Lecture 7: Normal Distribution

- Continous Distributions
- The Normal Distributions
- The Normal Approximation to the Binomial Distribution

Introduction

– In the last lecture we learned how to calculate expectation and variance.

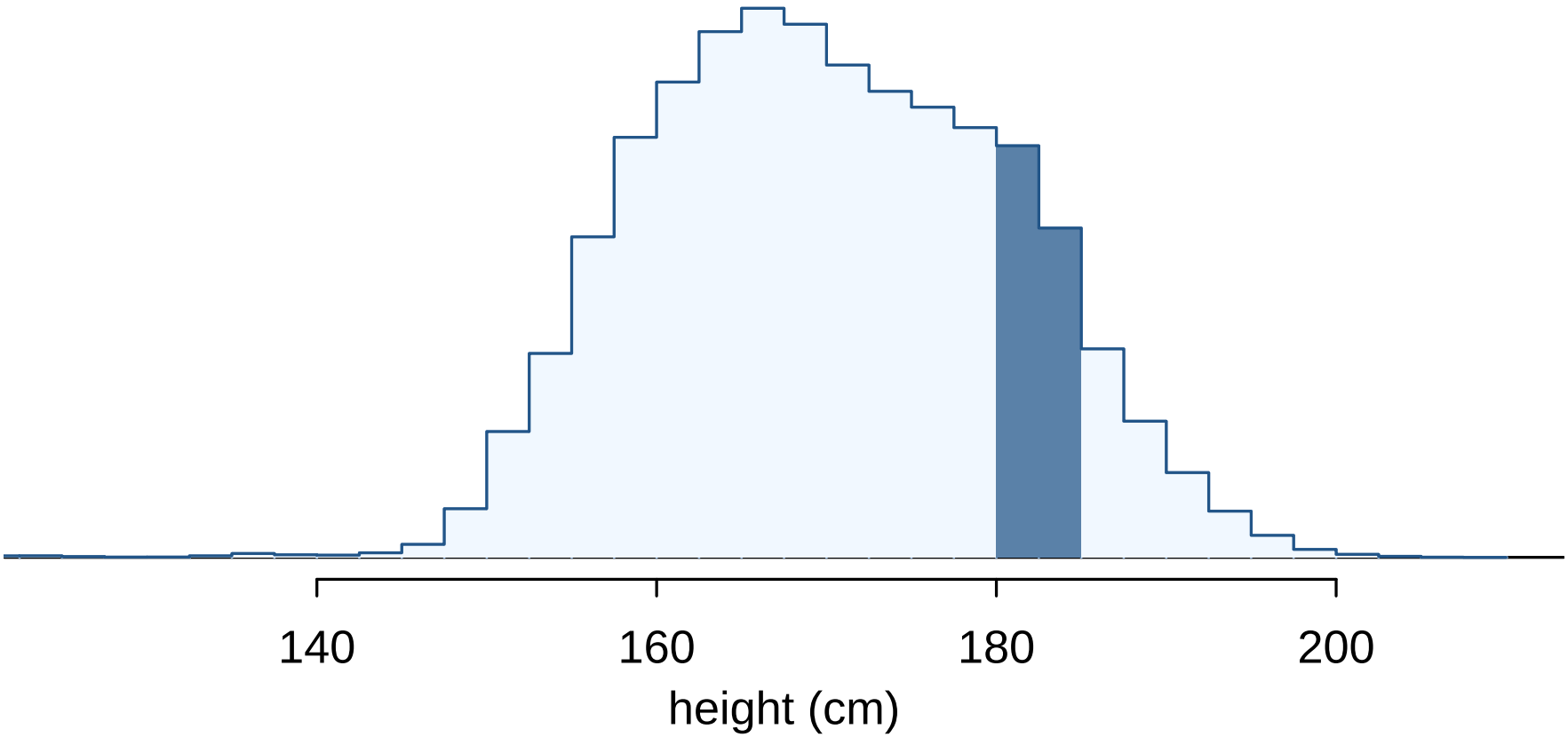
– We also learned about the standard normal distribution.

– Today we will first review continuous distributions, and then go over standard normal distribution and extend the discussion to the arbitrary case.

– Lastly, we will look at the normal approximation to the binomial distribution.

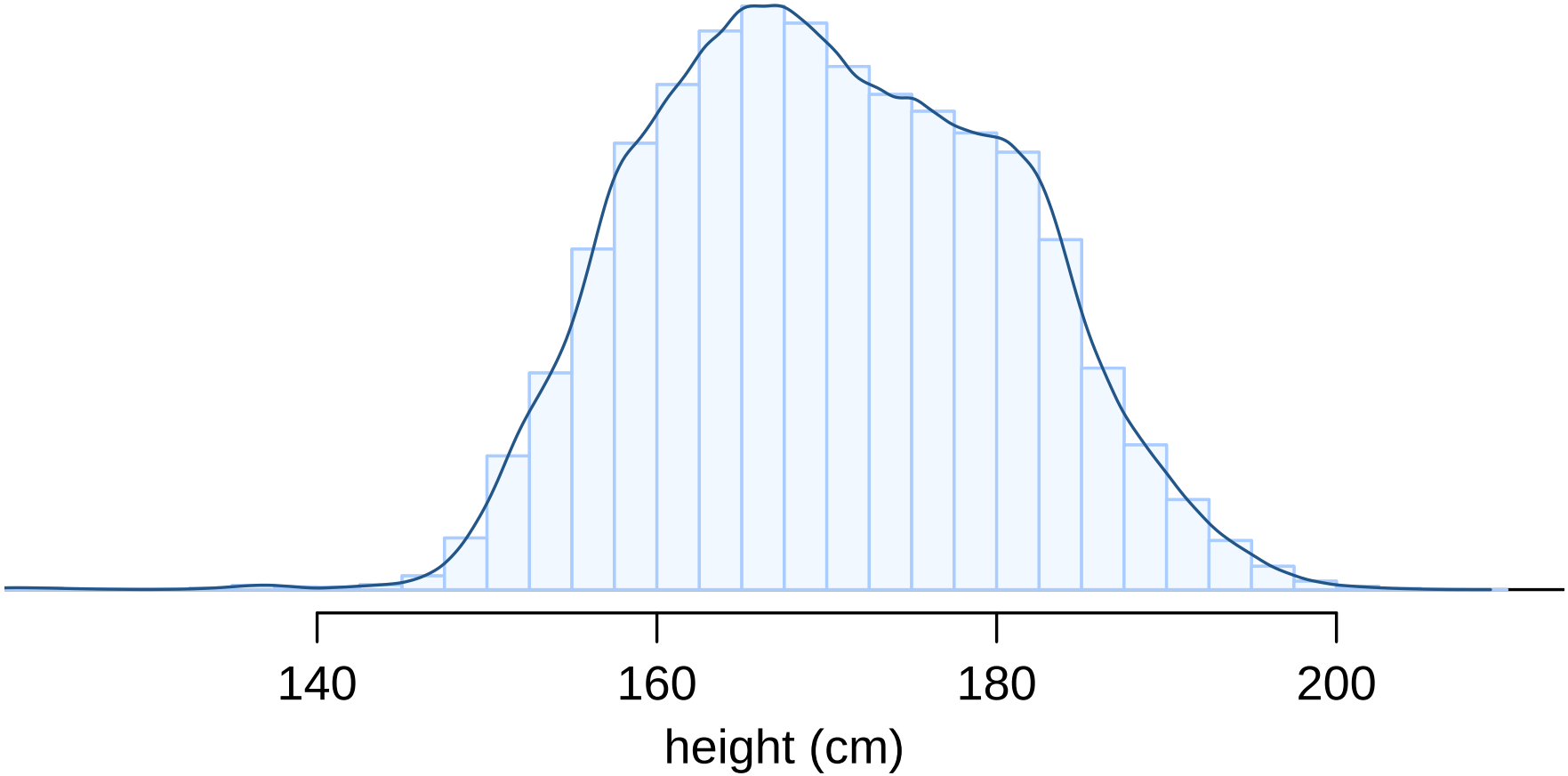
Continuous Distributions

- Below is a histogram of the distribution of heights of US adults.
- The proportion of data that falls in the shaded bins gives the probability that a randomly sampled US adult is between 180 cm and 185 cm (about 5'11" to 6'1").



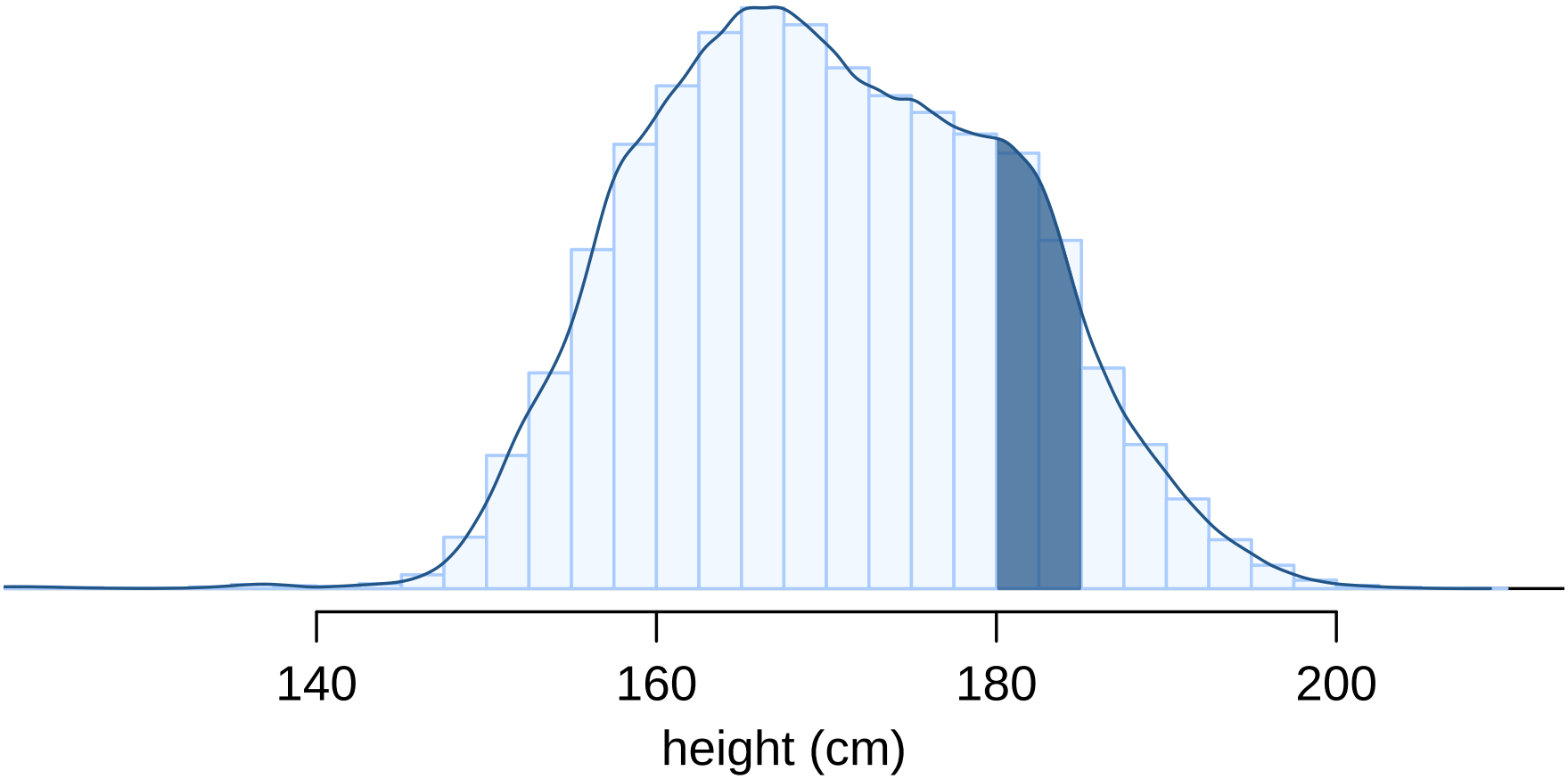
From histograms to continuous distributions

Since height is a continuous numerical variable, its **probability density function** is a smooth curve.



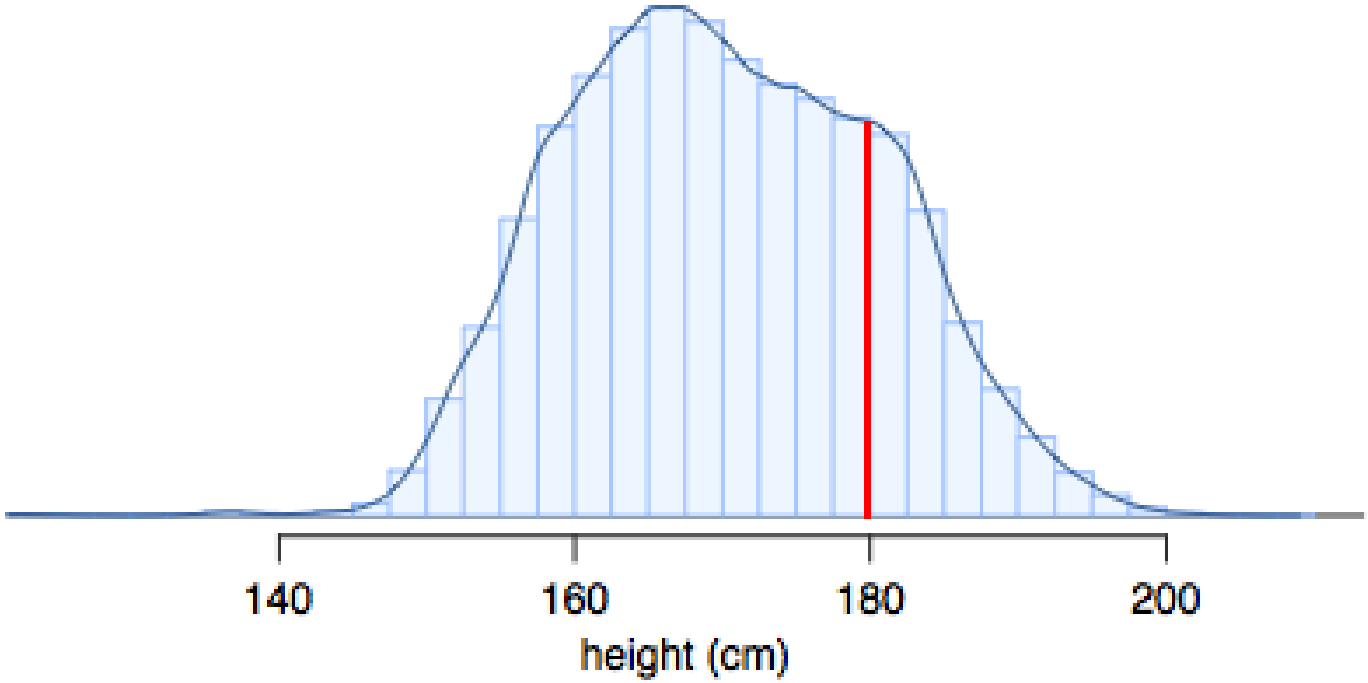
Probabilities from continuous distributions

Therefore, the probability that a randomly sampled US adult is between 180 cm and 185 cm can also be estimated as the shaded area under the curve.



