Probability, Sampling, and Experiments

Dr Andrew Mitchell 

a.j.mitchell@ucl.ac.uk

Lecturer in AI and Machine Learning for Sustainable Construction

Part 1: Introduction to Probability

What is Probability Theory?

- Branch of mathematics dealing with chance and uncertainty
- Foundation for statistics
- Provides tools to describe uncertain events
- Historical origins in games of chance
- Deep questions about meaning and interpretation

Probability theory is the branch of mathematics that deals with chance and uncertainty. It forms an important part of the foundation for statistics, because it provides us with the mathematical tools to describe uncertain events.

The study of probability arose in part due to interest in understanding games of chance, like cards or dice. These games provide useful examples of many statistical concepts, because when we repeat these games the likelihood of different outcomes remains (mostly) the same. However, there are deep questions about the meaning of probability that we will not address here.

Experiment, Sample Space, Events

- An experiment is any activity that produces or observes an outcome. Examples are flipping a coin, rolling a 6-sided die, or trying a new route to work to see if it's faster than the old route.

Experiment, Sample Space, Events

Experiment, Sample Space, Events

- An experiment is any activity that produces or observes an outcome. Examples are flipping a coin, rolling a 6-sided die, or trying a new route to work to see if it's faster than the old route.
  - → Coin flip: {heads, tails}
  - → Die roll: {1,2,3,4,5,6}
  - → Travel time: $(0,\infty)$
- The sample space is the set of possible outcomes for an experiment. We represent these by listing them within a set of squiggly brackets.

Experiment, Sample Space, Events

Experiment, Sample Space, Events

- An experiment is any activity that produces or observes an outcome. Examples are flipping a coin, rolling a 6-sided die, or trying a new route to work to see if it's faster than the old route.
  - → Coin flip: {heads, tails}
  - → Die roll: {1,2,3,4,5,6}
  - → Travel time: $(0,\infty)$
- The sample space is the set of possible outcomes for an experiment. We represent these by listing them within a set of squiggly brackets.
- An event is a subset of the sample space. In principle it could be one or more of possible outcomes in the sample space, but here we will focus primarily on elementary events which consist of exactly one possible outcome.
  - → Subset of sample space
  - → Can be elementary or compound
  - → Example: rolling a 4

Experiment, Sample Space, Events

To formalize probability theory, we first need to define a few terms:

- An experiment is any activity that produces or observes an outcome. Examples are flipping a coin, rolling a 6-sided die, or trying a new route to work to see if it's faster than the old route.
- The sample space is the set of possible outcomes for an experiment. We represent these by listing them within a set of squiggly brackets.
- An event is a subset of the sample space. In principle it could be one or more of possible outcomes in the sample space, but here we will focus primarily on elementary events which consist of exactly one possible outcome.

Kolmogorov's Axioms

For events $E_1, E_2, \ldots, E_N$ and random variable X:

1. Non-negativity:

   $$P(X = E_i) \geq 0$$

2. Normalization:

   $$\sum_{i=1}^{N} P(X = E_i) = 1$$

3. Boundedness:

   $$P(X = E_i) \leq 1$$

Implications:

- All probabilities are between 0 and 1
- Total probability must sum to 1

Kolmogorov's Axioms

- Individual probabilities ≤ 1

These are the features that a value has to have if it is going to be a probability, which were first defined by the Russian mathematician Andrei Kolmogorov.

The summation is interpreted as saying "Take all of the N elementary events, which we have labeled from 1 to N, and add up their probabilities. These must sum to one."

The third point is implied by the previous points; since they must sum to one, and they can't be negative, then any particular probability cannot exceed one.

Probability Rules and Classical Probability

Basic Rules

1. Rule of Subtraction:

   $P(\neg A) = 1 - P(A)$

   Example: P(not rolling a 1) = $1 - \frac{1}{6} = \frac{5}{6}$

2. Intersection Rule (independent events):

   $P(A \cap B) = P(A)*P(B)$

   Example: P(six on both rolls) = $\frac{1}{6}*\frac{1}{6} = \frac{1}{36}$

3. Addition Rule:

   $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

To understand de Méré's error, we need to introduce some of the rules of probability theory:

Basic Rules

1. The rule of subtraction says that the probability of some event A not happening is one minus the probability of the event happening
2. For independent events, we compute the probability of both occurring by multiplying their individual probabilities
3. The addition rule tells us that to obtain the probability of either of two events occurring, we add together the individual probabilities, but then subtract the likelihood of both occurring together

Classical Probability

Key Principles:

- Equal likelihood assumption
- Based on counting outcomes
- No experiments needed
- Common in games of chance

Basic Formula:

$$P(outcome_i) = \frac{1}{\text{number of possible outcomes}}$$

Examples:

- Fair coin: P(heads) = 1/2
- Fair die: P(6) = 1/6
- Two dice: P(double-six) = 1/36

Classical Probability

Classical probability arose from the study of games of chance such as dice and cards. In this approach, we compute the probability directly based on our knowledge of the situation.

We start with the assumption that all of the elementary events in the sample space are equally likely; that is, when you roll a die, each of the possible outcomes ({1,2,3,4,5,6}) is equally likely to occur.

de Méré's Problem

French gambler Chevalier de Méré played two games:

1. Bet on ≥1 six in 4 die rolls
2. Bet on ≥1 double-six in 24 rolls of two dice

He thought both had probability $\frac{2}{3}$ but...

- Won money on first bet
- Lost money on second bet

His reasoning:

For first bet:

$4 * \frac{1}{6} = \frac{2}{3}$

For second bet:

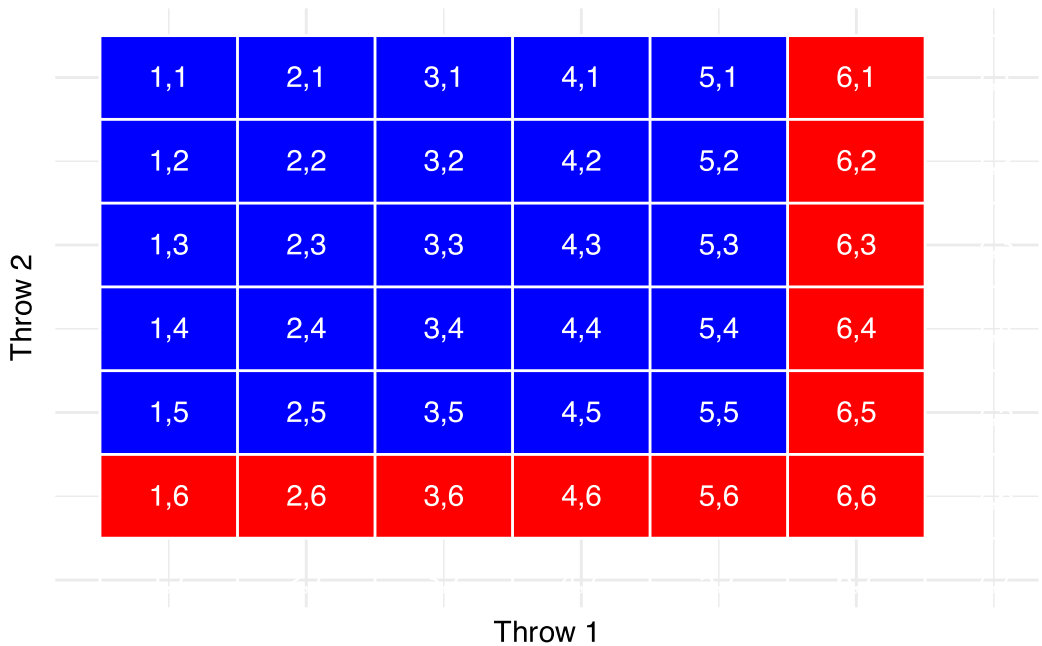
de Méré's Problem

$24 * \frac{1}{36} = \frac{2}{3}$

A famous example arose from a problem encountered by a French gambler who went by the name of Chevalier de Méré. de Méré played two different dice games: In the first he bet on the chance of at least one six on four rolls of a six-sided die, while in the second he bet on the chance of at least one double-six on 24 rolls of two dice. He expected to win money on both of these gambles, but he found that while on average he won money on the first gamble, he actually lost money on average when he played the second gamble many times.

