CAAP Statistics - Lec08

Jul 18, 2022

Review

- Definition of Probability
  - Frequentist vs Bayesian
  - Law of Large Numbers
- Probability Distribution
  - Independent vs Disjoint vs Complement
  - Product Rule: $P(A \cap B) = P(A) \times P(B)$
  - Addition Rule: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- Sampling from a small population
  - Sampling with & without replacement
- Random Variable
  - Expectation, Variance
- Continuous Distribution

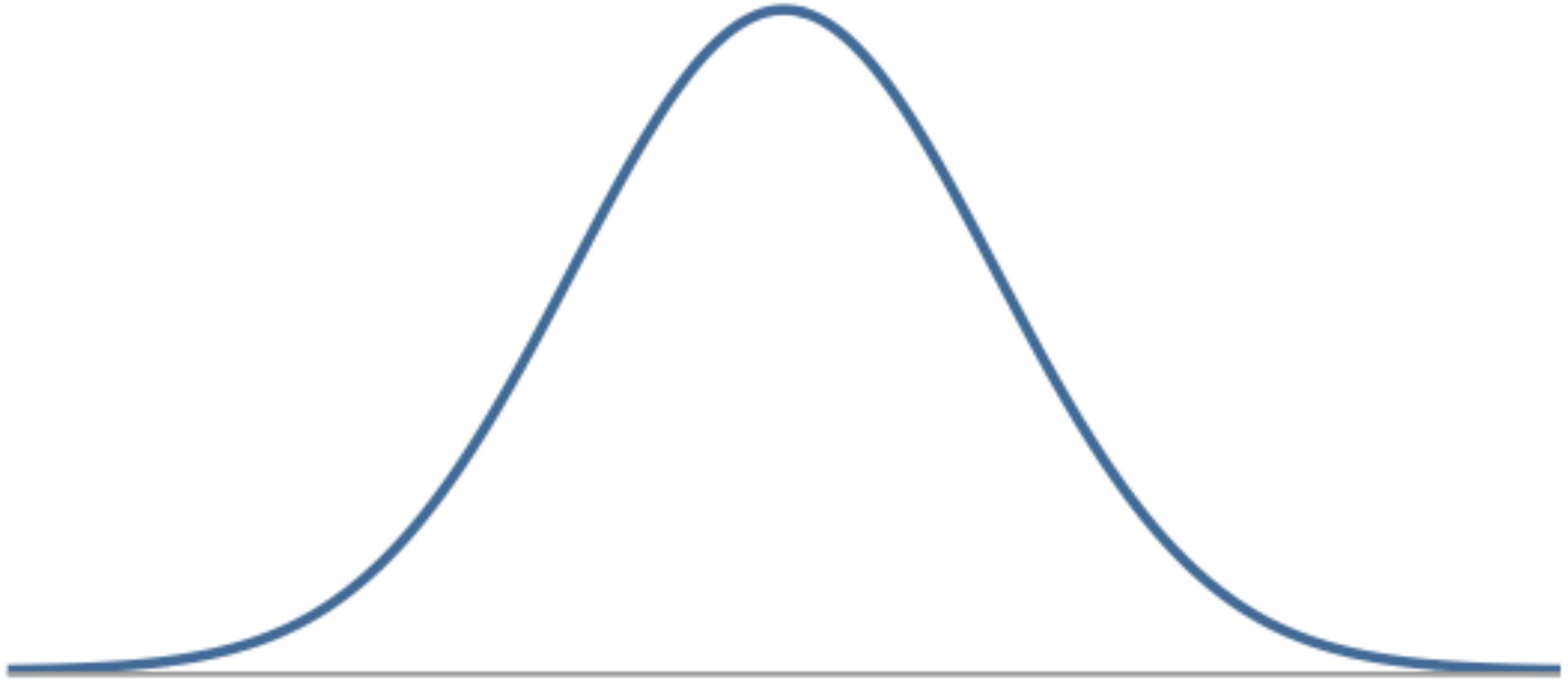
Learning Objectives

- Random Variables
  - Expectation and Variance
- Normal distribution
- Bernoulli distribution
- Binomial distribution
- Poisson distribution(if time allows)

Normal distribution

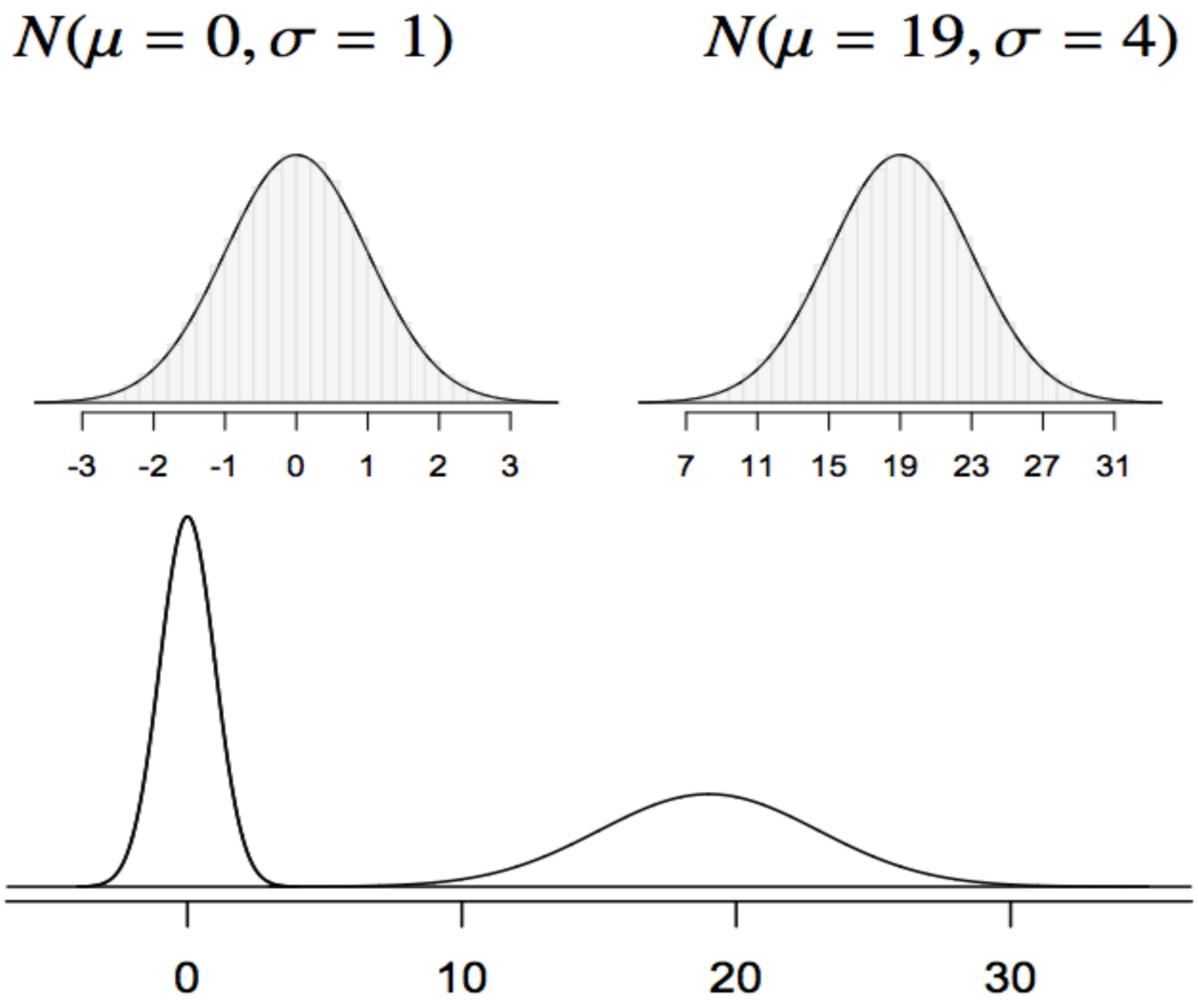
Normal Distribution

- 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

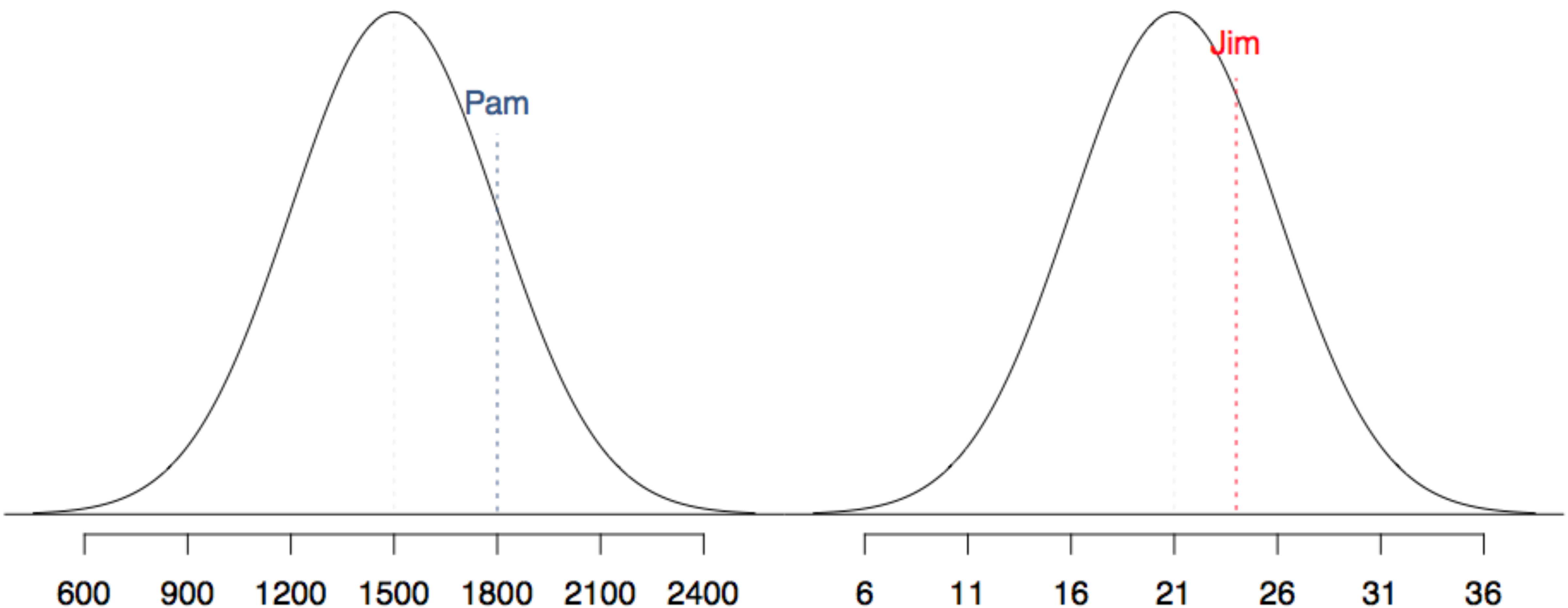


Normal distributions with different parameters

μ: mean, σ: standard deviation



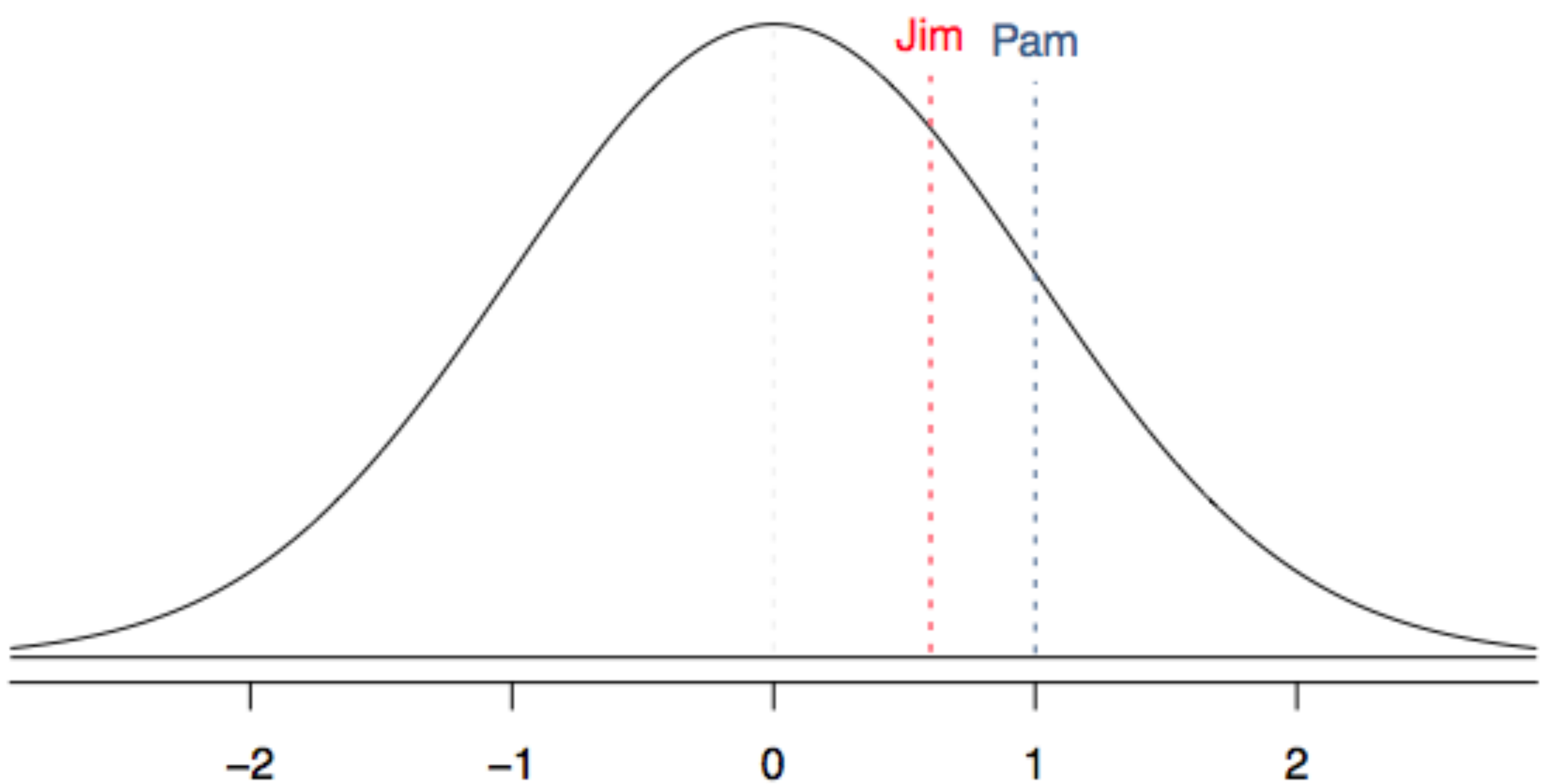
SAT scores are distributed nearly normally with mean 1500 and standard deviation 300. ACT scores are distributed nearly normally with mean 21 and standard deviation 5. A college admissions officer wants to determine which of the two applicants scored better on their standardized test with respect to 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*.

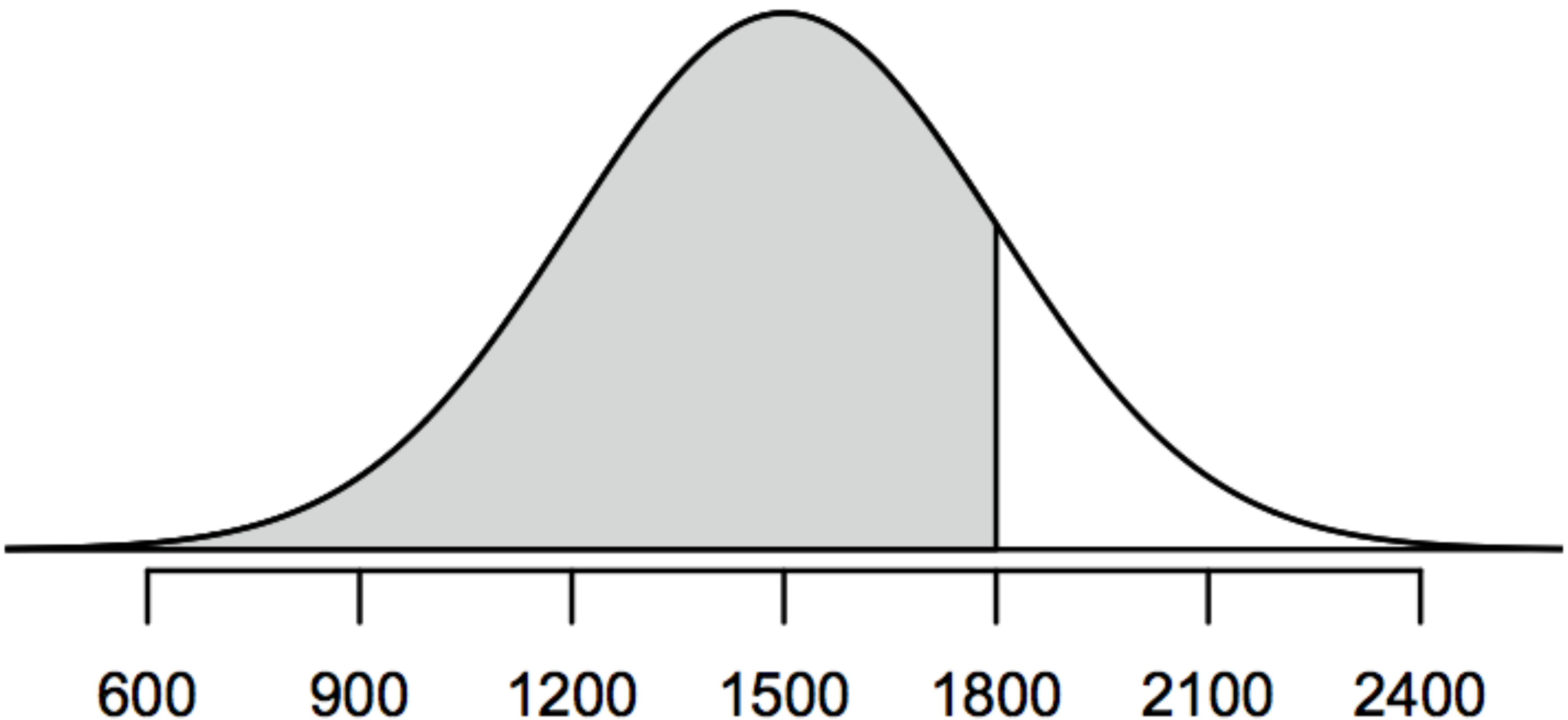
- Z score of an observation is the number of standard deviations it falls above or below the mean.

