DATA SCIENCE WITH R

Statistics Part 1 Review

1. **Two types of Statistics**
   - Descriptive
   - Inferential

2. **Random Variables**

3. **Probability Distributions – outcomes of random variables**
   - Discrete
   - Continuous

4. **Probability Distribution Functions**
   - Probability calculations using Tables – distances
   - Probability calculations using Formulae

DATA
SCIENCE
WITH R

STATISTICS

★ Hypothesis Testing ★

DATA SCIENCE WITH R

HYPOTHESIS TESTING

→ Introduction to Hypothesis Testing

Basic Framework of a Hypothesis Test

Distance Measures

Central Limit Theorem

Types of Hypothesis Tests

DATA
SCIENCE
WITH R

HYPOTHESIS TESTING

Introduction to Hypothesis Testing

➔ Basic Framework of a Hypothesis Test

Distance Measures

Central Limit Theorem

Types of Hypothesis Tests

HYPOTHESIS TESTING

Introduction to Hypothesis Testing

Basic Framework of a Hypothesis Test

→ Distance Measures

Central Limit Theorem

Types of Hypothesis Tests

HYPOTHESIS TESTING

Introduction to Hypothesis Testing

Basic Framework of a Hypothesis Test

Distance Measures

→ Central Limit Theorem

Types of Hypothesis Tests

HYPOTHESIS TESTING

Introduction to Hypothesis Testing

Basic Framework of a Hypothesis Test

Distance Measures

Central Limit Theorem

Types of Hypothesis Tests

HYPOTHESIS TESTING

➔ **Introduction to Hypothesis Testing**

Basic Framework of a Hypothesis Test

Distance Measures

Central Limit Theorem

Types of Hypothesis Tests

Introduction

- When dealing with random variables – you may expect ‘unexpected’ results!

Introduction

- When dealing with random variables – you may expect ‘unexpected’ results!

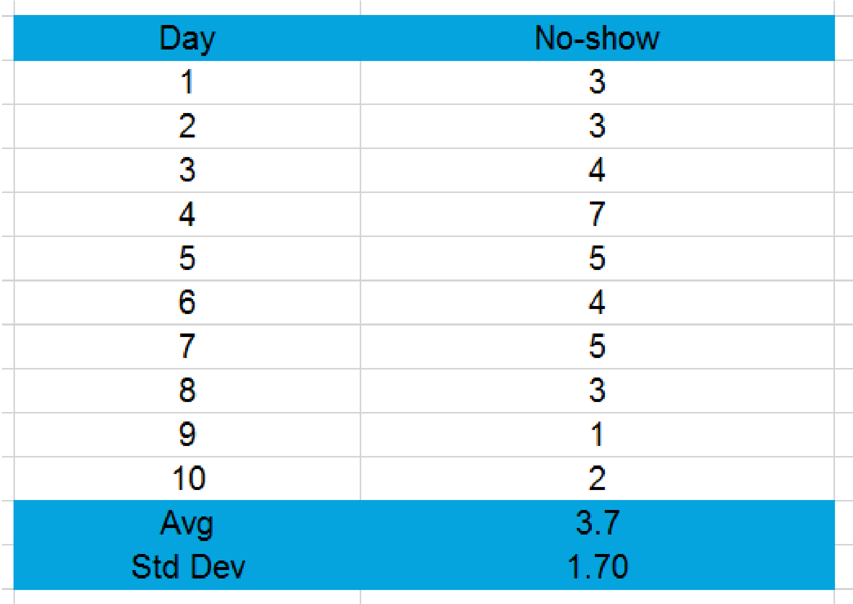
- In the airline no-show example, average no-shows based on 6 months of data is 5%

Introduction

- When dealing with random variables – you may expect ‘unexpected’ results!
- In the airline no-show example, average no-shows based on 6 months of data is 5%
- Your GM wants to check for the next 10 randomly chosen flights

Introduction

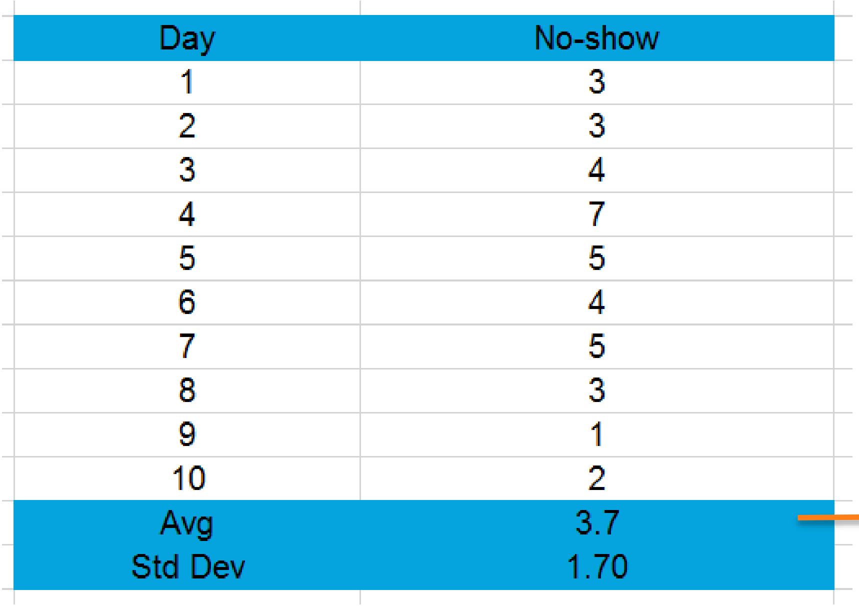
- When dealing with random variables – you may expect ‘unexpected’ results!
- In the airline no-show example, average no-shows based on 6 months of data is 5%
- Your GM wants to check for the next 10 randomly chosen flights