Visualizing Multiple Events

Matrix of Outcomes:



Key Points:

- Red cells: six on either throw

Visualizing Multiple Events

- Total red cells: 11
- Explains $\frac{11}{36}$ probability
- Shows de Méré's error

This matrix represents all possible combinations of results across two throws, and highlights the cells that involve a six on either the first or second throw. If you count up the cells in red you will see that there are 11 such cells. This shows why the addition rule gives a different answer from de Méré's; if we were to simply add together the probabilities for the two throws as he did, then we would count (6,6) towards both, when it should really only be counted once.

Pascal's Solution

First bet:

$$P(\text{no sixes}) = \left(\frac{5}{6}\right)^4 = 0.482$$

$$P(\geq 1 \text{ six}) = 1 - 0.482 = 0.517$$

Second bet:

$$P(\text{no double six}) = \left(\frac{35}{36}\right)^{24} = 0.509$$

$$P(\geq 1 \text{ double six}) = 1 - 0.509 = 0.491$$

Key Insights:

- Easier to compute complement
- First bet: P > 0.5
- Second bet: P < 0.5

Pascal's Solution

- Explains gambling results

Blaise Pascal used the rules of probability to solve de Méré's problem. First, he realized that computing the probability of at least one event out of a combination was tricky, whereas computing the probability that something does not occur across several events is relatively easy – it's just the product of the probabilities of the individual events.

The first bet has probability > 0.5, explaining why de Méré made money on this bet on average. The second bet has probability < 0.5, explaining why de Méré lost money on average on this bet.

Determining Probabilities

Three Approaches

1. **Personal Belief**
   - Subjective assessment
   - Based on knowledge/experience
   - Limited scientific validity
   - Often only available approach
2. **Empirical Frequency**
   - Based on repeated experiments
   - Law of large numbers
   - Real-world data collection
3. **Classical Probability**
   - Based on equally likely outcomes
   - Mathematical approach
   - Common in games of chance
   - No experiments needed

Three Approaches

Now that we know what a probability is, how do we actually figure out what the probability is for any particular event? There are three main approaches, each with their own strengths and limitations.

Personal Belief

Example Question:

What was the probability that Bernie Sanders would have won the 2016 presidential election if he had been the democratic nominee?

Key Points:

- Can't run this experiment
- People can still estimate based on knowledge
- Not scientifically satisfying
- Often the only available approach

Other Examples:

- Weather forecasts
- Sports predictions
- Economic forecasts

Personal Belief

- Personal decisions

Let's say that I asked you what the probability was that Bernie Sanders would have won the 2016 presidential election if he had been the democratic nominee instead of Hilary Clinton? We can't actually do the experiment to find the outcome. However, most people with knowledge of American politics would be willing to at least offer a guess at the probability of this event. In many cases personal knowledge and/or opinion is the only guide we have determining the probability of an event, but this is not very scientifically satisfying.

Empirical Frequency

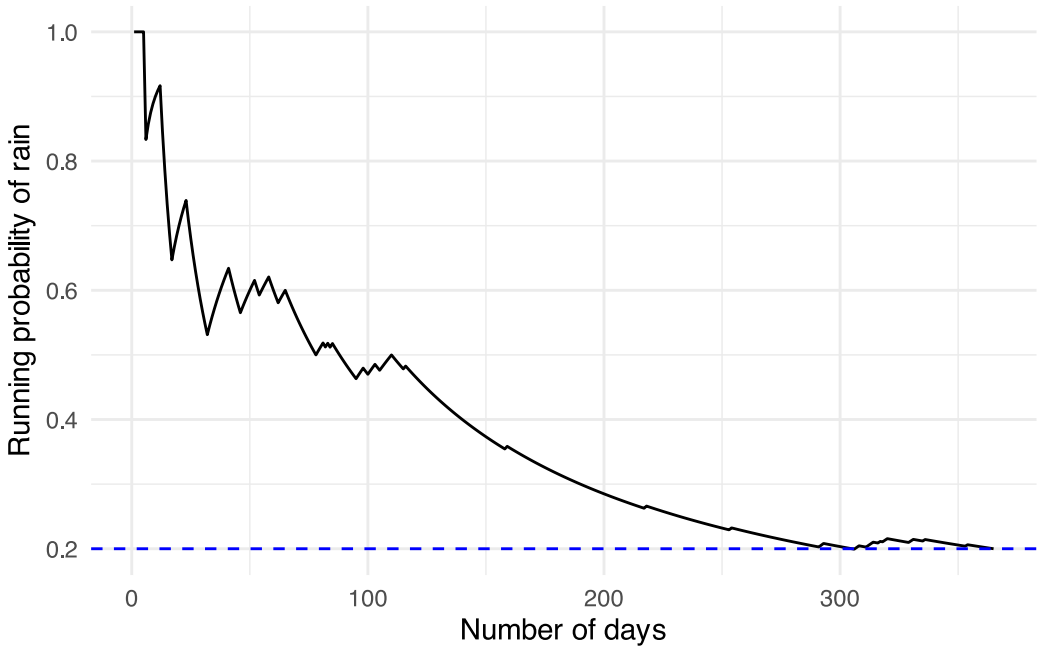
San Francisco Rain Example:

- Total days in 2017: 365
- Rainy days: 73
- P(rain in SF) = 73/365 = 0.2

Key Steps:

1. Define experiment clearly
2. Count occurrences
3. Divide by total trials

Empirical Frequency



Another way to determine the probability of an event is to do the experiment many times and count how often each event happens. From the relative frequency of the different outcomes, we can compute the probability of each outcome. For example, let's say that we

Empirical Frequency

are interested in knowing the probability of rain in San Francisco. We first have to define the experiment — let's say that we will look at the National Weather Service data for each day in 2017 and determine whether there was any rain at the downtown San Francisco weather station. According to these data, in 2017 there were 73 rainy days. To compute the probability of rain in San Francisco, we simply divide the number of rainy days by the number of days counted (365), giving P(rain in SF in 2017) = 0.2.

