Scores on first exam

Histogram of grades

$\bar{x} = 35.7$

$s = 12.0$

If a million of your peers took this test... (normal approx)

Histogram of fakegrades

$$\mu = 35.7$$
$$\sigma = 12.0$$
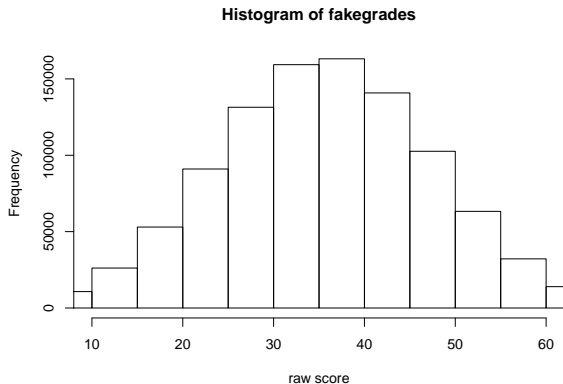
If a million of your peers took this test... (smaller bins)

Histogram of fakegrades

$\mu = 35.7$

$\sigma = 12.0$
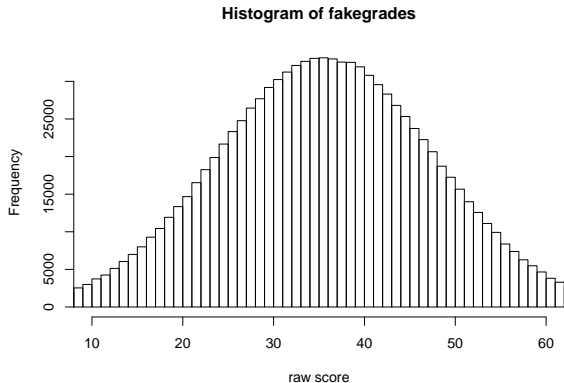
If a million of your peers took this test... with density

Histogram of fakegrades
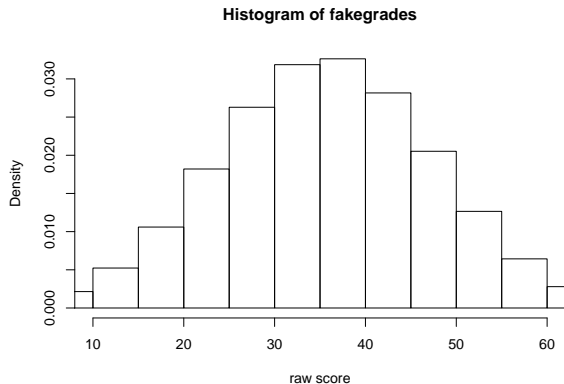
$\mu = 35.7$

$\sigma = 12.0$

Notice density stays constant with changes to bin widths

Histogram of fakegrades
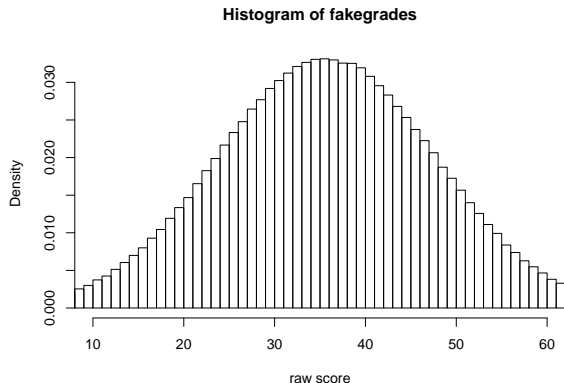
$$\mu = 35.7$$

$$\sigma = 12.0$$

With tiny widths (and many fake grades), we get a curve...

