Introduction to Probability

Lecture 3

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.2     v tibble    3.3.0
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.1.0
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
```

1 Outline

- Views of Probability
- Elementary Properties of Probability
- Calculating Probabilities
- Bayes' Theorem
- Screening Tests

2 Views of Probability

2.1 Objective Probability

Probability was thought of by statisticians and mathematicians only as an objective phenomenon derived from objective processes.

Note

There are two categories of objective probability:

- classical/a priori probability
- frequentist/a posteriori probability

2.2 Classical Probability

Probabilities for random events related to games of chance can be calculated by the processes of abstract reasoning. Therefore, it is not necessary for these events to happen to compute these probabilities.

Classical/A Priori Probability

Probability is defined as follows: If an event can occur in N mutually exclusive and equally likely ways, and if m of these possess a trait E, the probability of the occurrence of E is equal to m/N.

$$P(E) = m/N$$

Note

As an example, the probability of getting a three after tossing a **fair** six-sided die can be calculated by assuming that each of the six sides is equally likely to be observed.
The number of outcomes that include a three is 1 out of a total of 6 mutually exclusive outcomes.
Hence, the probability is 1/6.

2.3 Frequentist Probability

The frequentist approach to probability depends on the repeatability of some process and the ability to count the number of repetitions, as well as the number of times that some event of interest occurs. The probability describes the long-run relative frequency of occurrences from the process.

Frequentist/A Posteriori Probability

Probability is defined as follows: If some process is repeated a large number of times, n, and if some resulting event with the characteristic E occurs m times, the relative frequency of occurrence of E, m/n, will be *approximately equal* to the probability of E.

$$P(E) = m/n$$

Important

Note that m/n is only an estimate of the probability, and that n should be large to have a better estimate of $P(E)$.
Example: Monty-Hall Problem

2.4 Subjective Probability

This view holds that probability measures the confidence that a particular individual has in the truth of a particular proposition.

Note

The most famous example of subjective probability is *Bayesian*, which is a mathematically formal method to combine experimental data and expert information in producing probabilities.

3 Properties of Probability

3.1 Helpful Definitions

Sample Space

A sample space is a list of all possible outcomes that might be observed.

Event

An event is any collection of outcomes in the sample space.

Mutually Exclusive

If two events E_i and E_j contain no elements in common, these two events are said to be disjoint or mutually exclusive.

3.2 Axioms of Probability

The properties of probability was mathematically formalized by Russian mathematician A.N. Kolgomorov by defining the following axioms:

1. For each event E_i, the probability of E_i is non-negative.
2. The sum of the probabilities of the mutually exclusive events is equal to 1.
3. Consider any two mutually exclusive events, E_i and E_j. The probability of occurrence of either E_i or E_j is equal to the sum of their individual probabilities.

4 Calculating Probabilities

4.1 Calculating Marginal Probabilities

We can calculate the probability of occurrence for events using the sample point method.

Sample Point Method

Under the assumption that every outcome in the sample space is equally likely to occur, the probability that some event E_i occurs is defined as the number of outcomes corresponding to E_i divided by the total number of outcomes in the sample space.
In frequency distributions, this corresponds to the relative frequency of a specific outcome.

Important

Probabilities computed using this method are often referred to as **marginal probability**. Marginal probabilities are unconditional and are specific to an event.

4.2 Example

Out of 1,325 university students, 532 of them reported to consume caffeine through coffee. What is the marginal probability of selecting a university student who consumes coffee?

$$P(E) = \frac{532}{1325}$$

The resulting probability is 0.402.

4.3 Example 2

Consider the `AMSSurvey` data in the package `carData`. The data set includes the counts of new PhDs in the mathematical sciences for 2008-09 and 2011-12 categorized by type of institution, gender, and US citizenship status. You can learn more about the data set by typing `?carData::AMSsurvey` after installing the `car` package.

4.3.1 Question + Data

What is the probability of selecting a participant of the survey at random who is not a US citizen in 2008? Use the `count` variable.