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

- 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.

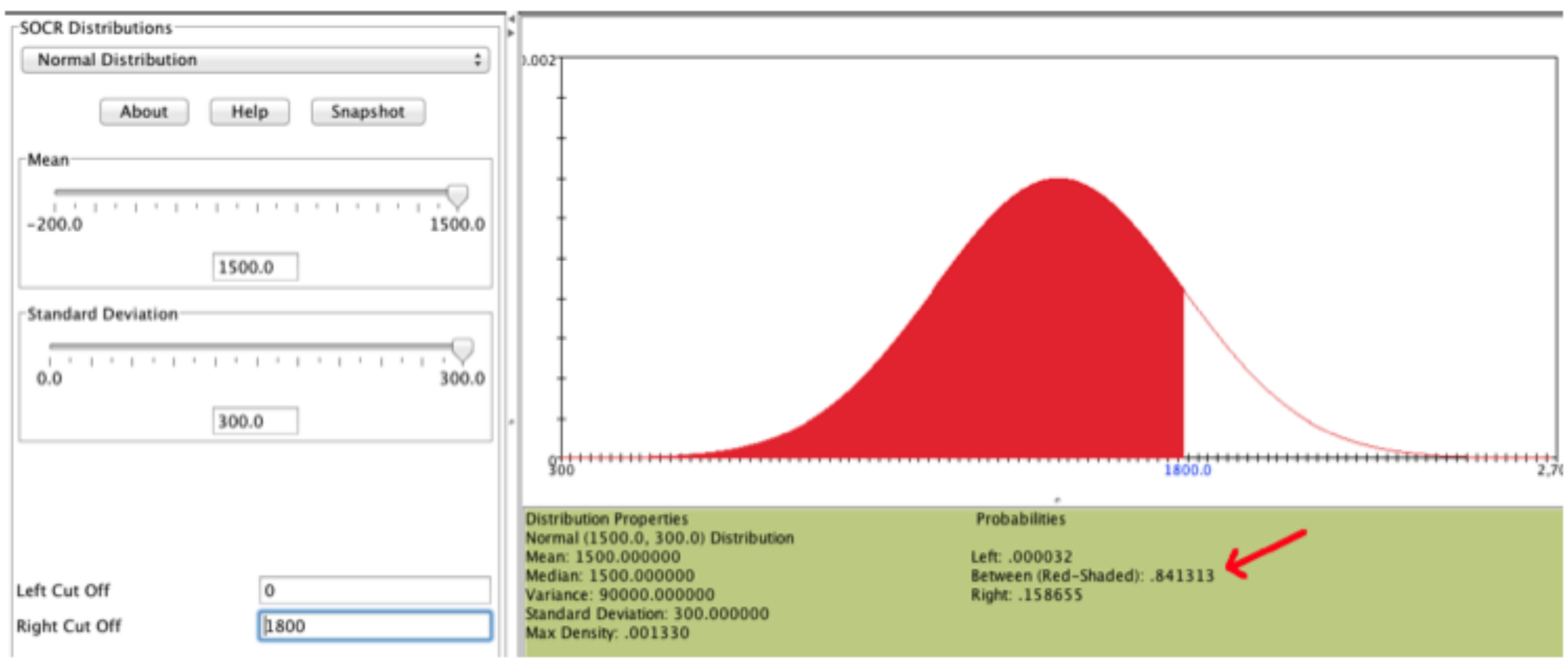


Calculating percentiles - using computation

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

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

Applet: www.socr.ucla.edu/htmls/SOCR_Distributions.html



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

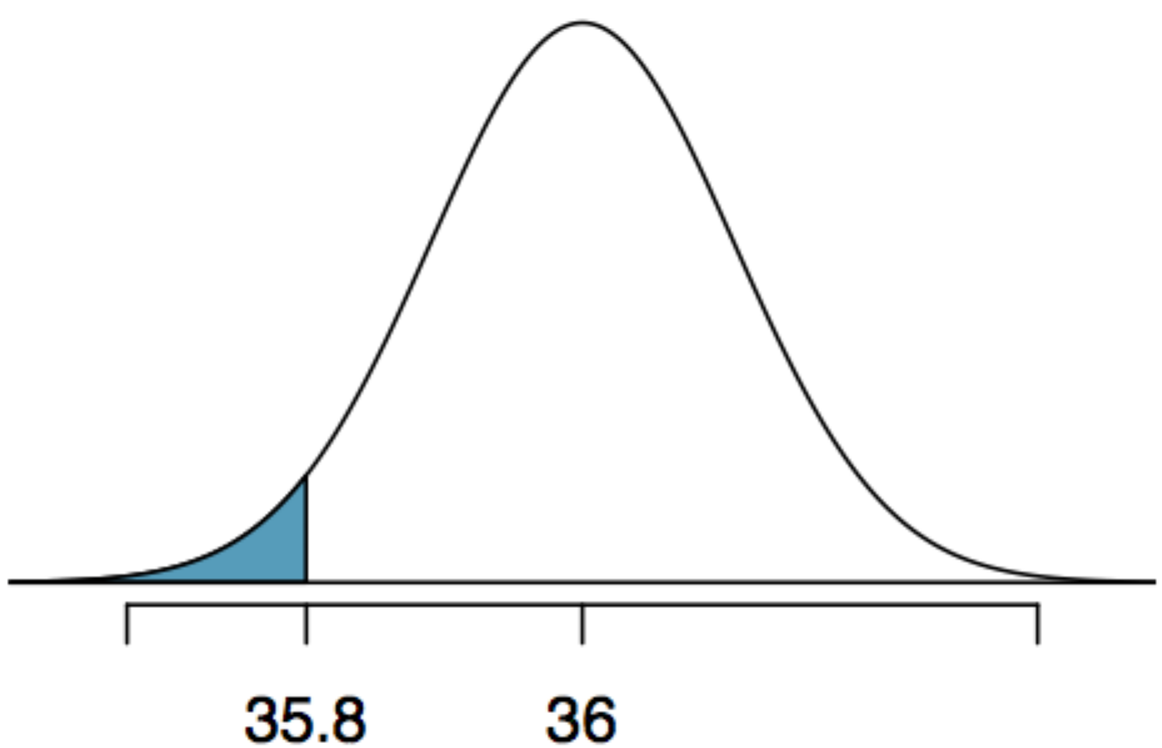
At Heinz ketchup factory the amounts which go into bottles of ketchup are supposed to be normally distributed with mean 36 oz. and standard deviation 0.11 oz. Once every 30 minutes a bottle is selected from the production line, and its contents are noted precisely. If the amount of ketchup in the bottle 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?

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

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?

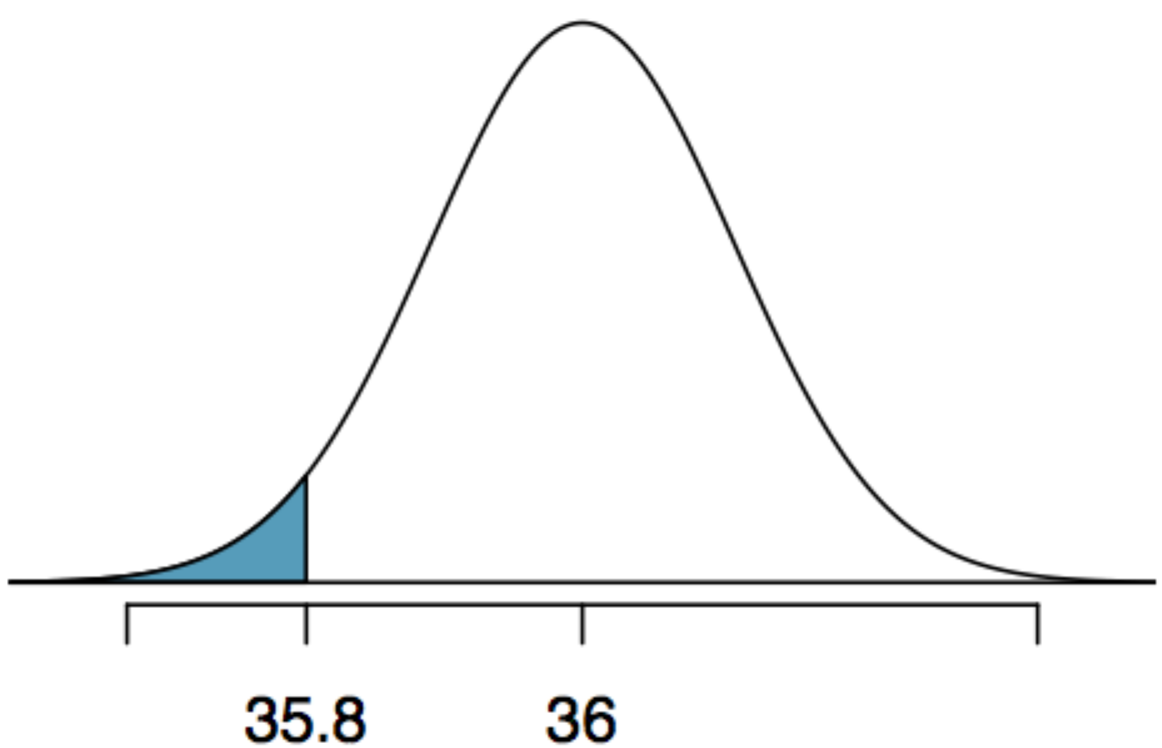
- *Let X = amount of ketchup in a bottle:* $X \sim N(\mu = 36, \sigma = 0.11)$



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 R

```
> pnorm(-1.82)
[1] 0.0344
```

OR

```
> pnorm(35.8, mean = 36, sd = 0.11)
[1] 0.0345
```

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

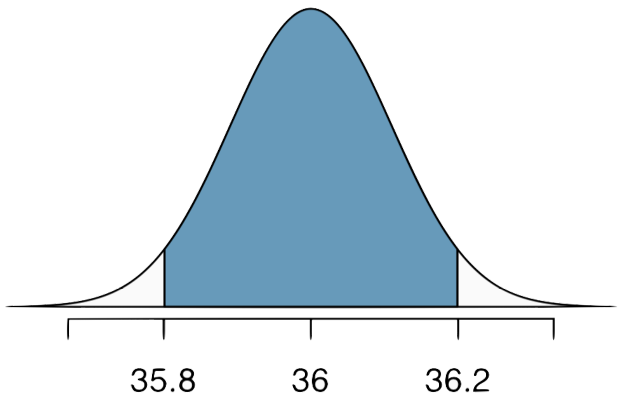
What percent of bottles pass the quality control inspection?

(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%

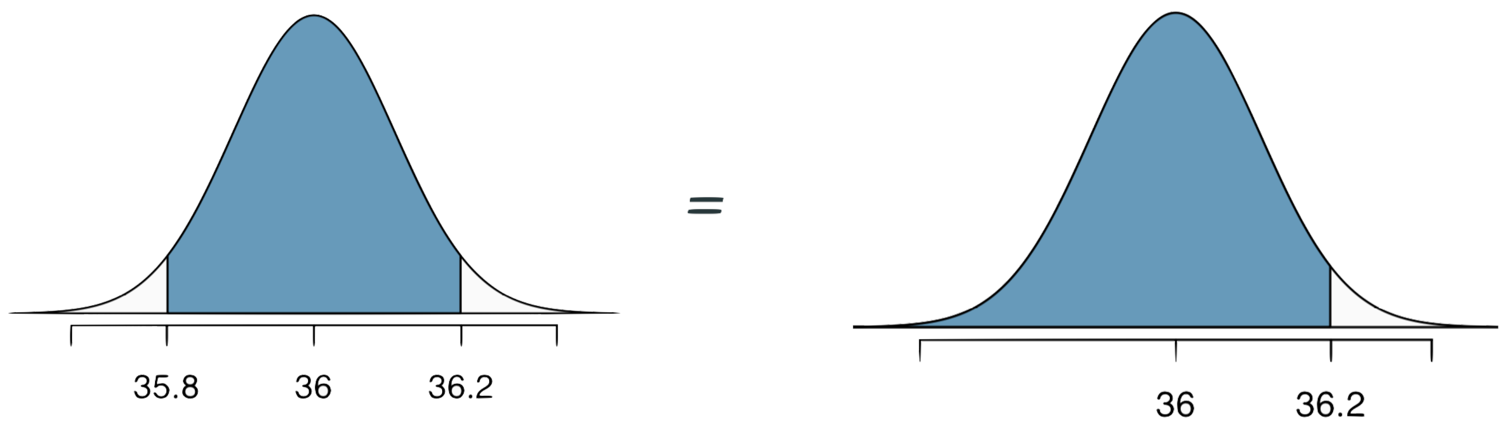


=

Practice

What percent of bottles pass the quality control inspection?

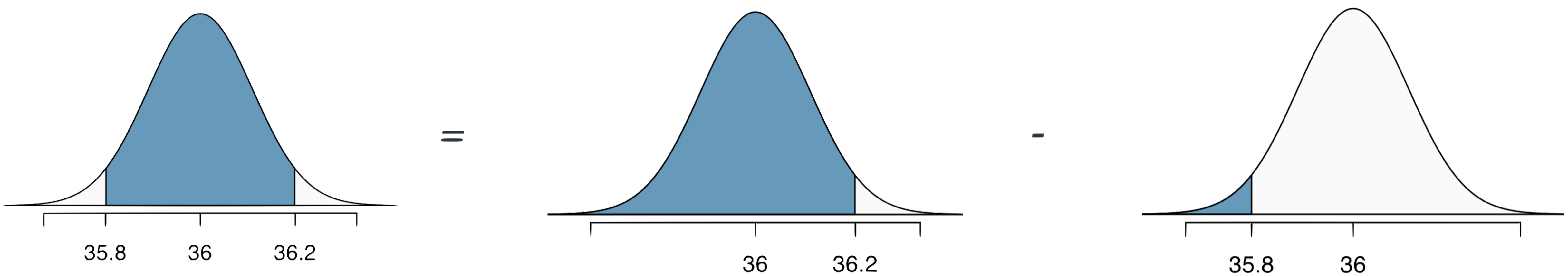
(a) 1.82%
(b) 3.44%
(c) 6.88%
(d) 93.12%
(e) 96.56%



Practice

What percent of bottles pass the quality control inspection?

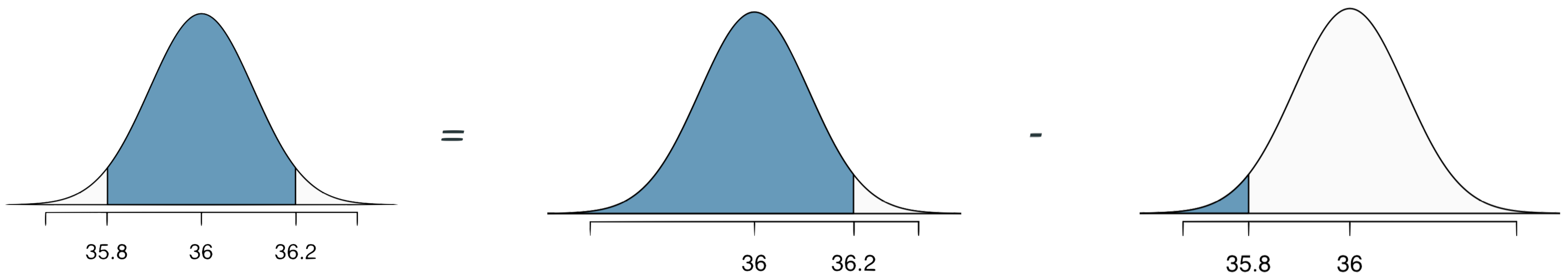
(a) 1.82%
(b) 3.44%
(c) 6.88%
(d) 93.12%
(e) 96.56%



