

Normal distribution

Normal Distributions

Normal distributions (aka. Gaussian distributions) are a family of *symmetric, bell-shaped* density curves defined by

a mean μ , and an SD σ

denoted as $N(\mu, \sigma)$. The formula for the $N(\mu, \sigma)$ curve is

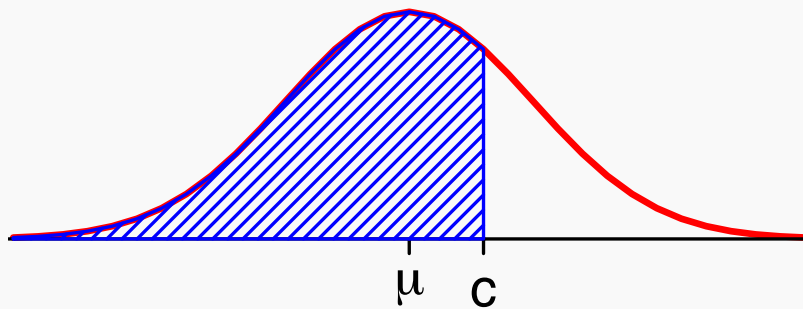
$$f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2}.$$



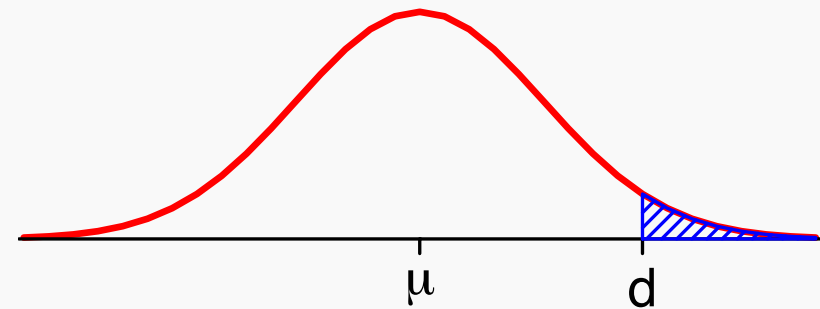
A normal distribution with $\mu = 0$, and $\sigma = 1$ is called the *standard normal distribution*, denoted as $N(0, 1)$.

Normal Probabilities

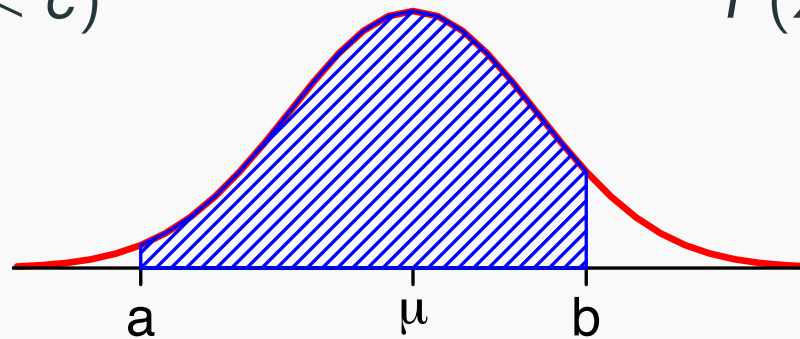
If X has a normal distribution, then to find probabilities about X is to find **areas** under a normal curve $N(\mu, \sigma)$.



$$P(X < c)$$



$$P(X > d)$$

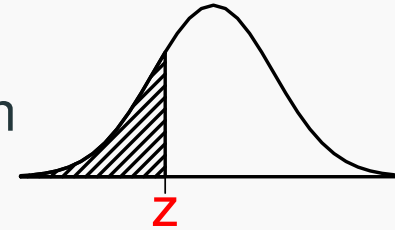


$$P(a < X < b)$$

But,... there is no simple formula to find areas under a Normal curve. Need to use softwares or the **normal probability table**.

