REVIEW OF DATA

ST101 – DR. ARIC LABARR

WHAT IS/ARE DATA?

data

noun

\ˈdā - tə\

factual **information** used as a basis for reasoning, discussion, or calculation

- Information – measurements or values describing an object, person, place, thing, etc.
- Inference – using information to come to some conclusion.
- Want to use the information to draw conclusions and make better decisions in the context of our problem.
- Who, what, where, when, why, how?

DATA TABLE

Date	Weekday	Season	Weather Type	Temperature (°F)	Humidity (%)	# Casual Users	# Registered Users
1/1/2011	Saturday	Winter	Misty	46.7	80.6	331	654
1/2/2011	Sunday	Winter	Misty	48.4	69.6	131	670
1/3/2011	Monday	Winter	Clear / Partly Cloudy	34.2	43.7	120	1229
1/4/2011	Tuesday	Winter	Clear / Partly Cloudy	34.5	59.0	108	1454
1/5/2011	Wednesday	Winter	Clear / Partly Cloudy	36.8	43.7	82	1518

Observations

⋮

DATA TABLE

Variables

Date	Weekday	Season	Weather Type	Temperature (°F)	Humidity (%)	# Casual Users	# Registered Users
1/1/2011	Saturday	Winter	Misty	46.7	80.6	331	654
1/2/2011	Sunday	Winter	Misty	48.4	69.6	131	670
1/3/2011	Monday	Winter	Clear / Partly Cloudy	34.2	43.7	120	1229
1/4/2011	Tuesday	Winter	Clear / Partly Cloudy	34.5	59.0	108	1454
1/5/2011	Wednesday	Winter	Clear / Partly Cloudy	36.8	43.7	82	1518

⋮

QUALITATIVE (CATEGORICAL) VARIABLES

Date	Weekday	Season	Weather Type	Temperature (°F)	Humidity (%)	# Casual Users	# Registered Users
1/1/2011	Saturday	Winter	Misty	46.7	80.6	331	654
1/2/2011	Sunday	Winter	Misty	48.4	69.6	131	670
1/3/2011	Monday	Winter	Clear / Partly Cloudy	34.2	43.7	120	1229
1/4/2011	Tuesday	Winter	Clear / Partly Cloudy	34.5	59.0	108	1454
1/5/2011	Wednesday	Winter	Clear / Partly Cloudy	36.8	43.7	82	1518

⋮

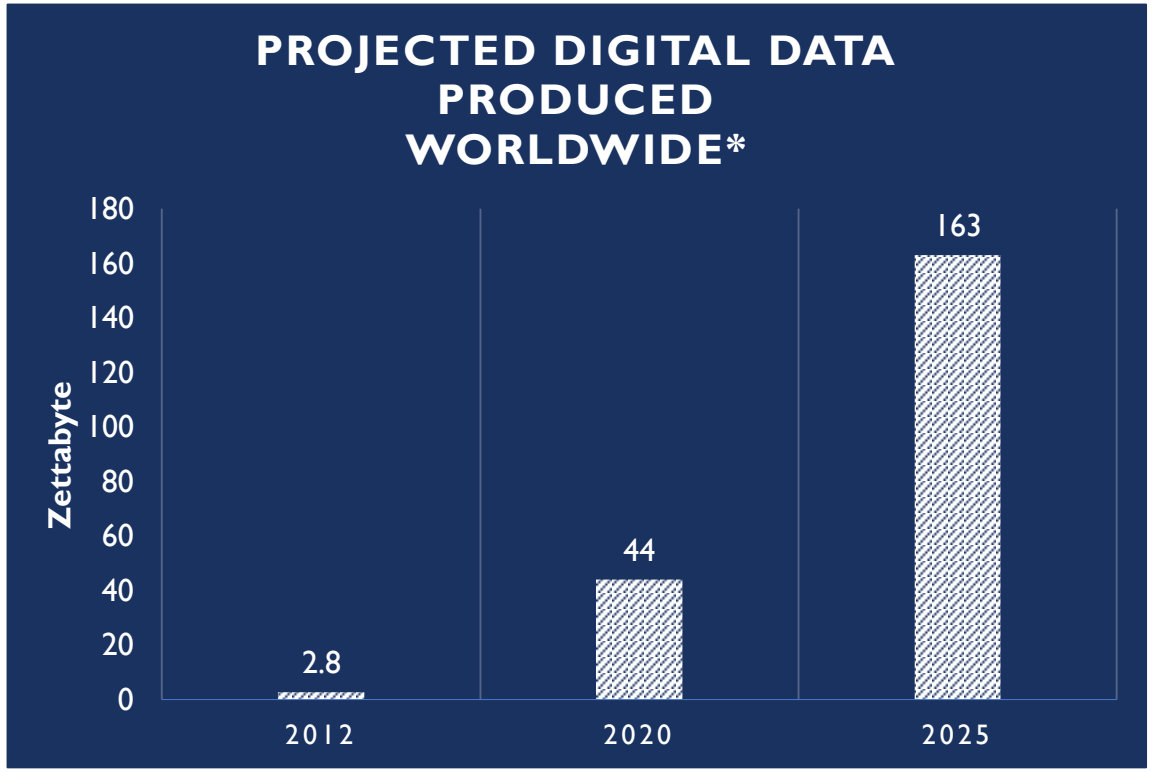
QUANTITATIVE (NUMERICAL) VARIABLES

Date	Weekday	Season	Weather Type	Temperature (°F)	Humidity (%)	# Casual Users	# Registered Users
1/1/2011	Saturday	Winter	Misty	46.7	80.6	331	654
1/2/2011	Sunday	Winter	Misty	48.4	69.6	131	670
1/3/2011	Monday	Winter	Clear / Partly Cloudy	34.2	43.7	120	1229
1/4/2011	Tuesday	Winter	Clear / Partly Cloudy	34.5	59.0	108	1454
1/5/2011	Wednesday	Winter	Clear / Partly Cloudy	36.8	43.7	82	1518

⋮

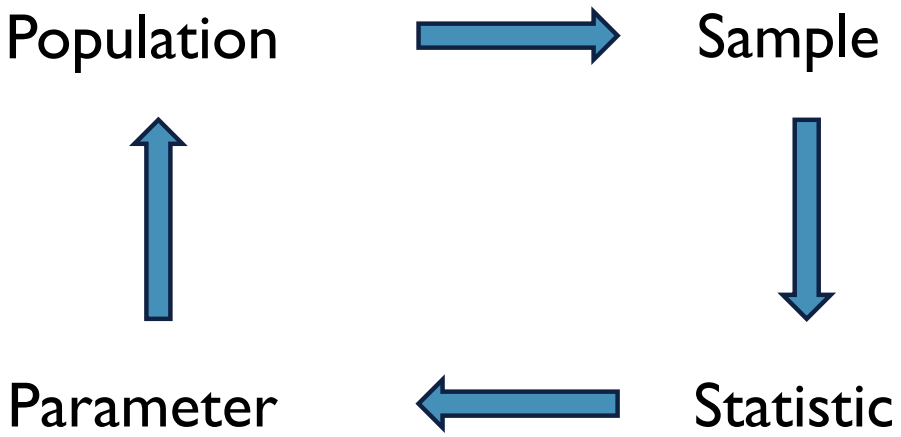
GATHERING DATA

- Data is everywhere.
- 1 zettabyte = 1,000,000,000,000 GB
- With all this data being gathered and stored, we need to understand good practices of gathering data.
- Data gathered without thinking ahead of time leaves itself open for problems later.



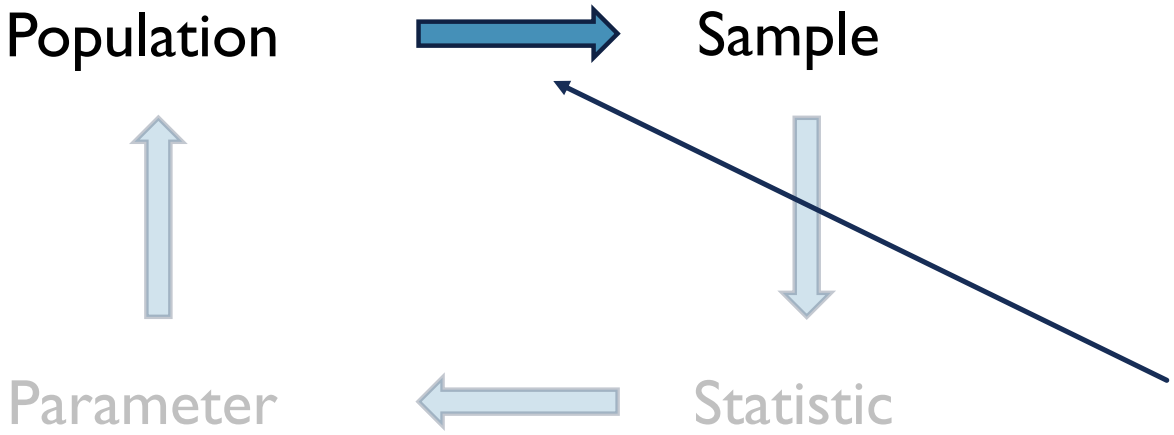
*IDC *Digital Universe*

GATHERING DATA



- Population – set of all objects/individuals of interest.
- Sample – subset of the population that information is actually obtained.
- Statistic – measures computed from a sample.
- Parameter – measures computed from a population.

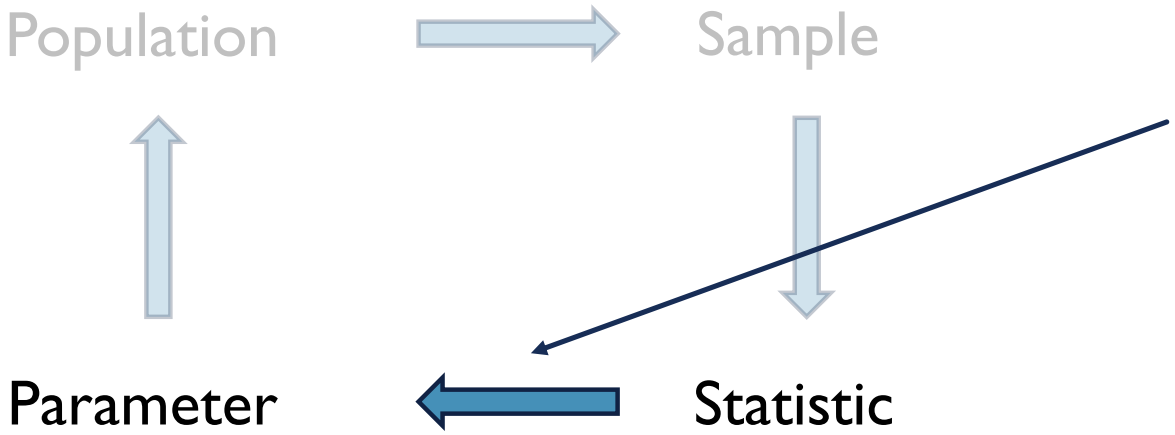
RANDOMNESS AND SAMPLING



Having randomness helps make the sample representative of the population.

Protects us from having certain pieces of information overly influence our sample.

