Announcements

- Lab 6 due tonight on Gradescope
- New homework will be posted today to be due next Monday
- Our next in-class lab will not be until next week
- Test 1 thoughts:
  - Will probably have results back to you on Friday
  - I think it was a bit long
  - Had one, maybe two questions with some wording accidentally open to alternative interpretations
  - Will be taking both into account when computing scores
- Finish looking over 2.6 and and read 2.7 for Wednesday
- Polling: `rembold-class.ddns.net`

Review Question!

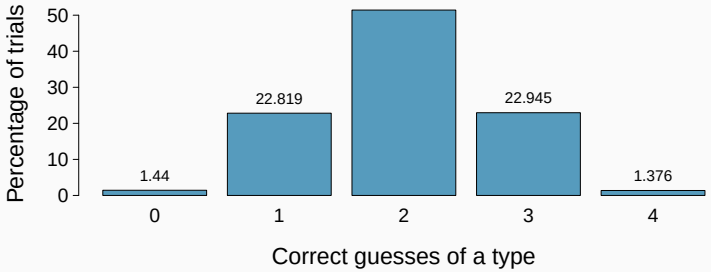
At the start of the semester we discussed the amazing tea-tasting lady who could differentiate between tea prepared milk first vs tea prepared tea first. In the trial to test her skills, of the 4 cups prepared tea first and the 4 cups prepared milk first, she had to choose 4 of any one type. The chart to the right shows the probability distribution assuming the null hypothesis over 100000 trials. The numbers across the bottom are the number of cups she correctly chose to match a single type in that trial. Given this information, what are the odds that she would have randomly succeeded at correctly guessing all four cups?

A) 1.44%

B) 1.38%

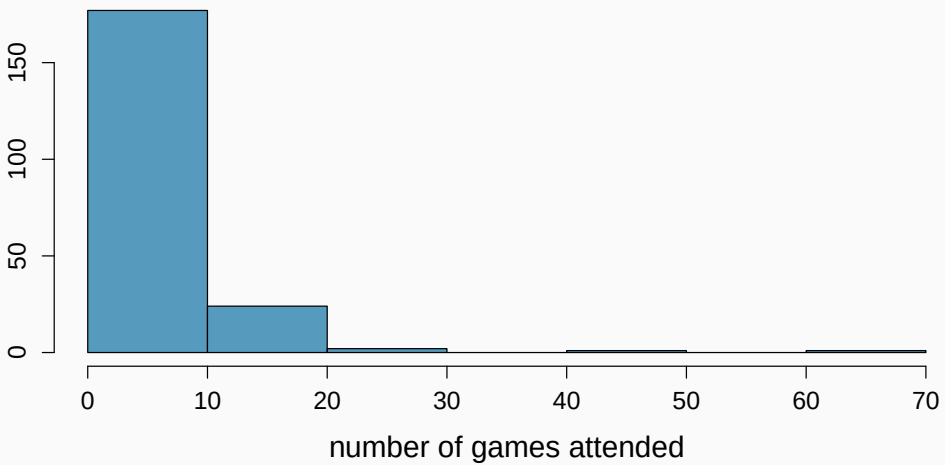
C) 2.82%

D) 0.06%



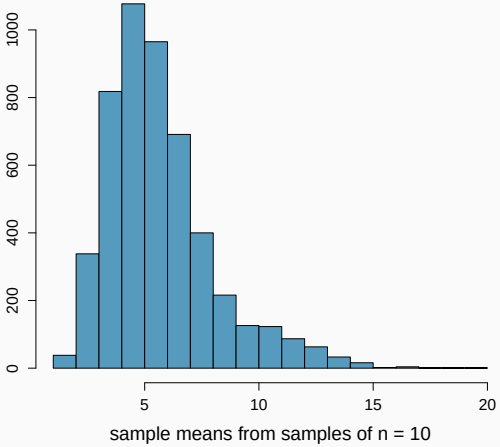
Average number of basketball games attended

Let's look at the population data (not sampled) for the number of basketball games attended by Duke students:



