Central to Stats: Sampling!

INTRODUCTION TO STATISTICS IN SPREADSHEETS

Ted Kwartler
Data Dude

datacamp

Lesson Overview

- Samples vs populations
- Central Limit Theorem (CLT)

So far, you have...

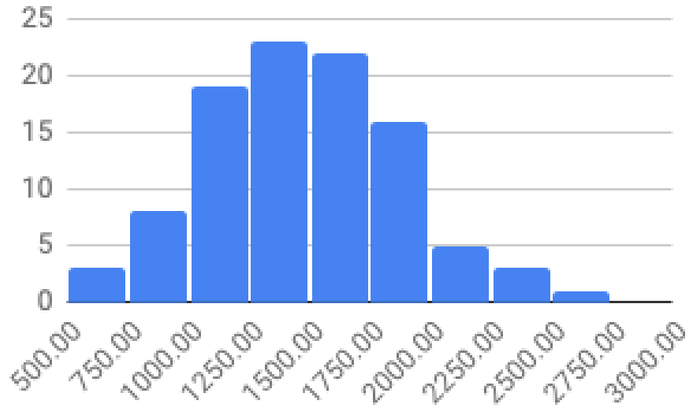
- Calculated descriptive statistics

Mean:	1,483.33
Median:	1,481.67
Mode:	1481.668
Standard Deviation:	400.2464

So far, you have...

- Calculated descriptive statistics
- Made data visualizations

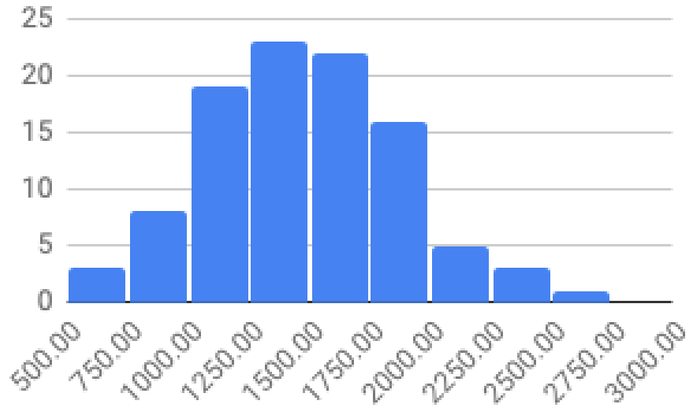
Mean:	1,483.33
Median:	1,481.67
Mode:	1481.668
Standard Deviation:	400.2464



So far, you have...

- Calculated descriptive statistics
- Made data visualizations
- Used **all** of the data

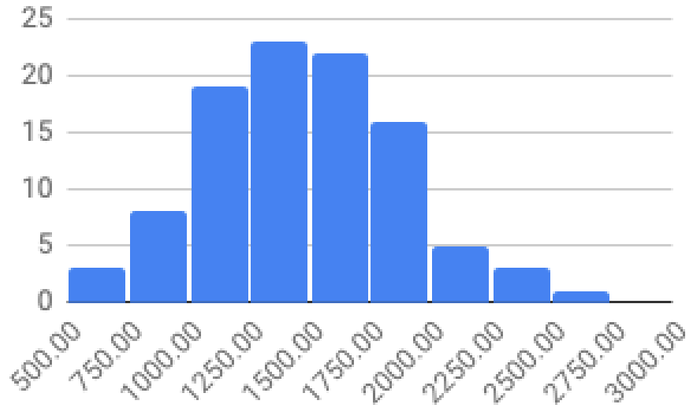
Mean:	1,483.33
Median:	1,481.67
Mode:	1481.668
Standard Deviation:	400.2464



So far, you have...

- Calculated descriptive statistics
- Made data visualizations
- Used **all** of the data
- Worked with "populations"

Mean:	1,483.33
Median:	1,481.67
Mode:	1481.668
Standard Deviation:	400.2464



What is a population?

