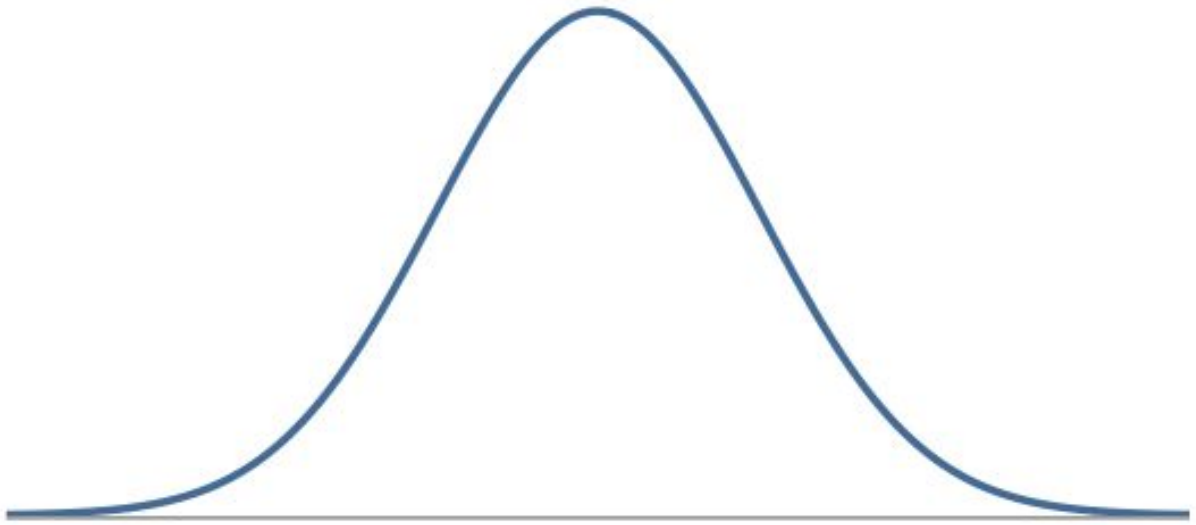
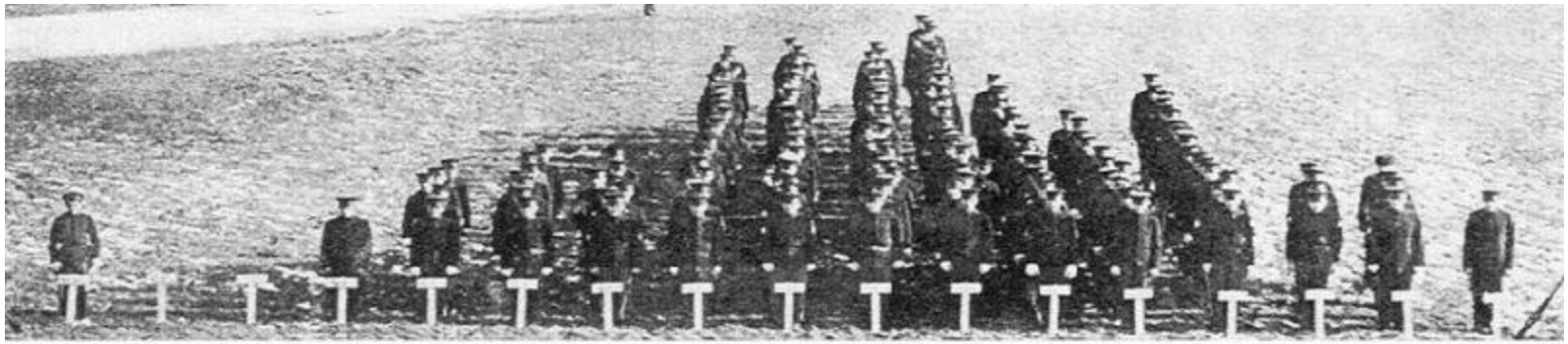
Normal distribution

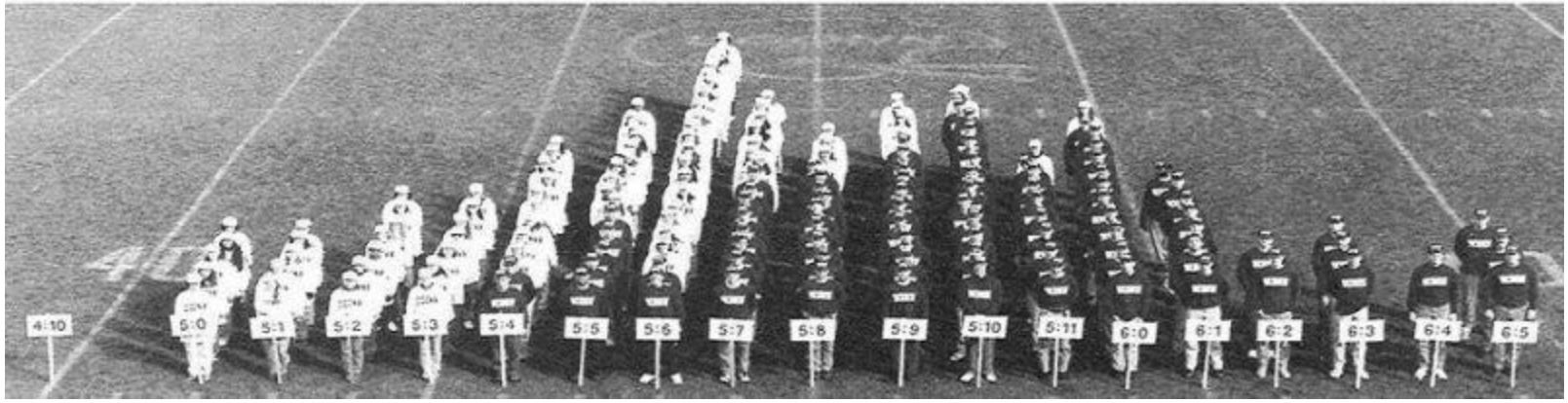
Normal Distribution

- 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

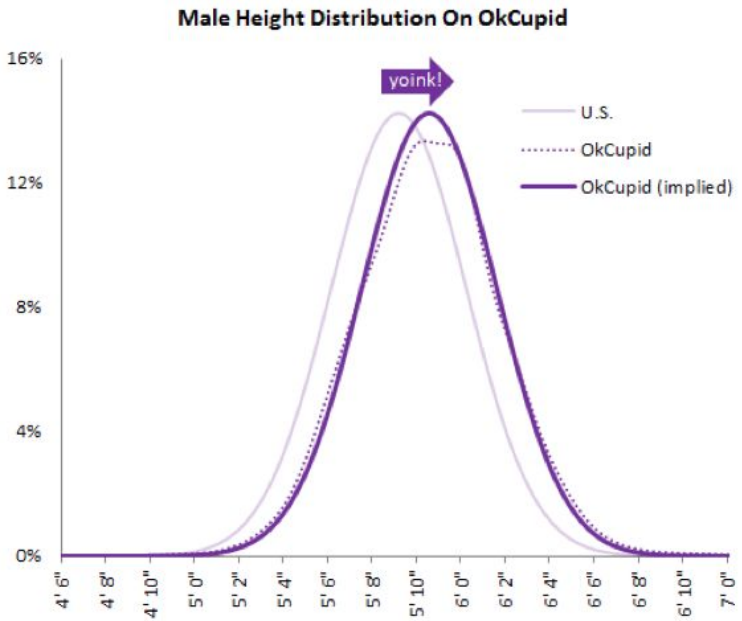




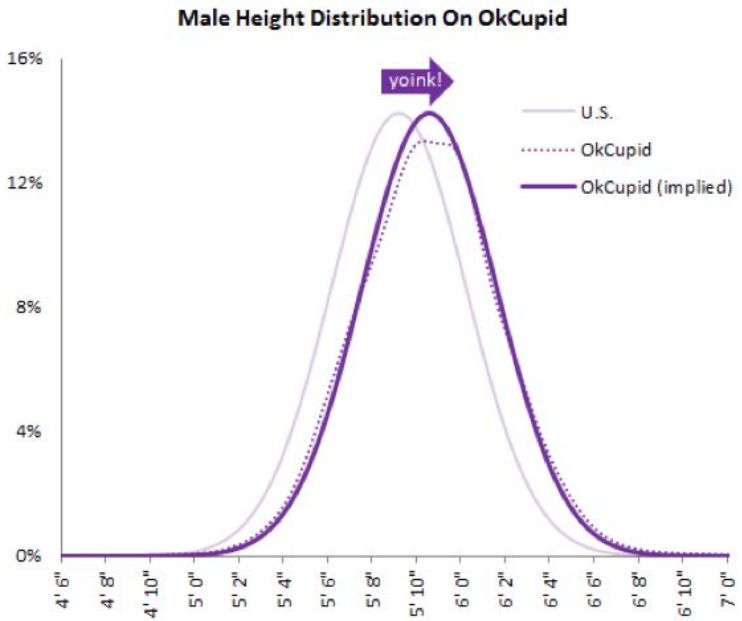
4:10 4:11 5:0 5:1 5:2 5:3 5:4 5:5 5:6 5:7 5:8 5:9 5:10 5:11 6:0 6:1 6:2