By definition...

Since continuous probabilities are estimated as “the area under the curve", the probability of a person being exactly 180 cm (or any exact value) is defined as 0.



Continuous Random Variable

– The following are true for any continuous random variable X and constants a and b:

1. $\mathbb{P}(X \leq b) = \mathbb{P}(X < b)$ and $\mathbb{P}(X \geq a) = \mathbb{P}(X > a)$. This is true because we assign zero probability to events such as $X = b$ for continuous random variables, that is $\mathbb{P}(X = b) = 0$
2. $\mathbb{P}(a \leq X \leq b) = \mathbb{P}(a \leq X < b) = \mathbb{P}(a < X \leq b) = \mathbb{P}(a < X < b)$ for the same reason as above.
3. Any of the probabilities in (2) above $= F(b) - F(a)$ where $F(x)$ is the cdf of X.

The Normal Distribution

The most famous continuous distribution is the **normal distribution** (also called the Gaussian distribution or the bell-shaped curve). It's p.d.f cannot be integrated in closed form, and hence tables of the c.d.f. or computer programs are necessary in order to compute probabilities and quantiles.

The distribution was named after Carl Friedrich Gauss, the greatest mathematician in history. He proved the fundamental theorem of algebra four ways, inventing a new branch of mathematics each time. He worked in number theory, co-invented the telegraph, and discovered non-Euclidean geometry, but did not publish, fearing controversy.

The Normal Distribution (Cont'd)

People believe the normal distribution describes IQ, height, rainfall, measurement error, and many other features. This is only approximately true. However, the random variables studied in various physical experiments often have distributions that are approximately normal.

A normal distribution with mean μ and standard deviation σ has the p.d.f:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right]$$

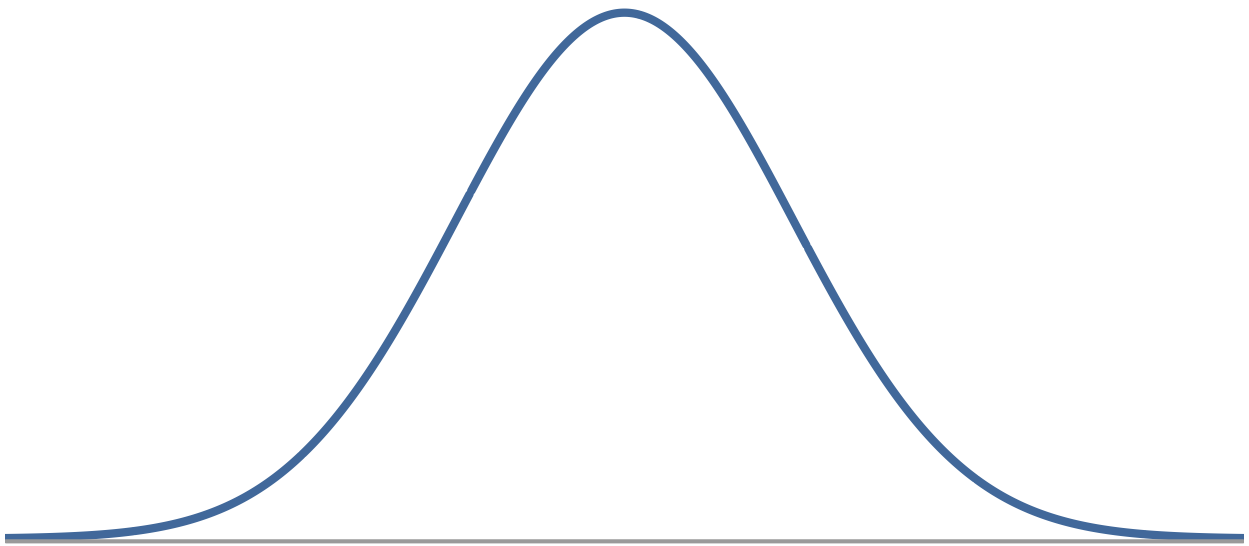
for $-\infty < \mu < \infty$ and $\sigma \geq 0$.

The μ is the mean of the entire population, whereas $\bar{X}$ is used to denote the mean of a sample from the population. Similarly, σ is the standard deviation of the entire population, whereas sd is used to denote the standard deviation of a sample.

The standard normal has $\mu = 0$ and $\sigma = 1$.

Normal distribution

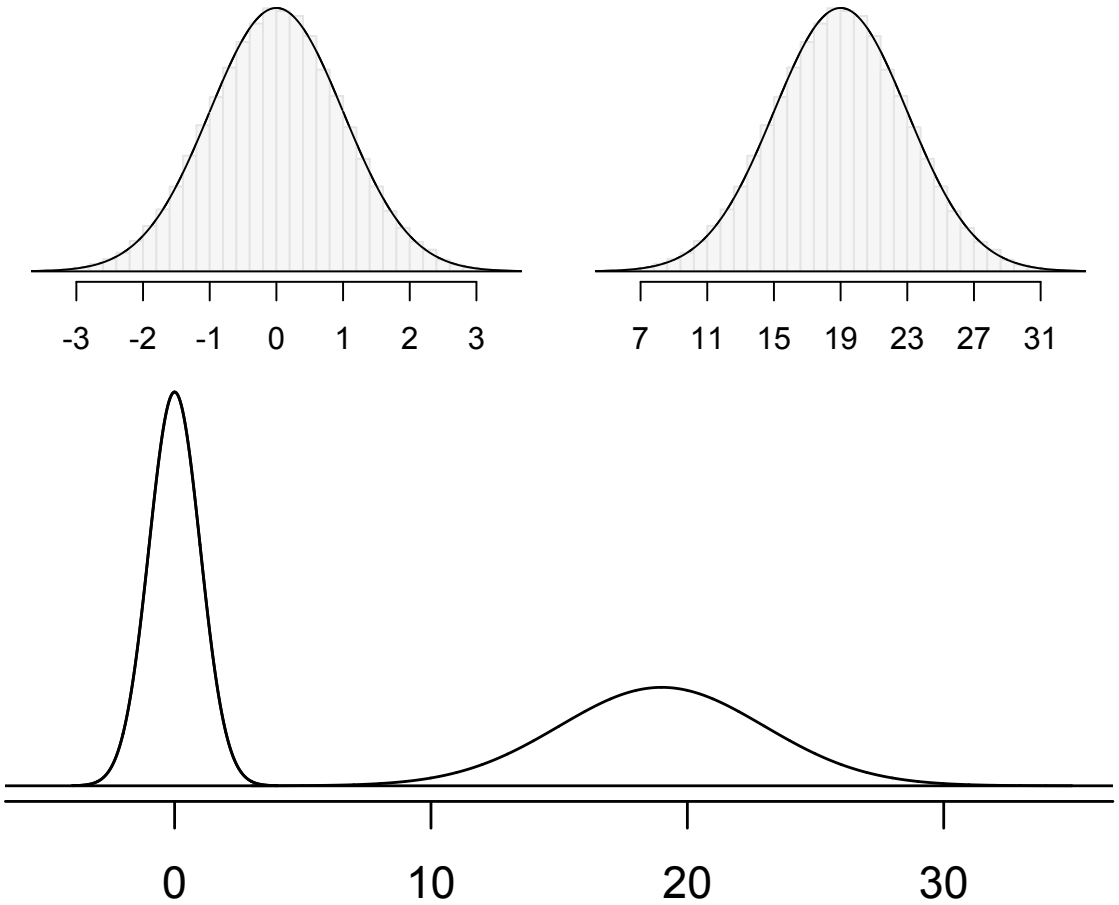
- Unimodal and symmetric, bell shaped curve
- Many variables are nearly normal, but none are exactly normal
- Denoted as $N(\mu, \sigma)$ (or $N(\mu, \sigma^2)$) $\rightarrow$ Normal with mean μ and standard deviation σ (or variance σ^2)



Normal Distributions with Different Parameters

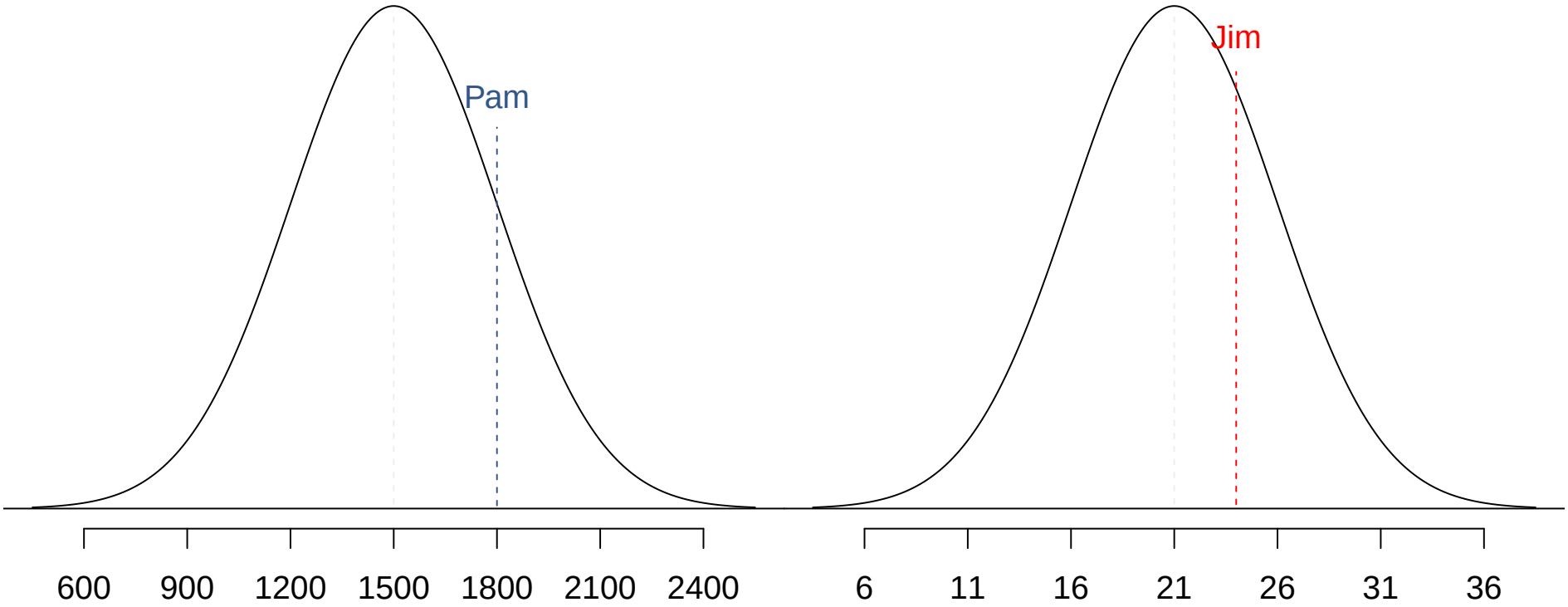
μ: mean, σ: standard deviation

$N(\mu = 0, \sigma = 1)$ $N(\mu = 19, \sigma = 4)$



SAT v.s. ACT

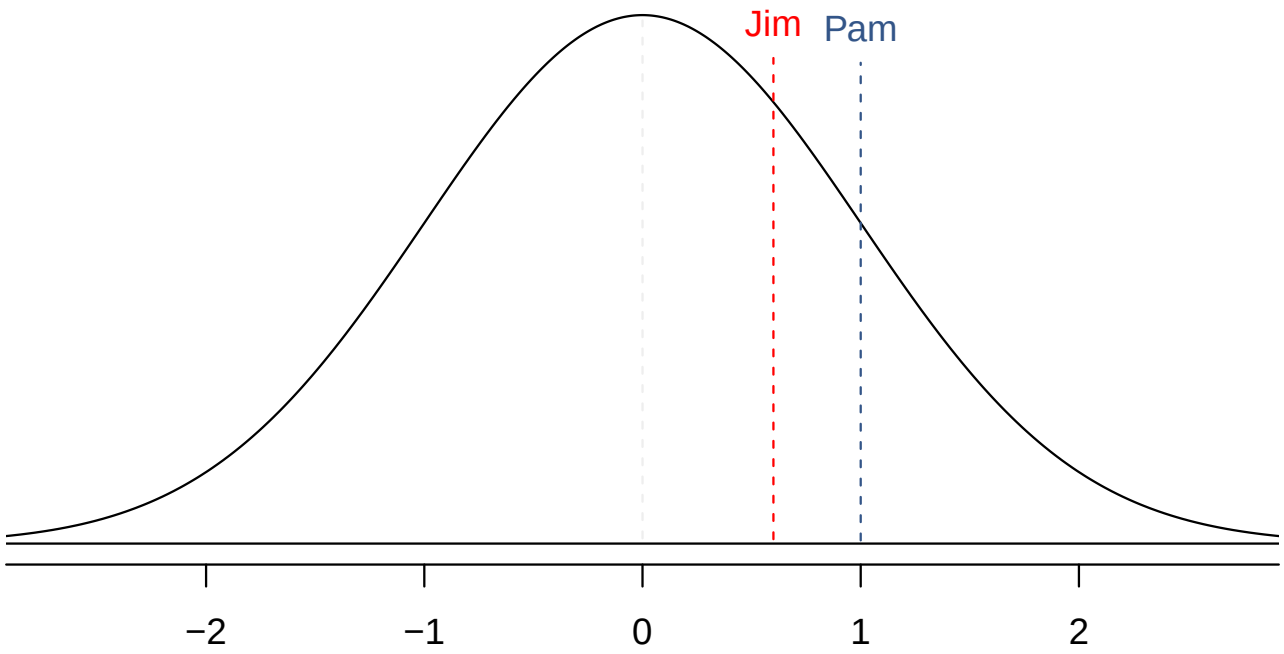
SAT scores are distributed nearly normally with mean 1500 and standard deviation 300. ACT scores are distributed nearly normally with mean 21 and standard deviation 5. A college admissions officer wants to determine which of the two applicants scored better on their standardized test with respect to 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 $\frac{1800-1500}{300} = 1$ standard deviation above the mean.
- Jim's score is $\frac{24-21}{5} = 0.6$ standard deviations above the mean.



Standardizing with Z scores (cont.)