- An *entire* distribution of observations/events
- Costly and time consuming to work with
- Better to "sample" the population

Sampling to the rescue

- A *subset* from a population
- Meant to represent the population
- The larger the sample size, the closer the statistics of the sample will emulate the statistics of the population

- If a sample size from an independent, random variable is **large enough**, then the sampling distribution will be normal or nearly normal

Central Limit Theorem (CLT)

- "Large enough" is vague...
- How accurate do you need to be? This affects the sample size needed
- The more closely the population follows a normal distribution, the fewer samples will be required
- **Minimum number needed is between 30 and 40**

Off to do some sampling!

INTRODUCTION TO STATISTICS IN SPREADSHEETS

Hypothesis Testing

INTRODUCTION TO STATISTICS IN SPREADSHEETS

Ted Kwartler
Data Dude

datacamp

Anatomy of hypothesis testing

"The average Ferrari price is higher than the average sports car price"

- Hypothesis: Any testable claim

Null Vs Alternate Testable Hypotheses

- **Null Hypothesis**
  - Represents the status quo (accepted fact)
  - Represented by `H0`
- Example `H0` : Average Ferrari Price **EQUALS** Average Sports Car Price
- **Alternate/Research Hypothesis**
  - Challenger statement
  - Represented by `H1`
- Example `H1` : Average Ferrari Price **DOES NOT EQUAL** Average Sports Car Price

Common sense testing

Null Hypothesis (`H0`):

> Average Ferrari Price equals Average Sports Car Price

Alternate Hypothesis (`H1`):

> Average Ferrari Price does not equal Average Sports Car Price

- Average Ferrari Price (Sample Size = 50) = $252,000
- Average non-Ferrari Sports Car Price (Sample Size = 50) = $85,000
- `REJECT` the Null Hypothesis (`H0`)

Another common sense test

Null Hypothesis (`H0`):

> Average Toyota Price equals Average Honda Price

Alternate Hypothesis (`H1`):

> Average Toyota Price does not equal Average Honda Price

- Average Toyota Price (Sample Size = 50) = $23,845
- Average Honda Price (Sample Size = 50) = $23,720
- `FAIL TO REJECT` the Null Hypothesis (`H0`)

Removing subjectivity in a test

- Hypothesis tests use "test statistics" to verify hypotheses
- Example: `t-test` (In Spreadsheets: `T.TEST(range1, range2, tails, type)`)
  - Produces a "p-value"
  - p-value: Probability that the results you see are due to chance/error
  - Choose a p-value cutoff (Example: 1%/0.01, or 5%/0.05)
  - If p-value Is less than the cutoff, `REJECT` the Null Hypothesis
    - Conclude that there **IS** a difference between the samples

- Spreadsheets Formula:

```
= T.TEST(range1, range2, tails, type)
```

- If testing **only** `GREATER THAN` / `>` or `LESS THAN` / `<`:
  - The test has `1` tail
- If the `H0` operator is `EQUALS` / `=`:
  - `2` tails, as the values can be above **or** below the mean
- If measuring the same observations at different times:
  - Type = `1`
- If measuring different observations with same variance:
  - Type = `2`
- If measuring different observations with different variances:
  - Type = `3`

Let's practice!

INTRODUCTION TO STATISTICS IN SPREADSHEETS

Hypothesis Testing with the Z-test

INTRODUCTION TO STATISTICS IN SPREADSHEETS

Ted Kwartler
Data Dude

datacamp

T-Tests

- Spreadsheets Formula:

```
T.TEST(range1, range2, tails, type)
```

- Infers whether there is a difference between two means

Comparing T-Tests and Z-Tests

Similarities

- Determine whether two population means are statistically different
- Select a p-value cutoff prior to the test (Usually `0.05`)
- If the resulting p-value is less than `0.05` , then `REJECT H0`
  - Else, `FAIL TO REJECT H0` .

