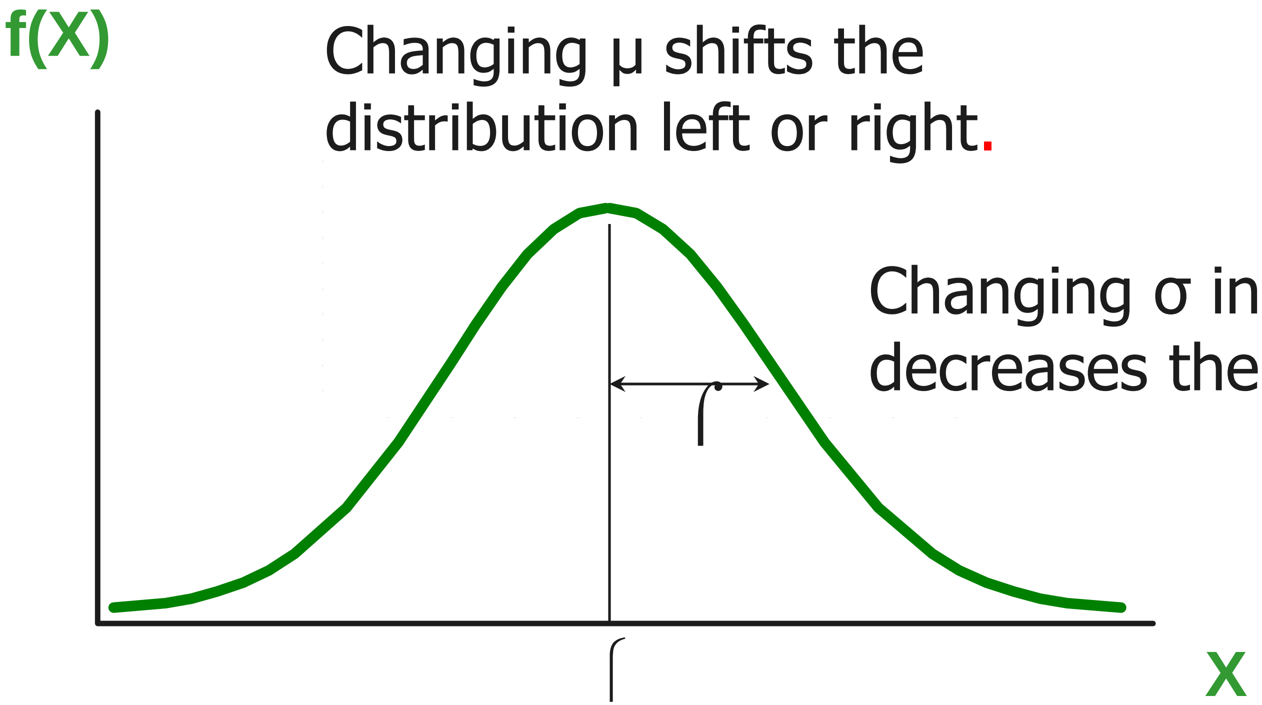
$$stddev(x) = \sqrt{1.16} = 1.08$$

Interpretation: On an average night, we expect Rohan to awaken 3 times, plus or minus 1.08. This gives you a feel for what would be considered an unusual night!

continuous probability(Gaussian) distributions:

The normal and standard normal

The Normal Distribution



Changing μ shifts the distribution left or right.

Changing σ increases or decreases the spread.

The Normal Distribution: as mathematical function (pdf)

$$f(x)=\frac{1}{\boxed{\sigma}\sqrt{2\pi}}\cdot e^{-\frac{1}{2}\left(\frac{x-\boxed{\mu}}{\sigma}\right)^{2}}$$

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 PDF

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

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

Normal distribution is defined by its mean and standard dev.