RANDOMNESS AND SAMPLING



Having a good representative sample means the inference we make from the statistic to the parameter is reasonable!

BAD SAMPLING METHODS LEAD TO BIAS

- Need good sampling to have good estimates.
- Bias – certain outcomes are favored over other outcomes in samples.
- 2 Common Types of Bias:
  1. Selection Bias
     a) Undercoverage
     b) Nonresponse
  2. Sampling Bias
     a) Convenience sampling
     b) Voluntary sampling

GOOD SAMPLING METHODS

- Need good sampling to have good estimates.
- 4 Common Techniques:
  1. Simple Random Sampling (SRS)
  2. Stratified Random Sampling (STS)
  3. Cluster Sampling
  4. Systematic Sampling

ETHICAL CONSIDERATIONS AROUND DATA

- The gathering of data leads to questions around the ethical collection and use of that data.
- As Christians we are held to an even higher standard around ethical considerations.
- In observational studies / experiments we must keep the interest of the subject we are collecting data from at the forefront.
- Collection of data:
  - Institutional review boards
  - Informed consent
  - Confidentiality

EXPLORING DIFFERENT TYPES OF VARIABLES

Date	Weekday	Season	Weather Type	Temperature (°F)	Humidity (%)	# Casual Users	# Registered Users
1/1/2011	Saturday	Winter	Misty	46.7	80.6	331	654
1/2/2011	Sunday	Winter	Misty	48.4	69.6	131	670
1/3/2011	Monday	Winter	Clear / Partly Cloudy	34.2	43.7	120	1229
1/4/2011	Tuesday	Winter	Clear / Partly Cloudy	34.5	59.0	108	1454
1/5/2011	Wednesday	Winter	Clear / Partly Cloudy	36.8	43.7	82	1518

Qualitative – explore within a category or across categories.

EXPLORING DIFFERENT TYPES OF VARIABLES

Date	Weekday	Season	Weather Type	Temperature (°F)	Humidity (%)	# Casual Users	# Registered Users
1/1/2011	Saturday	Winter	Misty	46.7	80.6	331	654
1/2/2011	Sunday	Winter	Misty	48.4	69.6	131	670
1/3/2011	Monday	Winter	Clear / Partly Cloudy	34.2	43.7	120	1229
1/4/2011	Tuesday	Winter	Clear / Partly Cloudy	34.5	59.0	108	1454
1/5/2011	Wednesday	Winter	Clear / Partly Cloudy	36.8	43.7	82	1518
⋮							

Quantitative – explore center, spread, and "look" of variables.

EXPLORING VARIABLES VISUALLY

Qualitative Variable Plots

- Pie chart – graph in which a circle is divided into sections that each represent a proportion of the whole.
- Bar chart – numerical values of variables are represented by the height or length of lines or rectangles of equal width.

Quantitative Variable Plots

- Line graph – uses lines to connect individual data points over time.
- Scatterplot – the values of two variables are plotted along two axes, the pattern of the resulting points revealing any relationship present.

EXPLORING VARIABLES NUMERICALLY

Measures of Center or "Typical"

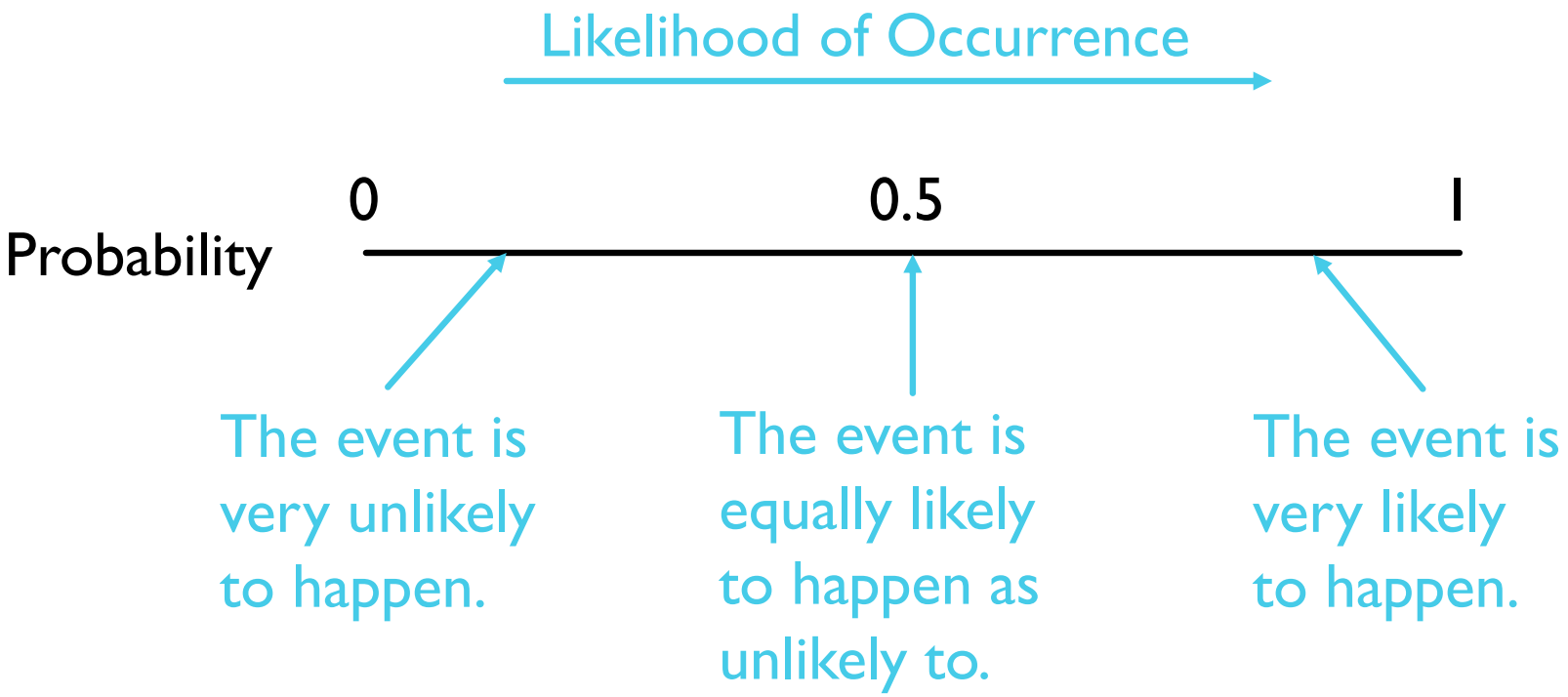
- Mode – the mode of a variable is the most common value.
- Mean – the mean of a variable is the sum of all the values divided by the number of values.
- Median – value in the middle when the data items are arranged in ascending order.

Measures of Spread or Variation

- Range – difference between the largest and smallest values.
- Variance – measure of dispersion around the mean of the data set.
- Standard deviation – the square root of the variance (helps with units of variance).

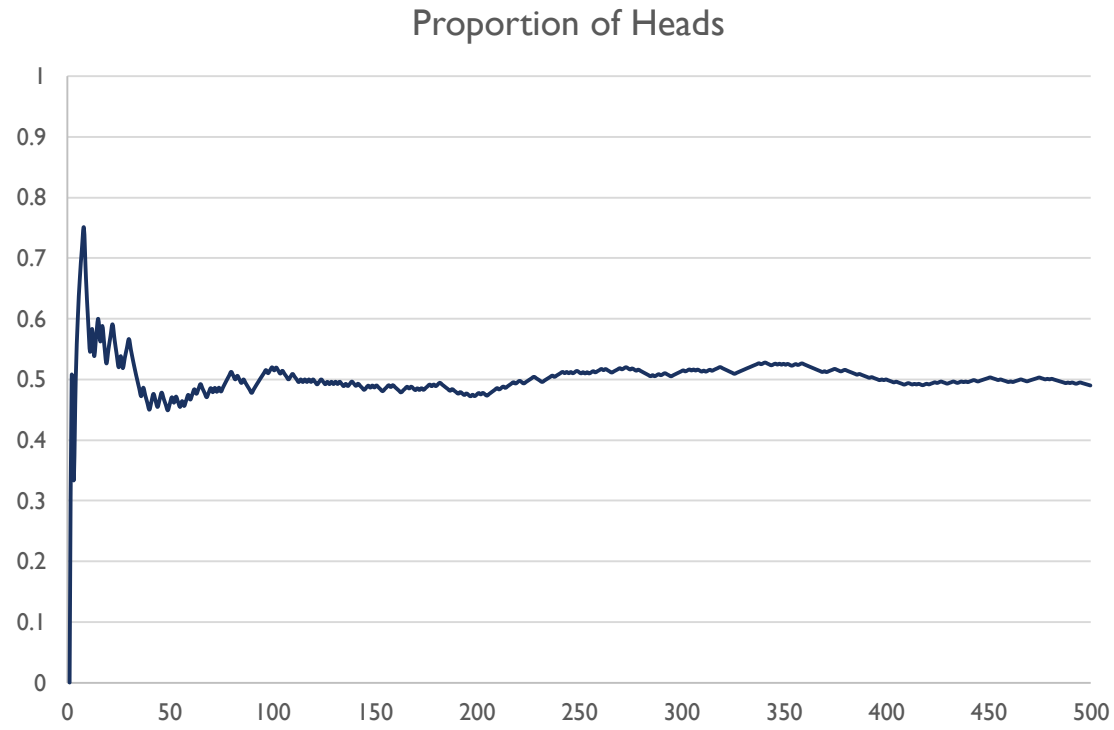
PROBABILITY

- The **probability** that an event happens is a numerical measure of the likelihood of that event's occurrence.



LAW OF LARGE NUMBERS

- The **law of large numbers** states that as the number of independent trials increases, *in the long run* the proportion for a certain event gets closer and closer to a single value (the probability of the event).



PROBABILITY DISTRIBUTION

- The **probability distribution** for a random variable describes how probabilities are distributed over the values of the random variable.
- Relative frequencies can be used as estimates to the probability of an event occurring.
- Probability distributions for discrete random variables are best described with tables, graphs, or equations.

DISCRETE PROBABILITY EXAMPLE

- Let x be the number of TV's sold at a small department store in one day where x can only take the values of {0, 1, 2, 3, 4, 5}
- We expect to sell 1.88 TV's per day on average with variance of 2.522.

