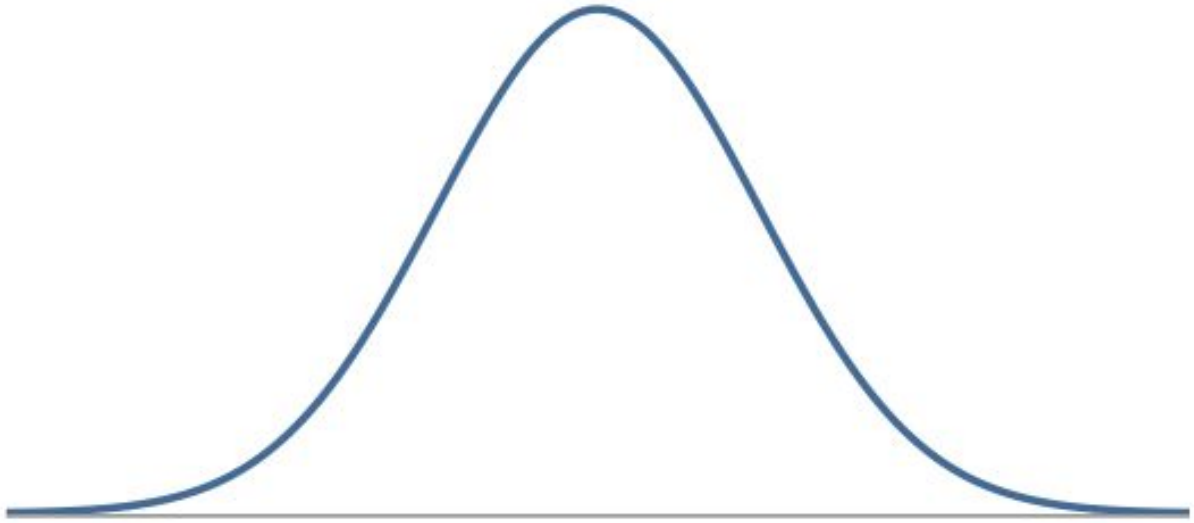
Section 4.1
Normal distribution

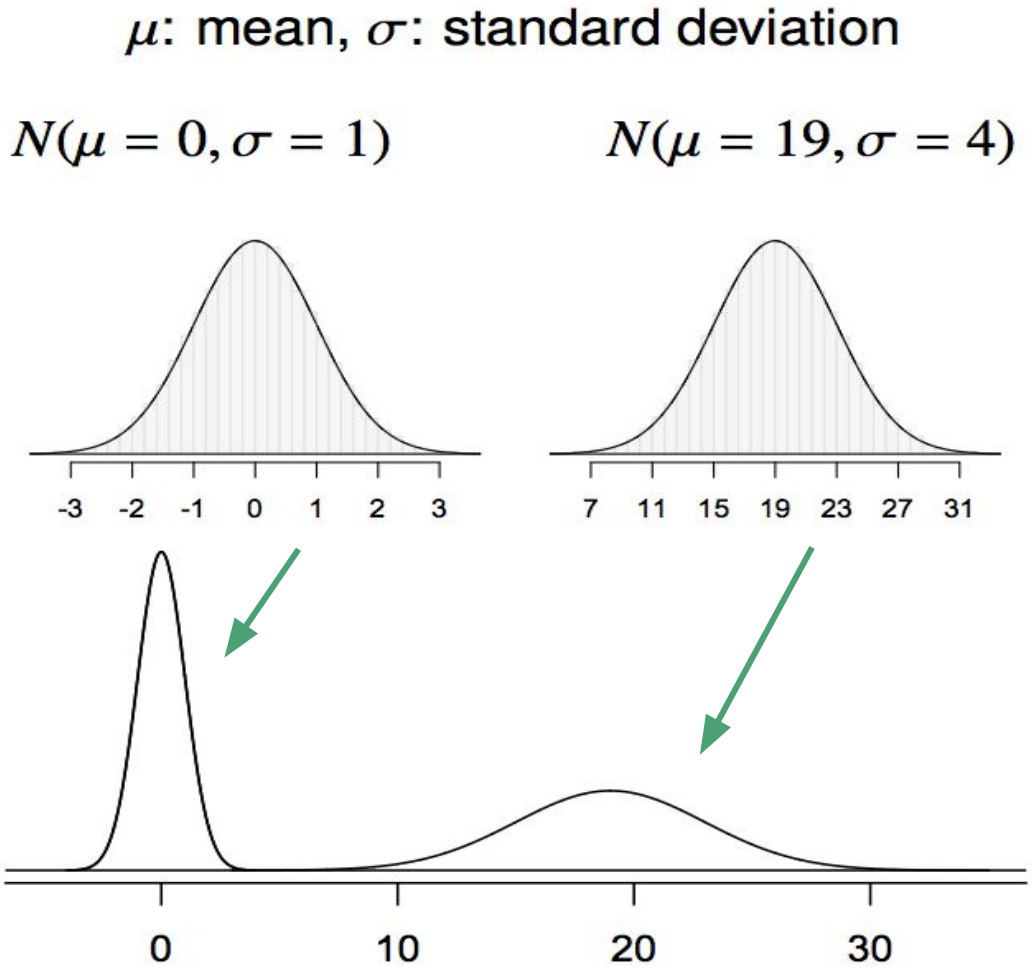
Stats 7 Summer Session II 2022

Normal Distribution

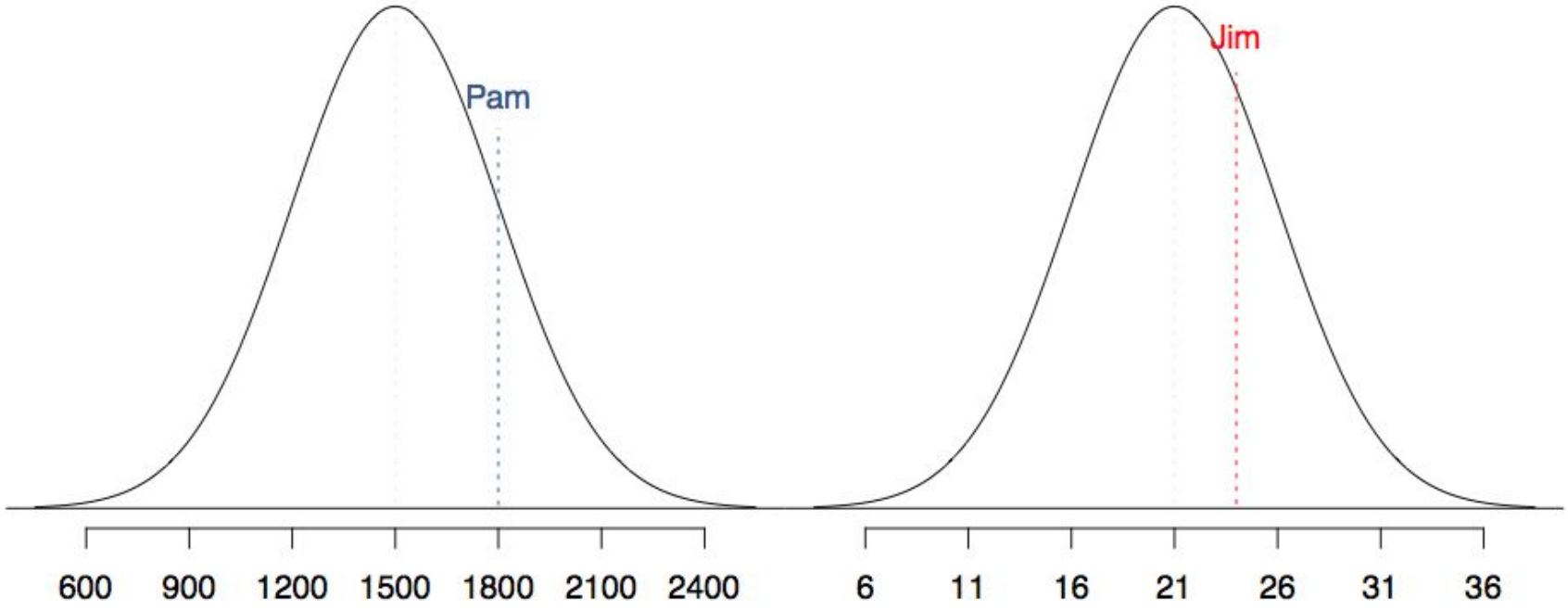
- Unimodal and symmetric, bell shaped curve
- Many variables are nearly normal, but none are exactly normal
- Denoted as $N(\mu, \sigma)$ ➔ Normal with mean μ and standard deviation σ



