Dr Andrew Mitchell

*a.j.mitchell@ucl.ac.uk*

*Lecturer in AI and Machine Learning for Sustainable Construction*

2025-01-30

# Part 2: Statistical Sampling

# Why Study Sampling?

**The Power of Sampling:**

Nate Silver's 2012 Election Prediction:

- Correctly predicted all 50 states
- Used only 21,000 people
- To predict 125 million votes
- Combined data from 21 polls

**Key Insights:**

1. Small samples can be powerful
2. Proper methodology is crucial
3. Combining data improves accuracy
4. Statistical rigor matters

# Why Study Sampling?

One of the foundational ideas in statistics is that we can make inferences about an entire population based on a relatively small sample of individuals from that population.

Anyone living in the United States will be familiar with the concept of sampling from the political polls that have become a central part of our electoral process. In some cases, these polls can be incredibly accurate at predicting the outcomes of elections. The best known example comes from the 2008 and 2012 US Presidential elections, when the pollster Nate Silver correctly predicted electoral outcomes for 49/50 states in 2008 and for all 50 states in 2012.

Silver did this by combining data from 21 different polls, which vary in the degree to which they tend to lean towards either the Republican or Democratic side. Each of these polls included data from about 1000 likely voters – meaning that Silver was able to almost perfectly predict the pattern of votes of more than 125 million voters using data from only about 21,000 people, along with other knowledge.

# Sampling Fundamentals

1. **Population vs Sample:**
   - Population: Entire group of interest
   - Sample: Subset used for measurement
   - Goal: Infer population parameters from sample statistics
2. **Representative Sampling:**
   - Equal chance of selection
   - Avoid systematic bias
   - Random selection crucial
3. **Types of Sampling:**
   - With replacement: Items can be selected multiple times
   - Without replacement: Items selected only once
   - Choice affects probability calculations
4. **Key Terms:**
   - Parameter: Population value (usually unknown)

# Sampling Fundamentals

- Statistic: Sample value (our estimate)
- Sampling Error: Difference between statistic and parameter

Our goal in sampling is to determine the value of a statistic for an entire population of interest, using just a small subset of the population. We do this primarily to save time and effort – why go to the trouble of measuring every individual in the population when just a small sample is sufficient to accurately estimate the statistic of interest?

In the election example, the population is all registered voters in the region being polled, and the sample is the set of 1000 individuals selected by the polling organization. The way in which we select the sample is critical to ensuring that the sample is representative of the entire population, which is a main goal of statistical sampling.

It's important to also distinguish between two different ways of sampling: with replacement versus without replacement. In sampling with replacement, after a member of the population has been sampled, they are put back into the pool so that they can potentially be sampled

again. In sampling without replacement, once a member has been sampled they are not eligible to be sampled again.

# Sampling Error & Distribution

*Concept*

**What is Sampling Error?**

- Difference between sample and population
- Varies across samples
- Affects measurement quality
- Can be quantified

