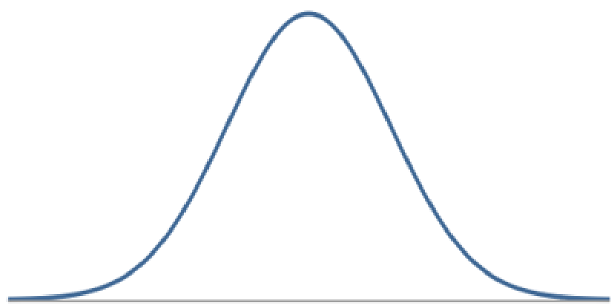
Central Limit Theorem and the Normal Distribution

Characteristics of the Normal Distribution

We run into the Normal distribution a lot in the real world and in statistics. What are some of the things that define a Normal distribution?



What is the shape of a Normal distribution?[1]

a) Skewed
b) Symmetric
c) Bimodal
d) Unimodal

What is the best numerical summary to use for the center of a Normal Distribution?[2]

a) Mean
b) Median
c) Standard deviation
d) IQR

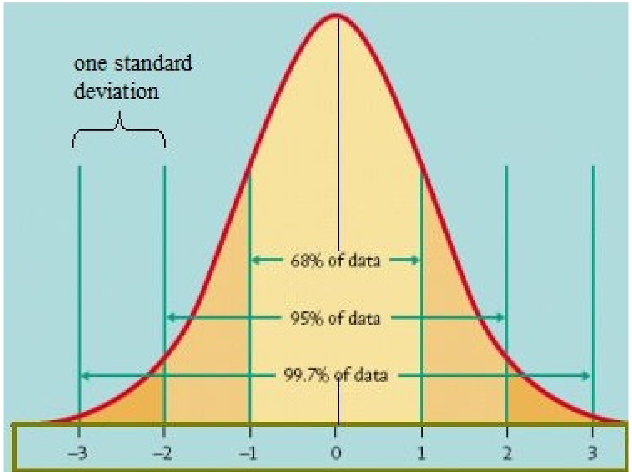
What is the best numerical summary to use for the spread of a Normal Distribution?[3]

a) Mean
b) Median
c) Standard deviation
d) IQR

Empirical Rule

The **Empirical Rule**, or **68-95-99.7 Rule** applies to any Normal distribution. This rule lets us understand what are the *most common* values observations from the population take. It is helpful for making quick estimates of what values observations in the population tend to take. useful to know what values occur most frequently and which ones are rare.

[1] b,d
[2] a
[3] c



Approximate the following using the Emirical Rule and Z ~ N(0, 1) What percentage of observations will be between $Z = -1$ and $Z = 1$? In shorthand: $P(-1 < Z < 1) =$ [4]

a) 100%

b) 99.7%

c) 95%

d) 68%

$P(Z > -2) =$[5]

a) 95%

b) 50%

c) 5%

d) 97.5%

$P(-2 < Z < 1) =$[6]

a) 81.5%

b) 68%

c) 95%

d) 78.5%

$P(-z^* < Z < z^*) = 0.95$[7]

a) $z^* = 1$

b) $z^* = 2$

c) $z^* = 3$

d) Impossible to know

---

[4] d
[5] d
[6] a
[7] b

Approximate the following using the Emirical Rule and X ~ N(100, 10). Hint: Draw a picture of the curve.
$P(X > 80) =$[8]

a) 97.5%

b) 95%

c) 75%

d) 2.5%

$P(90 < X < 120) =$[9]

a) 68%

b) 95%

c) 81.5%

d) 78.5%

$P(a < X < b) = 0.95$[10]

a) $a = 80$ and $b = 120$

b) $a = 90$ and $b = 110$

c) $a = 95$ and $b = 95$

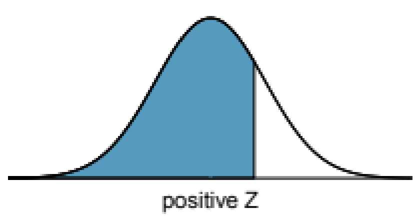
d) $a = -2$ and $b = 2$

Beyond the Empirical

Old way - if you're interested. . .