Normal distributions with different parameters



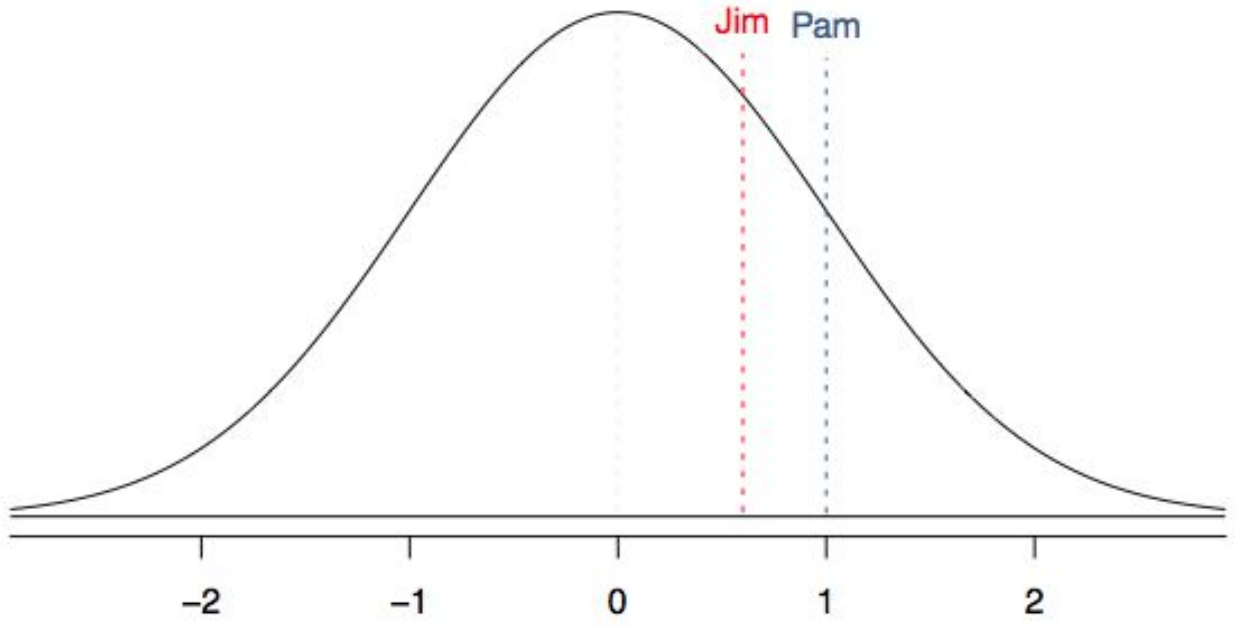
SAT scores are distributed nearly normally with mean 1500 and standard deviation 300. ACT scores are distributed nearly normally with mean 21 and standard deviation 5. A college admissions officer wants to determine which of the two applicants scored better on their standardized test with respect to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



Standardizing with Z scores

Since we cannot just compare these two raw scores, we instead compare how many standard deviations beyond the mean each observation is.

- Pam's score is (1800 - 1500) / 300 = 1 standard deviation above the mean.
- Jim's score is (24 - 21) / 5 = 0.6 standard deviations above the mean.



Standardizing with Z scores (cont.)