Day	No-show
1	3
2	3
3	4
4	7
5	5
6	4
7	5
8	3
9	1
10	2
Avg	3.7
Std Dev	1.70

Introduction

- When dealing with random variables – you may expect ‘unexpected’ results!
- In the airline no-show example, average no-shows based on 6 months of data is 5%
- Your GM wants to check for the next 10 randomly chosen flights



Day	No-show
1	3
2	3
3	4
4	7
5	5
6	4
7	5
8	3
9	1
10	2
Avg	3.7
Std Dev	1.70

→ **Not equal to 5%**

Introduction

Which number is now right - 5% or 3.7%?

Introduction

Which number is now right - 5% or 3.7%?

Remember:

No-shows is a random variable

With an expected value of 5 (per flight)

Introduction

Which number is now right - 5% or 3.7%?

Remember:

No-shows is a random variable

With an expected value of 5 (per flight)

Two possible explanations for 3.7% -

Introduction

Which number is now right - 5% or 3.7%?

Remember:

No-shows is a random variable

With an expected value of 5 (per flight)

Two possible explanations for 3.7% -

1. Your sample is different from the population

Introduction

Which number is now right - 5% or 3.7%?

Remember:

No-shows is a random variable

With an expected value of 5 (per flight)

Two possible explanations for 3.7% -

1. Your sample is different from the population

2. There is no difference between sample and population – what you are seeing is simply random chance outcome!

Introduction

What options do you have next?

1. Use another sample?

Introduction

What options do you have next?

1. Use another sample?
2. Increase sample size?

Introduction

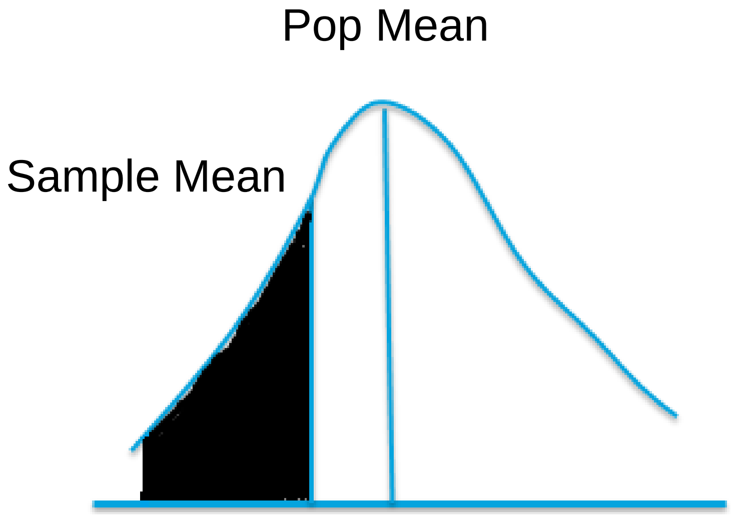
What options do you have next?

1. Use another sample?
2. Increase sample size?
3. **Calculate the random chance probability**

Introduction

What options do you have next?

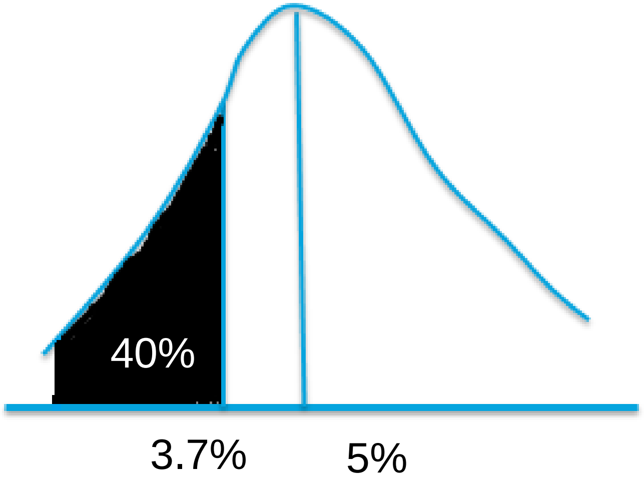
1. Use another sample?
2. Increase sample size?
3. **Calculate the random chance probability**



Introduction

Let's say that we calculate this probability (of seeing a sample mean of 3.7% or less) and find it is 40%

What does that imply?



