The Shape of Null Distributions

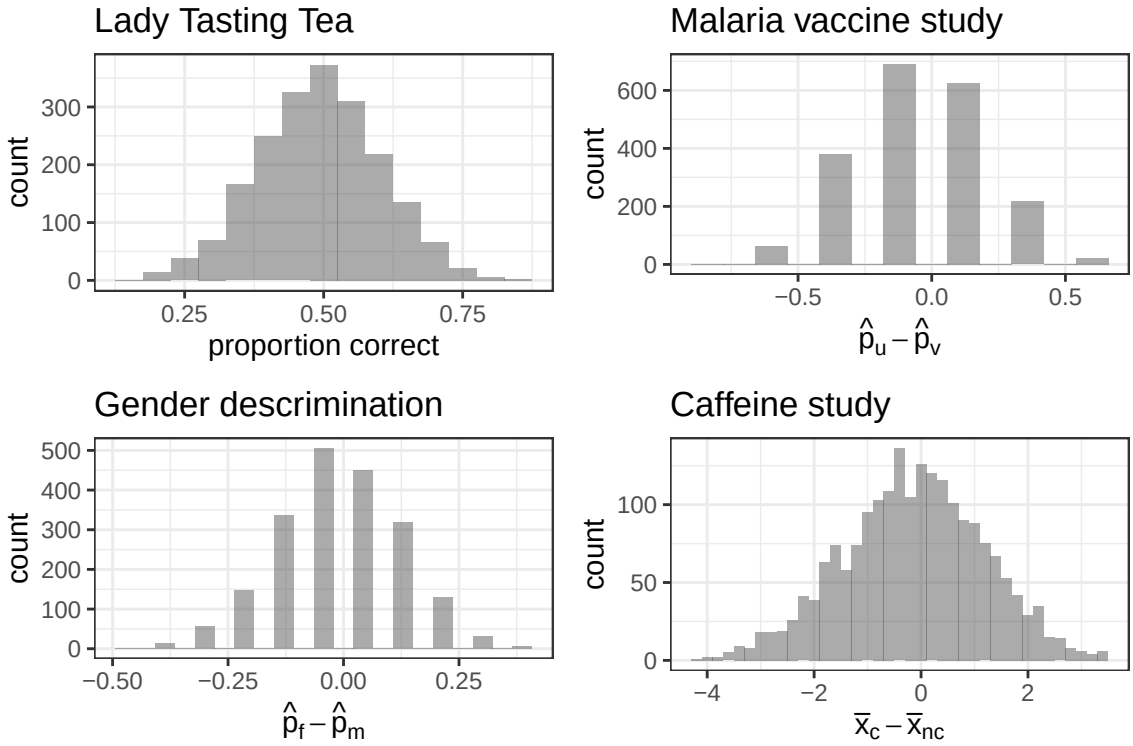
Stat 145

Examples we have seen

Here are four examples we have seen.

```
set.seed(123)
# Lady Tasting Tea
Tea_null <- do(2000) * rflip(20)
# Malaria vaccine study
Malaria_null <- do(2000) * diffprop(malaria ~ shuffle(group), data = Malaria)
# Gender discrimination in promotion decisions?
Promotion_null <- do(2000) * diffprop(decision ~ shuffle(gender), data = Promotion)
# Does caffeine increase finger tapping rate?
Taps_null <- do(2000) * diffmean(Taps ~ shuffle(Group), data = CaffeineTaps)
```

Let's look at the randomization distributions for each of these examples.



What do you notice about these distributions?