These are called *standardized* scores, or *Z scores*.

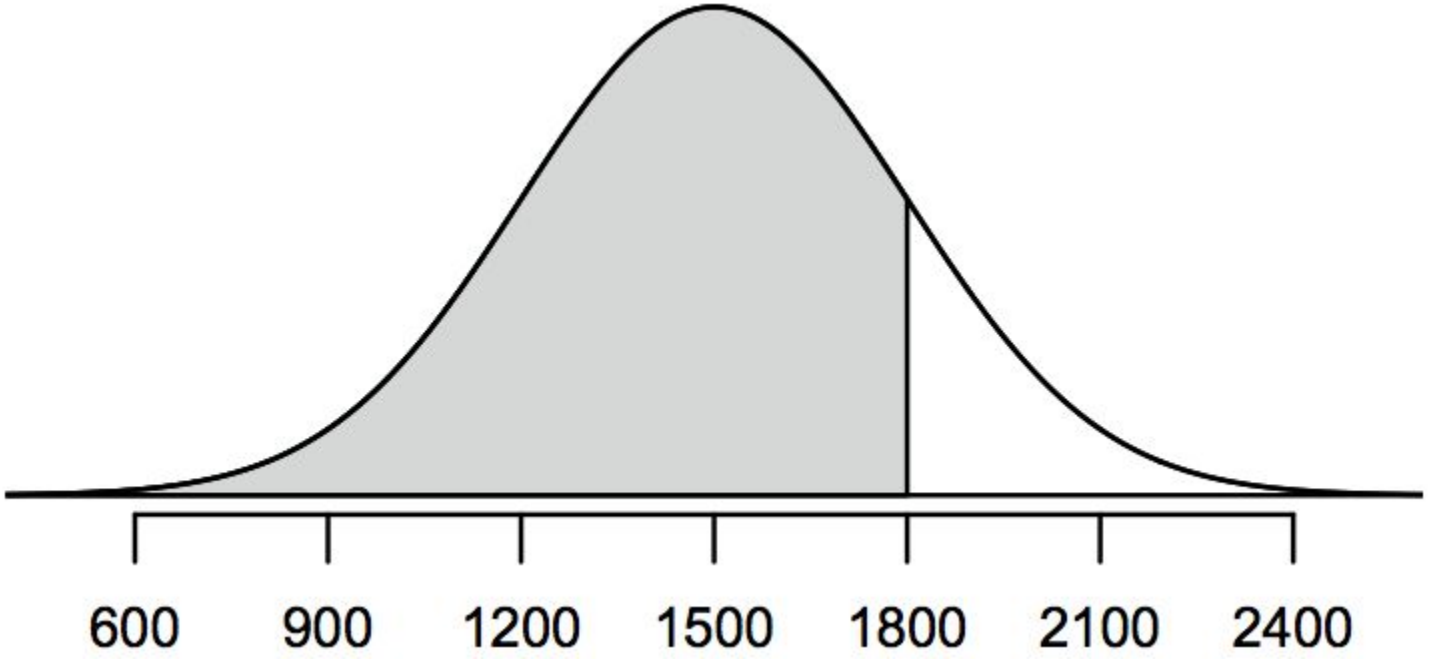
- Z score of an observation is the number of standard deviations it falls above or below the mean.

$$Z = \frac{observation - mean}{SD} \quad \text{aka} \quad Z = \frac{x - \mu}{\sigma}$$

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

Percentiles

- *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 that observation.
- Recall we call the 50th percentile the median



Calculating percentiles - using computation

To compute an area/percentile in R you can standardize before hand

$$Z = \frac{observation - mean}{SD}$$

```
> pnorm( (1800 - 1500) / 300, mean = 0, sd = 1)
[1] 0.8413447
```

Or you can compute directly

```
> pnorm(1800, mean = 1500, sd = 300)
[1] 0.8413447
```