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

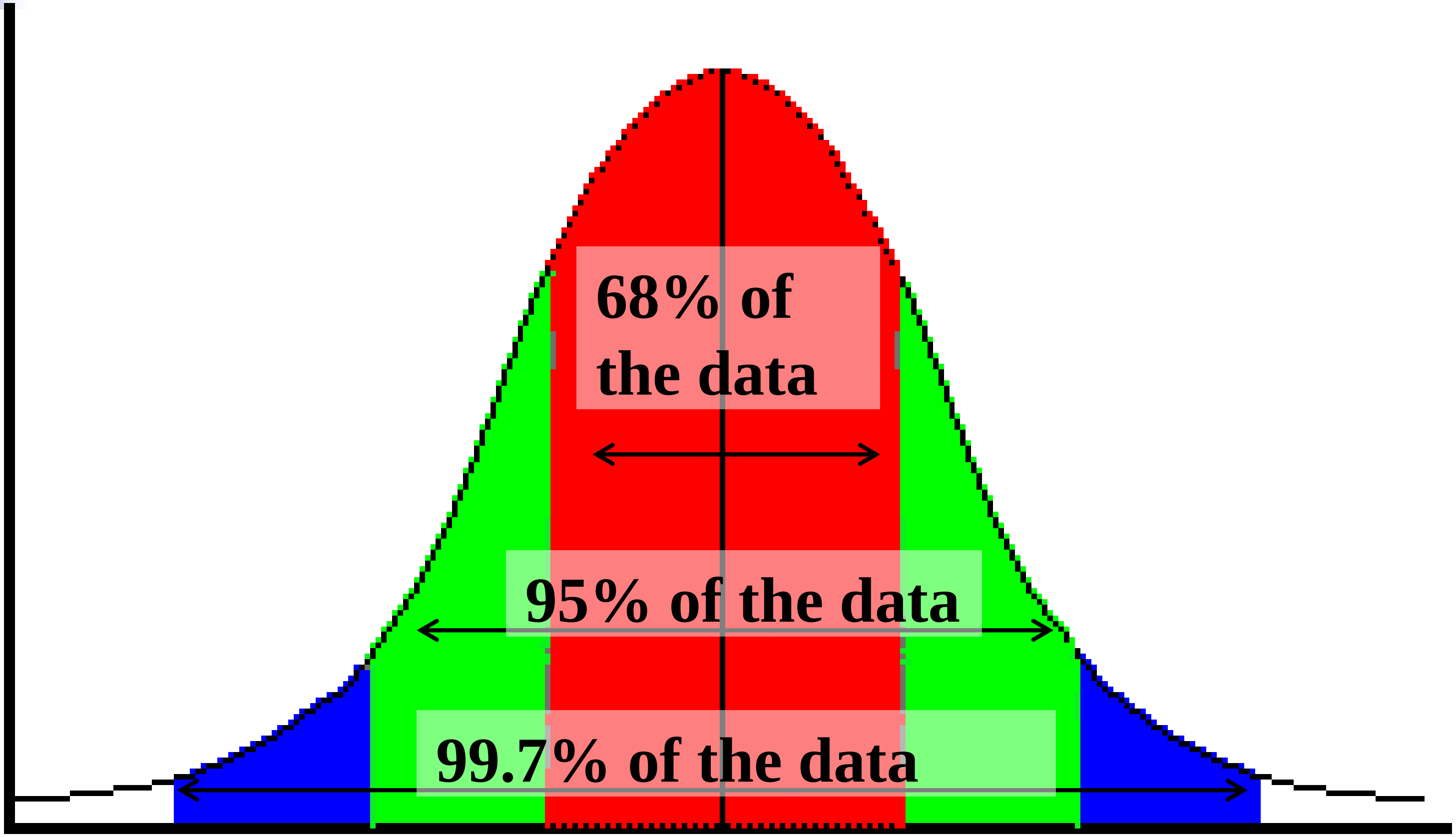
Var(X)=σ² = $\int_{-\infty}^{+\infty} x^2 \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx) - \mu^2$

Standard Deviation(X)=σ

**The beauty of the normal curve:

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

68-95-99.7 Rule



68-95-99.7 Rule in Math terms…

$$\int_{\mu-\sigma}^{\mu+\sigma} \frac{1}{\sigma\sqrt{2\pi}} \bullet e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx = .68$$

$$\int_{\mu-2\sigma}^{\mu+2\sigma} \frac{1}{\sigma\sqrt{2\pi}} \bullet e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx = .95$$

$$\int_{\mu-3\sigma}^{\mu+3\sigma} \frac{1}{\sigma\sqrt{2\pi}} \bullet e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx = .997$$

How good is rule for real data?