# Sampling Error & Distribution

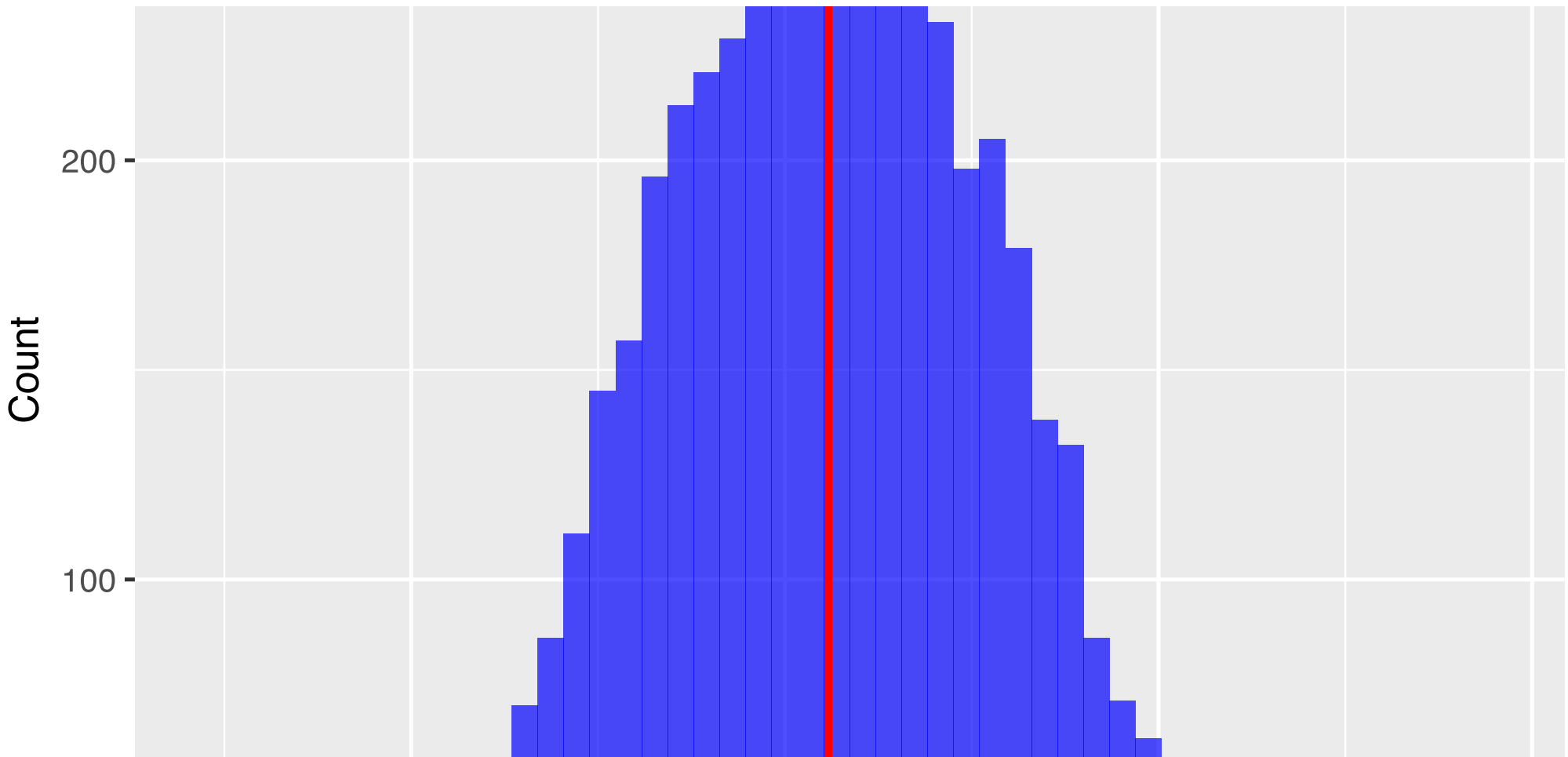
## *Concept*

```
# Take 5 samples of 50 adults each
set.seed(123)
samples <- map_df(
  1:5,
  ~{
    NHANES_adult |>
      sample_n(50) |>
      summarise(
        mean_height = mean(Height),
        sd_height = sd(Height)
      )
  }
)
samples
```

# Sampling Error & Distribution

```
# A tibble: 5 × 2
  mean_height sd_height
        <dbl>     <dbl>
1        169.     11.6
2        167.      9.13
3        169.     11.2
4        166.      9.62
5        169.     11.0
```

# Sampling Error & Distribution



# Sampling Error & Distribution

Regardless of how representative our sample is, it's likely that the statistic that we compute from the sample is going to differ at least slightly from the population parameter. We refer to this as sampling error. If we take multiple samples, the value of our statistical estimate will also vary from sample to sample; we refer to this distribution of our statistic across samples as the sampling distribution.

Sampling error is directly related to the quality of our measurement of the population. Clearly we want the estimates obtained from our sample to be as close as possible to the true value of the population parameter. However, even if our statistic is unbiased (that is, we expect it to have the same value as the population parameter), the value for any particular estimate will differ from the population value, and those differences will be greater when the sampling error is greater.

The visualization shows how sample means distribute around the true population mean (red line) when we take many samples.