Heights of males



Heights of males

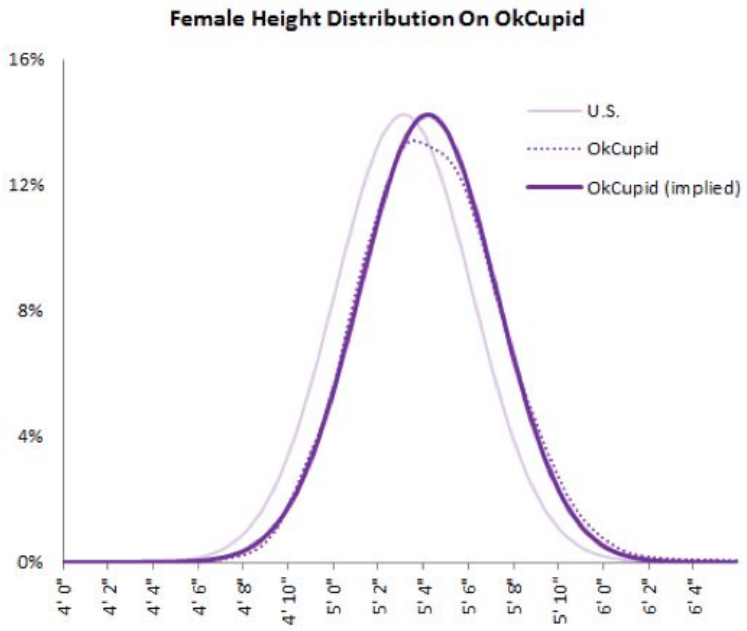


"The male heights on OkCupid very nearly follow the expected normal distribution -- except the whole thing is shifted to the right of where it should be. Almost universally guys like to add a couple inches."

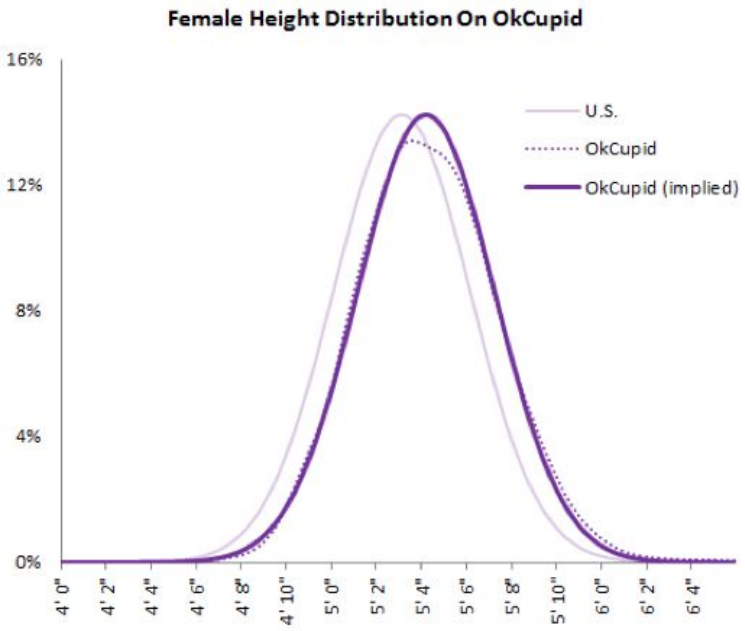
"You can also see a more subtle vanity at work: starting at roughly 5' 8", the top of the dotted curve tilts even further rightward. This means that guys as they get closer to six feet round up a bit more than usual, stretching for that coveted psychological benchmark."

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

Heights of females



Heights of females



"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

Why is the normal distribution ubiquitous?

Approximate normality of many biological variables of interest is likely consequence of the **Central Limit Theorem**:

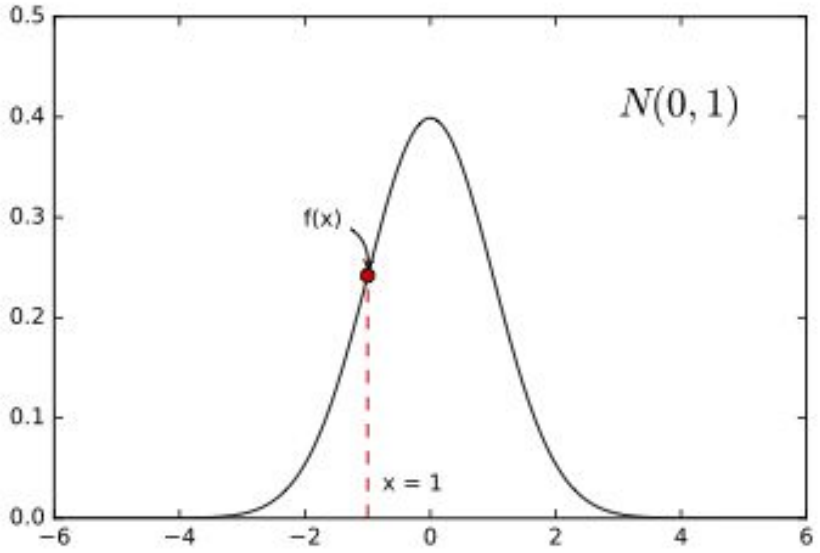
Distribution of the sum of a large number of independent, identically distributed variables is approximately normal regardless of the underlying distribution

See in class simulation of a multigenic trait

Mathematical Description of the Normal PDF

Mathematically, the probability density function (pdf) for the Normal distribution is defined as:

$$f(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



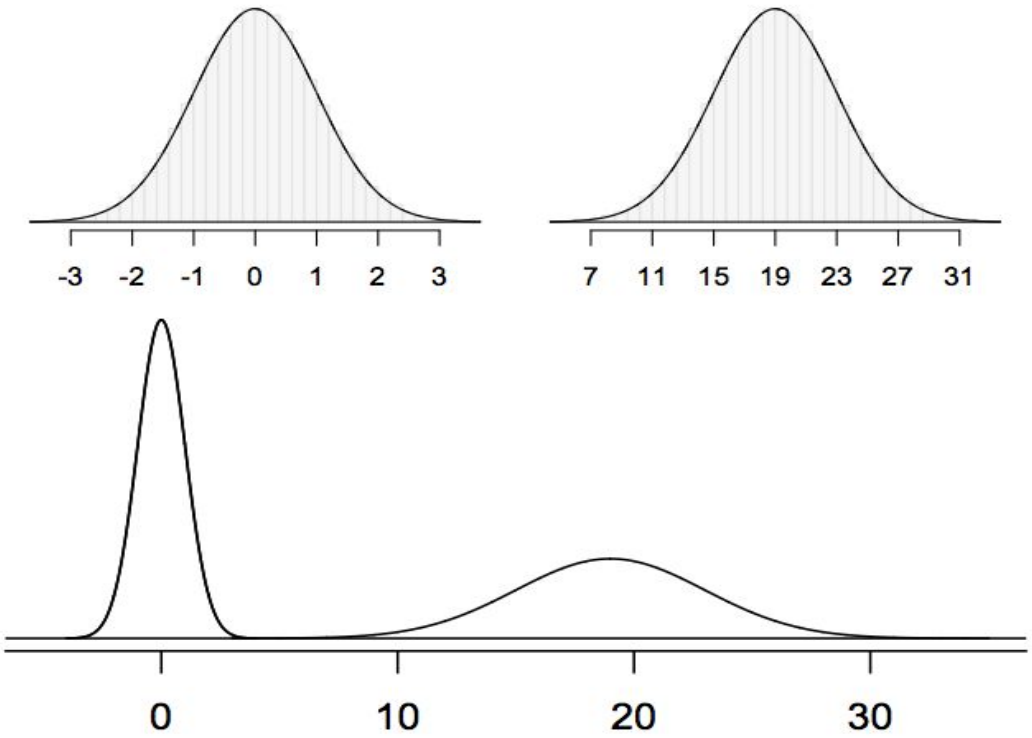
Normal density function in R

The function `dnorm` gives can be used to calculate the density function for a normal distribution of interest.