In most cases, we would like to use a Normal Distribution with more than "whole" standard deviations. It is difficult to find the proportion below -1.5 or above 2.33 standard deviations using the Empirical Rule.

In the past, we used a standard Normal table. But with the technology we have now, this method has gone the way of the dodo.



	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

[8] a
[9] c
[10] a

New way

Now we can use `R` to calculate and visualize area under the Normal curve. The function `xpnorm` in the `mosaic` package provides a nice way to calculate and visualize these areas. To use `xpnorm`, you need some basic argument inputs:

```
xpnorm(q, mean, sd, lower.tail)
```

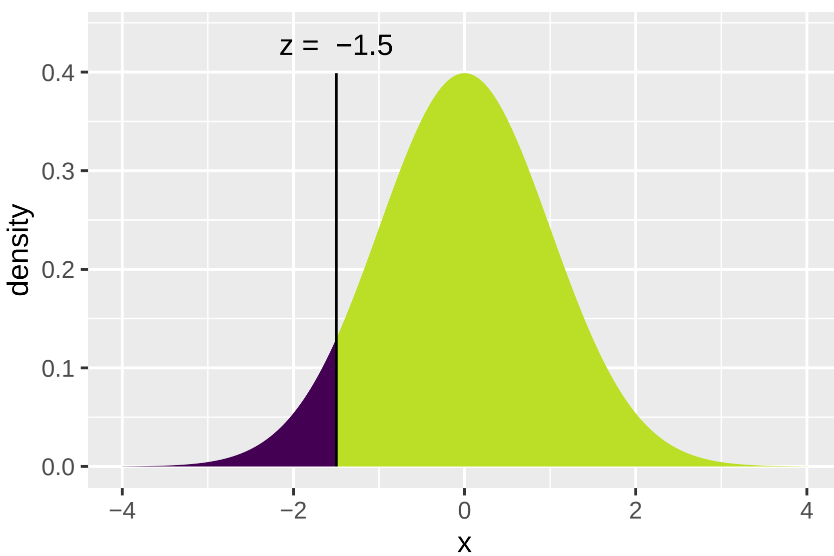
Argument	Purpose	Default
`q`	the quantile (`q`) value(s) on the x-axis we would like to use as a vertical cutoff. In other words: the value(s) on the Normal curve we are interested in	No default, must specify in code
`mean`	the mean value for the Normal distribution we are interested in	`mean = 0`
`sd`	the standard deviation for the Normal distribution we are interested in	`sd = 1`
`lower.tail`	specify the direction to calculate the proportion (the purple tail)	`lower.tail = TRUE`

On a Standard Normal Distribution ($Z \sim N(0, 1)$), to find:

$$P(Z < -1.5) =$$

```r
xpnorm(q = -1.5, mean = 0, sd = 1, lower.tail = TRUE)
```

```
##
## If X ~ N(0, 1), then
##  P(X <= -1.5) = P(Z <= -1.5) = 0.06681
##  P(X >  -1.5) = P(Z >  -1.5) = 0.9332
##
```




```
## [1] 0.0668072
```

We can see that $\approx 6.7\%$ of observations are more than 1.5 standard deviations below the mean on a Standard Normal curve.