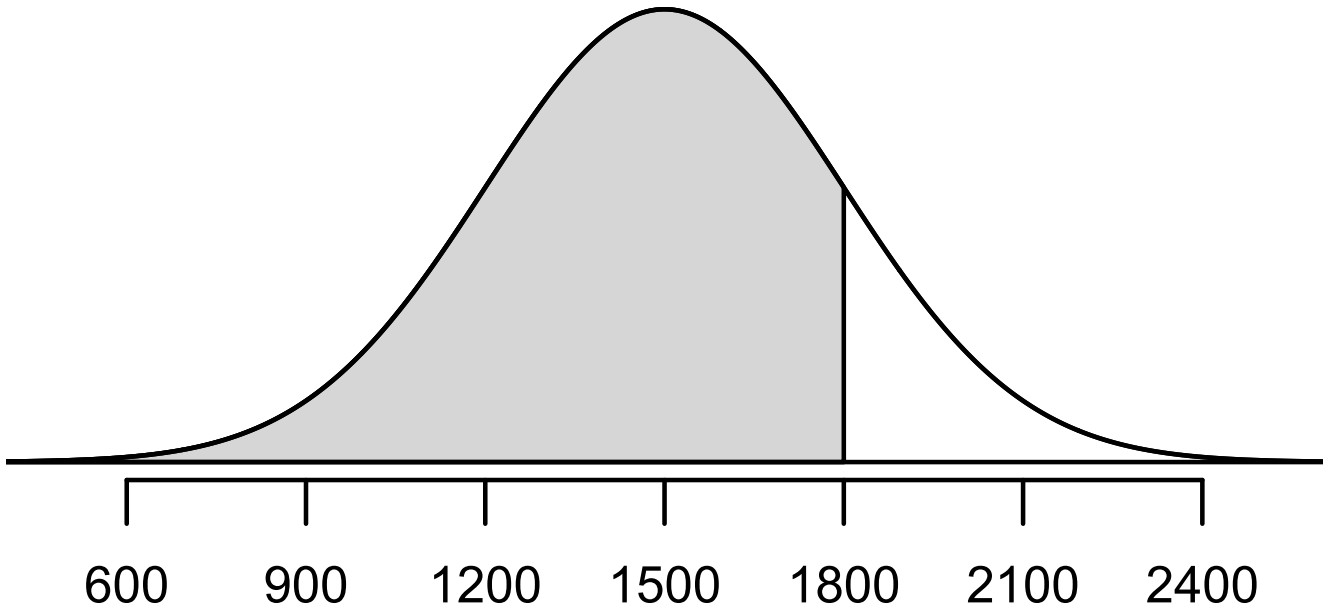
- These are called *standardized* scores, or *Z scores*.
- Z score of an observation is the number of standard deviations it falls above or below the mean.

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

- Z scores are defined for distributions of any shape
- 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.



Calculating percentiles - using computation

There are many ways to compute percentiles/areas under the curve:

- R:

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

- Applet: *https://gallery.shinyapps.io/dist_calc/*

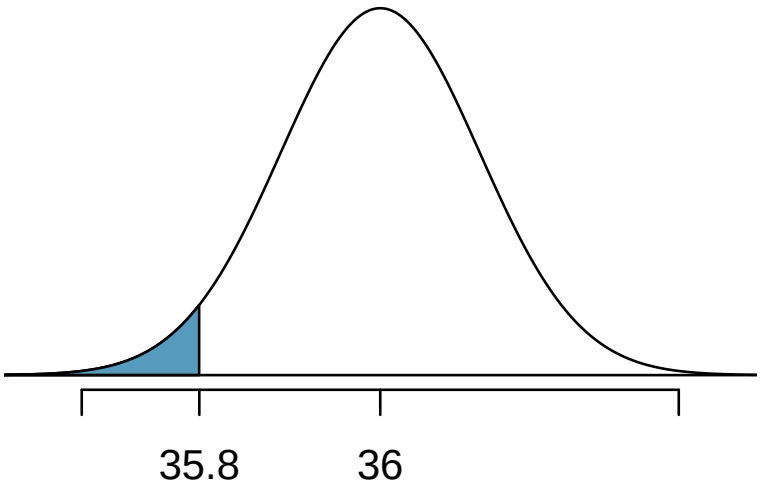
Calculating percentiles - using tables

	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?

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



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

Finding the exact probability - using the Z table

Second decimal place of Z										
0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01	0.00	Z
0.0014	0.0014	0.0015	0.0015	0.0016	0.0016	0.0017	0.0018	0.0018	0.0019	−2.9
0.0019	0.0020	0.0021	0.0021	0.0022	0.0023	0.0023	0.0024	0.0025	0.0026	−2.8
0.0026	0.0027	0.0028	0.0029	0.0030	0.0031	0.0032	0.0033	0.0034	0.0035	−2.7
0.0036	0.0037	0.0038	0.0039	0.0040	0.0041	0.0043	0.0044	0.0045	0.0047	−2.6
0.0048	0.0049	0.0051	0.0052	0.0054	0.0055	0.0057	0.0059	0.0060	0.0062	−2.5
0.0064	0.0066	0.0068	0.0069	0.0071	0.0073	0.0075	0.0078	0.0080	0.0082	−2.4
0.0084	0.0087	0.0089	0.0091	0.0094	0.0096	0.0099	0.0102	0.0104	0.0107	−2.3
0.0110	0.0113	0.0116	0.0119	0.0122	0.0125	0.0129	0.0132	0.0136	0.0139	−2.2
0.0143	0.0146	0.0150	0.0154	0.0158	0.0162	0.0166	0.0170	0.0174	0.0179	−2.1
0.0183	0.0188	0.0192	0.0197	0.0202	0.0207	0.0212	0.0217	0.0222	0.0228	−2.0
0.0233	0.0239	0.0244	0.0250	0.0256	0.0262	0.0268	0.0274	0.0281	0.0287	−1.9
0.0294	0.0301	0.0307	0.0314	0.0322	0.0329	0.0336	0.0344	0.0351	0.0359	−1.8
0.0367	0.0375	0.0384	0.0392	0.0401	0.0409	0.0418	0.0427	0.0436	0.0446	−1.7
0.0455	0.0465	0.0475	0.0485	0.0495	0.0505	0.0516	0.0526	0.0537	0.0548	−1.6
0.0559	0.0571	0.0582	0.0594	0.0606	0.0618	0.0630	0.0643	0.0655	0.0668	−1.5

Finding the exact probability - using the Z table

Second decimal place of Z										
0.09	*0.08*	*0.07*	*0.06*	*0.05*	*0.04*	*0.03*	*0.02*	*0.01*	*0.00*	*Z*
0.0014	*0.0014*	*0.0015*	*0.0015*	*0.0016*	*0.0016*	*0.0017*	*0.0018*	*0.0018*	*0.0019*	−2.9
0.0019	*0.0020*	*0.0021*	*0.0021*	*0.0022*	*0.0023*	*0.0023*	*0.0024*	*0.0025*	*0.0026*	−2.8
0.0026	*0.0027*	*0.0028*	*0.0029*	*0.0030*	*0.0031*	*0.0032*	*0.0033*	*0.0034*	*0.0035*	−2.7
0.0036	*0.0037*	*0.0038*	*0.0039*	*0.0040*	*0.0041*	*0.0043*	*0.0044*	*0.0045*	*0.0047*	−2.6
0.0048	*0.0049*	*0.0051*	*0.0052*	*0.0054*	*0.0055*	*0.0057*	*0.0059*	*0.0060*	*0.0062*	−2.5
0.0064	*0.0066*	*0.0068*	*0.0069*	*0.0071*	*0.0073*	*0.0075*	*0.0078*	*0.0080*	*0.0082*	−2.4
0.0084	*0.0087*	*0.0089*	*0.0091*	*0.0094*	*0.0096*	*0.0099*	*0.0102*	*0.0104*	*0.0107*	−2.3
0.0110	*0.0113*	*0.0116*	*0.0119*	*0.0122*	*0.0125*	*0.0129*	*0.0132*	*0.0136*	*0.0139*	−2.2
0.0143	*0.0146*	*0.0150*	*0.0154*	*0.0158*	*0.0162*	*0.0166*	*0.0170*	*0.0174*	*0.0179*	−2.1
0.0183	*0.0188*	*0.0192*	*0.0197*	*0.0202*	*0.0207*	*0.0212*	*0.0217*	*0.0222*	*0.0228*	−2.0
0.0233	*0.0239*	*0.0244*	*0.0250*	*0.0256*	*0.0262*	*0.0268*	*0.0274*	*0.0281*	*0.0287*	−1.9
0.0294	*0.0301*	*0.0307*	*0.0314*	*0.0322*	*0.0329*	*0.0336*	*0.0344*	*0.0351*	*0.0359*	−1.8
0.0367	*0.0375*	*0.0384*	*0.0392*	*0.0401*	*0.0409*	*0.0418*	*0.0427*	*0.0436*	*0.0446*	−1.7
0.0455	*0.0465*	*0.0475*	*0.0485*	*0.0495*	*0.0505*	*0.0516*	*0.0526*	*0.0537*	*0.0548*	−1.6
0.0559	*0.0571*	*0.0582*	*0.0594*	*0.0606*	*0.0618*	*0.0630*	*0.0643*	*0.0655*	*0.0668*	−1.5

Practice

What percent of bottles pass the quality control inspection?

(a) 1.82%

(b) 3.44%

(c) 6.88%

(d) 93.12%

(e) 96.56%

Practice

What percent of bottles pass the quality control inspection?

(a) 1.82%

(b) 3.44%

(c) 6.88%

(d) *93.12%*

(e) 96.56%

Practice

What percent of bottles pass the quality control inspection?

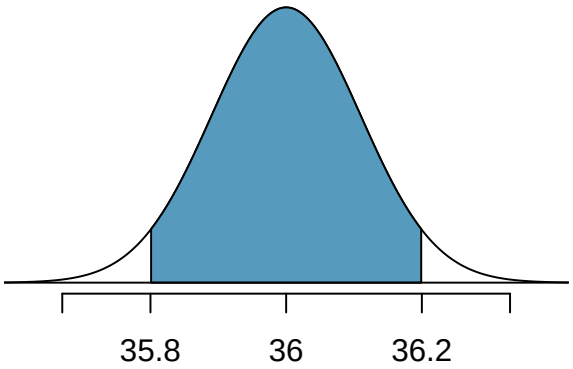
(a) 1.82%

(b) 3.44%

(c) 6.88%

(d) *93.12%*

(e) 96.56%



=

Practice

What percent of bottles pass the quality control inspection?

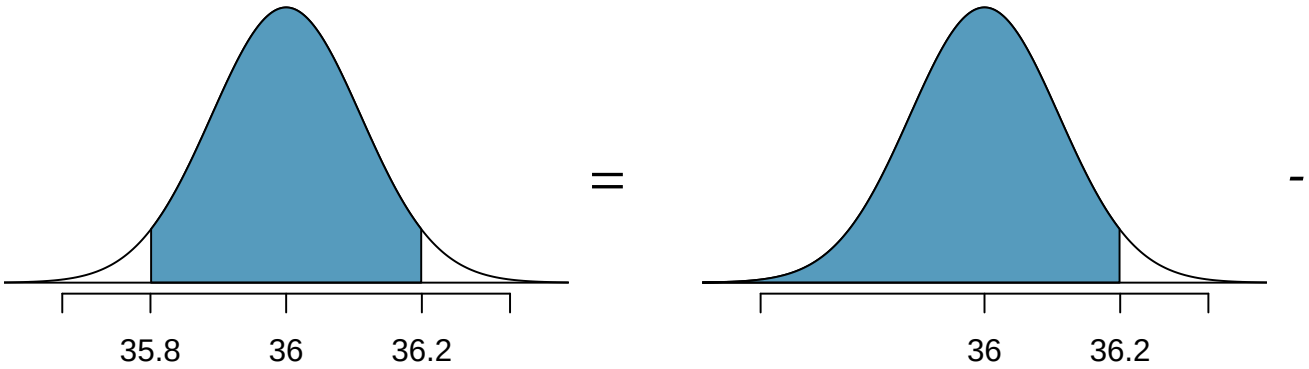
(a) 1.82%

(b) 3.44%

(c) 6.88%

(d) *93.12%*

(e) 96.56%



Practice

What percent of bottles pass the quality control inspection?

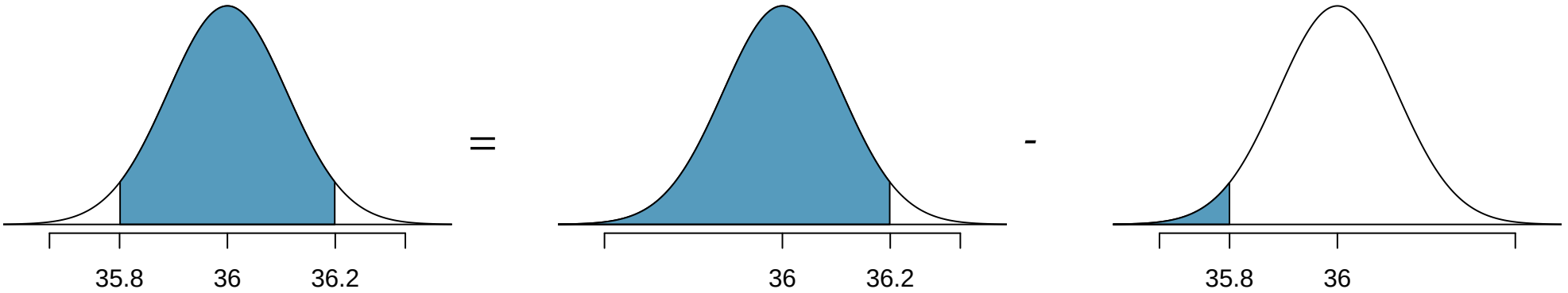
(a) 1.82%

(b) 3.44%

(c) 6.88%

(d) *93.12%*

(e) 96.56%



Practice

What percent of bottles pass the quality control inspection?

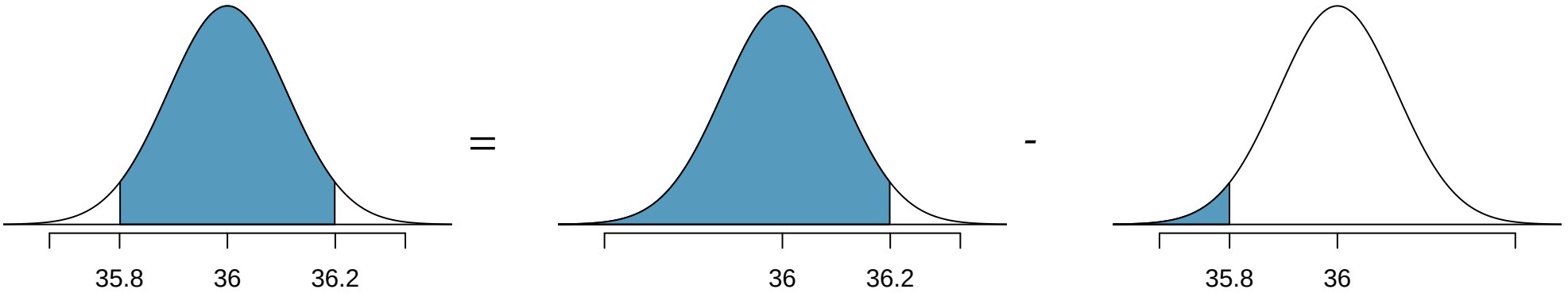
(a) 1.82%

(b) 3.44%

(c) 6.88%

(d) *93.12%*

(e) 96.56%



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

Practice

What percent of bottles pass the quality control inspection?

