Random Variables

Random Variables

- Functions from the sample space to the real line
- Facilitates talking about complex sample events
- We can define independence, conditional independence as usual
- Pairs (or more) of random variables have joint probability distributions
  - We can compute marginal, conditional distributions
- Expectation, Variance summarize features of the distribution
- Covariance, Correlation summarize linear dependence between variables

Practice

A casino game costs $5 to play. If you draw first a red card, then you get to draw a second card. If the second card is the ace of hearts, you win $500. If not, you don't win anything, i.e. lose your $5. What is your expected profits (or losses) from playing this game? Remember: profit (or loss) = winnings - cost.

(a) a loss of 10¢

(b) a loss of 25¢

(c) a loss of 30¢

(d) a profit of 5¢

Practice

A casino game costs $5 to play. If you draw first a red card, then you get to draw a second card. If the second card is the ace of hearts, you win $500. If not, you don't win anything, i.e. lose your $5. What is your expected profits (or losses) from playing this game? Remember: profit (or loss) = winnings - cost.

(a) a loss of 10¢

(c) a loss of 30¢

(b) a loss of 25¢

(d) a profit of 5¢

Event	*Win*	*Profit: X*	$P(X)$	$X \times P(X)$
Red, A♥	*500*	*500 - 5 = 495*	$\frac{25}{52} \times \frac{1}{51} = 0.0094$	$495 \times 0.0094 = 4.653$
Other	*0*	*0 - 5 = -5*	$1 - 0.0094 = 0.9906$	$-5 \times 0.9906 = -4.953$
				$E(X) = -0.3$

Fair game

A *fair* game is defined as a game that costs as much as its expected payout, i.e. expected profit is 0.

Do you think casino games in Vegas cost more or less than their expected payouts?

If those games cost less than their expected payouts, it would mean that the casinos would be losing money on average, and hence they wouldn't be able to pay for all this:

Image by Moyan_Brenn on Flickr
http://www.flickr.com/photos/aigle_dore/5951714693

Calculating the expectation of a linear combination

On average you take 10 minutes for each statistics homework problem and 15 minutes for each chemistry homework problem. This week you have 5 statistics and 4 chemistry homework problems assigned. What is the total time you expect to spend on statistics and physics homework for the week?

Calculating the expectation of a linear combination

On average you take 10 minutes for each statistics homework problem and 15 minutes for each chemistry homework problem. This week you have 5 statistics and 4 chemistry homework problems assigned. What is the total time you expect to spend on statistics and physics homework for the week?

$$E(S_1 + S_2 + S_3 + S_4 + S_5 + C_1 + C_2 + C_3 + C_4) = 5 \times E[S_1] + 4 \times E(C_1)$$
$$= 5 \times 10 + 4 \times 15$$
$$= 50 + 60$$
$$= 110\, min$$

Linear combinations

- A *linear combination* of random variables X and Y is given by

$$aX + bY$$

where a and b are some fixed numbers.

Linear combinations

- A *linear combination* of random variables X and Y is given by

$$aX + bY$$

  where a and b are some fixed numbers.

- The average value of a linear combination of random variables is given by

$$E(aX + bY) = a\,E(X) + b\,E(Y)$$

Linear combinations

- A *linear combination* of random variables X and Y is given by

$$aX + bY$$

where a and b are some fixed numbers.