Normal distributions with different parameters

μ: mean, σ: standard deviation

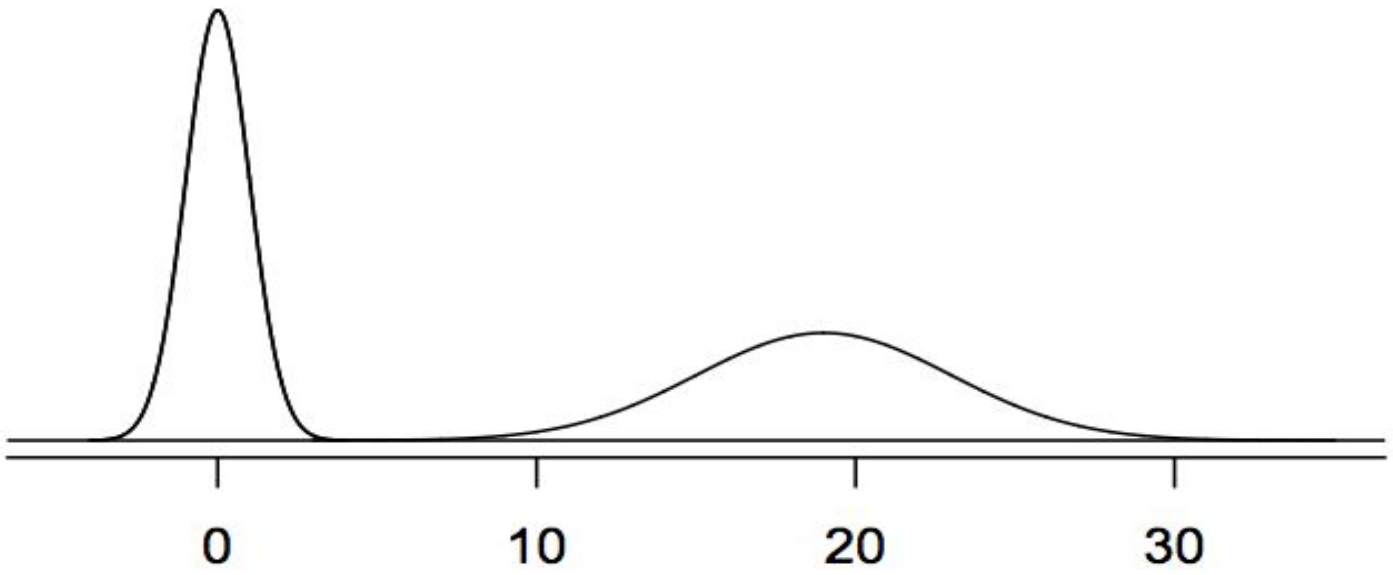
$N(\mu = 0, \sigma = 1)$ $N(\mu = 19, \sigma = 4)$



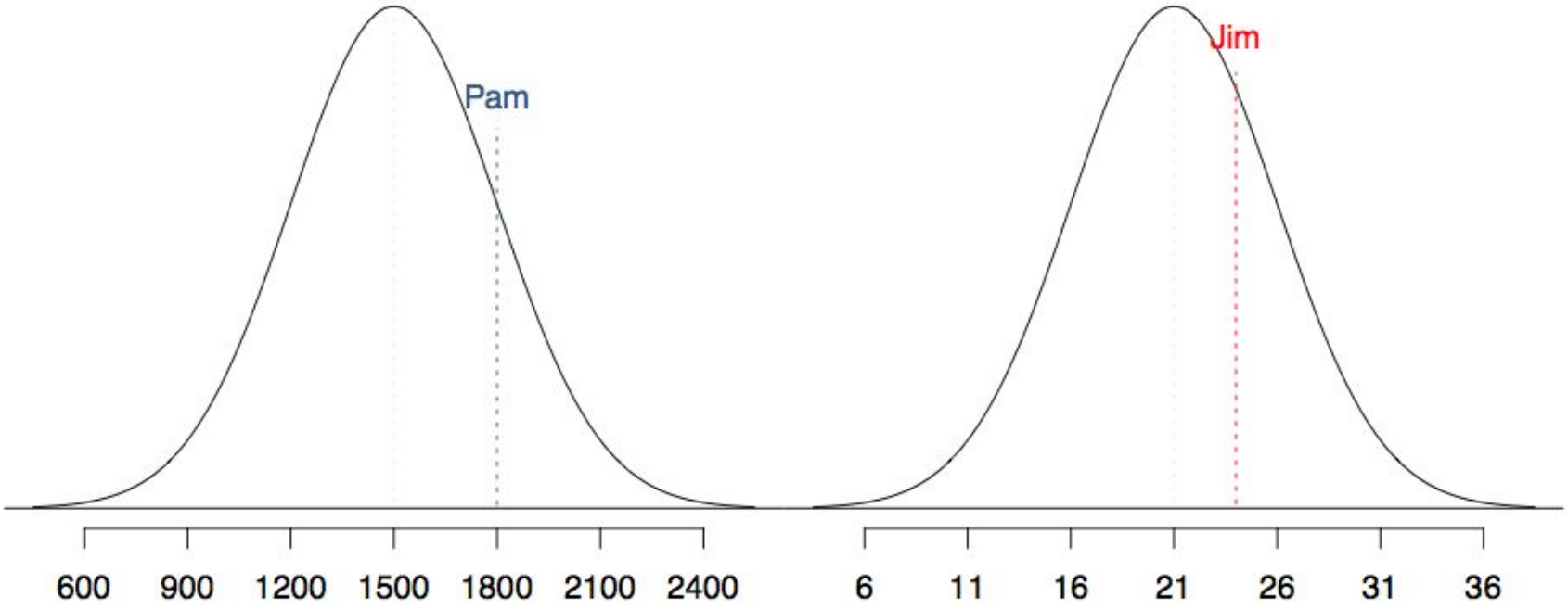
Draw these two normal distributions in R using ggplot functions

μ: mean, σ: standard deviation

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



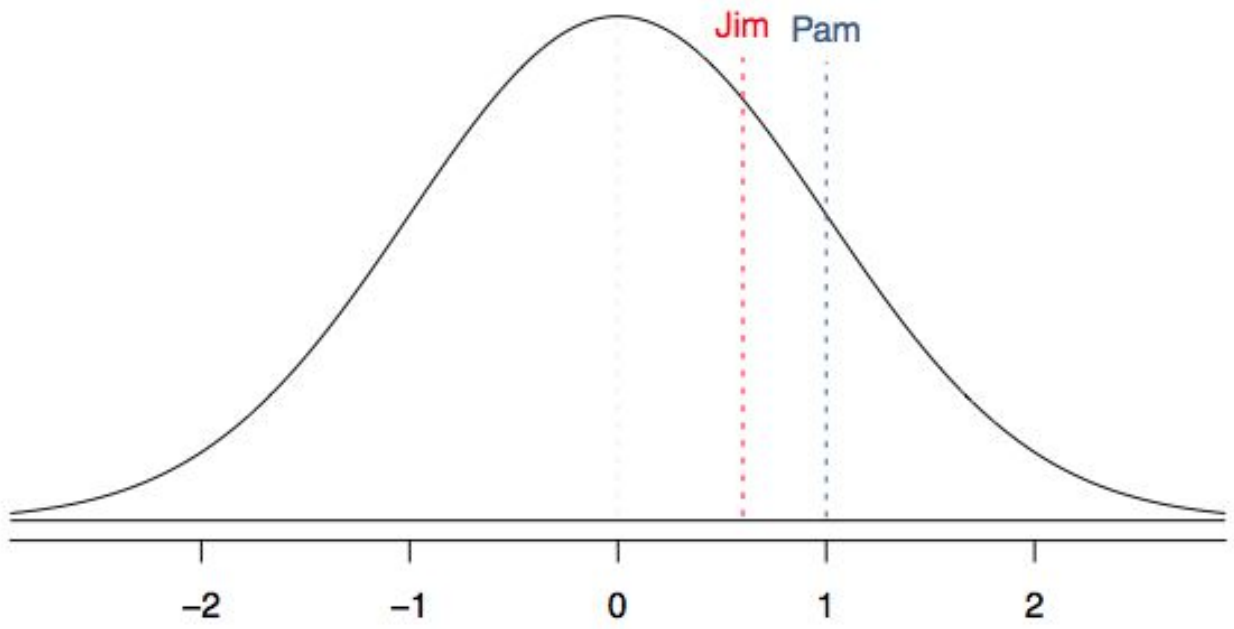
SAT scores are distributed nearly normally with mean 1500 and standard deviation 300. ACT scores are distributed nearly normally with mean 21 and standard deviation 5. A college admissions officer wants to determine which of the two applicants scored better on their standardized test with respect to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

