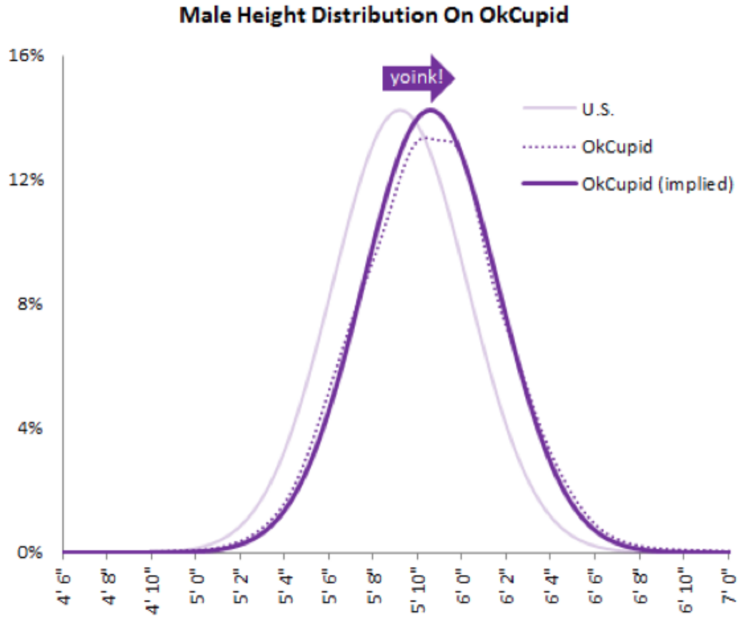
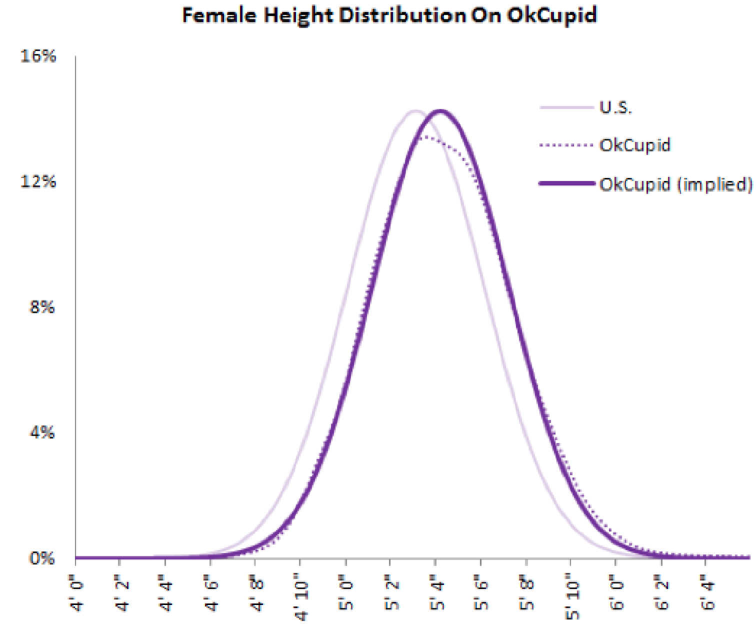
normal distribution

- normal distribution
- 68-95-99.7% rule
- standardized scores
- probabilities and percentiles