```
library(carData)
AMSsurvey
```

```
    type    sex citizen count count11
1  I(Pu)   Male      US   132     148
2  I(Pu) Female      US    35      40
3  I(Pr)   Male      US    87      63
4  I(Pr) Female      US    20      22
5     II   Male      US    96     161
6     II Female      US    47      53
7    III   Male      US    47      71
8    III Female      US    32      28
9     IV   Male      US    71      89
10    IV Female      US    54      55
11    Va   Male      US    34      42
12    Va Female      US    14      21
13 I(Pu)   Male  Non-US   130     136
14 I(Pu) Female  Non-US    29      32
15 I(Pr)   Male  Non-US    79      82
16 I(Pr) Female  Non-US    25      26
17    II   Male  Non-US    89     116
18    II Female  Non-US    50      56
19   III   Male  Non-US    53      61
```

```
20   III Female  Non-US    39      30
21    IV   Male  Non-US   122     153
22    IV Female  Non-US   105     115
23    Va   Male  Non-US    28      27
24    Va Female  Non-US    12      17
```

4.3.2 Answer

The total counts can be calculated using the `sum` function. Or simple addition.

```
sum(AMSsurvey$count)
```

```
[1] 1430
```

We can also count the total number of non-US citizens in the data set.

```
130+29+79+25+89+50+53+39+122+105+28+12
```

```
[1] 761
```

Thus, the probability can be calculated as 761/ 1430 = 0.532

4.4 Exercise

The "COVID-19 Effects on the Mental and Physical Health of Asian Americans & Pacific Islanders Survey Study II" (COMPASS II) data set is from a follow-up survey implemented from 12/2021 to 05/2022. The data set is composed of responses from 3,411 participants. (Do et al. 2025)

4.4.1 Exercise

The table below shows the breakdown of the census region reported by the participants. What is the probability that a randomly selected participant was from the Northeast census region?

Census Region	Frequency
Midwest	289
Northeast	344
South	472

Census Region	Frequency
West	2,303
Missing	3

4.4.2 Answer

The marginal probability is $344/3411 = 0.101$

4.5 Union of Events

Union means "together": the union of events A and B consists of all outcomes that are either in event A, event B, or both.

> **Note**
>
> The union of events is mathematically denoted by $\cup$ or "OR".
> As an example, consider Event A: drawing an ace from a standard deck of cards and Event B: drawing a diamond from a standard deck of cards. The union of events A and B can be interpreted as "the event where we draw an ace or a diamond card" or in mathematical terms $Ace \cup Diamond$.

4.6 Intersection of Events

The intersection of events A and B consists of all outcomes that are in event A **AND** event B.

> **Note**
>
> The intersection of events is mathematically denoted by $\cap$ or "AND".
> As an example, consider Event A: drawing an ace from a standard deck of cards and Event B: drawing a diamond from a standard deck of cards. The intersection of events A and B can be interpreted as "the event where we draw an ace AND a diamond card" or in mathematical terms $Ace \cap Diamond$. This simplifies to drawing of a specific card: the ace of diamonds.

4.7 Union vs. Intersection

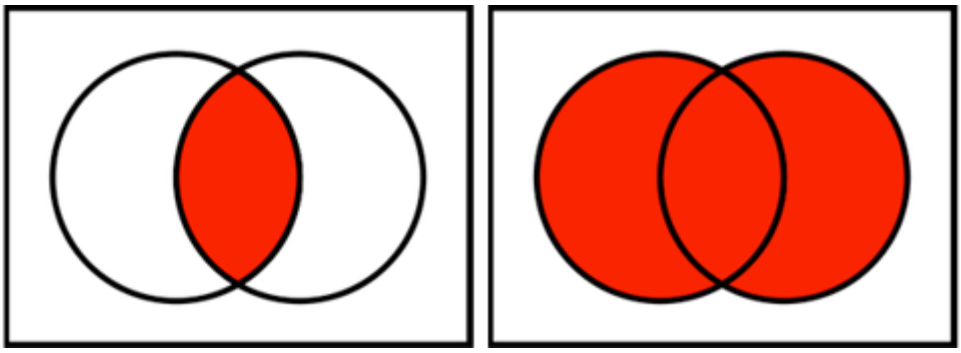

Figure 1: Or vs. And

4.8 Additive Law of Probability

For any two events A and B, the probability of A or B occurring is equal to:

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

> **!** Important
>
> Disjoint events cannot happen at the same time. For disjoint events C and D, $P(C \cap D) = 0$. Therefore, $P(C \cup D) = P(C) + P(D)$ is consistent with Axiom 3.

4.9 Example

What is the probability of rolling a 1 or 3 on a single roll of a fair six-sided die?

> **!** Important
>
> Rolling a 1 and a 3 is not possible. Thus, $P(roll1 \cap roll3) = 0$.

$$P(roll1 \cup roll3) = P(roll1) + P(roll3) - P(roll1 \cap roll3)$$
$$P(roll1 \cup roll3) = 1/6 + 1/6 - 0 = 1/3$$

Thus, $P(roll1 \cup roll3) = 0.3333$

4.10 Example 2

Consider a standard deck of playing cards. What is the probability of drawing an ace or a diamond card?

> **!** Important
