IMS 13 (ish): Normal Distribution

Packages Needed To Recreate Code on Slides

```
library(openintro)
```

Recall: FTS

Fundamental Theorem of Statistics (overview)

Suppose we have an independent random sample taken from a given population. Then estimated mean/proportion will converge to the true mean/proportion as the sample size (n) increases

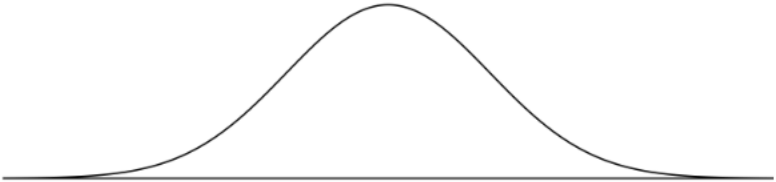
$$\overline{x} \to \mu$$

$$\hat{p} \to p$$

Furthermore, the sampling distribution of $\overline{x}/\hat{p}$ convervges to a **normal distribution** as n increases.

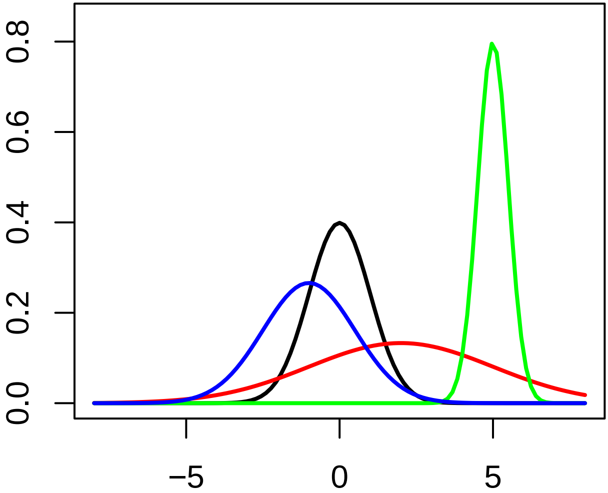
Normal Distribution

The normal distribution is a symmetric, unimodal, bell-shaped continuous probability distribution.



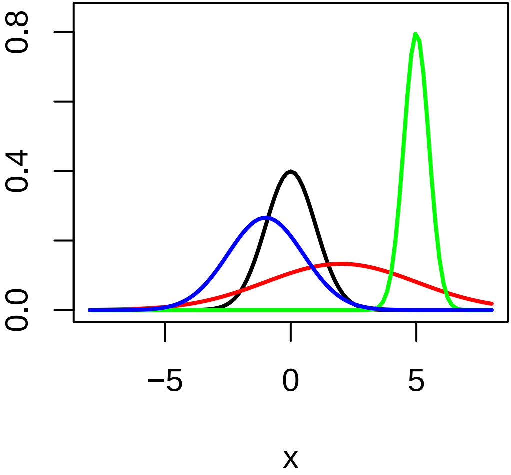
Normal Distribution

The normal distribution is characterized by its mean μ (expressing the center) and standard deviation σ (expressing the variability). All four of these curves are normal distributions!



Data that is Normal Distribution

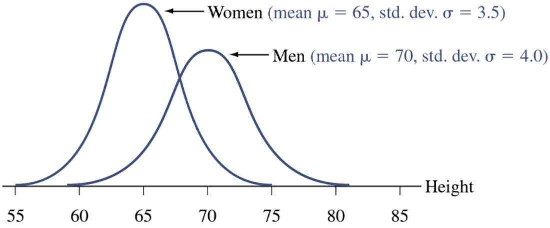
If a random variable has mean μ and standard deviation σ we may write the distribution as $N(\mu, \sigma)$



- $N(0, 1)$
- $N(2, 3)$
- $N(5, .5)$
- $N(-1.5, 1.5)$

Normal Distribution in the "Real World"