Histogram of fakegrades
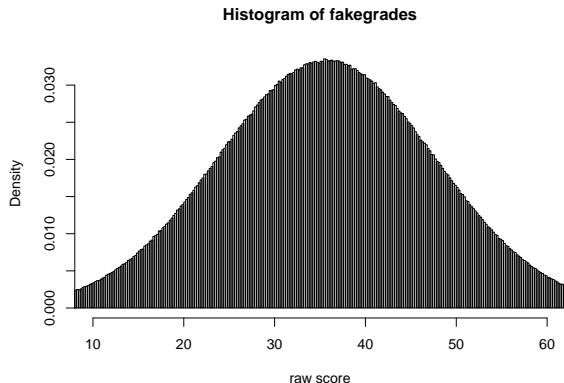
$$\mu = 35.7$$
$$\sigma = 12.0$$

What I consider fair for this exam...

| D- | D | D+ | C- | C | C+ | B- | B | B+ | A- | A | A+ |

$$\mu - \sigma \qquad \mu \qquad \mu + \sigma$$
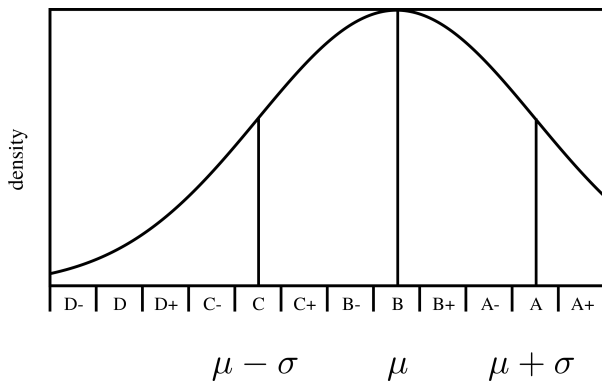
$$\mu = \text{B}$$

$\sigma =$ one letter grade

How to calculate your exam grade:

- Remember the sample mean and sample standard deviation.

$$\bar{x} = 35.7$$

$$s = 12.0$$

- Calculate your z-score (standardized score).

$$z = \frac{x - \bar{x}}{s}$$

- Calculate your exam grade:

$$\text{grade} = 10z + 85$$

How to estimate your percentile (find your theoretical percentile):

- Calculate your *z*-score (standardized score).

$$z = \frac{x - \bar{x}}{s}$$

- Find area under the standard normal curve from $-\infty$ to your *z*.

z=-3   z=-2   z=-1   z=0   z=1   z=2   z=3
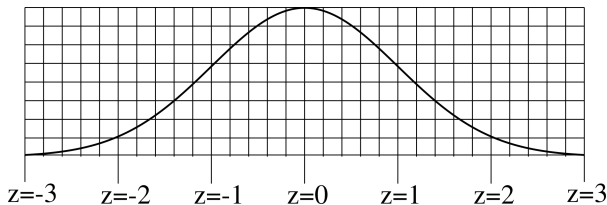
Normal distribution

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

Heights of males
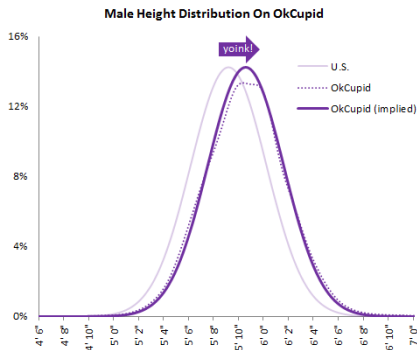

Male Height Distribution On OkCupid
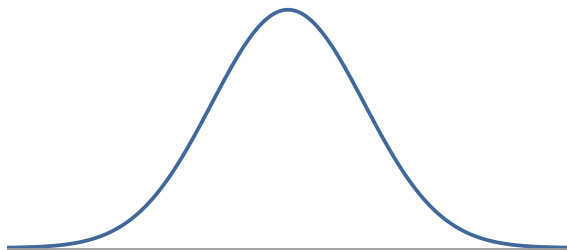
"The male heights on OkCupid very nearly follow the expected normal distribution – except the whole thing is shifted to the right of where it should be. Almost universally guys like to add a couple inches."

"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."