Average basketball games attended (sampled)

Sampling distribution: n=10

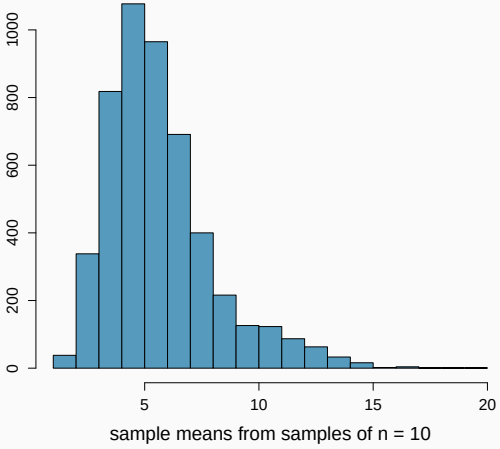


What does each observation in this distribution represent?

Is the variability in the sampling distribution smaller or larger than the variability in the population distribution?

Average basketball games attended (sampled)

Sampling distribution: n=10



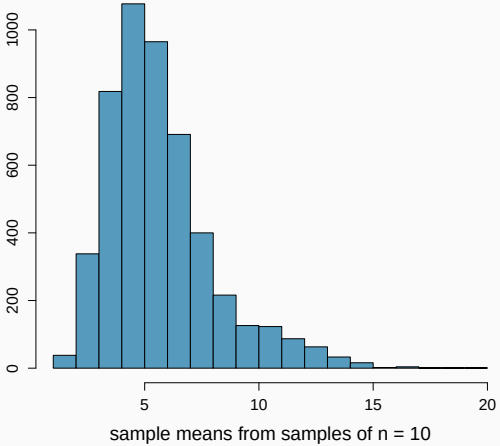
What does each observation in this distribution represent?

- The sample mean ($\bar{x}$) of samples of size $n = 10$

Is the variability in the sampling distribution smaller or larger than the variability in the population distribution?

Average basketball games attended (sampled)

Sampling distribution: n=10



What does each observation in this distribution represent?

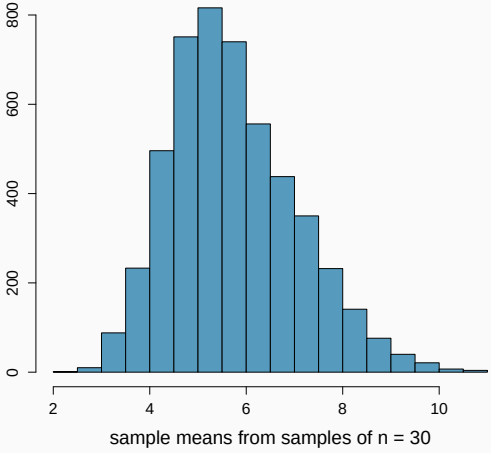
- The sample mean ($\bar{x}$) of samples of size $n = 10$

Is the variability in the sampling distribution smaller or larger than the variability in the population distribution?

- Smaller, as sample means will vary less than individual observations

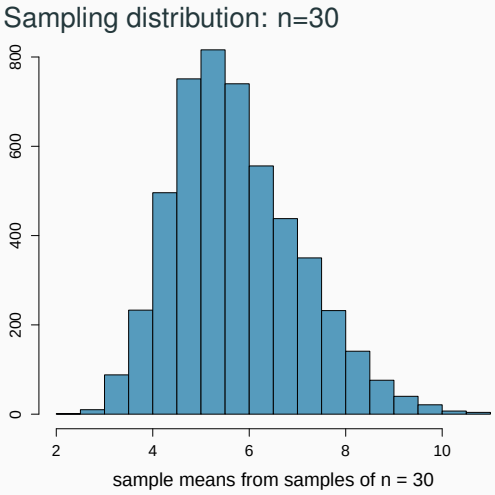
Average basketball games attended (sampled more)

Sampling distribution: n=30



How did the shape, center, and spread of the sampling distribution change going from $n = 10$ to $n = 30$?