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%

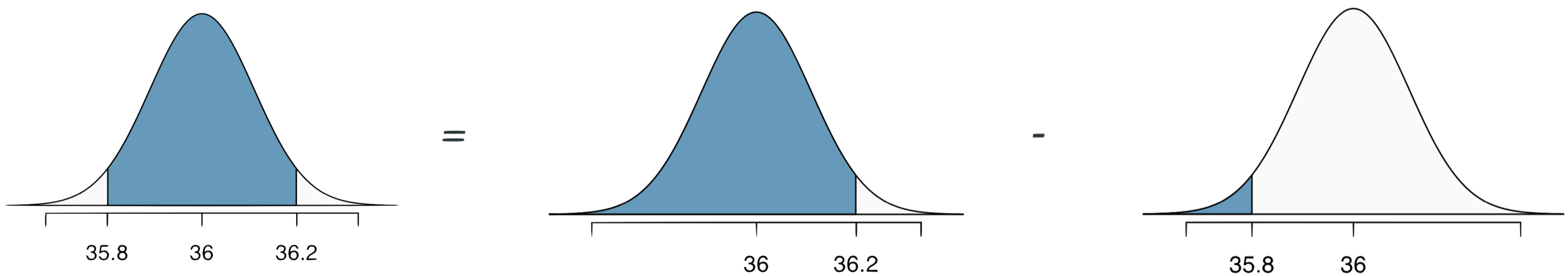


$$Z_{35.8} = \frac{35.8 - 36}{0.11} = -1.82$$

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%



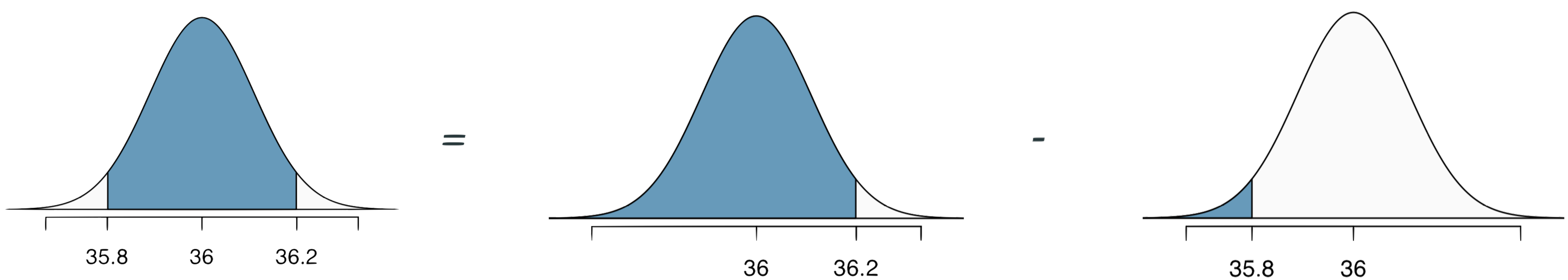
$$Z_{35.8} = \frac{35.8 - 36}{0.11} = -1.82$$

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

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%



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

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

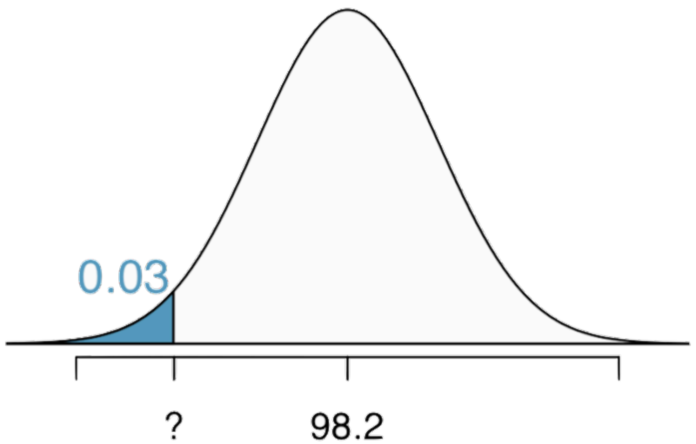
$$P(35.8 < X < 36.2) = P(-1.82 < Z < 1.82) = 0.9656 - 0.0344 = 0.9312$$

Finding cutoff points

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

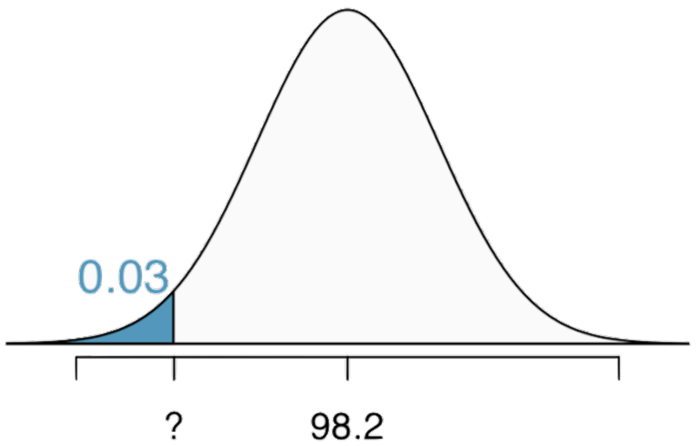
Finding cutoff points

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



Finding cutoff points

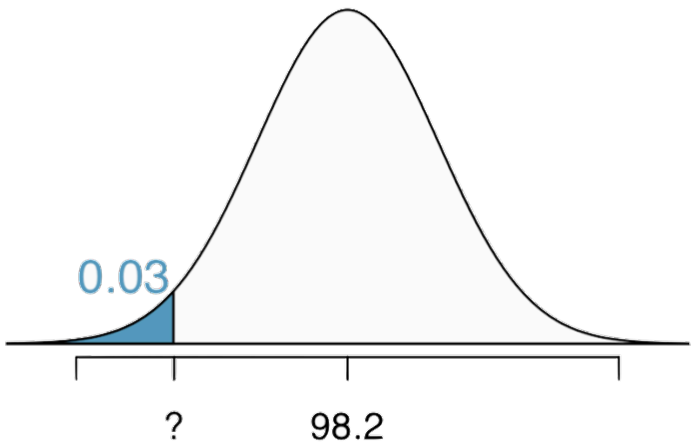
Body temperatures of healthy humans are distributed nearly normally with mean 98.2^{o}F and standard deviation 0.73^{o}F. What is the cutoff for the lowest 3% of human body temperatures?



$$P(X < x) \quad = \quad 0.03 \rightarrow P(Z < -1.88) = 0.03$$

Finding cutoff points

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

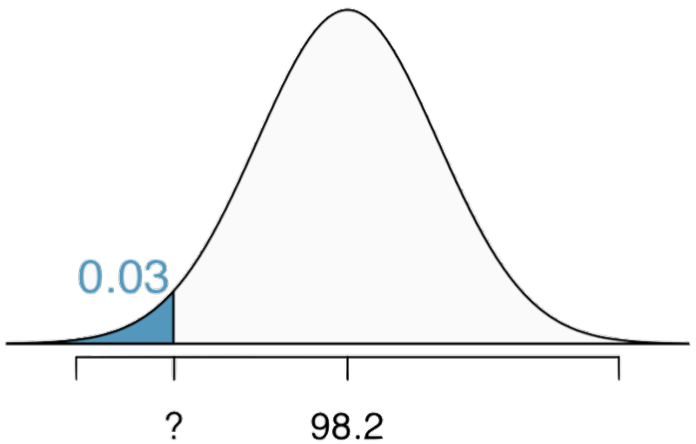


$$P(X < x) = 0.03 \rightarrow P(Z < -1.88) = 0.03$$

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

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?



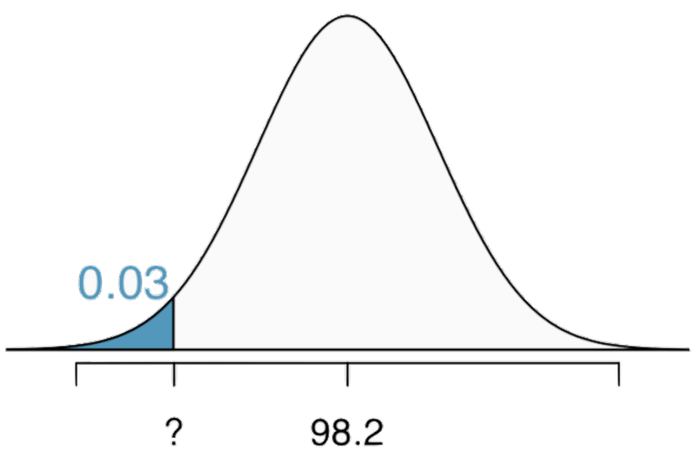
$$P(X < x) = 0.03 \rightarrow P(Z < -1.88) = 0.03$$

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

$$x = (-1.88 \times 0.73) + 98.2 = 96.8^{\circ}F$$

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?



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

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

$$x = (-1.88 \times 0.73) + 98.2 = 96.8^\circ F$$

```
> qnorm(0.03)
[1] -1.880794
```


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 nor- mally with mean 98.2^{o}F and standard deviation 0.73^{o}F. What is the cutoff for the highest 10% of human body temperatures?

(a) 97.3^{o}F
(b) 99.1^{o}F
(c) 99.4^{o}F
(d) 99.6^{o}F

Practice

Body temperatures of healthy humans are distributed nearly nor- mally with mean 98.2^{o}F and standard deviation 0.73^{o}F. What is the cutoff for the highest 10% of human body temperatures?

(a) 97.3^{o}F

(b) 99.1^{o}F

(c) 99.4^{o}F

(d) 99.6^{o}F

Practice

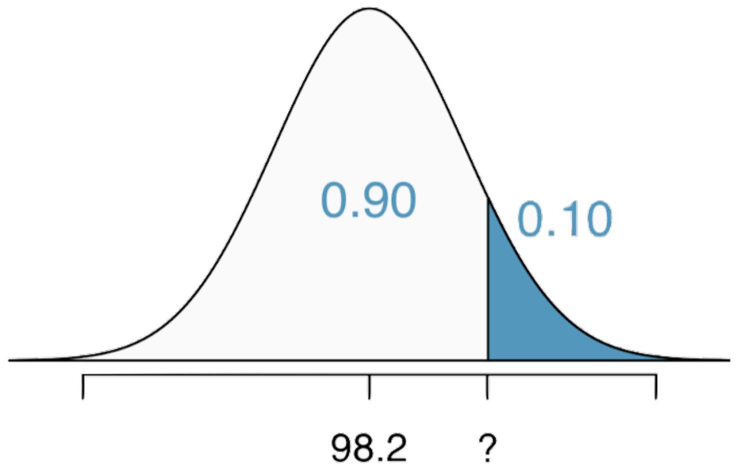
Body temperatures of healthy humans are distributed nearly nor- mally with mean 98.2^{o}F and standard deviation 0.73^{o}F. What is the cutoff for the highest 10% of human body temperatures?

(a) 97.3^{o}F

(b) 99.1^{o}F

(c) 99.4^{o}F

(d) 99.6^{o}F



Practice

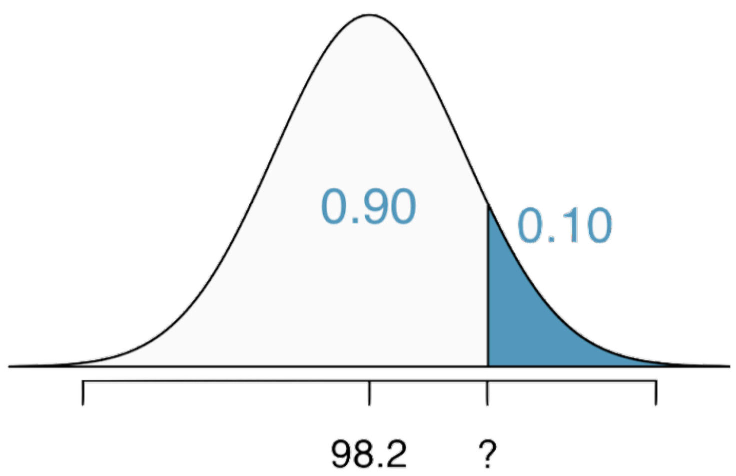
Body temperatures of healthy humans are distributed nearly nor- mally with mean 98.2°F and standard deviation 0.73°F. What is the cutoff for the highest 10% of human body temperatures?

(a) 97.3°F

(b) 99.1°F

(c) 99.4°F

(d) 99.6°F



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

Practice

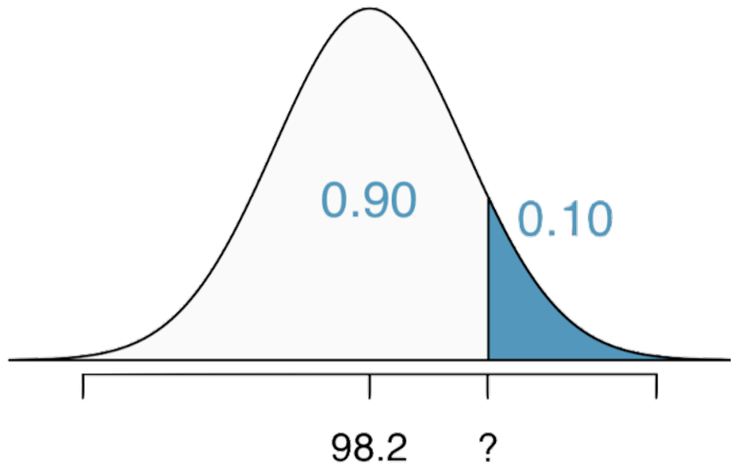
Body temperatures of healthy humans are distributed nearly nor- mally with mean 98.2^{o}F and standard deviation 0.73^{o}F. What is the cutoff for the highest 10% of human body temperatures?

(a) 97.3^{o}F

(b) 99.1^{o}F

(c) 99.4^{o}F

(d) 99.6^{o}F



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

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

Practice

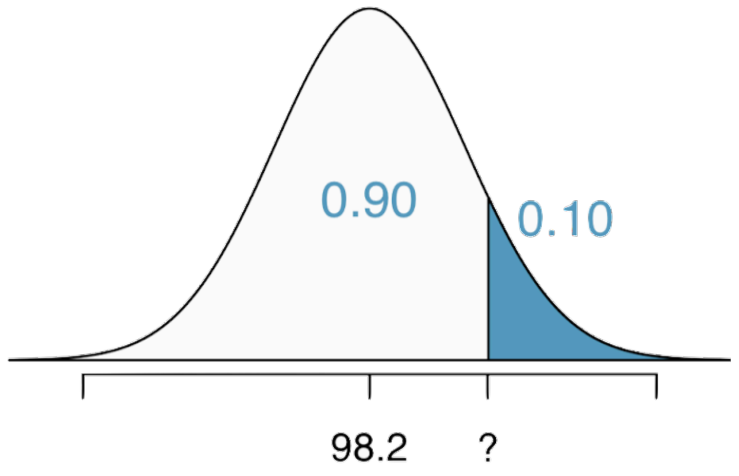
Body temperatures of healthy humans are distributed nearly nor- mally with mean 98.2^{o}F and standard deviation 0.73^{o}F. What is the cutoff for the highest 10% of human body temperatures?

(a) 97.3^{o}F

(b) 99.1^{o}F

(c) 99.4^{o}F

(d) 99.6^{o}F



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

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

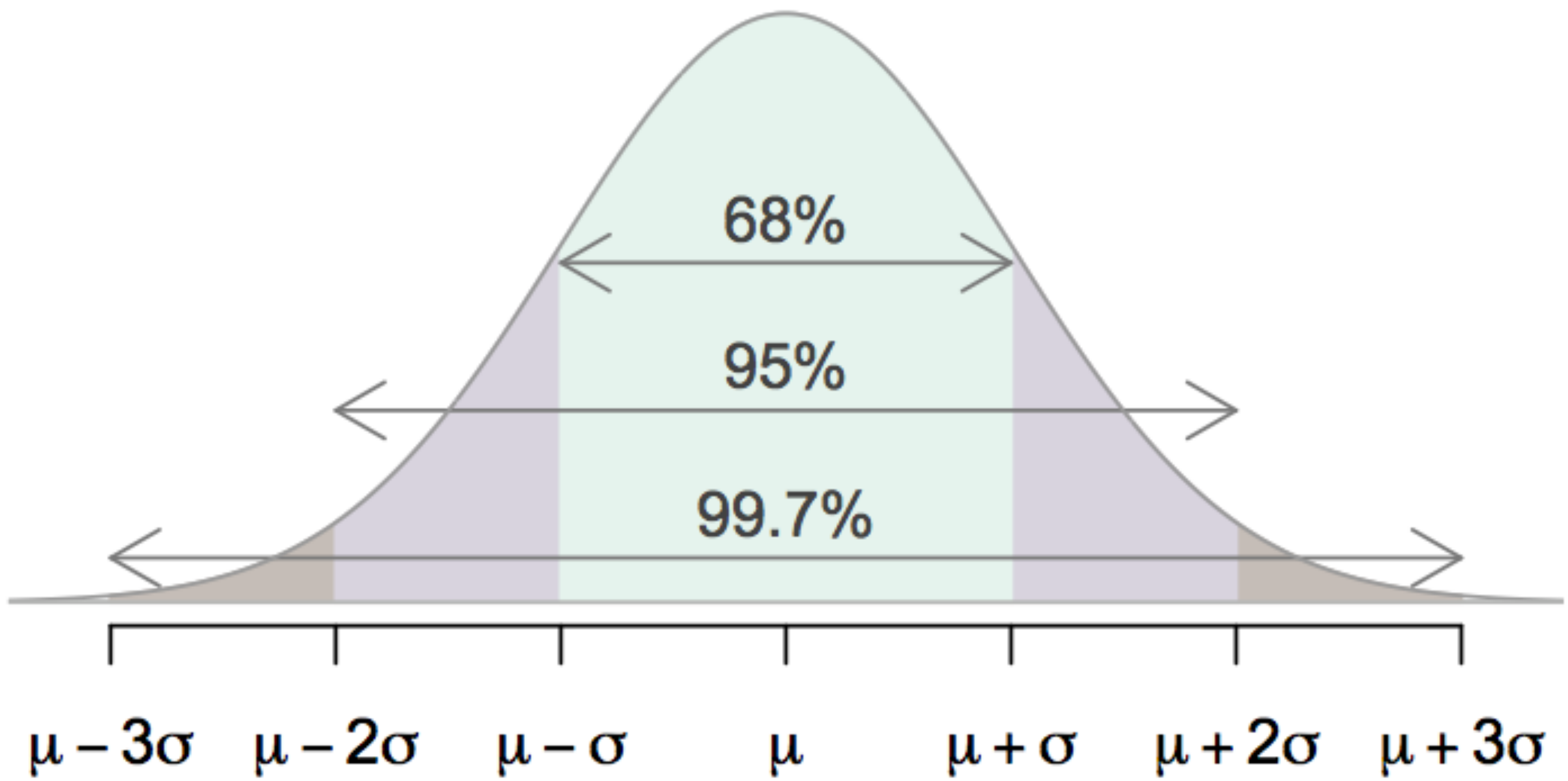
$$x = (1.28 \times 0.73) + 98.2 = 99.1$$

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.