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

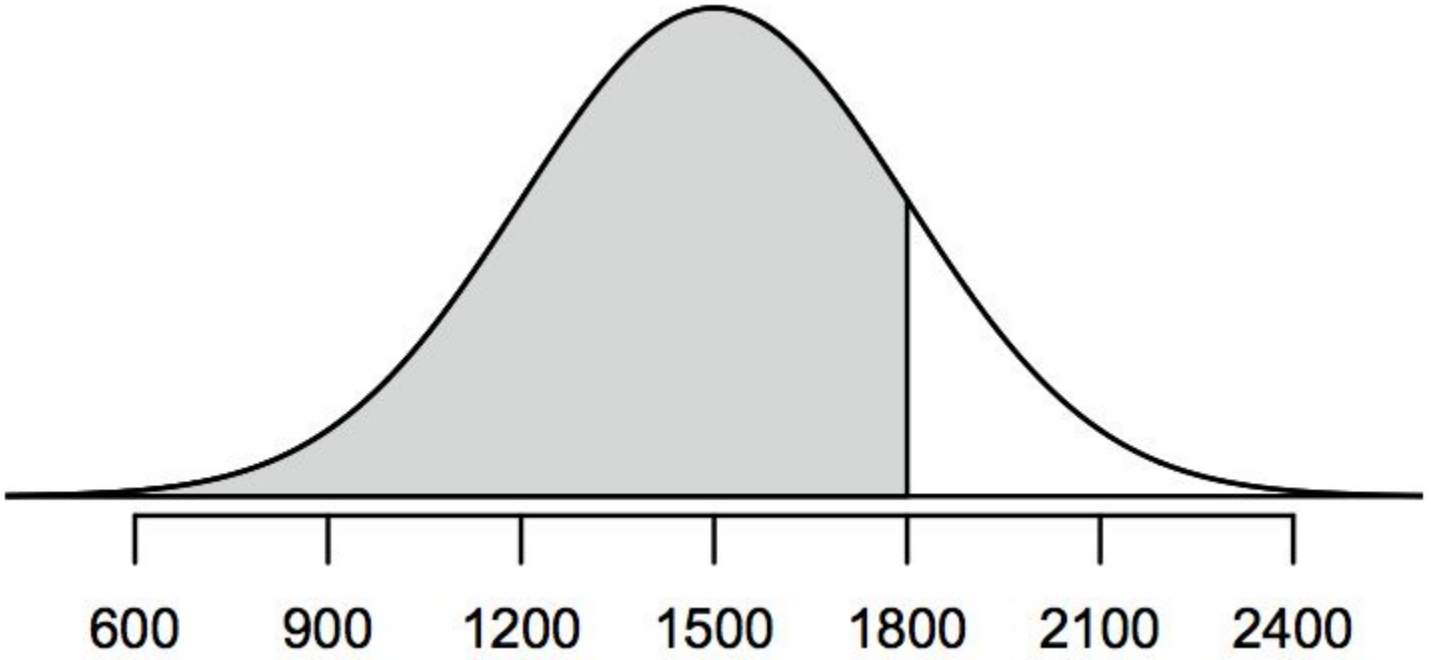
- Z scores are defined for distributions of any shape, but only when the distribution is normal can we use Z scores to calculate percentiles.
- Observations that are more than 2 SD away from the mean ($|Z| > 2$) are usually considered unusual.

Z scores in R

The function `scale` can be used to standardize a data set to have a mean of 0 and a standard deviation of 1. The standardized observations are therefore Z scores.

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 in R

There `pnorm` function can be used to compute percentiles/areas under the curve in R.

By setting the argument `lower.tail = FALSE`, pnorm can be used to calculate the area to the right of a given point (i.e. the upper tail of the distribution)

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

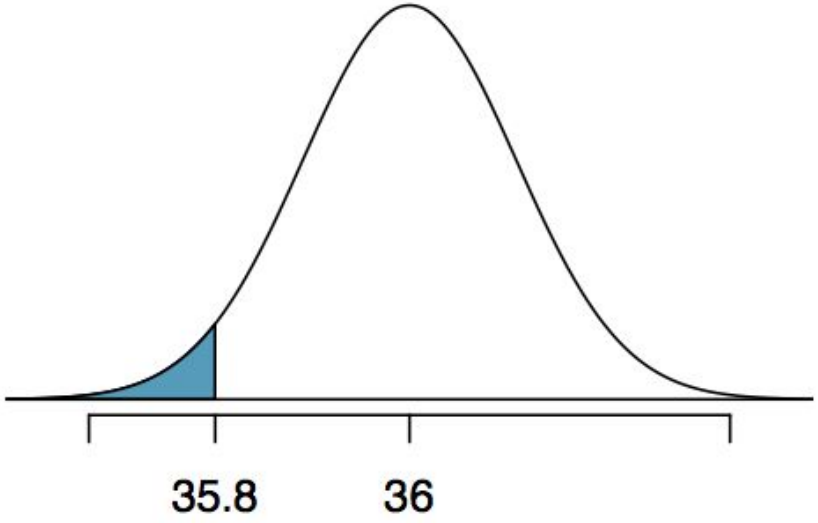
At Heinz ketchup factory the amounts which go into bottles of ketchup are supposed to be normally distributed with mean 36 oz. and standard deviation 0.11 oz. Once every 30 minutes a bottle is selected from the production line, and its contents are noted precisely. If the amount of ketchup in the bottle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

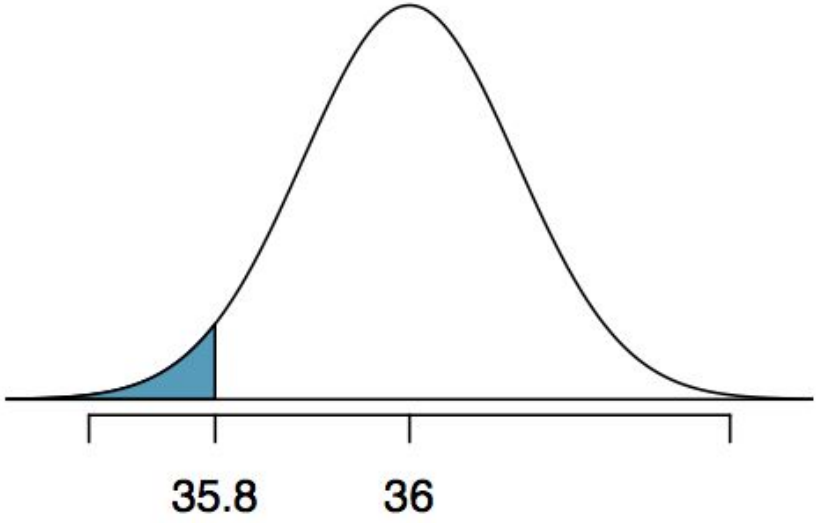
- *Let X = amount of ketchup in a bottle: X ~ N(μ = 36, σ = 0.11)*



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 ~ N(μ = 36, σ = 0.11)*



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

Practice

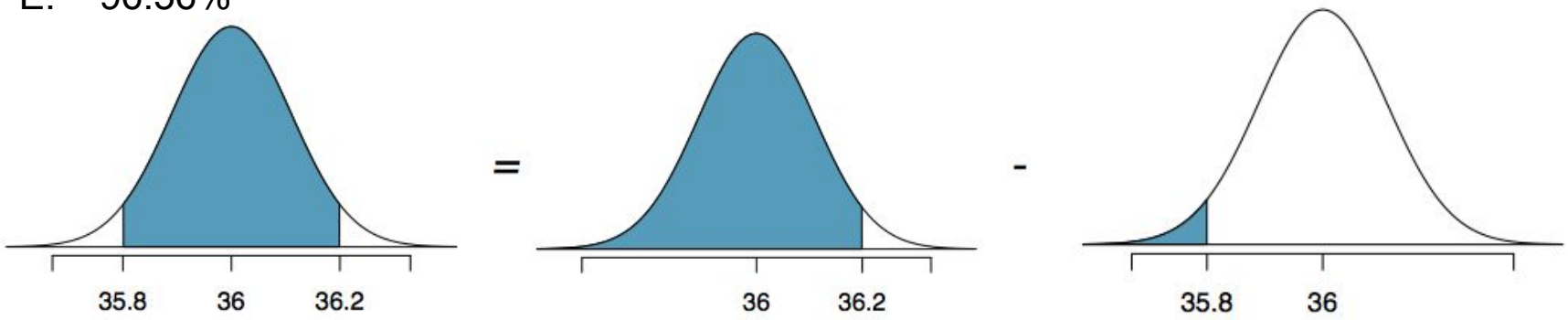
What percent of bottles pass the quality control inspection?