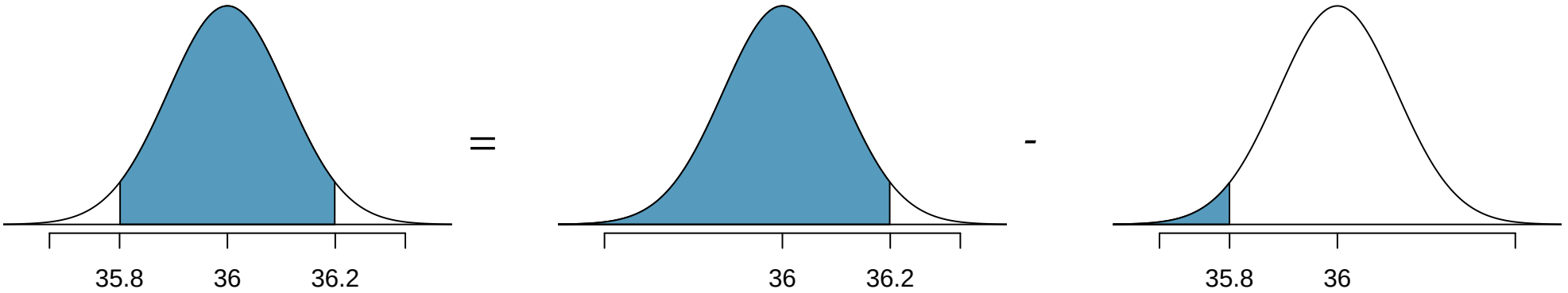
(a) 1.82%

(b) 3.44%

(c) 6.88%

(d) *93.12%*

(e) 96.56%



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

$$Z_{36.2} = \frac{36.2 - 36}{0.11} = 1.82$$

Practice

What percent of bottles pass the quality control inspection?

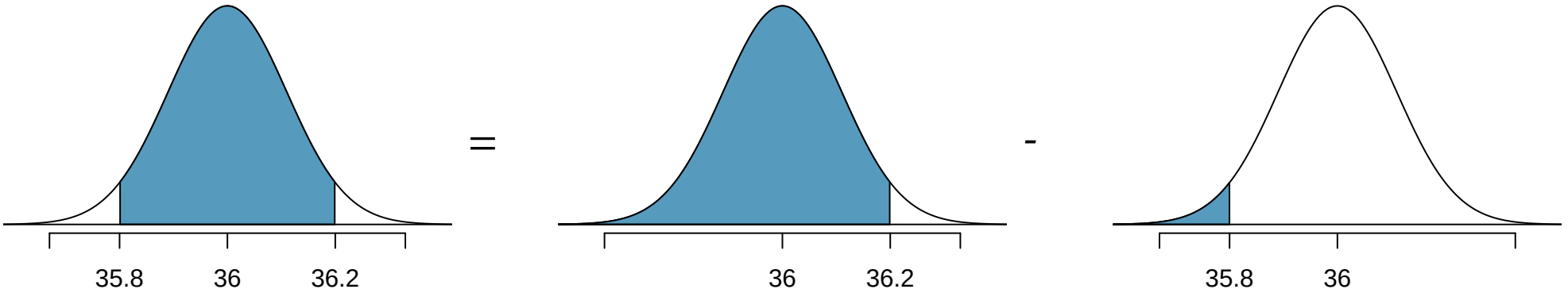
(a) 1.82%

(b) 3.44%

(c) 6.88%

(d) *93.12%*

(e) 96.56%



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

$$Z_{36.2} = \frac{36.2 - 36}{0.11} = 1.82$$

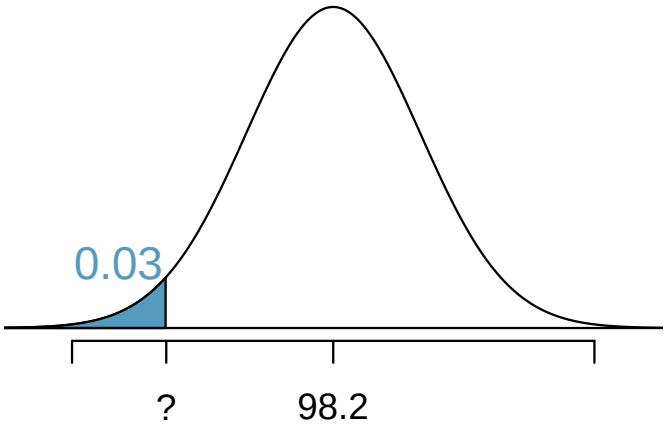
$$P(35.8 < X < 36.2) = P(-1.82 < Z < 1.82) = 0.9656 - 0.0344 = 0.9312$$

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?

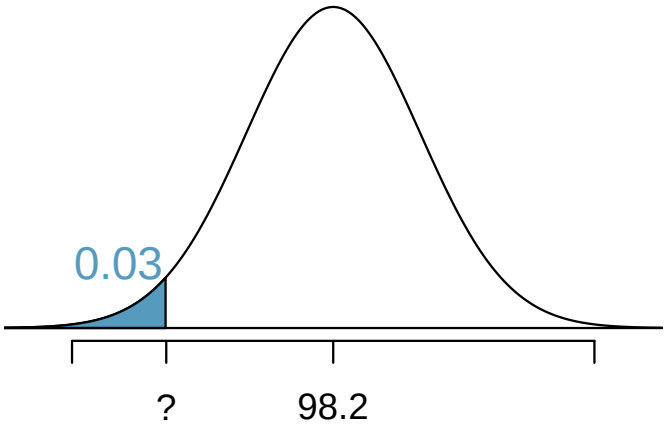
Finding cutoff points

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?



Finding cutoff points

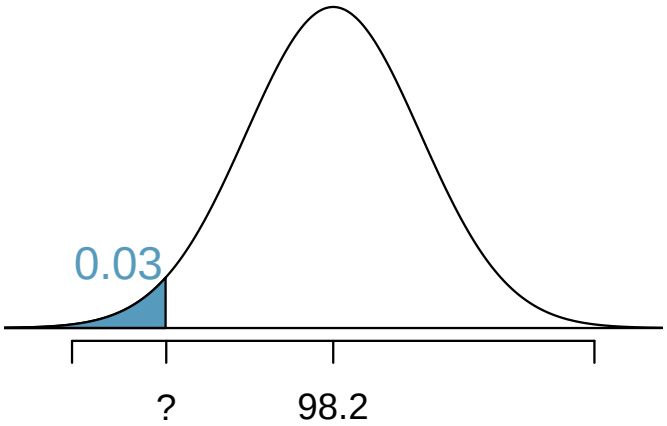
Body temperatures of healthy humans are distributed nearly normally with mean 98.2°F and standard deviation 0.73°F. What is the cutoff for the lowest 3% of human body temperatures?



0.09	0.08	0.07	0.06	0.05	Z
0.0233	0.0239	0.0244	0.0250	0.0256	-1.9
0.0294	*0.0301*	0.0307	0.0314	0.0322	-1.8
0.0367	0.0375	0.0384	0.0392	0.0401	-1.7

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?

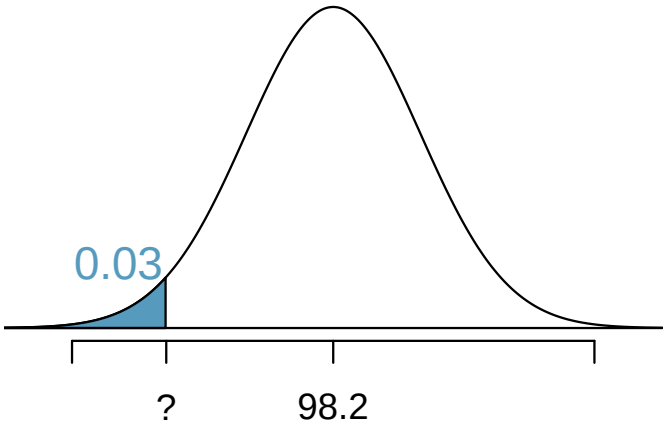


0.09	0.08	0.07	0.06	0.05	Z
0.0233	0.0239	0.0244	0.0250	0.0256	-1.9
0.0294	*0.0301*	0.0307	0.0314	0.0322	-1.8
0.0367	0.0375	0.0384	0.0392	0.0401	-1.7

$$P(X < x) = 0.03 \rightarrow P(Z < \textit{-1.88}) = 0.03$$

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?



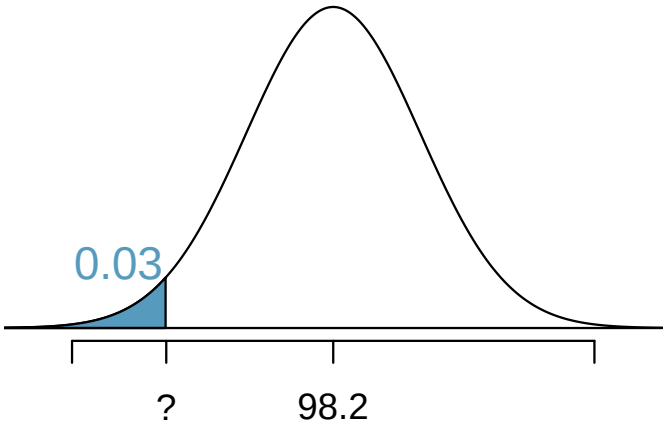
0.09	0.08	0.07	0.06	0.05	Z
0.0233	0.0239	0.0244	0.0250	0.0256	-1.9
0.0294	*0.0301*	0.0307	0.0314	0.0322	-1.8
0.0367	0.0375	0.0384	0.0392	0.0401	-1.7

$$P(X < x) = 0.03 \rightarrow P(Z < \textit{-1.88}) = 0.03$$

$$Z = \frac{obs - mean}{SD} \rightarrow \frac{x - 98.2}{0.73} = -1.88$$

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?



0.09	0.08	0.07	0.06	0.05	Z
0.0233	0.0239	0.0244	0.0250	0.0256	-1.9
0.0294	*0.0301*	0.0307	0.0314	0.0322	-1.8
0.0367	0.0375	0.0384	0.0392	0.0401	-1.7

$$
\begin{aligned}
P(X < x) &= 0.03 \rightarrow P(Z < \text{-}1.88) = 0.03 \\
Z &= \frac{obs \ - \ mean}{SD} \rightarrow \frac{x - 98.2}{0.73} = -1.88 \\
x &= (-1.88 \times 0.73) + 98.2 = 96.8^\circ F
\end{aligned}
$$

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

(b) 99.1°F

(c) 99.4°F

(d) 99.6°F

Practice

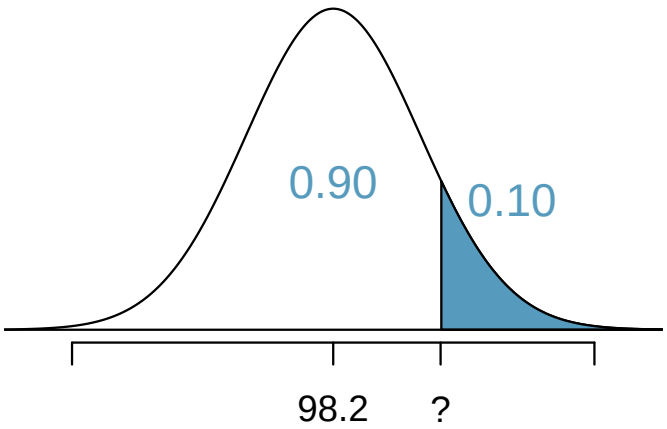
Body temperatures of healthy humans are distributed nearly normally with mean 98.2°F and standard deviation 0.73°F. What is the cutoff for the highest 10% of human body temperatures?

(a) 97.3°F

(b) *99.1°F*

(c) 99.4°F

(d) 99.6°F



Practice

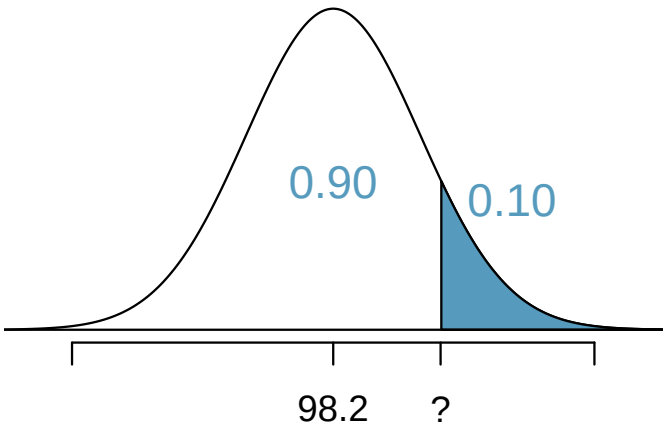
Body temperatures of healthy humans are distributed nearly normally with mean 98.2°F and standard deviation 0.73°F. What is the cutoff for the highest 10% of human body temperatures?

(a) 97.3°F

(b) *99.1°F*

(c) 99.4°F

(d) 99.6°F



Z	*0.05*	*0.06*	*0.07*	*0.08*	*0.09*
1.0	*0.8531*	*0.8554*	*0.8577*	*0.8599*	*0.8621*
1.1	*0.8749*	*0.8770*	*0.8790*	*0.8810*	*0.8830*
1.2	*0.8944*	*0.8962*	*0.8980*	*0.8997*	*0.9015*
1.3	*0.9115*	*0.9131*	*0.9147*	*0.9162*	*0.9177*

Practice

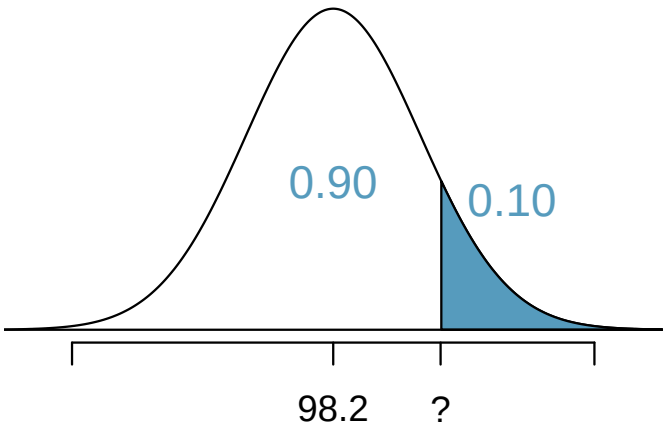
Body temperatures of healthy humans are distributed nearly normally with mean 98.2°F and standard deviation 0.73°F. What is the cutoff for the highest 10% of human body temperatures?