If we let X = height of a randomly selected adult female in America, and Y = height of a randomly selected adult male in America. We would see distributions that look like the following:



$$X \sim N(65, 3.5) \qquad Y \sim N(70, 4)$$

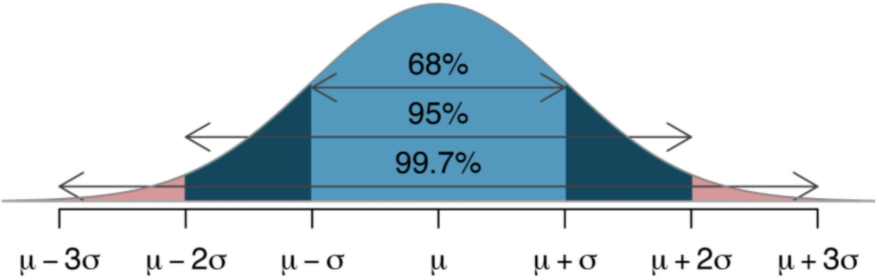
Which distribution has larger variability?

Normal Distribution and the FTS

- Individual random variables (like height) can follow a normal distribution.
- Typically, we are usually working with means ($\overline{x}$) or proportions ($\hat{p}$) when we think of the normal distribution.
- Regardless, the principles are the same!

Normal Distribution and Probability

For a normal distribution, the probability of falling within 1, 2, or 3 standard deviations of the mean equals 0.68, 0.95, and 0.997, respectively. This is true for any value of μ and any value of $\sigma > 0$.



Normal Distribution and Probability

The heights of adult women can be approximated by a normal distribution, μ =65.0 inches and σ =3.5 inches

- 68% are between 61.5 and 68.5 inches.
  - $[\mu \pm \sigma] = [65 \pm 3.5]$
- 95% are between 58.0 and 72.0 inches. -
  - $[\mu \pm 2\sigma] = [65 \pm (2)3.5]$
- 99.7% are between 54.5 and 75.5 inches.
  - $[\mu \pm 3\sigma] = [65 \pm (3)3.5]$

Normal Distribution and Probability

SAT scores closely follow the normal model with mean $\mu = 1500$ and standard deviation $\sigma = 300$.

1. Let X be an SAT score. Write down the short-hand for a normal distribution for X.
2. About what percent of test takers score 900 to 2100?
3. About what percent score between 1500 and 2100?

Z-Scores

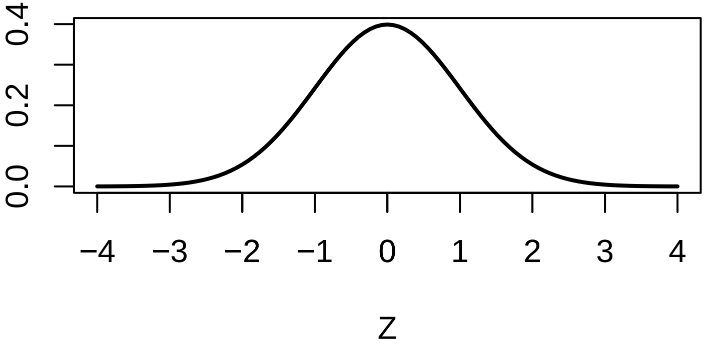
Z-Score

The **Z score** of an observation is the number of standard deviations it falls above or below the mean. We compute the Z score for an observation x that follows a distribution with mean μ and standard deviation σ using:

$$Z = \frac{x - \mu}{\sigma}$$

Z-Scores

- Z-scores help us compare observations from different distributions to each other.
- Z-scores transform the a random variable $X \sim N(\mu, \sigma)$ into a **standard normal distribution** where $Z = \frac{X-\mu}{\sigma} \sim N(0, 1)$.



