POLS201 Spring 2019

Central Limit Theorem and Midterm

March 11

Agenda

- Friday refresh and Central Limit Theorem
- Review the Midterm (1:15)
- Bring your laptops Wednesday and Friday!
- Tonight is the Building Assignment deadline.

The Kellstedt and Whitten Treatment of this topic. . .

- . . . is excellent.
- Between our discussion today and their description, you should have a good handle on it.

What was so special about the normal distribution?

- What is the utility of a Z-Table?
- What is one “Z”?
- What can you do with this formula?

$$Z = (x - \mu) \div \sigma$$

Practice Problem

- Suppose SAT scores are normally distributed with a mean of 1000 and standard deviation of 200.
- If you score 800, what percentile are you in?

$$Z = (x - \mu) \div \sigma$$

- x = your data point
- μ = mean of your data
- σ = standard deviaion of your data

And the answer is. . .

- $Z = (x - \mu) \div \sigma$

And the answer is. . .

- $Z = (x - \mu) \div \sigma$
- $Z = (800 - 1000) \div 200$

And the answer is. . .

- $Z = (x - \mu) \div \sigma$
- $Z = (800 - 1000) \div 200$
- $Z = -200 \div 200$

And the answer is. . .

- $Z = (x - \mu) \div \sigma$
- $Z = (800 - 1000) \div 200$
- $Z = -200 \div 200$
- $Z = -1$

And the answer is. . .

- $Z = (x - \mu) \div \sigma$
- $Z = (800 - 1000) \div 200$
- $Z = -200 \div 200$
- $Z = -1$
- A Z of -1 on Z-table is 0.1587

And the answer is. . .

- $Z = (x - \mu) \div \sigma$
- $Z = (800 - 1000) \div 200$
- $Z = -200 \div 200$
- $Z = -1$
- A Z of -1 on Z-table is 0.1587
- A score of 800 is in the 16th percentile

The Take Home Point

- When data are normally distributed (height, weight, SAT), we can easily describe how it falls in the distribution / what the distribution looks like.
- But many distributions are not normal. . . .so why is this idea so useful?

Because of the Central Limit Theorem

- The Central Limit Theorem states that the means of random samples drawn from ANY distribution will (skewed, binomial, or completely undefined) **form a normal distribution** as the number of samples increases.

Die Example

- Let's roll a six sided die.
- We'll simulate the rolling process on the computer.
- In R. To demonstrate, let's roll one die ten times.
- Does every number from 1 to 6 appear? Why or why not?

```
sample(1:6, size = 10, replace = TRUE)
```

```
##  [1] 4 5 4 5 1 1 5 1 6 4
```

We can “roll the die” as many times as we want

- So let’s roll the die six million times and store the result

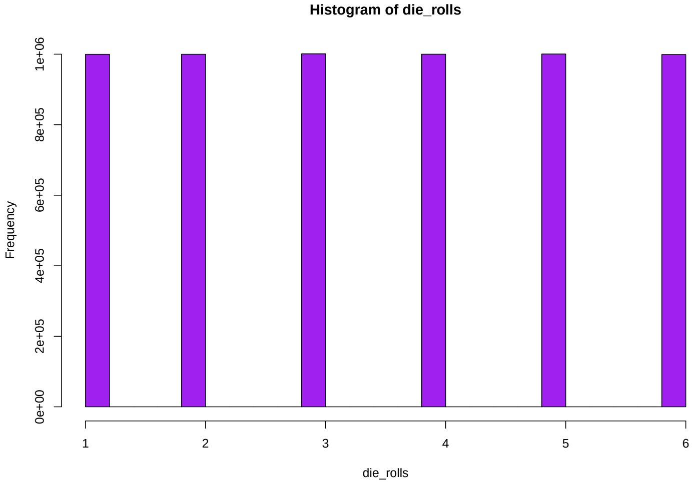
```
die_rolls <- sample(1:6,
                    size = 6000000,
                    replace = TRUE)

head(die_rolls)
```

```
## [1] 6 3 5 4 4 3
```

- Our six million rolls are stored in a data object called *die_rolls*

Each value should appear pretty close to a million times


```r
hist(die_rolls, col = "purple")
```



. . . and they do


```r
table(die_rolls)
```

```
## die_rolls
##       1       2       3       4       5       6
##  999740  999803 1000977  999955 1000444  999081
```

- So let's call our 6,000,000 set of rolls our "population"
- Easy to see this rendition of a "population", but we can get even fancier, purer, and more abstract.

We can assert theoretically that...