heights of males



heights of females



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

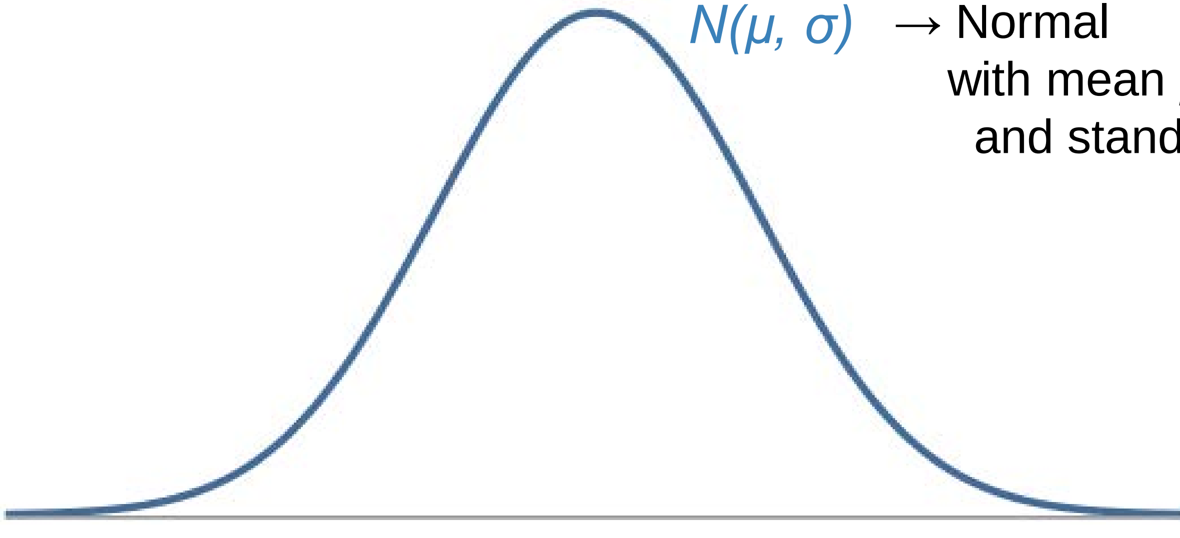
“You can also see a more subtle vanity at work: starting at roughly 5' 8", the top of the dotted curve tilts even further rightward. This means that guys as they get closer to six feet round up a bit more than usual, stretching for that coveted psychological benchmark.”

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

http://blog.okcupid.com/index.php/the-biggest-lies-in-online-dating

normal distribution

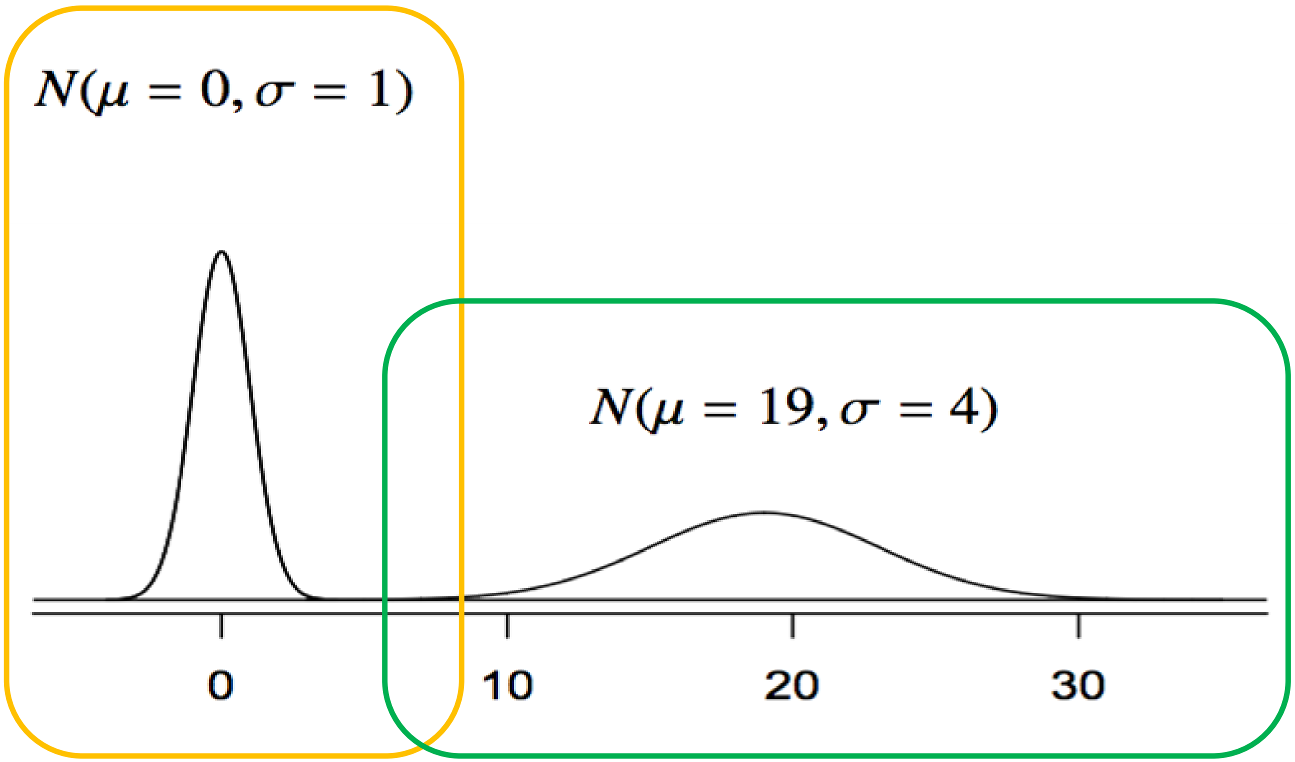
- unimodal and symmetric
    - bell shaped curve
- follows very strict guidelines about how variably around the mean
- many variables are nearly normal, but none are exactly normal