(a) 97.3°F

(b) *99.1°F*

(c) 99.4°F

(d) 99.6°F



Z	*0.05*	*0.06*	*0.07*	*0.08*	*0.09*
1.0	*0.8531*	*0.8554*	*0.8577*	*0.8599*	*0.8621*
1.1	*0.8749*	*0.8770*	*0.8790*	*0.8810*	*0.8830*
1.2	*0.8944*	*0.8962*	*0.8980*	*0.8997*	*0.9015*
1.3	*0.9115*	*0.9131*	*0.9147*	*0.9162*	*0.9177*

$$P(X > x) = 0.10 \rightarrow P(Z < 1.28) = 0.90$$

Practice

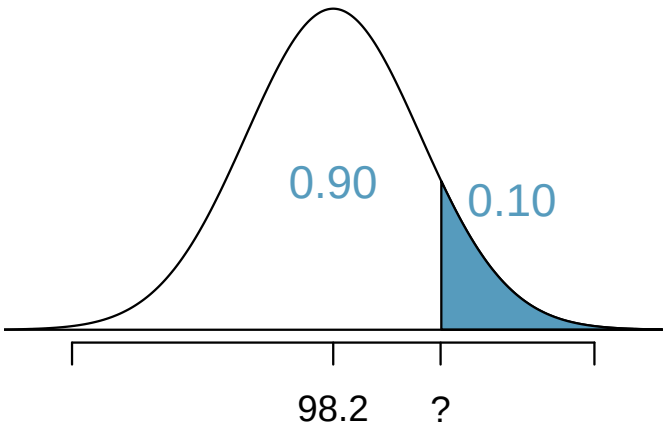
Body temperatures of healthy humans are distributed nearly normally with mean 98.2°F and standard deviation 0.73°F. What is the cutoff for the highest 10% of human body temperatures?

(a) 97.3°F

(b) *99.1°F*

(c) 99.4°F

(d) 99.6°F



Z	0.05	0.06	0.07	0.08	0.09
1.0	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9115	0.9131	0.9147	0.9162	0.9177

$$P(X > x) = 0.10 \rightarrow P(Z < 1.28) = 0.90$$

$$Z = \frac{obs - mean}{SD} \rightarrow \frac{x - 98.2}{0.73} = 1.28$$

Practice

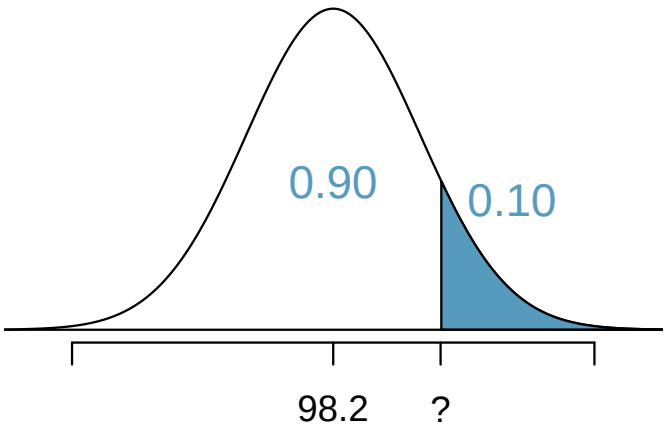
Body temperatures of healthy humans are distributed nearly normally with mean 98.2°F and standard deviation 0.73°F. What is the cutoff for the highest 10% of human body temperatures?

(a) 97.3°F

(b) *99.1°F*

(c) 99.4°F

(d) 99.6°F

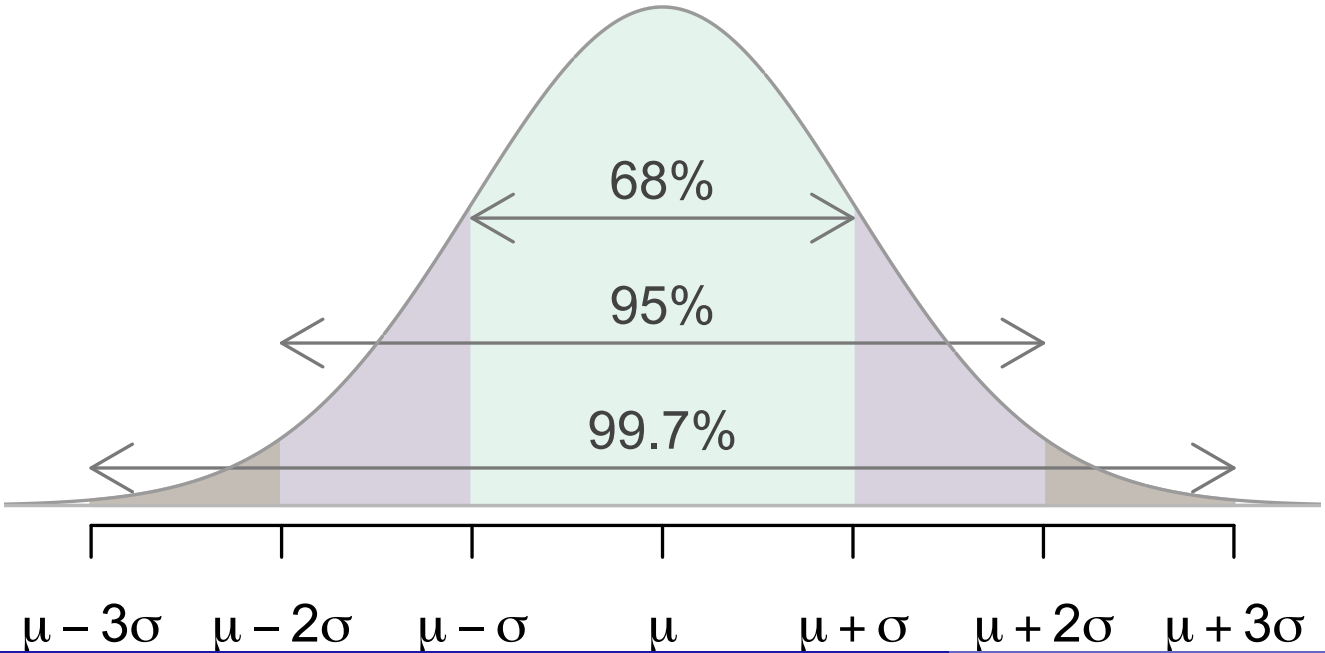


Z	0.05	0.06	0.07	0.08	0.09
1.0	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9115	0.9131	0.9147	0.9162	0.9177

$$
\begin{aligned}
P(X > x) &= 0.10 \rightarrow P(Z < 1.28) = 0.90 \\
Z &= \frac{obs - mean}{SD} \rightarrow \frac{x - 98.2}{0.73} = 1.28 \\
x &= (1.28 \times 0.73) + 98.2 = 99.1
\end{aligned}
$$

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.



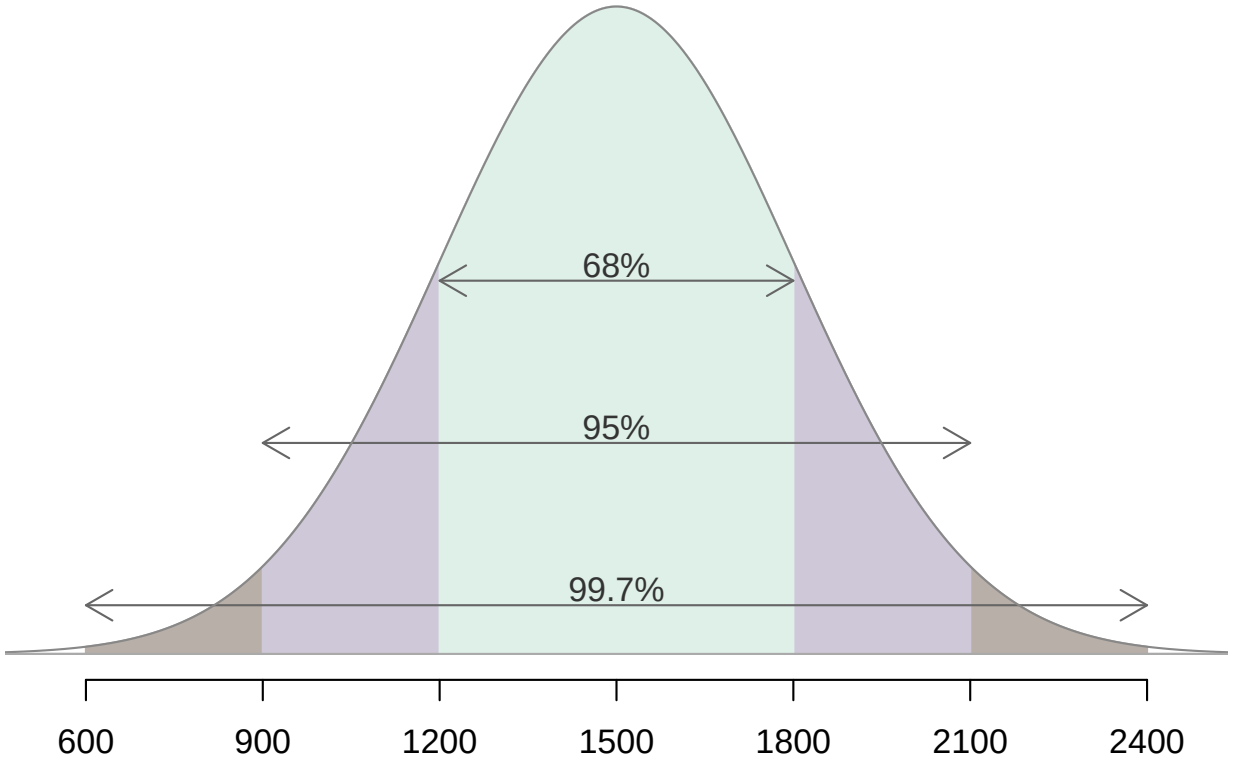
Describing variability using the 68-95-99.7 Rule

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

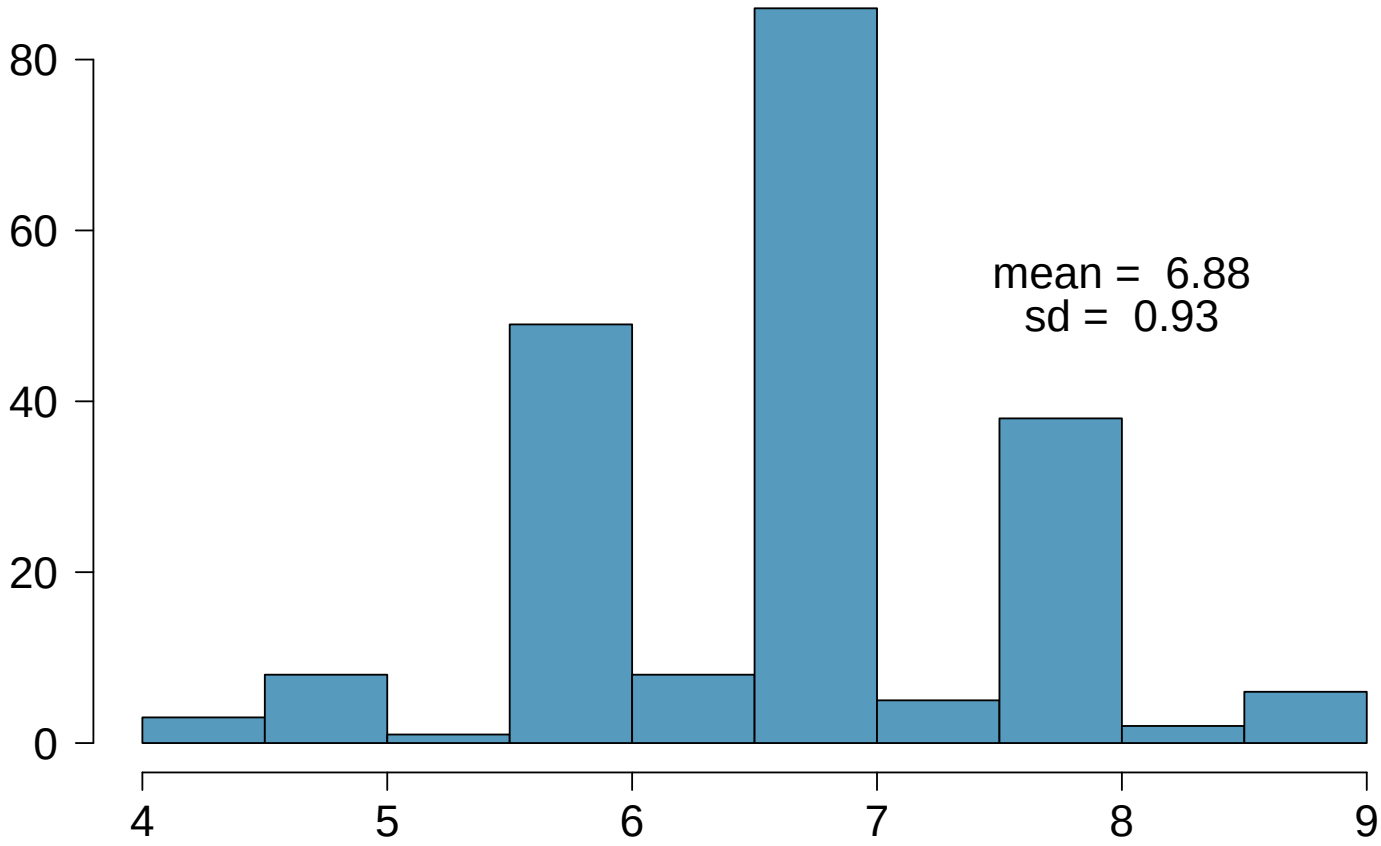
Describing variability using the 68-95-99.7 Rule

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.