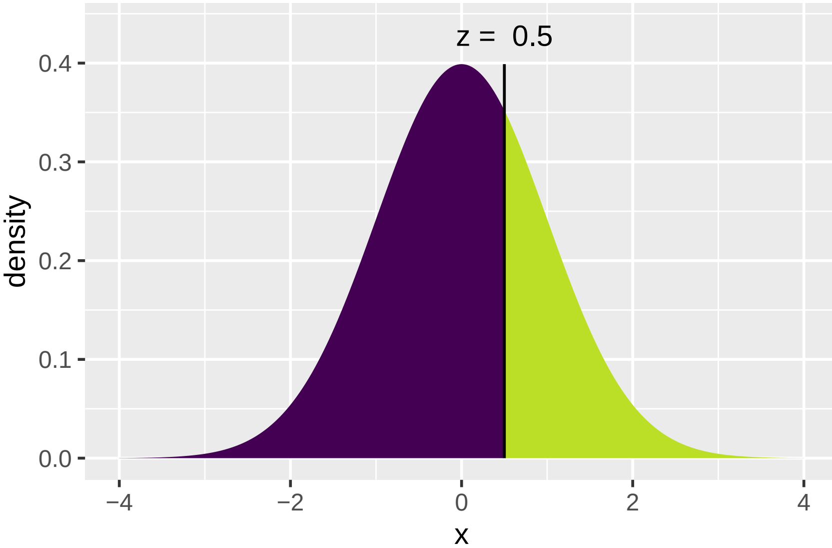
Your turn

Find the following probabilities in R from $Z \sim N(0, 1)$:

$P(Z < 0.5) =$

```r
xpnorm(q = 0.5, mean = 0, sd = 1)
```

```
##
## If X ~ N(0, 1), then
##  P(X <= 0.5) = P(Z <= 0.5) = 0.6915
##  P(X >  0.5) = P(Z >  0.5) = 0.3085
##
```

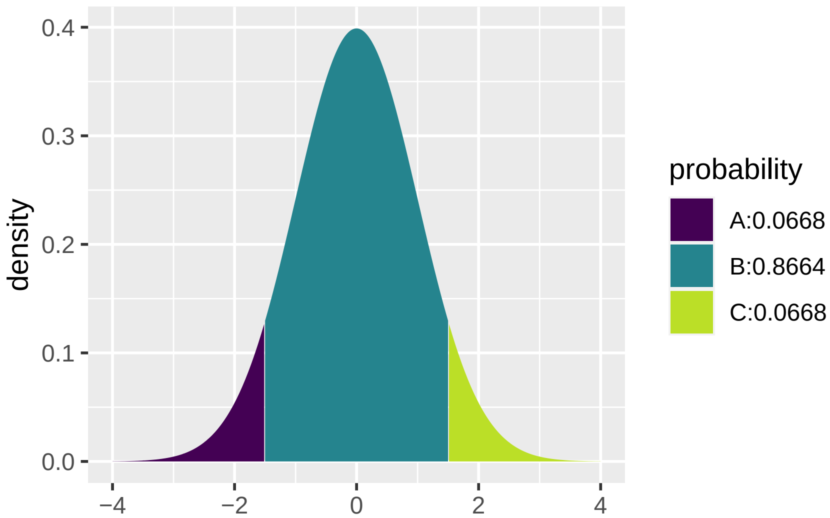



```
## [1] 0.6914625
```

$P(-1.5 < Z < 0.5) =$

```r
xpnorm(q = c(-1.5, 1.5), mean = 0, sd = 1)
```

```
##
## If X ~ N(0, 1), then
##  P(X <= -1.5) = P(Z <= -1.5) = 0.06681   P(X <=  1.5) = P(Z <=  1.5) = 0.93319
##  P(X >  -1.5) = P(Z >  -1.5) = 0.93319   P(X >   1.5) = P(Z >   1.5) = 0.06681
##
```



