W271 Summer 2022 Lecture Video Question Solutions Week 1

Contents

Week 1 Discrete Response Model Part 1 **1**

1.2 Introduction to Categorical Data, and the Bernoulli and Binomial Probability Models 1

1.3 Computing Probabilities of Binomial Probability Model 2

1.5 Maximum Likelihood Estimation 3

1.7 Wald Confidence Interval 3

1.11 Formulation of Contingency Table and Confidence Interval of Two Binary Variables 3

1.12 Relative Risk 5

1.13 Odds Ratios 6

Week 1 Discrete Response Model Part 1

1.2 Introduction to Categorical Data, and the Bernoulli and Binomial Probability Models

Q: Derive the mean of the binomial probability distribution.

Solution: Define

$n = \#$ of draws

$\pi =$ probability of success on a draw

$W = \#$ of successes in a sequence of n independent draws from a binomial distribution

$Y_i =$ indicator for whether the ith draw is a success (1 if yes; 0 if no)

Then W follows a discrete distribution where $P(W = k; n, \pi) = \binom{n}{k} \pi^k (1-\pi)^{n-k}$

We can write $W = Y_1 + ... + Y_n$

Then $E[W] = E[Y_1] + ... + E[Y_n] = \pi + ... + \pi = n\pi$

Since $E[Y_i] = 1 * P(Y_i = 1) + 0 * P(Y_i = 0) = 1 * \pi = \pi$ for all i

Q: Derive the variance of the binomial probability distribution.

Solution: Then $Var[W] = n\pi(1-\pi)$

Since $Var[W] = Var[Y_1 + ... + Y_n] = nVar[Y_i]$ due to independent and identical draws i.e $Cov[Y_i, Y_j] = 0$ for $i \neq j$

And $Var[Y_i] = E[Y_i^2] - E[Y_i]^2 = \pi - \pi^2$ where $E[Y_i^2] = 1^2 * \pi - 0^2 * (1-\pi) = \pi$

And using what we found for the mean of the binomial distribution

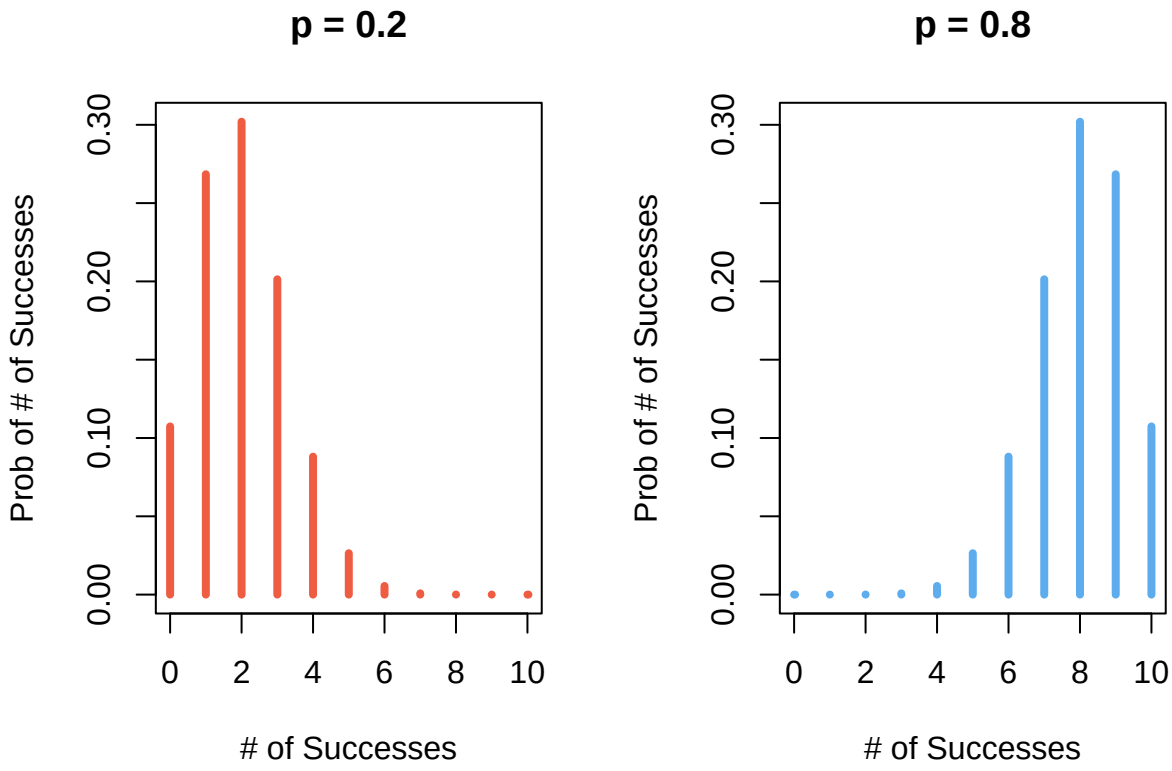
1.3 Computing Probabilities of Binomial Probability Model