Contrasting T-Tests and Z-Tests

Z-Test

Formula:

```
=Z.TEST(range1, testStatistic, StDev)
```

- Needs 1 range
- Needs a test statistic (i.e, population mean)
- Used with bigger datasets (n > 30)

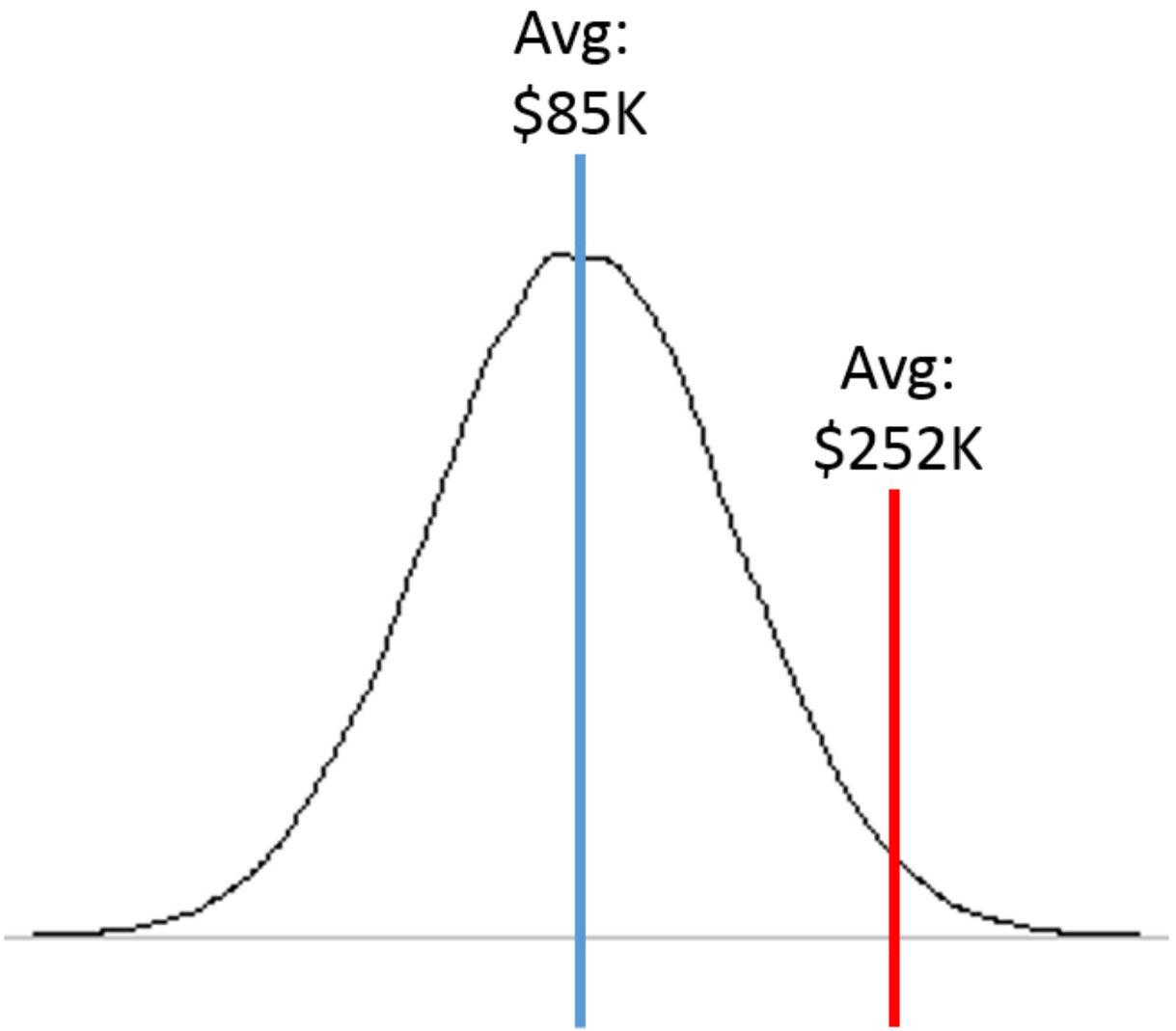
T-Test

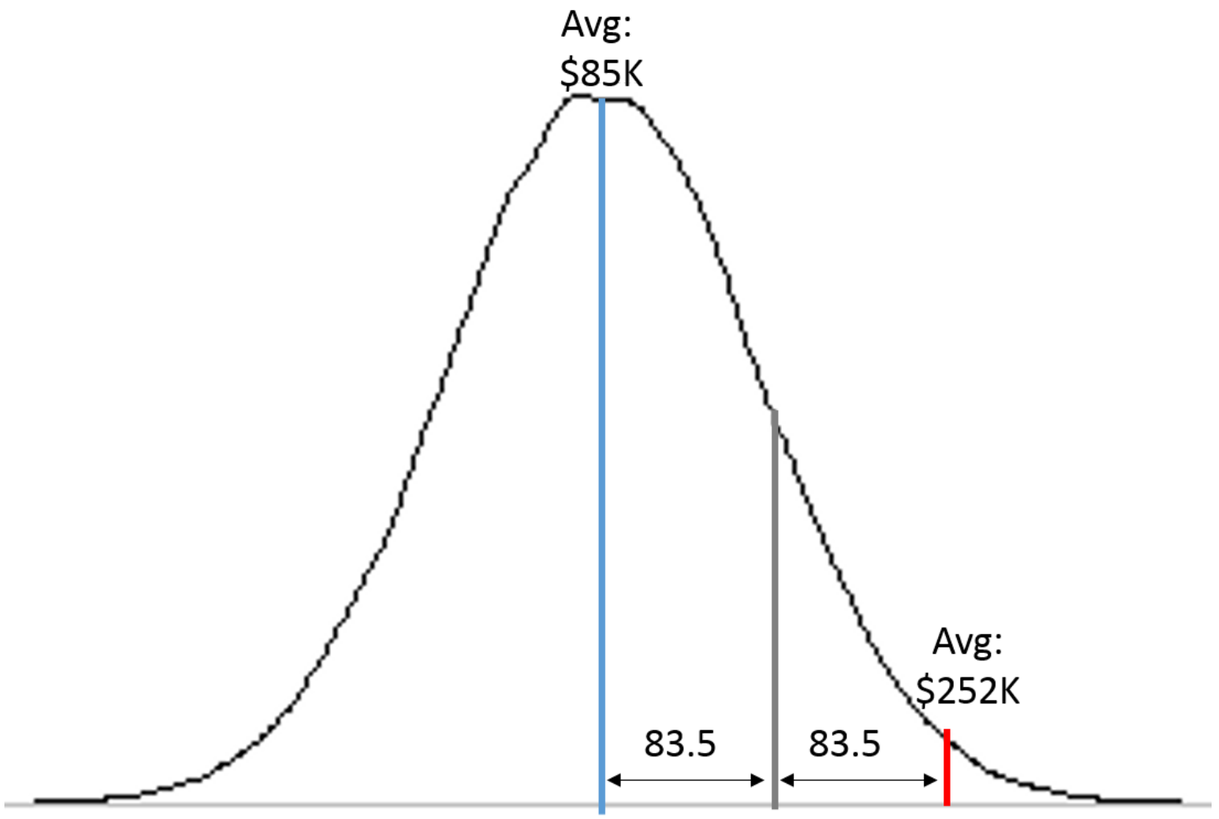
Formula:

```
T.TEST(range1, range2, tails, type)
```