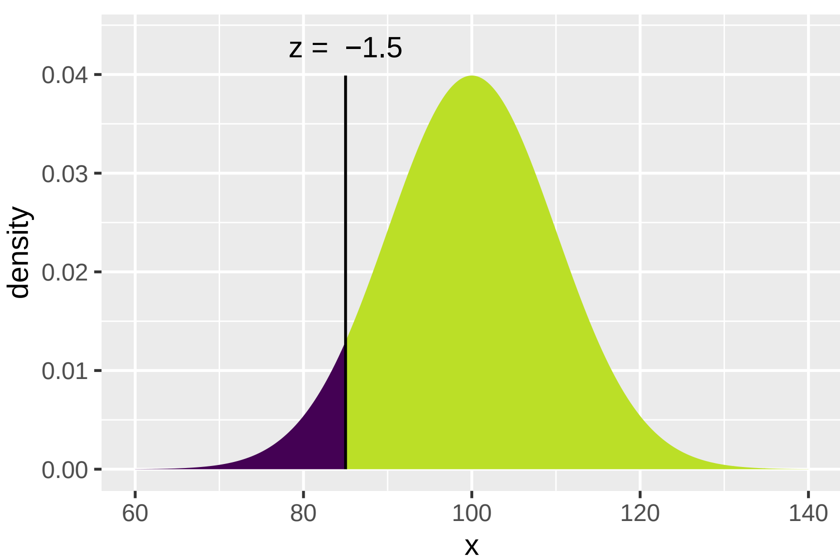

```
## [1] 0.0668072 0.9331928
```

Find the following probabilities from $X \sim N(100, 10)$:

$P(X > 85) =$

```r
xpnorm(q = 85, mean = 100, sd = 10, lower.tail = FALSE)
```

```
##
## If X ~ N(100, 10), then
##  P(X <= 85) = P(Z <= -1.5) = 0.06681
##  P(X >  85) = P(Z >  -1.5) = 0.9332
##
```



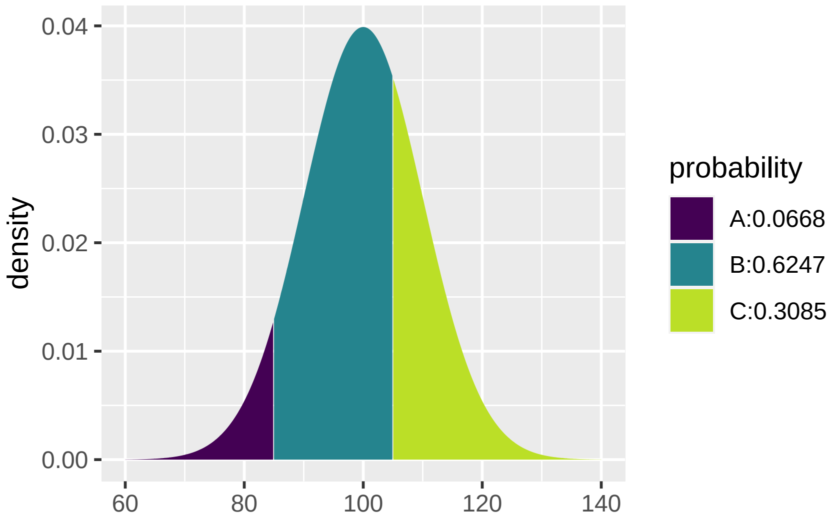

```
## [1] 0.9331928
```

$P(85 < X < 105) =$

```r
xpnorm(q = c(85, 105), mean = 100, sd = 10)
```

```
##
```

```
## If X ~ N(100, 10), then
##  P(X <=  85) = P(Z <= -1.5) = 0.06681    P(X <= 105) = P(Z <=  0.5) = 0.69146
##  P(X >   85) = P(Z >  -1.5) = 0.9332 P(X >  105) = P(Z >   0.5) = 0.3085
##
```




```
## [1] 0.0668072 0.6914625
```

These are all examples of finding the probability when we have a known quantile (reference point). We can also reverse the process.