Q: Repeat the implementation in R exercise using $\pi = 0.2$, n = 10. What about $\pi = 0.8$, n = 10?

```
par(mfrow=c(1,2))

n<-10; p<-0.2
binom.df1 <- data.frame("x" = 0:n, "p" = dbinom(x = 0:n, size = n, prob = p))
p1 <- plot(binom.df1$x, binom.df1$p,type="h", xlab="# of Successes", ylab = "Prob of # of Successes",
           col = "tomato2", main = paste("p = ", p, sep = ""), lwd = 4)

n<-10; p<-0.8
binom.df2 <- data.frame("x" = 0:n, "p" = dbinom(x = 0:n, size = n, prob = p))
p2 <- plot(binom.df2$x, binom.df2$p,type="h", xlab="# of Successes", ylab = "Prob of # of Successes",
           col = "steelblue2", main = paste("p = ", p, sep = ""), lwd = 4)
```



Solution:

Notice the shift to the right as p increases because the probability of success on any given trial increases, making more successes more likely.

1.5 Maximum Likelihood Estimation

Q: What assumptions are needed for us to derive the following likelihood function?

Solution: To apply the binomial probability model to a situation / set of data, we need the following five conditions to be satisfied:

(1) N identical trials i.e each trial in the data is conducted under the same exact circumstances

(2) Each trial only has two outcomes: success or failure

(3) The trials are independent of each other

(4) The probability of success and failure from trial to trial is constant

(5) Probability of success and failure together sum to one i.e. $P(Failure) = 1 - P(Success)$

If these conditions are satisfied then for a given set of data we can model the likelihood assuming a binomial probability distribution as:

$L(\pi|y_1, ..., y_n) = P(Y_1 = y_1, ..., Y_n = y_n)$

$= P(Y_1 = y_1) * ... * P(Y_n = y_n)$ (3) due to independence of trials

$= \Pi_{i=1}^{n} P(Y_i = y_i)$ (1) due to identical trials

$= \Pi_{i=1}^{n} \pi^w (1-\pi)^{n-w}$ due to the fact that (4) $P(success) = \pi$ i.e. is constant and (5) $P(Failure) = 1 - P(Success)$, assuming in the observed data $y_1, ..., y_n$ we have w successes and (2) each trial has only two outcomes (success and failure)

1.7 Wald Confidence Interval

Q: Interpret the confidence interval below. Please limit the response to one sentence.

Solution: Recall that the general form of a confidence interval for a parameter θ is $\hat{\theta} \pm z_{1-\frac{\alpha}{2}} SE(\hat{\theta})$.

For the example of the binomial probability distribution we have:

$\hat{\pi} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} = 0.4 \pm 1.96 \sqrt{\frac{\hat{0.4}(1-\hat{0.4})}{10}}$

Technically speaking, confidence intervals are related to the long run proportion of similarly computed intervals that will end up containing the true parameter value.

Any given interval either has the true parameter or not.

Practically speaking however, people often describe confidence intervals as containing the true parameter value with the specified probability.

So we can interpret the interval above as saying that we are 95% certain that π i.e. $P(success)$ in the data is almost certainly between $[0.01, 0.7]$.

1.11 Formulation of Contingency Table and Confidence Interval of Two Binary Variables

Q: Use the Larry Bird example and test the hypothesis that the probability of the first free throw made and that of the second free throw made is not different.

Solution: Note that the table shown in the lecture problem does not actually meet the definition of a contingency table. A contingency table for this problem would have free throw

attempt on the rows (group) and outcome success or failure on the columns (response). The lecture video computes the probabilities in the confidence interval incorrectly because it uses the wrong contingency table.

Define $\pi_1 = P(Success\ on\ free\ throw\ 1)$ and $\pi_2 = P(Success\ on\ free\ throw\ 2)$.

We want to test $H_0 : \pi_1 = \pi_2$ vs. $H_A : \pi_1 \neq \pi_2$.

Because of independence of trials and the fact that we assume probability of success on each free throw follows a normal distribution in the limit, we can model the difference $\pi_1 - \pi_2$ using a normal distribution as well i.e $d = \pi_1 - \pi_2 \sim N(0, (\frac{1}{n_1} + \frac{1}{n_2})\pi(1-\pi))$ under H_0 where $\pi = \frac{n_1\pi_1 + n_2\pi_2}{n_1+n_2}$.

Note that this distribution is using the pooled sample variance to test the hypothesis, which means we are effectively assuming equal variances for the probability of success in each free throw attempt. This leads to higher power in the hypothesis test, but it is an additional assumption we are making. Under the null hypothesis of equal proportions, we do have equal variances since for proportions the variance depends directly on π.