The *normal probability table* (on p.428-429 in Text) gives

$P(Z < z) =$ area of shaded region in

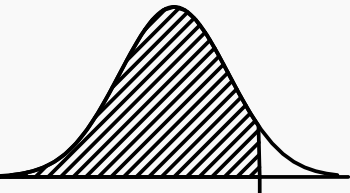


<i>z</i>	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
-0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
-0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
-0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

E.g., for $z = -0.83$, look at the row -0.8 and the column 0.03 .

$$P(Z < -0.83) = \text{area of shaded region in } \text{normal curve at } -0.83 = \underline{0.2033}$$

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

$$\begin{aligned}
 P(Z < 1.573) &= \text{area under the curve to the left of } 1.573 \\
 &= \text{between } P(Z < 1.57) \text{ and } P(Z < 1.58) \\
 &= \text{between } 0.9418 \text{ and } 0.9429
 \end{aligned}$$


Any value between 0.9418 and 0.9429 will be accepted in HWs and exams.

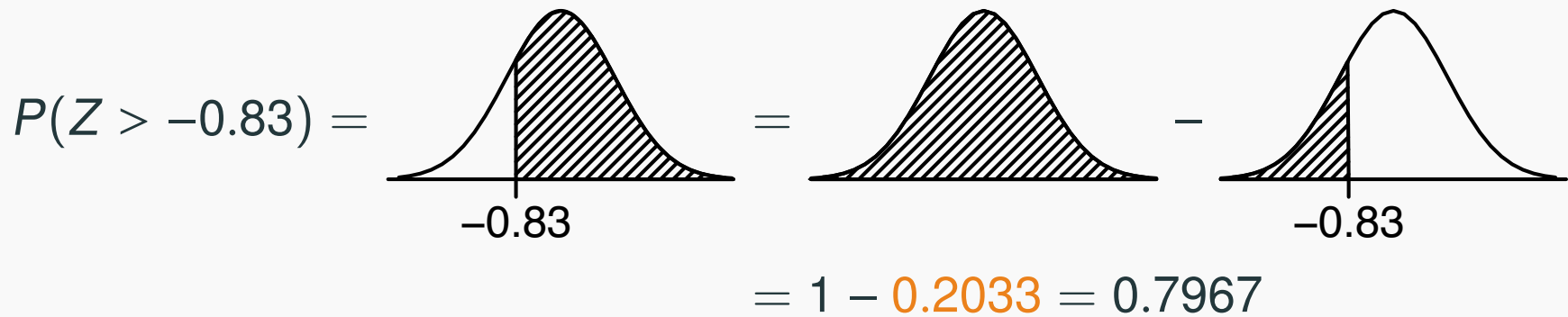
Find Normal Probabilities in R

The R command `pnorm()` can find areas under the standard normal $N(0, 1)$ curve

```
> pnorm(-0.83)  
[1] 0.2032694
```

```
> pnorm(1.573)  
[1] 0.9421406
```

Finding Upper Tail Probabilities



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148

```

> 1 - pnorm(-0.83)
[1] 0.7967306
> # another way to find upper tail area
> pnorm(-0.83, lower.tail=FALSE)
[1] 0.7967306
    
```

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
-0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857

$$\begin{aligned}
 P(-0.83 < Z < 2) &= \text{[Normal distribution graph with shaded area between -0.83 and 2]} \\
 &= \text{[Normal distribution graph with shaded area to the left of 2]} - \text{[Normal distribution graph with shaded area to the left of -0.83]} \\
 &= P(Z < 2) - P(Z < -0.83) \\
 &= 0.9772 - 0.2033 = 0.7739
 \end{aligned}$$

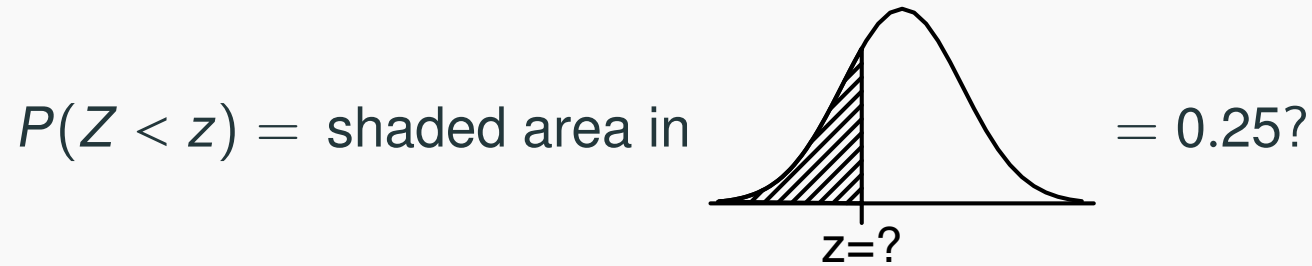
```

> pnorm(2) - pnorm(-0.83)
[1] 0.7739805

```


Finding z for a Given Probability

E.g, we want to find the first quartile of the standard normal, i.e., what's the z such that



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
-0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
-0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
-0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots

So the z is between -0.67 and -0.68 (about -0.675).