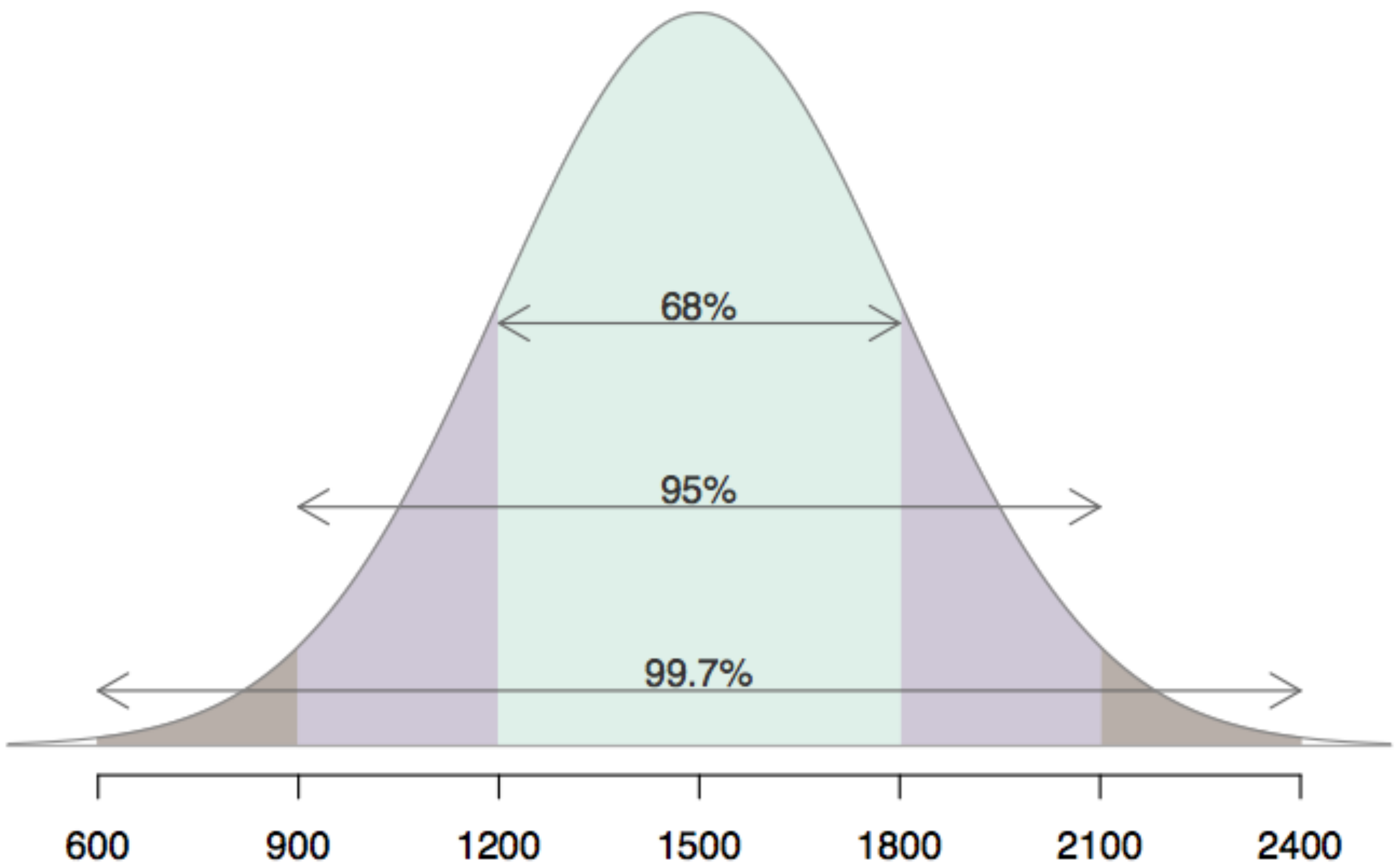
Describing variability using the 68-95-99.7 Rule

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

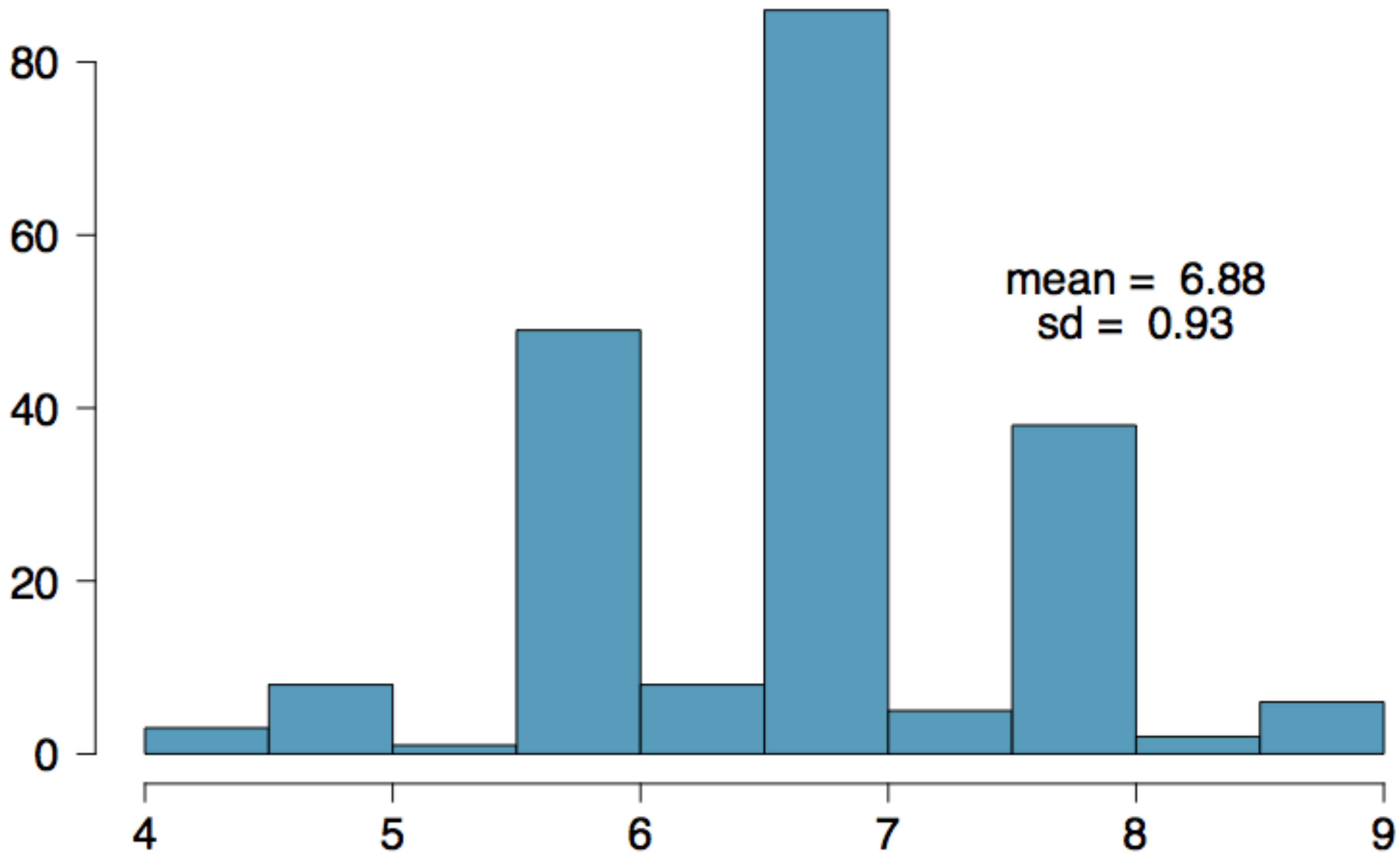
Describing variability using the 68-95-99.7 Rule

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.

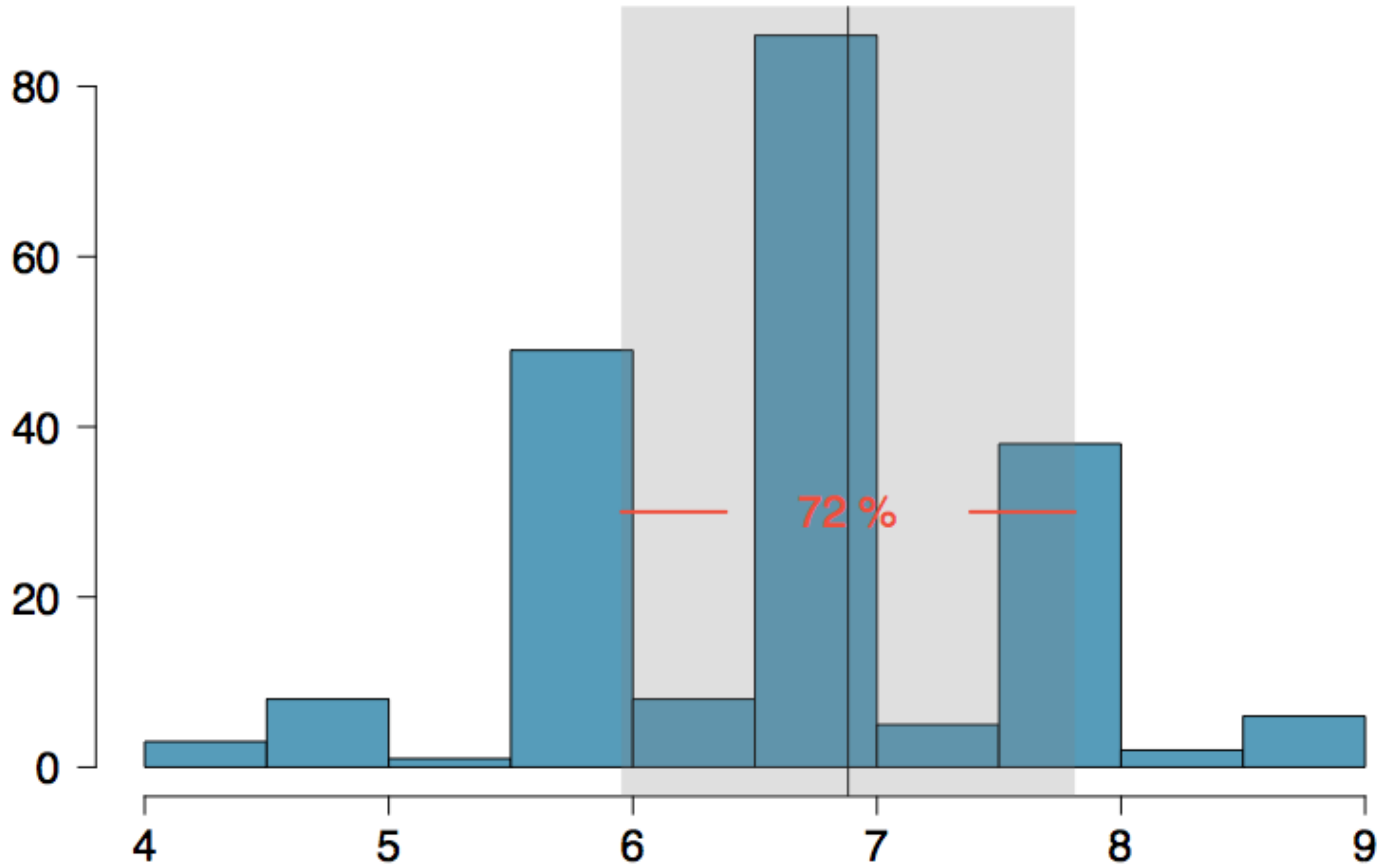


Number of hours of sleep on school nights



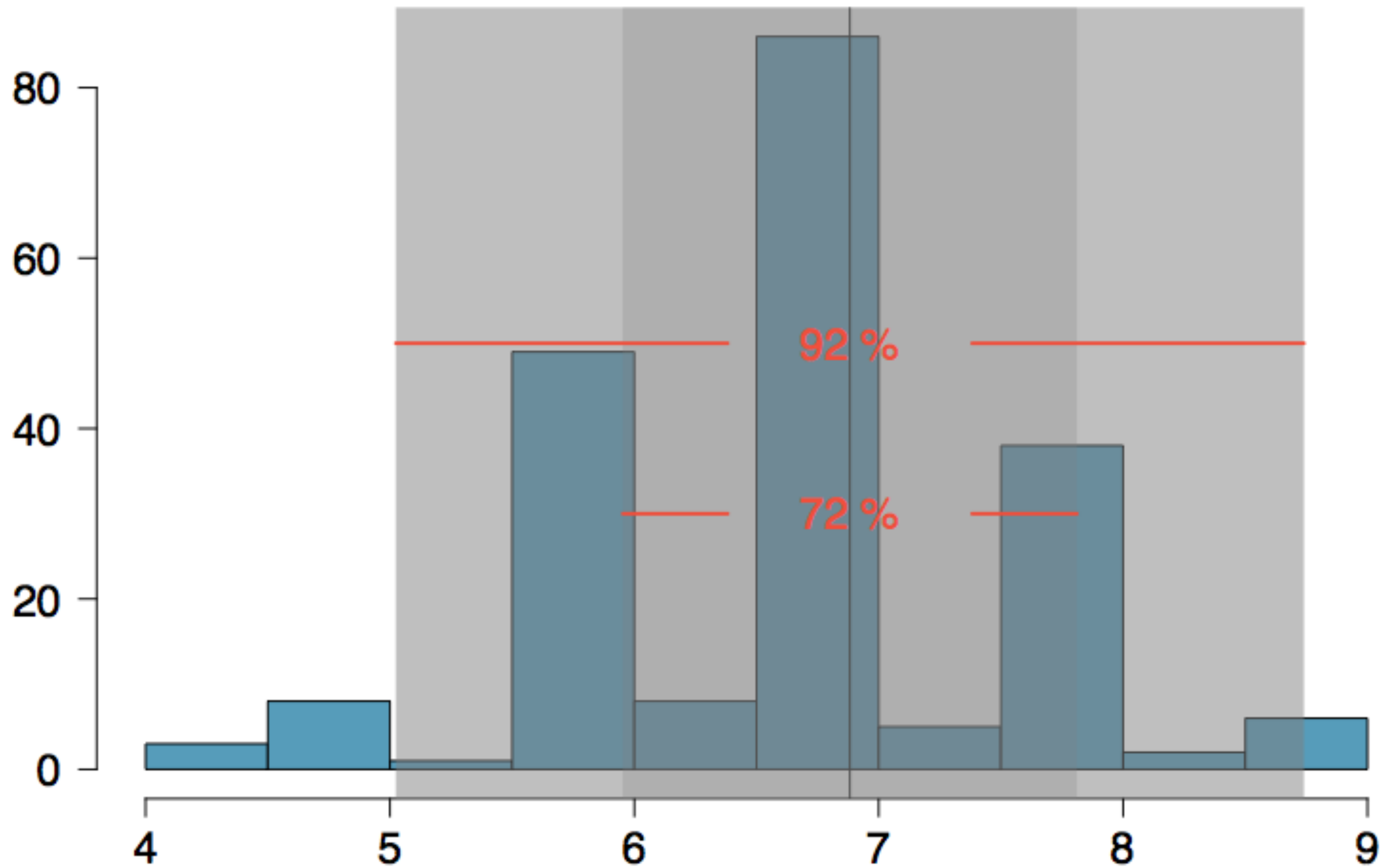
- Mean = 6.88 hours, SD = 0.92 hrs

Number of hours of sleep on school nights



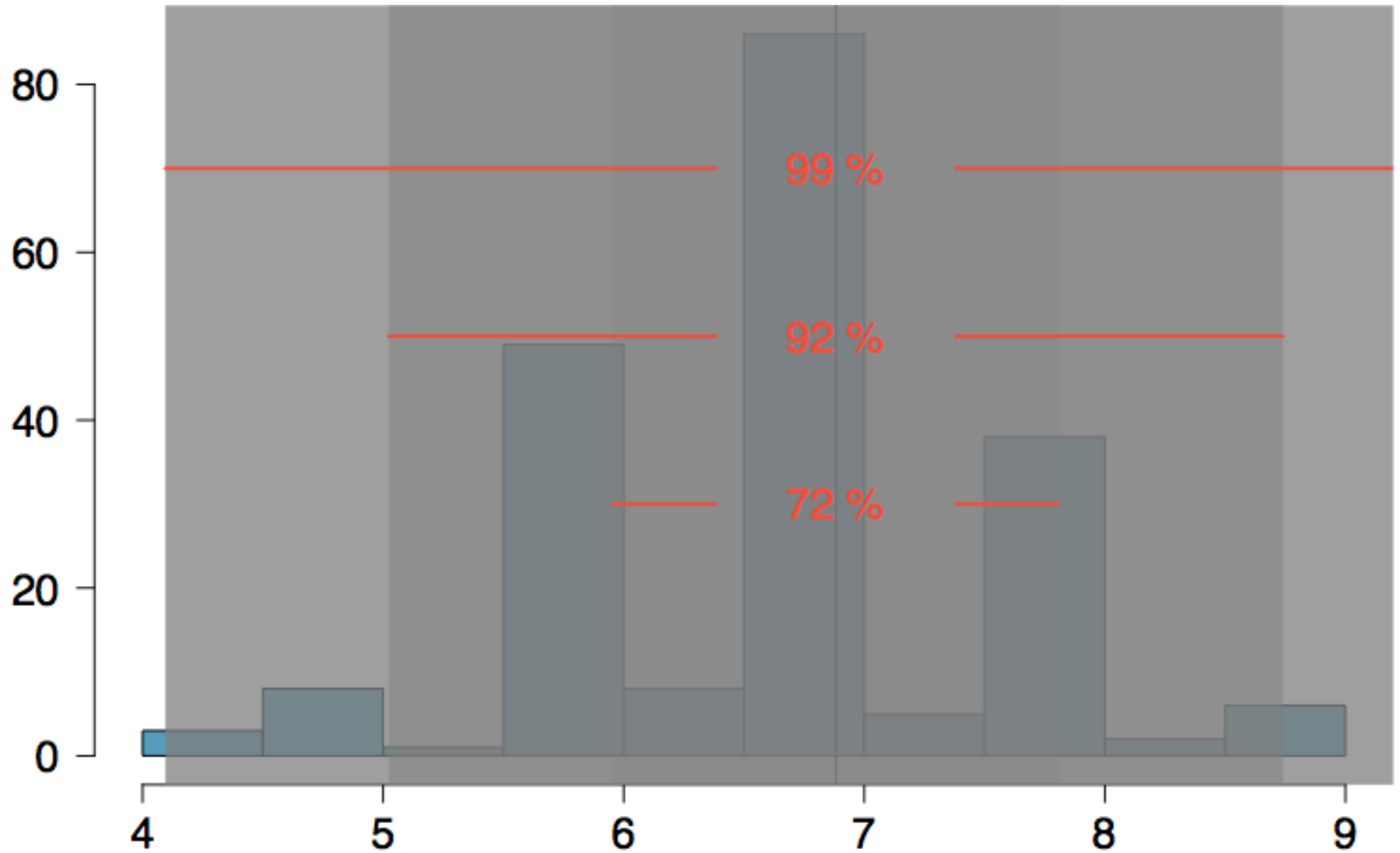
- Mean = 6.88 hours, SD = 0.92 hrs
- 72% of the data are within 1 SD of the mean: 6.88 ± 0.93

Number of hours of sleep on school nights



- Mean = 6.88 hours, SD = 0.92 hrs
- 72% of the data are within 1 SD of the mean: 6.88 ± 0.93
- 92% of the data are within 1 SD of the mean: 6.88 ± 2 x 0.93

Number of hours of sleep on school nights



- Mean = 6.88 hours, SD = 0.92 hrs
- 72% of the data are within 1 SD of the mean: 6.88 ± 0.93
- 92% of the data are within 1 SD of the mean: 6.88 ± 2 x 0.93
- 99% of the data are within 1 SD of the mean: 6.88 ± 3 x 0.93

Practice

Which of the following is false?