1. Where are they centered?
2. How would you describe their shape?

The shape of null distributions

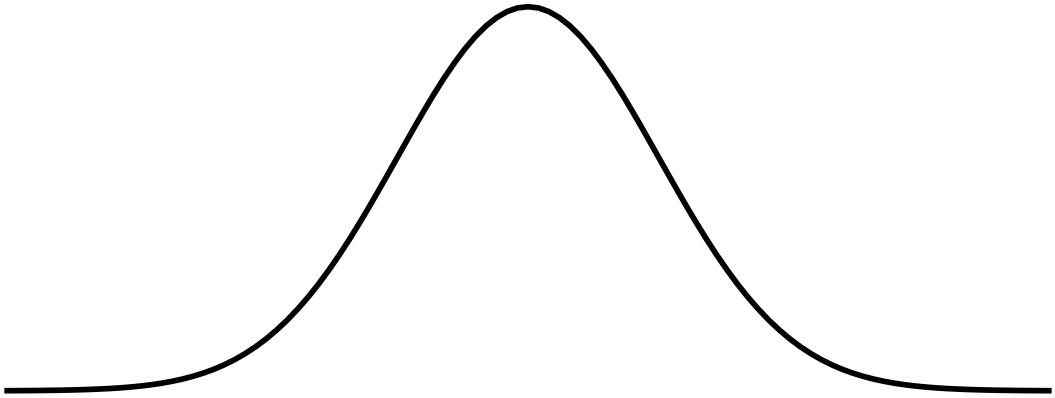
1. Null distributions will always be centered at the value the test statistic would have if the null hypothesis were true.
    - If this isn't true for your randomization distribution, that generally means that you have not randomized correctly.
2. Many (but not all) null distributions have a symmetric, bell-shaped distribution.
    - If we could know this shape without doing the randomization, we could get our p-values from this distribution without doing the randomization.
    - This approximation is better for larger sample sizes than for smaller sample sizes.

Many (again, not all) data distributions also have approximately this same shape as well. Since it shows up so often, we want to learn a bit about these bell-shaped distributions.

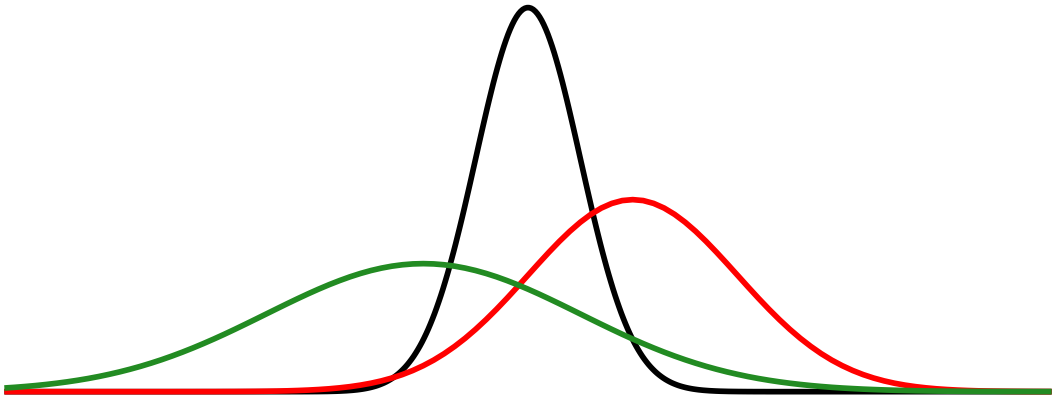
Normal Distributions

The Bell Curve

The bell-shaped curve that many distributions (approximately) have is called a **normal distribution**.