Finding Quantiles/Percentiles

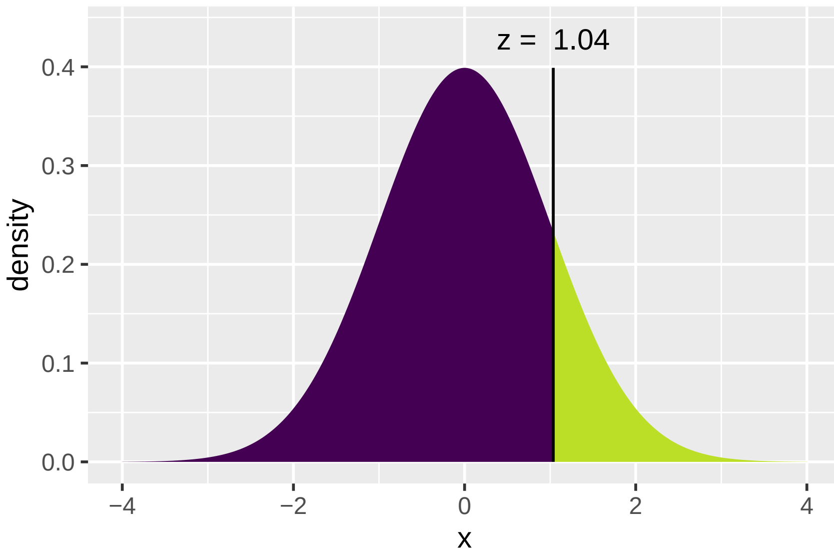
The `xqnorm` function

Explore the help documentation for `?xqnorm` or use your intuition to answer the following questions. (Advice: help documentation looks sort of weird in these tutorials, so you may want to switch to regular RStudio when using `?xqnorm`)

For $Z \sim N(0, 1)$, find the 85th percentile:

```
# In the previous examples we entered a quantile and
# got a proportion for the lower tail.
# Now try entering a proportion to get a
# quantile/percentile
xqnorm(p = 0.85, mean = 0, sd = 1, lower.tail = TRUE)
```

```
##
## If X ~ N(0, 1), then
##  P(X <= 1.036433) = 0.85
##  P(X >  1.036433) = 0.15
##
```

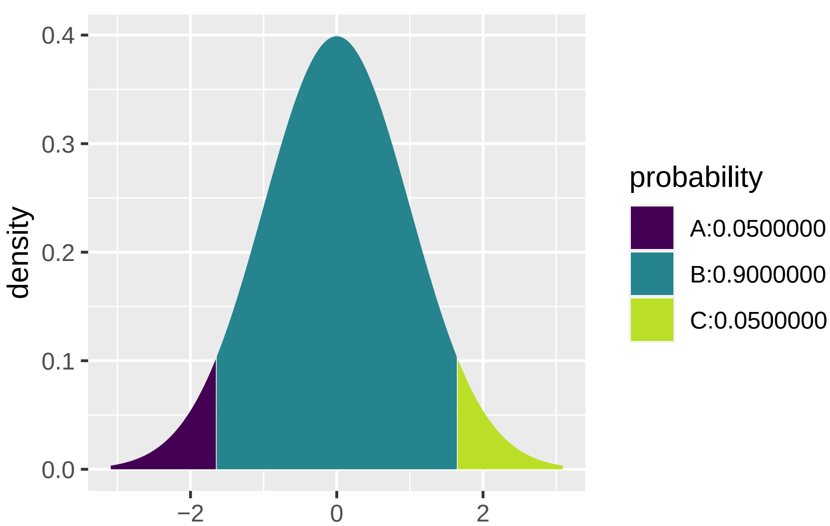



```
## [1] 1.036433
```

Find the cutoffs (quantiles) for the middle 90% ($P(-z^* < Z < z^*) = 0.90$) on a Standard Normal curve.

```
## Think about what two lower tail percentiles could
## be used to find the middle 90%
xqnorm(p = c(0.05, 0.95), mean = 0, sd = 1)
```

```
##
## If X ~ N(0, 1), then
##  P(X <= -1.644854) = 0.05    P(X <=  1.644854) = 0.95
##  P(X >  -1.644854) = 0.95    P(X >   1.644854) = 0.05
##
```