A. Majority of Z scores in a right skewed distribution are negative.
B. In skewed distributions the Z score of the mean might be different than 0.
C. For a normal distribution, IQR is less than 2 x SD.
D. Z scores are helpful for determining how unusual a data point is compared to the rest of the data in the distribution.

Practice

Which of the following is false?

A. Majority of Z scores in a right skewed distribution are negative.
B. In skewed distributions the Z score of the mean might be different than 0.
C. For a normal distribution, IQR is less than 2 x SD.
D. Z scores are helpful for determining how unusual a data point is compared to the rest of the data in the distribution.

Appendix

Probability Tables

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

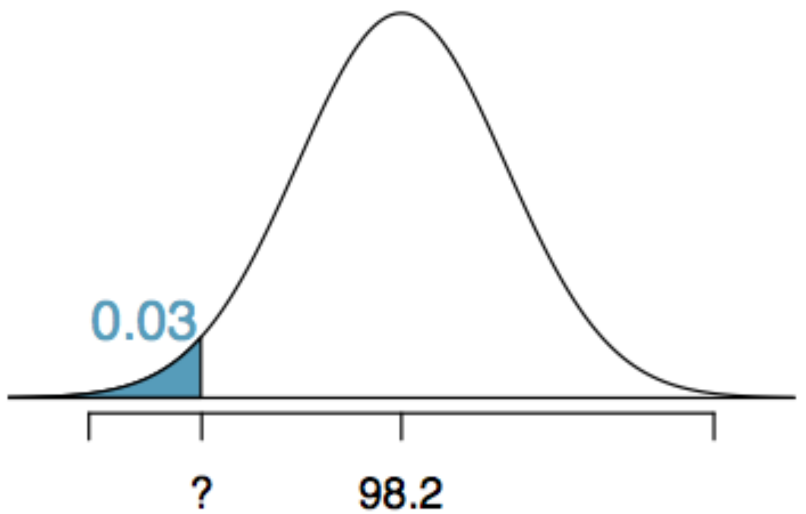
Second decimal place of Z										
0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01	0.00	Z
0.0014	0.0014	0.0015	0.0015	0.0016	0.0016	0.0017	0.0018	0.0018	0.0019	−2.9
0.0019	0.0020	0.0021	0.0021	0.0022	0.0023	0.0023	0.0024	0.0025	0.0026	−2.8
0.0026	0.0027	0.0028	0.0029	0.0030	0.0031	0.0032	0.0033	0.0034	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

Finding cutoff points

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

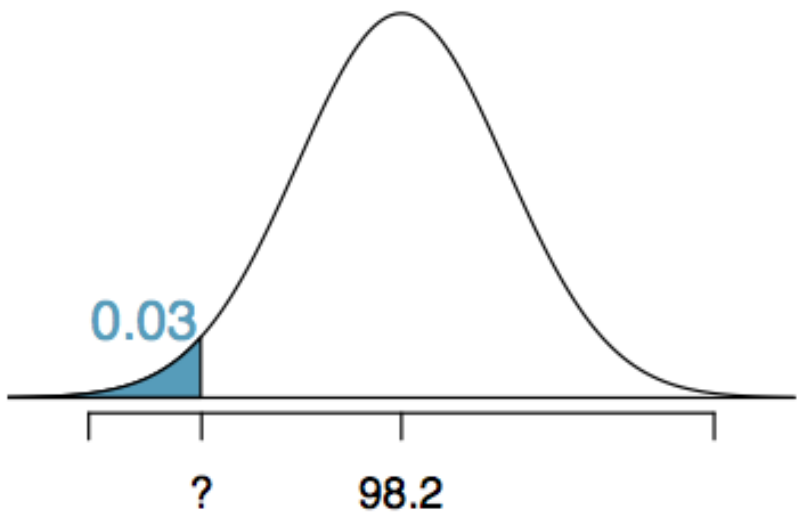
Finding cutoff points

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

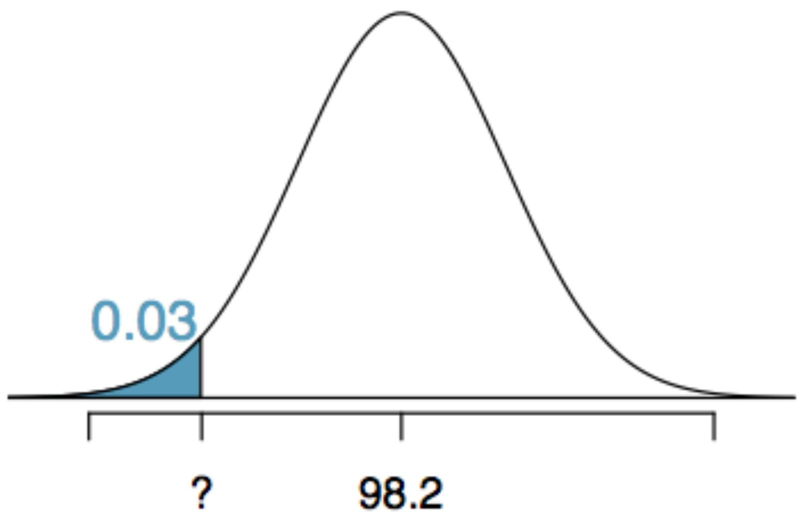
Body temperatures of healthy humans are distributed nearly normally with mean 98.2^{o}F and standard deviation 0.73^{o}F. What is the cutoff for the lowest 3% of human body temperatures?



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?

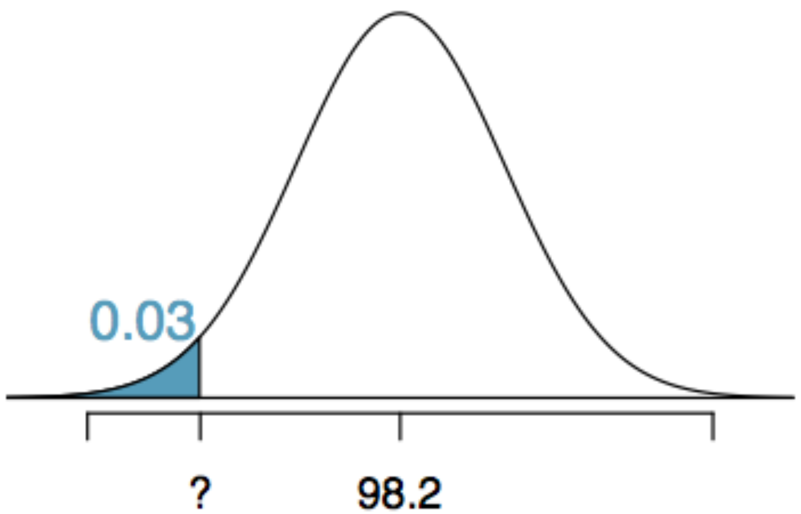


0.09	0.08	0.07	0.06	0.05	Z
0.0233	0.0239	0.0244	0.0250	0.0256	−1.9
0.0294	*0.0301*	0.0307	0.0314	0.0322	−1.8
0.0367	0.0375	0.0384	0.0392	0.0401	−1.7

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

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?



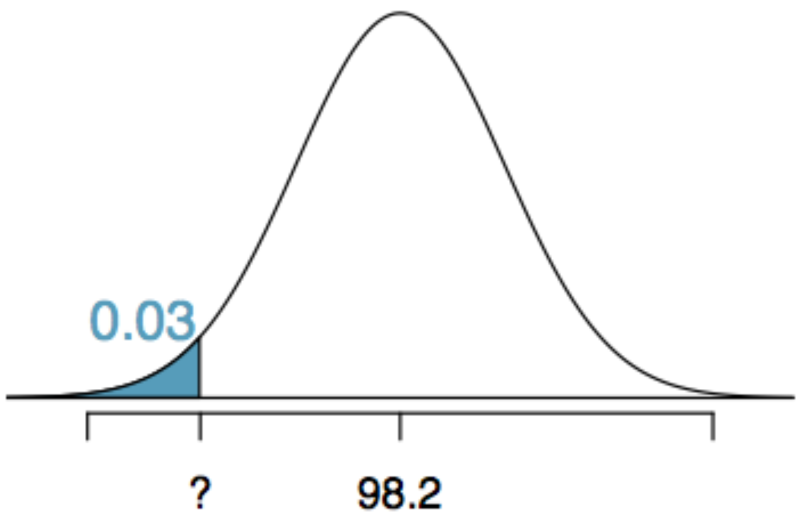
0.09	0.08	0.07	0.06	0.05	Z
0.0233	0.0239	0.0244	0.0250	0.0256	−1.9
0.0294	*0.0301*	0.0307	0.0314	0.0322	−1.8
0.0367	0.0375	0.0384	0.0392	0.0401	−1.7

$$P(X < x) \quad = \quad 0.03 \rightarrow P(Z < -1.88) = 0.03$$

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

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

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

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

$$x = (-1.88 \times 0.73) + 98.2 = 96.8^\circ F$$

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^{o}F and standard deviation 0.73^{o}F. What is the cutoff for the highest 10% of human body temperatures?

A. 97.3^{o}F

B. 99.1^{o}F

C. 99.4^{o}F

D. 99.6^{o}F

Practice

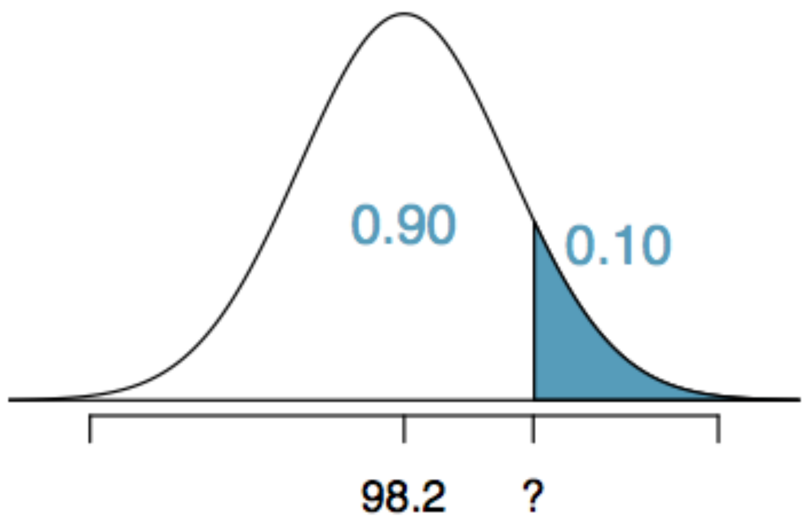
Body temperatures of healthy humans are distributed nearly normally with mean 98.2^{o}F and standard deviation 0.73^{o}F. What is the cutoff for the highest 10% of human body temperatures?

A. 97.3^{o}F

B. 99.1^{o}F

C. 99.4^{o}F

D. 99.6^{o}F



Practice

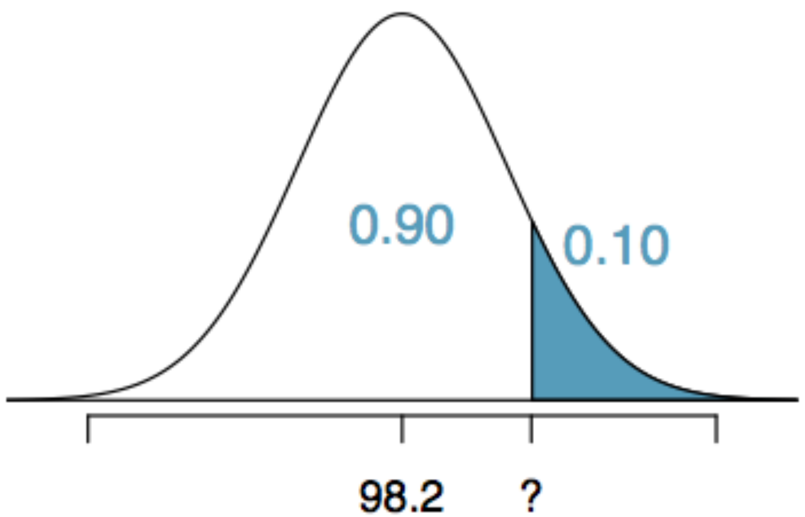
Body temperatures of healthy humans are distributed nearly normally with mean 98.2°F and standard deviation 0.73°F. What is the cutoff for the highest 10% of human body temperatures?

A. 97.3°F

B. 99.1°F

C. 99.4°F

D. 99.6°F

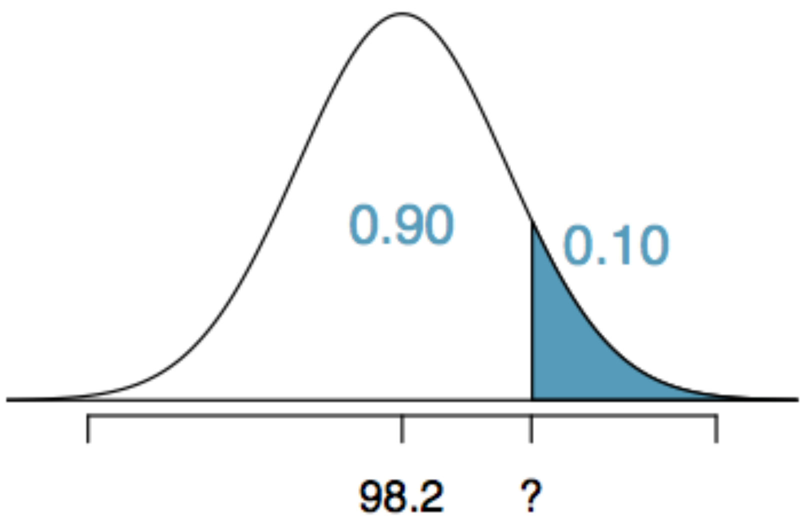


Z	0.05	0.06	0.07	0.08	0.09
1.0	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9115	0.9131	0.9147	0.9162	0.9177

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

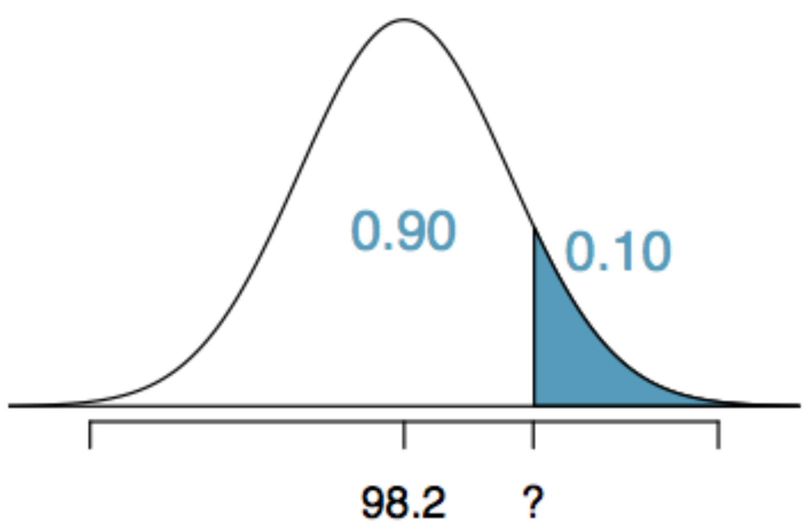
Body temperatures of healthy humans are distributed nearly normally with mean 98.2°F and standard deviation 0.73°F. What is the cutoff for the highest 10% of human body temperatures?