- Die rolls are generated from a universe where each of six numbers has an equal chance of appearing.
- Rolling the die is our "data generating process" or DGP
- Each value has exactly a $1/6$ probability.
- That kind of distribution has a name: *discrete uniform*.

The population mean of all die rolls

- For now, let's simplify things and suppose *die_rolls* is the population.

```
mean(die_rolls)
```

```
## [1] 3.499801
```

The population mean of all die rolls

- For now, let's simplify things and suppose *die_rolls* is the population.
- What do we expect the mean to be?

```
mean(die_rolls)
```

```
## [1] 3.499801
```

So what happens if we roll a die 16 times and compute the mean?

- Let's try a sequence of 16 rolls just once.

```
(example <- sample(die_rolls, 16))
```

```
##  [1] 4 1 6 1 4 4 1 6 1 3 5 5 6 1 6 5
```

```
mean(example)
```

```
## [1] 3.6875
```

How about five repetitions? Instead of just one?

- You can see a little variation

```
## # A tibble: 5 x 2
##    name value
##   <int> <dbl>
## 1     1  3.56
## 2     2  3.56
## 3     3  3.06
## 4     4  3.81
## 5     5  3.44
```

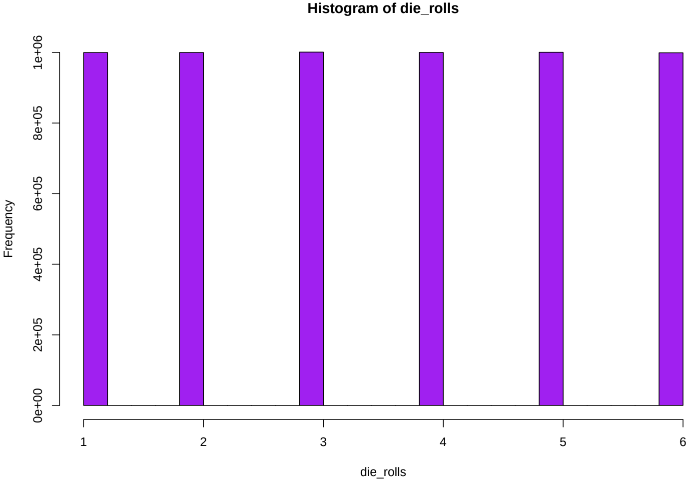
Now. . . let's increase the reps from five to. . .

- Oh, how about 10,000. Each time, we roll the die 16 times

```
## # A tibble: 10,000 x 2
##     name value
##    <int> <dbl>
##  1     1  3.75
##  2     2  4.31
##  3     3  3.62
##  4     4  4.62
##  5     5  4
##  6     6  3.94
##  7     7  3.12
##  8     8  4.12
##  9     9  3.12
## 10    10  3.25
## # ... with 9,990 more rows
```

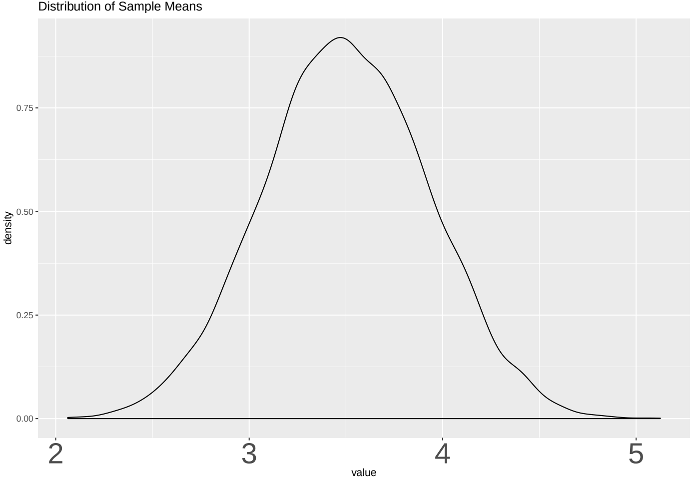
Remember this distribution of our population from before?

- This is something called a discrete uniform distribution
- It doesn't look normal, does it?



But behold the distribution of sample means

- It is converging to be perfectly normal
- The bigger the number, the more perfect it gets



Sampling Distribution of the Sample Mean

- The normal distribution produced by this exercise of drawing samples and plotting their means is called the **Sampling Distribution of the Sample Mean**

This gives us the Standard Error

- A standard error describes the standard deviation of some kind of sampling distibution
- Any time we generate a value based on a sample, we can generate a standard error
- In this case, standard error of our new normal distribution is computed like so:

$$\sigma \div \sqrt{n}$$

- Remember that σ is the population standard deviation
  - and the population distribution can be *anything*

We can derive the standard error with

- The standard deviation of the population $\div$ square root of the number of the rolls. In our case, we roll 16 times.
- Standard error $= \sigma \div \sqrt{n}$
- What is σ? What is the standard error?

```
sigma <- sd(die_rolls)
standard_error <- sigma / sqrt(16)
print(sigma)
```

```
## [1] 1.71
```

```
print(standard_error)
```

```
## [1] 0.427
```

We can double check this theoretical result. . .