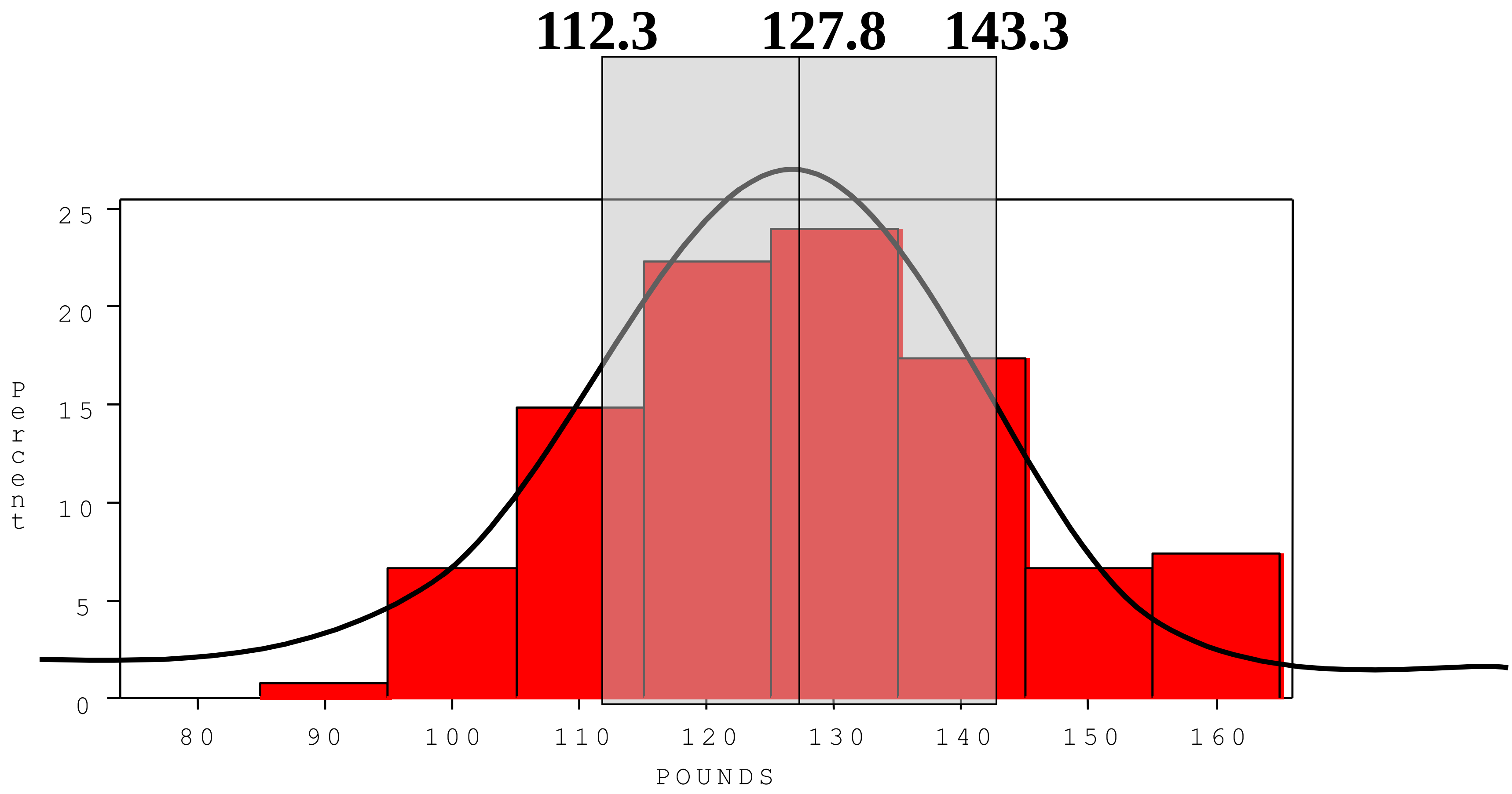
Check some example data:

The mean of the weight of the women = 127.8

The standard deviation (SD) = 15.5

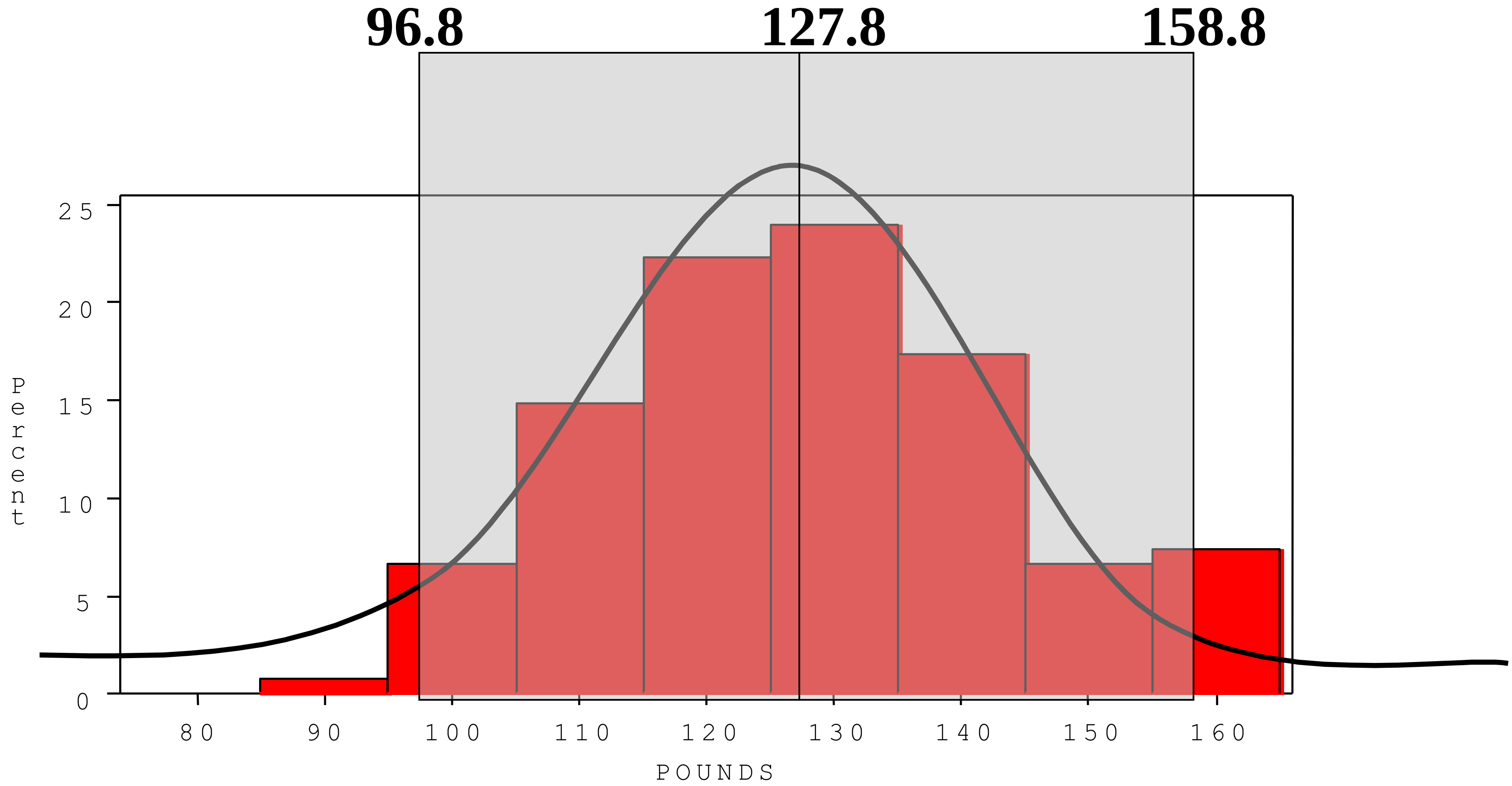
68% of 120 = .68x120 = ~ 82 runners

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



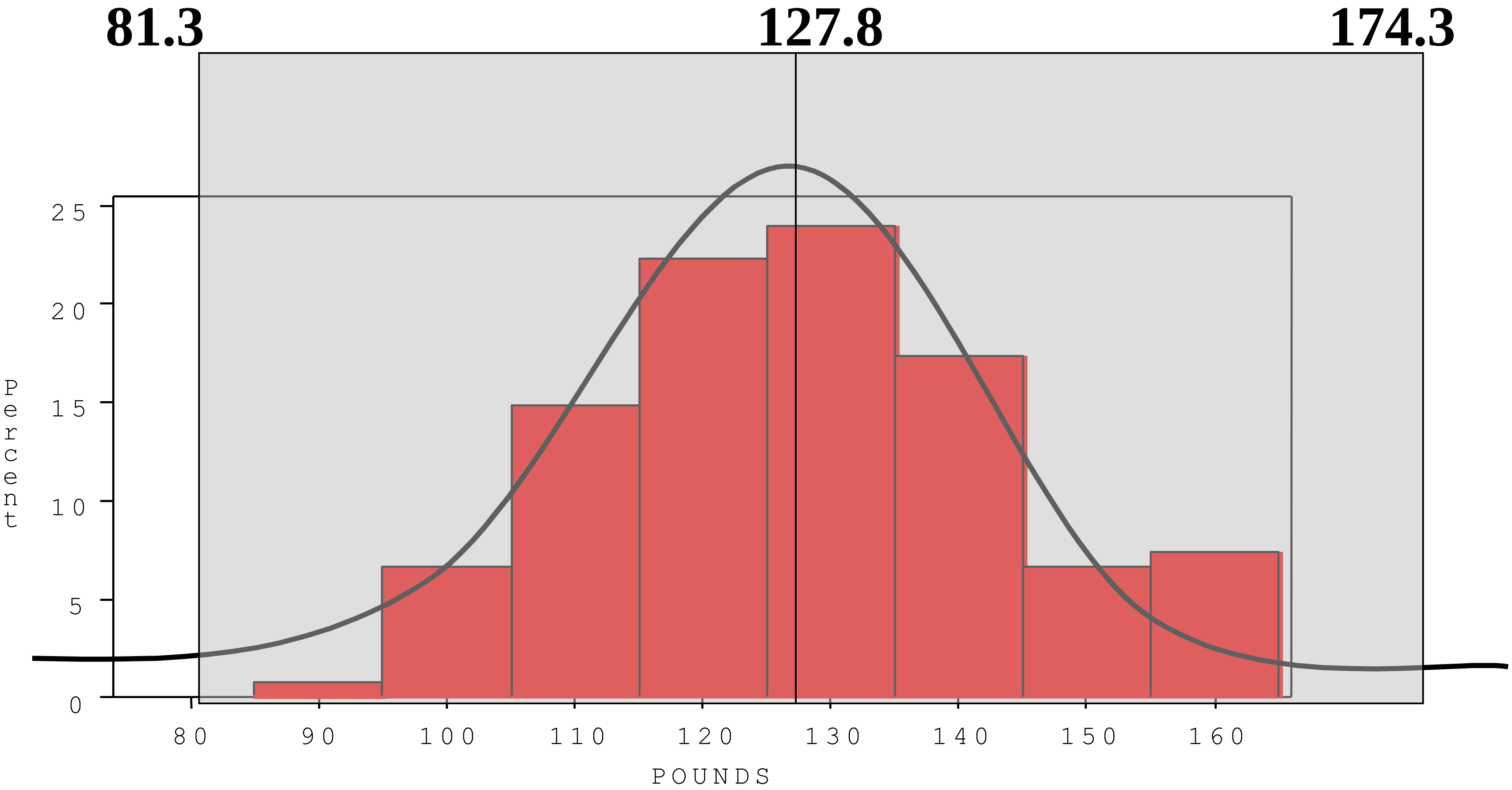
95% of 120 = .95 x 120 = ~ 114 runners

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



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

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



Example