A. 1.82%
B. 3.44%
C. 6.88%
D. 93.12%
E. 96.56%

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%

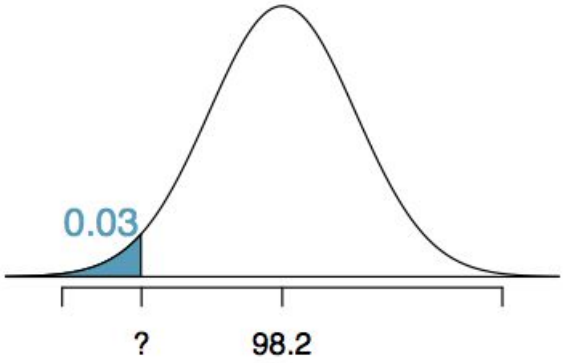


Finding cutoff points

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

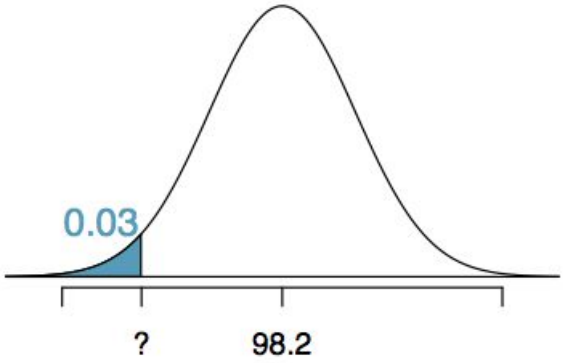
Finding cutoff points

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°F and standard deviation 0.73°F. What is the cutoff for the lowest 3% of human body temperatures?



The R function `qnorm` can be used to calculate a cutoff point corresponding to the given area under the curve.

Practice

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

A. 97.3ºF

B. 99.1ºF

C. 99.4ºF

D. 99.6ºF

Practice

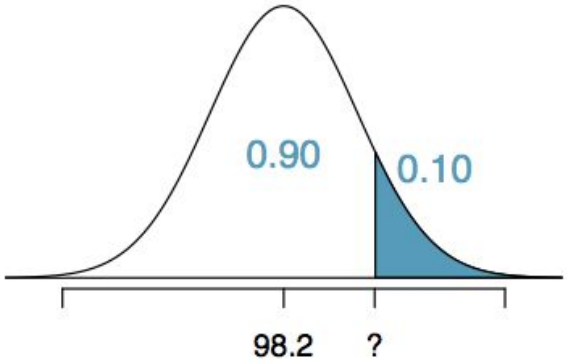
Body temperatures of healthy humans are distributed nearly normally with mean 98.2ºF and standard deviation 0.73ºF. What is the cutoff for the highest 10% of human body temperatures?

A. 97.3ºF

B. 99.1ºF

C. 99.4ºF

D. 99.6ºF

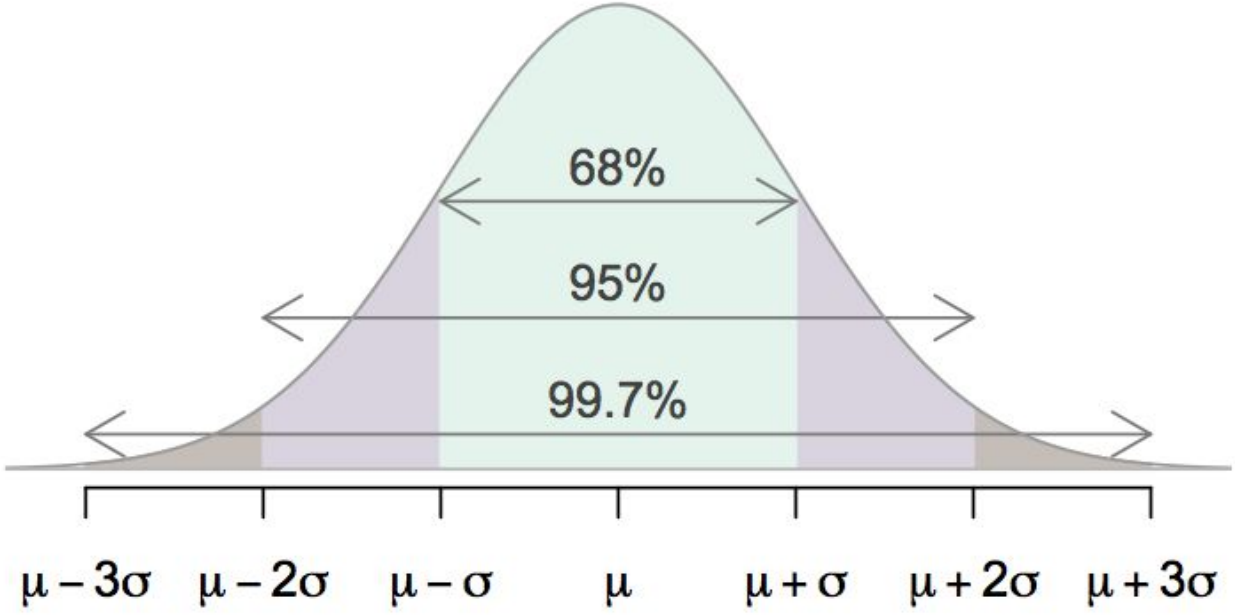


68-95-99.7 Rule

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.



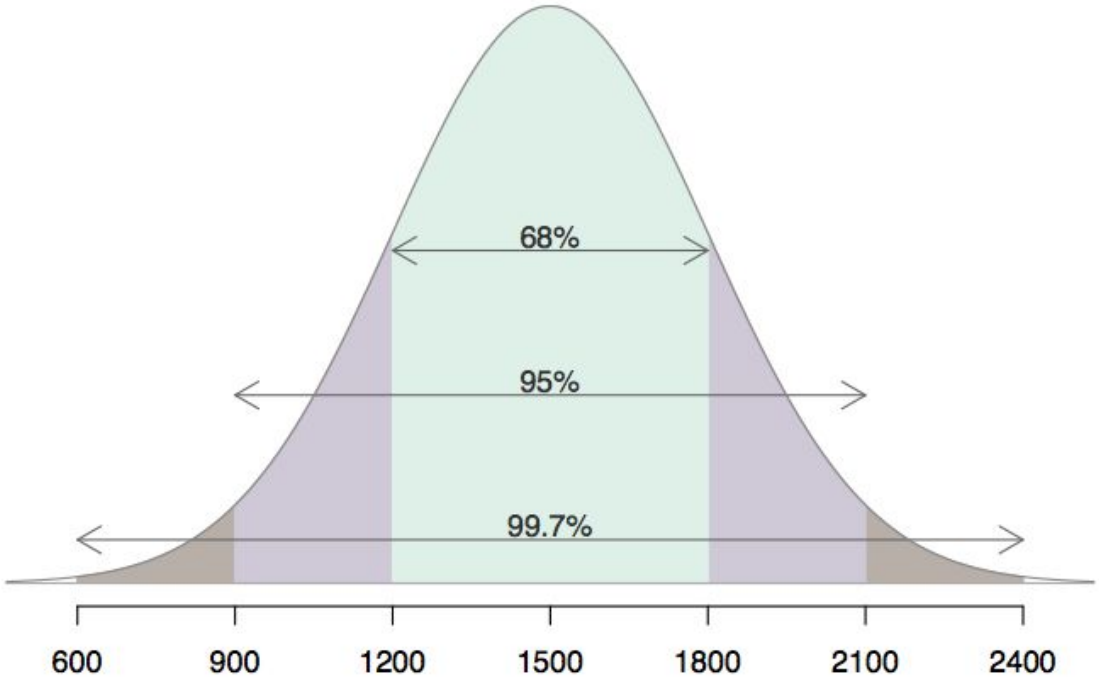
Describing variability using the 68-95-99.7 Rule