- . . . against the standard deviation of the sample means. We get

```
sd(sample_means$value)
```

```
## [1] 0.426
```

- Which ought to be very, very close to

```
print(standard_error)
```

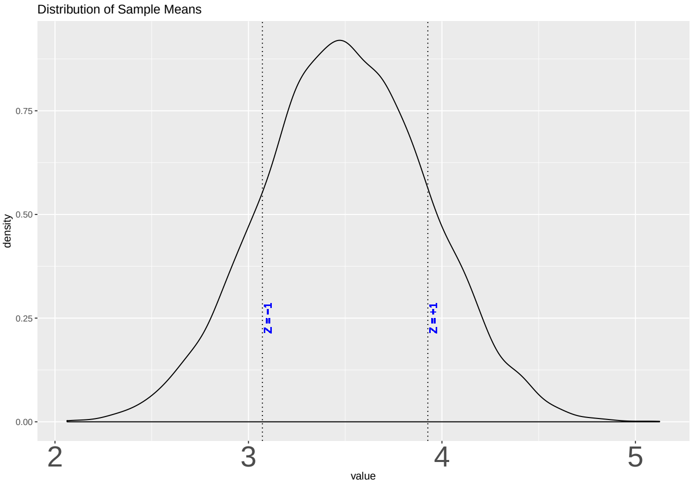
```
## [1] 0.427
```

Let's draw a line at the Z Scores -1 and +1

- Remember $Z = (x - \mu) \div \sigma$
- So we get $-1 = (x - 3.5) \div .427)$
- Leading us to $-.427 = x - 3.5$ and then
- $x = 3.5 - .427 = 3.072$

Let's draw a line at the Z Scores -1 and +1

- Remember $Z = (x - \mu) \div \sigma$
- So we get $+1 = (x - 3.5) \div .427)$
- Leading us to $+.427 = x - 3.5$ and then
- $x = 3.5 + .427 = 3.927$


Distribution of Sample Means
density
value
0.00
0.25
0.50
0.75
2
3
4
5
Z = −1
Z = +1

Distributions

	Population Distribution	Sample Distribution	Sampling Distribution of the sample mean
Shape	Irregular	Irregular	Normal
Mean	μ	$\bar{x}$	μ
Standard Deviation	σ	s	$\frac{\sigma}{\sqrt{n}}$

Politics Ain't Beanbag

- You are an experienced campaign staffer. You know that the average contributor gives \$100 (sd of \$50).
- You have one campaign event to raise funds. You invite a random sample of 100 pledged contributors and hope for the best.
- The campaign will cost \$9,000, which means you need your sample to average \$90 in contributions.
- What is the probability your campaign will go bankrupt?

You can treat this sample as if . . .

- . . . it came from a normal distribution.
- We solve for Z

$$Z = (x - \mu) \div \sigma$$

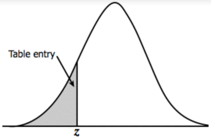
$$Z = (90 - 100) \div (50/sqrt(100))$$

$$Z = (90 - 100) \div (50/10)$$

$$Z = (-10) \div (5)$$

$$Z = -2$$

Don't worry, you probably won't go bankrupt



z	.00	.01	.02
−3.4	.0003	.0003	.0003
−3.3	.0005	.0005	.0005
−3.2	.0007	.0007	.0006
−3.1	.0010	.0009	.0009
−3.0	.0013	.0013	.0013
−2.9	.0019	.0018	.0018
−2.8	.0026	.0025	.0024
−2.7	.0035	.0034	.0033
−2.6	.0047	.0045	.0044
−2.5	.0062	.0060	.0059
−2.4	.0082	.0080	.0078
−2.3	.0107	.0104	.0102
−2.2	.0139	.0136	.0132
−2.1	.0179	.0174	.0170
−2.0	.0228	.0222	.0217
−1.9	.0287	.0281	.0274
−1.8	.0359	.0351	.0344

- Don't worry, it is unlikely you will go bankrupt!
- 2.28% chance

For Wednesday: Bring your computers

- Be sure to review Kellstedt and Whitten
- Let's review the midterm