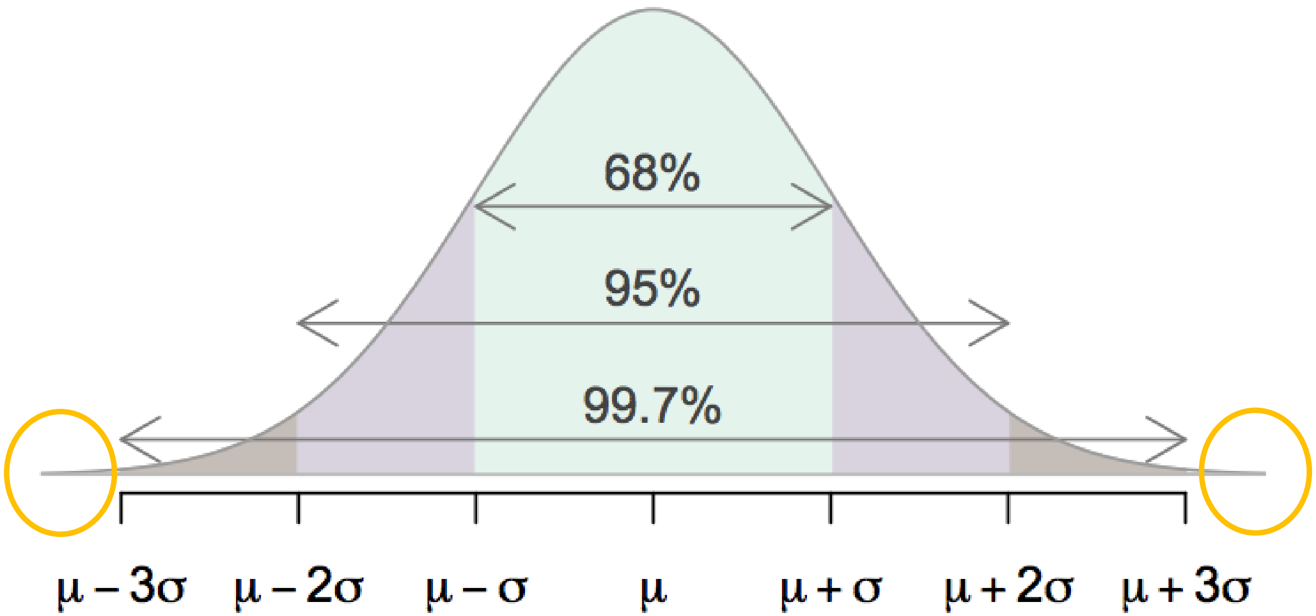
$N(\mu, \sigma)$ → Normal
with mean μ
and standard deviation σ

normal distributions with different parameters

μ: mean, σ: standard deviation



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.

Practice

Test scores are distributed nearly normally with mean 1500 and standard deviation 300. According to the 68-95-99.7% rule, which of the following is false?

a) Roughly 68% of students score between 1200 and 1800 on the Test

b) Roughly 95% of students score between 900 and 2100 on the Test

c) Roughly 99.7% of students score between 600 and 2400 on the Test

d) No students can score below 600 on the Test

Practice

A doctor collects a large set of heart rate measurements that approximately follow a normal distribution. He only reports 3 statistics the mean=110 beats per minutes, the minimum=65 beats per minute and the maximum =155 beats per minute. Which of the following is most likely to be the standard deviation of the distribution?