TV's Sold	Number of Days (Freq)	$P(X = x)$	$x_i \times P(X = x_i)$	$(x_i - \mu)^2 P(X = x_i)$
0	90	0.25	0.00	0.883
1	85	0.23	0.23	0.177
2	70	0.19	0.38	0.002
3	45	0.12	0.36	0.150
4	50	0.14	0.56	0.629
5	25	0.07	0.35	0.681
	365	**1.00**	**1.88**	**2.522**

BINOMIAL DISTRIBUTION

- The **binomial distribution** looks at the probabilities of the number of successes occurring in the *n* independent trials.
- The binomial probability function is comprised of two intuitive pieces:

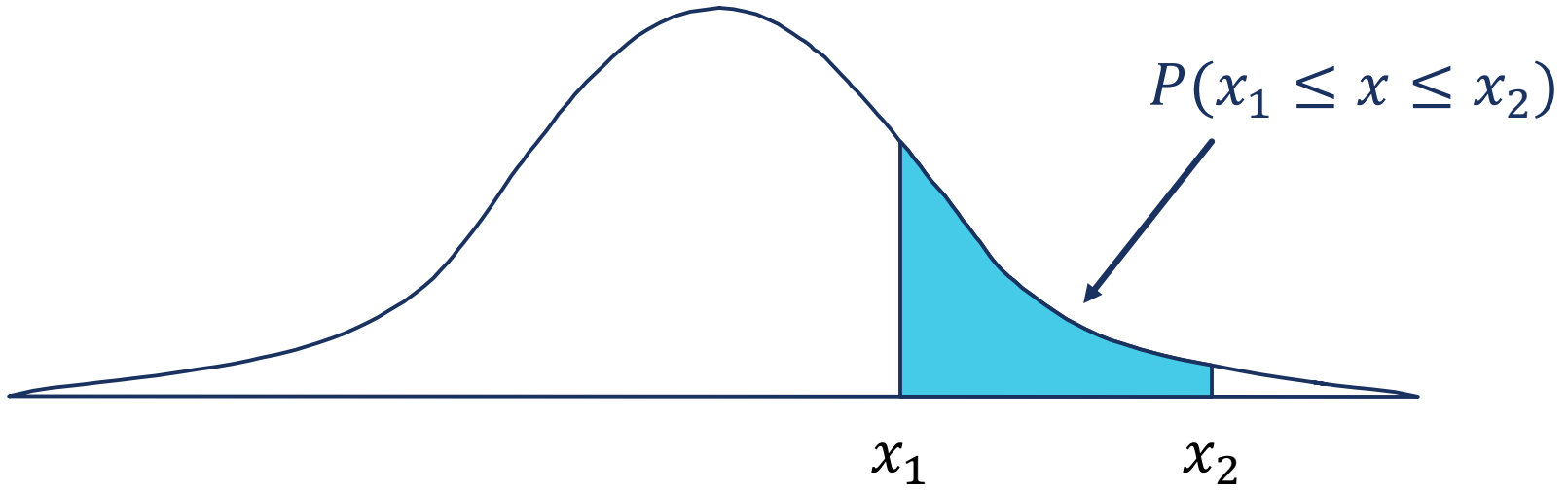
$$f(x) = \frac{n!}{x!\,(n-x)!} p^x (1-p)^{n-x}$$

Number of outcomes providing exactly x successes in *n* trials

Probability of a particular sequence of trial outcomes with x successes in *n* trials

PROBABILITIES ON INTERVALS

- A **continuous random variable** can assume any value in an interval on the real line or in a collection of intervals on the real line.
- The probability of the random variable assuming a value inside of a given interval from x_1 to x_2 is the **area under the graph** of the **probability density function** between x_1 and x_2.



UNIFORM PROBABILITY DISTRIBUTION

- A random variable follows a **uniform distribution** whenever the probability is proportional to the interval's length.
- In other words, every value has an equal probability of happening.
- The **probability density function** for the uniform distribution is:

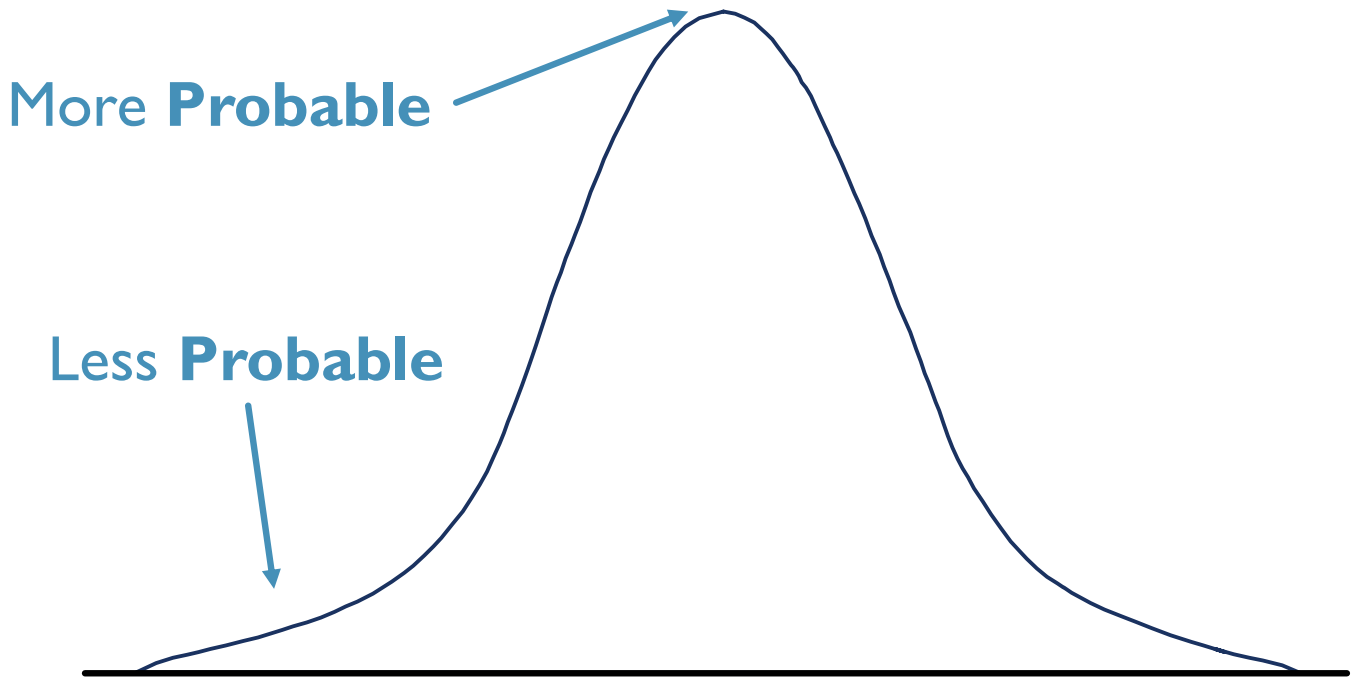
$$f(x) = \begin{cases} \dfrac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{elsewhere} \end{cases}$$

NORMAL PROBABILITY DISTRIBUTION

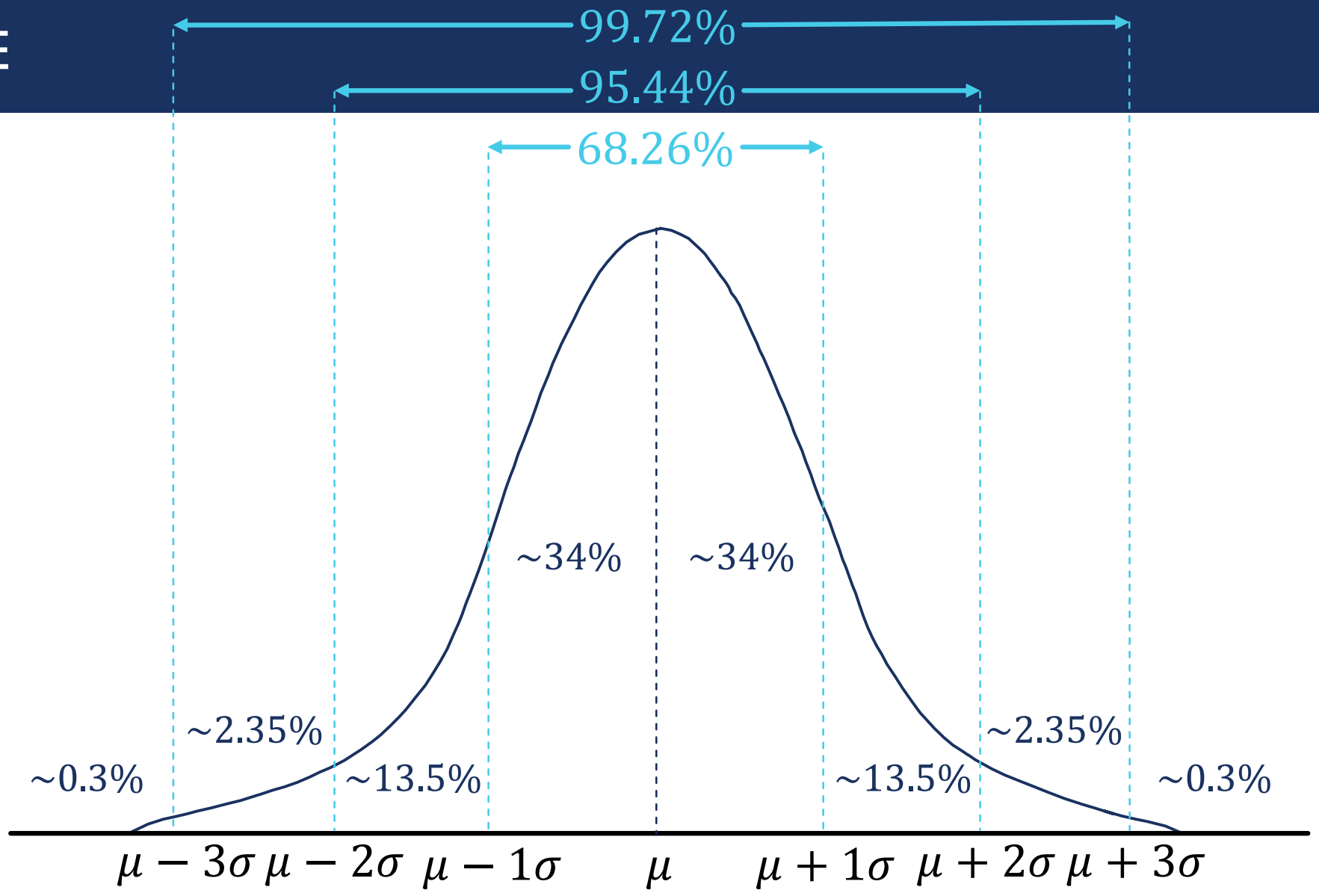
- The **Normal probability distribution** is one of the most common and important distributions for describing a continuous random variable.
- The Normal distribution is the foundation of statistical inference:
  - Hypothesis Testing
  - Confidence Intervals
  - Regression Analysis
- Appears in nature and real-world data.