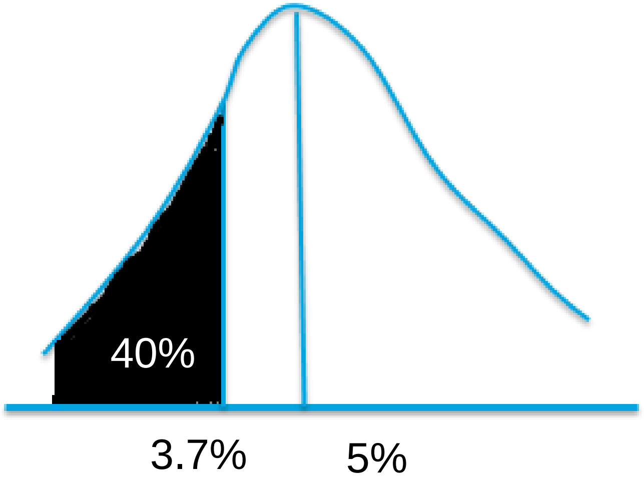
Introduction

Let's say that we calculate this probability (of seeing a sample mean of 3.7% or less) and find it is 40%

What does that imply?

There is a 40% chance that when you pick a random sample from a population with a mean of 5%, you get a sample mean of 3.7% or lower

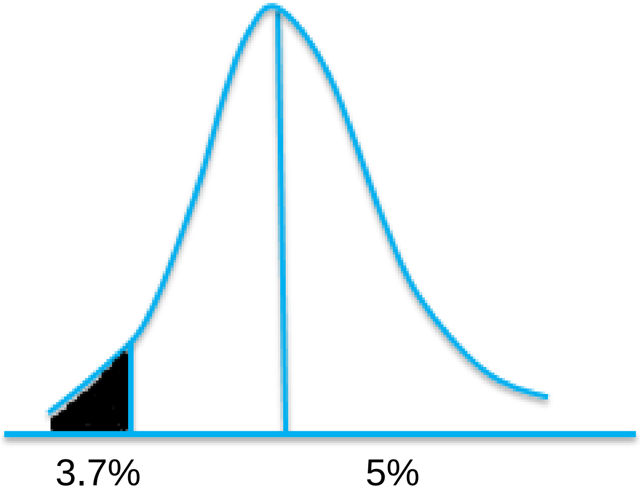
In other words, seeing a 3.7% or lower is pretty likely. You have very little reason to doubt that population average still applies, at 5%



Introduction

What is the probability of seeing 3.7% or less is not 40%, but lower, say 15%?

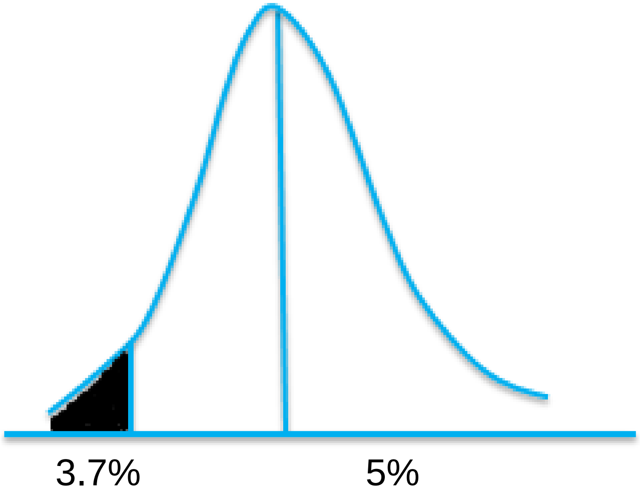
It implies that it is pretty unlikely that is the population mean was 5%, your sample mean would be 3.7% or lower simply because of random chance



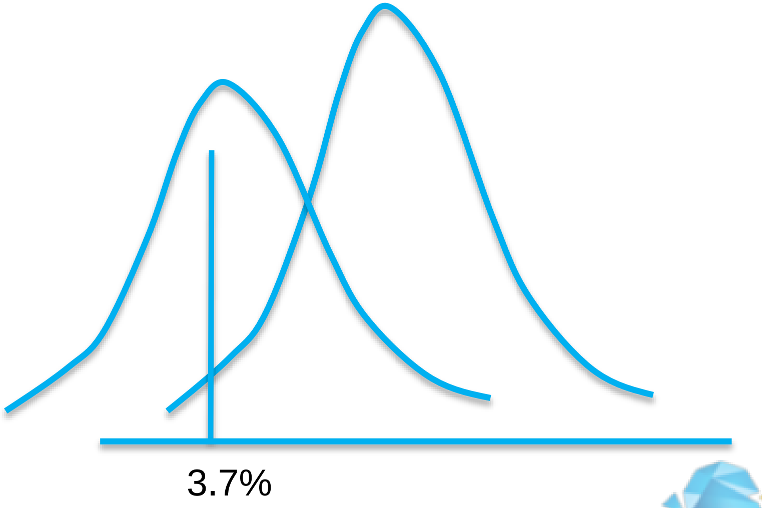
Introduction

What is the probability of seeing 3.7% or less is not 40%, but lower, say 15%?