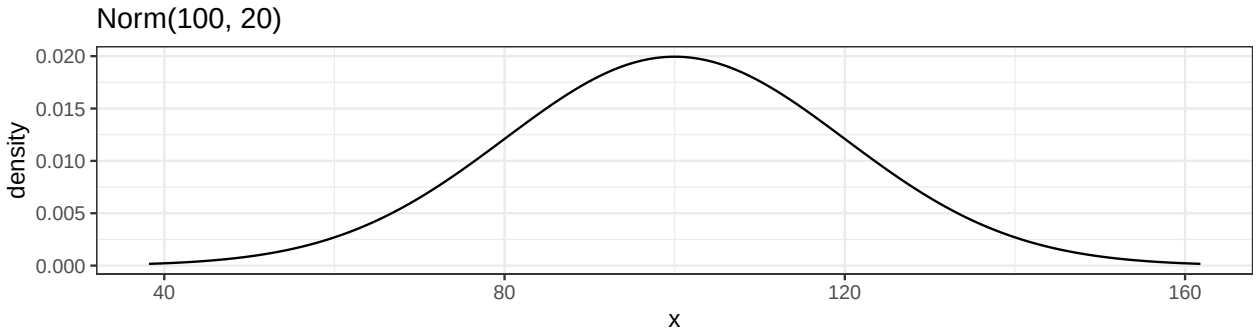


Actually, there are many normal distributions. They all have the same basic shape, but the can have different means (center) and standard deviations (spread).



We will denote a normal distribution with mean μ and standard deviation σ as $\mathsf{Norm}(\mu, \sigma)$.
For example, here is the Norm(100,20)-distribution.

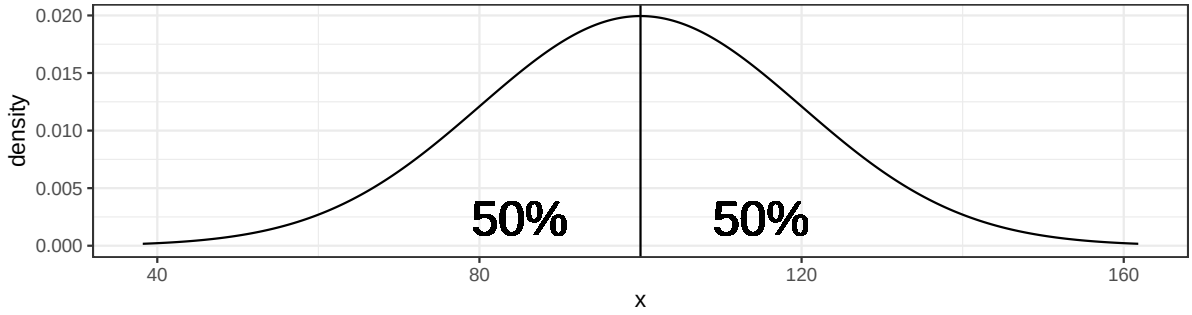
```
gf_dist("norm", mean = 100, sd = 20, title = "Norm(100, 20)")
```



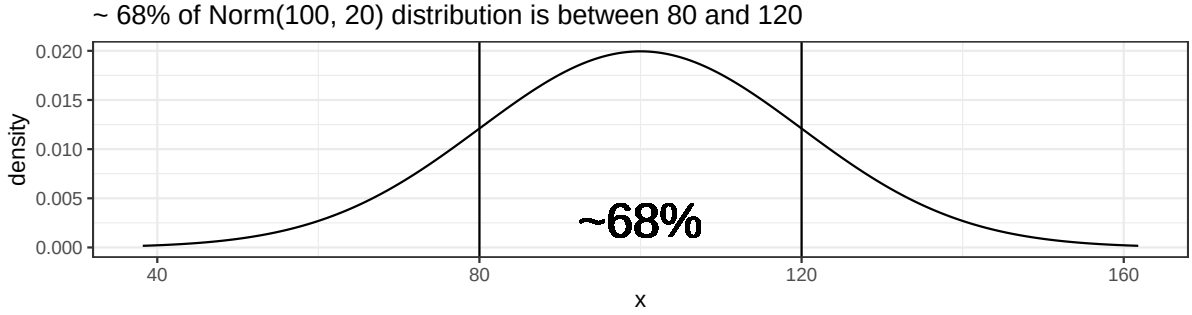
68- 95 - 99.7

These distributions have a number of special mathematical properties. Here are four important ones for us.