>
> Drawing an ace AND a diamond card is possible. This event corresponds to drawing the ace of diamonds. The probability of drawing the ace of diamonds from a standard deck of playing cards is $1/52$. Hence, $P(A \cap D) = 1/52$.

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

$$P(Ace \cup Diamond) = 4/52 + 13/52 - 1/52 = 16/52$$

Thus, $P(Ace \cup Diamond) = 0.3077$

4.11 Exercise

From a sample of 30 adults, nine reported cannabis use, eight reported cigarette use, and two reported both cigarette and cannabis use.

4.11.1 Question

What is the probability that a randomly selected adult from the sample reported cannabis or cigarette use?

4.11.2 Answer

$P(Cannabis) = 9/30$, $P(Cigarette) = 8/30$, $P(Cannabis \cap Cigarettes) = 2/30$. Therefore,

$$P(Cannabis \cup Cigarettes) = P(Can) + P(Cig) - P(Can \cap Cig)$$

$$P(Cannabis \cup Cigarettes) = 9/30 + 8/30 - 2/30 = 15/30$$

$P(Cannabis \cup Cigarettes) = 0.5$

4.12 Independent Events

Two events are independent if knowing the outcome for one of the events provides no information about the outcome of the other event. Mathematically,

> **Independent Events**
>
> Two events A and B are independent if $P(A \cap B) = P(A)P(B)$.

> **Important**
>
> The probability $P(A \cap B)$ is also referred to as the **joint probability** of A and B.

4.13 Example

What is the probability of getting two heads from two independent coin tosses?

$$P(H1 \cap H2) = P(H1)P(H2) = (1/2)(1/2)$$

$P(H1 \cap H2) = 0.25$

4.14 Exercise

According to the National Down Syndrome Society, the probability that a 35-year-old woman conceives a child with Down syndrome is approximately 0.28%.

4.14.1 Exercise

What is the probability that two independently sampled 35-year-old women will both conceive children with Down syndrome?

4.14.2 Answer

$$P(C1Down \cap C2Down) = P(C1Down)P(C2Down) = (0.0028)(0.0028)$$

$P(C1Down \cap C2Down) = 0.000784\%$

4.15 Complementary Events

The **complement** of an event A is defined to be the set of all elements in the sample space that are **NOT** A.

> **Tip**
>
> The complement of A is denoted by $\bar{A}$ or A^C.

> **Note**
>
> For any event A and its complement A^C,
>
> $$P(A^C) = 1 - P(A)$$

4.16 Example

What is the probability of NOT getting a 1 after rolling a fair six-sided die?

$$P([roll1]^C) = 1 - P(roll1) = 1 - \frac{1}{6} = \frac{5}{6}$$

4.17 Exercise

In a survey study, 10.7% reported to have taken undergraduate courses, 23.4% reported to have finished an undergraduate (Bachelor's/Associates) degree, 15.2% reported to have taken postgraduate credits, and 3.9% reported to have finished a graduate degree.

4.17.1 Question

What is the probability of randomly sampling a participant who had not received any college-level education?

4.17.2 Answer

Educational attainment are assumed to be disjoint because respondents cannot be in multiple levels. Hence,

$$P(College) = 0.107 + 0.234 + 0.152 + 0.039$$

The complement is what we're interested in.

$$P(NoCollege) = 1 - (0.107 + 0.234 + 0.152 + 0.039)$$

The resulting probability is 0.468

4.18 Conditional Probability

If two events are dependent, then knowing the outcome of one event provides information about the probability of the other event.

> **Note**
>
> In screening tests, the probability of a positive test depends on whether the subject has the condition or not.

> **Tip**
>
> The notation for conditional probabilities is $P(A|B)$, which is read as the "probability of A given B".

> **Important**
>
> Conditional probabilities are order-specific, i.e. $P(A|B) \neq P(B|A)$.

4.19 Example

Consider the `AMSSurvey` data in the package `carData`. The data set includes the counts of new PhDs in the mathematical sciences for 2008-09 and 2011-12 categorized by type of institution, gender, and US citizenship status.

4.19.1 Question + Data

What is the probability that a randomly selected participant is enrolled in a statistics/biostatistics program(`type=IV`) in 2008 (`count` variable) given that they are known to be a US citizen?