- The average value of a linear combination of random variables is given by

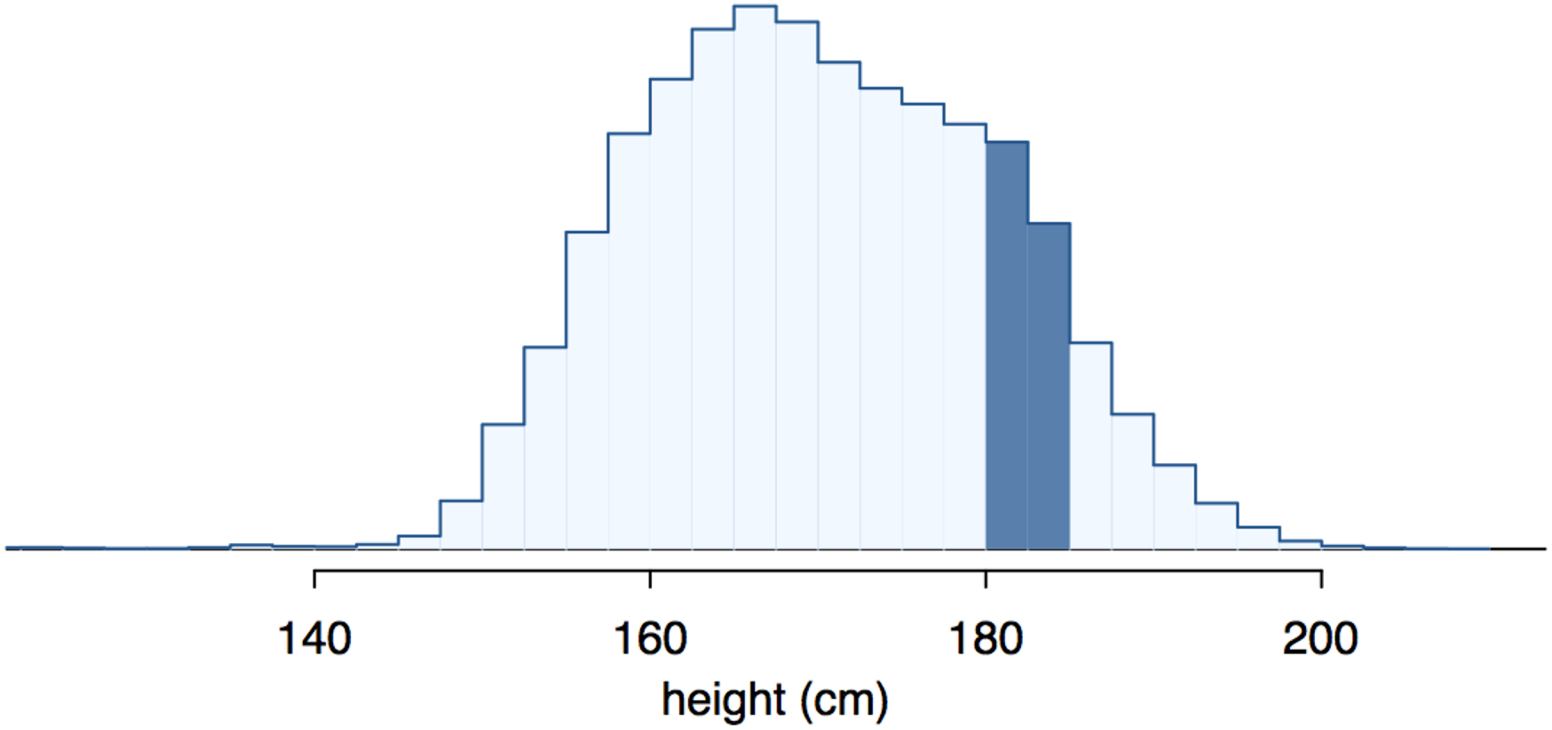
$$E(aX + bY) = a\,E(X) + b\,E(Y)$$

- The variance of a linear combination of random variables is given by

$$Var(aX + bY) = a^2\,Var(X) + b^2 Var(Y)$$

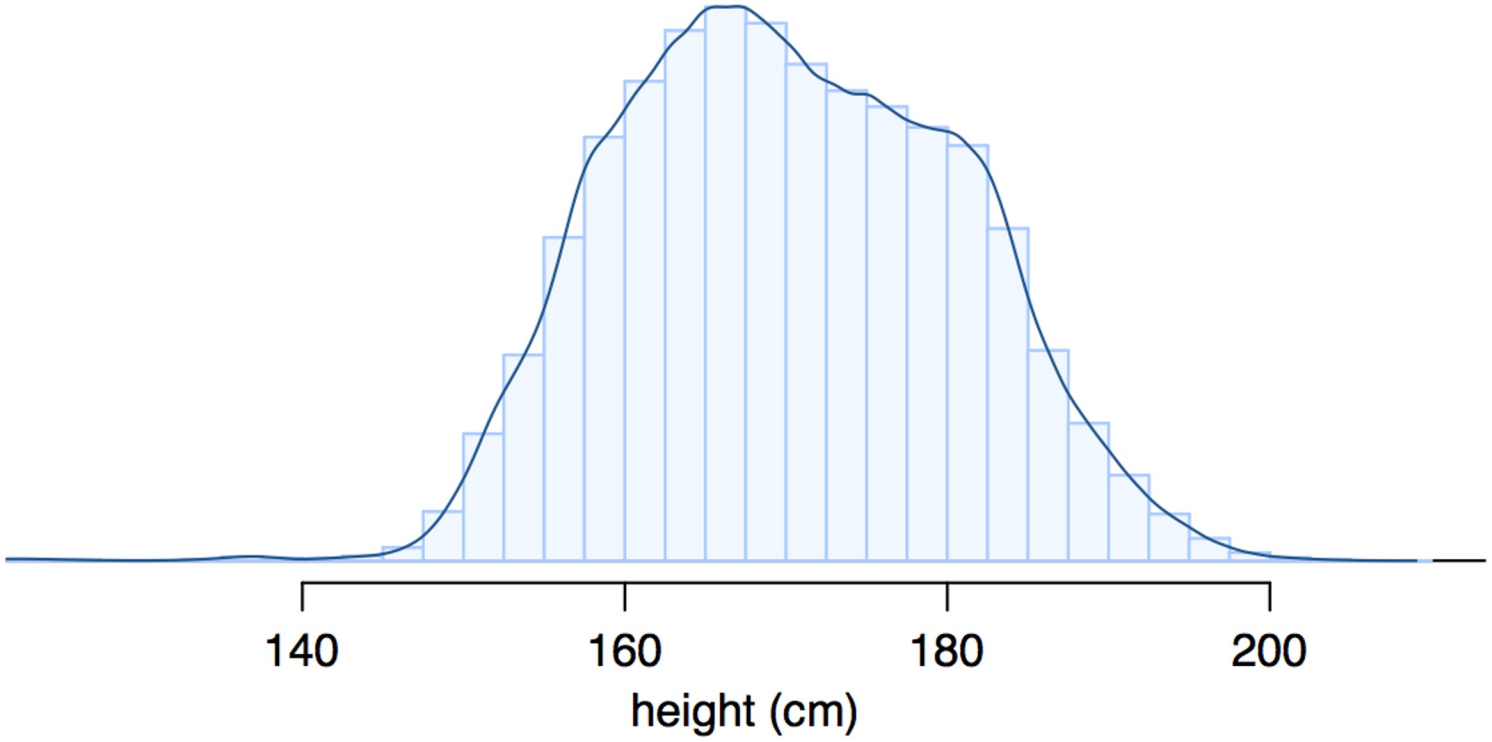
Continuous distributions

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



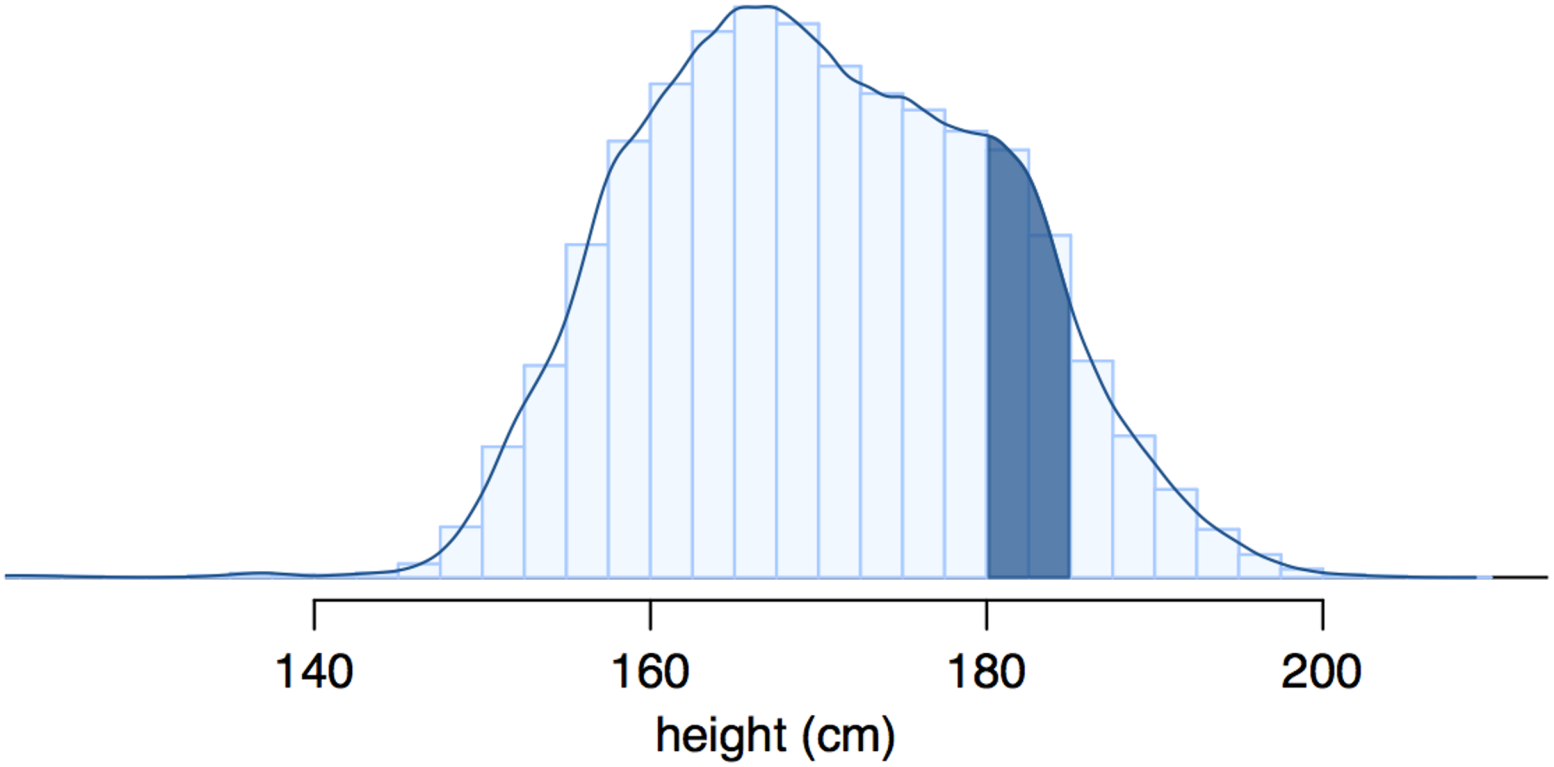
From histograms to continuous distributions