Average basketball games attended (sampled more)

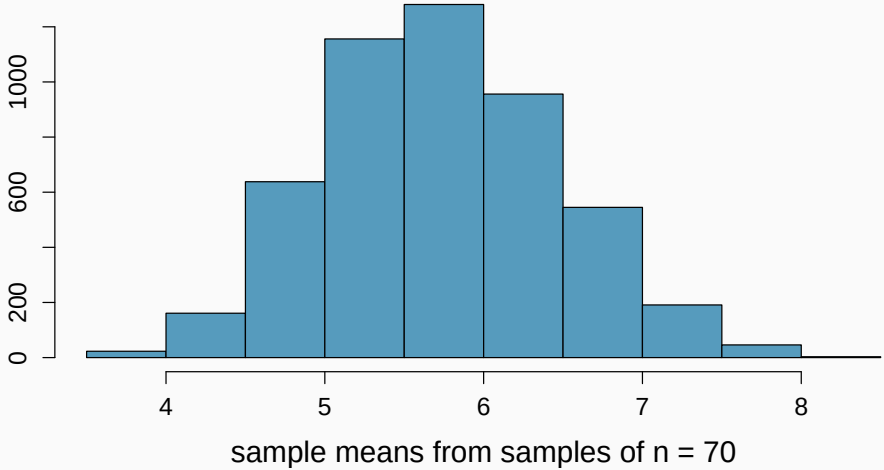


How did the shape, center, and spread of the sampling distribution change going from $n = 10$ to $n = 30$?

- Shape is more symmetric, center is about the same, spread is smaller

Average basketball games attended (sampled more[2])

Sampling distribution: n=30



Central Limit Theorem

Central Limit Theorem

Sample distributions of proportions (of differences in proportions) will appear to follow a symmetric, bell-shaped curved called the ***normal distribution*** provided that:

1. Observations in the sample are independent
2. The sample is large enough

- Gives us new methods to understand t-tests and the probability of something happening
- Will look more into what is “large enough” in the future

Understanding Check

Match each distribution to the corresponding plot:

(1) Single random sample of 100 observations of population

(2) Distribution of 100 samples means from random samples of size 7

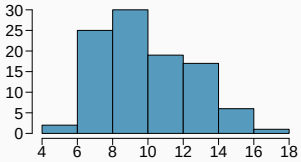
(3) Distribution of 100 sample means from random samples of size 49

A) A-(3); B-(2); C-(1)

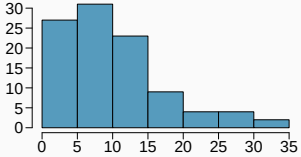
B) A-(2); B-(3); C-(1)

C) A-(2); B-(1); C-(2)

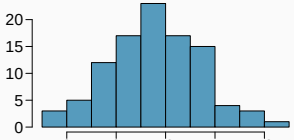
D) A-(1); B-(2); C-(3)