```
library(carData)
AMSsurvey
```

```
    type    sex citizen count count11
1  I(Pu)   Male      US   132     148
2  I(Pu) Female      US    35      40
3  I(Pr)   Male      US    87      63
4  I(Pr) Female      US    20      22
5     II   Male      US    96     161
6     II Female      US    47      53
7    III   Male      US    47      71
8    III Female      US    32      28
9     IV   Male      US    71      89
10    IV Female      US    54      55
11    Va   Male      US    34      42
12    Va Female      US    14      21
13 I(Pu)   Male  Non-US   130     136
14 I(Pu) Female  Non-US    29      32
15 I(Pr)   Male  Non-US    79      82
16 I(Pr) Female  Non-US    25      26
17    II   Male  Non-US    89     116
18    II Female  Non-US    50      56
19   III   Male  Non-US    53      61
20   III Female  Non-US    39      30
21    IV   Male  Non-US   122     153
22    IV Female  Non-US   105     115
23    Va   Male  Non-US    28      27
24    Va Female  Non-US    12      17
```

4.19.2 Answer

In total, there are 669 US citizens in the data set. Among them, there are 71+54 = 125 enrolled in statistics/biostatistics program (`type=IV`).

Therefore, the probability that a participant who is a known US citizen is enrolled in a statistics/biostatistics program is $\frac{71+54}{132+35+87+20+96+47+47+32+71+54+34+14} = 0.186846$

4.20 Multiplicative Law of Probability

Let A and B be events, and suppose $P(B) \neq 0$. Then, the joint, marginal, and conditional probabilities can be related by the following equation.

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

Using basic algebra, we can derive an expression for the conditional probability.

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

4.21 Example

Suppose the probability of traffic light failure is 0.0002. A local transportation office calculated that when there is a traffic light failure, there is a 20% chance of a vehicular accident occurring. What is the probability of a vehicular accident and a traffic light failure occurring at the same time?

- $P(Fail) = 0.0002$
- $P(Acc|Fail) = 0.20$

$$P(Acc \cap Fail) = P(Acc|Fail)P(Fail)$$

$P(Acc \cap Fail) = 0.0002 * 0.20 = 0.00004$

4.22 Example 2

Consider the `AMSSurvey` data in the package `carData`. The data set includes the counts of new PhDs in the mathematical sciences for 2008-09 and 2011-12 categorized by type of institution, gender, and US citizenship status.

4.22.1 Question + Data

Use the multiplication rule to calculate the probability that a randomly selected participant is enrolled in a statistics/biostatistics program(`type=IV`) in 2008 (`count` variable) given that they are known to be a US citizen.