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

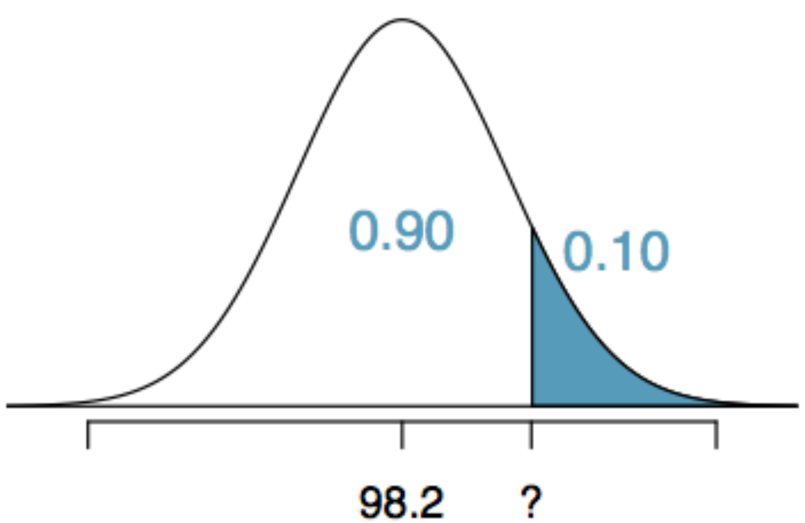
Body temperatures of healthy humans are distributed nearly normally with mean 98.2°F and standard deviation 0.73°F. What is the cutoff for the highest 10% of human body temperatures?

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

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

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

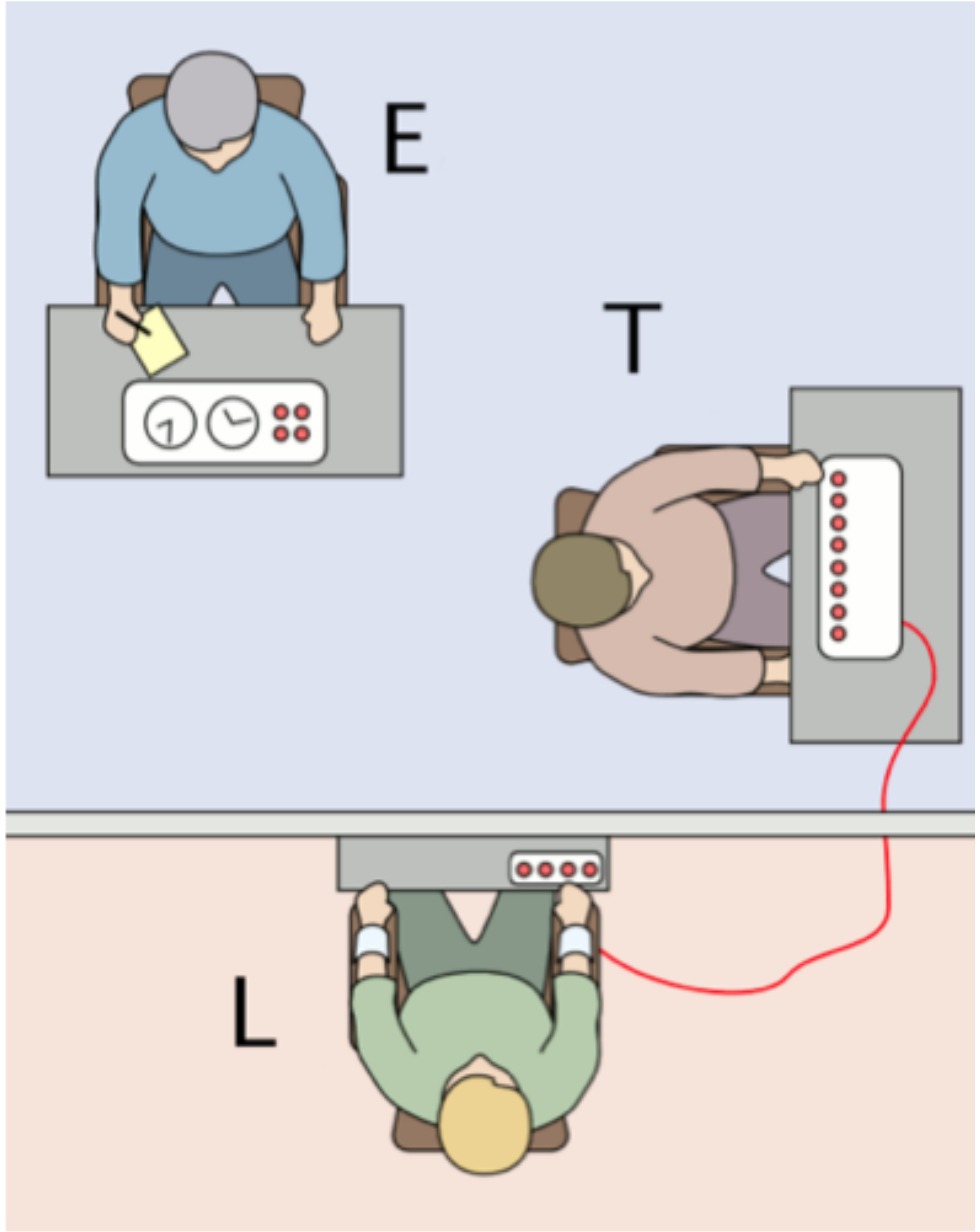
$$x = (1.28 \times 0.73) + 98.2 = 99.1$$

Binomial distribution

Milgram experiment

Stanley Milgram, a Yale University psychologist, conducted a series of experiments on obedience to authority starting in 1963.

- Experimenter (E) orders the teacher (T), the subject of the experiment, to give severe electric shocks to a learner (L) each time the learner answers a question incorrectly.
- The learner is actually an actor, and the electric shocks are not real, but a pre-recorded sound is played each time the teacher administers an electric shock.



http://en.wikipedia.org/wiki/
File:Milgram_Experiment_v2.png

Milgram experiment (cont.)

- These experiments measured the willingness of study participants to obey an authority figure who instructed them to perform acts that conflicted with their personal conscience.
- Milgram found that about 65% of people would obey authority and give such shocks.
- Over the years, additional research suggested this number is approximately consistent across communities and time.

Bernoulli random variables

- Each person in Milgram's experiment can be thought of as a *trial*.
- A person is labeled a *success* if she refuses to administer a severe shock, and *failure* if she administers such shock.
- Since only 35% of people refused to administer a shock, *probability of success* is p = 0.35.
- When an individual trial has only two possible outcomes, it is called a *Bernoulli random variable*.

Suppose we randomly select four individuals to participate in this experiment. What is the probability that exactly 1 of them will refuse to administer the shock?

Suppose we randomly select four individuals to participate in this experiment. What is the probability that exactly 1 of them will refuse to administer the shock?

Let's call these people Allen (A), Brittany (B), Caroline (C), and Damian (D). Each one of the four scenarios below will satisfy the condition of "exactly 1 of them refuses to administer the shock":

Scenario 1: $\underset{\text{(A) } \textit{refuse}}{\underline{0.35}} \times \underset{\text{(B) shock}}{\underline{0.65}} \times \underset{\text{(C) shock}}{\underline{0.65}} \times \underset{\text{(D) shock}}{\underline{0.65}} = 0.0961$

Scenario 2: $\underset{\text{(A) shock}}{\underline{0.65}} \times \underset{\text{(B) } \textit{refuse}}{\underline{0.35}} \times \underset{\text{(C) shock}}{\underline{0.65}} \times \underset{\text{(D) shock}}{\underline{0.65}} = 0.0961$

Scenario 3: $\underset{\text{(A) shock}}{\underline{0.65}} \times \underset{\text{(B) shock}}{\underline{0.65}} \times \underset{\text{(C) } \textit{refuse}}{\underline{0.35}} \times \underset{\text{(D) shock}}{\underline{0.65}} = 0.0961$

Scenario 4: $\underset{\text{(A) shock}}{\underline{0.65}} \times \underset{\text{(B) shock}}{\underline{0.65}} \times \underset{\text{(C) shock}}{\underline{0.65}} \times \underset{\text{(D) } \textit{refuse}}{\underline{0.35}} = 0.0961$

The probability of exactly one 1 of 4 people refusing to administer the shock is the sum of all of these probabilities.

$$0.0961 + 0.0961 + 0.0961 + 0.0961 = 4 \times 0.0961 = 0.3844$$

Binomial distribution

The question from the prior slide asked for the probability of given number of successes, *k*, in a given number of trials, *n*, (*k* = 1 success in *n* = 4 trials), and we calculated this probability as

# of scenarios x P(single scenario)

- # of scenarios: there is a less tedious way to figure this out, we'll get to that shortly...
- P(single scenario) = $p^k(1-p)^{n-k}$
  where p is the probability of success to the power of number of successes, probability of failure to the power of number of failures

The *Binomial distribution* describes the probability of having exactly *k* successes in *n* independent Bernoulli trials with probability of success *p*.

Computing the # of scenarios

Earlier we wrote out all possible scenarios that fit the condition of exactly one person refusing to administer the shock. If n was larger and/or k was different than 1, for example, n = 9 and k = 2:

*RR*SSSSSSS
S*RR*SSSSSS
SS*RR*SSSSS
. . .
SS*R*SS*R*SSS
. . .
SSSSSSS*RR*

writing out all possible scenarios would be incredibly tedious and prone to errors.

Computing the # of scenarios

Choose function

The *choose function* is useful for calculating the number of ways to choose k successes in n trials.

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Computing the # of scenarios

Choose function

The *choose function* is useful for calculating the number of ways to choose k successes in n trials.

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

$$k = 1, n = 4: \binom{4}{1} = \frac{4}{1} = 4$$

$$k = 2, n = 9: \binom{9}{2} = \frac{9 \times 8}{2 \times 1} = \frac{72}{2} = 36$$

Note: You can also use R for these calculations:

```
> choose(9,2)
[1] 36
```

Practice

Which of the following is false?

(a) There are n ways of getting 1 success in n trials, $\binom{n}{1} = n$.

(b) There is only 1 way of getting n successes in n trials, $\binom{n}{n} = 1$.

(c) There is only 1 way of getting n failures in n trials, $\binom{n}{0} = 1$.

(d) There are $n - 1$ ways of getting $n - 1$ successes in n trials, $\binom{n}{n-1} = n - 1$.

Practice

Which of the following is false?

(a) There are n ways of getting 1 success in n trials, $\binom{n}{1} = n$.

(b) There is only 1 way of getting n successes in n trials, $\binom{n}{n} = 1$.

(c) There is only 1 way of getting n failures in n trials, $\binom{n}{0} = 1$.

(d) *There are $n - 1$ ways of getting $n - 1$ successes in n trials, $\binom{n}{n-1} = n - 1$.*

Binomial distribution (cont.)

Binomial probabilities

If p represents probability of success, (1-p) represents probability of failure, n represents number of independent trials, and k represents number of successes

$$P(k\ successes\ in\ n\ trials) = \binom{n}{k} p^k\ (1-p)^{(n-k)}$$

Practice

Which of the following is not a condition that needs to be met for the binomial distribution to be applicable?

(a) the trials must be independent
(b) the number of trials, *n*, must be fixed
(c) each trial outcome must be classified as a *success* or a *failure*
(d) the number of desired successes, *k*, must be greater than the number of trials
(e) the probability of success, *p*, must be the same for each trial

Practice

Which of the following is not a condition that needs to be met for the binomial distribution to be applicable?

(a) the trials must be independent
(b) the number of trials, n, must be fixed
(c) each trial outcome must be classified as a *success* or a *failure*
(d) the number of desired successes, k, must be greater than the number of trials
(e) the probability of success, p, must be the same for each trial

Practice

A 2012 Gallup survey suggests that 26.2% of Americans are obese. Among a random sample of 10 Americans, what is the probability that exactly 8 are obese?

(a) pretty high
(b) pretty low

Gallup: http://www.gallup.com/poll/160061/obesity-rate-stable-2012.aspx, January 23, 2013.

Practice

A 2012 Gallup survey suggests that 26.2% of Americans are obese. Among a random sample of 10 Americans, what is the probability that exactly 8 are obese?

(a) pretty high
(b) pretty low

Gallup: http://www.gallup.com/poll/160061/obesity-rate-stable-2012.aspx, January 23, 2013.

Practice

A 2012 Gallup survey suggests that 26.2% of Americans are obese. Among a random sample of 10 Americans, what is the probability that exactly 8 are obese?

(a) $0.262^8 \times 0.738^2$

(b) $\binom{8}{10} \times 0.262^8 \times 0.738^2$

(c) $\binom{10}{8} \times 0.262^8 \times 0.738^2$

(d) $\binom{10}{8} \times 0.262^2 \times 0.738^8$

Practice

A 2012 Gallup survey suggests that 26.2% of Americans are obese. Among a random sample of 10 Americans, what is the probability that exactly 8 are obese?

(a) $0.262^8 \times 0.738^2$

(b) $\binom{8}{10} \times 0.262^8 \times 0.738^2$

(c) $\binom{10}{8} \times 0.262^8 \times 0.738^2 = 45 \times 0.262^8 \times 0.738^2 = 0.0005$

(d) $\binom{10}{8} \times 0.262^2 \times 0.738^8$

The birthday problem

What is the probability that 2 randomly chosen people share a birthday?

Pretty low, 1 / 365 ≈ 0.0027

What is the probability that at least 2 people out of 366 people share a birthday?

Exactly 1! (Excluding the possibility of a leap year birthday.)

The birthday problem (cont.)

What is the probability that at least 2 people (1 match) out of 121 people share a birthday?

Somewhat complicated to calculate, but we can think of it as the complement of the probability that there are no matches in 121 people.

$$
\begin{aligned}
P(no\ matches) &= 1 \times \left(1 - \frac{1}{365}\right) \times \left(1 - \frac{2}{365}\right) \times \cdots \times \left(1 - \frac{120}{365}\right) \\
&= \frac{365 \times 364 \times \cdots \times 245}{365^{121}} \\
&= \frac{365!}{365^{121} \times (365 - 121)!} \\
&= \frac{121! \times \binom{365}{121}}{365^{121}} \approx 0
\end{aligned}
$$

$$P(at\ least\ 1\ match) \approx 1$$

Expected value

A 2012 Gallup survey suggests that 26.2% of Americans are obese.

Among a random sample of 100 Americans, how many would you expect to be obese?

Expected value

A 2012 Gallup survey suggests that 26.2% of Americans are obese.

Among a random sample of 100 Americans, how many would you expect to be obese?

- Easy enough, $100 \times 0.262 = 26.2$.

Expected value

A 2012 Gallup survey suggests that 26.2% of Americans are obese.

Among a random sample of 100 Americans, how many would you expect to be obese?

- Easy enough, 100 x 0.262 = 26.2.
- Or more formally, $\mu = np = 100 \times 0.262 = 26.2$.

Expected value

A 2012 Gallup survey suggests that 26.2% of Americans are obese.

Among a random sample of 100 Americans, how many would you expect to be obese?

- Easy enough, 100 x 0.262 = 26.2.
- Or more formally, $\mu = np$ = 100 x 0.262 = 26.2.
- But this doesn't mean in every random sample of 100 people exactly 26.2 will be obese. In fact, that's not even possible. In some samples this value will be less, and in others more. How much would we expect this value to vary?

Expected value and its variability

Mean and standard deviation of binomial distribution

$$\mu = np \qquad \sigma = \sqrt{np(1-p)}$$

Expected value and its variability

Mean and standard deviation of binomial distribution

$$\mu = np \qquad \sigma = \sqrt{np(1-p)}$$

Going back to the obesity rate:

$$\sigma = \sqrt{np(1-p)} = \sqrt{100 \times 0.262 \times 0.738} \approx 4.4$$

Expected value and its variability

Mean and standard deviation of binomial distribution

$$\mu = np \qquad \sigma = \sqrt{np(1-p)}$$

Going back to the obesity rate:

$$\sigma = \sqrt{np(1-p)} = \sqrt{100 \times 0.262 \times 0.738} \approx 4.4$$

We would expect 26.2 out of 100 randomly sampled Americans to be obese, with a standard deviation of 4.4.

Note: Mean and standard deviation of a binomial might not always be whole numbers, and that is alright, these values represent what we would expect to see on average.

Unusual observations

Using the notion that *observations that are more than 2 standard deviations away from the mean are considered unusual* and the mean and the standard deviation we just computed, we can calculate a range for the plausible number of obese Americans in random samples of 100.

$$26.2 \pm (2 \times 4.4) \rightarrow (17.4, 35.0)$$

Practice

An August 2012 Gallup poll suggests that 13% of Americans think home schooling provides an excellent education for children. Would a random sample of 1,000 Americans where only 100 share this opinion be considered unusual?

(a) Yes

(b) No

	Excellent	Good	Only fair	Poor	Total excellent/ good
	%	%	%	%	%
Independent private school	31	47	13	2	78
Parochial or church-related schools	21	48	18	5	69
Charter schools	17	43	23	5	60
Home schooling	13	33	30	14	46
Public schools	5	32	42	19	37

Gallup, Aug. 9-12, 2012

http://www.gallup.com/poll/156974/private-schools-top-marks-educating-children.aspx

Practice

An August 2012 Gallup poll suggests that 13% of Americans think home schooling provides an excellent education for children. Would a random sample of 1,000 Americans where only 100 share this opinion be considered unusual?

(a) Yes

(b) No

$$\mu = np = 1,000 \times 0.13 = 130$$

$$\sigma = \sqrt{np(1-p)} = \sqrt{1,000 \times 0.13 \times 0.87} \approx 10.6$$

http://www.gallup.com/poll/156974/private-schools-top-marks-educating-children.aspx

Practice

An August 2012 Gallup poll suggests that 13% of Americans think home schooling provides an excellent education for children. Would a random sample of 1,000 Americans where only 100 share this opinion be considered unusual?