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



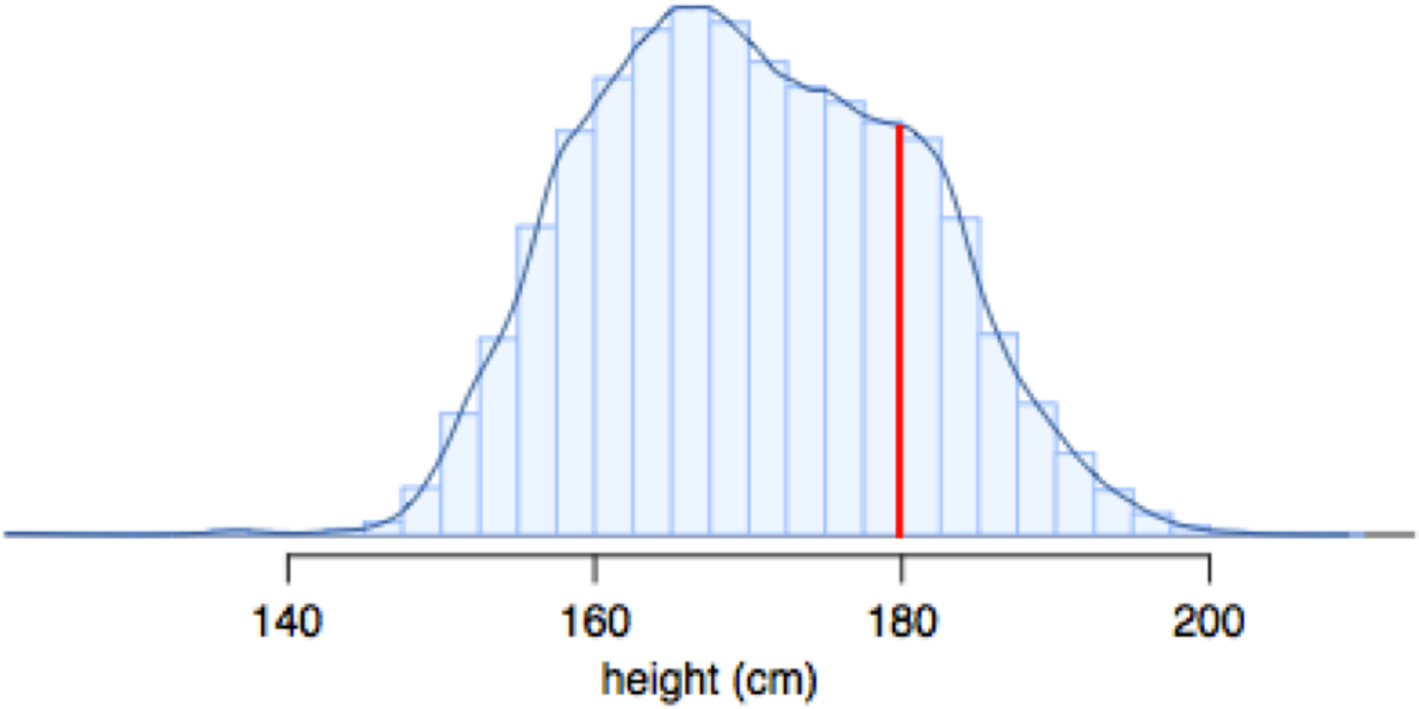
Probabilities from continuous distributions

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

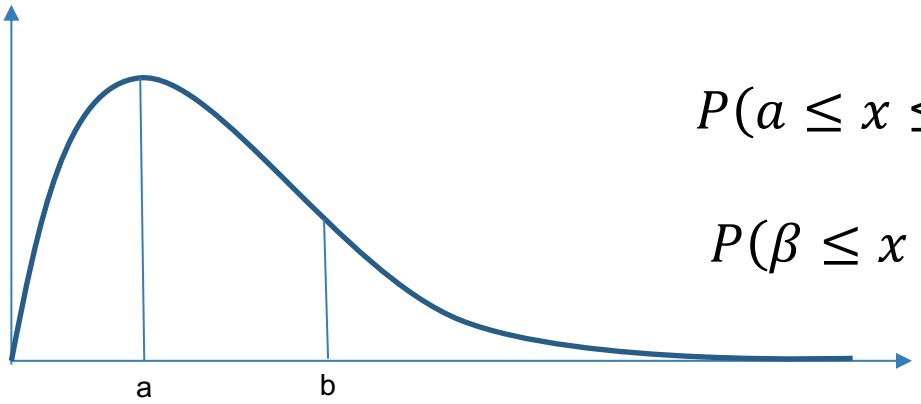


By definition...

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



Probability Density Function



$$P(a \leq x \leq a+\epsilon) = \int_{a}^{a+\epsilon} p(x)dx = p(a)\epsilon$$

$$P(\beta \leq x \leq \beta+\epsilon) = \int_{\beta}^{\beta+\epsilon} p(x)dx = p(\beta)\epsilon$$

- Events represented as intervals $\alpha \leq X < b$
- Probability $\int_a^b p(X = x)dx$
- Specific outcomes have zero probability, non-negative probability density

Continuous Probability

Most definitions for discrete RVs hold, replacing PMF with PDF

Two RVs X & Y are **independent** if and only if,

$$p(x, y) = p(x)p(y)$$

Conditionally independent given Z iff,

$$p(x, y \mid z) = p(x \mid z)p(y \mid z)$$

Probability chain rule,

$$p(x, y) = p(x)p(y \mid x)$$

Continuous Probability

…and by replacing summation with integration…

Law of Total Probability for continuous distributions,

$$p(x) = \int_{\mathcal{Y}} p(x, y)\, dy$$

Expectation of a continuous random variable,

$$\mathbf{E}[X] = \int_{\mathcal{X}} x \cdot p(x)\, dx$$

Covariance of two continuous random variables X & Y,

$$\mathbf{Cov}(X, Y) = \mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])] = \int_{\mathcal{X}} \int_{\mathcal{Y}} (x - \mathbf{E}[X])(y - \mathbf{E}[Y]) p(x, y)\, dx dy$$

Mixed Variables

- *Let X be the consumption of Soda (in ml)*
- *What does this look like?*
- *Is this a pdf or a pmf?*

CDF

Definition *The cumulative distribution function (CDF) of a real-valued continuous RV X is the function given by,*

$$F(x) = P(X \leq x)$$

- Can easily measure probability of closed intervals,

$$P(a \leq X < b) = F(b) - F(a)$$

- If *X* is *absolutely continuous* (i.e. differentiable) then,

$$f(x) = \frac{dF(x)}{dx} \qquad \text{and} \qquad F(t) = \int_{-\infty}^{t} f(x)\,dx$$

Where $f(x)$ is the *probability density function* (PDF)

- Typically use shorthand P for CDF and p for PDF instead of F and f

Quantile Function

Definition (The Quantile Function)