Z-Scores

Suppose

- X = height of adult women in America
- Y = height of adult men in America

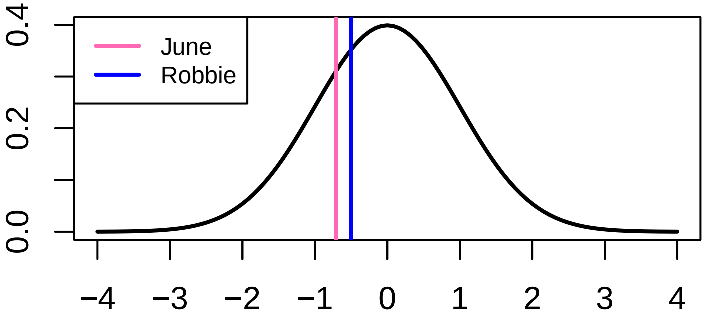
$$X \sim N(65, 3.5) \qquad Y \sim N(70, 4)$$

Further suppose that Robbie (male) is 68 inches tall and June (female) is 62.5 inches tall.

- Calculate their Z-scores.
- Who is shorter for their respective group?

Z-Scores

$$Z_{June} \approx -.71 \qquad Z_{Robbie} = -0.5$$



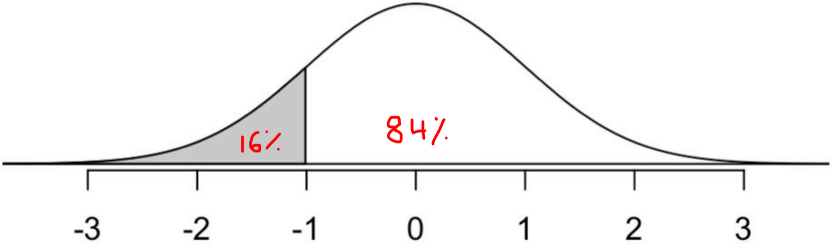
Probabilities and Percentiles

- We are frequently interested in finding the probability for a certain range of values.
- Calculating probabilities of these ranges is difficult and involves calculus...*messy calculus*.
- Instead we can use probability tables, and software to calculate percentiles.
- Recall: A k-th percentile, is a value at which a k percent of values falls below.

Probabilities and Percentiles

The 16^{th} percentile if the standard normal distribution -1.

That is $P(Z < -1) = 0.16$.



What is the 97.5% percentile of a standard normal distribution?

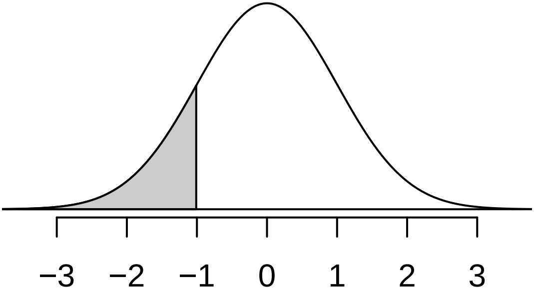
Probabilities and Percentiles

Tables and software typically return probabilities related percentiles (areas to the left of a point)

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

```
[1] 0.1586553
```

```
normTail(m = 0, s = 1, L = -1)
```



Probabilities and Percentiles

- To obtain the probability of observing something less than a particular value we can use the `pnorm()` function.
- In the previous slide:
  - For a random variable $X \sim N(0, 1)$, the probability of observing a random variable below -1 is approximately 0.16.
  - The `normTail()` graphed the distribution and highlighted the region of interest.

Probabilities and Percentiles

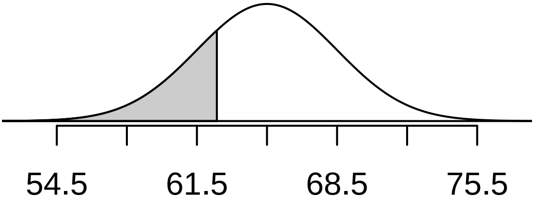
Find the probability of observing a height less than Junes. Draw the region below June's height (62.5) using the original distribution, recall $X \sim N(65, 3.5)$.