Finding the z such that $P(Z < z)$ equals a specific probability in R:

```
> qnorm(0.25)
[1] -0.6744898
```

Quartiles of the Standard Normal Distribution

The quartiles of the standard normal distributions are:

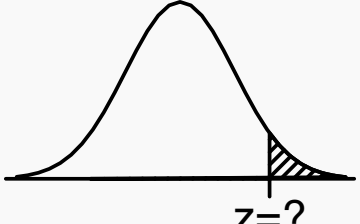
$$Q_1 \approx -0.675 \dots\dots \text{(from the previous slide)}$$

$$Q_2 = 0 \dots\dots \text{(why?)}$$

$$Q_3 \approx 0.675 \dots\dots \text{(why?)}$$

The interquartile range (IQR) for the standard normal curve is

$$\text{IQR} = Q_3 - Q_1 \approx 0.675 - (-0.675) \approx 1.35$$

If $P(Z > z) =$  $= 0.05$, then $z = ?$

This implies  $= 0.95$, so $z \approx 1.645$ (between 1.64 and 1.65).

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633

```
> qnorm(1-0.05)
```

```
[1] 1.644854
```

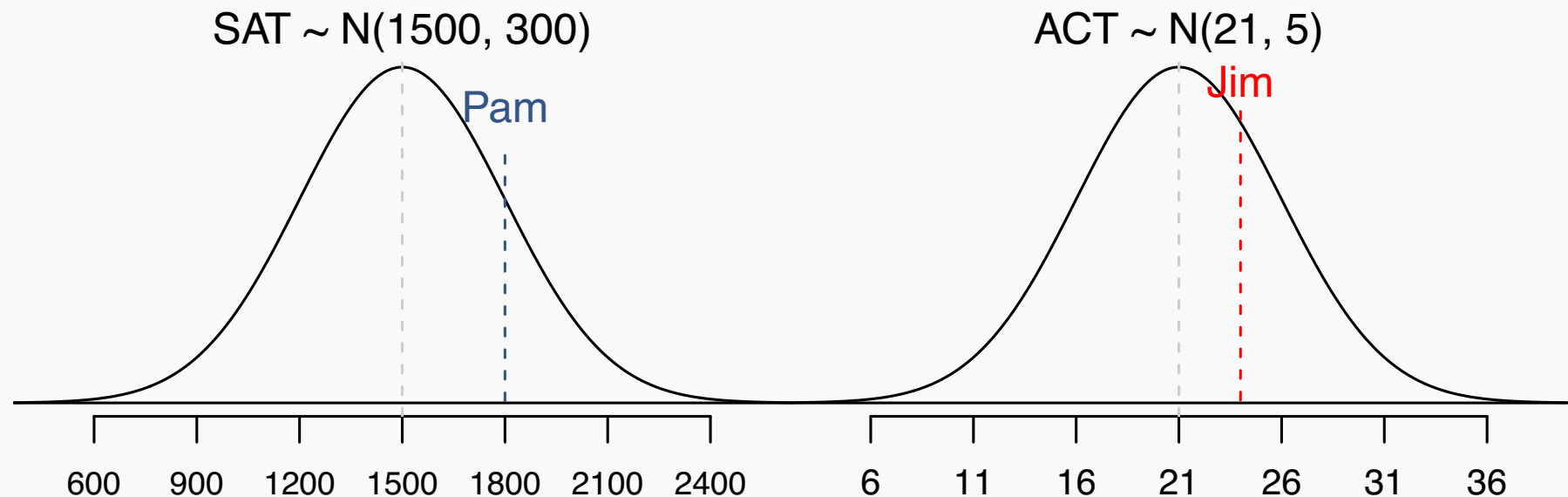
```
> qnorm(0.05, lower.tail=F) # alternative way
```

```
[1] 1.644854
```

Now we've learned how to find probabilities about the standard normal $N(0, 1)$. To compute probability about general normal distribution $N(\mu, \sigma)$, we need to know about the *Z score*.