It implies that it is pretty unlikely that is the population mean was 5%, your sample mean would be 3.7% or lower simply because of random chance



In other words, your sample is more likely to have come from a population with different (lower) mean than the one you are looking at -



Or, your sample is **different** from your population!

Introduction

Another example: Quality Control

Introduction

Another example: Quality Control

At your manufacturing unit, the mean weight of part A produced is expected to be 2.5 lbs, with a std dev of 0.12 lbs, and is normally distributed.

Introduction

Another example: Quality Control

At your manufacturing unit, the mean weight of part A produced is expected to be 2.5 lbs, with a std dev of 0.12 lbs, and is normally distributed.

You take a random sample of 45 units and find that mean weight is 2.68 lbs.

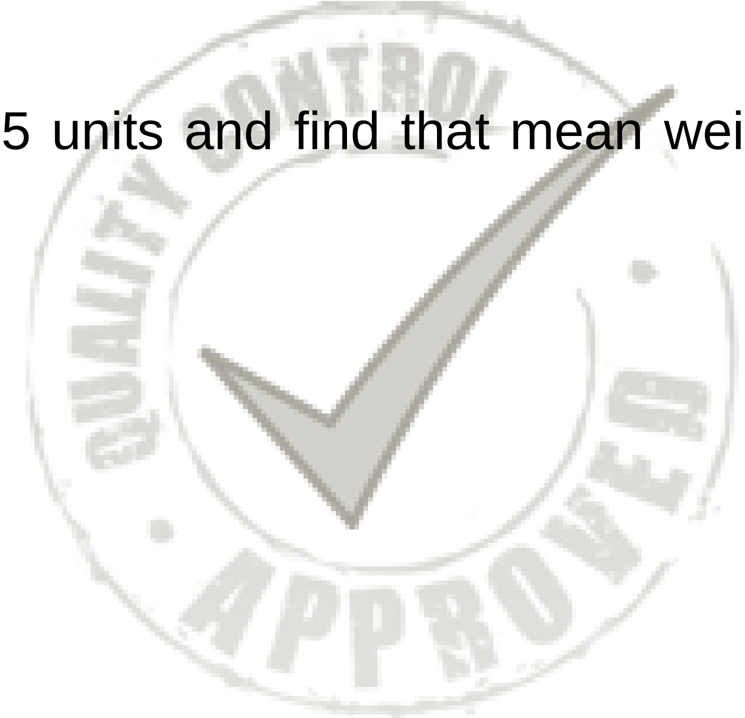
Introduction

Another example: Quality Control

At your manufacturing unit, the mean weight of part A produced is expected to be 2.5 lbs, with a std dev of 0.12 lbs, and is normally distributed.

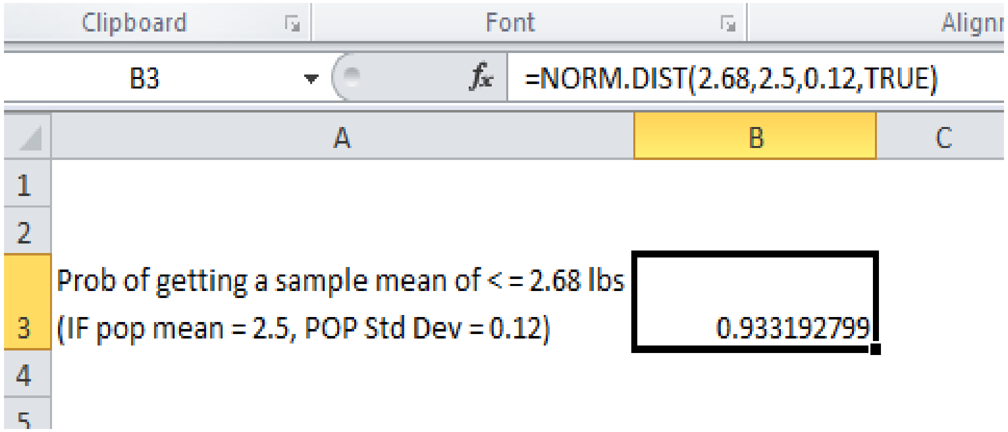
You take a random sample of 45 units and find that mean weight is 2.68 lbs.

Is there an issue?



Introduction

Calculate: Probability of seeing a sample mean of 2.68 if true population mean was 2.5