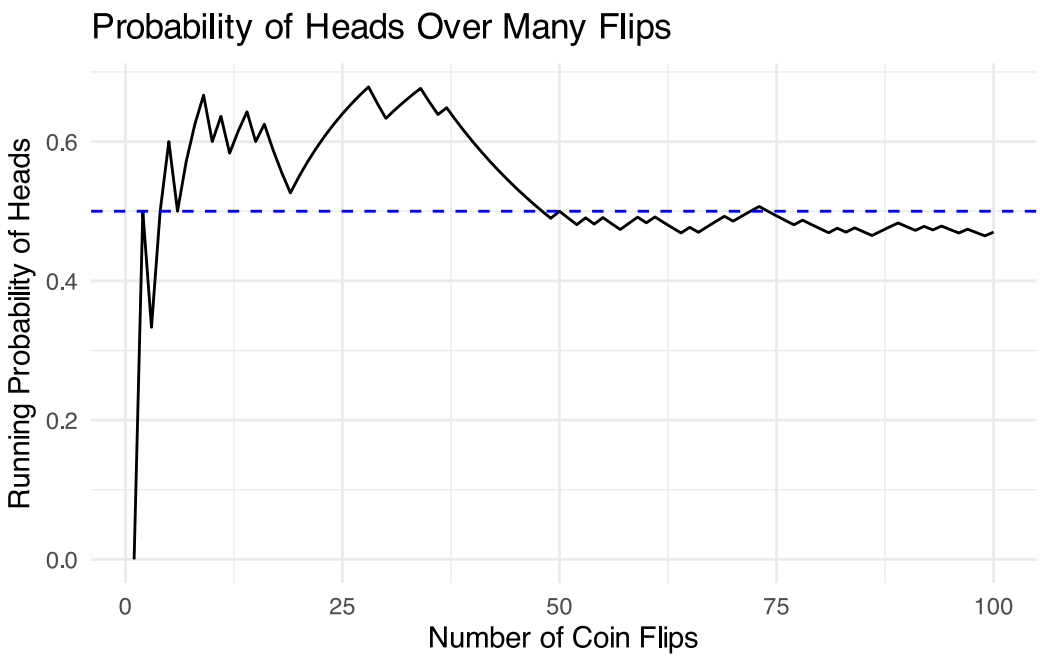
The graph shows how the empirical probability of rain converges to 0.2 as we accumulate more days of data throughout the year.

Law of Large Numbers

Coin Flip Example:

- True probability of heads = 0.5
- Small samples vary widely
- More flips = better estimate
- Converges to true probability
- “Law of small numbers” fallacy

Law of Large Numbers



The graph shows how early results from coin flips can be highly variable and unrepresentative of the true value. Even though we know a fair coin has a probability of 0.5

Law of Large Numbers

for heads, small samples can give very different results. This demonstrates how small samples can give misleading results.

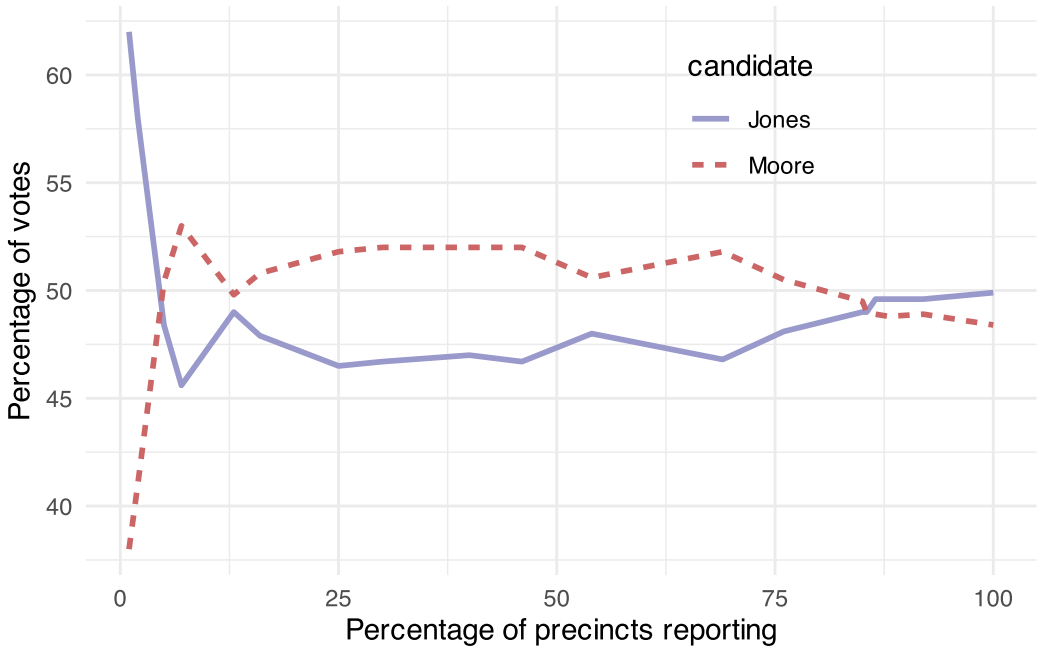
This was referred to as the “law of small numbers” by psychologists Danny Kahneman and Amos Tversky, who showed that people (even trained researchers) often behave as if the law of large numbers applies even to small samples, leading them to draw strong conclusions from insufficient data.

Real-World Example: Alabama Election

2017 Senate Race:

- Roy Moore vs Doug Jones
- Early results volatile
- Final outcome different
- Small sample warning

Real-World Example: Alabama Election



A real-world example of this was seen in the 2017 special election for the US Senate in Alabama. Early in the evening the vote counts were especially volatile, swinging from a large

Real-World Example: Alabama Election

initial lead for Jones to a long period where Moore had the lead, until finally Jones took the lead to win the race.

This demonstrates how small samples can give misleading results. Unfortunately, many people forget this and overinterpret results from small samples.

Conditional Probability and Independence

What is Conditional Probability?

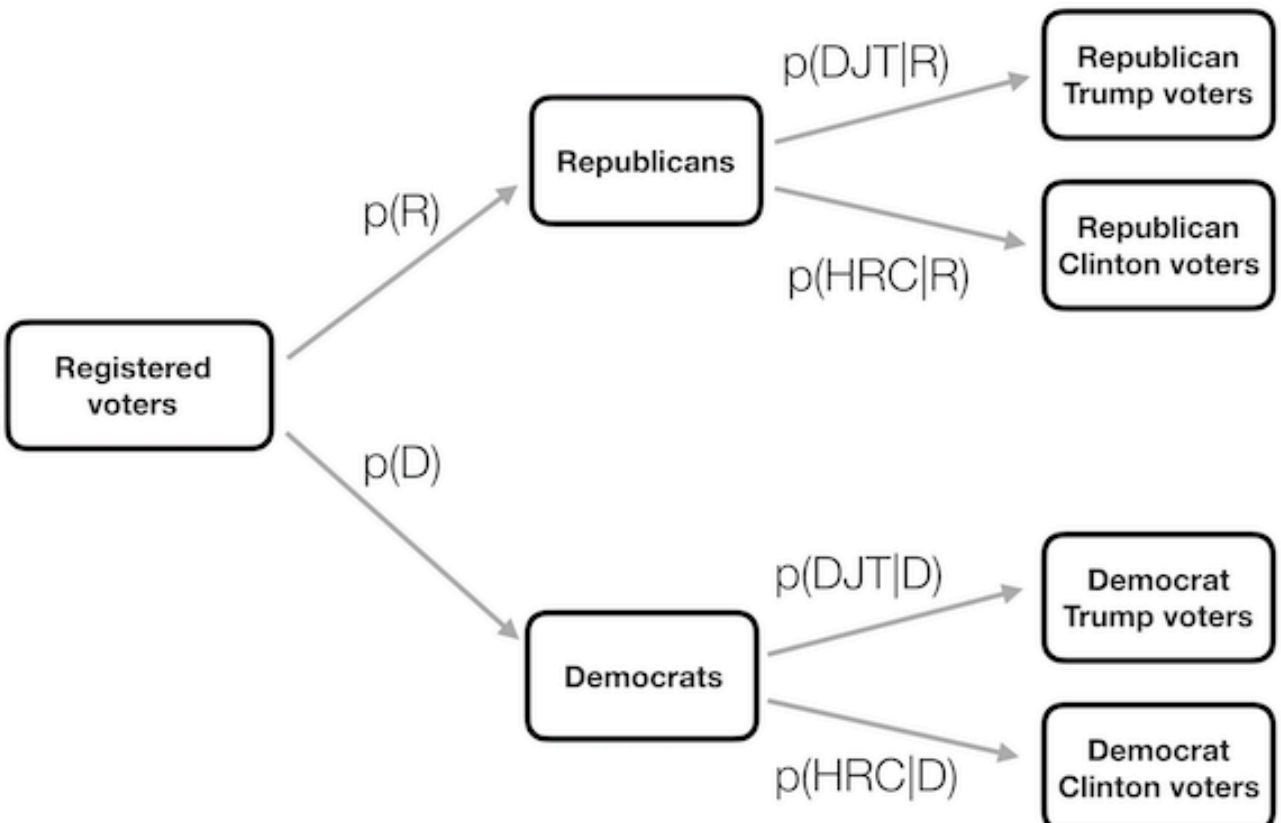
Definition:

- Probability of A given B occurred
- Written as $P(A \mid B)$
- Updates probability based on new information

Formula:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

What is Conditional Probability?