Example: SAT vs. ACT

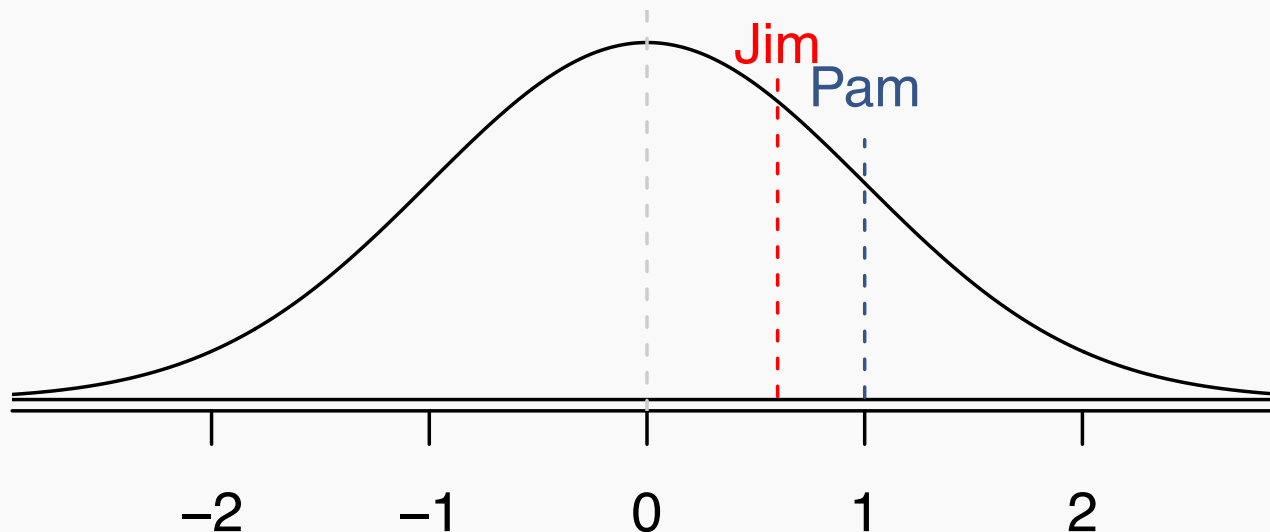
SAT scores are distributed nearly normally with mean 1500 and standard deviation 300. ACT scores are distributed nearly normally with mean 21 and standard deviation 5. A college admissions officer wants to determine which of the two applicants scored better on their standardized test with respect to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

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

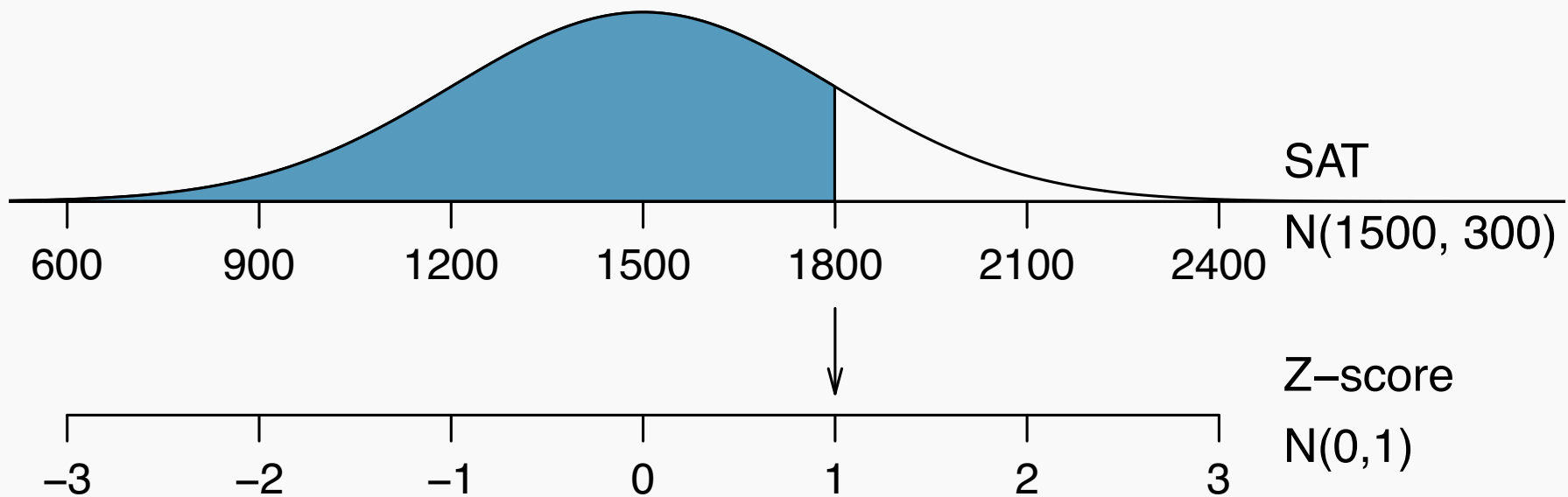
- Z scores are defined for distributions of any shape, but only when the distribution is normal can we use Z scores to calculate normal probabilities.
- Observations that are more than 3 SD away from the mean ($|Z| > 3$) are usually considered unusual.