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

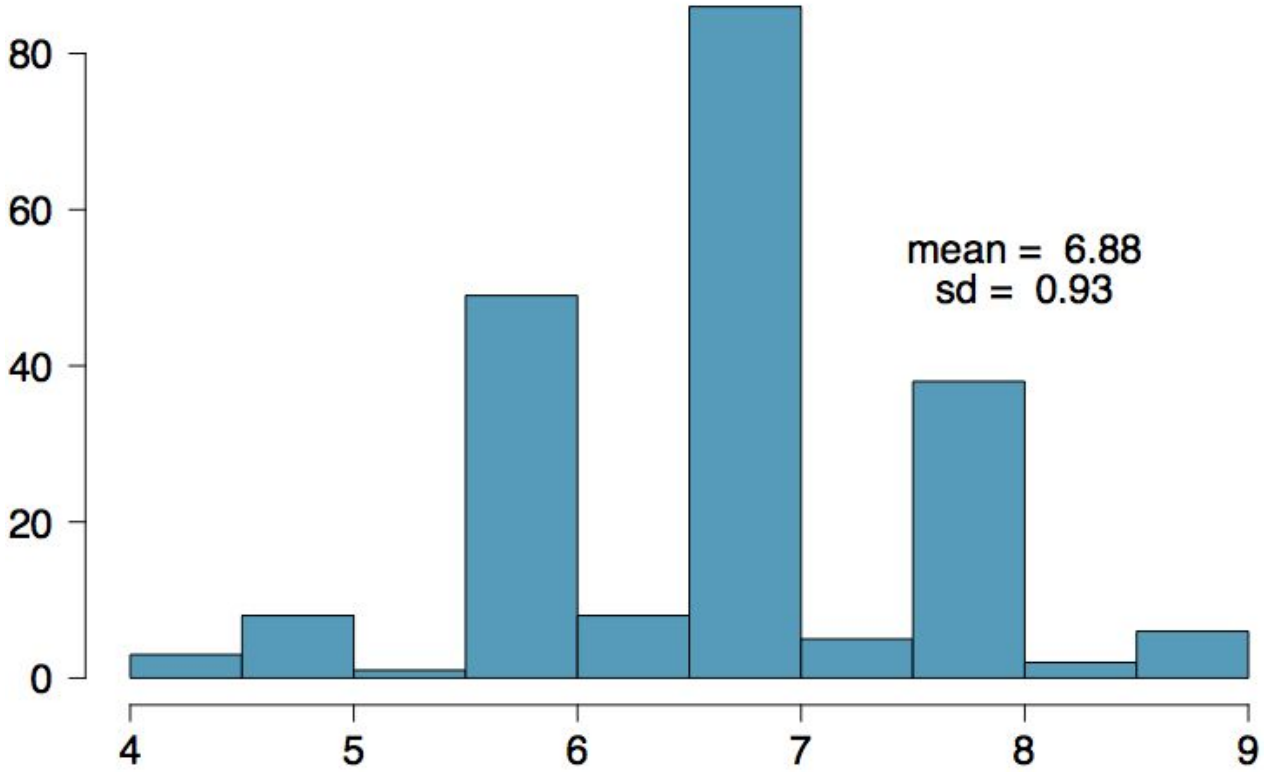
Describing variability using the 68-95-99.7 Rule

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.