We are also using the Wald interval here, but we could repeat the test using alternative forms of variance for the normal distribution such as the Agresti and Caffo adjustment.

```
# incorrect contingency table from lecture
dat1 <- array(data = c(251, 48, 34, 5), dim = c(2, 2),
              dimnames = list(First = c("Made", "Missed"),
                              Second = c("Made", "Missed")))
dat1
```

```
##         Second
## First    Made Missed
##   Made    251     34
##   Missed   48      5
```

```
prob1 <- dat1 / rowSums(dat1)
prob1
```

```
##         Second
## First         Made     Missed
##   Made   0.8807018 0.11929825
##   Missed 0.9056604 0.09433962
```

```
# correct contingency table
dat2 <- array(data = c(285, 299, 53, 39), dim = c(2, 2),
              dimnames = list(FreeThrow = c("First", "Second"),
                              Outcome = c("Made", "Missed")))
dat2
```

```
##          Outcome
## FreeThrow Made Missed
##    First   285     53
##    Second  299     39
```

```
prob2 <- dat2 / rowSums(dat2)
prob2
```

```
##          Outcome
## FreeThrow      Made    Missed
##    First  0.8431953 0.1568047
##    Second 0.8846154 0.1153846
```

```r
# hypothesis test
p1 <- prob2[1,1]
p2 <- prob2[2,1]
n1 <- unname(rowSums(dat2)[1])
n2 <- unname(rowSums(dat2)[2])
p <- (n1 * p1 + n2 * p2) / (n1 + n2)
alpha <- 0.05

test.stat <- abs((p1 - p2) / sqrt( (1/n1 + 1/n2) * p * (1 - p)))
print("test statistic"); test.stat
```

```
## [1] "test statistic"
```

```
## [1] 1.570367
```

```r
print("p-value"); pnorm(test.stat, lower.tail = F)
```

```
## [1] "p-value"
```

```
## [1] 0.05816493
```

Since the p-value of the test statistic is above $\alpha = 0.05$, we fail to reject the null hypothesis. There is not sufficient statistical evidence to conclude that the probability of Larry Bird making his first free throw is different than the probability of him making his second free throw.

1.12 Relative Risk

Q: What does a relative risk of 1 mean? What does a relative risk of 0.2 mean?

Solution: Relative risk or RR is defined as $\frac{\pi_1}{\pi_2}$ or the ratio is group probabilities of success. Group 2 in this case is considered the reference group, and the relative risk represents the relative increase in likehood of a success / outcome vs. failure / no outcome for group 1 over group 2.

Let's assume that for this example that group 1 received a drug and group 2 received a placebo where outcomes are defined as having an adverse reaction or not to match the lecture material.

A relative risk of 1 implies that $\pi_1 = \pi_2$ so that the likelihood of an outcome between the two groups is the same. In the lecture example, this would mean that patients who got the drug were as likely to develop adverse reactions compared to patients who got the placebo.

A relative risk of 0.2 implies that $\pi_1 = 0.2\pi_2$ so that the likelihood of an outcome between the two groups is lower for group 1. In the lecture example, this would mean that patients who got the drug were much less likely (20% times) to develop adverse reactions compared to patients who got the placebo.

Side note, often in experiments at a company, understanding both absolute risk and relative risk are useful to set context. Absolute risk defines the actual value from a feature / decision while relative risk reveals how much you are improving / changing things compared to the status quo.

1.13 Odds Ratios

What is the numerical range of an odds? What does it mean for an odds to be 1?

Solution: Odds is defined as the ratio of probability of success to failure i.e. $Odds = \frac{\pi}{1-\pi}$. You can think of it as the relative risk of success to failure.

Given that $\pi \in [0,1]$ inclusive we have $Odds \in [\frac{0}{1-0}, \frac{1}{1-1}] = [0,\infty)$ i.e. odds is bounded below by zero and strictly positive.

An odds of one means that $Odds = \frac{\pi}{1-\pi} = 1$ or that $\pi = 1-\pi$ so that $P(Success) = P(Failure)$ and that both outcomes are equally likely. The relative risk of a success compared to a failure is one.

What is the numerical range of OR? What does it mean for OR to be 1? What does it mean for OR > 1? What does it mean for OR < 1?

Solution: Odds ratio or OR is the ratio of odds between two groups i.e $OR = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$. It represents the chance of success relative to failure in group 1 compared to group 2.

Because $Odds \in [0,\infty)$, we have $OR \in [0,\infty]$ as well.

An odds ratio of 1 means that $OR = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = 1$ so that $\pi_1/(1-\pi_1) = \pi_2/(1-\pi_2)$ or that the relative chance of success in group 1 is the same as in group 2. The odds in both groups are same.

An odds ratio above 1 means that $OR = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} > 1$ so that $\pi_1/(1-\pi_1) > \pi_2/(1-\pi_2)$ or that the relative chance of success in group 1 is higher than in group 2. The odds in group 1 are greater than the odds in group 2.

An odds ratio below 1 means that $OR = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} < 1$ so that $\pi_1/(1-\pi_1) < \pi_2/(1-\pi_2)$ or that the relative chance of success in group 1 is lower than in group 2. The odds in group 1 are lower than the odds in group 2.