# Standard Error of the Mean

**Definition:**

$SEM = \frac{\hat{\sigma}}{\sqrt{n}}$

Where:

- $\hat{\sigma}$ is estimated standard deviation
- $n$ is sample size

**Key Properties:**

- Measures sampling distribution variability
- Decreases with larger samples
- Increases with population variability

**Example with NHANES:**

# Standard Error of the Mean

```r
# Population SEM
pop_sd <- sd(NHANES_adult$Height)
n <- 50
sem_theoretical <- pop_sd / sqrt(n)

# Observed SEM from samples
sem_observed <- sd(samples_large$mean_height)

cat("Theoretical SEM:", round(sem_theoretical, 2), "\n")
```

```
Theoretical SEM: 1.44
```

```r
cat("Observed SEM:", round(sem_observed, 2))
```

# Standard Error of the Mean

```
Observed SEM: 1.42
```

Later in the course it will become essential to be able to characterize how variable our samples are, in order to make inferences about the sample statistics. For the mean, we do this using a quantity called the standard error of the mean (SEM), which one can think of as the standard deviation of the sampling distribution of the mean.

The formula for the standard error of the mean implies that the quality of our measurement involves two quantities: the population variability, and the size of our sample. Because the sample size is the denominator in the formula for SEM, a larger sample size will yield a smaller SEM when holding the population variability constant.

We have no control over the population variability, but we do have control over the sample size. Thus, if we wish to improve our sample statistics (by reducing their sampling variability) then we should use larger samples. However, the formula also tells us something

# Standard Error of the Mean

very fundamental about statistical sampling – namely, that the utility of larger samples diminishes with the square root of the sample size.