Recap: Ways to Detect Outliers

- 1.5 IQR rule
- Observations with $|Z\text{-scores}| > 3$ (or sometimes > 2)
- Histograms
- Scatterplots

Calculating Normal Probabilities

Approximately what percent of students score below 1800 on the SAT? Recall that $SAT \sim N(\mu = 1500, \sigma = 300)$



The Z-score of 1800 is $Z = (1800 - 1500)/300 = 1$.

From the table (next slide), we can see that $P(Z < 1) = 0.8413$.

So about 84% of students score below 1800 on the SAT.

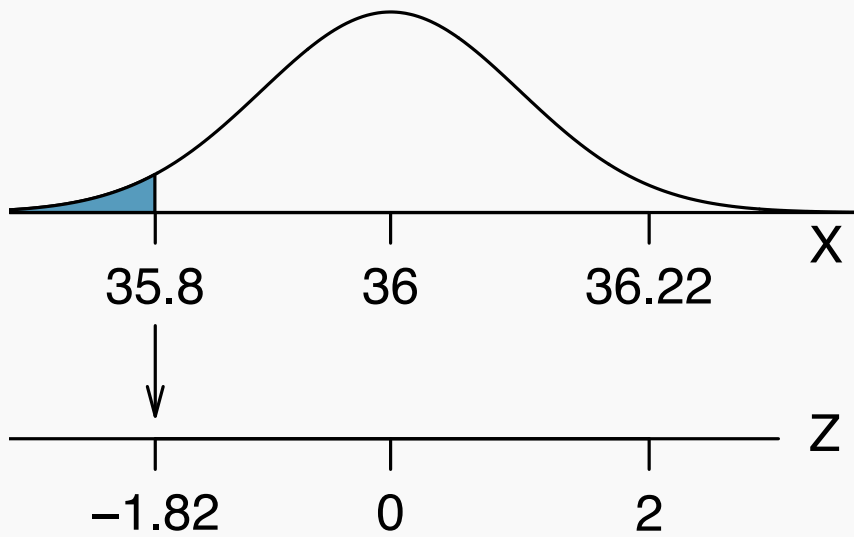
In R:

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


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:

$$X \sim N(\mu = 36, \sigma = 0.11)$$

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

Second decimal place of Z										Z
0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01	0.00	
.0183	.0188	.0192	.0197	.0202	.0207	.0212	.0217	.0222	.0228	-2.0
.0233	.0239	.0244	.0250	.0256	.0262	.0268	.0274	.0281	.0287	-1.9
.0294	.0301	.0307	.0314	.0322	.0329	.0336	.0344	.0351	.0359	-1.8
.0367	.0375	.0384	.0392	.0401	.0409	.0418	.0427	.0436	.0446	-1.7
.0455	.0465	.0475	.0485	.0495	.0505	.0516	.0526	.0537	.0548	-1.6
.0559	.0571	.0582	.0594	.0606	.0618	.0630	.0643	.0655	.0668	-1.5