What is Conditional Probability?

So far we have limited ourselves to simple probabilities - that is, the probability of a single event or combination of events. However, we often wish to determine the probability of some event given that some other event has occurred, which are known as conditional probabilities.

NHANES Example: Physical Activity

Question:

What is P(diabetes|inactive)?

total	inactive	diabetes	diabetes_given_inactive
5443	0.454	0.101	0.141

Joint Probabilities:

Diabetes	PhysActive	n	prob
No	No	2123	0.3900423
No	Yes	2770	0.5089105
Yes	No	349	0.0641191
Yes	Yes	201	0.0369282

We can compute conditional probabilities directly from data. Let's say that we are interested in the following question: What is the probability that someone has diabetes, given that they

NHANES Example: Physical Activity

are not physically active? The NHANES dataset includes two variables that address the two parts of this question: Diabetes and PhysActive.

Independence

Statistical Independence:

$$P(A \mid B) = P(A)$$

Key Points:

- B tells us nothing about A
- Different from everyday usage
- Must check with data

Example: Jefferson State

- P(Jeffersonian) = 0.014
- P(Californian) = 0.986
- Not independent!
- Mutually exclusive

Independence

The term “independent” has a very specific meaning in statistics, which is somewhat different from the common usage of the term. Statistical independence between two variables means that knowing the value of one variable doesn’t tell us anything about the value of the other.

For example, there is currently a move by a small group of California citizens to declare a new independent state called Jefferson. The new states might be politically independent, but they would not be statistically independent, because if we know that a person is Jeffersonian, then we can be sure they are not Californian!

Mental Health and Physical Activity

Question: Are physical and mental health independent?

Variables:

- PhysActive: physically active?
- DaysMentHlthBad: bad mental health days
- Threshold: >7 days = bad mental health

PhysActive	Bad Mental Health	Good Mental Health	Total
No	629	2510	3139
Yes	471	3095	3566
Total	1100	5605	6705

Let's look at another example, using the NHANES data: Are physical health and mental health independent of one another? To determine whether mental health and physical

Mental Health and Physical Activity

activity are independent, we would compare the simple probability of bad mental health to the conditional probability of bad mental health given that one is physically active.

Bayes' Rule and Learning from Data

The Basic Formula

When we know $P(A \mid B)$ but want $P(B \mid A)$:

$$P(B \mid A) = \frac{P(A \mid B)*P(B)}{P(A)}$$

Alternative Form:

$$P(B \mid A) = \frac{P(A \mid B)*P(B)}{P(A \mid B)*P(B)+P(A \mid \neg B)*P(\neg B)}$$

Components:

- Prior: $P(B)$
- Likelihood: $P(A \mid B)$
- Marginal likelihood: $P(A)$
- Posterior: $P(B \mid A)$

The Basic Formula

In many cases, we know P(A|B) but we really want to know P(B|A). This commonly occurs in medical screening, where we know P(positive test result| disease) but what we want to know is P(disease|positive test result).

If we have only two outcomes, we can express Bayes' rule in a somewhat clearer way, using the sum rule to redefine P(A).

Putting Bayes into Practice

Construction company drug testing

A major construction company conducts mandatory random drug and alcohol screening using rapid saliva tests. Consider the following scenario:

- In the UK construction industry during 2023, the prevalence of substance use affecting workplace safety was estimated at 2.5% of the workforce

Putting Bayes into Practice

Construction company drug testing

A major construction company conducts mandatory random drug and alcohol screening using rapid saliva tests. Consider the following scenario:

- In the UK construction industry during 2023, the prevalence of substance use affecting workplace safety was estimated at 2.5% of the workforce
- The rapid saliva test used has a sensitivity (true positive rate) of 85% when conducted according to protocol

Putting Bayes into Practice

Construction company drug testing

A major construction company conducts mandatory random drug and alcohol screening using rapid saliva tests. Consider the following scenario:

- In the UK construction industry during 2023, the prevalence of substance use affecting workplace safety was estimated at 2.5% of the workforce
- The rapid saliva test used has a sensitivity (true positive rate) of 85% when conducted according to protocol
- The specificity (true negative rate) of these tests is 99.2%

Let's consider a specific example. Suppose that a worker is selected for a random drug screening. The test result is positive. What is the probability that this worker is actually positive for substances?

Putting Bayes into Practice

Construction company drug testing

Let's consider a specific example. Suppose that a worker is selected for a random drug screening. The test result is positive. What is the probability that this worker is actually positive for substances?

Context: The company's current policy is immediate suspension without pay following a positive test result, pending a more accurate laboratory confirmation test that takes 48 hours.

- Mandatory screening
- Rapid saliva test
- Safety-critical roles
- Immediate consequences

Putting Bayes into Practice

Construction company drug testing

Construction Site Testing:

- Sensitivity: P(positive|substance) = 0.85
- Specificity: P(negative|no substance) = 0.992
- Base rate: P(substance) = 0.025

Key Values:

- P(S) = 0.025 (prevalence)
- P(P|S) = 0.85 (sensitivity)
- P(P|not S) = 0.008 (1 - specificity)

A major construction company conducts mandatory random drug and alcohol screening using rapid saliva tests. In the UK construction industry during 2023, the prevalence of substance use affecting workplace safety was estimated at 2.5% of the workforce. The rapid

Putting Bayes into Practice

saliva test used has a sensitivity of 85% when conducted according to protocol, and a specificity of 99.2%.

Let's Work Through It

Using Bayes' Theorem, calculate the probability that this worker is actually positive for substances given their positive test result.

- P(S) = 0.025 (prevalence)
- P(P|S) = 0.85 (sensitivity)
- P(P|not S) = 0.008 (1 - specificity)

A construction worker is randomly selected for testing at the start of their shift. Their saliva test comes back positive. Using Bayes' Theorem, calculate the probability that this worker is actually positive for substances given their positive test result.

Solution

Calculate P(substance|positive):

Interpretation:

- ~73.1% chance true positive
- ~26.9% chance false positive
- Much higher than 2.5% base rate
- Still significant uncertainty

Using Bayes' Theorem, we find that given a positive test result, there is a 73.1% probability that the worker actually has substances present. This is much higher than the base rate of 2.5%, but still leaves significant uncertainty with a 26.9% false positive rate.

Discussion: The Real-world Implications

The company's current policy is immediate suspension without pay following a positive test result.

What do these results mean for this business policy? Is it fair to immediately suspend workers without pay for a positive test result?

The company's current policy is immediate suspension without pay following a positive test result, pending a more accurate laboratory confirmation test that takes 48 hours.

Given that approximately 26.9% of positive test results may be false positives, an immediate suspension without pay could unfairly penalize innocent workers; however, the high stakes of construction safety and the 73.1% probability of a true positive suggest that temporary removal from safety-critical roles is prudent while awaiting confirmation.