a) 5

b) 15

c) 35

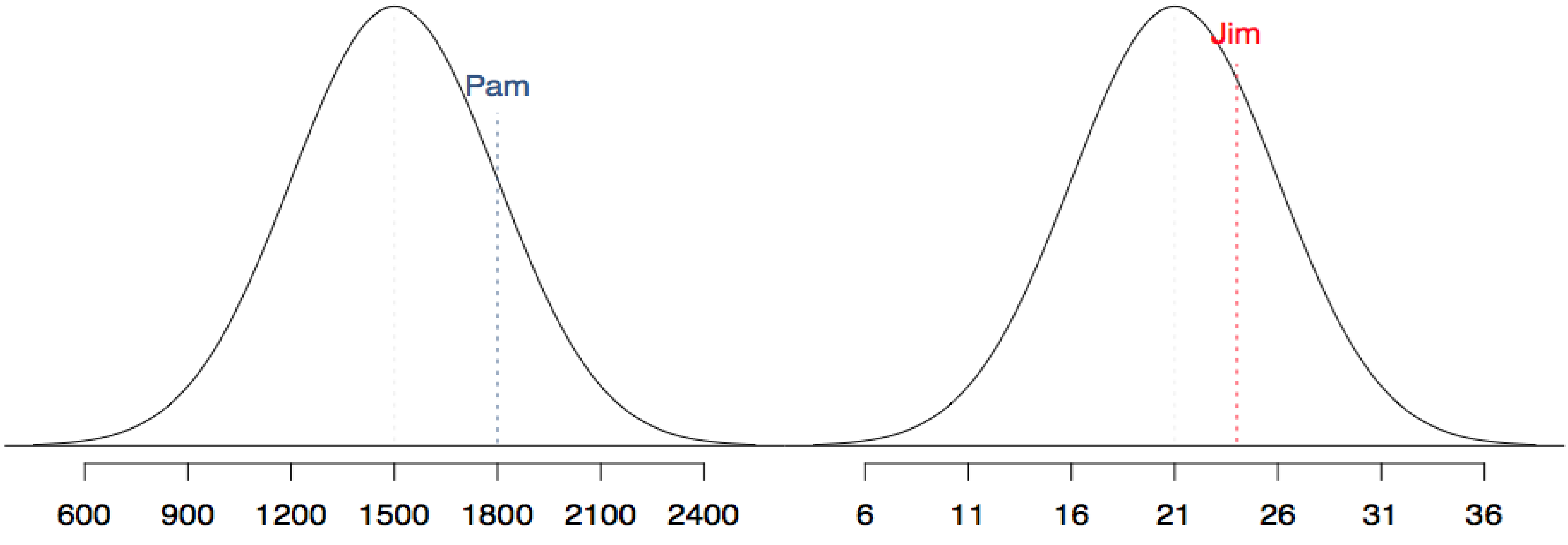
d) 90

Practice

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?

SAT scores ~ N(mean=1500, SD=300)

ACT scores ~ N(mean=21, SD=5)



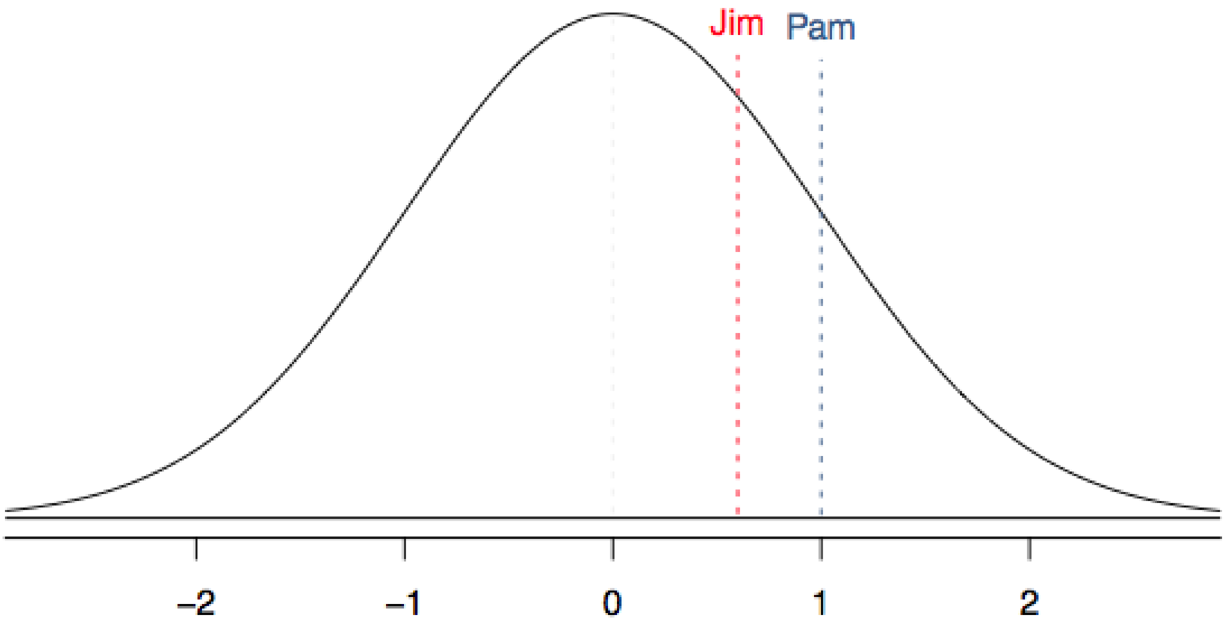
Since we cannot just compare these two raw scores, we instead compare how many standard deviations beyond the mean each observation is.