CHARACTERISTICS OF NORMAL DISTRIBUTION

- The Normal distribution has some useful characteristics:
  - Perfectly Symmetric (Skewness = 0)
  - Unimodal
  - Mean = Median = Mode
  - Asymptotic to x-axis (Can take any value from $-\infty$ to ∞)
  - Completely Defined by Mean and Standard Deviation

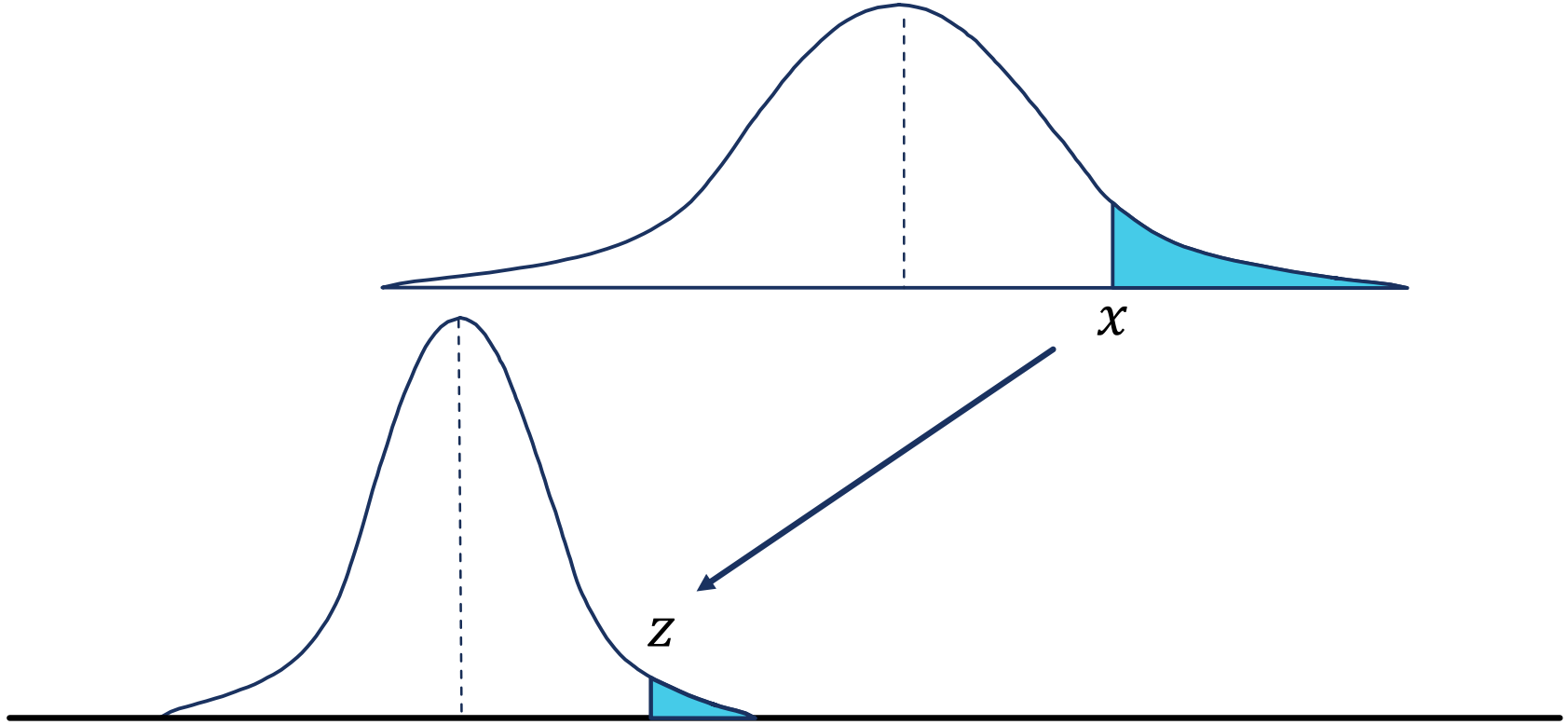


EMPIRICAL RULE



CONVERSION OF NORMAL DISTRIBUTIONS

- All Normal distributions can be converted into standard Normal distributions for ease of computing probabilities under the curve.



SAMPLING ERROR

Population: 1, 3, 5, 5, 7, 9, 4, 6, 10, 2 $\quad \mu = 5.2$

Sample 1: 1, 10, 6, 9 $\quad \bar{x}_1 = 6.5 \quad \bar{x}_1 - \mu = 6.5 - 5.2 = 1.3$

Sample 2: 1, 3, 2, 5 $\quad \bar{x}_2 = 2.75 \quad \bar{x}_2 - \mu = 2.75 - 5.2 = -2.45$

If sample statistics (like the sample mean) had a predictable pattern, then the errors would have a typical pattern as well!

CENTRAL LIMIT THEOREM

- If we use a large sample ($n \geq 50$), the **Central Limit Theorem (CLT)** states that the sampling distribution of $\bar{x}$ is approximately Normally distributed, **regardless of the population distribution**.
- If we use a small sample ($n < 50$), the sampling distribution of $\bar{x}$ is approximately Normally distributed **only if the population distribution is Normal**.

SAMPLING DISTRIBUTION OF $\bar{x}$

- The **sampling distribution of $\bar{x}$** is the probability distribution of all the possible values of the sample mean $\bar{x}$.
- The **sampling distribution of $\bar{x}$** has a mean (expected value) and variance as well.

$$E(\bar{x}) = \mu_{\bar{x}} = \mu$$

$$SD(\bar{x}) = \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

SAMPLING DISTRIBUTION OF $\hat{p}$

- The **sampling distribution of $\hat{p}$** is the probability distribution of all the possible values of the sample proportion $\hat{p}$.
- The **sampling distribution of $\hat{p}$** has a mean (expected value) and variance as well.

$$E(\hat{p}) = \mu_{\hat{p}} = p$$

$$SD(\hat{p}) = \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

MARGIN OF ERROR

- A point estimator cannot be expected to provide the exact value of the population parameter.
- An **interval estimate** can be computed by adding and subtracting a **margin or error** to the point estimate:

$$\text{Point Estimate} \pm \text{Margin of Error}$$

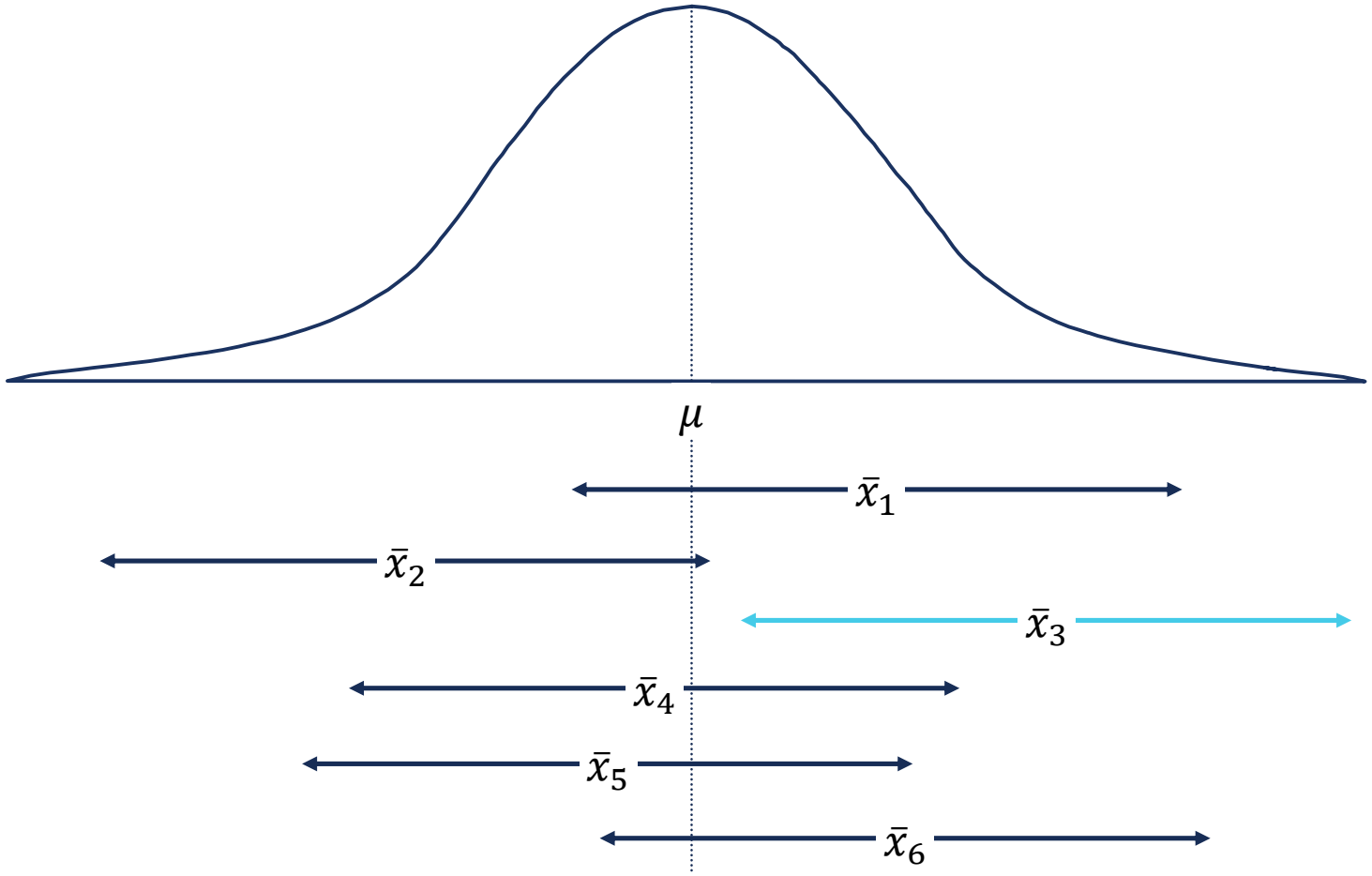
- The purpose of an interval estimate is to provide information about how close the point estimate is to the value of the parameter.

CONFIDENCE INTERVALS

- **Confidence Intervals** are interval estimates where we say we have a certain level of **confidence** in the interval.
- For example, we are **95% confident** that the population average daily number of total users of the bike rental company is between 4,000 and 5,000.

If we were to take many samples (same size) that each produced different confidence intervals, then 95% of them would contain the true parameter.