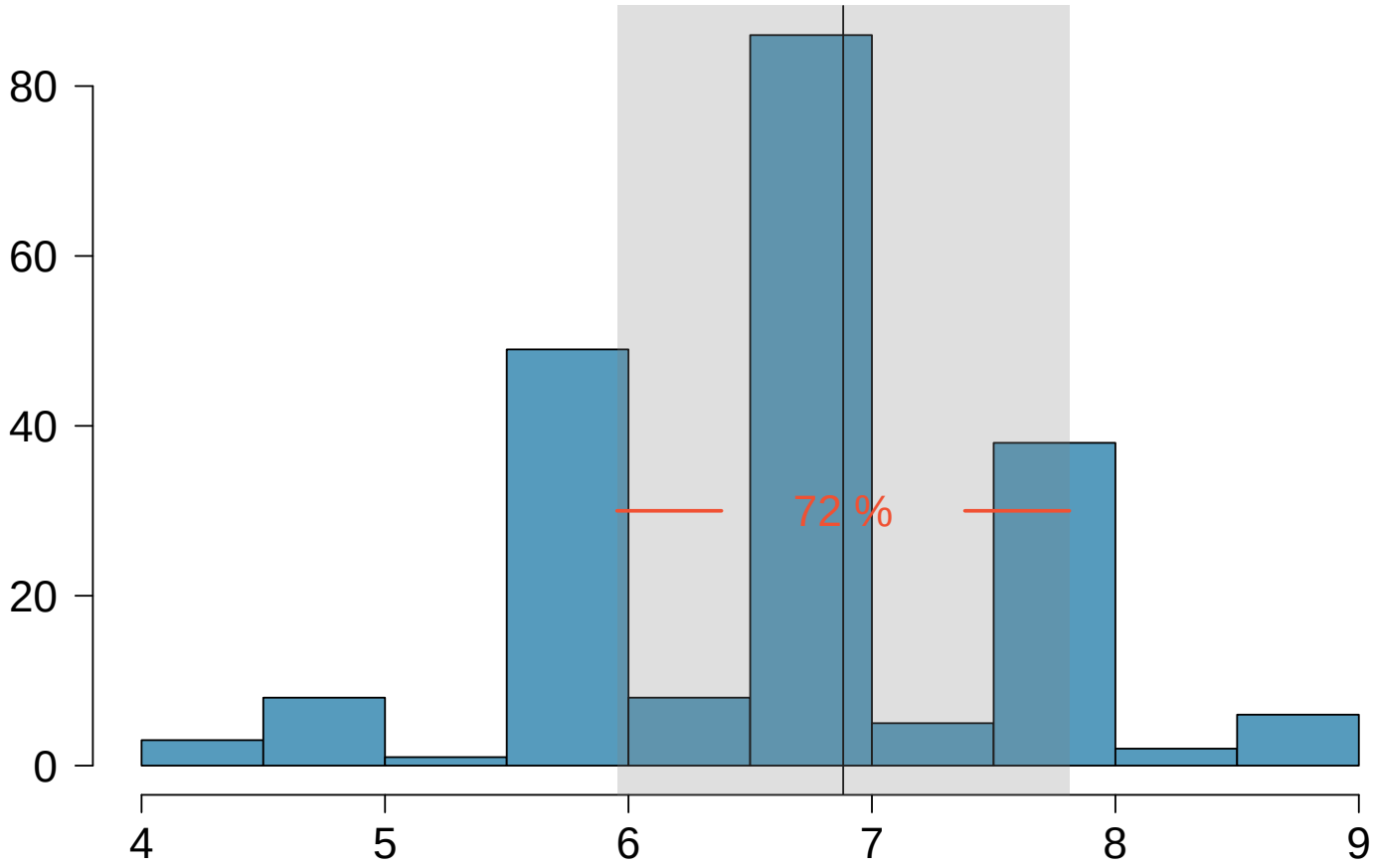


Number of hours of sleep on school nights



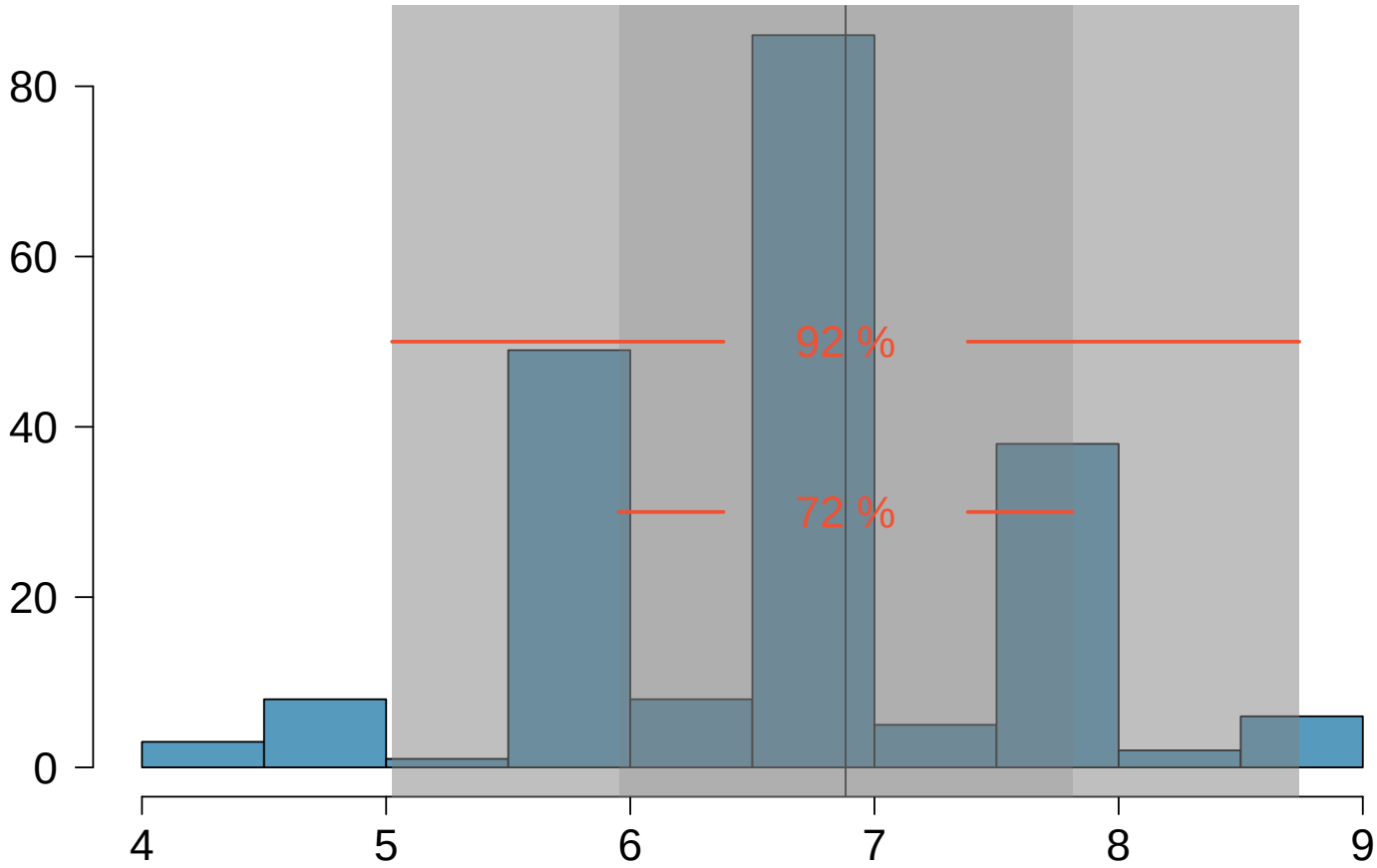
- Mean = 6.88 hours, SD = 0.93 hrs

Number of hours of sleep on school nights



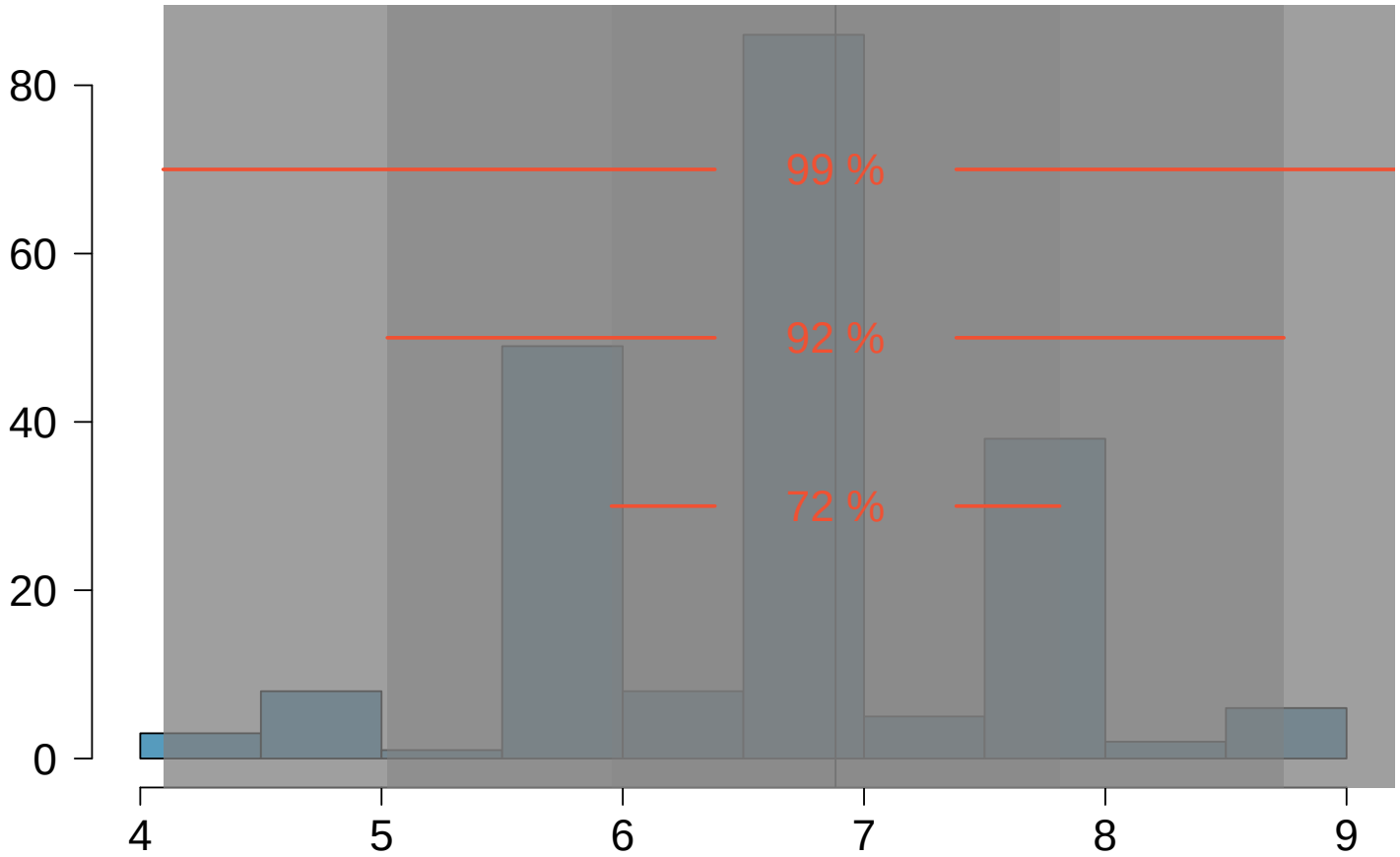
- Mean = 6.88 hours, SD = 0.93 hrs
- 72% of the data are within 1 SD of the mean: 6.88 ± 0.93

Number of hours of sleep on school nights



- Mean = 6.88 hours, SD = 0.93 hrs
- 72% of the data are within 1 SD of the mean: 6.88 ± 0.93
- 92% of the data are within 2 SD of the mean: $6.88 \pm 2 \times 0.93$

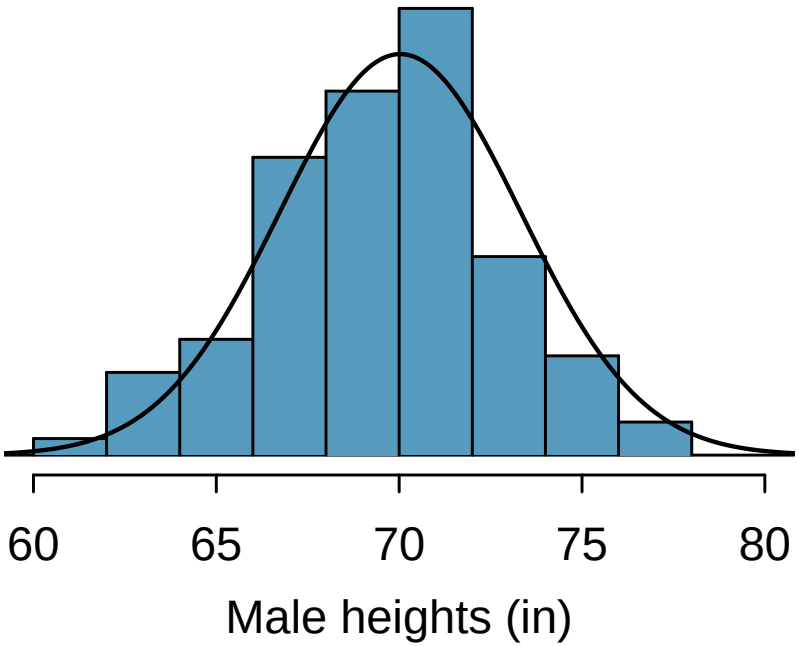
Number of hours of sleep on school nights

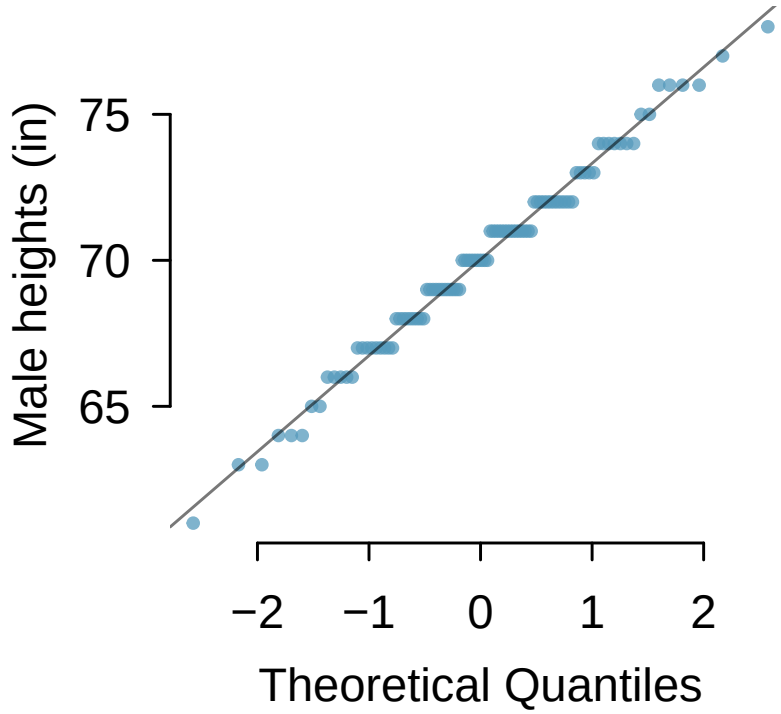


- Mean = 6.88 hours, SD = 0.93 hrs
- 72% of the data are within 1 SD of the mean: 6.88 ± 0.93
- 92% of the data are within 2 SD of the mean: $6.88 \pm 2 \times 0.93$
- 99% of the data are within 3 SD of the mean: $6.88 \pm 3 \times 0.93$

Normal probability plot

A histogram and *normal probability plot* of a sample of 100 male heights.