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

Heights of females

Female Height Distribution On OkCupid

Legend:
- U.S.
- OkCupid
- OkCupid (implied)
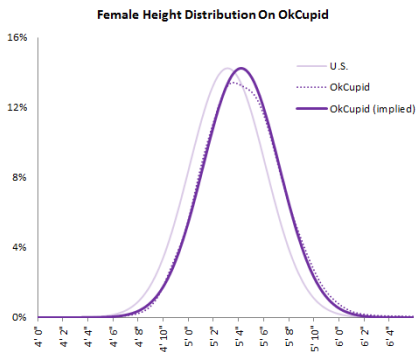
"When we looked into the data for women, we were surprised to see height exaggeration was just as widespread, though without the lurch towards a benchmark height."

Normal distributions with different parameters

μ: mean, σ: standard deviation

$N(\mu = 0, \sigma = 1)$ $N(\mu = 19, \sigma = 4)$
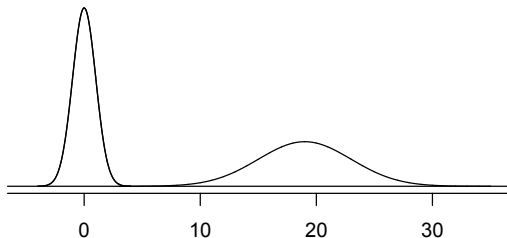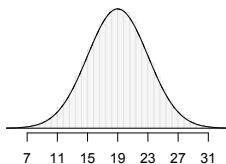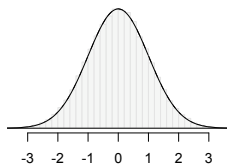
SAT scores are distributed nearly normally with mean 1500 and standard deviation 300. ACT scores are distributed nearly normally with mean 21 and standard deviation 5. A college admissions officer wants to determine which of the two applicants scored better on their standardized test with respect to the other test takers: Pam, who earned an 1800 on her SAT, or Jim, who scored a 24 on his ACT?
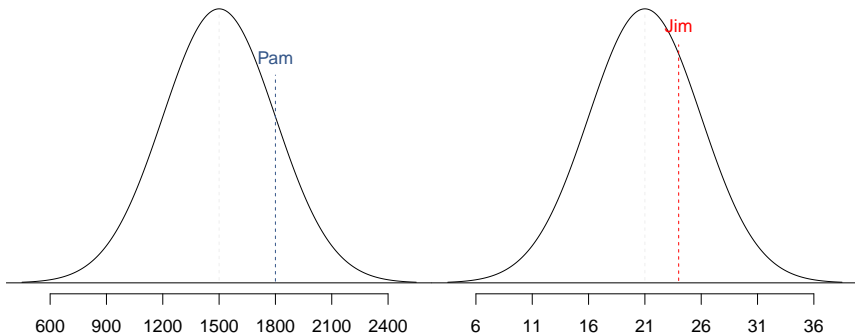
Standardizing with Z scores

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.
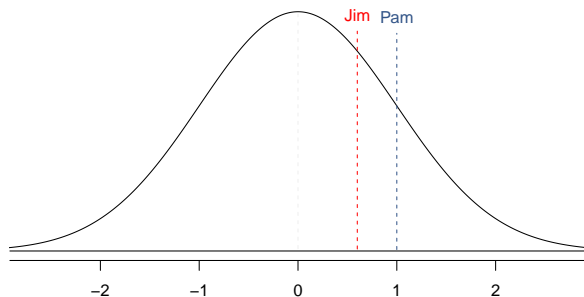
Standardizing with Z scores (cont.)