CONFIDENCE INTERVALS EXAMPLE



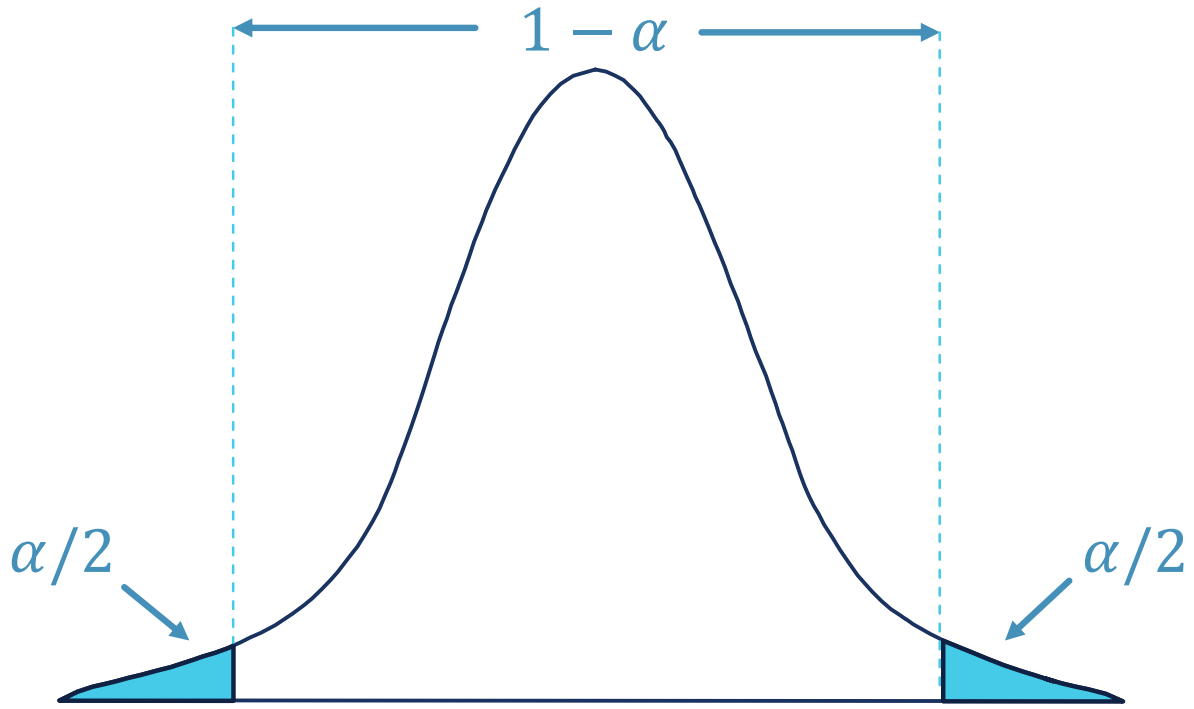
CONFIDENCE INTERVALS FOR MEANS AND PROPORTIONS

- Proportions:

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- Means:

$$\bar{x} \pm t_{\alpha/2} \times \frac{s}{\sqrt{n}}$$



CONFIDENCE INTERVALS FOR MEANS AND PROPORTIONS

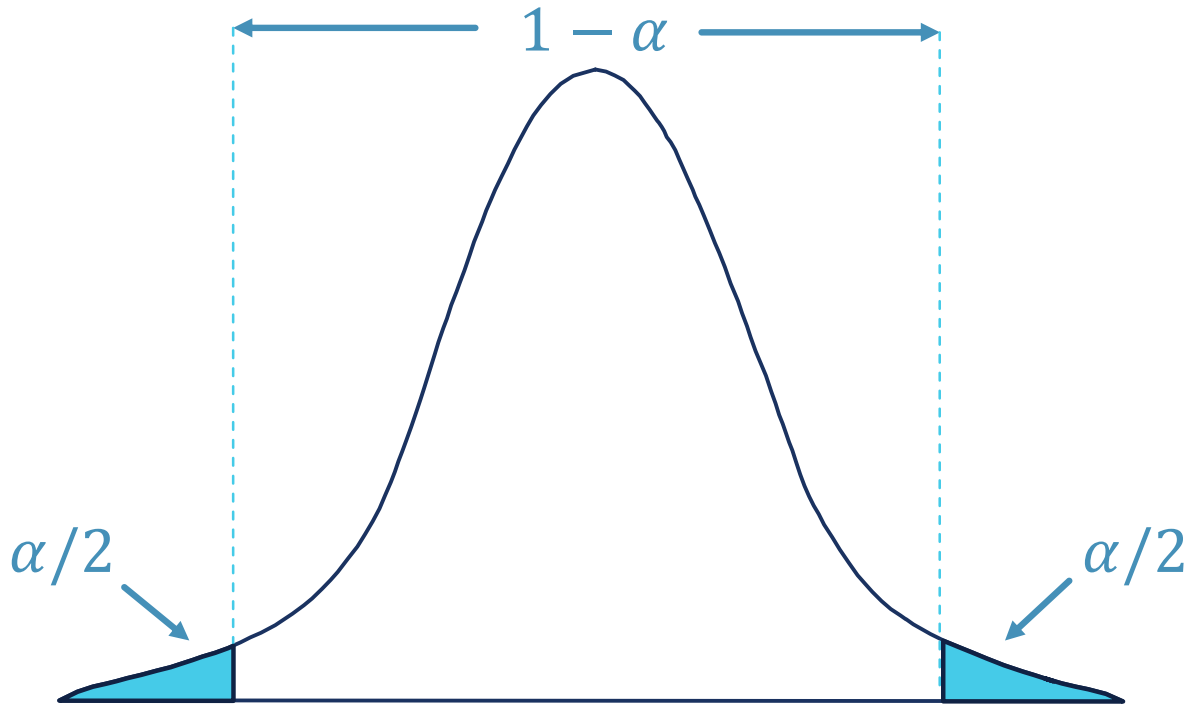
- Proportions:

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- Means:

$$\bar{x} \pm t_{\alpha/2} \times \frac{s}{\sqrt{n}}$$

Need t-distributions since we are now estimating both μ and σ.



HYPOTHESIS TESTING THROUGH EXAMPLE

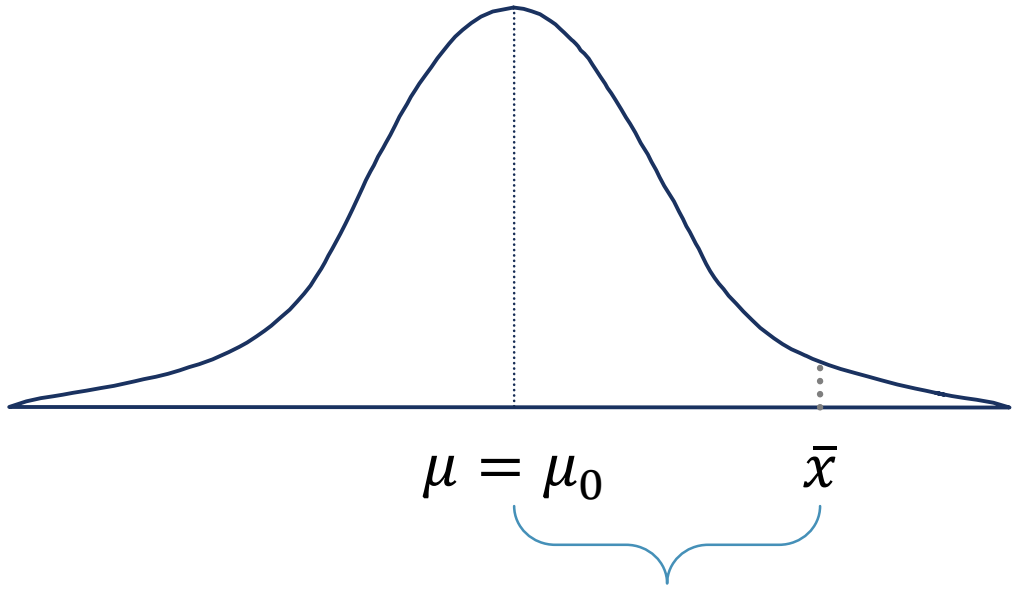
- I have a coin that you believe is fair to start. NULL Hypothesis
- To test if this coin is fair, you ask me to flip the coin repeatedly and record the results. Test Statistic

Flip Number	Result	P-value
1	Heads	0.50
2	Heads	0.25
3	Heads	0.125
4	Heads	0.0625
5	Heads	0.03125

- No longer believe the coin is fair. Decision on NULL Hypothesis

HYPOTHESIS TESTING

- A hypothesis test uses data to help evaluate an initial claim about a parameter from the population.
- There are 4 main steps to hypothesis testing:
  1. State the hypotheses
  2. Test statistic
  3. P-value
  4. Decision on null hypothesis



How likely is this to happen?

TYPE I VS. TYPE II ERRORS

CHOICE	TRUTH: H_0 True	TRUTH: H_0 False
Do Not Reject H_0	Correct	Type II
Reject H_0	Type I	Correct

SUMMARY

- Data is everywhere.
- We can do amazing things with data – explore it, understand it, make inferences from it.
- God gives us glimpses of this world through data.
- Must use the information we find wisely and ethically.

COMPLETE EXAMPLE WITH BIKE DATA

REVIEW OF DATA

OVERVIEW OF PROBLEM

- You have been hired as a data analyst by *IAA BikeSharing Inc.* (the "Customer") and your task is to provide key insights and recommendations about their business.
- The Customer is a bike rental business, operating in Washington DC and Arlington, VA.

MISSION STATEMENT

- "Bike sharing systems are a new generation of traditional bike rentals where the whole process from membership to rental and return has become automatic. Through these systems, a user can easily rent a bike from a particular position and return it back to another position. Currently, there are over 500 bike-sharing programs around the world which are composed of over 500 thousand bicycles. Today, there exists great interest in these systems due to their important role in traffic, environment and health issues…
- Compared to other transport services such as bus or subway, the duration of travel from the departure to the arrival position is explicitly recorded in these systems. This feature turns bike sharing systems in a virtual sensor network that can be used for sensing mobility in the city…"

DATA

- The company has provided detailed rental and environmental data for a two-year period (2011 and 2012).
- The data are based on their Washington DC operations and cover measures such as daily rental counts, precipitation, day of week, season, and other variables which might have a potential impact on rental behavior.

KEY GOALS OF ANALYSIS – PART 1

- Describe the key statistical measures of registered and casual users.
- Identify any extreme observations in the counts of registered and casual users.
  - Do they tend to appear on certain days?
  - Certain seasons?
- Describe any changes in registered users from 2011 to 2012.
  - Do you see any improvements?
  - Are these improvements evident across all months?
  - Are there any months that stand out, in terms of over- or under- performance?

KEY GOALS OF ANALYSIS – PART 2

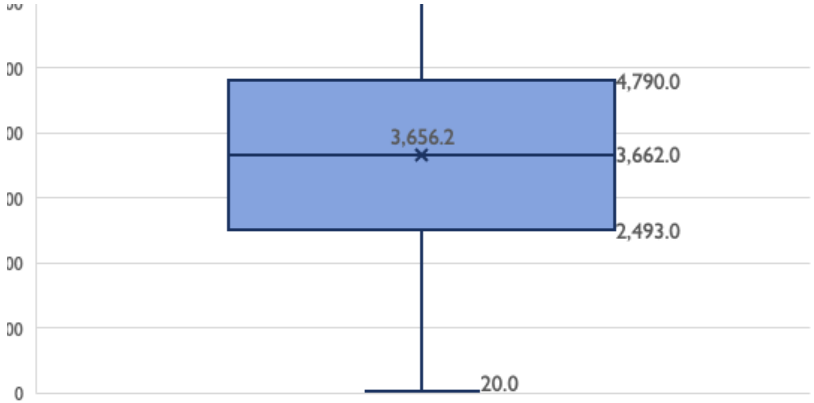
- Identify key probabilities around customers:
  - Probability a customer is a registered user given the season is Fall
  - Probability a customer is a registered user given the season is Summer
  - Provide interval estimates for these probabilities
- The Customer's Marketing Division also has preconceived notions on the average number of total users for the 2012 seasons as follows to help develop their marketing budgets:
  - The average number of total users in the Summer is no less than 6500.
  - The average number of total users in the Fall is no more than 6500.
  - Validate the above claims using the appropriate statistical tests.