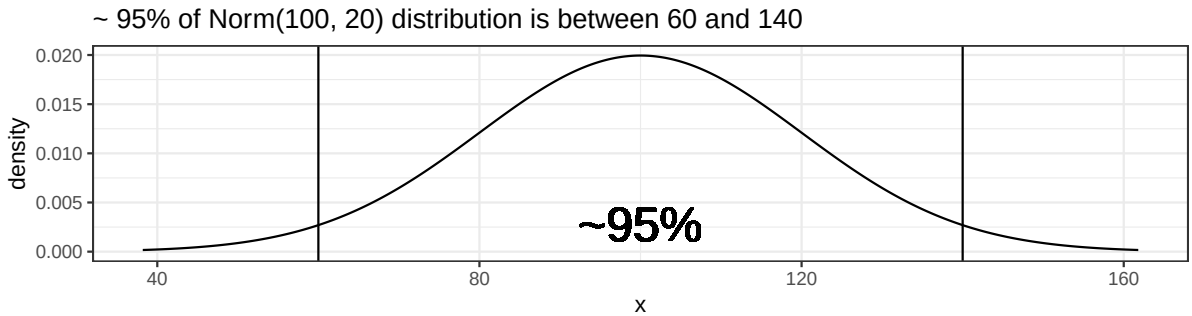
- Since they are symmetric, the mean, median, and mode are all the same. Half of the distribution is below the mean and half is above.



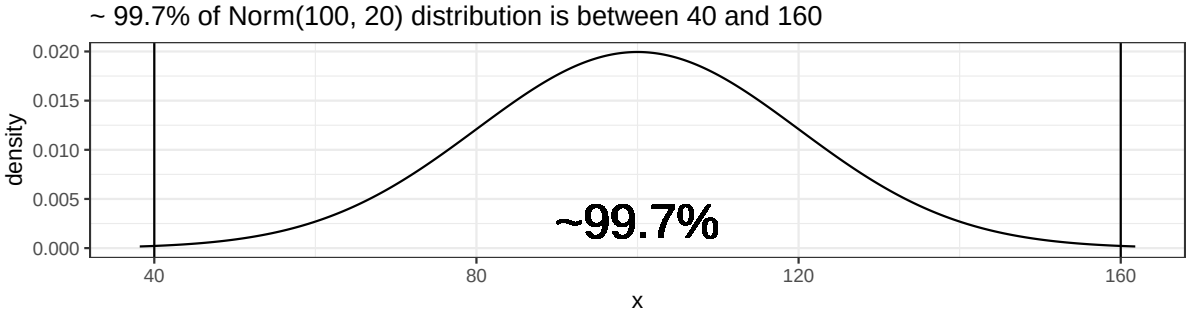
- ~68% of the distribution is between 1 standard deviation below the mean and 1 standard deviation above the mean.



- ~95% of the distribution is between 2 standard deviations below the mean and 2 standard deviations above the mean.



- ~99.7% of the distribution is between 3 standard deviations below the mean and 3 standard deviations above the mean.



- The same sort of thing is true for any number of standard deviations up and down from the mean, but we won't memorize any more of these. Instead, we will let R compute the values for us. (Stay tuned for how this is done.)

Examples

1. A certain IQ test has scores that are approximately normally distributed with a mean of 100 and a standard deviation of 15.
    a. What percent of people have an IQ above 100?
    b. What percent of people have an IQ between 85 and 115?
    c. What percent of people have an IQ above 115?
    d. What percent of people have an IQ below 85?
    e. What percent of people have an IQ above 85?
    f. What percent of people have an IQ abovce 130?
    g. What percent of people have an IQ between 85 and 130?
2. In a normal distribution with mean 20 and standard deviation 2,
    a. What proportion of the distribution is below 20?
    b. What proportion of the distribution is between 18 and 22?
    c. What proportion of the distribution is above 22?
    d. What proportion of the distribution is between 18 and 24?
    e. What proportion of the distribution is below 16?

Z-scores

The most important thing to know about a value in a normal distribution is how many standard deviations above or below the mean it is. This number is called its **z-score** or **standardized score**.

$$z = \frac{\text{value} - \text{mean}}{\text{standarddeviation}}$$

Examples

1. A certain IQ test has scores that are approximately normally distributed with a mean of 100 and a standard deviation of 15.
    a. Alice's IQ score is 105. What is her z-score?
    b. Barb's IQ score is 125. What is her z-score?
    c. Carla's IQ score is 90. What is her z-score?