The following code finds $P(X < 62.5) \approx 0.24$

```
pnorm(62.5, mean = 65, sd = 3.5)
```

```
[1] 0.2375253
```

```
normTail(m = 65, s = 3.5, L = 62.5)
```



Probabilities and Percentiles

Find the probability of observing a height larger than Robbies. Draw the region above Robbie's height (68) using the original distribution, recall $Y \sim N(70, 4)$.

$$
\begin{aligned}
P(Y > 68) =& 1 - P(Y < 68) \\
\approx& 1 - .31 \\
=& .69
\end{aligned}
$$

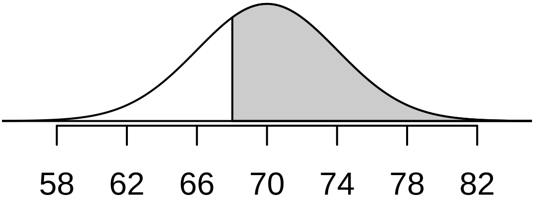
Probabilities and Percentiles

Find the probability of observing a height larger than Robbies. Draw the region above Robbie's height (68) using the original distribution, recall $Y \sim N(70, 4)$.

```
1 - pnorm(68, mean = 70, sd = 4)
```

```
[1] 0.6914625
```

```
normTail(m = 70, s = 4, U = 68)
```



Probabilities and Percentiles

Find the probability of observing a z that is within one standard deviation of the mean of a standard normal distribution. Draw the region within one standard deviation of the mean of a standard normal distribution.

$$\begin{aligned} P(|Z| < 1) =& P(Z < 1) - P(Z < -1) \\ \approx& .84 - .16 \\ =& .68 \end{aligned}$$

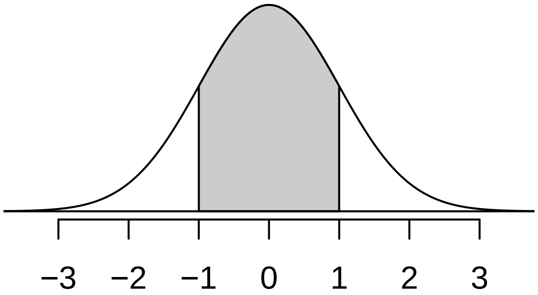
Probabilities and Percentiles

Find the probability of observing a z that is within one standard deviation of the mean of a standard normal distribution. Draw the region within one standard deviation of the mean of a standard normal distribution.

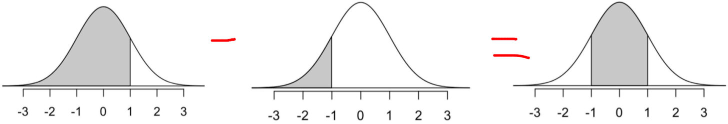
```
pnorm(1, mean = 0, sd = 1) - pnorm(-1, mean = 0, sd = 1)
```

```
[1] 0.6826895
```

```
normTail(m = 0, s = 1, M = c(-1, 1))
```



Probabilities and Percentiles



Probabilities and Percentiles

Find the probability of observing a z that is outside of one standard deviation of the mean of a standard normal distribution. Draw the region outside of one standard deviation of the mean of a standard normal distribution.

$$
\begin{aligned}
P(|Z| > 1) =& P(Z > 1) + P(Z < -1) \\
\approx& P(Z < -1) + P(Z < -1) \qquad \text{(By symmetry)} \\
=& .32
\end{aligned}
$$