- Pam's score is

$\frac{1800-1500}{300} = 1$ standardrd deviation above the mean.

- Jim's score is

$\frac{24-21}{5} = 0.6$ standard deviation above the mean



standardizing with Z scores

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

- standardized (Z) scores of an observation is the number of standard deviations it falls above or below the mean.

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

- Z score of mean =0
- unusual observation: |Z|>2
- defined for distributions of any shape, but only when the distribution is normal can we use Z scores to calculate percentiles.

Practice

Scores on a standardized test are nearly normally distributed with a mean of 100 and a standard deviation of 20. If these scores are converted to standard normal Z scores, which of the following statements will be correct?

a) The mean will be 0, and the median should be roughly 0 as well

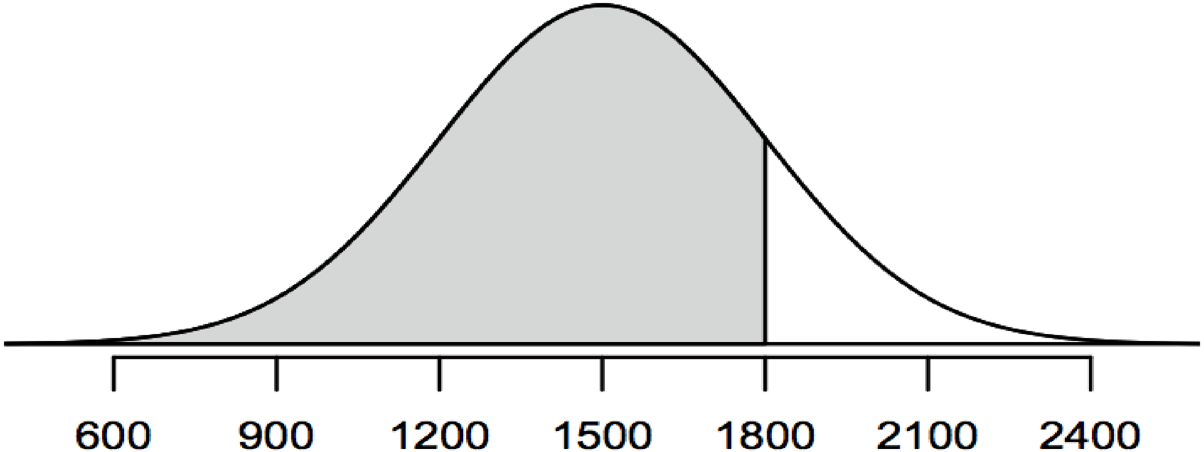
b) The mean will be 0, and the median cannot be determined

c) The mean of the standardized Z scores will be equal 100

d) The mean of the standardized Z scores will be equal 5

percentiles

- when the distribution is normal, Z scores can be used to calculate percentile
- percentile is the percentage of observations that fall below a given data point.
- graphically, percentile is the area below the probability distribution curve to the left of that observation.



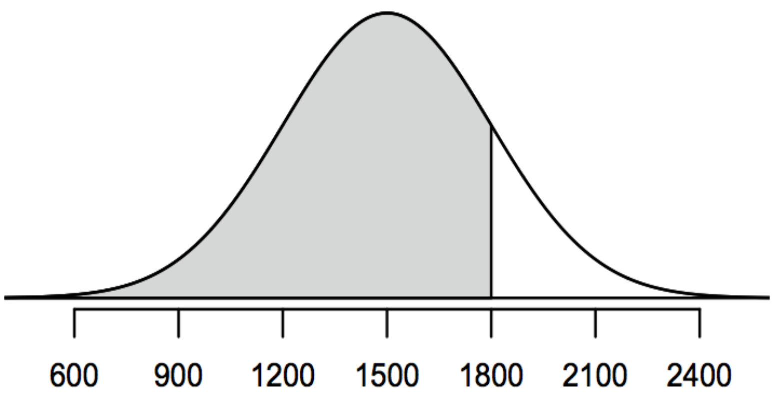
There are many ways to compute percentiles/areas under the curve.