They are equivalent

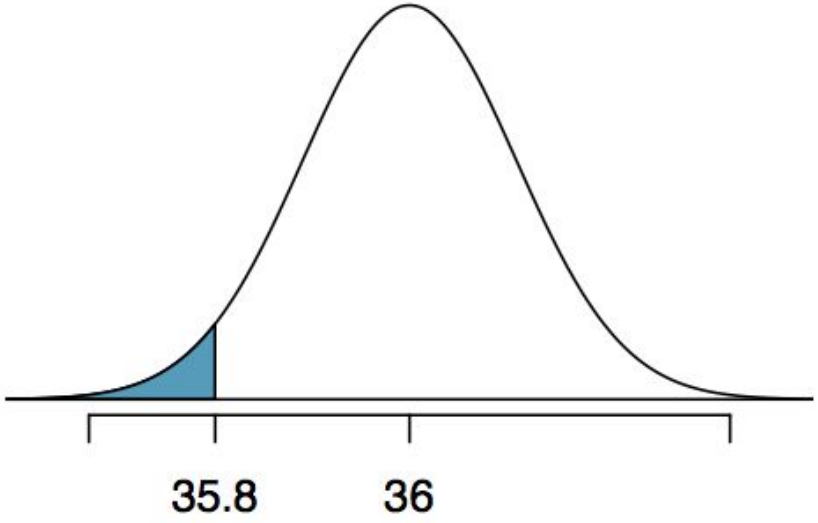
This means the probability of someone scoring lower than 1800 on the SAT was 0.84

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?

- *Let X = amount of ketchup in a bottle: $X \sim N(\mu = 36, \sigma = 0.11)$*

This is how we write that we are modeling with a normal distribution



$$Z = \frac{35.8 - 36}{0.11} = -1.82$$

Finding the exact probability - using R

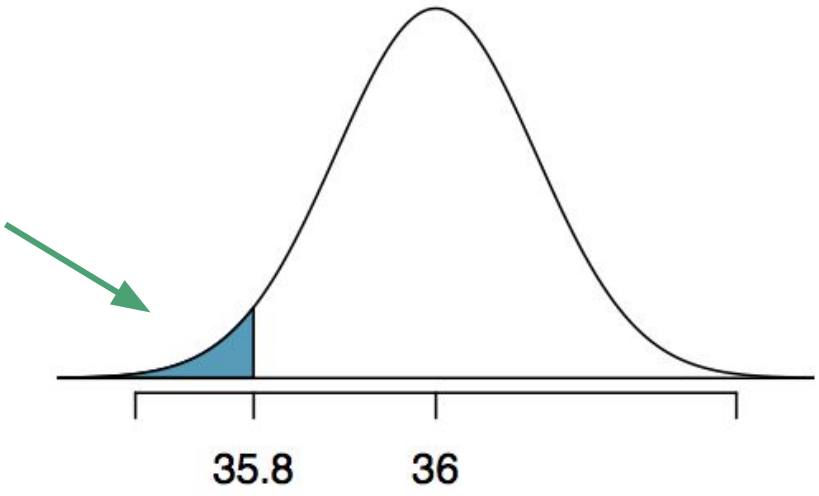
The code below calculates the probability of the Z-score

```
> pnorm( (35.8 - 36) / 0.11, mean = 0, sd = 1)
[1] 0.03451817
```

Equivalently, we could have directly computed without first standardizing

```
> pnorm(35.8, mean = 36, sd = 0.11)
[1] 0.03451817
```

This means we estimate there to be 0.0345 probability of a can being below 35.8 oz



Practice

We just found the proportion of bottles that have less than 35.8 ounces of ketchup to be 0.0345 (3.45%).

What percent of bottles have more than 35.8 ounces of ketchup?