# Sample Size Effects

## *Theory*

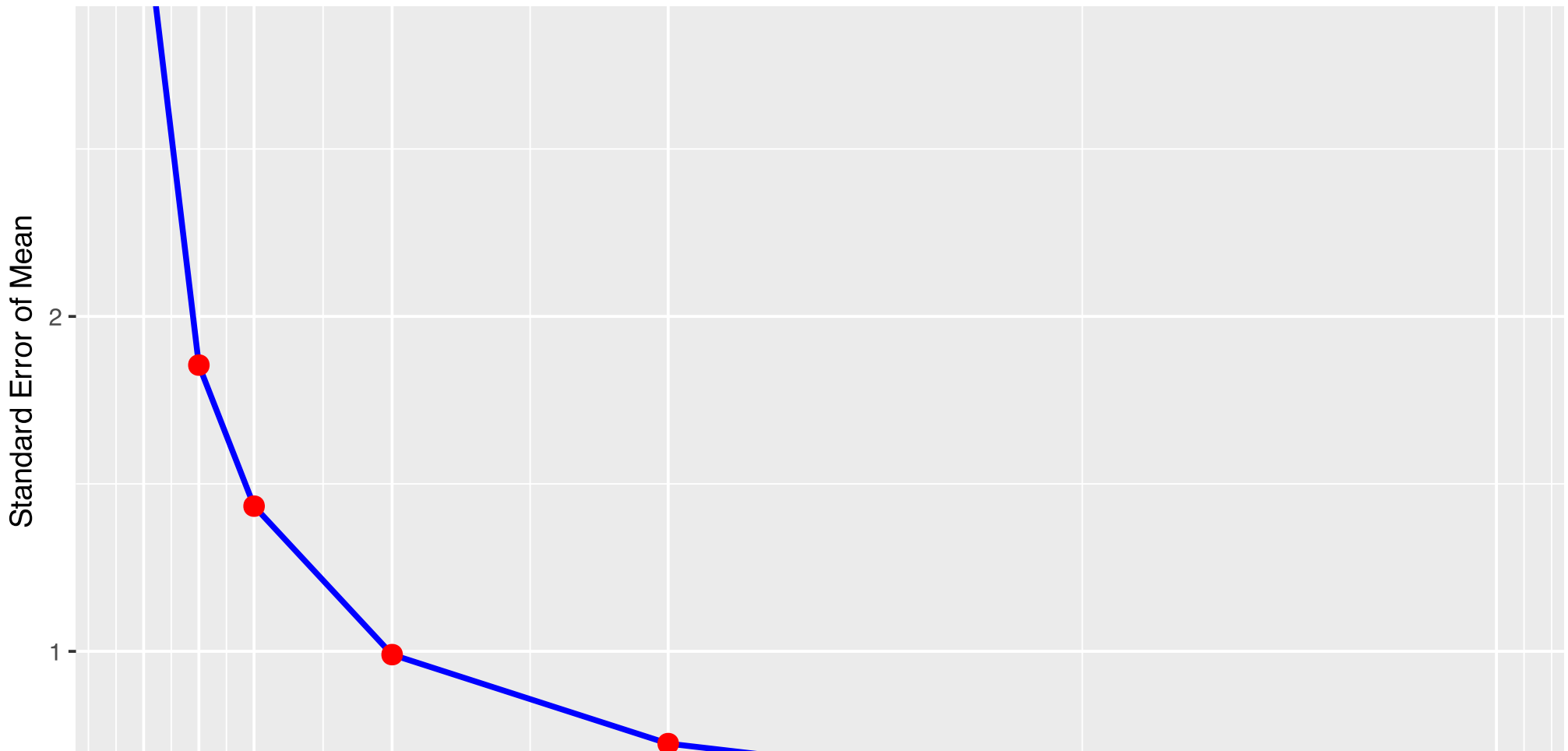
**Impact of Sample Size:**

- Larger n $\longrightarrow$ Smaller SEM
- Relationship is not linear
- Diminishing returns
- Square root relationship

# Sample Size Effects

*Visualization*

# Sample Size Effects



# Sample Size Effects

*Code*

```
# Compare SEM for different sample sizes
n1 <- 50
n2 <- 200 # 4 times larger

sem1 <- pop_sd / sqrt(n1)
sem2 <- pop_sd / sqrt(n2)

# Improvement factor
improvement <- sem1 / sem2
cat("Improvement factor:", round(improvement, 2))
```

The relationship between sample size and standard error is not linear. Doubling the sample size will not double the quality of the statistics; rather, it will improve it by a factor of $\sqrt{2}$. This has important implications for study design and resource allocation.

# Sample Size Effects

The visualization shows how the standard error decreases as sample size increases, but with diminishing returns. This means that after a certain point, increasing sample size may not be worth the additional cost and effort.

This relationship is fundamental to statistical power, which we will discuss in later sections. Understanding this relationship helps researchers make informed decisions about sample size requirements for their studies.

# The Central Limit Theorem

**Key Points:**

1. As sample size increases:
   - Sampling distribution becomes normal
   - Regardless of population distribution
   - Mean approaches population mean
   - Variance decreases
2. Implications:
   - Enables statistical inference
   - Justifies normal approximation
   - Explains real-world patterns

The Central Limit Theorem tells us that as sample sizes get larger, the sampling distribution of the mean will become normally distributed, even if the data within each sample are not