KEY GOALS OF ANALYSIS – PART 1

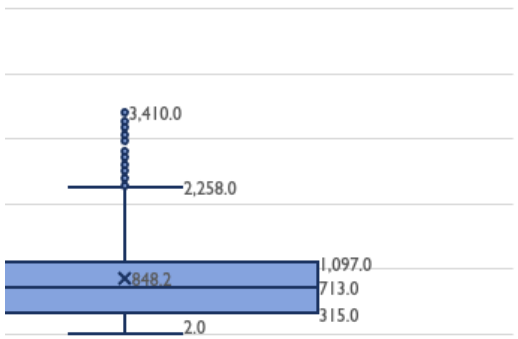
- Describe the key statistical measures of registered and casual users.
- Identify any extreme observations in the counts of registered and casual users.
  - Do they tend to appear on certain days?
  - Certain seasons?
- Describe any changes in registered users from 2011 to 2012.
  - Do you see any improvements?
  - Are these improvements evident across all months?
  - Are there any months that stand out, in terms of over- or under- performance?

REGISTERED VS. CASUAL USERS

Registered Users



Casual Users



MEASURES OF CENTER / TYPICAL

Registered Users

- Mean = 3,656.2 users per day
- Median = 3,662 users per day

Casual Users

- Mean = 848 users per day
- Median = 713 users per day

MEASURES OF VARIABILITY

Registered Users

- Range = 6,946 – 20
  = 6,926 users per day
- Standard Deviation = 686.6 users per day
- IQR = 4,783.3 – 2,488.5
  = 2,294.8 users per day

Casual Users

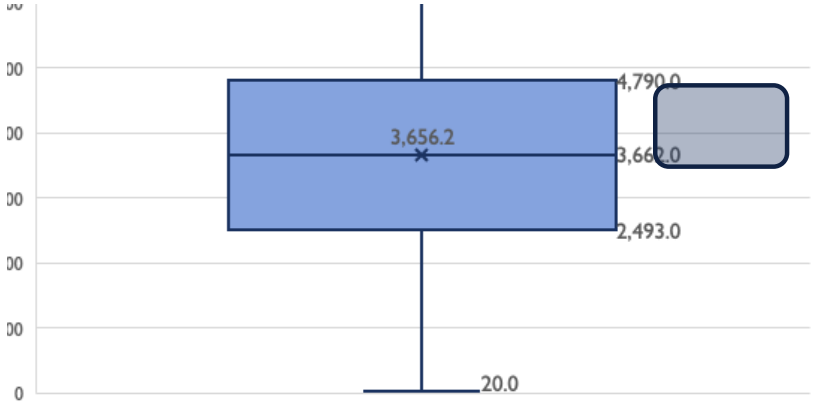
- Range = 3,410 – 2
  = 3,408 users per day
- Standard Deviation = 1,560.3 users per day
- IQR = 1,096.5 – 315.3
  = 781.2 users per day

KEY GOALS OF ANALYSIS – PART 1

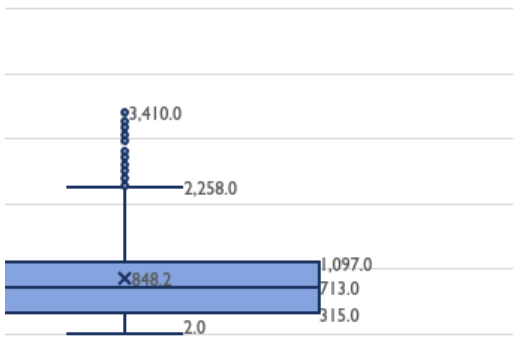
- Describe the key statistical measures of registered and casual users.
- Identify any extreme observations in the counts of registered and casual users.
  - Do they tend to appear on certain days?
  - Certain seasons?
- Describe any changes in registered users from 2011 to 2012.
  - Do you see any improvements?
  - Are these improvements evident across all months?
  - Are there any months that stand out, in terms of over- or under- performance?

REGISTERED VS. CASUAL USERS

Registered Users



Casual Users



5 HIGHEST REGISTERED USER DAYS

Date	Weekday	Season	Weather Type	Temperature (°F)	Humidity (%)	# Registered Users
9/26/2012	Wednesday	Fall	Clear / Partly Cloudy	71.3	63.1	6,946
9/21/2012	Friday	Summer	Clear / Partly Cloudy	68.3	66.9	6,917
10/10/2012	Wednesday	Fall	Clear / Partly Cloudy	61.1	63.1	6,911
10/24/2012	Wednesday	Fall	Clear / Partly Cloudy	67.3	63.6	6,898
10/3/2012	Wednesday	Fall	Misty	73.2	79.4	6,844

5 HIGHEST REGISTERED USER DAYS

Date	Weekday	Season	Weather Type	Temperature (°F)	Humidity (%)	# Registered Users
9/26/2012	Wednesday	Fall	Clear / Partly Cloudy	71.3	63.1	6,946
9/21/2012	Friday	Summer	Clear / Partly Cloudy	68.3	66.9	6,917
10/10/2012	Wednesday	Fall	Clear / Partly Cloudy	61.1	63.1	6,911
10/24/2012	Wednesday	Fall	Clear / Partly Cloudy	67.3	63.6	6,898
10/3/2012	Wednesday	Fall	Misty	73.2	79.4	6,844

5 HIGHEST REGISTERED USER DAYS

Date	Weekday	Season	Weather Type	Temperature (°F)	Humidity (%)	# Registered Users
9/26/2012	Wednesday	Fall	Clear / Partly Cloudy	71.3	63.1	6,946
9/21/2012	Friday	Summer	Clear / Partly Cloudy	68.3	66.9	6,917
10/10/2012	Wednesday	Fall	Clear / Partly Cloudy	61.1	63.1	6,911
10/24/2012	Wednesday	Fall	Clear / Partly Cloudy	67.3	63.6	6,898
10/3/2012	Wednesday	Fall	Misty	73.2	79.4	6,844

5 HIGHEST REGISTERED USER DAYS

Date	Weekday	Season	Weather Type	Temperature (°F)	Humidity (%)	# Registered Users
9/26/2012	Wednesday	Fall	Clear / Partly Cloudy	71.3	63.1	6,946
9/21/2012	Friday	Summer	Clear / Partly Cloudy	68.3	66.9	6,917
10/10/2012	Wednesday	Fall	Clear / Partly Cloudy	61.1	63.1	6,911
10/24/2012	Wednesday	Fall	Clear / Partly Cloudy	67.3	63.6	6,898
10/3/2012	Wednesday	Fall	Misty	73.2	79.4	6,844

5 HIGHEST CASUAL USER DAYS

Date	Weekday	Season	Weather Type	Temperature (°F)	Humidity (%)	# Casual Users
5/19/2012	Saturday	Spring	Clear / Partly Cloudy	68.4	45.6	3,410
5/27/2012	Sunday	Spring	Clear / Partly Cloudy	76.0	69.7	3,283
4/7/2012	Saturday	Spring	Clear / Partly Cloudy	54.6	25.4	3,252
9/15/2012	Saturday	Summer	Clear / Partly Cloudy	69.1	50.1	3,160
3/17/2012	Saturday	Winter	Misty	61.1	75.6	3,155

5 HIGHEST CASUAL USER DAYS

Date	Weekday	Season	Weather Type	Temperature (°F)	Humidity (%)	# Casual Users
5/19/2012	Saturday	Spring	Clear / Partly Cloudy	68.4	45.6	3,410
5/27/2012	Sunday	Spring	Clear / Partly Cloudy	76.0	69.7	3,283
4/7/2012	Saturday	Spring	Clear / Partly Cloudy	54.6	25.4	3,252
9/15/2012	Saturday	Summer	Clear / Partly Cloudy	69.1	50.1	3,160
3/17/2012	Saturday	Winter	Misty	61.1	75.6	3,155

5 HIGHEST CASUAL USER DAYS

Date	Weekday	Season	Weather Type	Temperature (°F)	Humidity (%)	# Casual Users
5/19/2012	Saturday	Spring	Clear / Partly Cloudy	68.4	45.6	3,410
5/27/2012	Sunday	Spring	Clear / Partly Cloudy	76.0	69.7	3,283
4/7/2012	Saturday	Spring	Clear / Partly Cloudy	54.6	25.4	3,252
9/15/2012	Saturday	Summer	Clear / Partly Cloudy	69.1	50.1	3,160
3/17/2012	Saturday	Winter	Misty	61.1	75.6	3,155

5 HIGHEST CASUAL USER DAYS

Date	Weekday	Season	Weather Type	Temperature (°F)	Humidity (%)	# Casual Users
5/19/2012	Saturday	Spring	Clear / Partly Cloudy	68.4	45.6	3,410
5/27/2012	Sunday	Spring	Clear / Partly Cloudy	76.0	69.7	3,283
4/7/2012	Saturday	Spring	Clear / Partly Cloudy	54.6	25.4	3,252
9/15/2012	Saturday	Summer	Clear / Partly Cloudy	69.1	50.1	3,160
3/17/2012	Saturday	Winter	Misty	61.1	75.6	3,155

5 LOWEST REGISTERED USER DAYS

Date	Weekday	Season	Weather Type	Temperature (°F)	Humidity (%)	# Registered Users
10/29/2012	Monday	Fall	Rain or Snow	54.8	88.0	20
1/27/2011	Thursday	Winter	Clear / Partly Cloudy	34.1	68.8	416
12/26/2012	Wednesday	Winter	Rain or Snow	38.2	82.3	432
12/25/2011	Sunday	Winter	Clear / Partly Cloudy	40.8	68.1	451
1/26/2011	Wednesday	Winter	Rain or Snow	36.0	86.3	472