Learning from Data

Bayes' Rule as Learning:

$$P(B \mid A) = \frac{P(A \mid B)}{P(A)} * P(B)$$

Components:

- Prior belief: $P(B)$
- Evidence strength: $\frac{P(A \mid B)}{P(A)}$
- Updated belief: $P(B \mid A)$

Key Insights:

- Updates prior knowledge
- Evidence can strengthen/weaken
- Systematic way to learn
- Combines knowledge & data

Learning from Data

Another way to think of Bayes' rule is as a way to update our beliefs on the basis of data. The different parts of Bayes' rule have specific names, that relate to their role in using Bayes' rule to update our beliefs.

The part on the left tells us how much more or less likely the data A are given B, relative to the overall likelihood of the data, while the part on the right side tells us how likely we thought B was before we knew anything about the data.

Odds and Odds Ratios

Converting to Odds:

$$\text{odds of A} = \frac{P(A)}{P(\neg A)}$$

Example:

Drug test odds:

- Prior: $\frac{0.025}{0.975} = 0.026$
- Posterior: $\frac{0.7314974}{0.2685026} = 2.724$

Odds Ratio:

$$\frac{\text{posterior odds}}{\text{prior odds}} = \frac{2.724}{0.026} = 106.25$$

Interpretation:

- Odds increased 105×

Odds and Odds Ratios

- Much stronger evidence
- Shows test's power
- Despite false positives

We can convert probabilities into odds which express the relative likelihood of something happening or not. An odds ratio is an example of what we will later call an effect size, which is a way of quantifying how relatively large any particular statistical effect is.

First, remember the rule for computing a conditional probability. We can rearrange this to get the formula to compute the joint probability using the conditional. Using this we can compute the inverse probability.

Probability Distributions

What is a Probability Distribution?

Definition:

- Describes all possible outcomes
- Assigns probability to each
- Different types for different data
- Mathematical formulation

Examples:

- Binomial (success/failure)
- Normal (continuous)
- Poisson (counts)

```r
# Create example distributions
x <- seq(-4, 4, length.out = 100)
normal_df <- data.frame(
```

What is a Probability Distribution?

```
  x = x,
  y = dnorm(x),
  type = "Normal"
)

x <- 0:10
poisson_df <- data.frame(
  x = x,
  y = dpois(x, lambda = 3),
  type = "Poisson"
)

colors <- c(
  "Normal" = "blue",
  "Poisson" = "red"
```

What is a Probability Distribution?

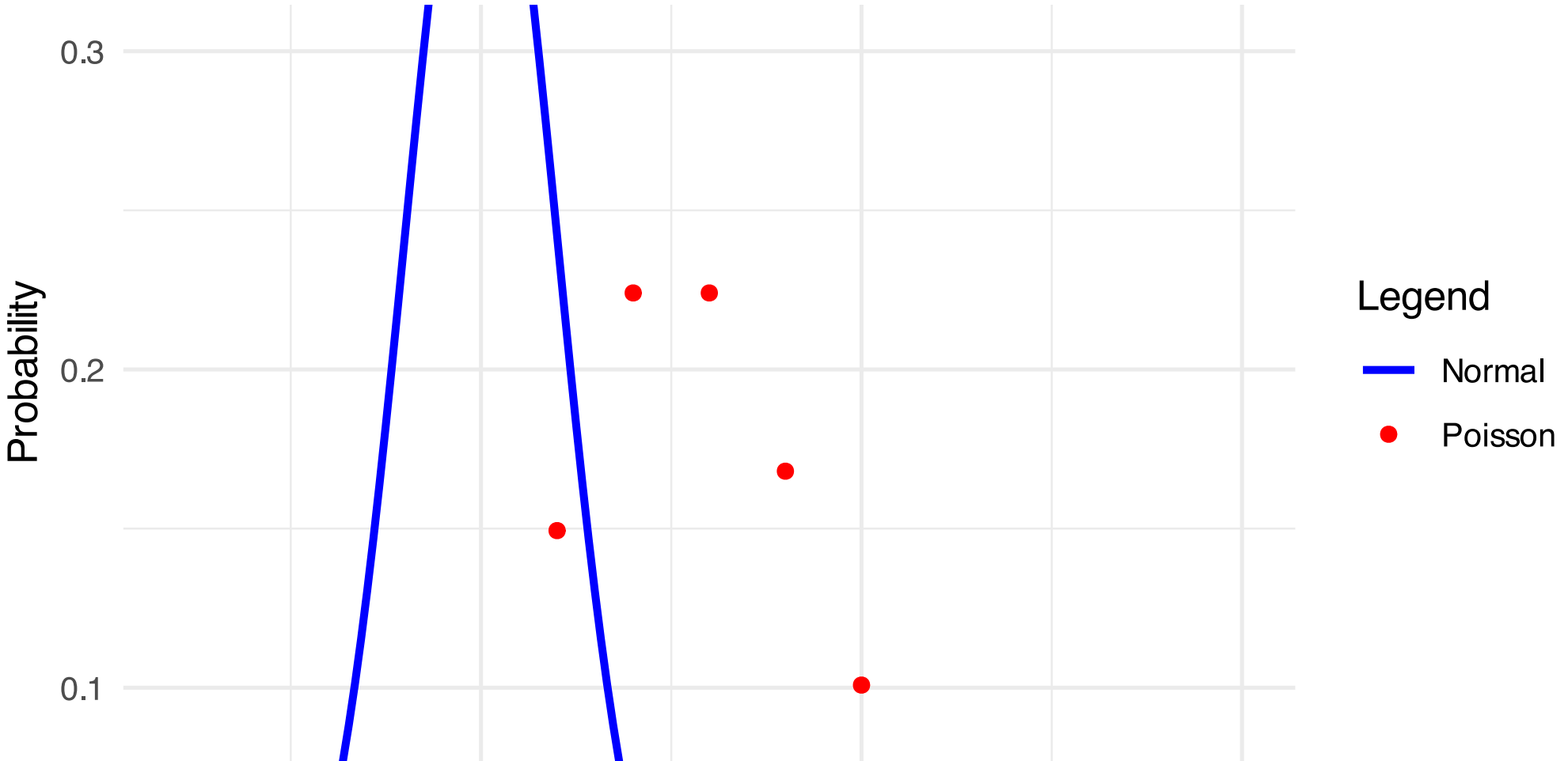
```
)

# Plot distributions
ggplot() +
  geom_line(data = normal_df, aes(x = x, y = y, color = "Normal"), size = 1)
+
  geom_point(
    data = poisson_df,
    aes(x = x, y = y, color = "Poisson"),
    size = 1.5
  ) +
  labs(
    title = "Example Distributions",
    x = "Value",
    y = "Probability",
```

What is a Probability Distribution?

```
    color = "Legend"
  ) +
  theme(legend.position = "top") +
  scale_color_manual(values = colors) +
  theme_minimal()
```

What is a Probability Distribution?



What is a Probability Distribution?

A probability distribution describes the probability of all of the possible outcomes in an experiment. Throughout this section we will encounter a number of these probability distributions, each of which is appropriate to describe different types of data.