```
library(carData)
AMSsurvey
```

```
    type    sex citizen count count11
1  I(Pu)   Male      US   132     148
2  I(Pu) Female      US    35      40
3  I(Pr)   Male      US    87      63
4  I(Pr) Female      US    20      22
5     II   Male      US    96     161
```

```
6     II Female      US    47      53
7    III   Male      US    47      71
8    III Female      US    32      28
9     IV   Male      US    71      89
10    IV Female      US    54      55
11    Va   Male      US    34      42
12    Va Female      US    14      21
13 I(Pu)   Male  Non-US   130     136
14 I(Pu) Female  Non-US    29      32
15 I(Pr)   Male  Non-US    79      82
16 I(Pr) Female  Non-US    25      26
17    II   Male  Non-US    89     116
18    II Female  Non-US    50      56
19   III   Male  Non-US    53      61
20   III Female  Non-US    39      30
21    IV   Male  Non-US   122     153
22    IV Female  Non-US   105     115
23    Va   Male  Non-US    28      27
24    Va Female  Non-US    12      17
```

4.22.2 Answer

The marginal probability of selecting a US citizen is $\frac{669}{669+761} = 0.4678$.

The probability of selecting a US citizen and `type=IV` is $\frac{(71+54)}{(669+761)} = 0.0874$

Therefore, the probability of selecting a participant in `type=IV` given that they are a US citizen is:

$$P(IV|US) = \frac{P(IV \cap US)}{P(US)} = \frac{\frac{(71+54)}{(669+761)}}{\frac{669}{669+761}}$$

Resulting in the following value: 0.1868, which was the same value as in the previous example.

4.23 Exercise

In 2016, the global incidence rate of the Zika virus is 0.0017. Suppose a screening test was developed such that the probability of a positive test result given that the participant had the Zika virus is 0.90.

4.23.1 Question

What is the probability that a randomly selected individual tested positive and had Zika virus?

4.23.2 Answer

- $P(Z) = 0.0017$
- $P(PT|Z) = 0.90$
- $P(PT \cap Z) = P(PT|Z)P(Z) = (0.90)(0.0017) = 0.0015$

4.24 Law of Total Probability

The law of probability states that for events $B_1, B_2, ..., B_n$ that are disjoint outcomes that span the sample space, and with $P(B_j) \neq 0$ for all j, then for any event A,

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + ... = \sum_{i=1}^{n} P(A|B_i)P(B_i)$$

5 Bayes' Theorem

5.1 Bayes' Theorem

Recall that $P(B|A) \neq P(A|B)$. Bayes' theorem provides a way to invert conditional probabilities.

Bayes' Theorem

Let A and $B_1, B_2, ..., B_n$ be disjoint events that span the whole sample space, and with $P(B_j) \neq 0$ for all j,

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)}$$

Tip

If we know the marginal probability of A,

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{P(A)}$$

6 Screening Tests

6.1 Screening Tests

When screening for diseases, we are interested in the true positive (positive tests for people who have the disease) and true negative rates (negative tests for people who do not have the disease).

	AN (+)	AN(-)
Test (+)	TP	FP
Test (-)	FN	TN

6.1.1 Sensitivity

Sensitivity

The probability that a person randomly selected from the population tests positive given that they have the disease is called the *sensitivity* of a test.

$$Sens = P(PT|D) = \frac{TP}{TP + FN}$$

6.1.2 Specificity

Specificity

The probability that a person randomly selected from the population tests negative given that they do not have the disease is called the *specificity* of a test.

$$Spec = P(NT|ND) = 1 - P(PT|ND) = \frac{TN}{TN + FP}$$

6.1.3 Base Rate

Base Rate

The probability that a person randomly selected from the population has the disease is called the *base rate* of a test.

6.2 Positive Predictive Value

We are also interested in the conditional probabilities of having the disease based on the test for a random member of the population.

Positive Predictive Value (PPV)

The probability of having the disease given a positive test (P(D|PT)) is called the positive predictive value (PPV).

$$P(D|PT) = \frac{P(PT|D)P(D)}{P(PT)} = \frac{sens * BR}{sens * BR + (1 - spec) * (1 - BR)}$$

Warning

The positive predictive value can be calculated using a contingency table based on the results of the screening tests, but this might not reflect the true predictive value for a randomly selected member of the population.

6.3 Negative Predictive Value

Negative Predictive Value (NPV)

The probability of not having the disease given a negative test (P(ND|NT)) is called the negative predictive value (NPV).

$$P(ND|NT) = \frac{P(NT|ND)P(ND)}{P(NT)} = \frac{spec * (1 - BR)}{spec * (1 - BR) + (1 - sens) * (BR)}$$

> **Warning**
>
> The negative predictive value can be calculated using a contingency table based on the results of the screening tests, but this might not reflect the true predictive value for a randomly selected member of the population.

6.4 Example

Suppose a screening test designed for virus X has a sensitivity of 0.64 and a specificity of 0.98.

6.4.1 Question

If the base rate of virus X is 0.002,

- What is the negative predictive value of the test?
- What is the positive predictive value of the test?

6.4.2 Answer

- NPV

