Central Limit Theorem / Normal Approximation

- Applies to SUMS and to AVERAGES
- Suppose the observations X in some population have mean μ_X and standard deviation σ_X.
- Let S be the SUM of observations in a random sample; let $\overline{X}$ be the AVERAGE of the observations.
- The CLT tells us that the Z-score from either S or $\overline{X}$ will be approximately normal:

  $$\frac{S - n\mu_X}{\sqrt{n}\sigma_X} \sim Normal,\ Mean\ =\ 0, SD\ =\ 1.$$

  and

  $$\frac{\overline{X} - \mu_X}{\sigma_X/\sqrt{n}} \sim Normal,\ Mean\ =\ 0, SD\ =\ 1.$$

- **Amazing Fact:** This works, *no matter what the shape of the population distribution* (this is "swamping" again).

We will look at some examples of what calculations this lets us do (you already know <u>*how*</u> to do the calculations, using the 68–95–99.7 rule!).

From `http://palottery.com/`*: "The daily number"*

- Pick any three digit number
- Perfect match pays 500 to 1, so if you bet \$1.00 a perfect match pays \$500.
- Prob of an exact match on three digits:

$$\frac{1}{10} \cdot \frac{1}{10} \cdot \frac{1}{10} = 0.001, \text{ or 1 in 1000.}$$

- For a \$1.00 bet, the expected net gain is

$$\$499 \cdot 0.001 - \$1.00 \cdot 0.999 = -\$0.50$$

  and the SD is

$$\begin{aligned} &\sqrt{(\$499 - [-\$0.50])^2 \cdot 0.001 + (-\$1.00 - [-\$0.50])^2 \cdot 0.999} \\ &= \sqrt{249.75} \\ &= 15.80 \end{aligned}$$

Some people like to "cover their bets" by buying lots of tickets at once, figuring that they have a greater chance of winning on one of 100 tickets, say, than on just one.

Suppose I buy 100 randomly-chosen daily number tickets for \$1.00 each. Using the CLT, approximately what is the probability that I will break even, i.e. the sum of my 100 wins and losses will be greater than 0?

- Mean of the sum $= 100 \cdot (-\$0.50) = -50.00.$
- SD of the sum $= \sqrt{100} \cdot 15.80 = 150.8.$

$$
\begin{aligned}
P[SUM > 0] &= P\left[\frac{SUM - (-\$50.00)}{150.8} > \frac{0 - (-\$50.00)}{150.8}\right] \\
&= P[Z > 0.33]
\end{aligned}
$$

This can be calculated either

(a) *Exact:* By looking at the table on p. 262 of Siegel and Morgan, or on p. 519 of Moore, we see that $P[Z \leq 0.33] = 0.629$, so

$$P[Z > 0.33] = 1 - 0.629 = 0.371.$$

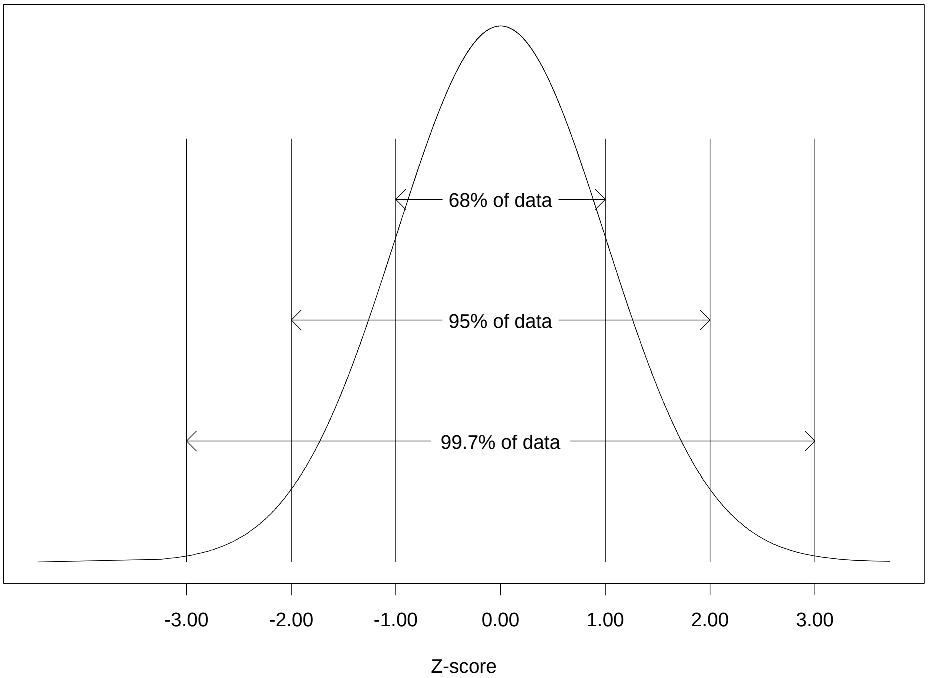
or…

(b) *Rough and ready.* We know from the 68–95–99.7 rule that $P[Z \leq 0] = 0.50$, and $P[Z \leq 1] = 0.68 + 0.16 = 0.84$, so by interpolation

$$P[Z \leq 0.33] \approx 0.50 + \frac{0.84 - 0.50}{1 - 0}(0.33 - 0.00) = 0.612$$

and so, approximately,

$$P[Z > 0.33] = 1 - 0.612 = 0.388.$$



The mean and the SD can help you decide whether a bet is "worth the risk". Some examples:

- Would you rather have

  (a) \$1 million, for sure; or

  (b) \$2 million, if a fair coin comes up heads, and otherwise nothing?

  Most people would choose (a) [SD=0] even though (b) [SD=\$1 million] also has expected value \$1 million, since there is less "risk" (less variability, smaller SD) in the outcome.

- If two bets have the same mean net gain but different risks (SD), we might prefer

  - The bet with the lower risk, if our mean net gain is positive, or
  - The bet with the higher risk, if our mean net gain is negative.

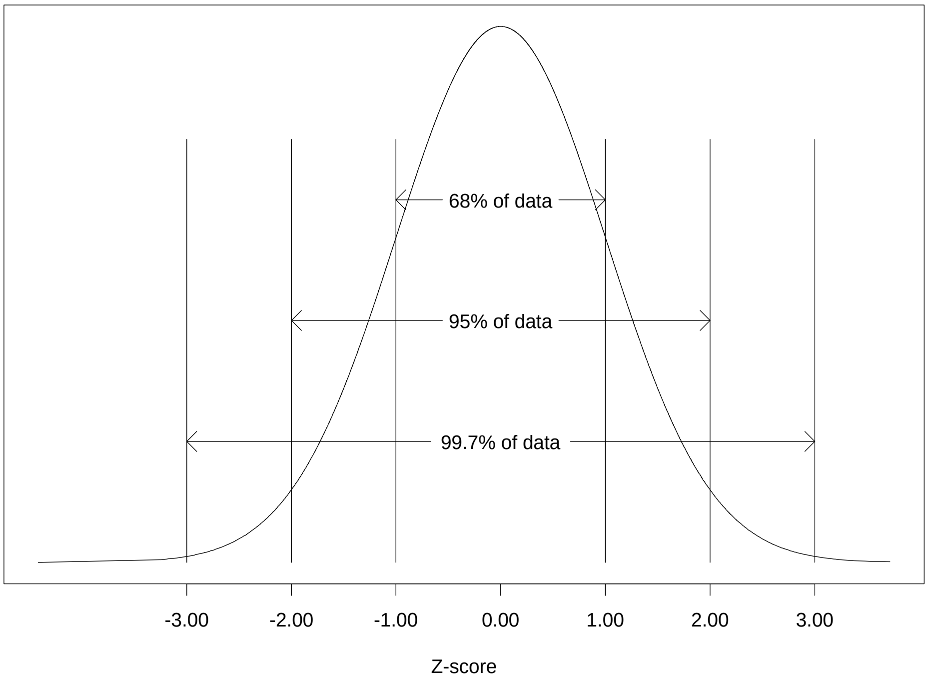
- If two bets have the same risk (SD) but different mean net gains, we might prefer the bet with the higher mean.