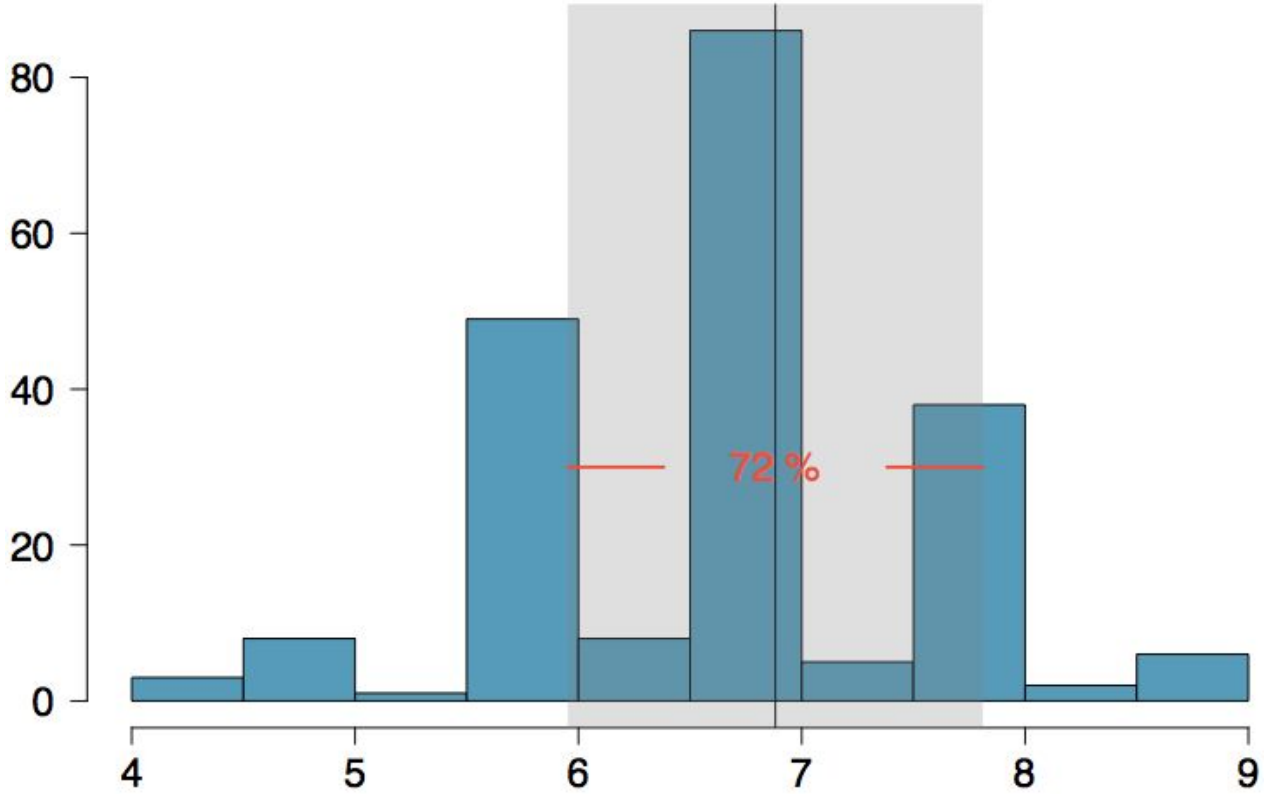


Number of hours of sleep on school nights



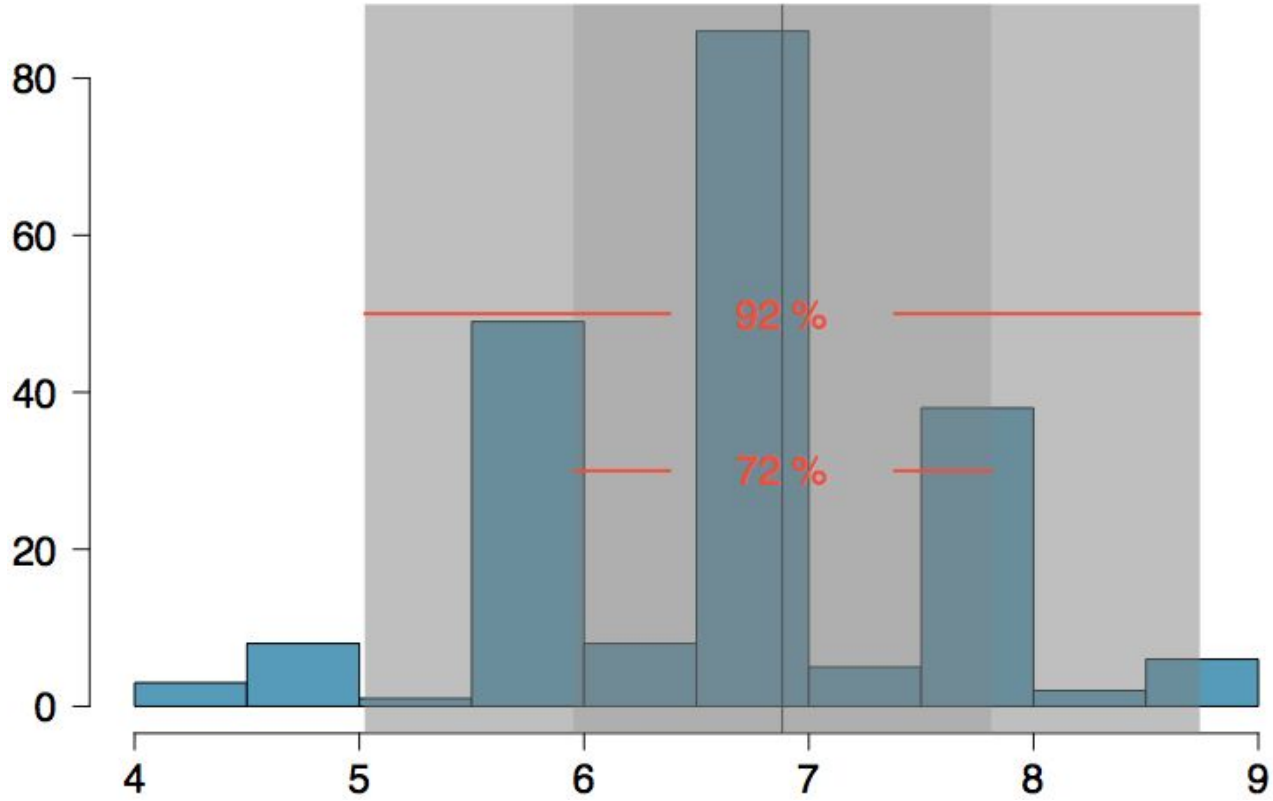
- Mean = 6.88 hours, SD = 0.92 hrs

Number of hours of sleep on school nights



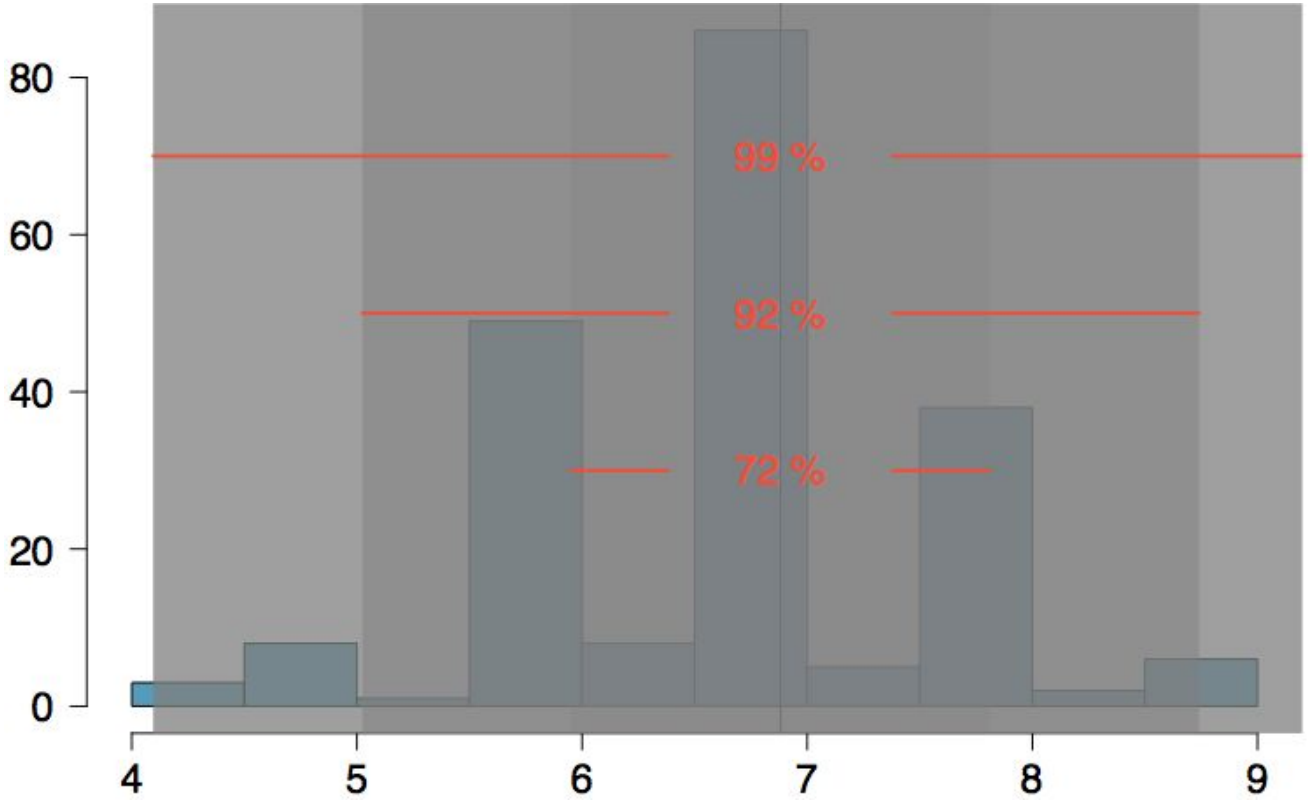
- Mean = 6.88 hours, SD = 0.92 hrs
- 72% of the data are within 1 SD of the mean: 6.88 ± 0.93

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

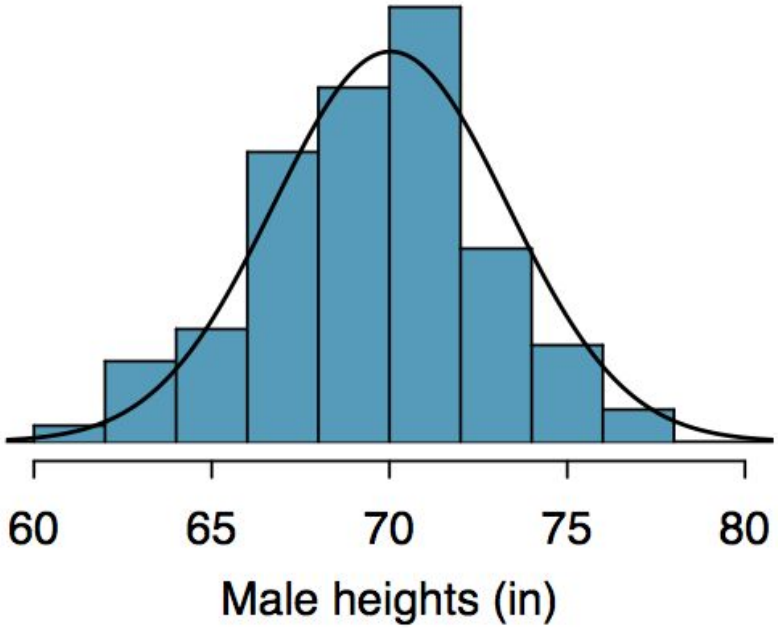


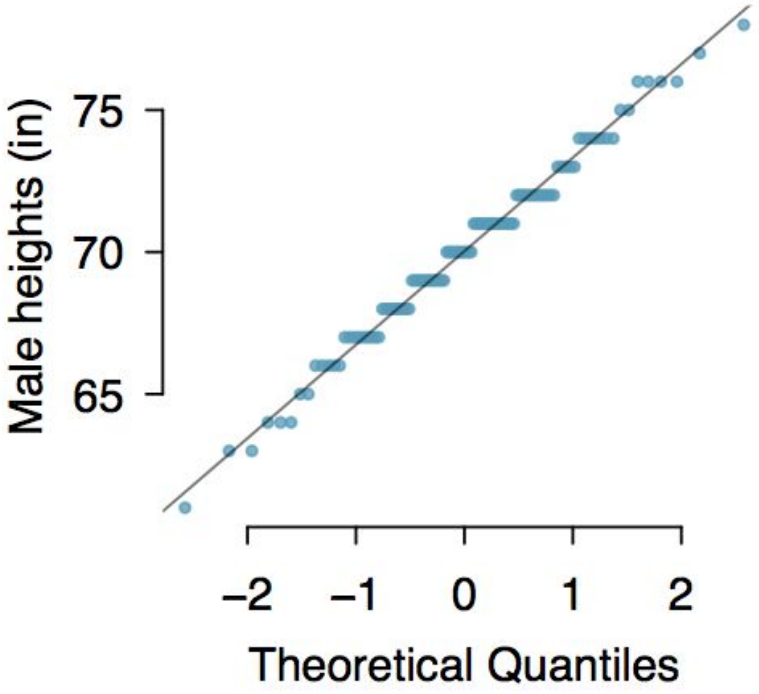
- Mean = 6.88 hours, SD = 0.92 hrs
- 72% of the data are within 1 SD of the mean: 6.88 ± 0.93
- 92% of the data are within 1 SD of the mean: 6.88 ± 2 x 0.93
- 99% of the data are within 1 SD of the mean: 6.88 ± 3 x 0.93

Evaluating the normal approximation

Normal probability plot

A histogram and *normal probability plot* of a sample of 100 male heights.



