Chapter 10: Discrete Data Analysis

In Chapters 8, we looked at confidence intervals and hypothesis testing, which are ways to determine the accuracy of a point estimate of a parameter. In this chapter, we'll continue this goal, but applied specifically to some discrete probability distributions, which we studied in Chapter 3. In the language of Chapter 6, we note that discrete distributions are much better suited to handle categorical data, in which we sort data into types rather than make measurements, and then count frequencies. There are two discrete distributions that are particularly well-suited for this. The binomial distribution, in which we break everything into two categories, will be the subject of Section 10.1. The multinomial distribution, in which we break our data into several categories, will be the subject of Section 10.3. (We'll skip the other sections in this chapter for time.)

Section 10.1: Inferences on a Population Proportion

In this section, the unknown parameter is the proportion of a population that possess a particular characteristic. Consider a Bernoulli experiment with a success probability "p" and a random sample of n observations obtained from the population. Let X be the random variable representing the number of successes in n observations. Then we know from Chapter 3 that $X \sim Bin(n, p)$. In my notes in Chapter 3, I liked to use the letter k to denote the number of successes, just to stress that this must be an integer and not any real number. However, following the book here, I'll bend and use x to represent the number of successes. That is, x is the number of observations with the characteristic in a random sample of size n, and thus $n - x$ equals the number of observations without the characteristic, or "failures". We want to use this data to make a guess at what the actual probability of success is. That is, we guess what the actual proportion of the population is that has the desired characteristic.

The value $\frac{x}{n}$ is the sample proportion $\hat{p}$. We previously saw that $E(\hat{p}) = p$ and $Var\left(\hat{p}\right) = \frac{p(1-p)}{n}$. Furthermore, by the central limit theorem, we know that as long as x and $n - x$ are reasonably large (let's say both are greater than 5) that we can approximate the distribution of $\hat{p}$ using a normal random variable by $\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$, where the approximation gets better as n increases. Note that converting to a standard normal means that $\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0, 1)$. To find a $1-\alpha$ confidence interval, we solve $-z_{\alpha/2} \leq \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \leq z_{\alpha/2}$ for p to get the two-sided confidence interval. In particular, we get a two-sided $1-\alpha$ level confidence interval of $\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$.

There's nothing really new here in the math, but now that the parameter we're finding a confidence interval for is a probability, we can add in some common sense guidelines. In particular, even though we don't know the value of *p,* we know that it must be between 0 and 1. However, it's possible that the number $\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ could be less than 0, or that $\hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ could be bigger than 1. Since we know that *p* can't actually be less than 0 or bigger than 1, we can just truncate the lower end at 0 or the upper end at 1. In a sense, this is good news, because it decreases the size of the confidence interval in this case. We use the same logic when we look at one-sided intervals. We summarize the information in the following box.

Two-sided Confidence Intervals for a Population Proportion

If the random variable *X* has a *B(n, p)* distribution, then an approximate two-sided 1 – α confidence level confidence interval for the success probability *p* based on an observed value of the random variable *x* is

$$\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

where the estimated success probability is $\hat{p} = \frac{x}{n}$. This confidence interval can also be written as

$$\left(\hat{p} - \frac{z_{\alpha/2}}{n}\sqrt{\frac{x(n-x)}{n}},\ \hat{p} + \frac{z_{\alpha/2}}{n}\sqrt{\frac{x(n-x)}{n}}\right).$$

If a random sample of *n* observations is taken from a population and *x* of the observations are of a certain type, then this expression provides a confidence interval for the proportion *p* of the population of that type.

If it turns out that the upper end of this confidence interval is larger than 1, then we truncate it at 1. If the lower end of the confidence interval is smaller than 0, then we truncate it at 0.

One-sided Confidence Intervals for a Population Proportion

If the random variable *X* has a *B(n, p)* distribution, then approximate one-sided 1 – α confidence level confidence intervals for the success probability *p* based on an observed value *x* of the random variable are

$$p \in \left(\hat{p} - z_{\alpha}\sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}},\ 1\right) \text{ and } p \in \left(0,\ \hat{p} + z_{\alpha}\sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}}\right),$$

where the estimated success probability is $\hat{p} = \frac{x}{n}$. As above these simplify to

$$p \in \left(\hat{p} - \frac{z_{\alpha}}{n}\sqrt{\frac{x(n-x)}{n}},\ 1\right) \text{ and } p \in \left(0,\ \hat{p} + \frac{z_{\alpha}}{n}\sqrt{\frac{x(n-x)}{n}}\right).$$

These confidence intervals provide a lower bound and an upper bound (respectively) on the probability *p*.