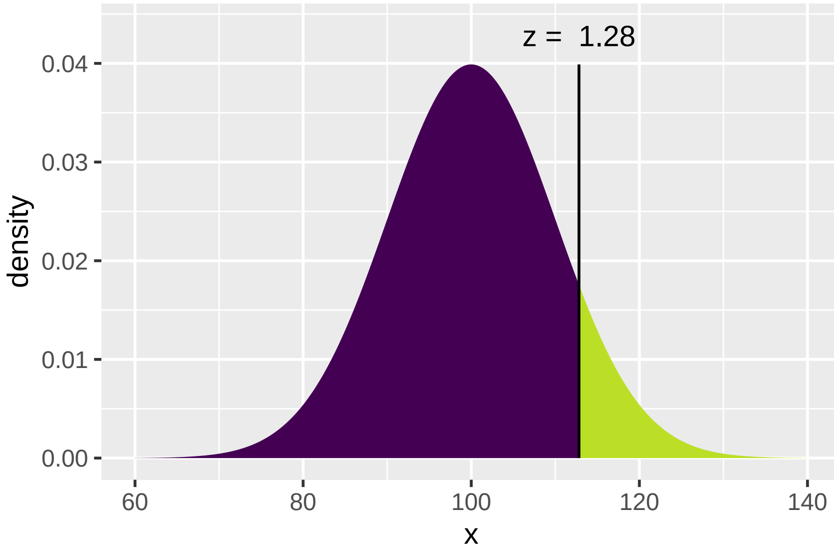

```
## [1] -1.644854  1.644854
```

Find the quantile for the *upper* 10% of $X \sim N(100, 10)$:

```
xqnorm(p = 0.10, mean = 100, sd = 10, lower.tail = FALSE)
```

```
##
```

```
## If X ~ N(100, 10), then
##  P(X <= 112.8155) = 0.9
##  P(X >  112.8155) = 0.1
##
```

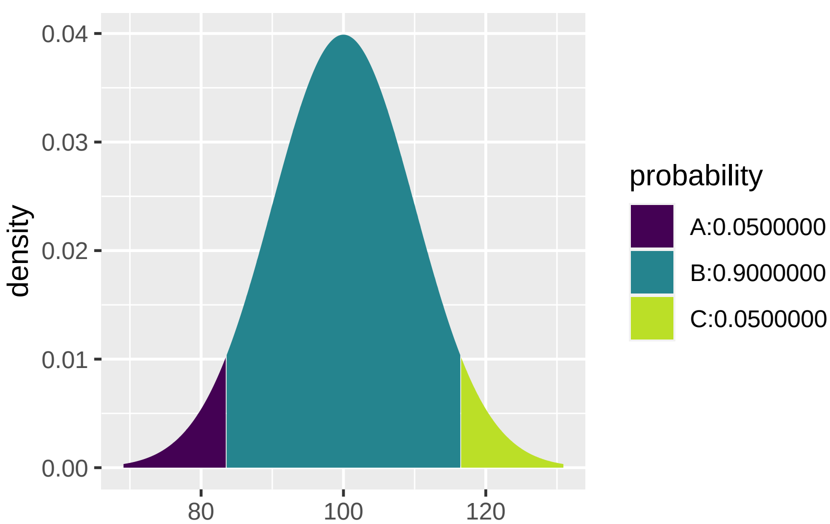



```
## [1] 112.8155
```

Find the cutoff values for the middle 90% of $X \sim N(100, 10)$:

```
xqnorm(p = c(0.05, 0.95), mean = 100, sd = 10)
```

```
##
## If X ~ N(100, 10), then
##  P(X <=  83.55146) = 0.05    P(X <= 116.44854) = 0.95
##  P(X >   83.55146) = 0.95    P(X >  116.44854) = 0.05
##
```




```
## [1]  83.55146 116.44854
```

Recap

Bringing it all together

So `xpnorm` gives us a proportion (`p`) under the Normal curve when we **already know** the quantile(s) (`q`).

And `xqnorm` gives us a quantile(s) (`q`) for a proportion (`p`) under the Normal curve we are **already interested in**.

Quantiles and area under the Normal curve allow us to answer questions about populations and samples. Recall that our simulated null distributions all looked very similar to the Normal curve. And remember that when we are calculating p-values to measure strength of sample evidence, we needed to know the proportion of simulated samples more extreme than our observed sample. This is a preview of what we use the normal distribution for…

Hopefully the Normal distribution is a little less scary now.