Practice : computing percentiles

SAT scores are distributed nearly normally with mean 1500 and standard deviation 300. Pam earned an 1800 on her SAT. What is Pam's percentile score?

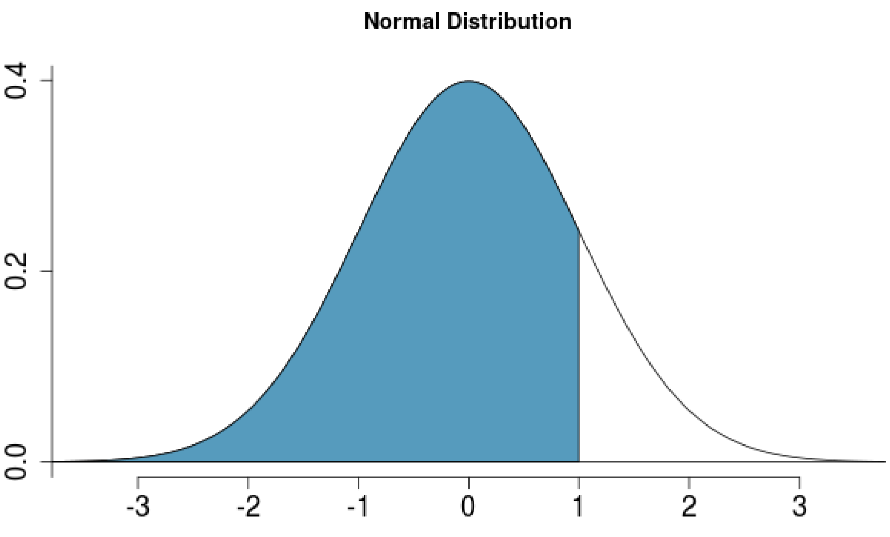
1. using R

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



2. using Applet

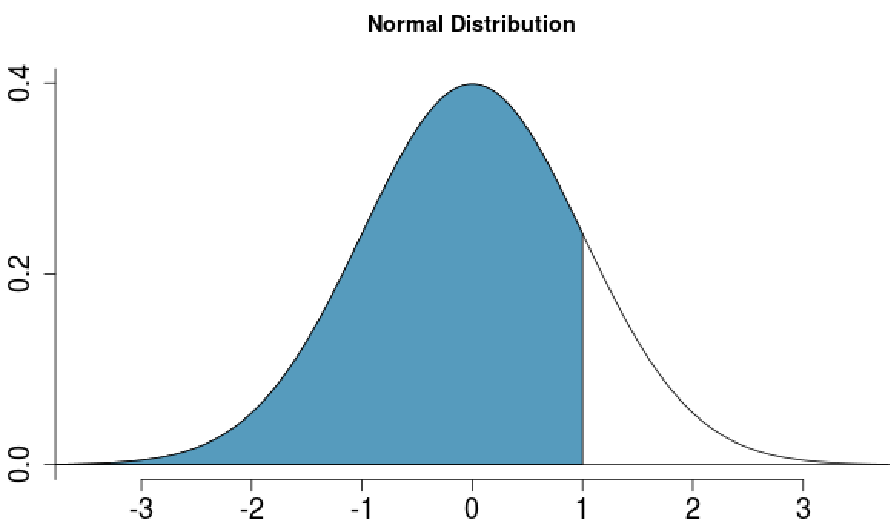
https://gallery.shinyapps.io/dist_calc/



P(X < 1) = 0.841

Practice : computing percentiles

SAT scores are distributed nearly normally with mean 1500 and standard deviation 300. Pam earned an 1800 on her SAT. What is Pam's percentile score?