In both cases, these confidence intervals are reasonable as long as both *x* and *n* – *x* are more than 5.

Hypothesis Tests

It should come as no surprise that we can translate the math behind these confidence intervals into hypothesis tests, just as we did in moving from Section 8.1 to Section 8.2. The null hypothesis will then be a guess about the population proportion *p*. Using the null hypothesis $H_0 : p = p_0$ will give a two-sided hypothesis test, while using a null hypothesis of the form $H_0 : p \geq p_0$ or $H_0 : p \leq p_0$ will result in one-sided hypothesis tests. Since the confidence intervals for population proportion all involve the standard normal distribution, we expect that the *p*-values for our hypothesis tests also involve the standard normal distribution. However, it's important to remember that the confidence intervals we got are really an estimation using the central limit theorem. We can use the same logic to get estimations for the *p*-values in our hypothesis tests, but we could also compute them exactly, since we know the probability mass function for the binomial distribution quite well by now.

Recall that the *p*-value is defined as the probability of getting the data set or "worse" given that the null hypothesis is true. The data set we got gives us a value of $\hat{p}$ which is either bigger or smaller than p_0. A "worse" data set would be one in which the value of $\hat{p}$ was even farther away from p_0. In other words, if the null hypothesis is true, then the random variable representing the observed number of successes X has the distribution $X \sim B(n, p_o)$. We then would expect to see np_0 successes. However, we let x be the number of successes we do get. Then $P(X = x)$ is the probability that we get exactly this proportion if the null hypothesis is true. But the *p*-value also takes into consideration the chance that the observed number of successes in a data set is even farther away from np_0 than the number in our data set! That is, if $x \geq np_0$, then the *p*-value is $2 \times P(X \geq x)$, and if $x \leq np_0$, then the *p*-value is $2 \times P(X \leq x)$. (Note that we multiply by 2 because this is a two-sided test, so we have to consider error on both sides.)

So, we can compute *p*-values for hypothesis tests this way. However, we've done enough computations with binomial random variables to know that getting exact probabilities that range over many cases is difficult, since there's no closed form of a cumulative distribution function to use. This was one of the motivations behind using a normal random variable to approximate a binomial random variable in the first place! So even though we CAN get exact *p*-values, when *n* is large, we don't want to because it's a pain.

So we'll bring back our normal approximation to the binomial, and we define a *z*-statistic to be $z = \dfrac{x - np_0}{\sqrt{np_0(1 - p_0)}}$. Then for a size α test, we'll accept the null hypothesis if $|z| \leq z_{\alpha/2}$, and we'll reject the null hypothesis if $|z| > z_{\alpha/2}$.

But wait! What about the continuity correction we've used? Well, just as before, it will give a slightly better estimation, particularly when *n* is relatively small. So we can always choose to incorporate that if we want. We'll include the details about that in the following box summarizing this information.

Two-Sided Hypothesis Tests for a Population Proportion

If the random variable X has a $B(n, p)$ distribution, then the p-value for the two-sided hypothesis testing problem $H_0: p = p_0$ versus $H_A: p \neq p_0$ based upon an observed value x of the random variable is computed exactly by assuming X has a $B(n, p_0)$ distribution and using the following:

- If $\hat{p} = \frac{x}{n}$ is bigger than p_0, then the p-value is $2 \times P(X \geq x)$.
- If $\hat{p} = \frac{x}{n}$ is less than p_0 then the p-value is $2 \times P(X \leq x)$.

When np_0 and $n(1 - p_0)$ are both larger than 5 a normal approximation can be used.

In this case the test value is the z-statistic, $z = \dfrac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \dfrac{x - np_0}{\sqrt{np_0(1-p_0)}}$, and the p-value is $2 \times \Phi(-|z|)$ where Φ is the standard normal cumulative distribution function. We can improve the normal approximation with a continuity correction by using a numerator of $x - np_0 - 0.5$ when $x - np_0 > 0.5$ and a numerator of $x - np_0 + 0.5$ when $x - np_0 < -0.5$.

For a test of size α, we accept the null hypothesis when $|z| \leq z_{\alpha/2}$
and reject the null hypothesis when $|z| > z_{\alpha/2}$

There isn't too much to mention in the move to one-sided hypothesis tests, as no new problems occur, so we summarize that information in the box on the next page.

One-Sided Hypothesis Tests for a Population Proportion (left tailed)

If the random variable X has a $B(n, p)$ distribution, then the p-value for the one-sided hypothesis testing problem $H_0 : p \geq p_0$ versus $H_A : p < p_0$ based upon an observed value x of the random variable is usually computed by assuming X has a $B(n, p_0)$ distribution and the p-value is $P(X \leq x)$.

