**Chapter 1 – Analyzing a binary response, part 1**

What are categorical (qualitative) variables?

* Field goal result – success or failure
* Disease infection – yes or no
* Criminal offense convictions – murder, robbery, assault, …
* Highest attained education level – HS, BS, MS, PhD (ordinal properties)
* Food for breakfast – cereal, bagel, eggs,…
* Annual income – <25,000, 25,000 to <50,000,
50,000 to <100,000, ≥ 100,000 (ordinal properties)

We live in a categorical world!

Chapters 1 and 2 focus on binary response categorical variables.

Our goal in Section 1.1 is to estimate the overall probability of observing one of two possible outcomes for this random variable.

* This is often equated with the “probability of success” for an individual item in the population.
* Equivalently, this is the overall prevalence of successes in the population because each item has the same probability of success.

Future sections will extend this to a heterogeneous setting where individual items come from one of two categories. Chapter 2 completes the extension to a heterogeneous population where we use a regression model to estimate the binary response probability.

**Section 1.1.1 – Bernoulli and binomial probability distributions**

Suppose Y = 1 is a success where the probability of a success is P(Y = 1) = π. Also, suppose Y = 0 is a failure. Our goal will be to estimate π. We need to understand first the probability mechanisms behind observing values of Y.

Bernoulli probability mass function (PMF)

P(Y = y) =  for y = 0 or 1

Notice that P(Y = 1) = π and P(Y = 0) = 1 – π

Often, you observe multiple success/failure observations. Let Y1, …, Yn denote random variables for these observations. If the random variables are independent and have the same probability of success π, then we can use a binomial PMF for .

Binomial PMF

P(W = w) =  for w = 0, 1, …, n

Notes:

*  = n choose w
* W is a random variable denoting the number of “successes” out of n trials
* W has a fixed number of possibilities – 0, 1, …, n
* n is a fixed constant
* π is a parameter denoting the probability of a “success” with values between 0 and 1.

Question: Why examine probability distributions?

They can be used to help model real life events. Remember we are making ASSUMPTIONS about the population. Rarely (if ever) will these assumptions be totally satisfied! Often, these assumptions will be satisfied "close enough" to justify their use.

Example: Field goal kicking

Suppose a field goal kicker attempts 5 field goals during a game and each field goal has the same probability of being successful (the kick is made). Also, assume each field goal is attempted under similar conditions; i.e., distance, weather, surface,….

Below are the characteristics that must be satisfied in order for the binomial distribution to be used.

1. There are n identical trials.

n = 5 field goals attempted under the exact same conditions

1. Two possible outcomes of a trial. These are typically referred to as a success or failure.

Each field goal can be made (success) or missed (failure)

1. The trials are independent of each other.

The result of one field goal does not affect the result of another field goal.

1. The probability of success, denoted by π, remains constant for each trial. The probability of a failure is 1-π.

Suppose the probability a field goal is good is 0.6; i.e., P(success) = π = 0.6.

1. The random variable, W, represents the number of successes.

Let W = number of field goals that are good. Thus, W can be 0, 1, 2, 3, 4, or 5.

Because these 5 items are satisfied, the binomial probability mass function can be used and W is called a binomial random variable.

Mean and variance for Binomial random variable

E(W) = nπ

Var(W) = nπ(1-π)

Proofs would be covered in a mathematical statistics course.

Example: Field goal kicking (Binomial.R)

Suppose π = 0.6, n = 5. What are the probabilities for each possible value of w?

P(W = 0) = 

=  ≈ 0.0102

For w = 0, …, 5:

|  |  |
| --- | --- |
| w | P(W = w) |
| 0 | 0.0102 |
| 1 | 0.0768 |
| 2 | 0.2304 |
| 3 | 0.3456 |
| 4 | 0.2592 |
| 5 | 0.0778 |

E(W) = nπ = 5×0.6 = 3 and

Var(W) = nπ(1-π) = 5×0.6×(1-0.6) = 1.2

R code and output:

> dbinom(x = 1, size = 5, prob = 0.6)

[1] 0.0768

> dbinom(x = 0:5, size = 5, prob = 0.6)

[1] 0.01024 0.07680 0.23040 0.34560 0.25920 0.07776

> pmf <- dbinom(x = 0:5, size = 5, prob = 0.6)

> save <- data.frame(w = 0:5, prob = round(x = pmf, digits = 4))

> save

 w prob

1 0 0.0102

2 1 0.0768

3 2 0.2304

4 3 0.3456

5 4 0.2592

6 5 0.0778

> #While not necessary, a new plotting window can be opened

> dev.new(width = 6, height = 6, pointsize = 12)

> plot(x = save$w, y = save$prob, type = "h", xlab = "w", ylab = "P(W=w)", main = "Plot of a binomial PMF for n=5, pi=0.6", panel.first = grid(col = "gray", lty = "dotted"), lwd = 3)

> abline(h = 0)

 

Example: Simulating observations characterized by a binomial PMF (Binomial.R)

The purpose of this example is to show how one can “simulate” observing a random sample of observations from a population characterized by a binomial distribution.

Why would someone want to do this?

Many statistical procedures are based upon particular probability distribution assumptions. We can evaluate how well the procedures do, by simulating data under the correct assumption.

Use the rbinom() function in R.

> # Generate observations from a Binomial distribution

> set.seed(4848)

> bin5 <- rbinom(n = 1000, size = 5, prob = 0.6)

> head(bin5)

[1] 3 2 4 1 3 1

> bin5[1:10]

[1] 3 2 4 1 3 1 3 3 3 4

> mean(bin5)

[1] 2.991

> var(bin5)

[1] 1.236155

> table(x = bin5)

x

0 1 2 3 4 5

12 84 215 362 244 83

> hist(x = bin5, main = "Binomial with n=5, pi=0.6, 1000

 bin. observations", probability = TRUE, breaks = -

 0.5:5.5, ylab = "Relative frequency")

> -0.5:5.5

[1] -0.5 0.5 1.5 2.5 3.5 4.5 5.5



Notes:

* The shape of the histogram looks similar to the shape of the actual binomial distribution.
* The mean and variance are close to what we expect them to be!