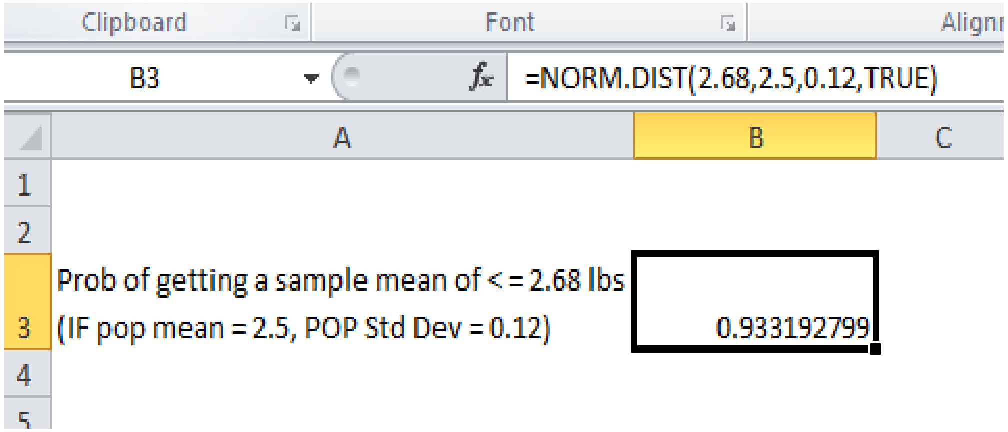


B3 =NORM.DIST(2.68,2.5,0.12,TRUE)

	A	B	C
1			
2			
3	Prob of getting a sample mean of < = 2.68 lbs (IF pop mean = 2.5, POP Std Dev = 0.12)	0.933192799	
4			
5			

Introduction

Calculate: Probability of seeing a sample mean of 2.68 if true population mean was 2.5



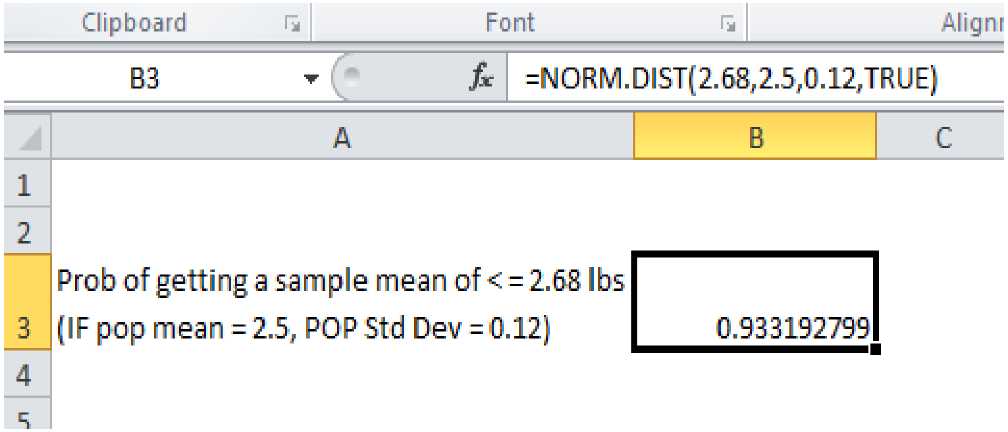
B3 =NORM.DIST(2.68,2.5,0.12,TRUE)

	A	B
3	Prob of getting a sample mean of < = 2.68 lbs (IF pop mean = 2.5, POP Std Dev = 0.12)	0.933192799

i.e - there is a 93% chance that you could get a sample mean of 2.68 or less from a population with mean 2.5

Introduction

Calculate: Probability of seeing a sample mean of 2.68 if true population mean was 2.5



=NORM.DIST(2.68,2.5,0.12,TRUE)

	A	B	C
1			
2			
3	Prob of getting a sample mean of < = 2.68 lbs (IF pop mean = 2.5, POP Std Dev = 0.12)	0.933192799	
4			
5			

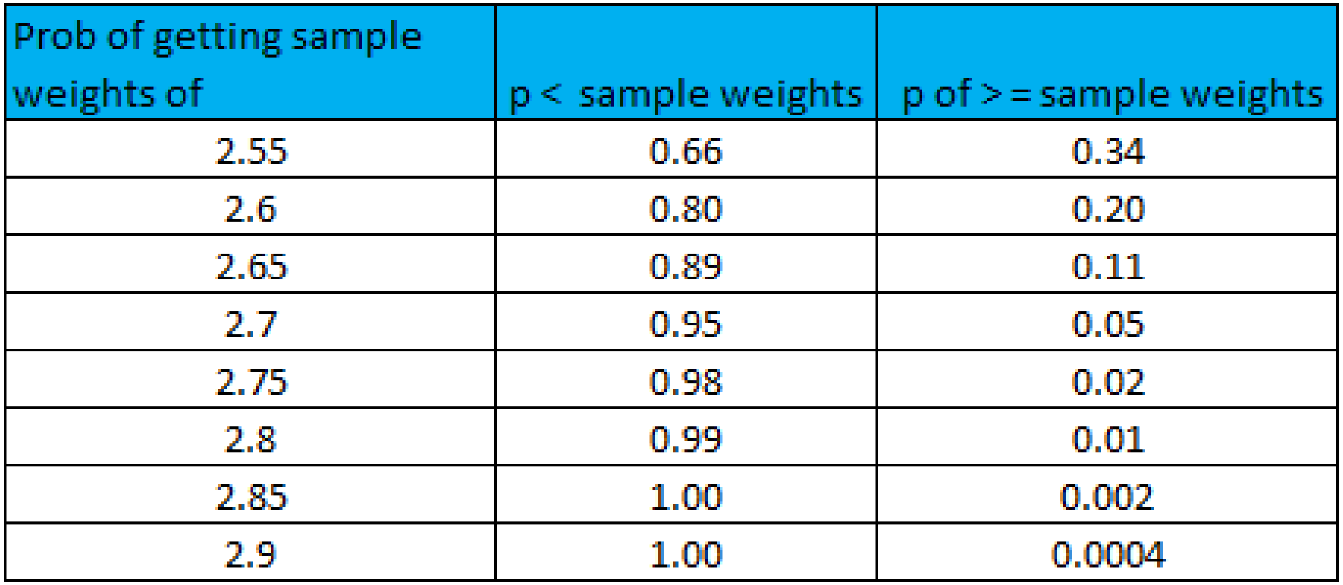
i.e - there is a 93% chance that you could get a sample mean of 2.68 or less from a population with mean 2.5