- Suppose SAT scores roughly follows a normal distribution in the U.S. population of college-bound students (with range restricted to 200-800), and the average math SAT is 500 with a standard deviation of 50, then:
  - 68% of students will have scores between 450 and 550
  - 95% will be between 400 and 600
  - 99.7% will be between 350 and 650

Example

- BUT…
- What if you wanted to know the math SAT score corresponding to the 90th percentile (=90% of students are lower)?

P(X≤Q) = .90 →

$$\boxed{\int_{200}^{Q} \frac{1}{(50)\sqrt{2\pi}} \bullet e^{-\frac{1}{2}\left(\frac{x-500}{50}\right)^2} dx = .90}$$

The Standard Normal (Z): "Universal Currency"

The formula for the standardized normal probability density function is

$$p(Z)=\frac{1}{(1)\sqrt{2\pi}}\cdot e^{-\frac{1}{2}(\frac{Z-0}{1})^2}=\frac{1}{\sqrt{2\pi}}\cdot e^{-\frac{1}{2}(Z)^2}$$

The Standard Normal Distribution (Z)

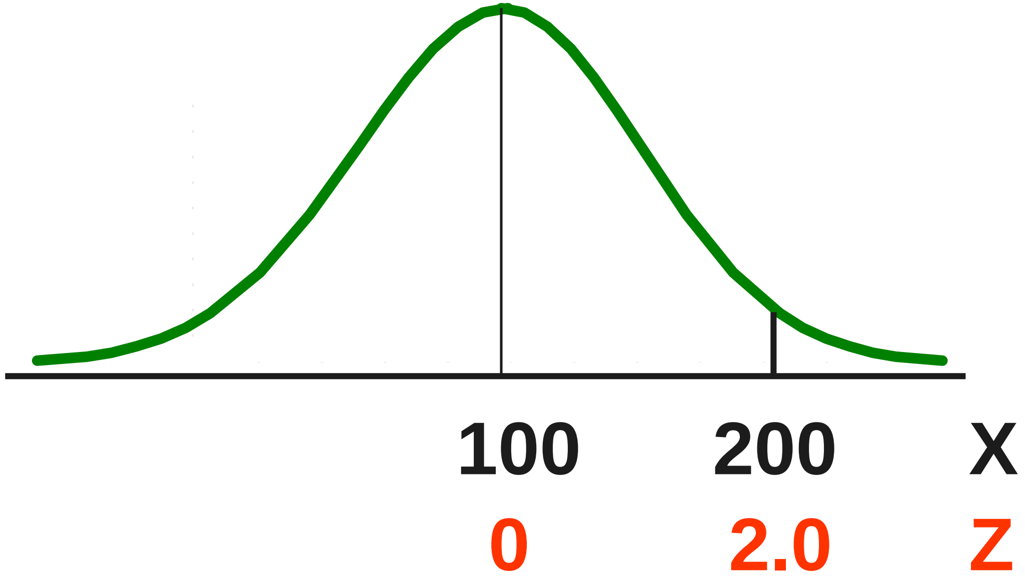
All normal distributions can be converted into the standard normal curve by subtracting the mean and dividing by the standard deviation:

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

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.