# The Central Limit Theorem

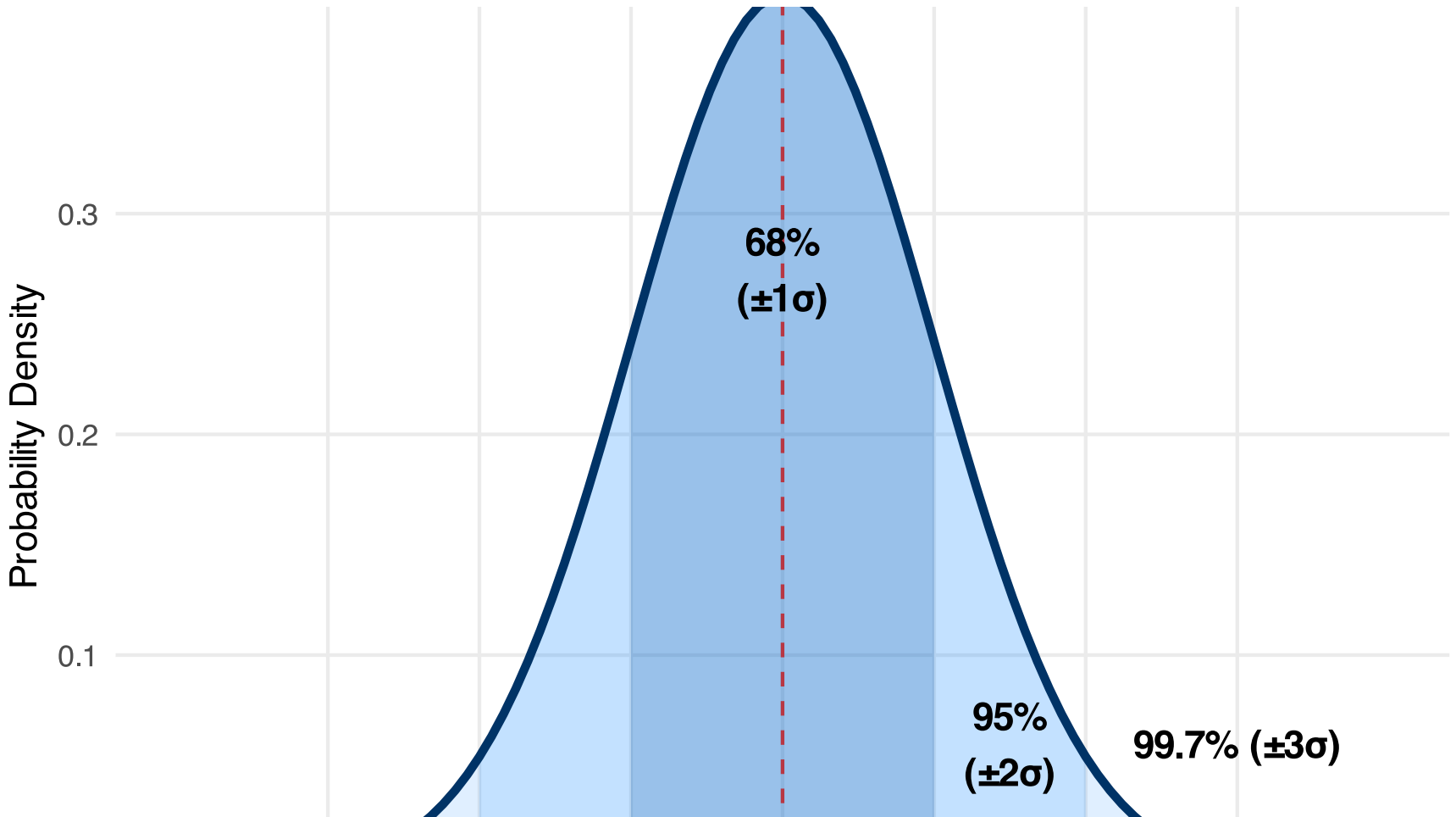
normally distributed. **This is a powerful result that allows us to make inferences about population parameters based on sample statistics.**

# The Central Limit Theorem

*Normal Distribution:*

- Bell-shaped curve
- Defined by mean ($\mu$) and SD ($\sigma$)
- Symmetric around mean

# The Central Limit Theorem



# The Central Limit Theorem

The Central Limit Theorem tells us that as sample sizes get larger, the sampling distribution of the mean will become normally distributed, even if the data within each sample are not normally distributed.

The normal distribution is described in terms of two parameters: the mean (which you can think of as the location of the peak), and the standard deviation (which specifies the width of the distribution). The bell-like shape of the distribution never changes, only its location and width.