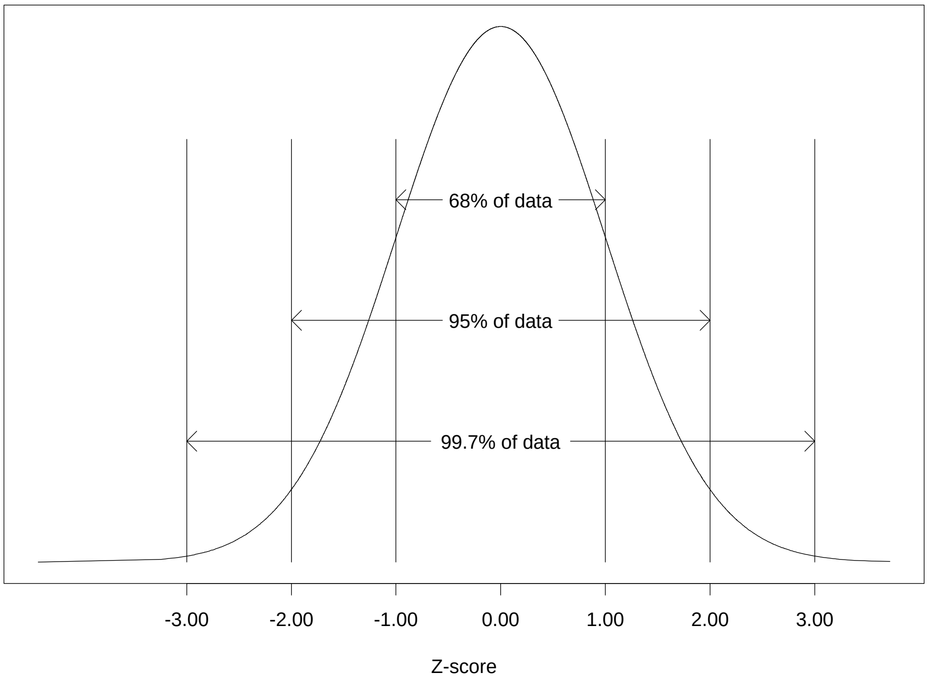
A Survey Example

The Survey of Study Habits and Attitudes (SSHA) is a psychological test that measures the attitude toward school and study habits of studfents. Scores range from 0 to 200. The mean score for U.S. students of college age is 115 with a SD of 30, and SSHA scores are approximately normally distributed.

a. What is the probability that a single randomly chosen college student would score 106 or lower?

b. What is the probability that the average of 100 randomly chosen college students would be 106 or lower?


68% of data
95% of data
99.7% of data
-3.00
-2.00
-1.00
0.00
1.00
2.00
3.00
Z-score

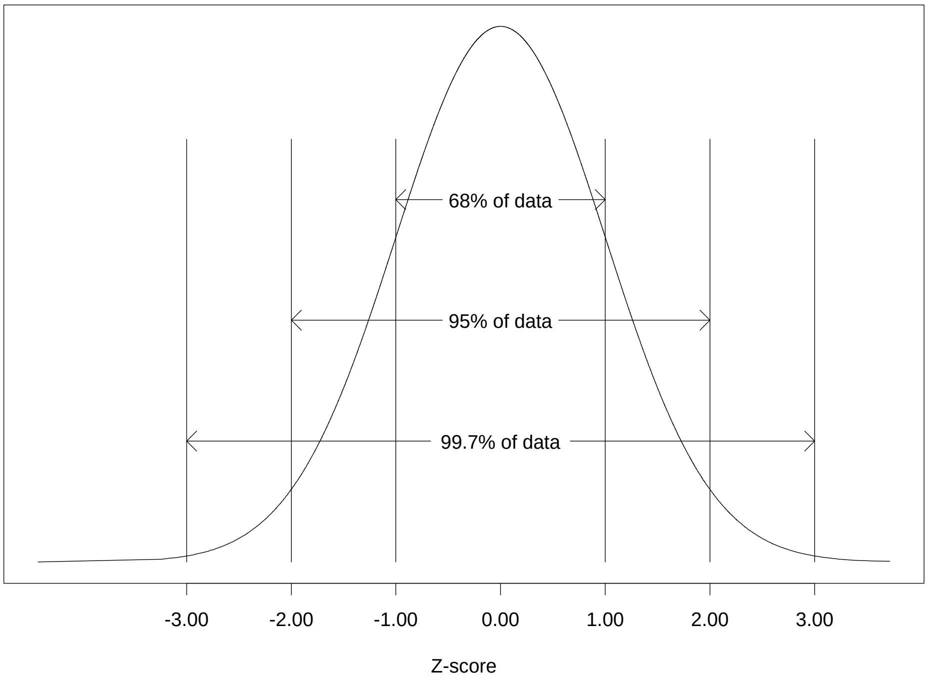

68% of data
95% of data
99.7% of data
-3.00
-2.00
-1.00
0.00
1.00
2.00
3.00
Z-score

c. How do the answers to (a) and (b) change, if the SSHA distribution is strongly skewed to the right?

d. An educator claims that older student have better attitudes toward school. She gives the SSHA to a randomly selected group of 36 students who are at least 30 years old, and finds an average score of 136.

   - How surprising would it be to get an average of 136 or higher from a random sample of 36 college-age students?

   - What does this suggest about the educator's claim?


68% of data
95% of data
99.7% of data
-3.00
-2.00
-1.00
0.00
1.00
2.00
3.00
Z-score

Square root law

We've seen this before...

- For an average $\overline{X}$ from a random sample, we know that the Z-score

  $$\frac{\overline{X} - \mu_X}{\sigma_X / \sqrt{n}}$$

  is approximately Normally distriuted, mean 0 and variance 1.

- So the "typical" values for $\overline{X}$ are all like

  $$\overline{X} \approx \mu_X \pm \sigma_X / \sqrt{n}$$

  note that there is less variability in $\overline{X}$ as n grows; this is the *square root law*.

- The square root law just says that $\overline{X}$ becomes more and more precise as a measure of μ_X, as sample size grows.

Pennsylvania Lottery "Lotto"

- More complicated than I thought!
  - Basically I forgot to include the fact that prize categories typically have multiple winners, who split the "pot" for that prize category.
  - Figuring out how many multiple winners there are per prize category is a bit like the "birthday problem."
- *Maybe I'll try a new analysis over spring break…*