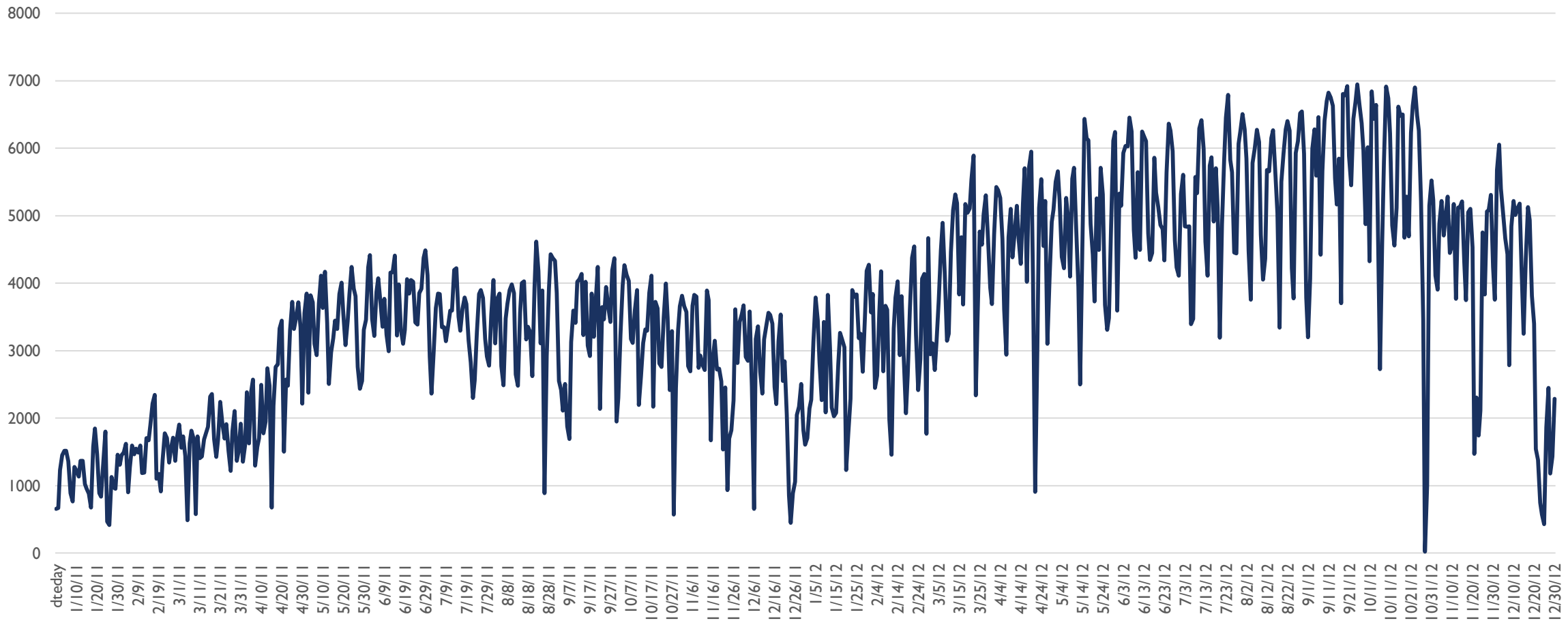
5 LOWEST CASUAL USER DAYS

Date	Weekday	Season	Weather Type	Temperature (°F)	Humidity (%)	# Casual Users
10/29/2012	Monday	Fall	Rain or Snow	54.8	88	2
12/26/2012	Wednesday	Winter	Rain or Snow	38.2	82.3	9
1/18/2011	Tuesday	Winter	Misty	35.9	86.2	9
1/27/2011	Thursday	Winter	Clear / Partly Cloudy	34.1	68.8	15
1/12/2011	Wednesday	Winter	Clear / Partly Cloudy	32.2	60.0	25

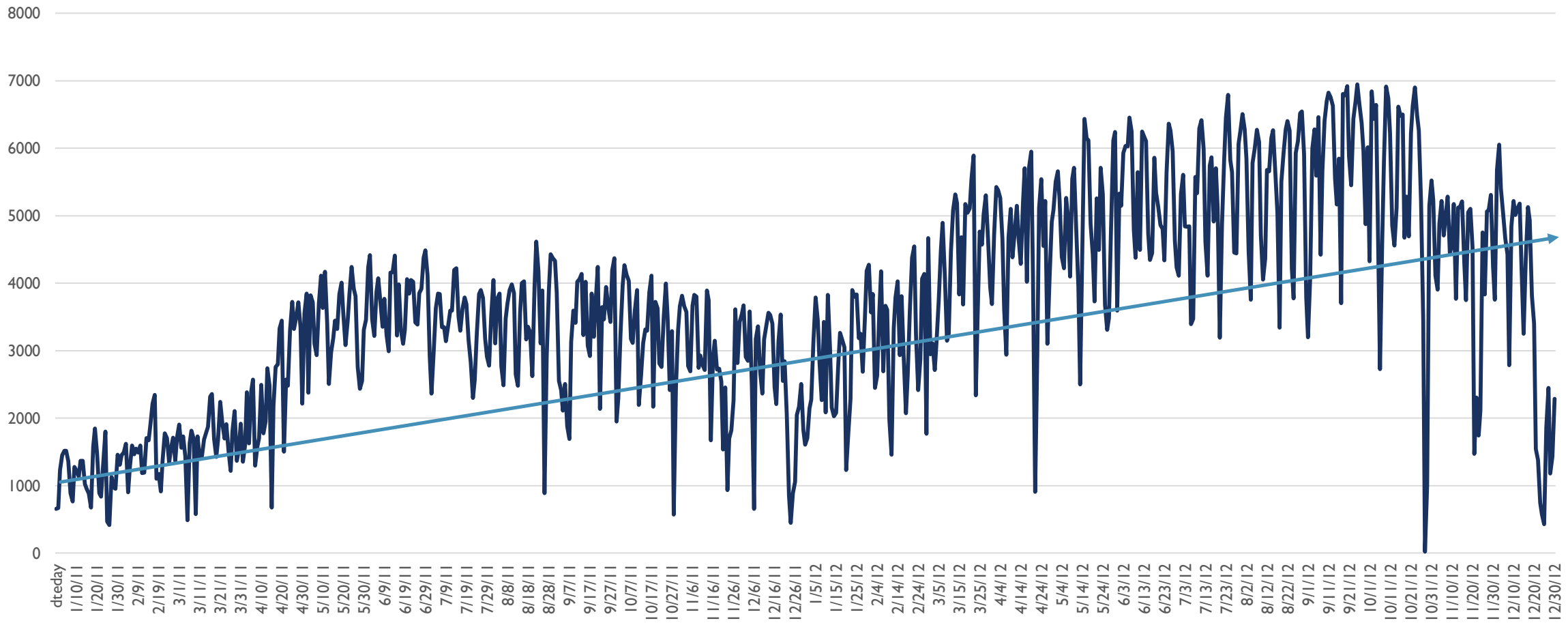
KEY GOALS OF ANALYSIS – PART 1

- Describe the key statistical measures of registered and casual users.
- Identify any extreme observations in the counts of registered and casual users.
  - Do they tend to appear on certain days?
  - Certain seasons?
- Describe any changes in registered users from 2011 to 2012.
  - Do you see any improvements?
  - Are these improvements evident across all months?
  - Are there any months that stand out, in terms of over- or under- performance?