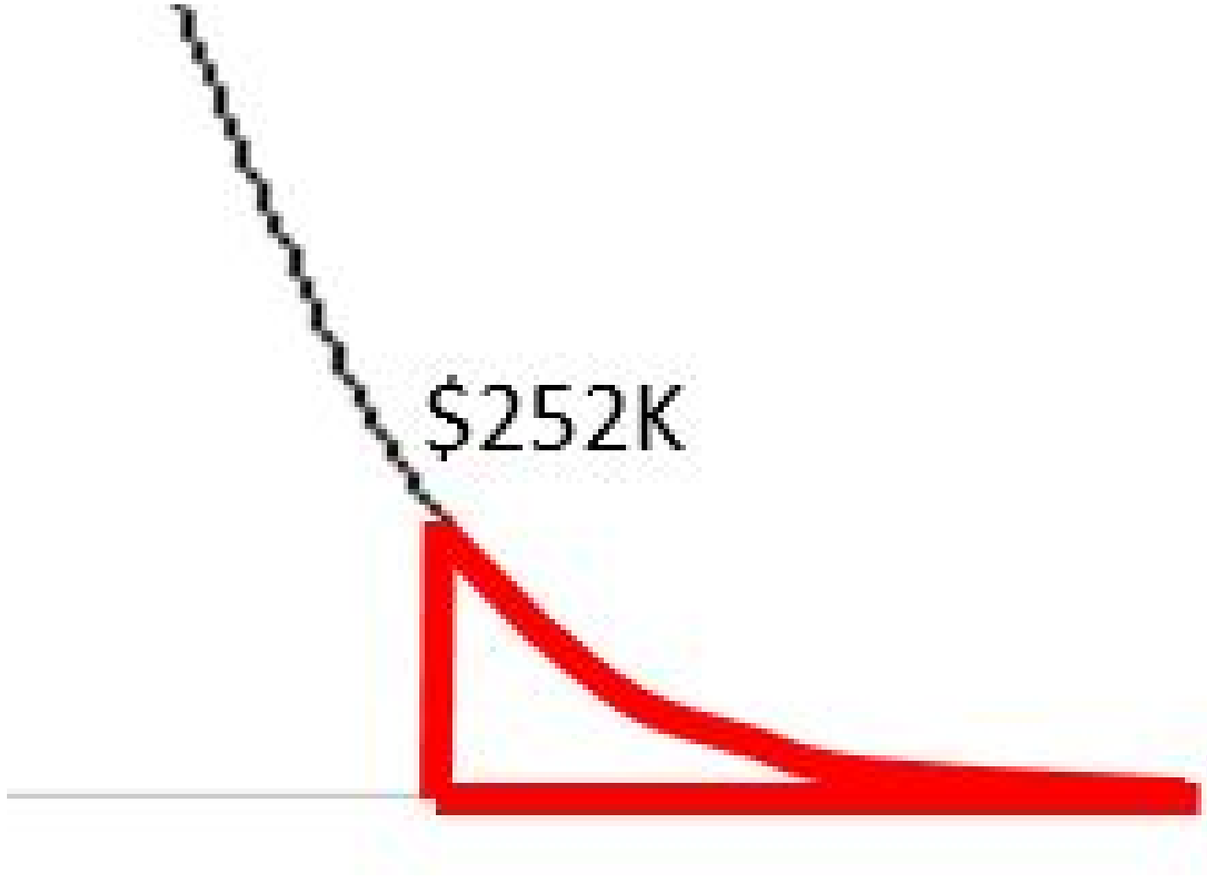
- Needs 2 ranges
- Works when variance is unknown
- Used with smaller datasets (n < 30)

[1] https://keydifferences.com/difference-between-t-test-and-z-test.html


Avg:
$85K
Avg:
$252K


Avg:
$85K
Avg:
$252K
83.5
83.5

Z-Tests calculate the probability



Z-Tests in Spreadsheets

```
Z.TEST(data, value, [standard_deviation])
```

- `FAIL TO REJECT` `H0` if the -value is greater than `0.05`
- `REJECT` `H0` if the p-value is less than `0.05`

Let's practice!

INTRODUCTION TO STATISTICS IN SPREADSHEETS

Hypothesis Testing with the Chi-squared test

INTRODUCTION TO STATISTICS IN SPREADSHEETS

Ted Kwartler
Data Dude

Applications of the chi-squared test

- Comparing samples for meaningful differences
- Example: Did the treatment actually work?

[1] https://www.pexels.com/photo/20-mg-label-blister-pack-208512/

Testing before and after

Two possibilities:

- The difference in the new group stems from random sampling (`H0`)
- There is in fact a meaningful difference (`H1`)

Chi-squared tests also provide a p-value

Chi-squared test conditions

For a chi-squared test to be useful:

- Data has to be in groups (e.g: "Old treatment", "New treatment")
- Avoid really small expected values (`< 5`)

Testing the independence of two groups

- "Clinically proven"
- "Lab tested"
- Experiment needs to control for lifestyle factors
- Compare weight within two groups
  - One treated with the supplement, and the other not

[1] https://www.pexels.com/photo/woman-girl-jeans-clothes-53528/

Chi-squared test in Spreadsheets

```
CHITEST(observed_range, expected_range)
```

- `FAIL TO REJECT` `H0` if the p-value is greater than `0.05`
- `REJECT` `H0` if the p-value is less than `0.05`

Let's practice!

INTRODUCTION TO STATISTICS IN SPREADSHEETS