Plot A



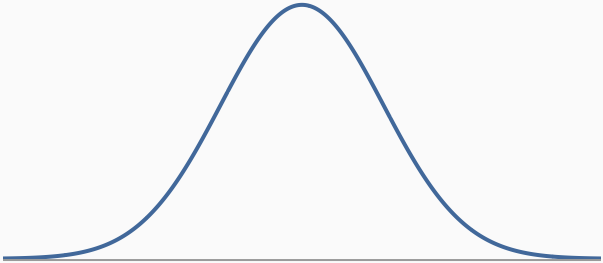
Plot B



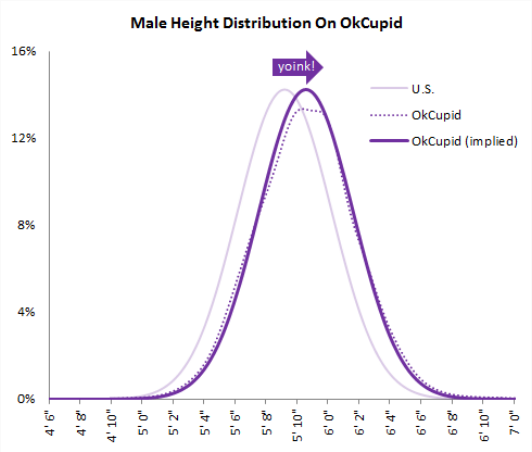
Plot C

The Normal Distribution

- Unimodal and symmetric, bell shaped curve
- Many variables are nearly normal in their distribution, but none exactly normal
  - This makes it a very flexible tool for us to apply in a wide variety of instances
- Denoted as $N(\mu, \sigma)$
  - Gives a normal distribution with a mean of μ and standard deviation of σ

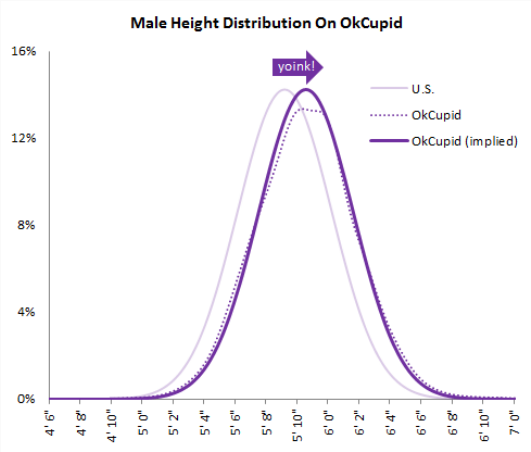


Statistics on fudging statistics (in the name of love)



"The male heights on OkCupid very nearly follow the expected normal distribution – except the whole thing is shifted to the right of where it should be. Almost universally guys like to add a couple inches."

Statistics on fudging statistics (in the name of love)

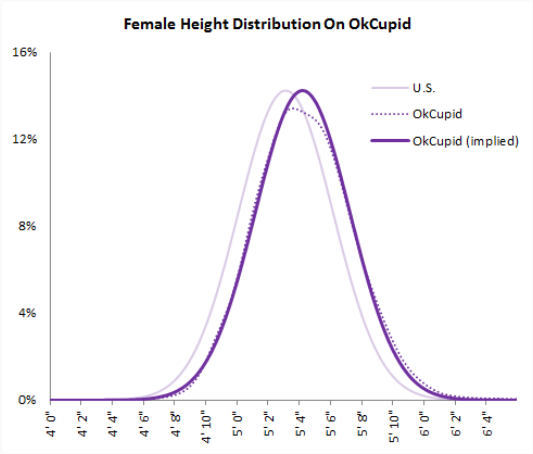


"The male heights on OkCupid very nearly follow the expected normal distribution – except the whole thing is shifted to the right of where it should be. Almost universally guys like to add a couple inches."

"You can also see a more subtle vanity at work: starting at roughly 5' 8", the top of the dotted curve tilts even further rightward. This means that guys as they get closer to six feet round up a bit more than usual, stretching for that coveted psychological benchmark."

Source: here

Women too!

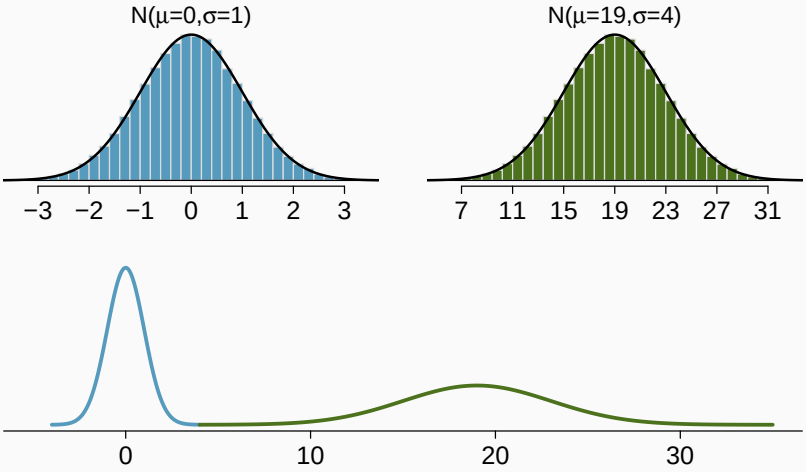


"When we looked into the data for women, we were surprised to see height exaggeration was just as widespread, though without the lurch towards a benchmark height."