```
sens <- 0.64
spec <- 0.98
br <- 0.002

(spec*(1-br))/(spec*(1-br) + (1-sens)*br)
```

```
[1] 0.9992644
```

- PPV

```
(sens*br)/(sens*br + (1-spec)*(1-br))
```

```
[1] 0.06026365
```

A negative test would most likely mean the subject does not have the disease, but a positive test will most likely need other confirmatory tests to confirm if the subject has the disease.

6.5 Example

Suppose a screening test designed for virus X has a sensitivity of 0.64 and a specificity of 0.98.

6.5.1 Question

If the base rate of virus X is 0.002,

- What is the negative predictive value of the test?
- What is the positive predictive value of the test?

6.5.2 Answer

- NPV

```
sens <- 0.64
spec <- 0.98
br <- 0.002

(spec*(1-br))/(spec*(1-br) + (1-sens)*br)
```

```
[1] 0.9992644
```

- PPV

```
(sens*br)/(sens*br + (1-spec)*(1-br))
```

```
[1] 0.06026365
```

A negative test would most likely mean the subject does not have the disease, but a positive test will most likely need other confirmatory tests to confirm if the subject has the disease.

6.6 Exercise

Consider a screening instrument developed to detect symptoms of anorexia nervosa in adolescents. The instrument was validated against the Eating Attitudes Test (EAT) results. The results are shown below.

	AN (+)	AN(-)
Test (+)	250	32
Test (-)	68	784

6.6.1 Question

Assuming the prevalence rate of anorexia nervosa in adolescents is 7%,

- What is the sensitivity of the test?
- What is the specificity of the test?
- What is the probability that a randomly screened adolescent has anorexia nervosa given they tested positive?
- What is the probability that a randomly screened adolescent does not have anorexia nervosa given they tested negative?

6.6.2 Answer

- What is the sensitivity of the test?

```
250/(250+68)
```

```
[1] 0.7861635
```

- What is the specificity of the test?

```
784/(784+32)
```

```
[1] 0.9607843
```

- What is the probability that a randomly screened adolescent has anorexia nervosa given they tested positive? → PPV

```
sens <- 250/(250+68)
spec <- 784/(784+32)
br <- 0.07

(sens*br)/(sens*br + (1-spec)*(1-br))
```

```
[1] 0.6014232
```

- What is the probability that a randomly screened adolescent does not have anorexia nervosa given they tested negative? → NPV

```
(spec*(1-br))/(spec*(1-br) + (1-sens)*br)
```

```
[1] 0.9835238
```

A positive test has a moderate likelihood that the subject has the disease, and a negative test has a high likelihood that the subject does not have the disease.

6.7 Notes on Screening Tests

> **Important**
>
> There are a lot of available functions in `R` that calculate sensitivity, specificity, NPV, and PPV. However, you need to be careful about interpreting the PPV and NPV values provided. Sometimes, these values do not account for the base rate.

> **Tip**
>
> Packages like `yardstick` and `epiR` calculate screening test characteristics and relevant confidence intervals.

References

Do, Vuong Van, Van My Ta Park, Nhung Nguyen, Pamela May Ling, Marian Tzuang, Bora Nam, Marcelle M. Dougan, Oanh L. Meyer, and Janice Y. Tsoh. 2025. "Use of Cigarettes, Cannabis, and Alcohol Among Asian American, Native Hawaiian, and Pacific Islander Adults: Community-Based National Survey Analysis." *JMIR Public Health and Surveillance* 11 (1): e76465. https://doi.org/10.2196/76465.