2. The heights of men and women in the US ages 18–24 are approximately normal. For men, the mean is 70 inches with a standard deviation of 2.8 inches. For women, the mean is 64.3 inches with a standard deviation of 2.6 inches.

    a. What is your z-score? (If you only know your height in cm, you can convert your height to inches by dividing by 2.54, or you can convert the mean and standard deviation to cm by multiplying by 2.54.)
    b. Denise is 5 foot 10 inches tall. What is her z-score?
    c. Edward is 6 foot 4 inches tall. What is his z-score?
    d. What is more unusual, for a woman to be at least 5 foot 8 or for a man to be at least 6 foot 2?

3. Use the information below to compute the z-score for each test statistic.

```
diffprop(malaria ~ group, data = Malaria)
```

```
##  diffprop
## 0.6428571
```

```
df_stats(~diffprop, data = Malaria_null)
```

```
##   response        min          Q1      median        Q3       max        mean
## 1 diffprop -0.7857143 -0.07142857 -0.07142857 0.1666667 0.6428571 0.001190476
##          sd    n missing
## 1 0.2464614 2000       0
```

```
diffprop(decision ~ gender, data = Promotion)
```

```
##   diffprop
## -0.2916667
```

```
df_stats(~diffprop, data = Promotion_null)
```

```
##   response        min     Q1      median         Q3   max         mean       sd
## 1 diffprop -0.4583333 -0.125 -0.04166667 0.04166667 0.375 -0.009458333 0.131084
##      n missing
## 1 2000       0
```

```
diffmean(Taps ~ Group, data = CaffeineTaps)
```

```
## diffmean
##     -3.5
```

```
df_stats(~diffmean, data = Taps_null)
```

```
##   response  min   Q1 median  Q3 max    mean       sd    n missing
## 1 diffmean -4.3 -0.9   -0.1 0.9 3.5 -0.0412 1.318792 2000       0
```

4. What does the z-score of a test-statistic tell you about the p-value (assuming the null distribution is approximately normal)?

    a. Roughly how large must the z-score be to get a 2-sided p-value below 0.05?
    b. If a 1-sided test as that same z-score will the p-value be larger or smaller than 0.05?

When z-scores aren't integers

Of course, not every value is 0, 1, 2, or 3 standard deviations away from the mean. We could memorize even more values, but that would quickly get tedious. Instead we will let R compute these values for us as we need them.

The one slightly tricky thing is that R works with "below" rather than "between". (For most uses this is actually easier.) The two functions we need are `pnorm()` and `qnorm()`.

- `pnorm(x, mean, sd)` computes the **proportion** of the normal distribution with mean `mean` and standard deviation `sd` that is **below** `x`.
- `qnorm(p, mean, sd)` computes the **value** in the normal distribution with mean `mean` and standard deviation `sd` **that has proportion** `p` **below it**.

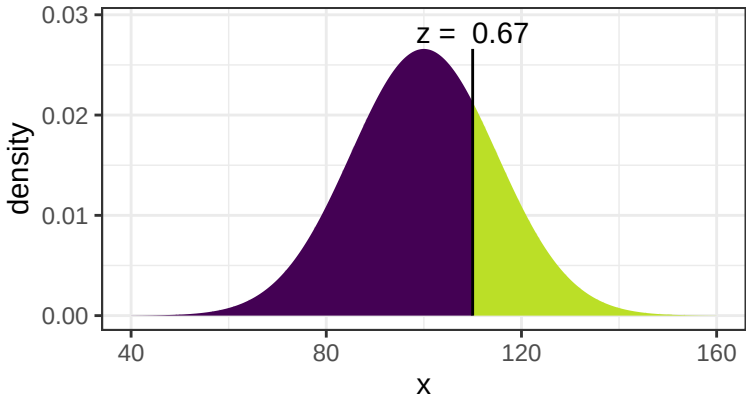
If you want to get fancy, you can use `xpnrom()` and `qpnorm()`. These functions will additionally draw a picture of the normal distribution for you.