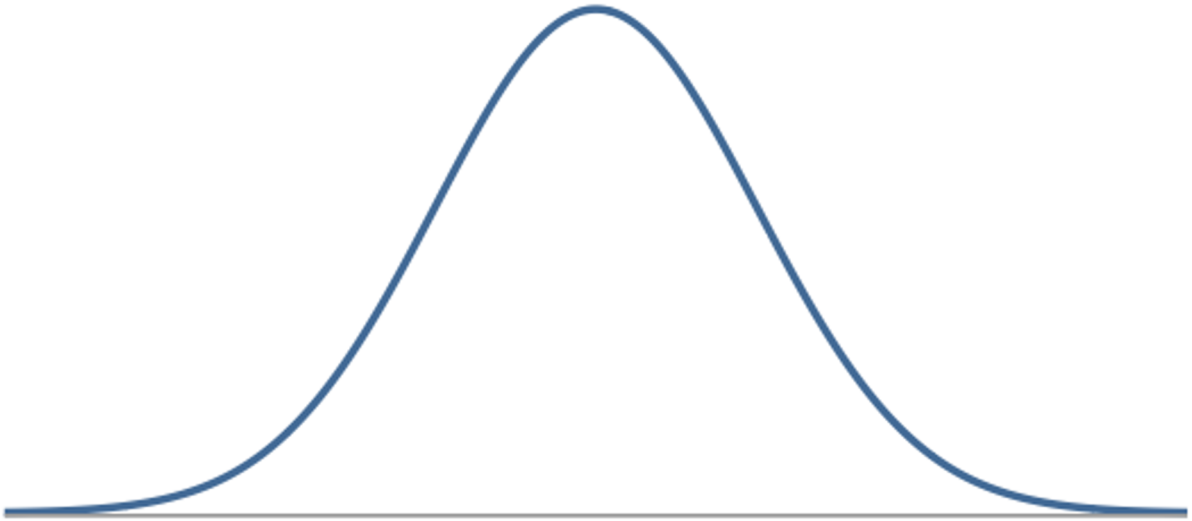
Let X be a random variable with CDF F. For each p strictly between 0 and 1, define $F^{-1}(p)$ *to be the smallest value* x *for which* $F(X) \geq p$. *Then* $F^{-1}(p)$ *is called the* p*-quantile of* X *or the* $100p$ *percentile of* X. *The function* F^{-1} *defined on (0,1) is called the quantile function.*

Example: Compute the quantile function for the uniform distribution.

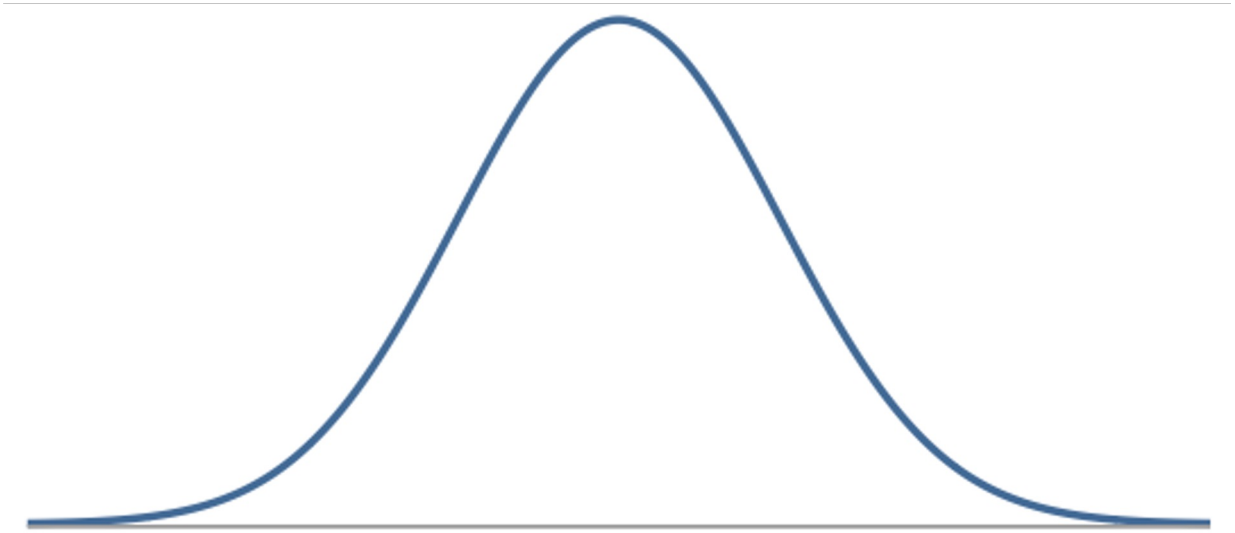
Normal distribution

Normal Distribution

- Unimodal and symmetric, bell-shaped curve
- Many variables are nearly normal
- Denoted as $N(\mu, \sigma^2)$ → Normal with mean μ and variance σ^2



Normal Distribution



Standard normal: $\mathcal{N}(0,1)$: $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$

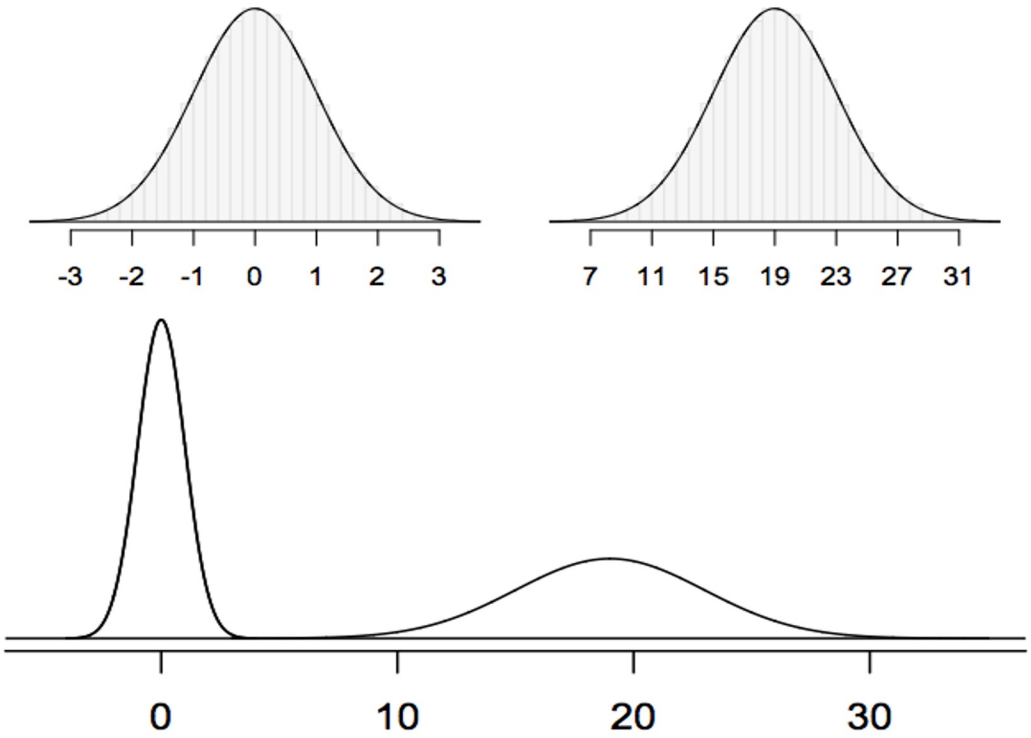
Normal with mean μ and variance σ^2 $\mathcal{N}(\mu, \sigma^2)$: $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$