(a) 1.82%

(b) 3.44%

(c) 6.88%

(d) 93.10%

(e) 96.55%

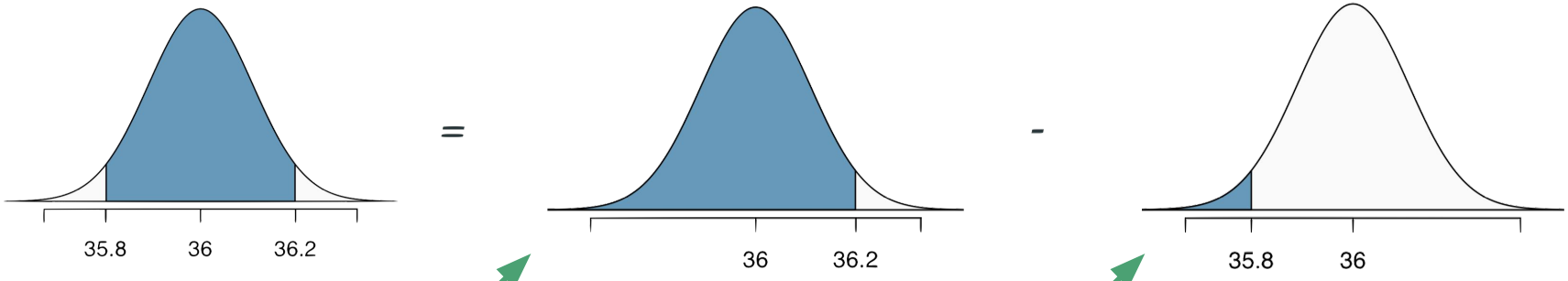
Practice

What percent of bottles pass the quality control inspection (i.e. are between 35.8 oz. or above 36.2 oz)?

(a) 1.82% (c) 6.88% (e) 96.55%

(b) 3.44% *(d) 93.10%*

pnorm() always computes the area to the left so we can compute the blue shaded area by subtracting the tail area from the overall area




```
> pnorm(36.2, mean = 36, sd = 0.11)
[1] 0.9654818
```

```
> pnorm(35.8, mean = 36, sd = 0.11)
[1] 0.03451817
```

```
> pnorm(36.2, mean = 36, sd = 0.11) - pnorm(35.8, mean = 36, sd = 0.11)
[1] 0.9309637
```

pnorm() versus qnorm()

P(Z < (observation - mean) / sd) = probability

- We know pnorm() will give us the probability of getting a value less than that observed
  - So we give it an observation and get a probability

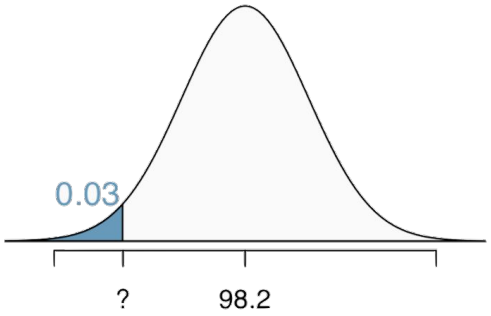
```
> pnorm(96.82702, mean = 98.2, sd = 0.73)
[1] 0.02999994
```

- We also have qnorm() to give us the value of the observation given a probability
  - So we give it a probability to get an observation

```
> qnorm(0.03, mean = 98.2, sd = 0.73)
[1] 96.82702
```

Finding cutoff points

Body temperatures of healthy humans are distributed nearly normally with mean 98.2°F and standard deviation 0.73°F. What is the cutoff for the lowest 3% of human body temperatures?



What observation has 0.03 probability of a value being less than?

```
> qnorm(0.03, mean = 98.2, sd = 0.73)
[1] 96.82702
```

The lowest 3% of body temperatures are below 96.83°F

Mackowiak, Wasserman, and Levine (1992), A Critical Appraisal of 98.6 Degrees F, the Upper Limit of the Normal Body Temperature, and Other Legacies of Carl Reinhold August Wunderlick.