Source: here

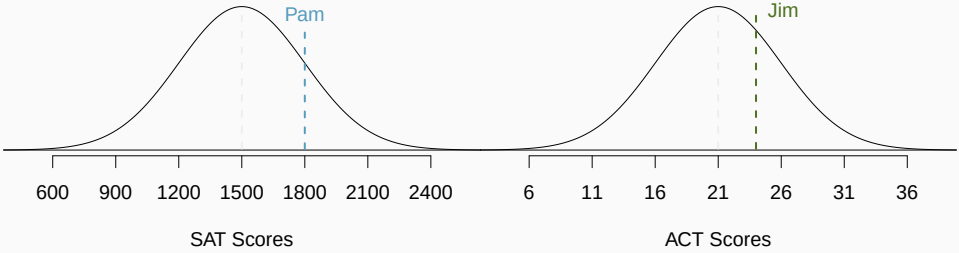
Parameters of Normal Distributions

Recall that μ = mean and σ = standard deviation



Making Comparisons

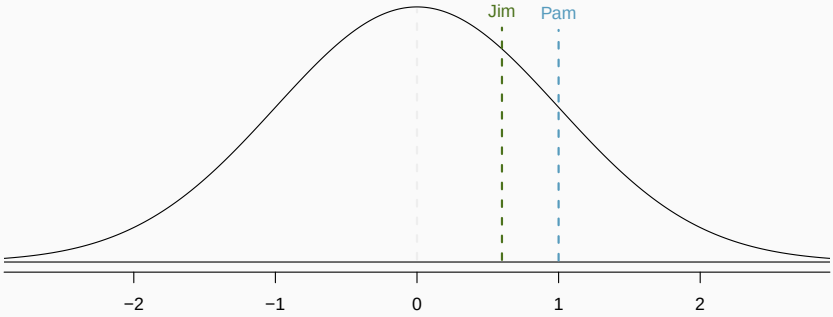
SAT scores are distributed nearly normally with mean 1500 and standard deviation 300. ACT scores are distributed nearly normally with mean 21 and standard deviation 5. A college admissions officer wants to determine which of the two applicants scored better on their standardized test with respect to the other test takers: Pam, who earned an 1800 on her SAT, or Jim, who scored a 24 on his ACT?



Enter the Z-Score

The two scores are not directly comparable, but we can instead compare how many standard deviations beyond the mean each observation is.

- Pam's score is $\frac{1800-1500}{300} = 1$ standard deviation above the mean.
- Jim's score is $\frac{24-21}{5} = 0.6$ standard deviations above the mean.



Standardizing with Z scores

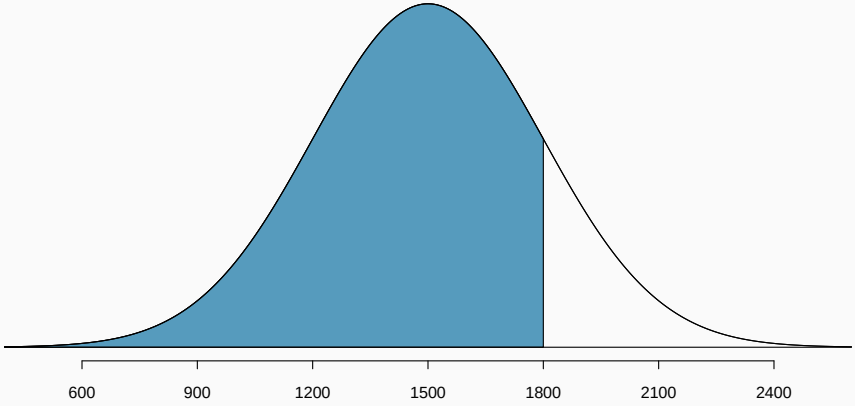
- These are called ***standardized*** scores, or ***Z scores***.
- The Z score of an observation is the number of standard deviations it falls above or below the mean.

$$Z = \frac{\text{observation} - \text{mean}}{\text{standard deviation}} = \frac{x - \mu}{\sigma}$$

- Z scores are defined for distributions of any shape, but we can only use Z scores to calculate percentiles when the distribution is normal.
- Observations more than 2 standard deviations from the mean ($|Z| > 2$) are usually considered unusual.