(a) Yes (b) No

$$\mu = np = 1,000 \times 0.13 = 130$$

$$\sigma = \sqrt{np(1-p)} = \sqrt{1,000 \times 0.13 \times 0.87} \approx 10.6$$

Method 1: *Range of usual observations:* $130 \pm 2 \times 10.6 = (108.8, 151.2)$

100 is outside this range, so would be considered unusual.

http://www.gallup.com/poll/156974/private-schools-top-marks-educating-children.aspx

Practice

An August 2012 Gallup poll suggests that 13% of Americans think home schooling provides an excellent education for children. Would a random sample of 1,000 Americans where only 100 share this opinion be considered unusual?

(a) Yes

(b) No

$$\mu = np = 1,000 \times 0.13 = 130$$

$$\sigma = \sqrt{np(1-p)} = \sqrt{1,000 \times 0.13 \times 0.87} \approx 10.6$$

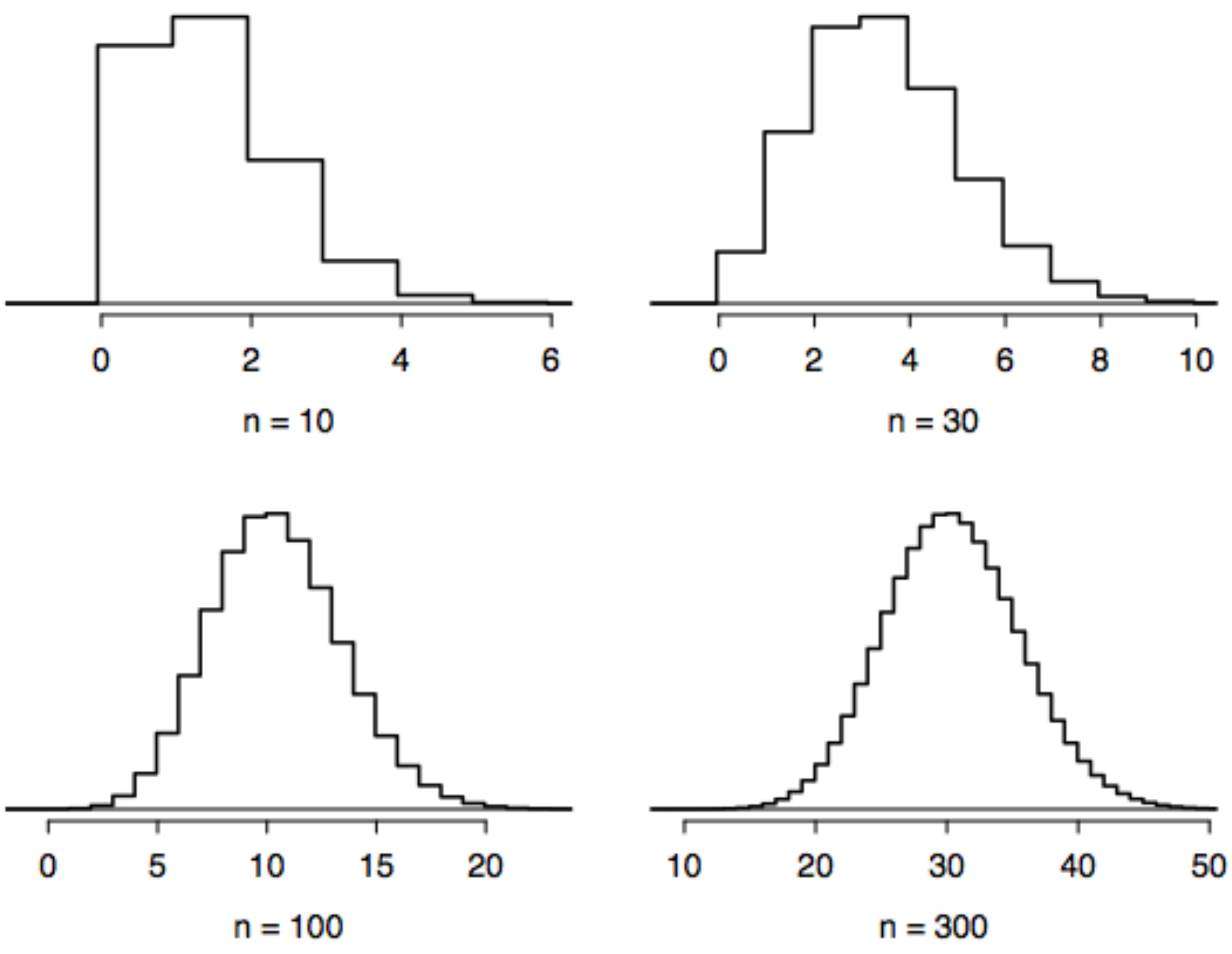
Method 1: *Range of usual observations:* $130 \pm 2 \times 10.6 = (108.8, 151.2)$
100 is outside this range, so would be considered unusual.

Method 2: *Z-score of observation:* $Z = \frac{x - mean}{SD} = \frac{100-130}{10.6} = -2.83$
100 is more than 2 SD below the mean, so would be considered unusual.

http://www.gallup.com/poll/156974/private-schools-top-marks-educating-children.aspx

Distributions of number of successes

Hollow histograms of samples from the binomial model where p = 0.10 and n = 10, 30, 100, and 300. What happens as n increases?



How large is large enough?

The sample size is considered large enough if the expected number of successes and failures are both at least 10.

$np \geq 10$ and $n(1 - p) \geq 10$

How large is large enough?

The sample size is considered large enough if the expected number of successes and failures are both at least 10.

$np \geq 10$ and $n(1 - p) \geq 10$

$$10 \times 0.13 \approx 1.3$$

$$10 \times (1 - 0.13) = 8.7$$

Practice

Below are four pairs of Binomial distribution parameters. Which distribution can be approximated by the normal distribution?

(a) $n = 100$, $p = 0.95$
(b) $n = 25$, $p = 0.45$
(c) $n = 150$, $p = 0.05$
(d) $n = 500$, $p = 0.015$

Practice

Below are four pairs of Binomial distribution parameters. Which distribution can be approximated by the normal distribution?

(a) $n = 100$, $p = 0.95$
(b) $n = 25$, $p = 0.45 \rightarrow 25 \times 0.45 = 11.25$, $25 \times 0.55 = 13.75$
(c) $n = 150$, $p = 0.05$
(d) $n = 500$, $p = 0.015$

An analysis of Facebook users

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

1. 40% of Facebook users in our sample made a friend request, but 63% received at least one request
2. 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
3. Users sent 9 personal messages, but received 12
4. 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:

1. 40% of Facebook users in our sample made a friend request, but 63% received at least one request
2. 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
3. Users sent 9 personal messages, but received 12
4. 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

Practice

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 $P(K \geq 70)$. To proceed, we need independence, which we'll assume but could check if we had access to more Facebook data.

Practice

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 $P(K \geq 70)$. To proceed, we need independence, which we'll assume but could check if we had access to more Facebook data.

$$P(X \geq 70) = P(K = 70 \text{ or } K = 71 \text{ or } K = 72 \text{ or } \dots \text{ or } K = 245)$$
$$= P(K = 70) + P(K = 71) + P(K = 72) + \dots + P(K = 245)$$

Practice

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 $P(K \geq 70)$. To proceed, we need independence, which we'll assume but could check if we had access to more Facebook data.

$$P(X \geq 70) = P(K = 70 \text{ or } K = 71 \text{ or } K = 72 \text{ or } \ldots \text{ or } K = 245)$$
$$= P(K = 70) + P(K = 71) + P(K = 72) + \ldots + P(K = 245)$$

This seems like an awful lot of work...

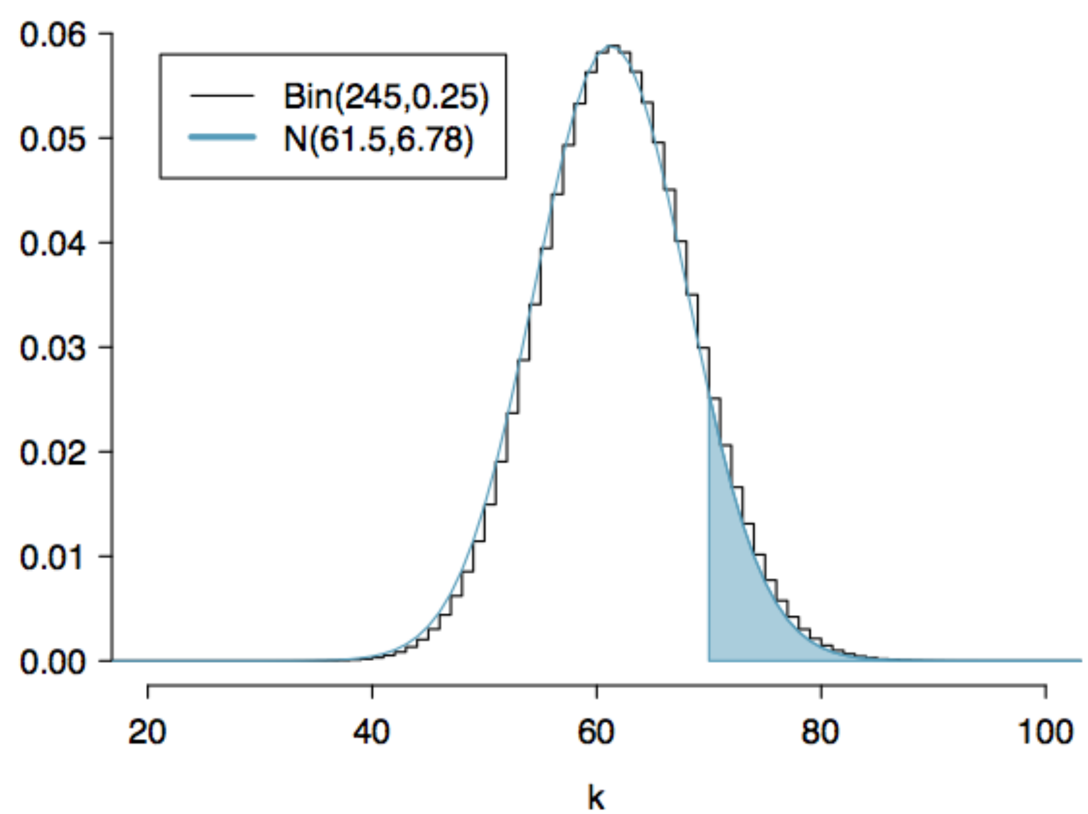
Normal approximation to the binomial

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 $p = 0.25$.