Practice

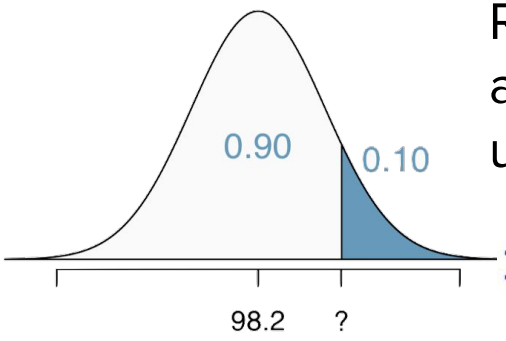
Body temperatures of healthy humans are distributed nearly normally with mean 98.2°F and standard deviation 0.73°F. What is the cutoff for the highest 10% of human body temperatures?

(a) 97.3°F

(c) 99.4°F

(b) 99.1°F

(d) 99.6°F



Remember that by default we are always computing the area to the left. We have to subtract from 1 to get the upper area.

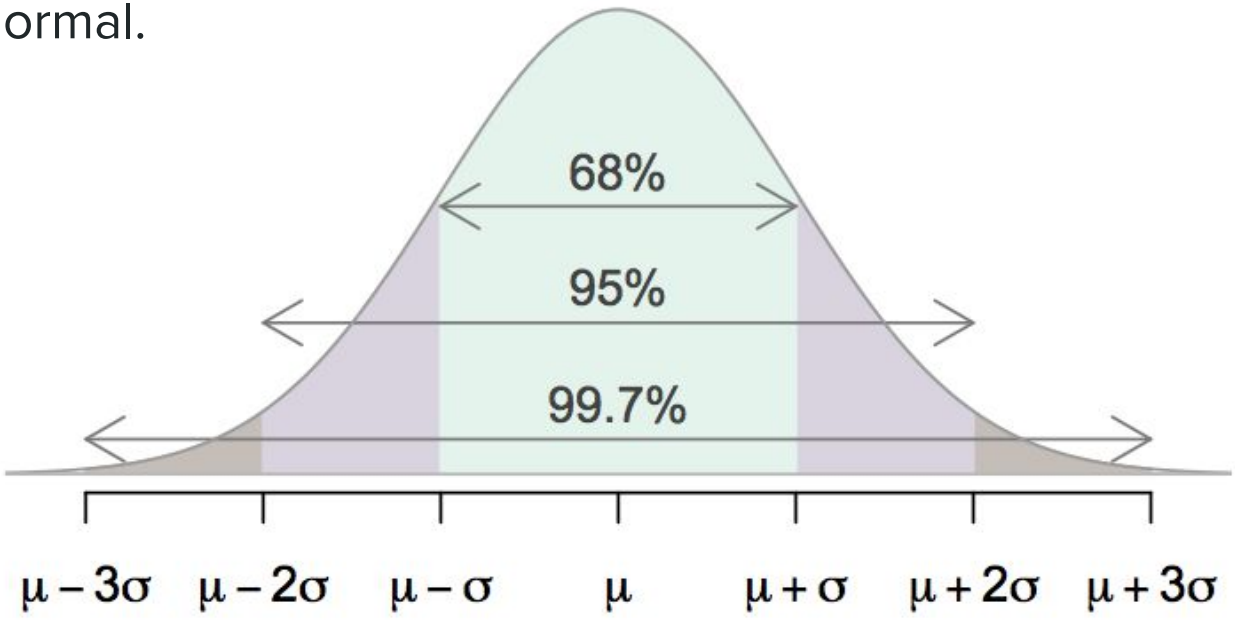
```
> qnorm(1 - 0.1, mean = 98.2, sd = 0.73)
[1] 99.13553
```

68-95-99.7 Rule

For nearly normally distributed data,

- about 68% falls within 1 SD of the mean,
- about 95% falls within 2 SD of the mean,
- about 99.7% falls within 3 SD of the mean.

It is possible for observations to fall 4, 5, or more standard deviations away from the mean, but these occurrences are very rare if the data are nearly normal.



Describing variability using the 68-95-99.7 Rule

SAT scores are distributed nearly normally with mean 1500 and standard deviation 300.

- ~68% of students score between 1200 and 1800 on the SAT.
- ~95% of students score between 900 and 2100 on the SAT.
- ~99.7% of students score between 600 and 2400 on the SAT.