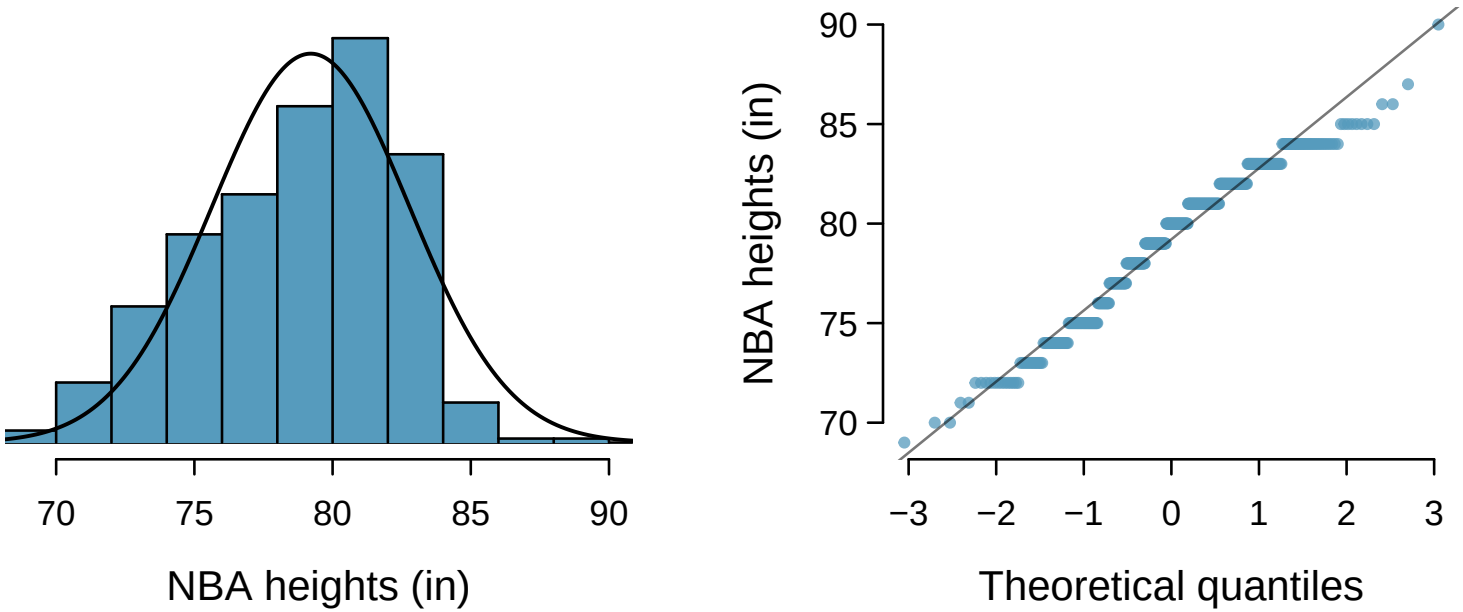


Anatomy of a normal probability plot

- Data are plotted on the y-axis of a normal probability plot, and theoretical quantiles (following a normal distribution) on the x-axis.
- If there is a linear relationship in the plot, then the data follow a nearly normal distribution.
- Constructing a normal probability plot requires calculating percentiles and corresponding z-scores for each observation, which is tedious. Therefore we generally rely on software when making these plots.

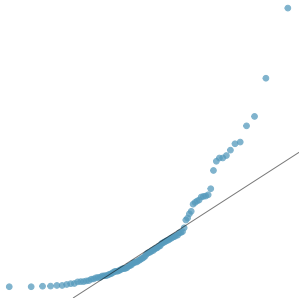
NBA Heights

Below is a histogram and normal probability plot for the NBA heights from the 2008-2009 season. Do these data appear to follow a normal distribution?

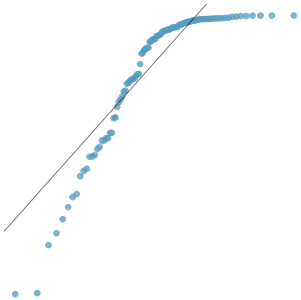


The histogram in the left panel is slightly left skewed, which contrasts with the symmetric normal distribution. The points in the normal probability plot do not appear to closely follow a straight line but show what appears to be a "wave". NBA player heights do not appear to come from a normal distribution.

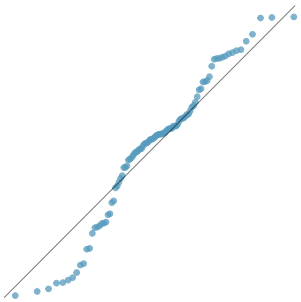
Normal probability plot and skewness



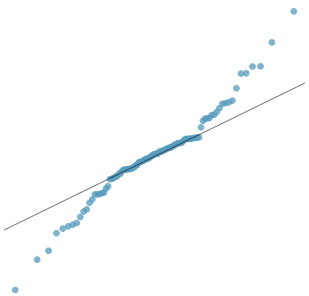
Right skew - Points bend up and to the left of the line.



Left skew- Points bend down and to the right of the line.



Short tails (narrower than the normal distribution) - Points follow an S shaped-curve.



Long tails (wider than the normal distribution) - Points start below the line, bend to follow it, and end above it.

Practice More with the Table

You may practice reading the standard normal table, and use it at home.

What proportion of a standard normal population has values between -1.5 and 1.5? From the table, the proportion is 1 - 2 * 0.067 = 0.866 (i.e., the total area under the curve is one, and we subtract the upper tail area and the symmetric lower tail area).

Now go the other way. About 80% of the population lies between what two values that are centered at 0? The answer is about -1.28 and 1.28 (the table shows that 10% of the area is above 1.28, and symmetrically, 10% is below -1.28).
In this class, use the nearest value in the table. Do not use a number from your calculator!

Z transformation

– We now show how to convert a question about an arbitrary normal distribution into an equivalent question about the standard normal, and vice-versa. Thus we can use the table to answer questions about all normal distributions, not just the standard normal.

– Define a new random variable $Z = \frac{X - \mu}{\sigma}$. Then it turns out that $Z \sim N(0, 1)$. This transformation from X to Z is called the z-transformation.

– Well, this is great! To find probabilities under any normal distribution, we simply have to do the z-transformation to use the standard normal table

– To go the other way, we convert the standard normal value to an arbitrary normal distribution by solving for X. So that $X = \mu + Z\sigma$.

Reggie Jackson's IQ

Example 3: Reggie Jackson, the famous baseball player, has an IQ of 140. What percentage of people are smarter?
Assume that IQs are normally distributed with mean 100 and standard deviation 16.

We want to find $\mathbb{P}(X > 140)$ where $X \sim N(100, 16^2)$. That is, we want the area under the normal distribution for IQ that lies to the right of 140. By the z-transformation, this is equivalent to the area under the standard normal distribution that lies to the right of

$$z = \frac{X - \mu}{\sigma} = \frac{140 - 100}{16} = 2.5.$$

From the normal table, the area above 2.5 is 0.006. Thus about 0.6% of people are smarter than Reggie Jackson.

Top 2%

– Now we go the other way. We find the X value that corresponds to a given percentage.

– *Example 4:* What IQ score do you need to be in the top 2% of the IQ distribution?

– In the body of the normal table, look up 2%, or 0.02. That gives the z-value of approximately 2.05.

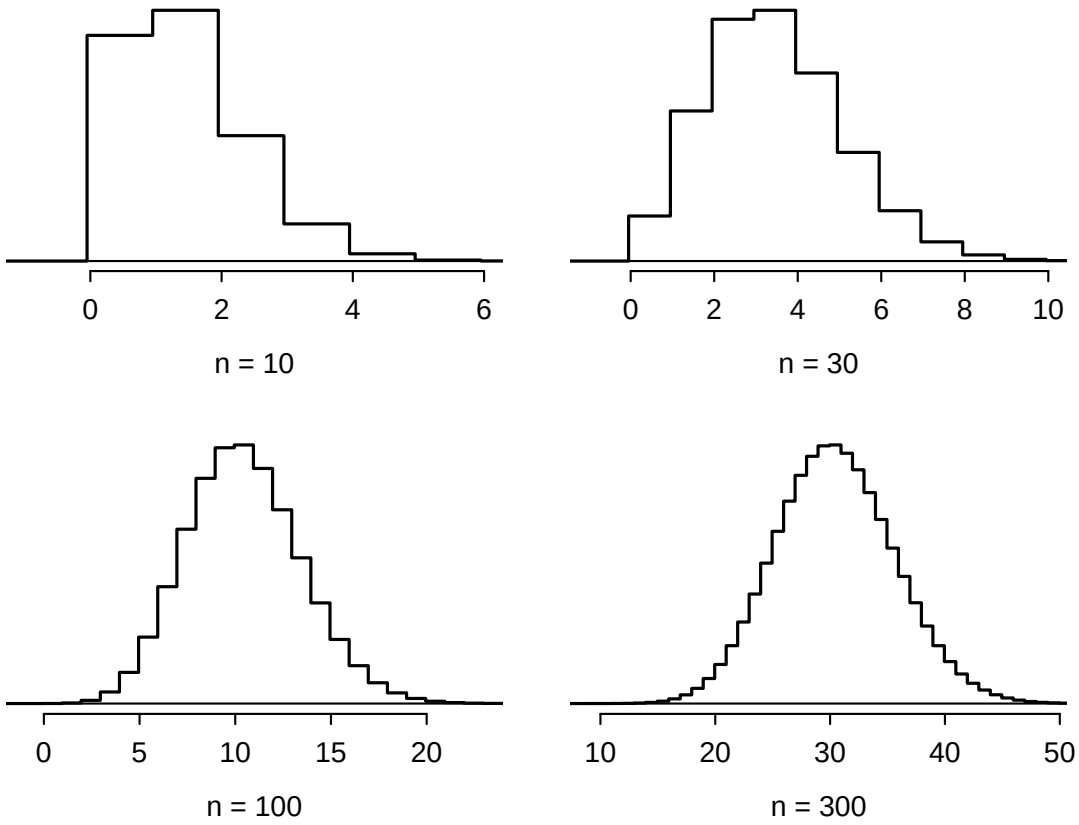
Now we use the inverse z-transformation:

$$X = \mu + Z\sigma = 100 + (2.05)(16) = 132.8.$$

One needs an IQ score of at least 132.8 (i.e., 133).

Shapes of binomial distributions

For this activity you will use a web applet. Go to *https://gallery.shinyapps.io/dist_calc/* and choose Binomial coin experiment in the drop down menu on the left. Hollow histograms of samples from the binomial model where $p = 0.10$ and $n = 10$, 30, 100, and 300. What happens as n increases?



An analysis of Facebook users

A recent study found that “Facebook users get more than they give". For example:

- 40% of Facebook users in our sample made a friend request, but 63% received at least one request
- Users in our sample pressed the like button next to friends’ content an average of 14 times, but had their content “liked" an average of 20 times
- Users sent 9 personal messages, but received 12
- 12% of users tagged a friend in a photo, but 35% were themselves tagged in a photo

Any guesses for how this pattern can be explained?

http://www.pewinternet.org/Reports/2012/Facebook-users/Summary.aspx

An analysis of Facebook users

A recent study found that “Facebook users get more than they give". For example:

- 40% of Facebook users in our sample made a friend request, but 63% received at least one request
- Users in our sample pressed the like button next to friends’ content an average of 14 times, but had their content “liked" an average of 20 times
- Users sent 9 personal messages, but received 12
- 12% of users tagged a friend in a photo, but 35% were themselves tagged in a photo

Any guesses for how this pattern can be explained?

Power users contribute much more content than the typical user.

http://www.pewinternet.org/Reports/2012/Facebook-users/Summary.aspx

This study also found that approximately 25% of Facebook users are considered power users. The same study found that the average Facebook user has 245 friends. What is the probability that the average Facebook user with 245 friends has 70 or more friends who would be considered power users? Note any assumptions you must make.

We are given that $n = 245$, $p = 0.25$, and we are asked for the probability $P(X \geq 70)$. To proceed, we need independence, which we'll assume but could check if we had access to more Facebook data.

This study also found that approximately 25% of Facebook users are considered power users. The same study found that the average Facebook user has 245 friends. What is the probability that the average Facebook user with 245 friends has 70 or more friends who would be considered power users? Note any assumptions you must make.

We are given that $n = 245$, $p = 0.25$, and we are asked for the probability $P(X \geq 70)$. To proceed, we need independence, which we'll assume but could check if we had access to more Facebook data.

$$\begin{aligned} P(X \geq 70) &= P(X = 70 \text{ or } X = 71 \text{ or } X = 72 \text{ or } \cdots \text{ or } X = 245) \\ &= P(X = 70) + P(X = 71) + P(X = 72) + \cdots + P(X = 245) \end{aligned}$$

This study also found that approximately 25% of Facebook users are considered power users. The same study found that the average Facebook user has 245 friends. What is the probability that the average Facebook user with 245 friends has 70 or more friends who would be considered power users? Note any assumptions you must make.

We are given that $n = 245$, $p = 0.25$, and we are asked for the probability $P(X \geq 70)$. To proceed, we need independence, which we'll assume but could check if we had access to more Facebook data.

$$\begin{aligned} P(X \geq 70) &= P(X = 70 \text{ or } X = 71 \text{ or } X = 72 \text{ or } \cdots \text{ or } X = 245) \\ &= P(X = 70) + P(X = 71) + P(X = 72) + \cdots + P(X = 245) \end{aligned}$$

This seems like an awful lot of work...