Prob of getting sample mean of 2.68 or higher = 0.07

Introduction

Conclusion?

There is a 7% chance that simply due to random chance your sample shows a mean weight of 2.68 lbs or greater, even if sample came from a population with a mean of 2.5 lbs



Prob of getting sample weights of	p < sample weights	p of > = sample weights
2.55	0.66	0.34
2.6	0.80	0.20
2.65	0.89	0.11
2.7	0.95	0.05
2.75	0.98	0.02
2.8	0.99	0.01
2.85	1.00	0.002
2.9	1.00	0.0004

Introduction

How would you choose between the two possible explanations?

Outcome	Probability	Conclusion
Random Chance of seeing different sample mean from population	High	Cannot conclude that there is a difference between sample and population
	Low	Conclude that sample is different from population

Introduction

How would you choose between the two possible explanations?

Outcome	Probability	Conclusion
Random Chance of seeing different sample mean from population	High	Cannot conclude that there is a difference between sample and population
	Low	Conclude that sample is different from population

But what would you consider high or low probability?

Introduction

The boundary between high and low tends to be subjective:

Introduction

The boundary between high and low tends to be subjective:

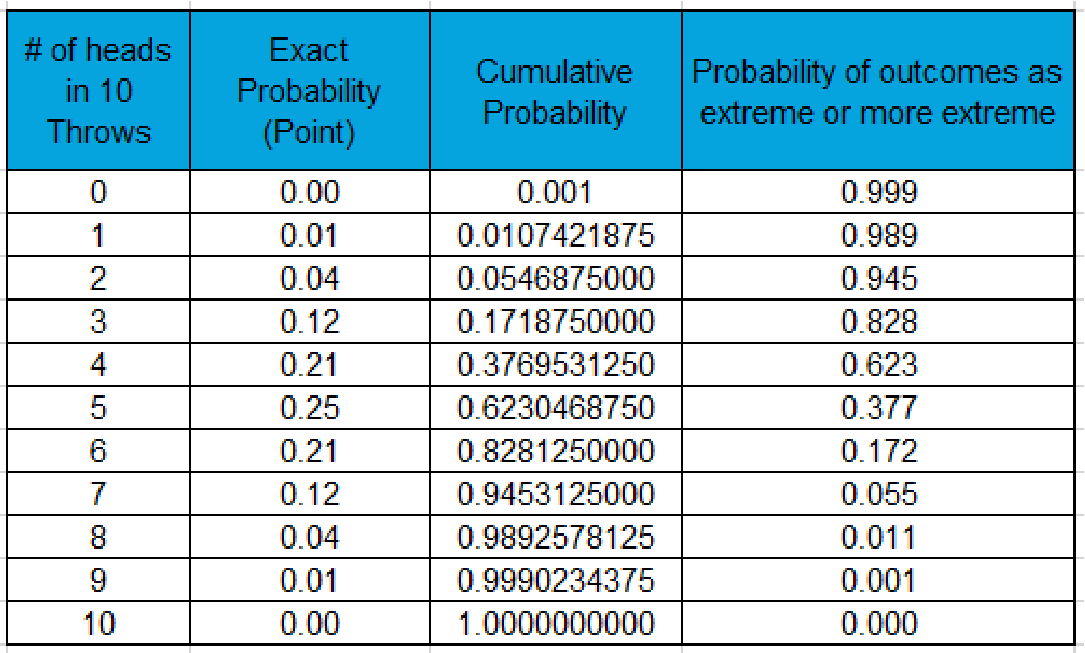
You may decide that 70% is high probability, while another person may decide 90% is high probability. The same applies to low probability

Introduction

The boundary between high and low tends to be subjective:

You may decide that 70% is high probability, while another person may decide 90% is high probability. The same applies to low probability

See table of # of *heads* in 10 flips of a ‘fair’ coin:



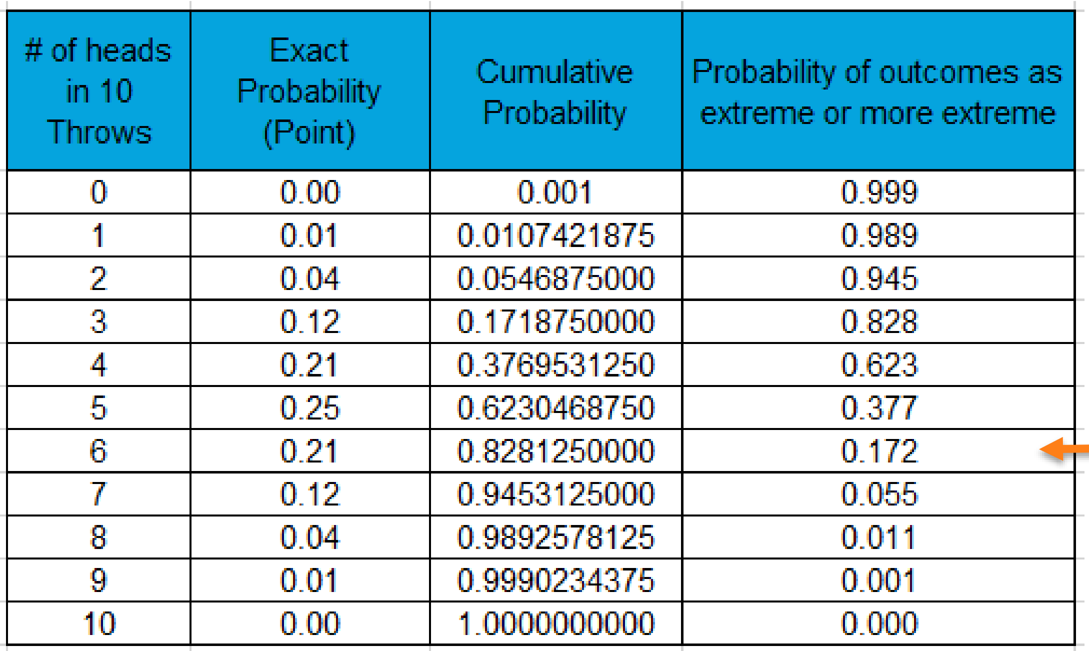
# of heads in 10 Throws	Exact Probability (Point)	Cumulative Probability	Probability of outcomes as extreme or more extreme
0	0.00	0.001	0.999
1	0.01	0.0107421875	0.989
2	0.04	0.0546875000	0.945
3	0.12	0.1718750000	0.828
4	0.21	0.3769531250	0.623
5	0.25	0.6230468750	0.377
6	0.21	0.8281250000	0.172
7	0.12	0.9453125000	0.055
8	0.04	0.9892578125	0.011
9	0.01	0.9990234375	0.001
10	0.00	1.0000000000	0.000

Introduction

The boundary between high and low tends to be subjective:

You may decide that 70% is high probability, while another person may decide 90% is high probability. The same applies to low probability

See table of # of *heads* in 10 flips of a ‘fair’ coin:



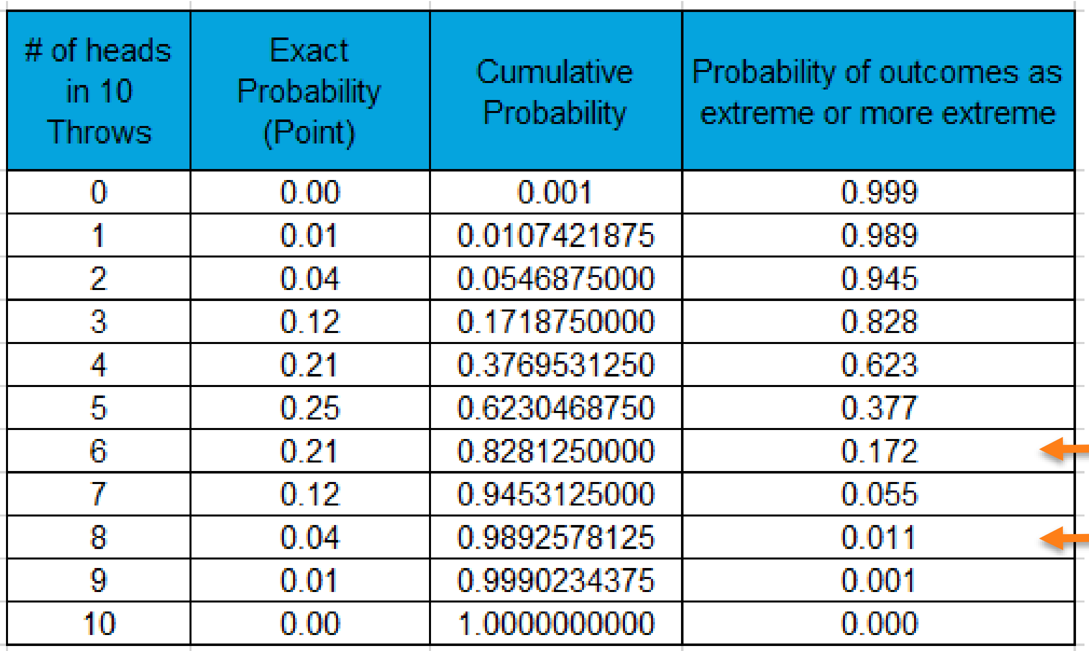
# of heads in 10 Throws	Exact Probability (Point)	Cumulative Probability	Probability of outcomes as extreme or more extreme
0	0.00	0.001	0.999
1	0.01	0.0107421875	0.989
2	0.04	0.0546875000	0.945
3	0.12	0.1718750000	0.828
4	0.21	0.3769531250	0.623
5	0.25	0.6230468750	0.377
6	0.21	0.8281250000	0.172
7	0.12	0.9453125000	0.055
8	0.04	0.9892578125	0.011
9	0.01	0.9990234375	0.001
10	0.00	1.0000000000	0.000