$E(X) = \mu, \mathrm{Var}(X) = \sigma^2$

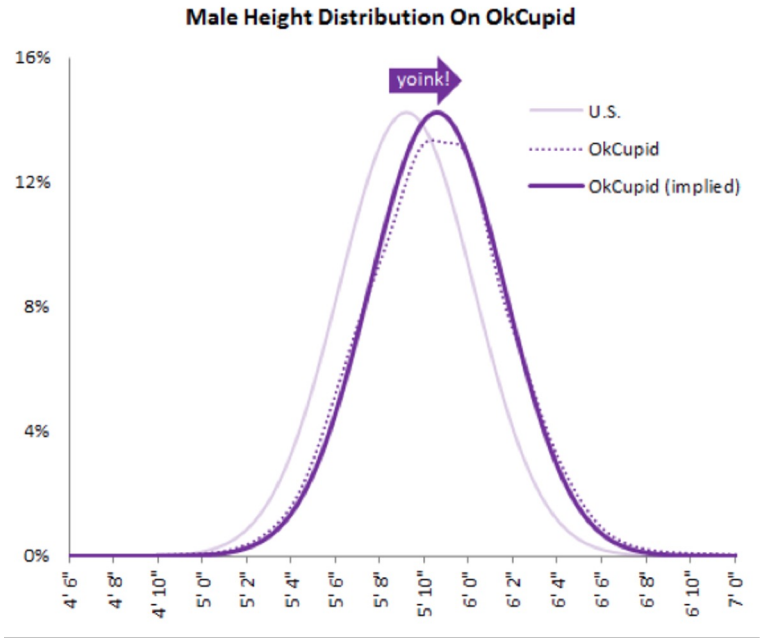
Normal distributions with different parameters

μ: mean, σ: standard deviation

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



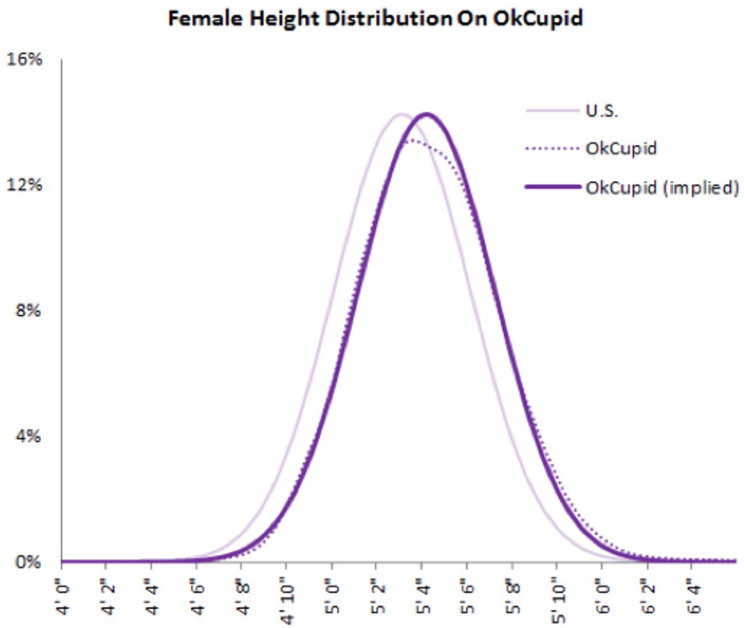
Heights of males



"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."

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

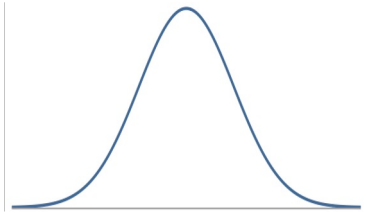
Heights of females



"When we looked into the data for women, we were surprised to see height exaggeration was just as widespread."

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

Normal Distribution



Standard normal: $\mathcal{N}(0,1)$: $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$

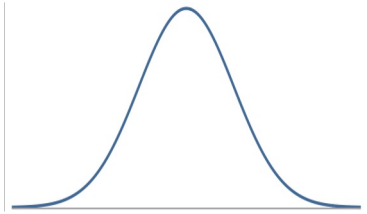
Normal with mean μ and variance σ^2 $\quad \mathcal{N}(\mu, \sigma^2)$: $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$

$$E(X) = \mu, \mathrm{Var}(X) = \sigma^2$$

Theorem (Linear transformations of a normal are normal)
If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $\alpha X + \beta \sim N(?, ?)$

Theorem (The sum of independent normals is normal)
If the random variables $X_1, \ldots, X_k$ are independent and if $X_i \sim N(\mu_i, \sigma_i^2)$ then
$X_1 + \cdots + X_k \sim N(?, ?)$

Normal Distribution



Standard normal: $\mathcal{N}(0,1)$: $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$

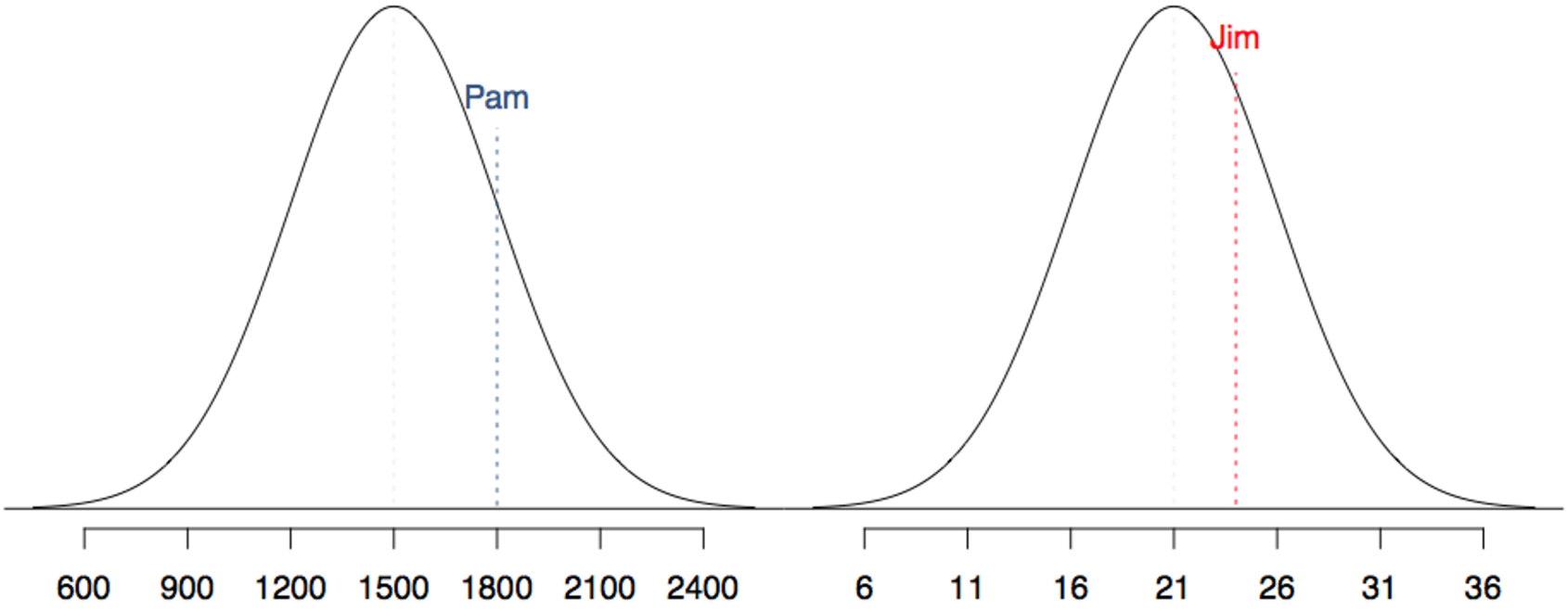
Normal with mean μ and variance σ^2 $\quad \mathcal{N}(\mu, \sigma^2)$: $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$

$E(X) = \mu, \mathrm{Var}(X) = \sigma^2$

Theorem (Linear transformations of a normal are normal)
If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $\alpha X + \beta \sim N(\alpha\mu + \beta, \alpha^2\sigma^2)$