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

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

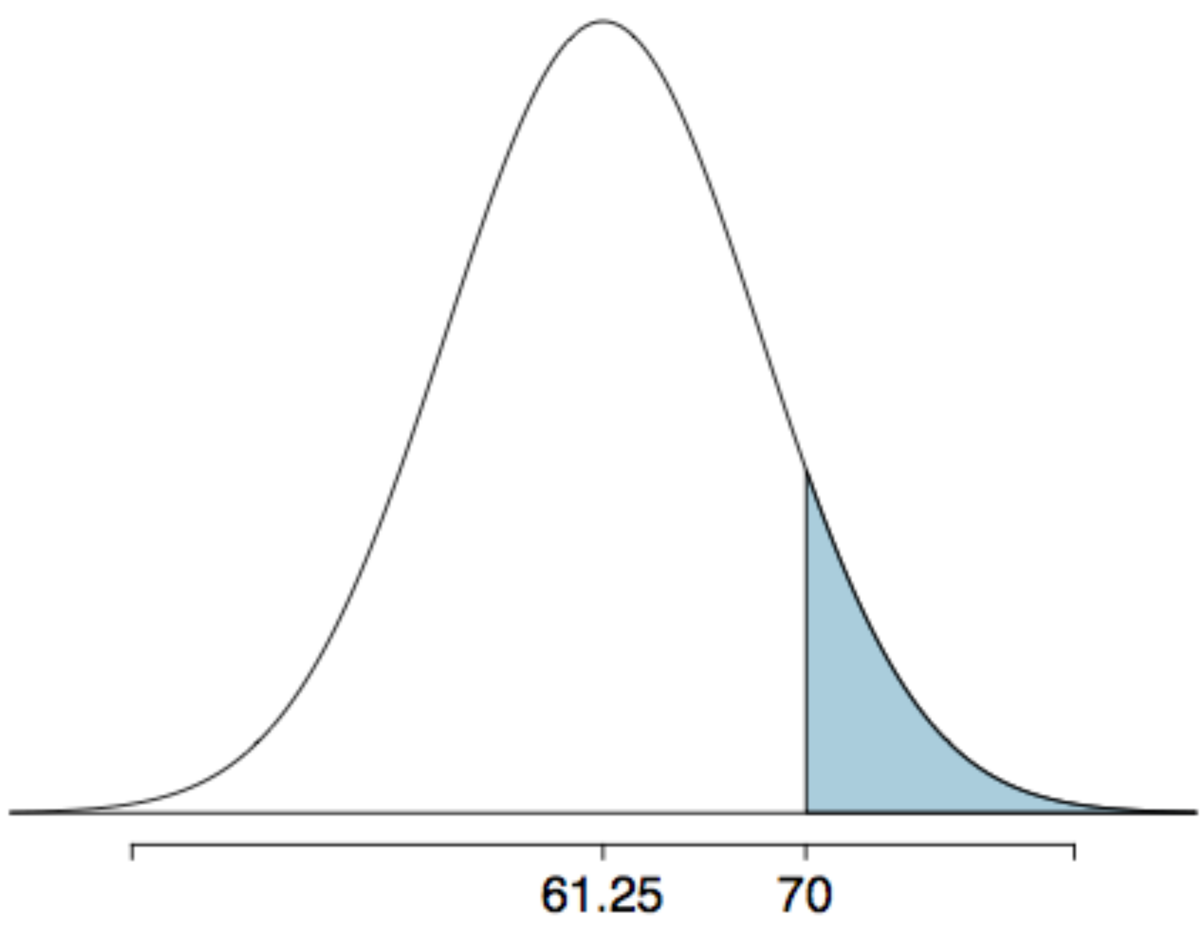


Practice

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

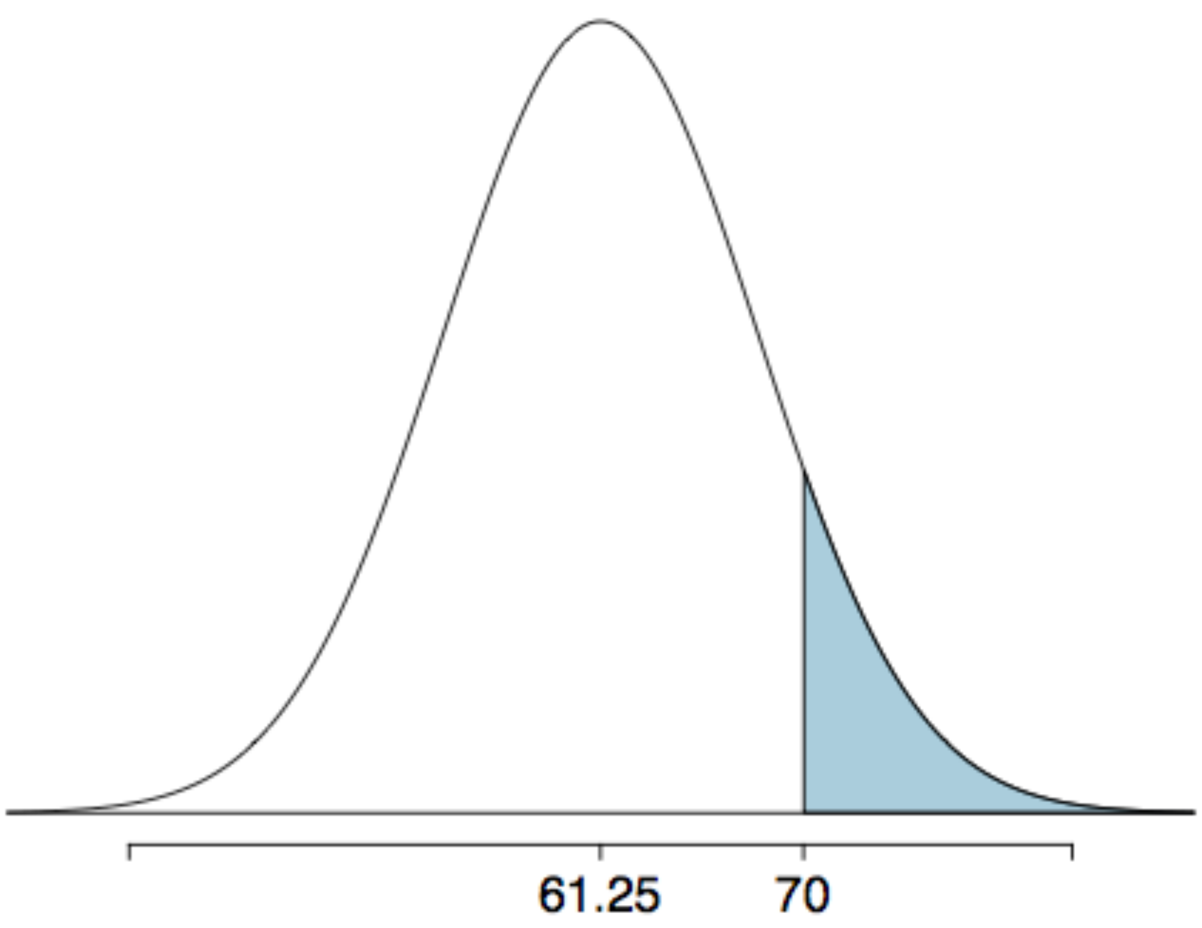
Practice

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



Practice

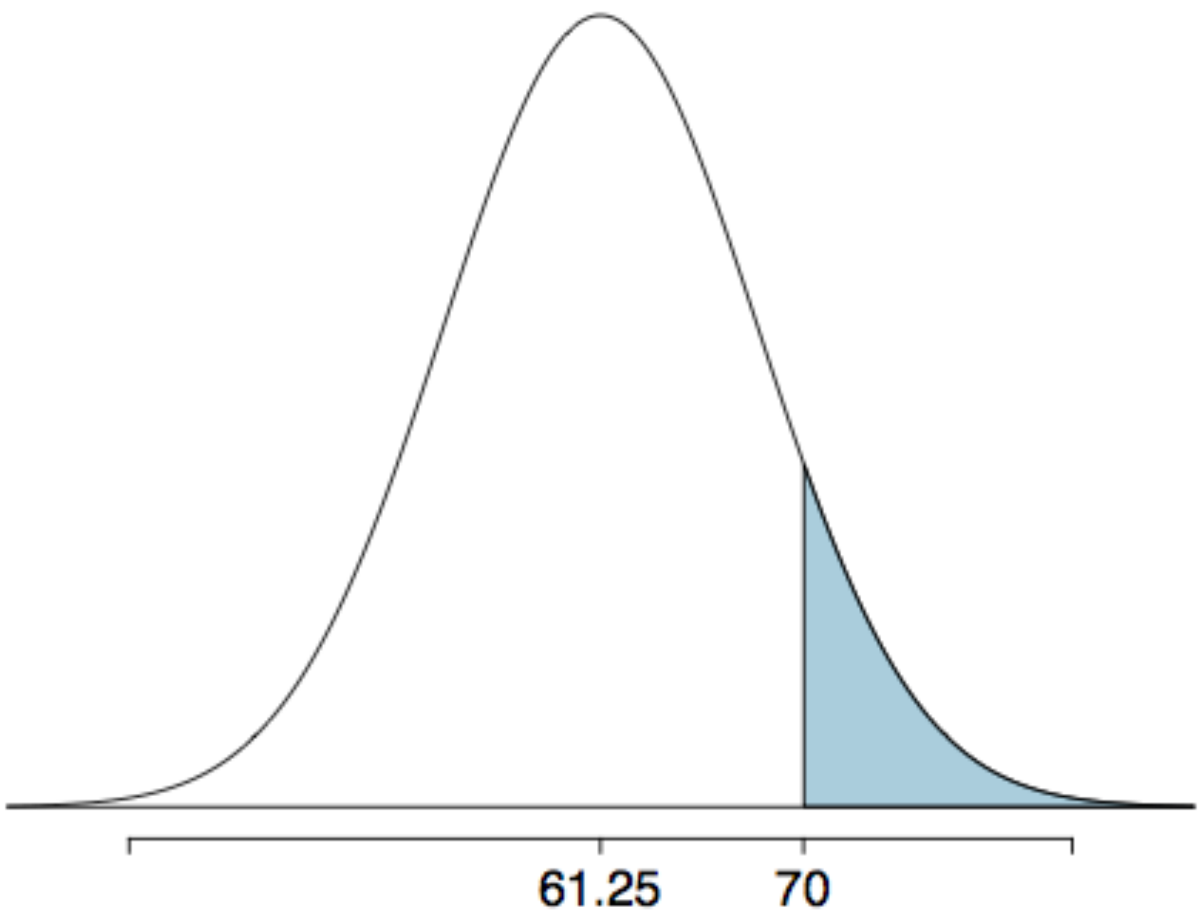
What is the probability that the average Facebook user with 245 friends has 70 or more friends who would be considered power users?



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

Practice

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



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

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

The normal approximation breaks down on small intervals

- The normal approximation to the binomial distribution tends to perform poorly when estimating the probability of a small range of counts, even when the conditions are met.

- This approximation for intervals of values is usually improved if cutoff values are extended by 0.5 in both directions.

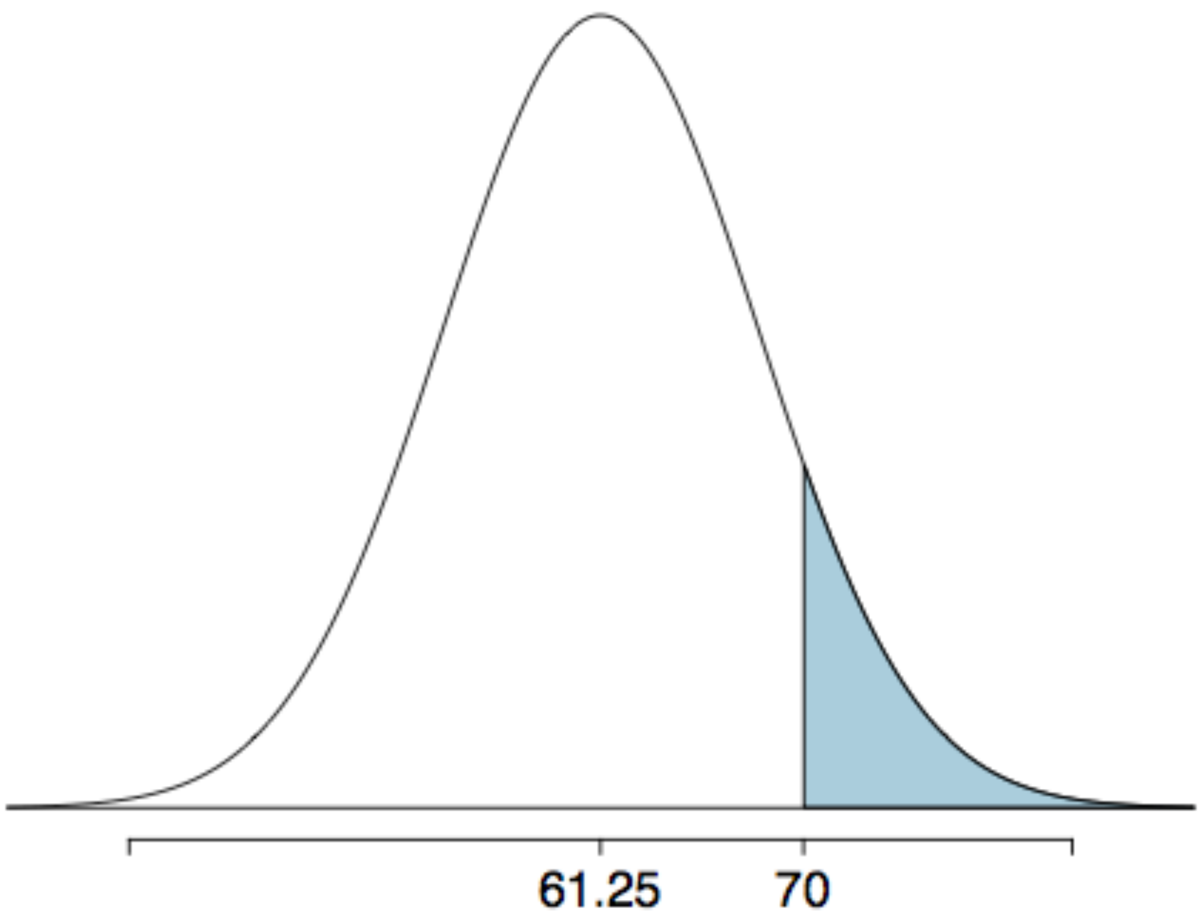
- The tip to add extra area when applying the normal approximation is most often useful when examining a range of observations. While it is possible to also apply this correction when computing a tail area, the benefit of the modification usually disappears since the total interval is typically quite wide.

Appendix

Probability Tables

Practice

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

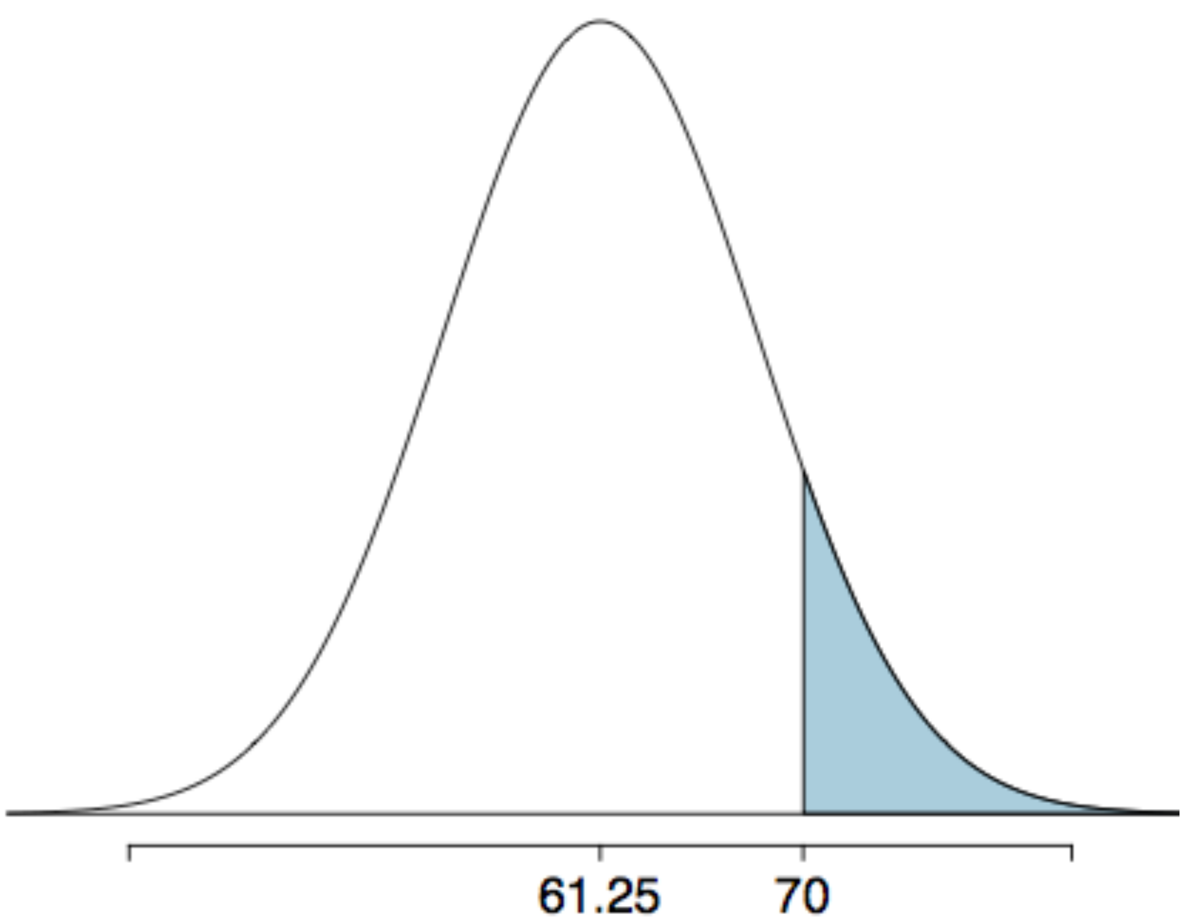


$$Z = \frac{obs - mean}{SD} = \frac{70 - 61.25}{6.78} = 1.29$$

	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*

Practice

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*

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

Poisson Distribution

Poisson distribution

- The Poisson distribution is often useful for estimating the number of rare events in a large population over a short unit of time for a fixed population if the individuals within the population are independent.
- The **rate** for a Poisson distribution is the average number of occurrences in a mostly-fixed population per unit of time, and is typically denoted by λ.
- Using the rate, we can describe the probability of observing exactly *k* rare events in a single unit of time.

$$\text{P(observe } k \text{ rare events)} = \frac{\lambda^k e^{-\lambda}}{k!}$$

where k may take a value 0, 1, 2, and so on, and *k!* represents *k*-factorial. The letter *e* ≈ 2.718 is the base of the natural logarithm.
57 The mean and standard deviation of this distribution are λ and √λ, respectively

Example #1

Suppose that in a rural region of a developing country electricity power failures occur following a Poisson distribution with an average of 2 failures every week. Calculate the probability that in a given week the electricity fails only once.

Example #1

Suppose that in a rural region of a developing country electricity power failures occur following a Poisson distribution with an average of 2 failures every week. Calculate the probability that in a given week the electricity fails only once.

Given $\lambda = 2$.

Example #1

Suppose that in a rural region of a developing country electricity power failures occur following a Poisson distribution with an average of 2 failures every week. Calculate the probability that in a given week the electricity fails only once.

Given $\lambda = 2$.

$$P(\text{only 1 failure in a week}) = \frac{2^1 \times e^{-2}}{1!}$$

Example #1

Suppose that in a rural region of a developing country electricity power failures occur following a Poisson distribution with an average of 2 failures every week. Calculate the probability that in a given week the electricity fails only once.

Given λ = 2.

$$
\begin{aligned}
P(\text{only 1 failure in a week}) &= \frac{2^1 \times e^{-2}}{1!} \\
&= \frac{2 \times e^{-2}}{1}
\end{aligned}
$$

Example #1

Suppose that in a rural region of a developing country electricity power failures occur following a Poisson distribution with an average of 2 failures every week. Calculate the probability that in a given week the electricity fails only once.

Given λ = 2.

$$
\begin{aligned}
P(\text{only 1 failure in a week}) &= \frac{2^1 \times e^{-2}}{1!} \\
&= \frac{2 \times e^{-2}}{1} \\
&= 0.27
\end{aligned}
$$

Example #2

Suppose that in a rural region of a developing country electricity power failures occur following a Poisson distribution with an average of 2 failures every week. Calculate the probability that on a given day the electricity fails three times.

Example #2

Suppose that in a rural region of a developing country electricity power failures occur following a Poisson distribution with an average of 2 failures every week. Calculate the probability that on a given day the electricity fails three times.

We are given the weekly failure rate, but to answer this question we need to first calculate the average rate of failure on a given day: λ_{day} = 2 = 0.2857. Note that we are assuming that the probability 7 of power failure is the same on any day of the week, i.e. we assume independence.

Example #2

Suppose that in a rural region of a developing country electricity power failures occur following a Poisson distribution with an average of 2 failures every week. Calculate the probability that on a given day the electricity fails three times.

We are given the weekly failure rate, but to answer this question we need to first calculate the average rate of failure on a given day:

λ_{day} = 2 = 0.2857. Note that we are assuming that the probability 7

of power failure is the same on any day of the week, i.e. we assume independence.

$$P(\text{3 failures on a given day}) = \frac{0.2857^{1} \times e^{-0.2857}}{3!}$$

Example #2

Suppose that in a rural region of a developing country electricity power failures occur following a Poisson distribution with an average of 2 failures every week. Calculate the probability that on a given day the electricity fails three times.

We are given the weekly failure rate, but to answer this question we need to first calculate the average rate of failure on a given day: λ_{day} = 2 = 0.2857. Note that we are assuming that the probability 7 of power failure is the same on any day of the week, i.e. we assume independence.

$$P(\text{3 failures on a given day}) = \frac{0.2857^1 \times e^{-0.2857}}{3!}$$

$$= \frac{0.2857 \times e^{-0.2857}}{6}$$

Example #2

Suppose that in a rural region of a developing country electricity power failures occur following a Poisson distribution with an average of 2 failures every week. Calculate the probability that on a given day the electricity fails three times.

We are given the weekly failure rate, but to answer this question we need to first calculate the average rate of failure on a given day: λ_{day} = 2 = 0.2857. Note that we are assuming that the probability 7 of power failure is the same on any day of the week, i.e. we assume independence.

$$\begin{aligned} P(\text{3 failures on a given day}) &= \frac{0.2857^1 \times e^{-0.2857}}{3!} \\ &= \frac{0.2857 \times e^{-0.2857}}{6} \\ &= 0.0358 \end{aligned}$$

Is it Poisson?

- A random variable may follow a Poisson distribution if the event being considered is rare, the population is large, and the events occur independently of each other
- However we can think of situations where the events are not really independent. For example, if we are interested in the probability of a certain number of weddings over one summer, we should take into consideration that weekends are more popular for weddings.
- In this case, a Poisson model may sometimes still be reasonable if we allow it to have a different rate for different times; we could model the rate as higher on weekends than on weekdays.
- The idea of modeling rates for a Poisson distribution against a second variable (day of the week) forms the foundation of some more advanced methods called *generalized linear models*. There are beyond the scope of this course, but we will discuss a foundation of linear models in Chapters 7 and 8.

Practice

A random variable that follows which of the following distributions can take on values other than positive integers?

(a) Poisson

(b) Negative binomial

(c) Binomial

(d) Normal

(e) Geometric

Practice

A random variable that follows which of the following distributions can take on values other than positive integers?

(a) Poisson

(b) Negative binomial

(c) Binomial

(d) *Normal*

(e) Geometric

Let’s discuss!

Auto Insurance premiums

Suppose a newspaper article states that the distribution of auto insurance premiums for residents of California is approximately normal with a mean of $1,650. The article also states that 25% of California residents pay more than $1,800.

(a) What is the Z-score that corresponds to the top 25% (or the 75th percentile) of the standard normal distribution?

(b) What is the mean insurance cost? What is the cutoff for the 75th percentile?

(c) Identify the standard deviation of insurance premiums in California.

Survey response rate

Pew Research reported that the typical response rate to their surveys is only 9%. If for a particular survey 15,000 households are contacted, what is the probability that at least 1,500 will agree to respond?

Quiz on Thursday

- Tomorrow is R Session.
- **Quiz**
  - OpenIntro Chapter 1-4
  - Lecture 1-9
  - Code from R sessions will be on the exam. You need to know the meaning of each code — eg: what does `sample(1:6, size =2)` do?
- We will cover the questions from practice exam during OH on Monday and Tuesday