Examples Here are some examples. The first one is worked out for you.

1. Going back to the IQ test with mean 100 and standard deviation 15,
   a. What proportion of IQ tests are below 110?

```
xpnorm(110, mean = 100, sd = 15)
```

```
##
## If X ~ N(100, 15), then
##  P(X <= 110) = P(Z <= 0.6667) = 0.7475
##  P(X >  110) = P(Z >  0.6667) = 0.2525
##
```




```
## [1] 0.7475075
```

   b. What proportion of IQ tests are between 90 and 110?

```
# We can get this be subtracting the part that is below 90
pnorm(110, mean = 100, sd = 15) - pnorm(90, mean = 100, sd = 15)
```

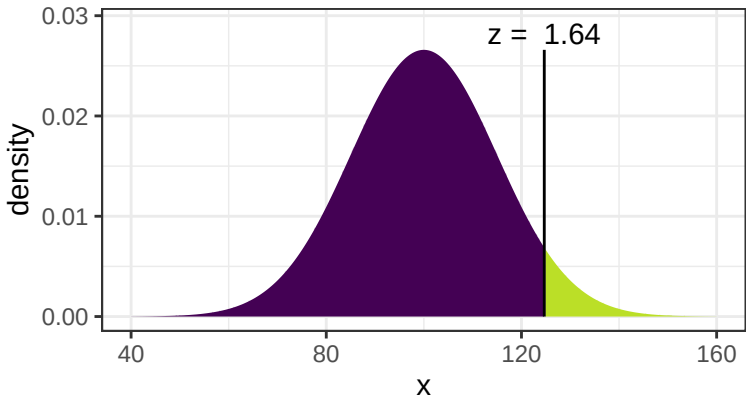
```
## [1] 0.4950149
```

   c. Don scores at the 95th percentile. What is his IQ score?

```
xqnorm(0.95, mean = 100, sd = 15)
```

```
##
```

```
## If X ~ N(100, 15), then
##  P(X <= 124.6728) = 0.95
##  P(X >  124.6728) = 0.05
##
```




```
## [1] 124.6728
```

2. More IQ

    a. What proportion of these IQ tests are below 80?
    b. What proportion of these IQ tests are above 125?
    c. Fred's IQ is at the 75th percentile, what is his score? What proportion of IQ tests are lower than Fred's score?

3. https://nces.ed.gov/programs/digest/d17/tables/dt17_226.40.asp lists the mean and standard deviation for SAT scores in each state. For Michigan, the mean is 1005 and the standard deviation is 195. The distibution is approximately normally distributed.

    a. Approximately what proportion of Michigan students score below 1200 on the SAT?
    b. Approximately what proportion of Michigan students score above 1200 on the SAT?
    c. Approximately what proportion of Michigan students score above 1400 on the SAT?
    d. What score is the 90th percentile for this test?

4. We can ask questions about z-scores by using normal distribution with mean 0 and standard deviation 1. (These are the default values, so you can actually omit them altogether if you like.)

    a. What proportion of a normal distribution has a z-score below 1.5?
    b. What proportion of a normal distribution has a z-score between -1.5 and 1.5?
    c. What is the z-score for the 85th percentile?

    This gives us a second way to work with normal distribuitons: Translate the question to one about z-scores, and work with a Norm(0, 1) distribuiton. This important example of a normal distribution is called the **standard normal** distribution.

5. Z-score review. Fill in the blank.

    a. value = 125; mean = 100; sd = 10; z-score = _____
    b. value = 37; mean = 30; sd = 5; z-score = _____
    c. value = _____; mean = 250; sd = 20; z-score = 1.5
    d. value = _____; mean = 200; sd = 30; z-score = -2.5