Theorem (The sum of independent normals is normal)
If the random variables $X_1, \ldots, X_k$ are independent and if $X_i \sim N(\mu_i, \sigma_i^2)$ then
$X_1 + \cdots + X_k \sim N(\mu_1 + \cdots + \mu_k, \sigma_1^2 + \cdots + \sigma_k^2)$

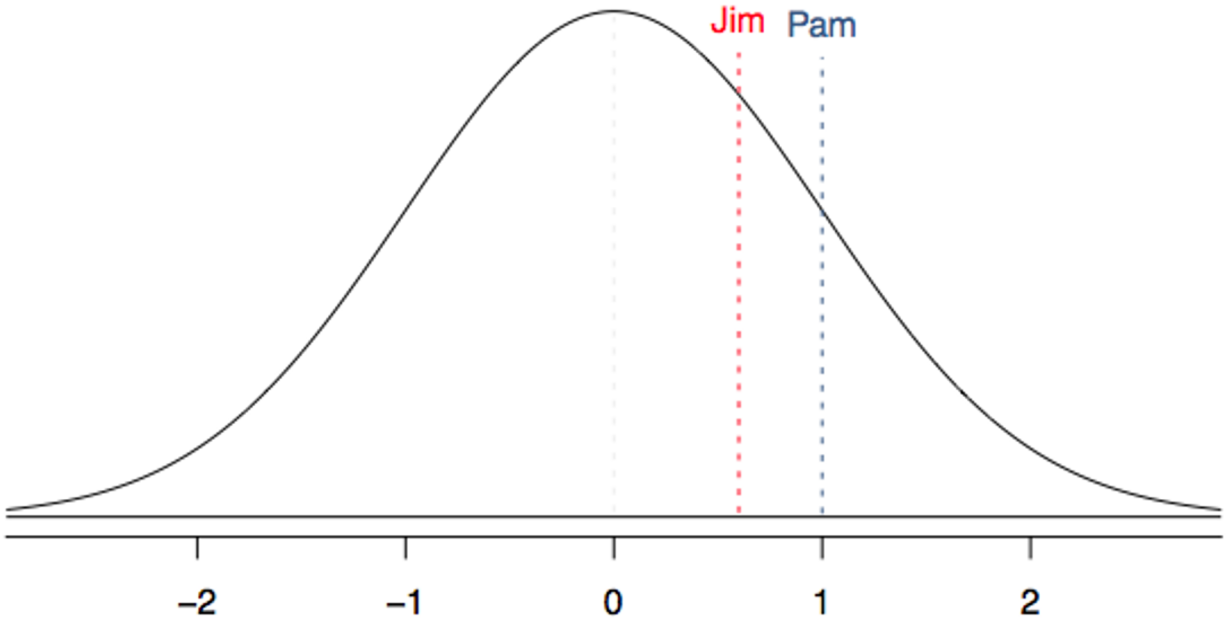
SAT scores are distributed nearly normally with mean 1500 and standard deviation 300. ACT scores are distributed nearly normally with mean 21 and standard deviation 5. A college admissions officer wants to determine which of the two applicants scored better on their standardized test with respect to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

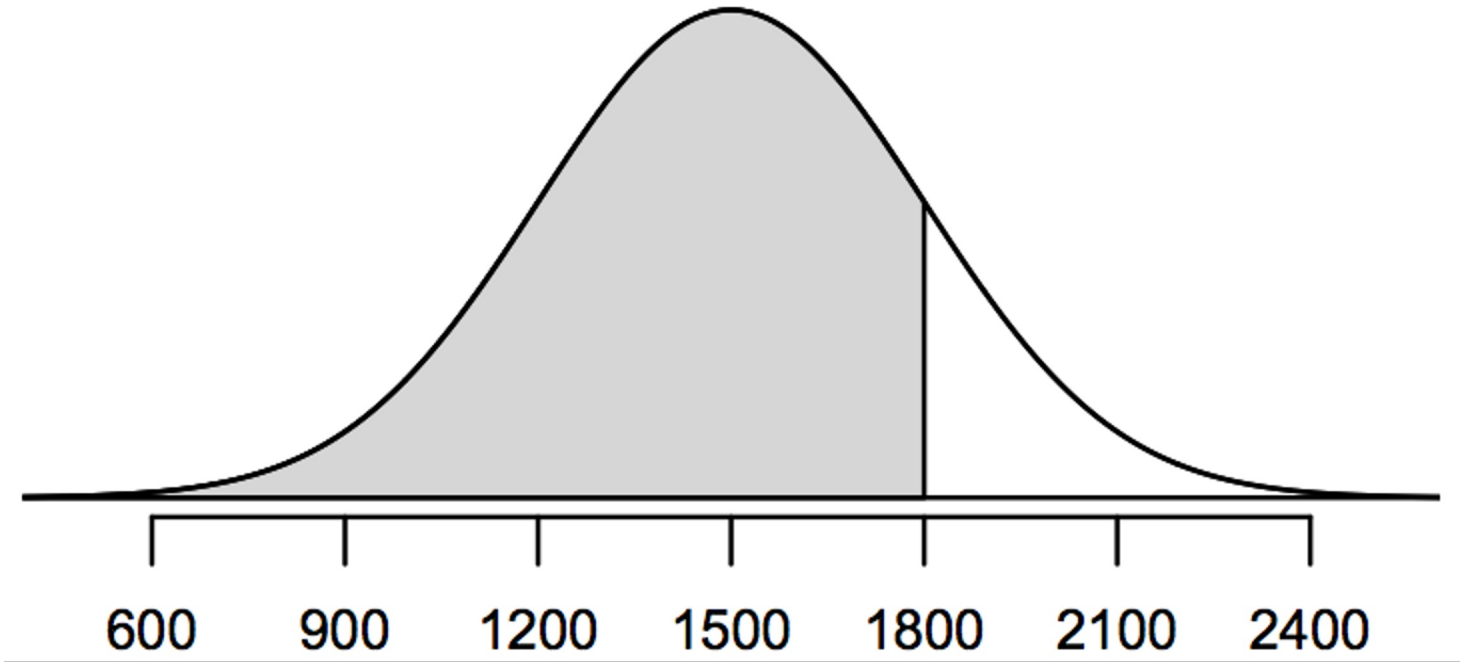
- Z score of an observation is the number of standard deviations it falls above or below the mean.