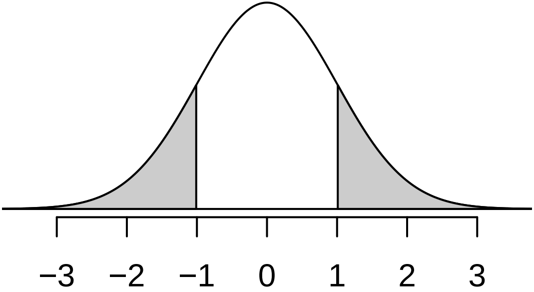
Probabilities and Percentiles

Find the probability of observing a z that is outside of one standard deviation of the mean of a standard normal distribution. Draw the region outside of one standard deviation of the mean of a standard normal distribution.

```
2*pnorm(-1, mean = 0, sd = 1)
```

```
[1] 0.3173105
```

```
normTail(m = 0, s = 1, L = -1, U = 1)
```



Probabilities and Percentiles

- It is STRONGLY recommended to always draw the picture when finding probabilities and percentiles.
- Using R (and tables) is much preferable to calculus.
- We often have to use "tricks", i.e. take advantage of symmetry, to find probabilities.

Finding Percentiles

Recall the height for adult males in America is $Y \sim N(70, 4)$. Suppose you learn that Jake (male) is equal to the 40^{th} percentile for height. What is Jake's height?

In other words, we are finding y_{Jake} such that,

$$.40 = P(Y < y_{Jake})$$

Finding Percentiles

We can answer this question using R as well! For normally distributed random variables the `qnorm()` function returns the corresponding percentile for a probability.

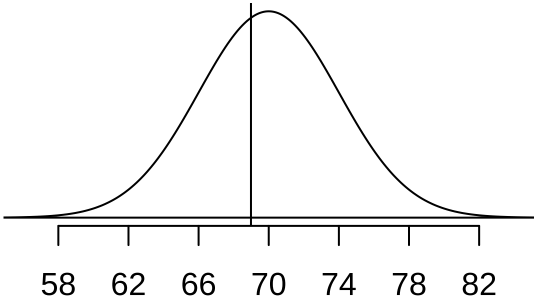
```
jake <- qnorm(.4, mean = 70, sd = 4)
jake
```

```
[1] 68.98661
```

Finding Percentiles

Graphing Jake's height:

```
normTail(m = 70, s = 4)
abline(v = jake)
```



Practice Questions

[OI 4.2] Area under the curve, Part II. What percent of a standard normal distribution $N(\mu = 0, \sigma = 1)$ is found in each region? Draw the graph.

a) Z > -1.13

b) Z < 0.18

c) Z > 8

d) $|$Z$| < 0.5$

[OI 4.4] Triathlon Times. In triathlons, it is common for racers to be placed into age and gender groups. Suppose two friends competed in the above below. Leo completed the race in 4948 seconds, while Mary completed the race in 5513 seconds. Here is some information on the performance of two groups for a particular race:

- Men, Ages 30 - 34 group: mean of 4313 seconds with a standard deviation of 583 seconds.
- Women, Ages 25 - 29 group: mean of 5261 seconds with a standard deviation of 807 seconds.
- The distributions of finishing times for both groups are approximately Normal.

a) Write down the short-hand for these two normal distributions.

b) What are the Z-scores for Leo's and Mary's finishing times? What do these Z-scores tell you?

c) Did Leo or Mary rank better in their respective groups? Explain your reasoning.

Practice Questions

[OI 4.4] Triathlon Times. In triathlons, it is common for racers to be placed into age and gender groups. Suppose two friends competed in the above below. Leo completed the race in 4948 seconds, while Mary completed the race in 5513 seconds. Here is some information on the performance of two groups for a particular race:

- Men, Ages 30 - 34: mean of 4313 seconds with a standard deviation of 583 seconds.
- Women, Ages 25 - 29: mean of 5261 seconds with a standard deviation of 807 seconds.
- The distributions of finishing times for both groups are approximately Normal.

d) What percent of the triathletes did Leo finish faster than in his group?

e) What percent of the triathletes did Mary finish faster than in her group?

f) If the distributions of finishing times are not nearly normal, would your answers to parts (b) - (e) change?