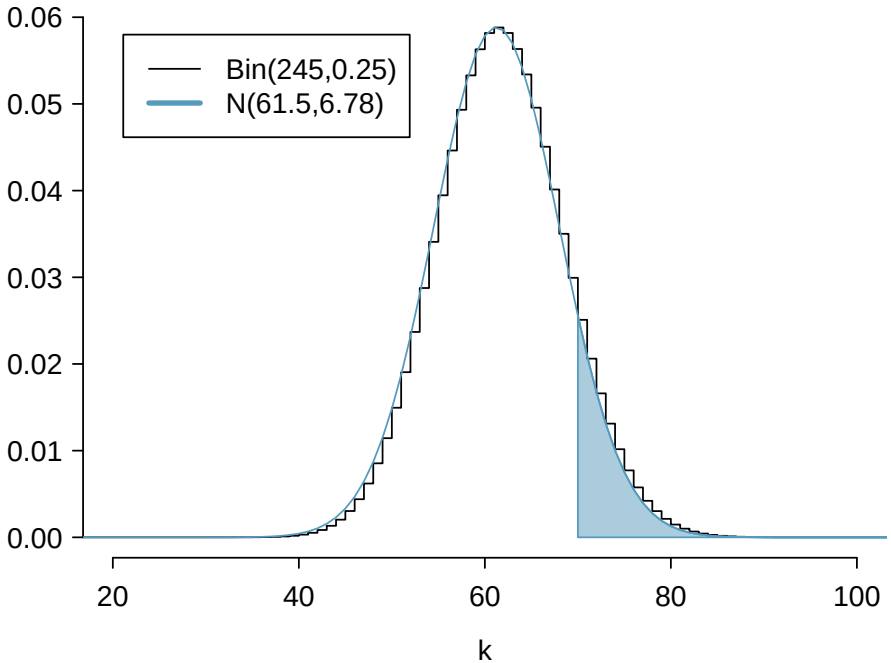
Normal approximation to the binomial

When the sample size is large enough, the binomial distribution with parameters n and p can be approximated by the normal model with parameters $\mu = np$ and $\sigma = \sqrt{np(1-p)}$.

- In the case of the Facebook power users, $n = 245$ and $p = 0.25$.

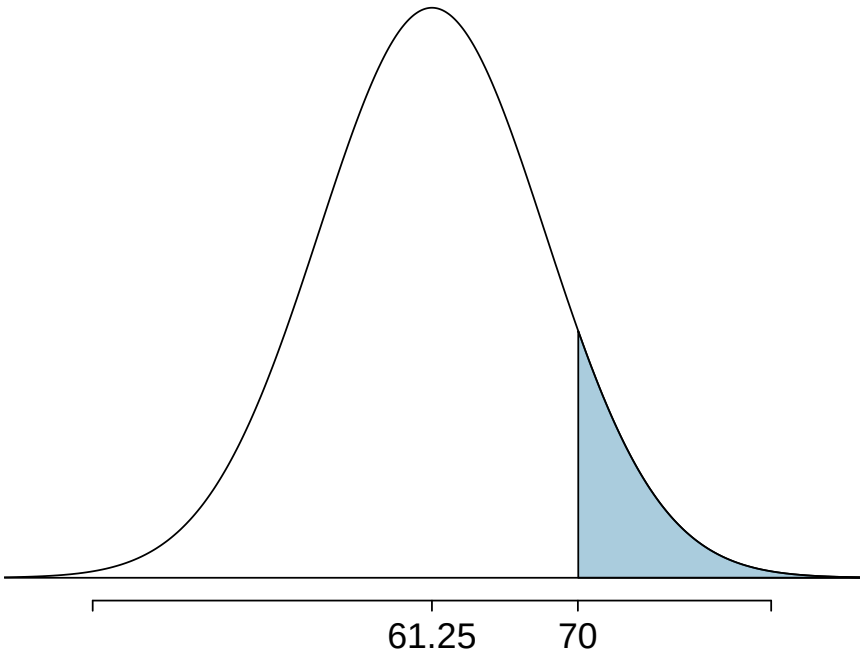
$$\mu = 245 \times 0.25 = 61.25 \qquad \sigma = \sqrt{245 \times 0.25 \times 0.75} = 6.78$$

- $Bin(n = 245, p = 0.25) \approx N(\mu = 61.25, \sigma = 6.78)$.

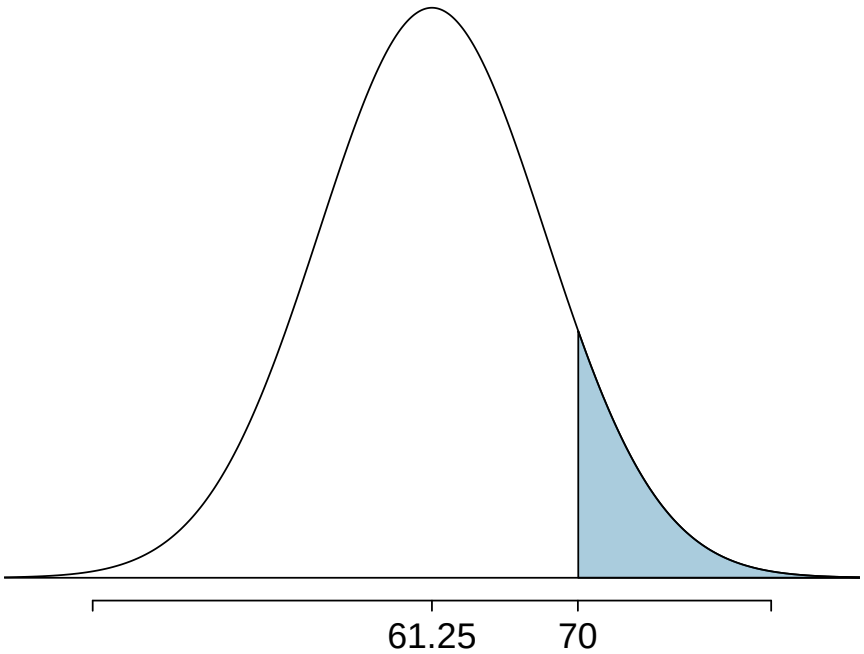


What is the probability that the average Facebook user with 245 friends has 70 or more friends who would be considered power users?

What is the probability that the average Facebook user with 245 friends has 70 or more friends who would be considered power users?

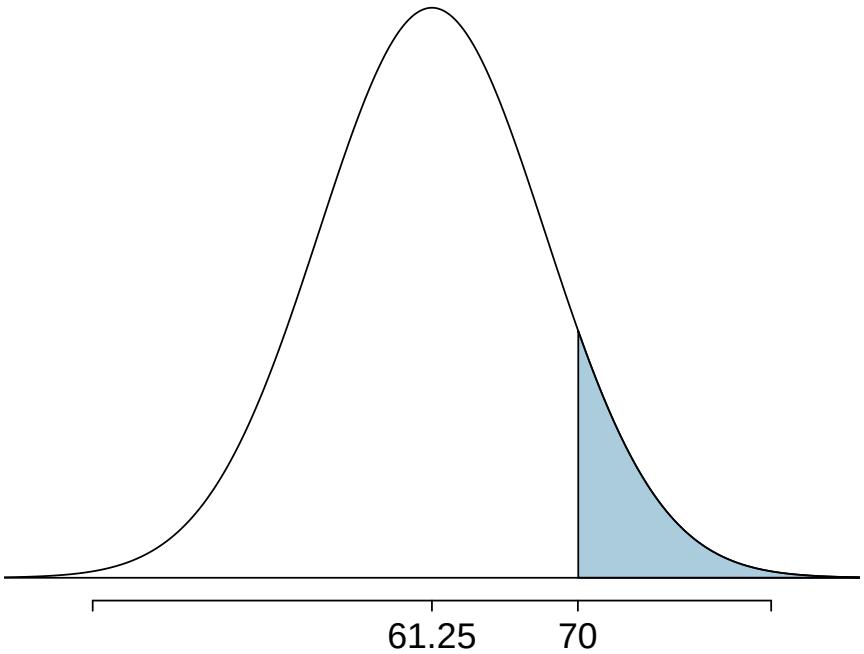


What is the probability that the average Facebook user with 245 friends has 70 or more friends who would be considered power users?



$$Z = \frac{obs - mean}{SD} = \frac{70 - 61.25}{6.78} = 1.29$$

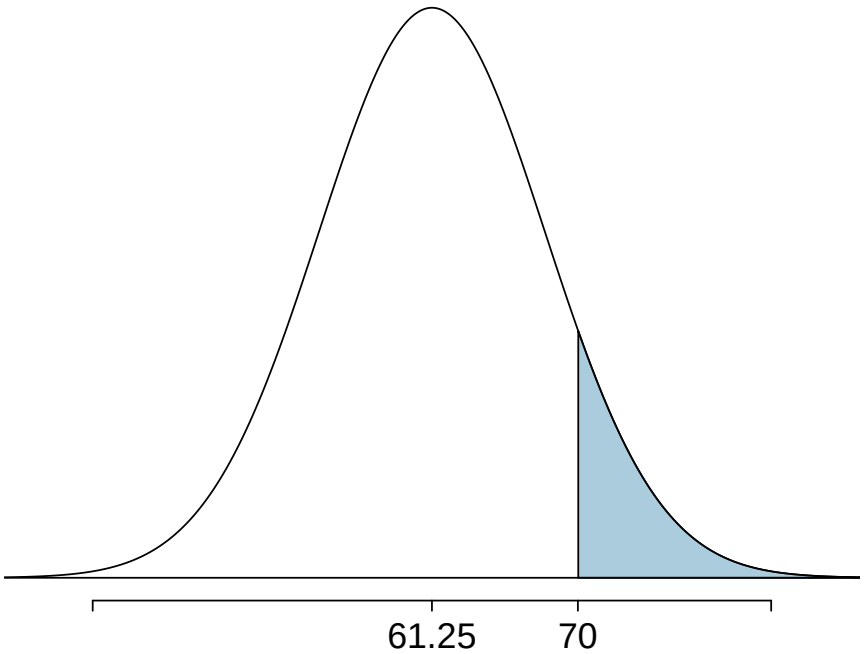
What is the probability that the average Facebook user with 245 friends has 70 or more friends who would be considered power users?



$$Z = \frac{obs - mean}{SD} = \frac{70 - 61.25}{6.78} = 1.29$$

	Second decimal place of Z				
Z	0.05	0.06	0.07	0.08	0.09
1.0	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8944	0.8962	0.8980	0.8997	*0.9015*

What is the probability that the average Facebook user with 245 friends has 70 or more friends who would be considered power users?



$$Z = \frac{obs - mean}{SD} = \frac{70 - 61.25}{6.78} = 1.29$$

	Second decimal place of Z				
Z	0.05	0.06	0.07	0.08	0.09
1.0	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8944	0.8962	0.8980	0.8997	*0.9015*

$$P(Z > 1.29) = 1 - 0.9015 = 0.0985$$

Continuity Correction

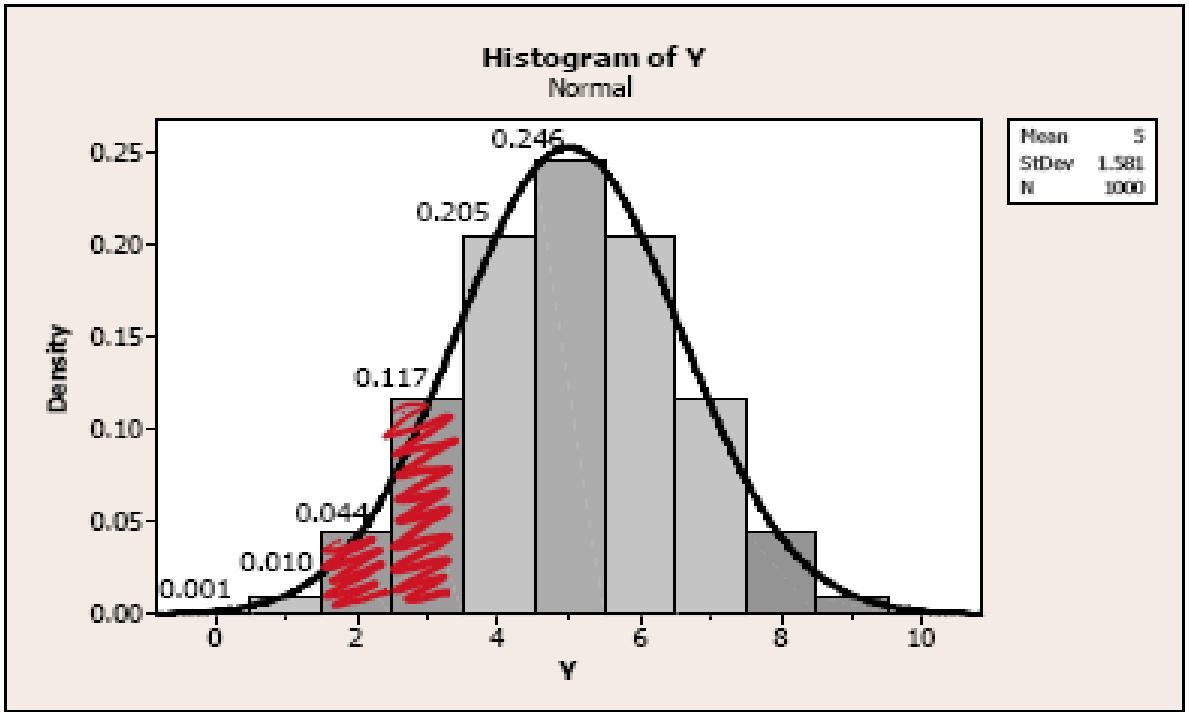
– A normal distribution describes data that can take any possible value (integers, fractions, irrationals, etc.), but often data can only take non-negative integer values.

– In a class of twelve students, each tosses a fair coin to decide whether to attend class. So class attendance is a random variable that has the Bin(12, 0.5) distribution. Its mean is $np = 6$ and the standard deviation is $\sqrt{n * p * (1-p)} = 1.732$.

– We can use the normal distribution to estimate the approximate probability that, say, 3 or fewer students will attend tomorrow's lecture. But because only integers are possible, we can improve the accuracy of the normal approximation by using the **continuity correction**.

Continuity Correction (Cont'd)

– We approximate the binomial by a normal distribution with the same mean and standard deviation.



– The bad approximation uses the z-transformation $z = (3-6)/1.732 = -1.732$, and finds the area under the N(0,1) curve that lies below -1.732 as 0.0418.

Continuity Correction (Cont'd)

– The good way handles the area between 3 and 4 appropriately, to take account of the fact that the histogram bar is centered at 3 and we want to include the area up to 3.5. We use the z-transformation $z = (\mathbf{3.5} - 6)/1.732 = -1.443$, and find the probability as 0.0749.

– The normal approximation to the binomial is helpful when n is very large. For example, suppose we wanted to find the probability that more than 20,000 of the 228,330 residents of Durham are unemployed, when the unemployment rate in NC is 10.1%. To use the binomial, we would have to calculate

$$1 - \sum_{x=0}^{20{,}000} \binom{228,330}{x} (0.101)^x (1 - 0.101)^{228{,}300 - x}.$$

This is intractable, but the normal approximation is not.

– The normal approximation is accurate when $np > 10$ and $n(1-p) > 10$.

Recap

We discussed the following:

- The normal distribution
- Normal approximation for the binomial distribution

Suggested reading:

- D.S. Sec. 5.6
- OpenIntro3: Sec. 2.5, 3.1, 3.2, 3.4.2, 3.4.3