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

- 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

- $a-th$ *Percentile* is the number (score) under which $a\%$ of the observations fall.
- Graphically, the $a-th$ percentile is the value for which below the probability distribution curve to the left of that

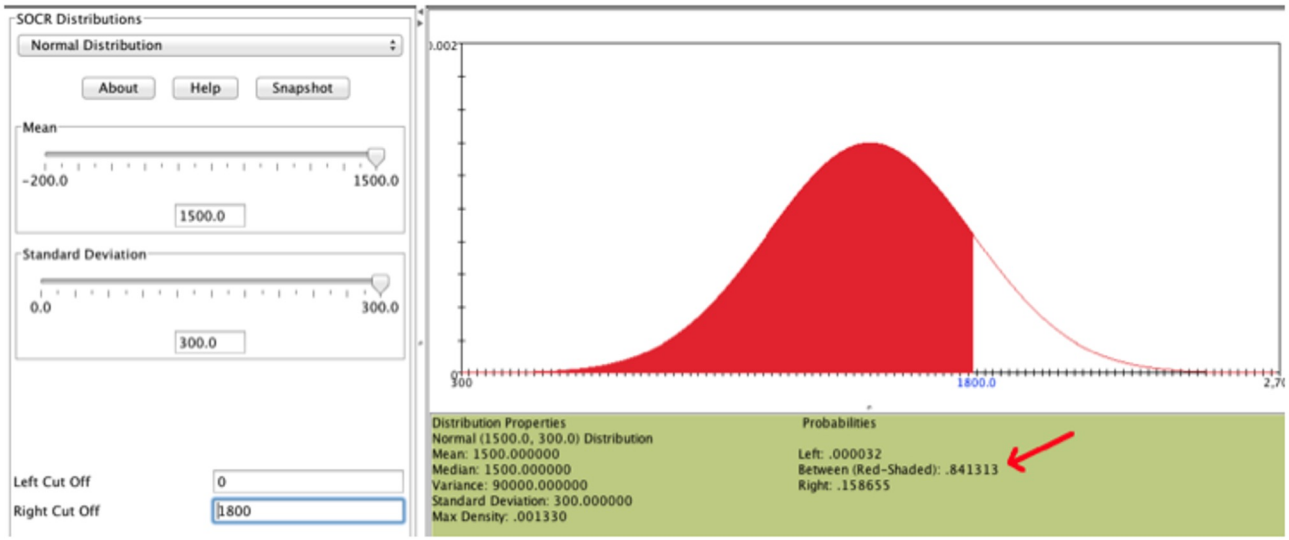


Calculating CDF/percentiles - using computation

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

```
> 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

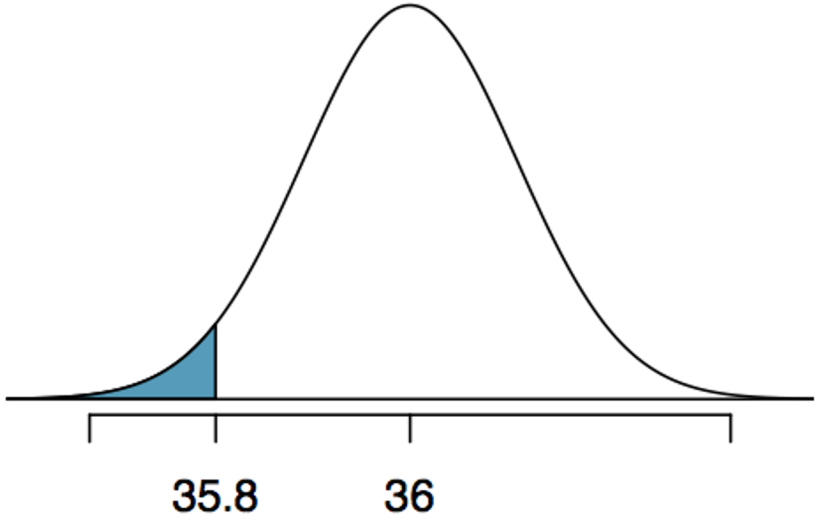
Quality control

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 ~ N(μ = 36, σ = 0.11)*



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

Finding the exact probability - using R

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

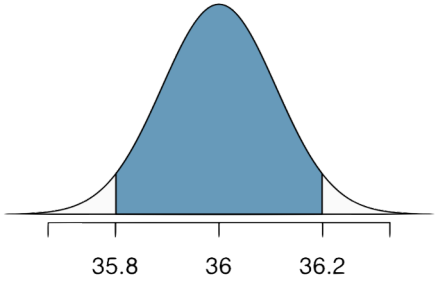
b) 3.44%

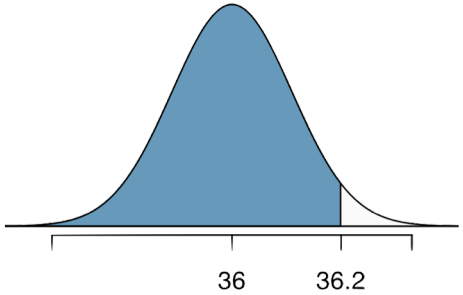
c) 6.88%

d) *93.12%*

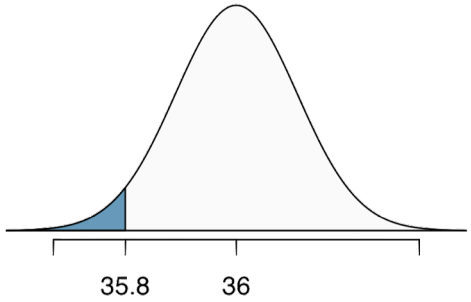
e) 96.56%

Practice

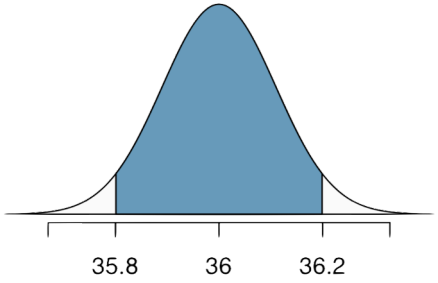




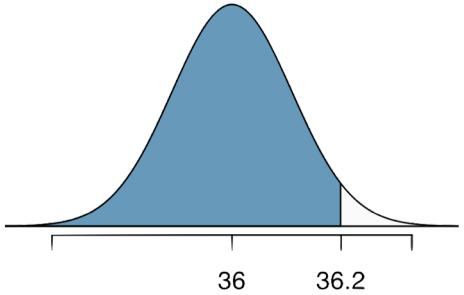
-



Practice

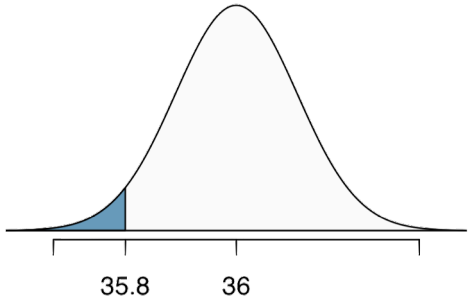


$$Z_{36.2} = \frac{36.2 - 36}{0.11} = 1.82$$



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

-



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

Practice

What percent of bottles pass the quality control inspection?

a) 1.82%

b) 3.44%

c) 6.88%

d) *93.12%*

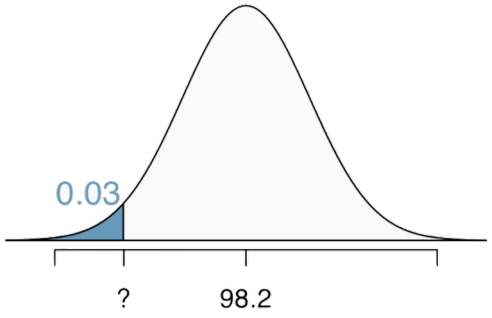
e) 96.56%

Practice: Percentiles

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?



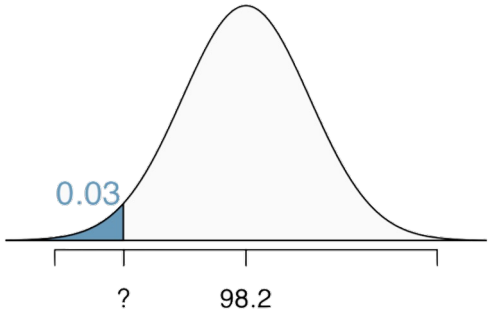
$$P(X < x) = 0.03 \rightarrow P(Z < -1.88) = 0.03$$

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

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

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?