The Binomial Distribution

Properties:

- Independent trials
- Two outcomes
- Fixed probability
- Order doesn't matter

Formula:

$$P(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k}$$

Where:

- k = successes
- n = trials
- p = probability per trial

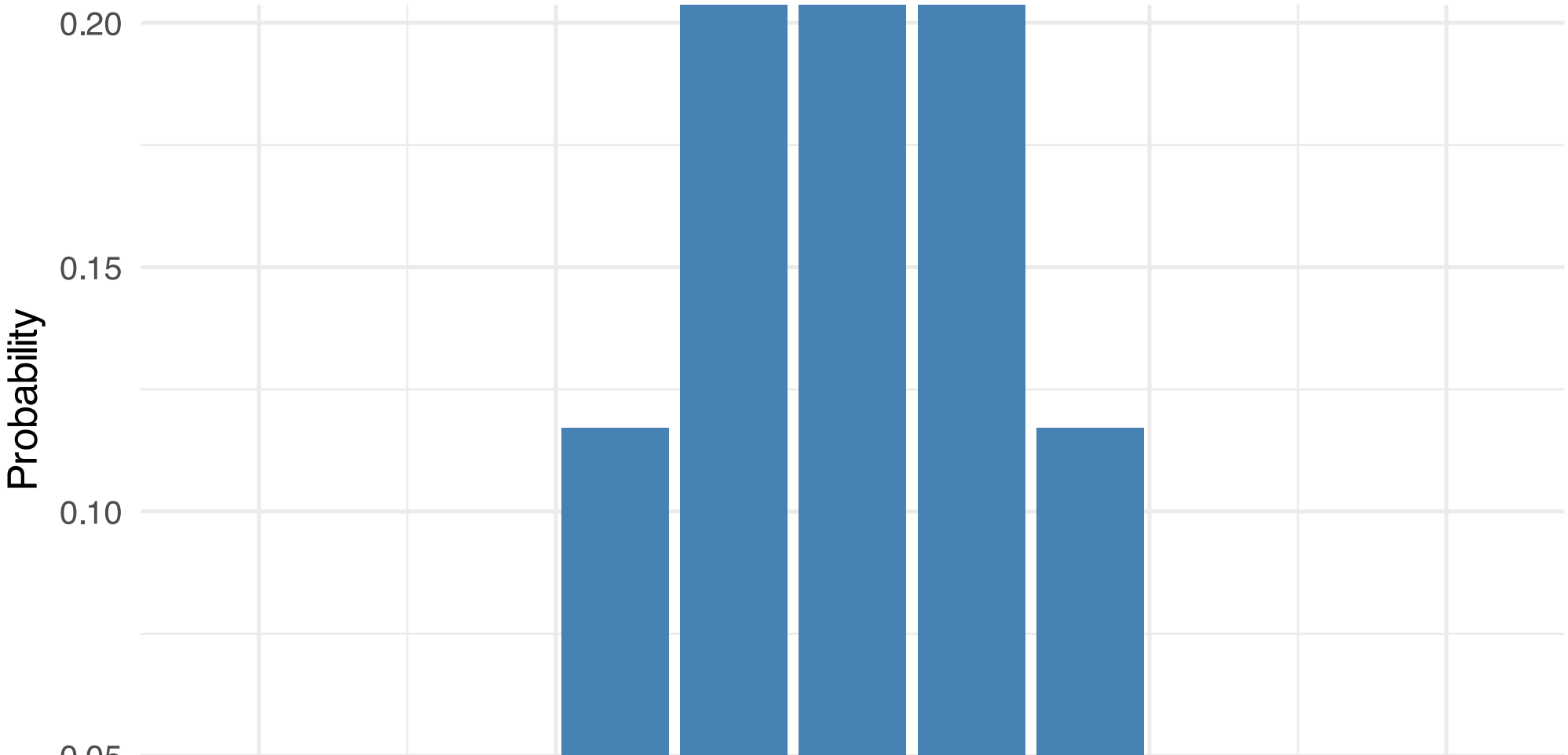
The Binomial Distribution

Binomial Coefficient:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

```r
# Create binomial distribution plot
x <- 0:10
n <- 10
p <- 0.5
binom_df <- data.frame(
  x = x,
  y = dbinom(x, size = n, prob = p)
)
```

The Binomial Distribution



The Binomial Distribution

The binomial distribution provides a way to compute the probability of some number of successes out of a number of trials on which there is either success or failure and nothing in between (known as “Bernoulli trials”), given some known probability of success on each trial.

Example: Steph Curry's Free Throws

Scenario:

- Steph Curry hits 91% of his free throws
- In a game in Jan, 2018, he hit only **2 out of 4** free throws
- It seems pretty unlikely that he would hit only 50% of his free throws in a game, but exactly how unlikely is it?

Calculation:

$P(2; 4, 0.91) = \binom{4}{2} 0.91^2 (1 - 0.91)^2$

$= 6*0.8281*0.0081$

$= 0.040$

Interpretation:

- Very unlikely (4%)

Example: Steph Curry's Free Throws

- Yet it happened
- Rare events do occur
- Don't overinterpret

On Jan 20 2018, the basketball player Steph Curry hit only 2 out of 4 free throws in a game against the Houston Rockets. We know that Curry's overall probability of hitting free throws across the entire season was 0.91, so it seems pretty unlikely that he would hit only 50% of his free throws in a game, but exactly how unlikely is it?

Cumulative Distributions

Often we want to know not just how likely a specific value is, but how likely it is to find a value that is as extreme or more than a particular value?

Definition:

- Probability of value ≤ x
- Accumulates probabilities
- Often more useful
- Important for testing

Example:

$$P(k \leq 2) = P(k = 2) + P(k = 1) + P(k = 0)$$

Often we want to know not just how likely a specific value is, but how likely it is to find a value that is as extreme or more than a particular value?

Cumulative Distributions

```
# curry_df <- tibble(
#   numSuccesses = seq(0, 4)
# ) %>%
#   mutate(
#     Probability = dbinom(numSuccesses, size = 4, prob = 0.91),
#     CumulativeProbability = pbinom(numSuccesses, size = 4, prob = 0.91)
#   )
# Create data for Curry's free throw distributions
n_throws <- 4
curry_prob <- 0.91
x <- 0:n_throws

curry_dist_df <- data.frame(
  x = x,
  Simple = dbinom(x, size = n_throws, prob = curry_prob),
```

Cumulative Distributions

```
  Cumulative = pbinom(x, size = n_throws, prob = curry_prob)
)

kable(
  curry_dist_df,
  caption = "Simple and cumulative probability distributions",
  digits = 3
)
```

x	Simple	Cumulative
0	0.000	0.000
1	0.003	0.003
2	0.040	0.043
3	0.271	0.314
4	0.686	1.000

Cumulative Distributions

Cumulative Distributions

The binomial distribution is a discrete probability distribution that describes the number of successes in a sequence of independent experiments, each of which has a constant probability of success. In this example, we are looking at the probability of Steph Curry making a certain number of free throws out of 4 attempts, given that his overall success rate is 91%.

This visualization shows both the probability of making exactly k free throws (blue bars) and the probability of making k or fewer free throws (red line) for Curry's specific scenario of 4 attempts with a 91% success rate.