P(X < 1) = 0.841

3. using Z table

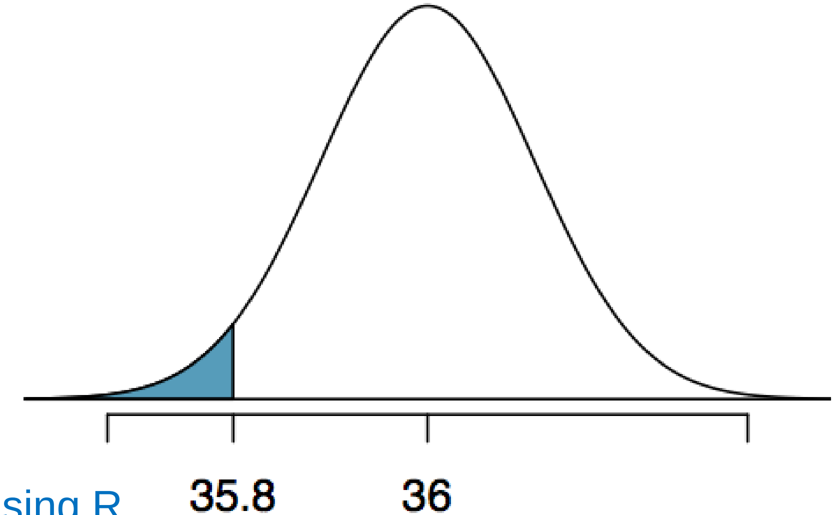
	Second decimal place of Z				
Z	0.00	0.01	0.02	0.03	0.04
0.0	0.5000	0.5040	0.5080	0.5120	0.5160
0.1	0.5398	0.5438	0.5478	0.5517	0.5557
0.2	0.5793	0.5832	0.5871	0.5910	0.5948
0.3	0.6179	0.6217	0.6255	0.6293	0.6331
0.4	0.6554	0.6591	0.6628	0.6664	0.6700
0.5	0.6915	0.6950	0.6985	0.7019	0.7054
0.6	0.7257	0.7291	0.7324	0.7357	0.7389
0.7	0.7580	0.7611	0.7642	0.7673	0.7704
0.8	0.7881	0.7910	0.7939	0.7967	0.7995
0.9	0.8159	0.8186	0.8212	0.8238	0.8264
1.0	*0.8413*	0.8438	0.8461	0.8485	0.8508
1.1	0.8643	0.8665	0.8686	0.8708	0.8729
1.2	0.8849	0.8869	0.8888	0.8907	0.8925

Practice

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.

1. What percent of bottles have less than 35.8 ounces of ketchup?

- *Let X = amount of ketchup in a bottle: X ~ N(μ = 36, σ = 0.11)*



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

using R

```
> pnorm(-1.82, mean = 0, sd = 1)
[1] 0.0344
```

using R

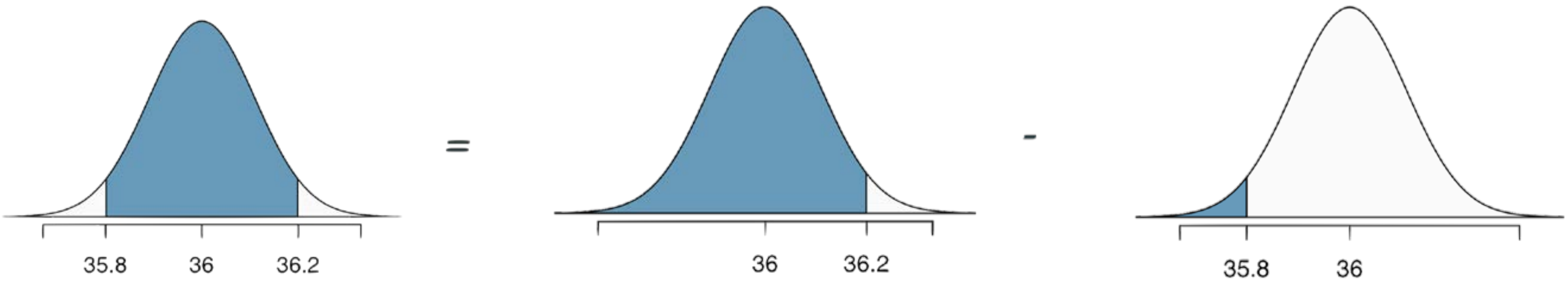
```
> pnorm(35.8, mean = 36, sd = 0.11)
[1] 0.0345
```