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 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

Practice

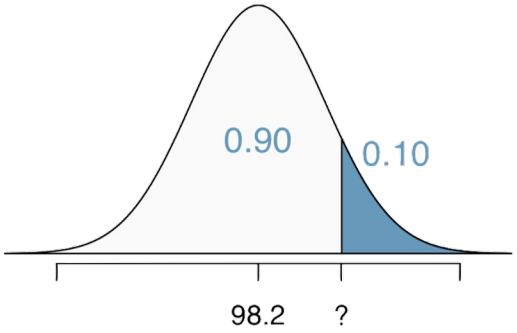
Body temperatures of healthy humans are distributed nearly nor- mally with mean 98.2°F and standard deviation 0.73°F. What is the cutoff for the highest 10% of human body temperatures?

(a) 97.3°F

(b) 99.1°F

(c) 99.4°F

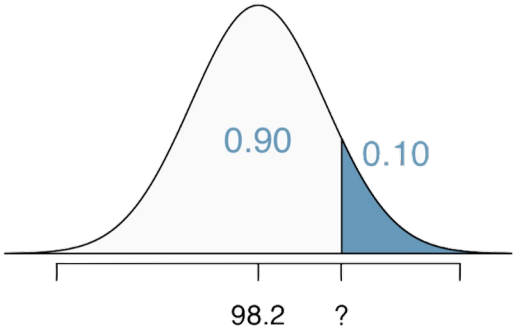
(d) 99.6°F



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



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

Practice

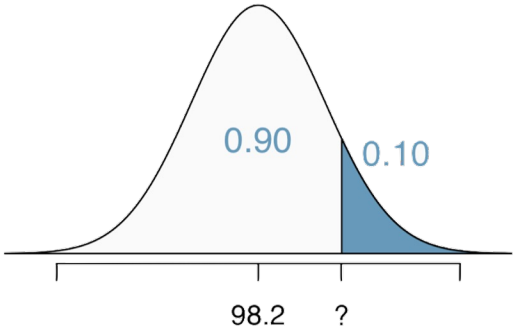
Body temperatures of healthy humans are distributed nearly nor- mally with mean 98.2°F and standard deviation 0.73°F. What is the cutoff for the highest 10% of human body temperatures?

(a) 97.3°F

(b) 99.1°F

(c) 99.4°F

(d) 99.6°F



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

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

Practice

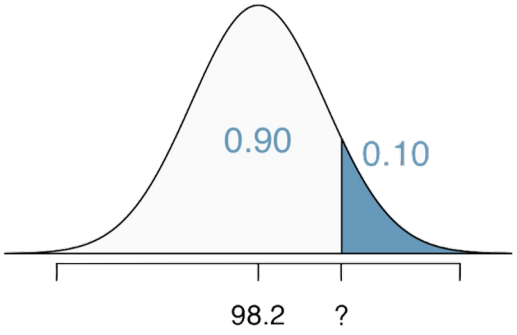
Body temperatures of healthy humans are distributed nearly nor- mally with mean 98.2°F and standard deviation 0.73°F. What is the cutoff for the highest 10% of human body temperatures?

(a) 97.3°F

(b) 99.1°F

(c) 99.4°F

(d) 99.6°F



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

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

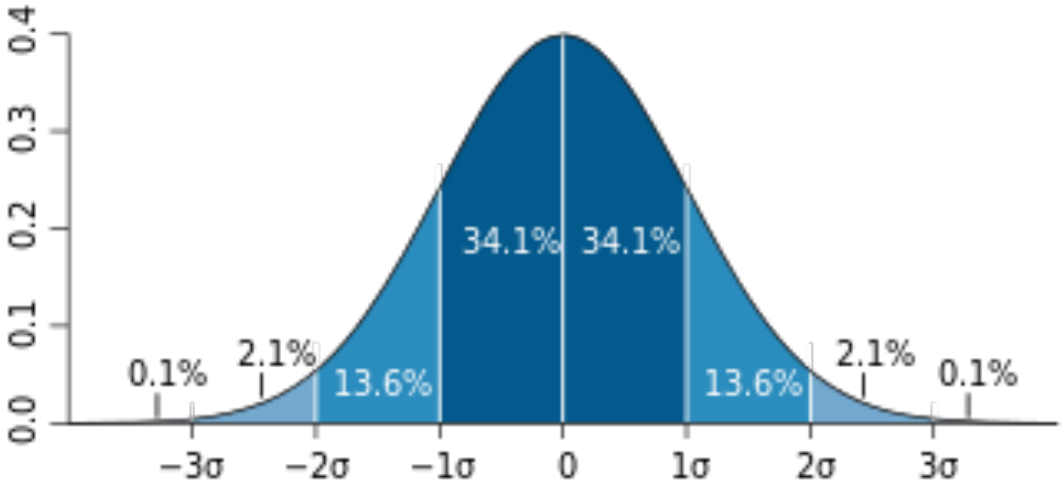
$$x = (1.28 \times 0.73) + 98.2 = 99.1$$

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.

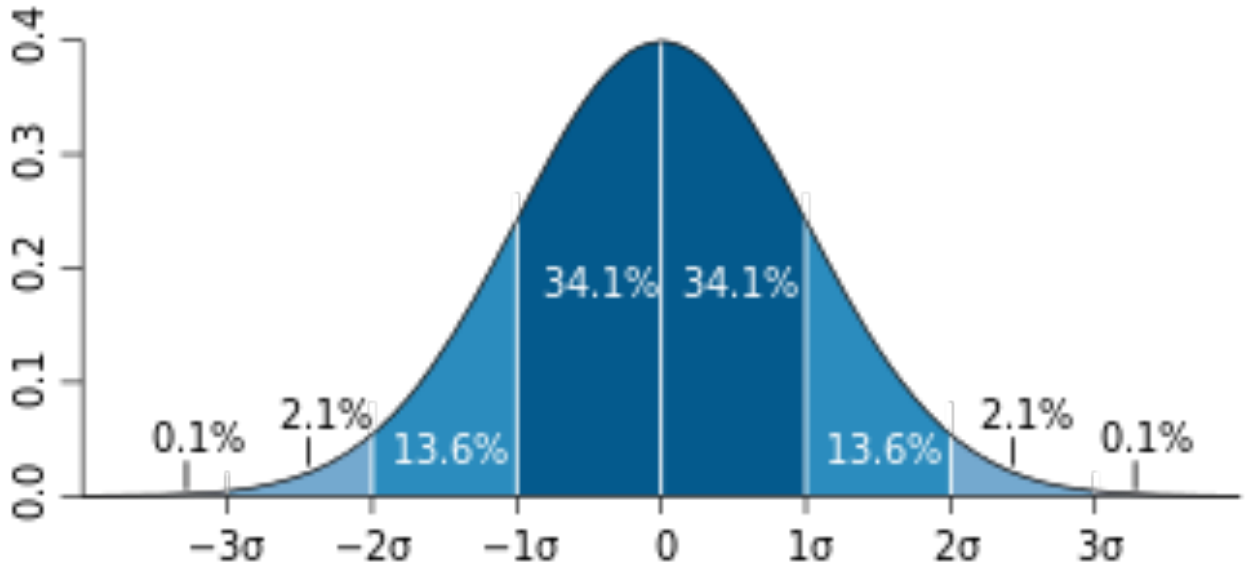
It is possible for observations to fall 4, 5, or more standard deviations away from the mean, but these occurrences are very rare if the data are nearly normal.



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.



Practice

Are the statements true or 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, Interquanile Range 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.