Comparing X and Z units



(μ = 100, σ = 50)

(μ = 0, σ = 1)

Example

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

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

- i.e., A score of 575 is 1.5 standard deviations above the mean

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

But to look up Z= 1.5 in standard normal chart (or enter into SAS)→ no problem! = .9332

Answer

a. What is the chance of obtaining a birth weight of 141 oz *or heavier* when sampling birth records at random?

$$\boxed{Z = \frac{141 - 109}{13} = 2.46}$$

From the chart or SAS → Z of 2.46 corresponds to a right tail (greater than) area of: P(Z≥2.46) = 1-(.9931)= .0069 or .69 %

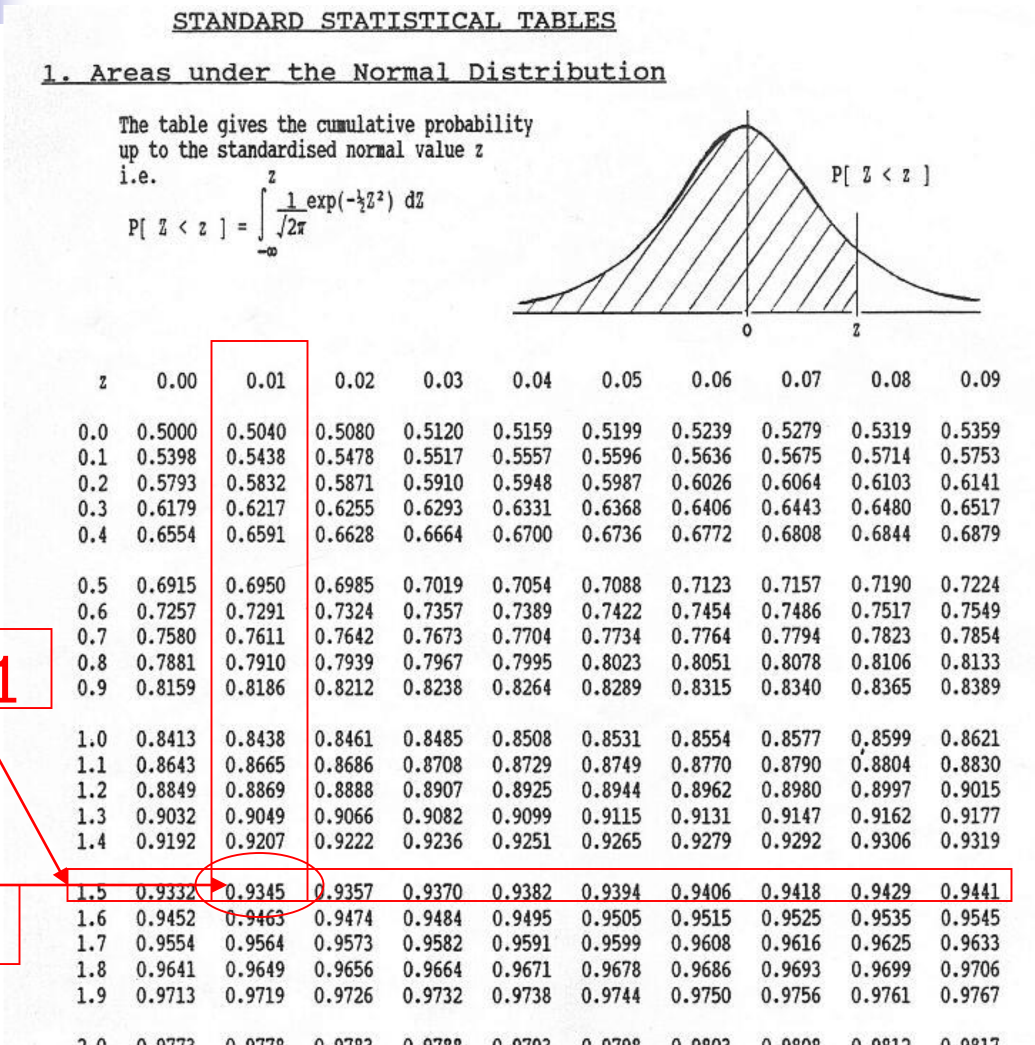
Answer

b. What is the chance of obtaining a birth weight of 120 *or lighter*?

$$\boxed{Z = \frac{120 - 109}{13} = .85}$$

From the chart or SAS → Z of .85 corresponds to a left tail area of:
P(Z≤.85) = .8023= 80.23%

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
2.0	0.9773	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817

What is the area to the left of Z=1.51 in a standard normal curve?

Z=1.51

Z=1.51

Area is 93.45%