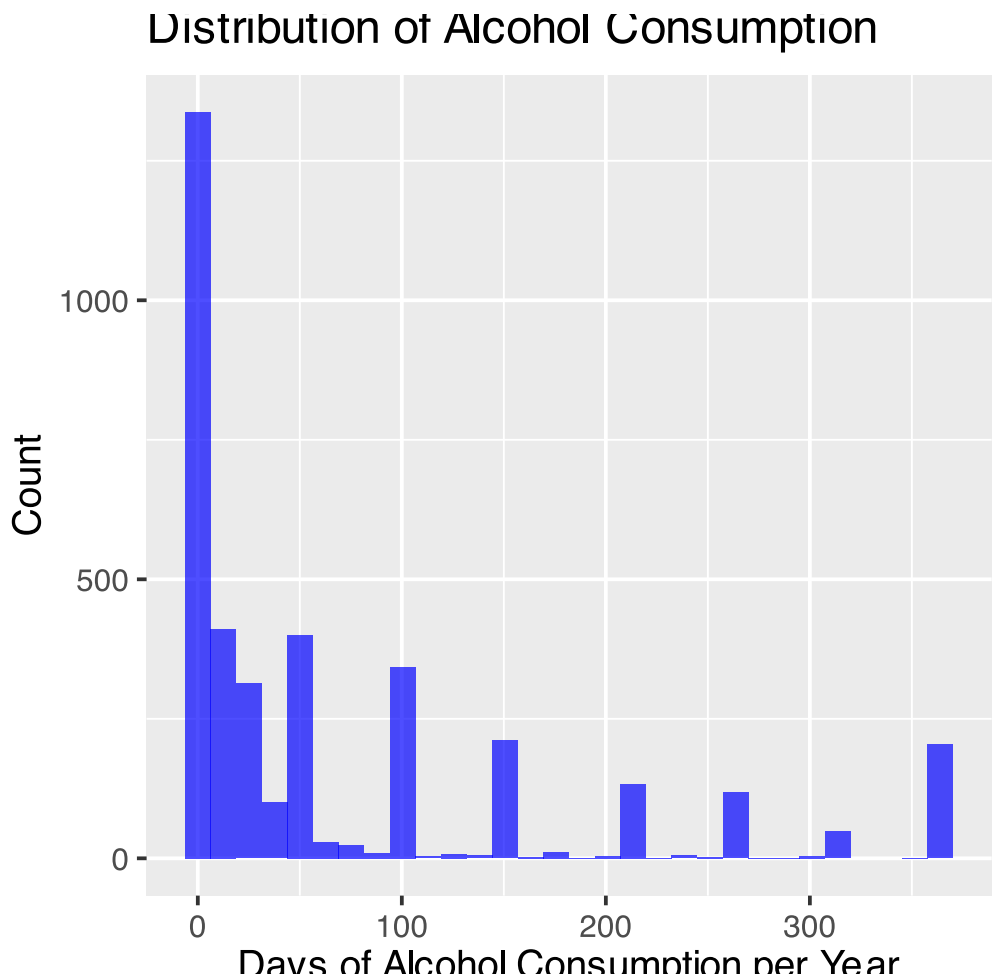
The normal distribution is commonly observed in data collected in the real world – and the central limit theorem gives us some insight into why that occurs. For example, the height of any adult depends on a complex mixture of their genetics and experience; even if those individual contributions may not be normally distributed, when we combine them the result is a normal distribution.