Summary

Core Concepts:

1. Probability measures uncertainty
2. Three approaches:
   - Personal belief
   - Empirical frequency
   - Classical probability
3. Fundamental rules:
   - Addition
   - Multiplication
   - Subtraction

Advanced Topics:

1. Conditional probability
2. Independence

Summary

3. Bayes' rule
4. Probability distributions

Applications:

- Medical screening
- Data analysis
- Decision making
- Statistical inference

These concepts form the foundation for statistical inference, which we will explore in later chapters. Having read this chapter, you should be able to:

- Describe the sample space for a selected random experiment
- Compute relative frequency and empirical probability
- Compute probabilities of single events, complementary events, and unions/intersections

Summary

- Describe the law of large numbers
- Understand conditional probability and independence
- Use Bayes' theorem

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

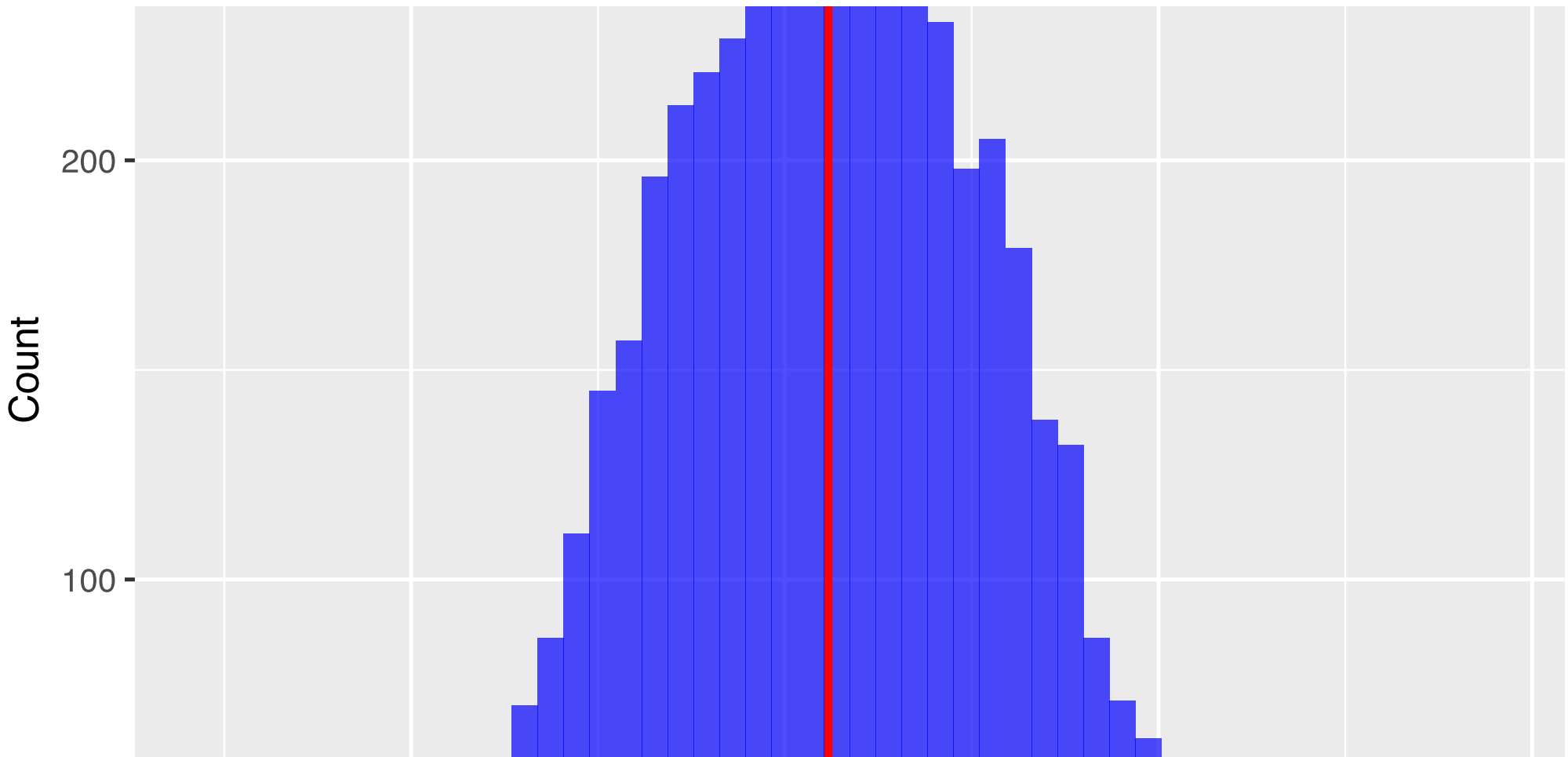
Concept

```
# 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

```
# 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
```

```
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

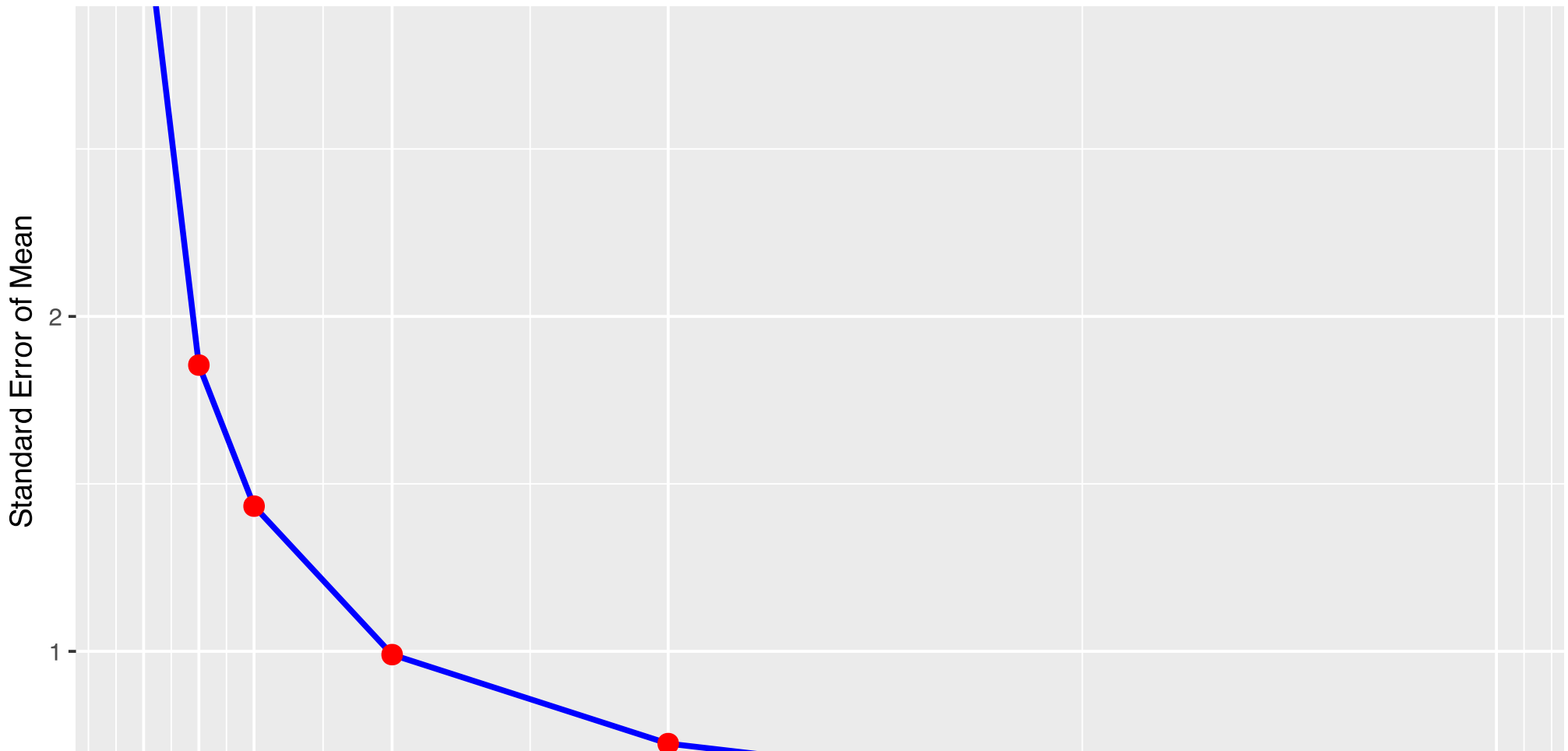
Impact of Sample Size:

- 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

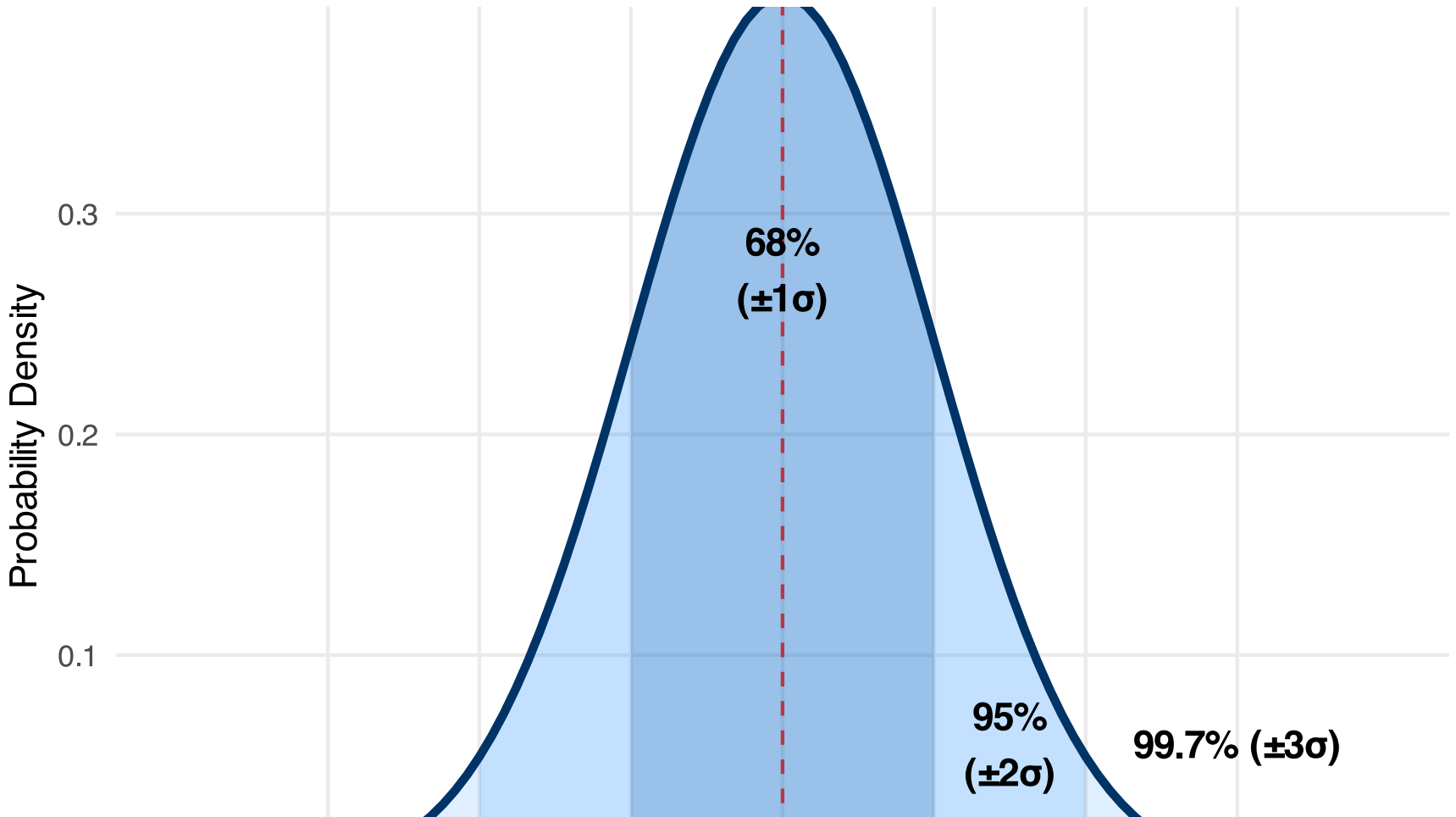
normally distributed. **This is a powerful result that allows us to make inferences about population parameters based on sample statistics.**

The Central Limit Theorem

Normal Distribution:

- Bell-shaped curve
- Defined by mean (μ) and SD (σ)
- 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.

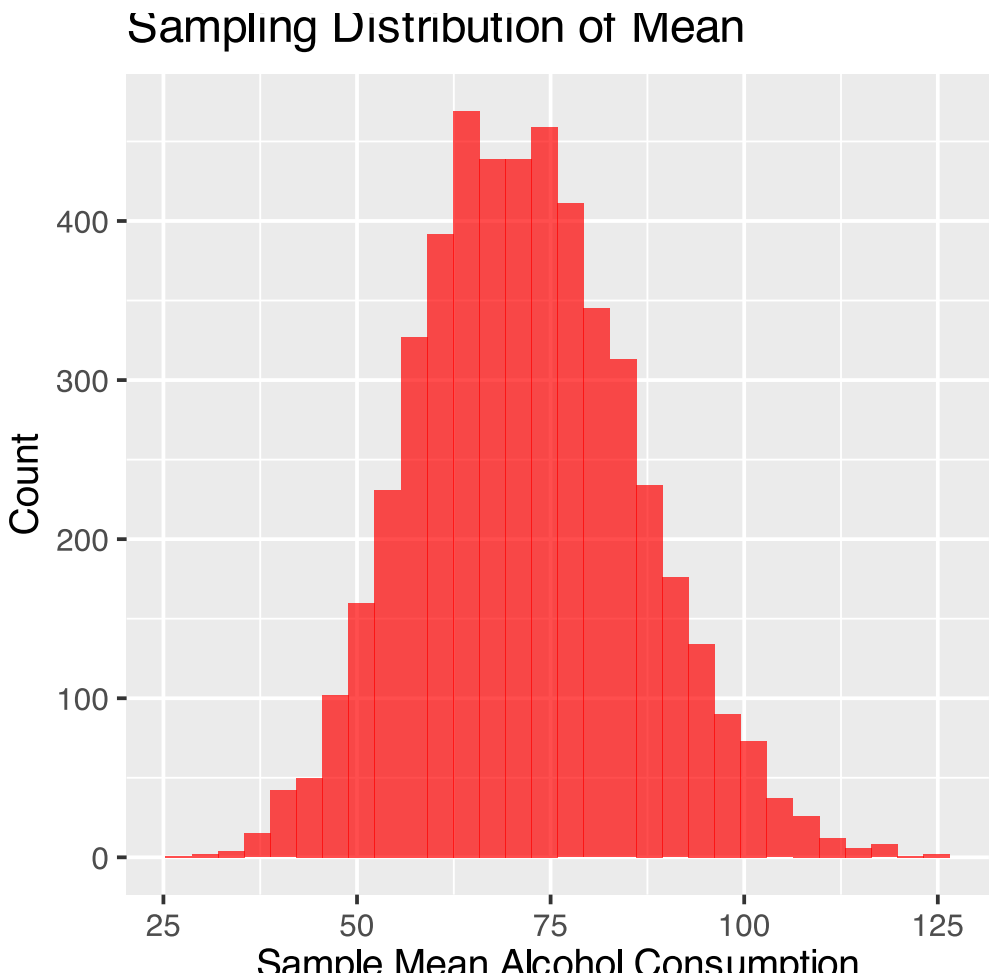
The normal distribution is commonly observed in data collected in the real world – and the central limit theorem gives us some insight into why that occurs. For example, the height of any adult depends on a complex mixture of their genetics and experience; even if those individual contributions may not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

```
# 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