Answer: $0.0344 = 3.44\%$

In R:

```
> pnorm(-1.82, mean = 0, sd = 1)
```

```
[1] 0.0343795
```

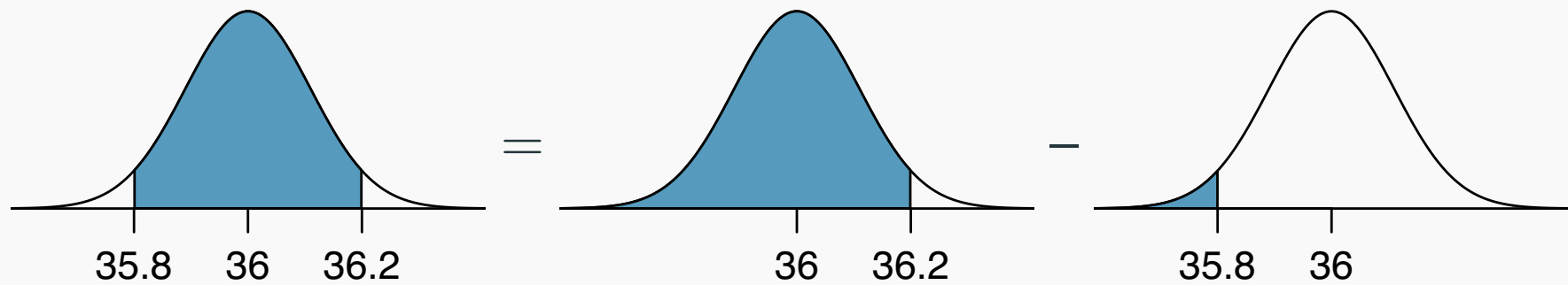
or

```
> pnorm(35.8, mean = 36, sd = 0.11)
```

```
[1] 0.03451817
```

Practice

What percent of bottles pass the quality control inspection (i.e., between 35.8 oz. and 36.2 oz.)?



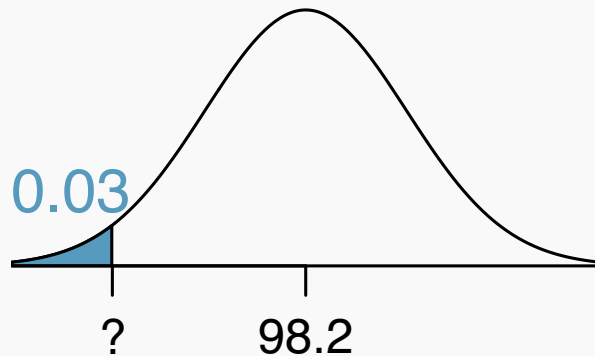
$$Z_{35.8} = \frac{35.8 - 36}{0.11} = -1.82, \quad Z_{36.2} = \frac{36.2 - 36}{0.11} = 1.82$$

$$\begin{aligned} P(35.8 < X < 36.2) &= P(-1.82 < Z < 1.82) \\ &= P(Z < 1.82) - P(Z < -1.82) \\ &= 0.9656 - 0.0344 = 0.9312 \end{aligned}$$

Answer: 93.12%.

Finding Cutoff Points For A Percentile

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

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

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

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

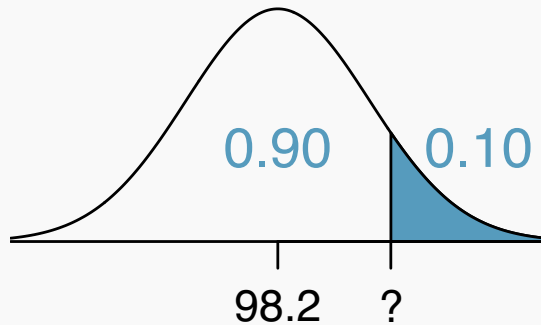
In R:

```
> qnorm(0.03, m = 98.2, s = 0.73)
```

```
[1] 96.82702
```

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?