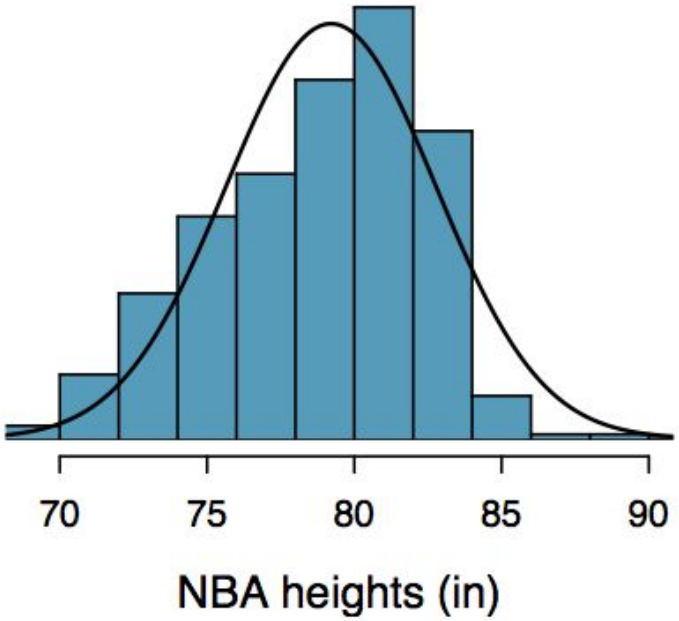


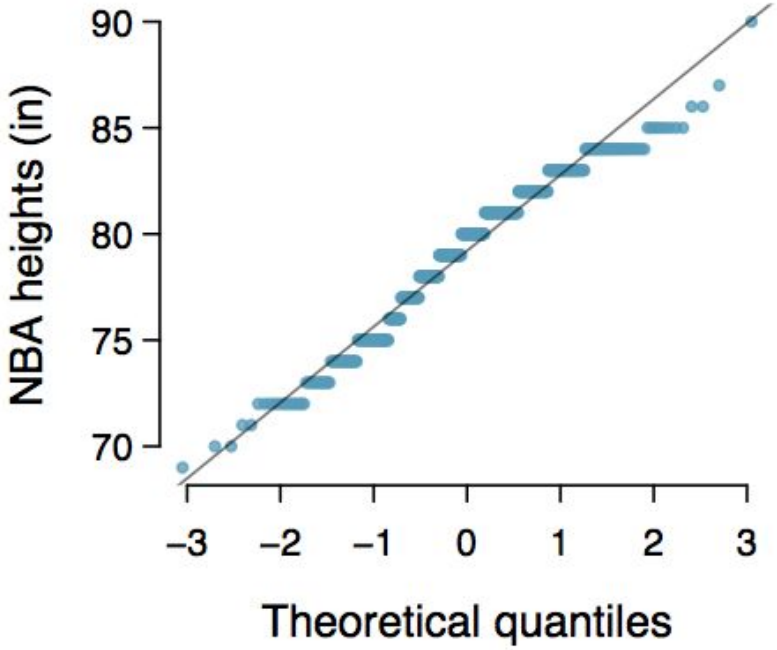
Anatomy of a normal probability plot

- Data are plotted on the y-axis of a normal probability plot, and theoretical quantiles (following a normal distribution) on the x-axis.
- If there is a linear relationship in the plot, then the data follow a nearly normal distribution.
- Constructing a normal probability plot requires calculating percentiles and corresponding z-scores for each observation, which is tedious. Therefore we generally rely on software when making these plots.

Practice

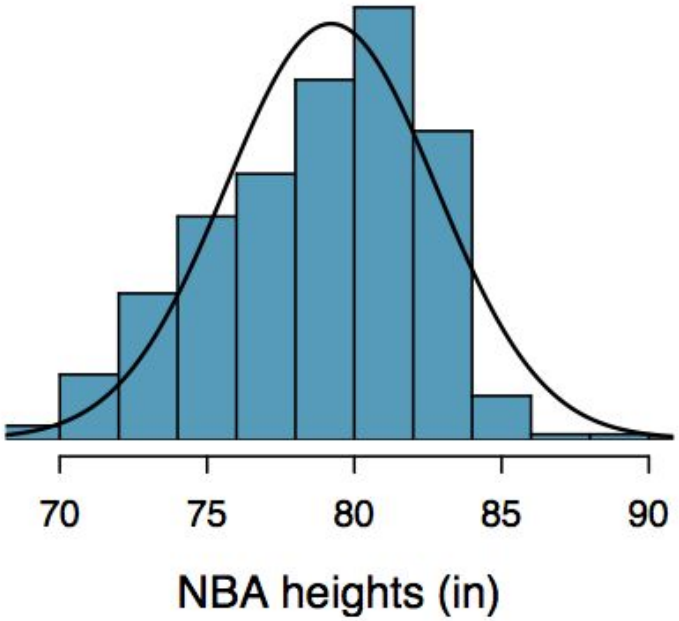
Below is a histogram and normal probability plot for the NBA heights from the 2008-2009 season. Do these data appear to follow a normal distribution?

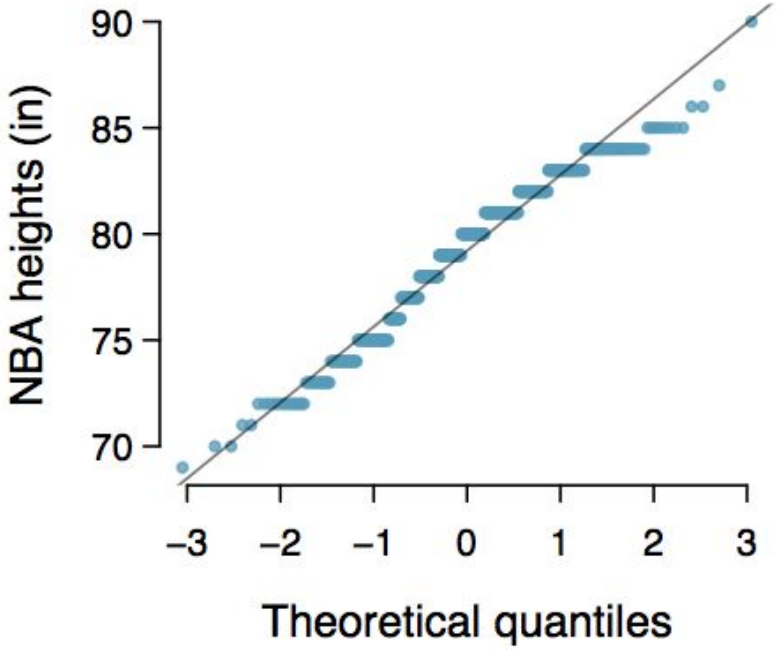




Practice

Below is a histogram and normal probability plot for the NBA heights from the 2008-2009 season. Do these data appear to follow a normal distribution?





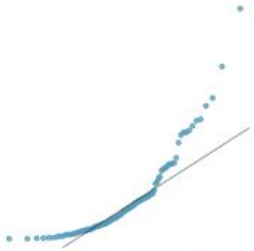
Why do the points on the normal probability have jumps?

Creating a normal probability plot in R

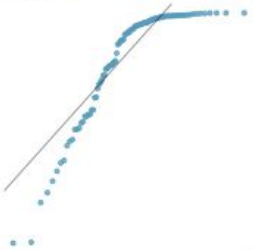
There are several ways to create normal probability plots in R.

- `qqnorm` and `qqline`
- Using ggplot
  - `geom_qq`, explicitly calculate expected slope and intercept of line and use `geom_abline`

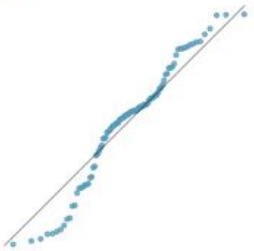
Normal probability plot and skewness



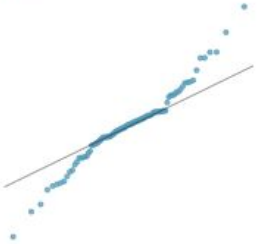
Right skew - Points bend up and to the left of the line.



Left skew - Points bend down and to the right of the line.



Short tails (narrower than the normal distribution) - Points follow an S shaped-curve.



Long tails (wider than the normal distribution) - Points start below the line, bend to follow it, and end above it.