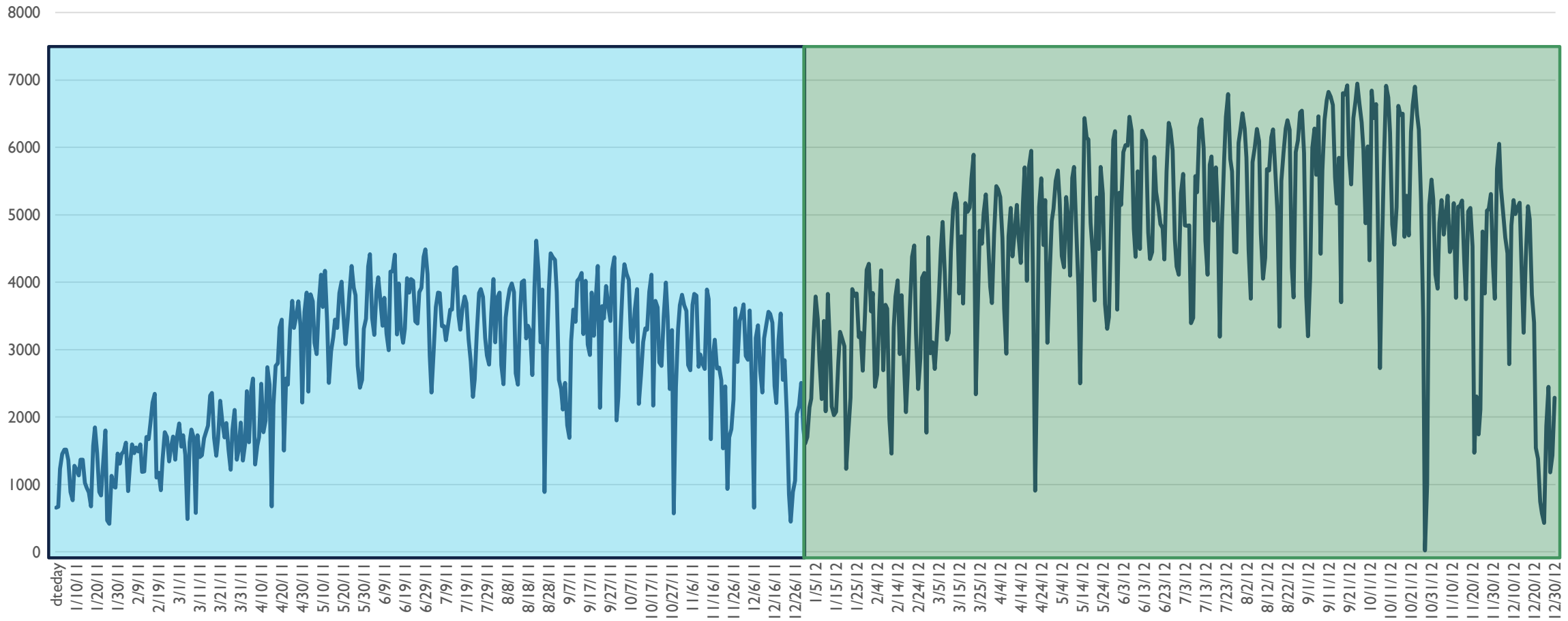
REGISTERED USERS OVER TIME



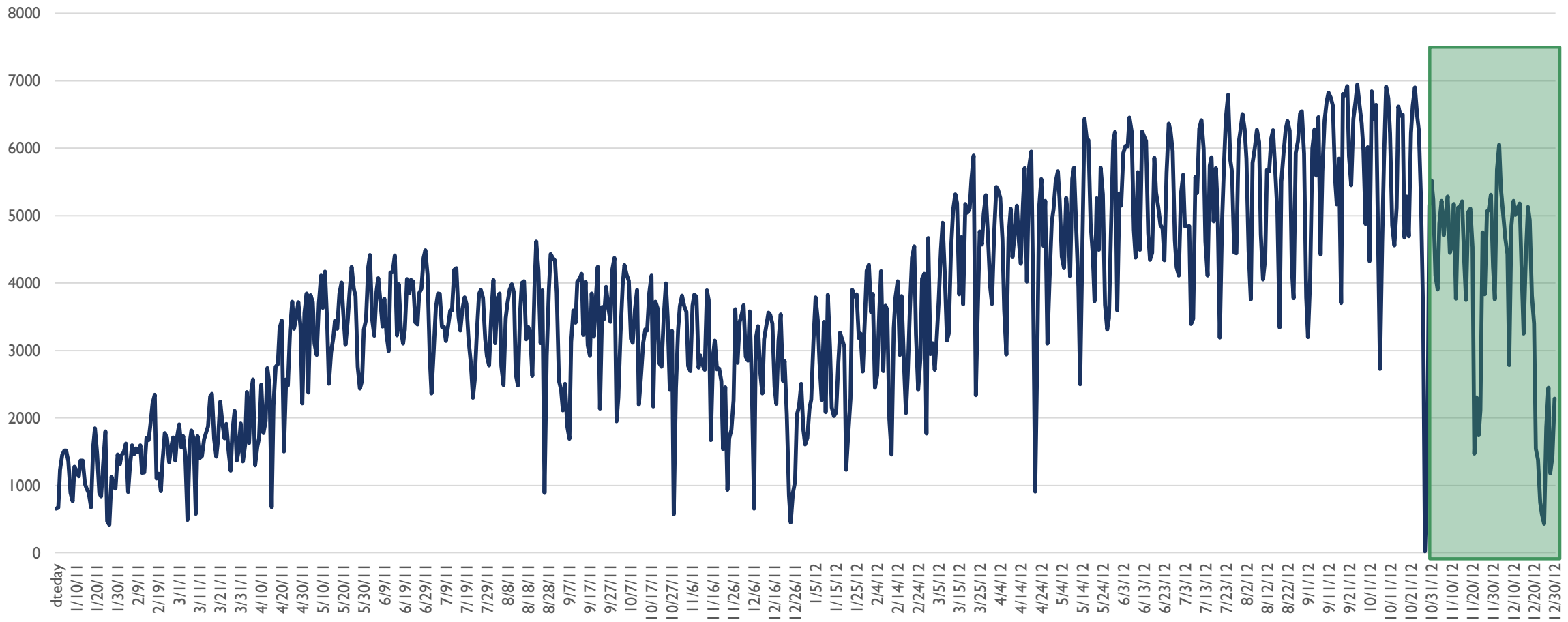
REGISTERED USERS OVER TIME



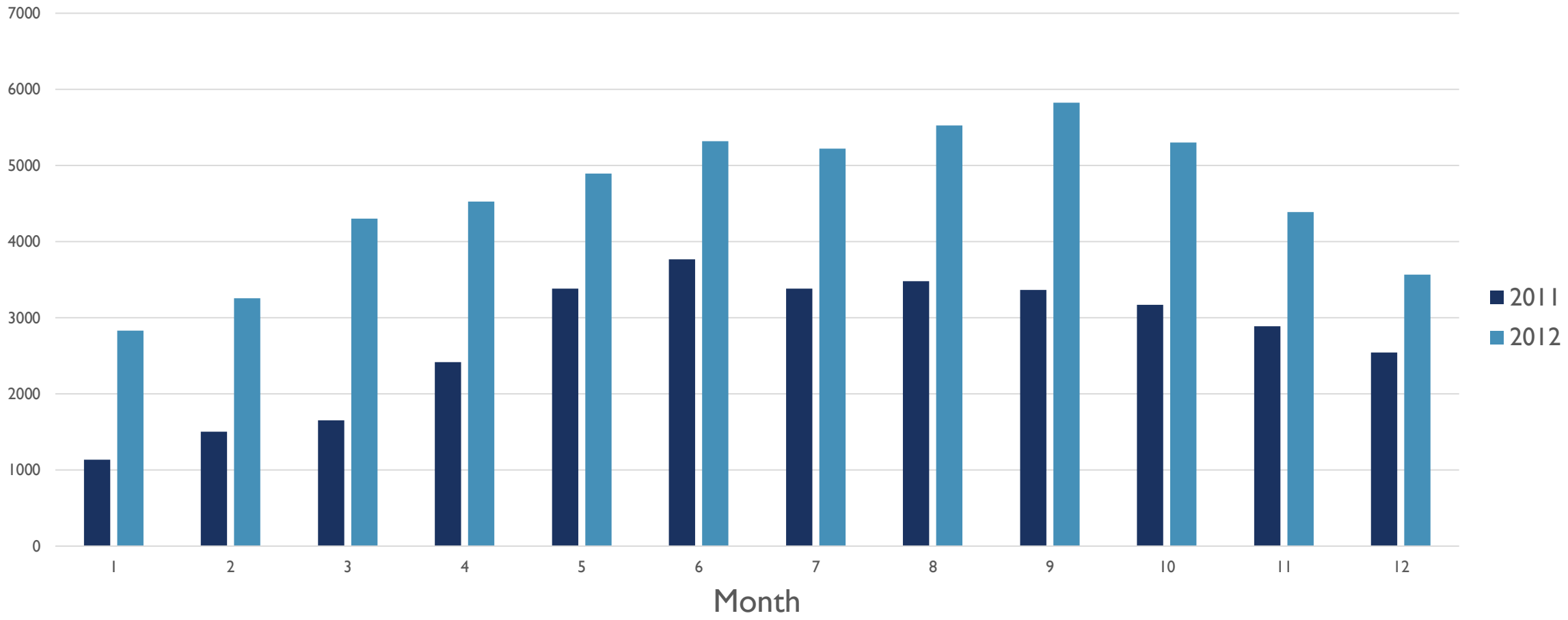
REGISTERED USERS OVER TIME



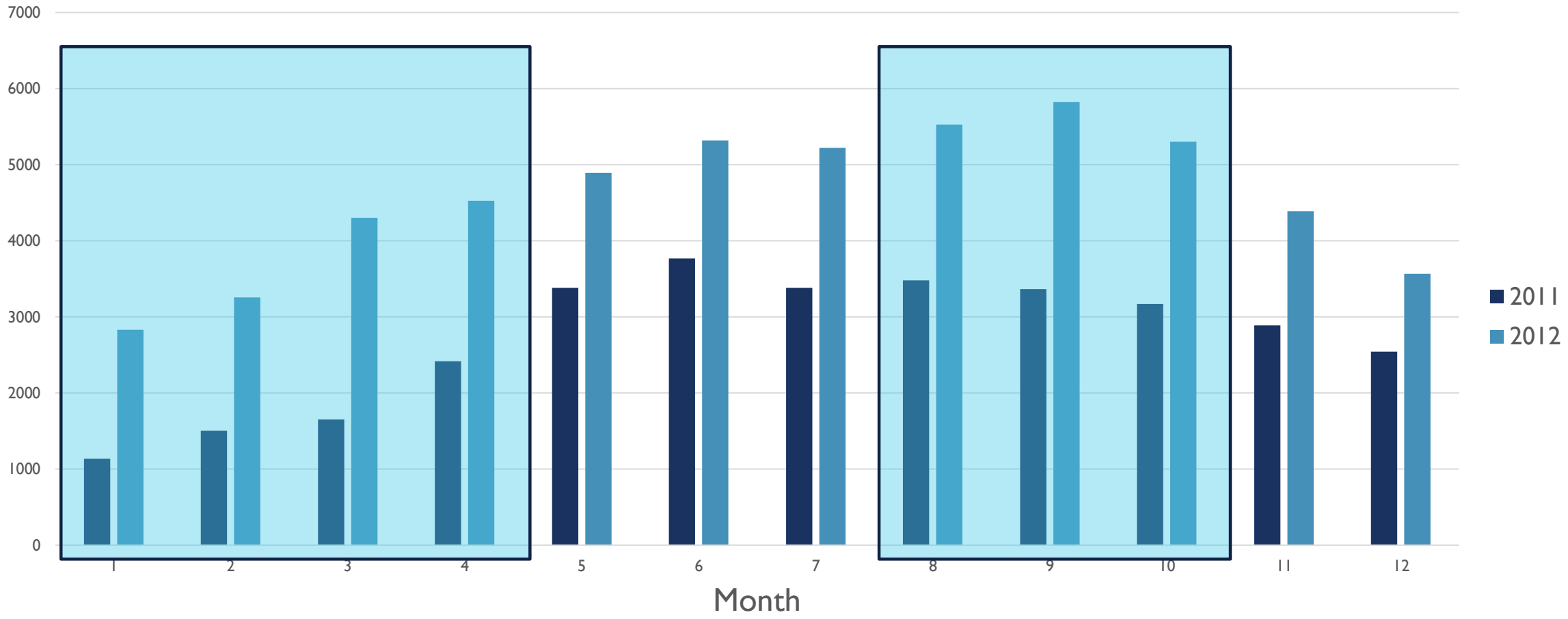
REGISTERED USERS OVER TIME



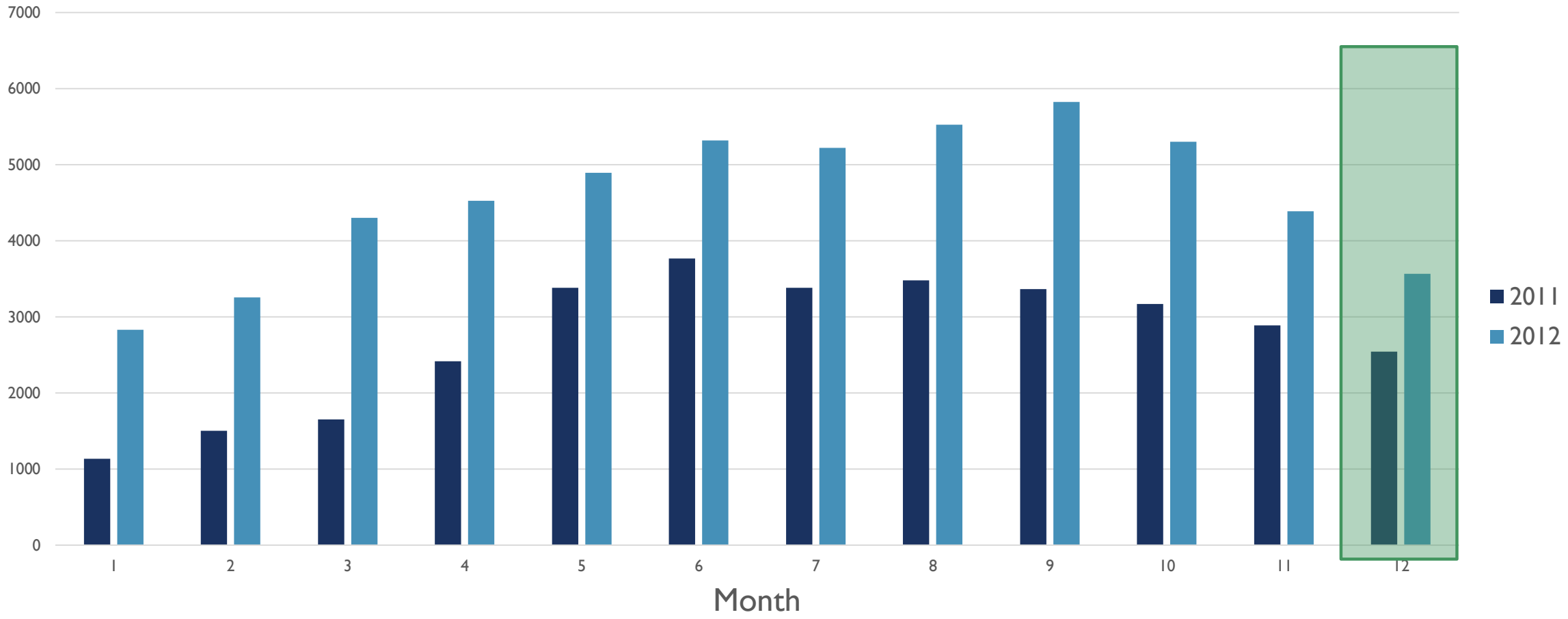
AVERAGE REGISTERED USERS YEAR OVER YEAR



AVERAGE REGISTERED USERS YEAR OVER YEAR



AVERAGE REGISTERED USERS YEAR OVER YEAR



KEY GOALS OF ANALYSIS – PART 2

- Identify key probabilities around customers:
  - Probability a customer is a registered user given the season is Fall
  - Probability a customer is a registered user given the season is Summer
  - Provide interval estimates for these probabilities
- The Customer's Marketing Division also has preconceived notions on the average number of total users for the 2012 seasons as follows to help develop their marketing budgets:
  - The average number of total users in the Summer is no less than 6500.
  - The average number of total users in the Fall is no more than 6500.
  - Validate the above claims using the appropriate statistical tests.