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

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

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

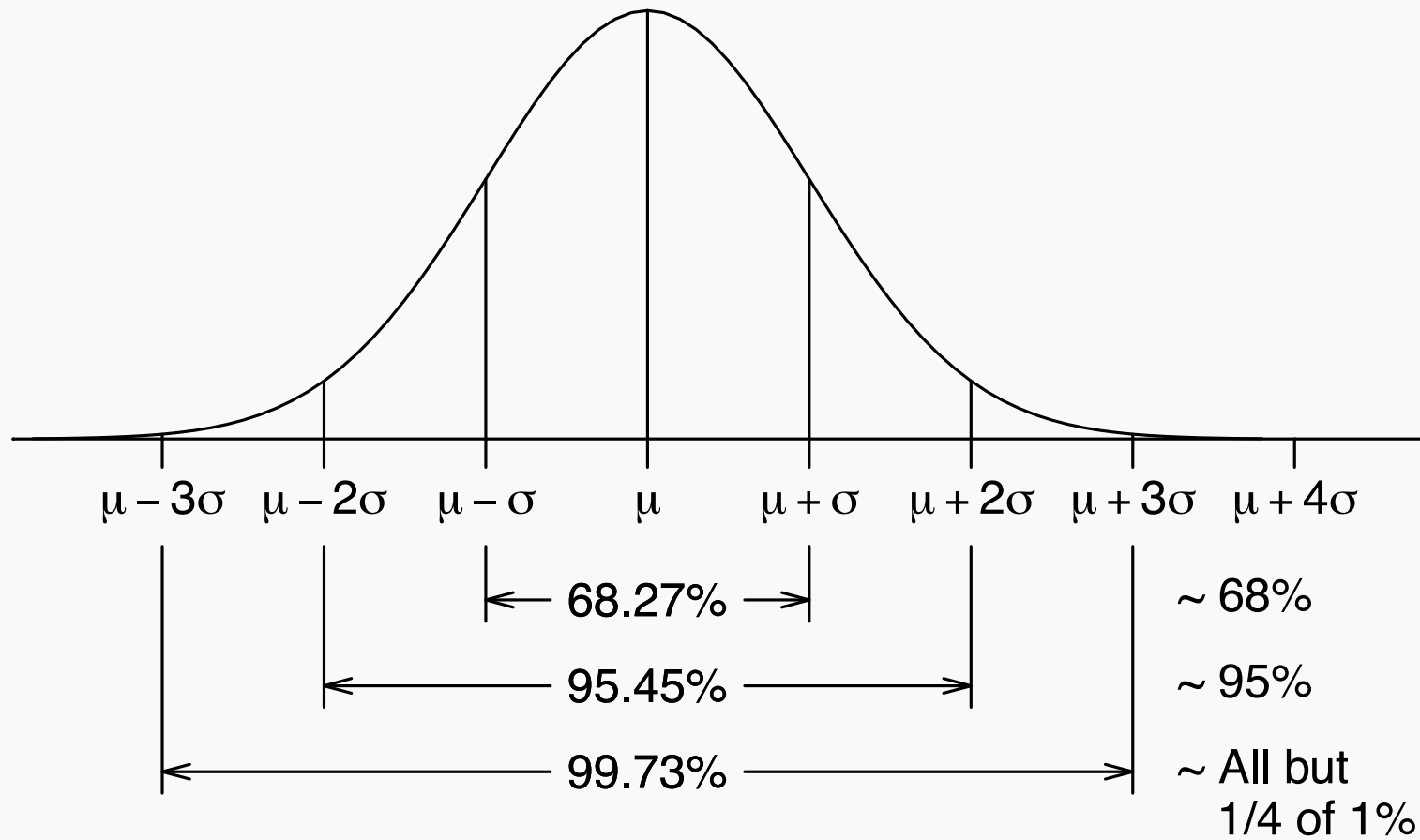
$$x = (1.28 \times 0.73) + 98.2 = 99.1$$

Answer: 99.1°F

```
> qnorm(0.9, m = 98.2, s = 0.73)
```

```
[1] 99.13553
```

68-95-99.7% Rule for Normal Distributions



```
> pnorm(1) - pnorm(-1)
```

```
[1] 0.6826895
```

```
> pnorm(2) - pnorm(-2)
```

```
[1] 0.9544997
```

```
> pnorm(3) - pnorm(-3)
```

```
[1] 0.9973002
```