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

2. What percent of bottles pass the quality control inspection?



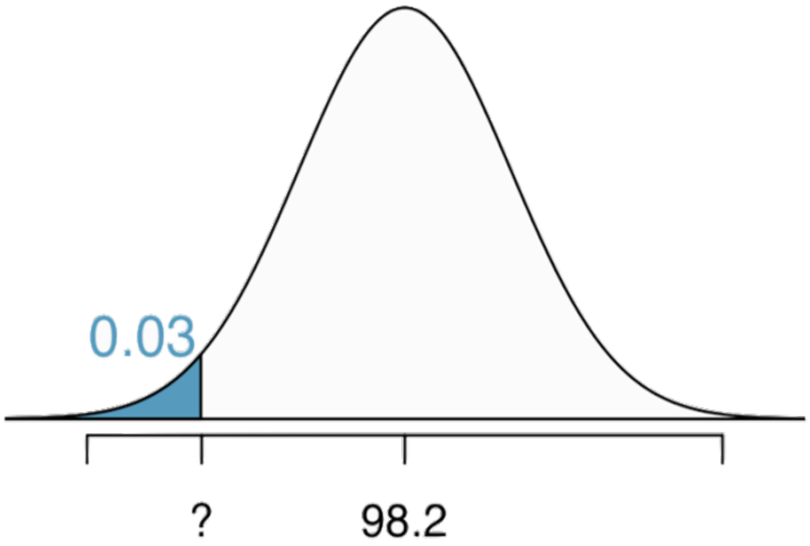
$$Z = \frac{35.8 - 36}{0.11} = -1.82 \qquad Z = \frac{36.2 - 36}{0.11} = 1.82$$

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

Practice: finding cutoff points

Body temperatures of healthy humans are distributed nearly normally with mean 98.2ºF and standard deviation 0.73ºF.

1. What is the cutoff for the lowest 3% of human body temperatures?



using R: qnorm

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

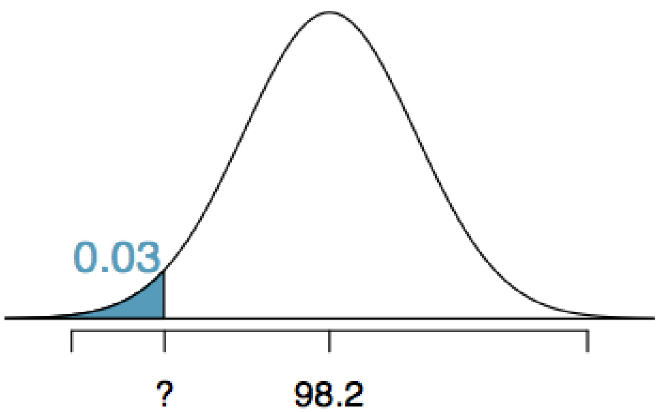
$$P(X < x) = 0.03 \rightarrow P(Z < -1.88) = 0.03$$

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

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

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.

finding cutoff points : using Z table



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

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

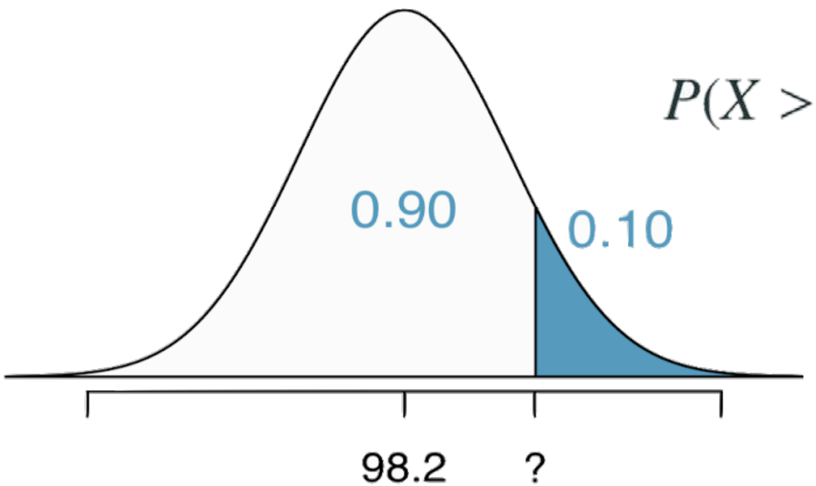
$$Z = \frac{obs - mean}{SD} \rightarrow \frac{x - 98.2}{0.73} = -1.88$$

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

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

2. What is the cutoff for the highest 10% of human body temperatures?



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

Homework Week 6

Problem 1: State the appropriate null hypothesis H_0 and alternative hypothesis H_A . The mean area of the several thousand apartment in a new development is advertised to be 1250 square feet. A tenant group thinks that the apartments are smaller than advertised. They hire an engineer to measure a sample of apartment to test their suspicion.

Problem 2: Find a news article or research report that conducts and describes a hypothesis test

Homework Week 6

Problem 3: The movie lengths are supposed to be normally distributed with mean 110.5 minutes and standard deviation 22.4 minutes. (You can us any of R, Applet and Z table.)

a) What fraction of movies are more than 2 hours long?
b) What fraction of movies are between one and a half hours and two hours long?
c) What is the cutoff for the lowest 10% of movie lengths?