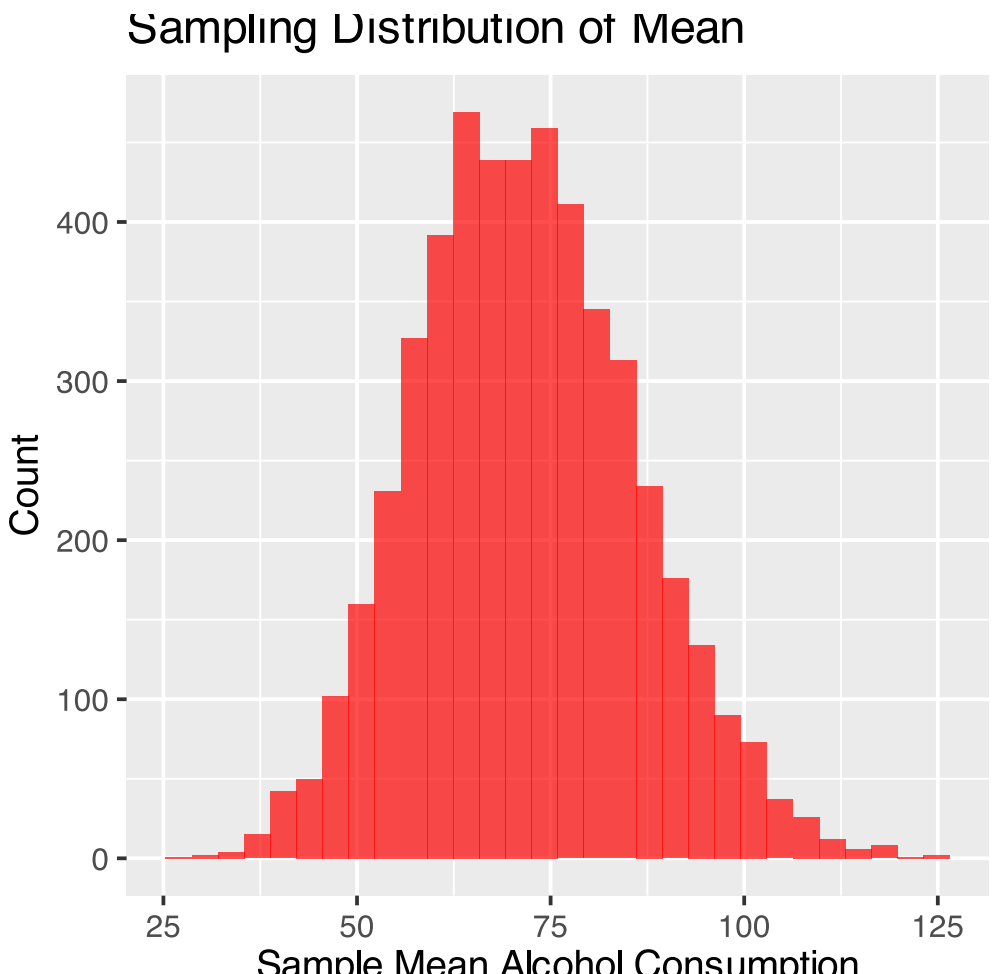
# The Central Limit Theorem

# CLT in Action: NHANES Example

*Original Distribution*

# CLT in Action: NHANES Example



# CLT in Action: NHANES Example

## *Code Example*

```r
# Compare skewness
library(moments)
original_skew <- skewness(NHANES_clean$AlcoholYear)
sampling_skew <- skewness(samples_alc$mean_alcohol)

cat("Original Distribution Skewness:", round(original_skew, 2), "\n")
cat("Sampling Distribution Skewness:", round(sampling_skew, 2))
```

## *Key Insights*

1. Original data is highly skewed
2. Sampling distribution is nearly normal
3. CLT works even with:
   - Non-normal data
   - Skewed distributions

# CLT in Action: NHANES Example

   - Discrete values
4. Sample size of 50 is sufficient

Let's work with the variable AlcoholYear from the NHANES dataset, which is highly skewed. This distribution is, for lack of a better word, funky – and definitely not normally distributed.

Now let's look at the sampling distribution of the mean for this variable. Despite the clear non-normality of the original data, the sampling distribution is remarkably close to the normal.

The Central Limit Theorem is important for statistics because it allows us to safely assume that the sampling distribution of the mean will be normal in most cases. This means that we can take advantage of statistical techniques that assume a normal distribution.

# Summary

1. **Sampling Fundamentals:**
   - Population vs Sample
   - Representative sampling
   - With/without replacement
   - Sampling error
2. **Standard Error:**
   - Measures sampling variability
   - Decreases with $\sqrt{n}$
   - Guides sample size decisions
   - Quantifies precision
3. **Central Limit Theorem:**
   - Sampling distribution normality
   - Independent of original distribution
   - Enables statistical inference

# Summary

   - Foundation for hypothesis testing
4. **Applications:**
   - Political polling
   - Clinical trials
   - Quality control
   - Research design

In this lecture, we covered: - The fundamentals of statistical sampling and why it works - How to characterize sampling error and the sampling distribution - The standard error of the mean and its relationship with sample size - The Central Limit Theorem and its importance in statistical inference - Real-world applications and examples using the NHANES dataset