- ► 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, 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.
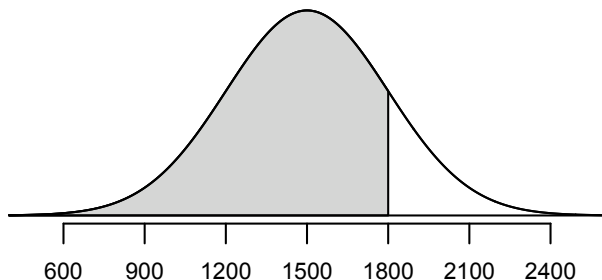
Calculating percentiles - using computation

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

- R:

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

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

Calculating percentiles - using tables

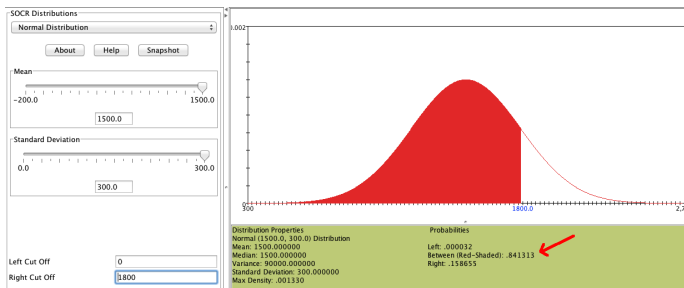| Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
				Second decimal place of Z						
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	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

Six sigma

"The term *six sigma process* comes from the notion that if one has six standard deviations between the process mean and the nearest specification limit, as shown in the graph, practically no items will fail to meet specifications."

$$6\sigma$$

http://en.wikipedia.org/wiki/Six_Sigma

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?

Finding the exact probability - using the Z table

| 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	
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%
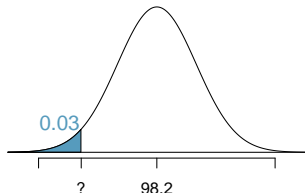
(d) 93.12%

(e) 96.56%

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

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

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

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

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.
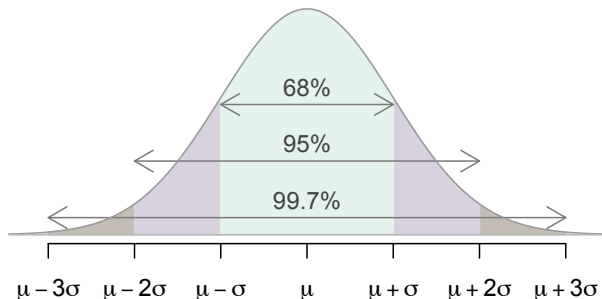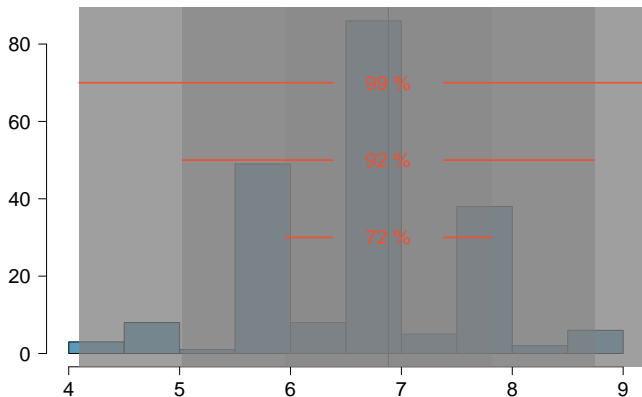
Describing variability using the 68-95-99.7 Rule

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

- ~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.
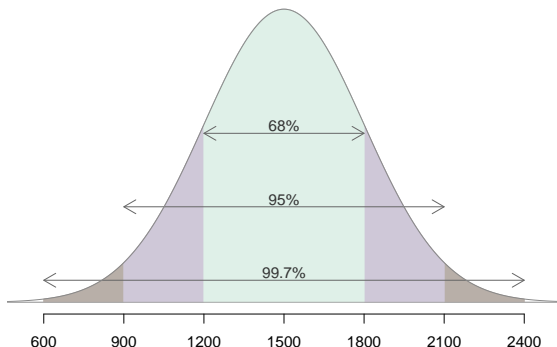
Number of hours of sleep on school nights

- Mean = 6.88 hours, SD = 0.92 hrs
- 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: $6.88 \pm 3 \times 0.93$

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 $2 \times SD$.
(d) Z scores are helpful for determining how unusual a data point is compared to the rest of the data in the distribution.