In the top percentile

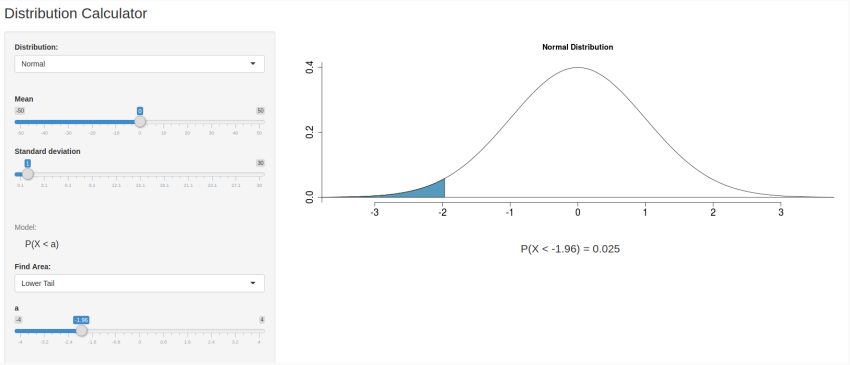
- ***Percentile*** is the percentage of observations that fall below a given data point.
- Graphically, percentile is the area below the probability distribution curve to the left of the observation.



Calculating Percentiles

There are a wide variety of ways to determine percentiles or the area under the curve:

- R: `> pnorm(1800, mean = 1500, sd = 300)`
- Applet: https://gallery.shinyapps.io/dist_calc/



Calculating percentiles - using tables

	Second decimal place of Z									
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	0.5160	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	0.7852
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	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015

Quality Control

At Heinz ketchup factory the amounts which go into bottles of ketchup are supposed to be normally distributed with mean 36 oz. and standard deviation 0.11 oz. Once every 30 minutes a bottle is selected from the production line, and its contents are noted precisely. If the amount of ketchup in the bottle is below 35.8 oz. or above 36.2 oz., then the bottle fails the quality control inspection. What percent of bottles have less than 35.8 ounces of ketchup?

Quality Control

At Heinz ketchup factory the amounts which go into bottles of ketchup are supposed to be normally distributed with mean 36 oz. and standard deviation 0.11 oz. Once every 30 minutes a bottle is selected from the production line, and its contents are noted precisely. If the amount of ketchup in the bottle is below 35.8 oz. or above 36.2 oz., then the bottle fails the quality control inspection. What percent of bottles have less than 35.8 ounces of ketchup?

- Draw a picture

Quality Control

At Heinz ketchup factory the amounts which go into bottles of ketchup are supposed to be normally distributed with mean 36 oz. and standard deviation 0.11 oz. Once every 30 minutes a bottle is selected from the production line, and its contents are noted precisely. If the amount of ketchup in the bottle is below 35.8 oz. or above 36.2 oz., then the bottle fails the quality control inspection. What percent of bottles have less than 35.8 ounces of ketchup?

- Draw a picture
- Compute Z score

Quality Control

At Heinz ketchup factory the amounts which go into bottles of ketchup are supposed to be normally distributed with mean 36 oz. and standard deviation 0.11 oz. Once every 30 minutes a bottle is selected from the production line, and its contents are noted precisely. If the amount of ketchup in the bottle is below 35.8 oz. or above 36.2 oz., then the bottle fails the quality control inspection. What percent of bottles have less than 35.8 ounces of ketchup?

- Draw a picture
- Compute Z score
- Compute percentile