Introduction

The boundary between high and low tends to be subjective:

You may decide that 70% is high probability, while another person may decide 90% is high probability. The same applies to low probability

See table of # of *heads* in 10 flips of a ‘fair’ coin:



# of heads in 10 Throws	Exact Probability (Point)	Cumulative Probability	Probability of outcomes as extreme or more extreme
0	0.00	0.001	0.999
1	0.01	0.0107421875	0.989
2	0.04	0.0546875000	0.945
3	0.12	0.1718750000	0.828
4	0.21	0.3769531250	0.623
5	0.25	0.6230468750	0.377
6	0.21	0.8281250000	0.172
7	0.12	0.9453125000	0.055
8	0.04	0.9892578125	0.011
9	0.01	0.9990234375	0.001
10	0.00	1.0000000000	0.000

Introduction

To avoid subjectivity, both in academic studies and business applications, a cut-off of 5% for low probability is commonly used

Introduction

To avoid subjectivity, both in academic studies and business applications, a cut-off of 5% for low probability is commonly used

What does 5% imply?

Introduction

To avoid subjectivity, both in academic studies and business applications, a cut-off of 5% for low probability is commonly used

What does 5% imply?

Only if random chance probability of seeing sample means as extreme or more extreme than is observed is < 5%, will you conclude that sample is really different from the population

Is this a strong criteria or relaxed criteria?

Introduction

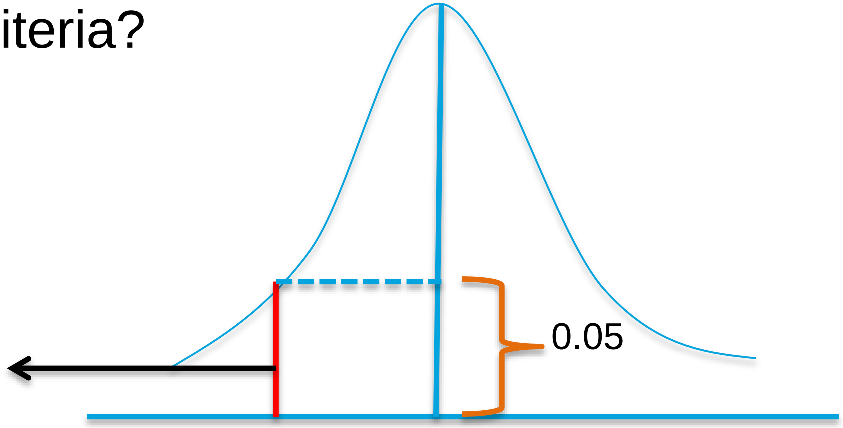
To avoid subjectivity, both in academic studies and business applications, a cut-off of 5% for low probability is commonly used

What does 5% imply?

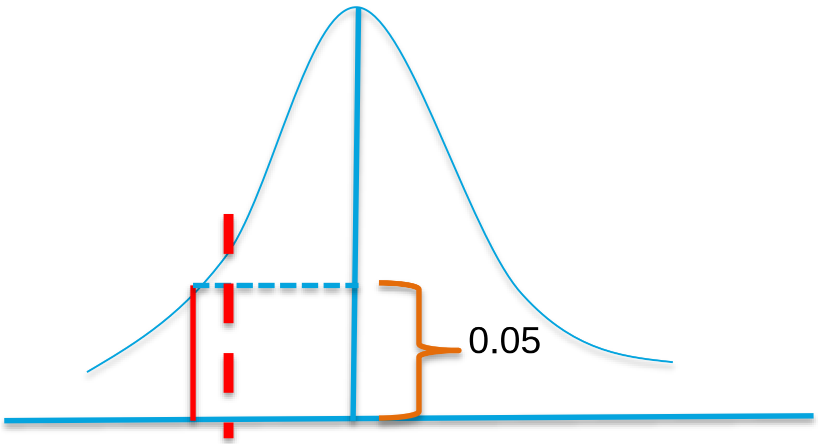
Only if random chance probability of seeing sample means as extreme or more extreme than is observed is < 5%, will you conclude that sample is really different from the population

Is this a strong criteria or relaxed criteria?

Conclude sample is different from population



Introduction



In the quality control example, we got a p-value of 0.07

What would that imply?

Recap

- Introduction to Hypothesis Tests

THANK YOU