When np_0 and $n(1 - p_0)$ are both larger than 5 a normal approximation can be used. In this case the test value is $z = \dfrac{x - np_0 + 0.5}{\sqrt{np_0(1 - p_0)}}$ and the p-value is $\Phi(z)$.

For a traditional hypothesis test of size α, we accept the null hypothesis when $z \geq -z_\alpha$ and reject the null hypothesis when $z < -z_\alpha$

One-Sided Hypothesis Tests for a Population Proportion (right tailed)

If the random variable X has a $B(n, p)$ distribution, then the p-value for the one-sided hypothesis testing problem $H_0 : p \leq p_0 \ versus \ H_A : p > p_0$ based upon an observed value x of the random variable is usually computed by assuming X has a $B(n, p_0)$ distribution and the p-value is $P(X \geq x)$.

When np_0 and $n(1 - p_0)$ are both larger than 5 a normal approximation can be used. In this case the test value is $z = \dfrac{x - np_0 - 0.5}{\sqrt{np_0(1 - p_0)}}$ and the p-value is $1 - \Phi(z)$.

For a traditional hypothesis test of size α, we accept the null hypothesis when $z \leq z_\alpha$ and reject the null hypothesis when $z > z_\alpha$

Sample Size Calculations for Confidence intervals.

We've already seen a lot of problems ask how large the sample size must be to ensure a sufficiently small confidence interval and a sufficiently high level of confidence. Although this involved a small amount of guesswork for t-intervals, we can take that part of the guesswork out of it here, since our confidence intervals are computed with the standard normal, whose critical points do not depend on n.

In particular, the length of a confidence interval with confidence level $1 - \alpha$ is $L = 2z_{\alpha/2}\sqrt{\dfrac{\hat{p}(1 - \hat{p})}{n}}$.

But then solving this for n, we get $n = \dfrac{4z_{\alpha/2}^2 \hat{p}(1 - \hat{p})}{L^2}$. Note that as we would expect, as the desired length L gets smaller, the size n required of a sample to meet that length gets larger. In addition, as the confidence level increases, the size n also must increase.

There is a bit of a wrinkle with the logic here though: if we're computing n to figure out how big our sample must be, that means we haven't collected the sample yet, so we haven't computed $\hat{p}$ yet, so we can't compute n in the first place! So we have a different sort of guesswork to deal with here, in estimating $\hat{p}$. We could just think about the worst case scenario, since using calculus, we know that the maximum value of $\hat{p}\left(1-\hat{p}\right)$ is (0.5)(0.5) = 1/4. Plugging this in, it gives a bound on a value of n that is guaranteed to ensure we get the desired confidence interval length: $n > \frac{z_{\alpha/2}^2}{L^2}$. However, this bound may be larger than we want to deal with. Therefore, what we really want is a more reasonable maximum for $\hat{p}\left(1-\hat{p}\right)$. If we can clearly see $\hat{p} < p^* < 0.5$ then we may use p^* in the formula. If we can clearly see $\hat{p} > p^* > 0.5$ then we may use this p^* in the formula. For example, if we need find the sample size needed for a small confidence interval, and we're sure that $\hat{p} < 0.4$, then we can use 0.4 in the formula in place of $\hat{p}$. Similarly, if we're sure that $\hat{p} > 0.75$, then we can use 0.75 in the formula in place of $\hat{p}$.

Section 10.3: Goodness of Fit Tests

Everything we did in 10.1 is for splitting our population into two categories, which we can think of as "success" or "failure". But a lot of categorical data has more categories than that. So instead of modeling the situation with a binomial distribution, we want to model with a multinomial distribution.

So what the parameter we're looking to estimate? Let's motivate this by a simple example. Suppose we give a poll with three options, which I'll creatively call Option 1, Option 2, and Option 3. If all we want is to get an estimate for the proportion of the population that favors Option 1, we can sort of lump the other two options together as "failure" and turn it into a binomial, which enables us to use the methods of Section 10.1. Similarly, we could do the same thing for Option 2 by combining Option 1 and Option 3 as "failure", etc. So Section 10.1 techniques will let us get estimates for the proportions for each option INDIVIDUALLY. But what we really want is to analyze the estimates for each option TOGETHER. That is, we're trying to analyze a collection of estimates consisting of an estimate for each proportion that all sum to 1.

To be precise, suppose we have k options or categories, where the kth category represents an actual proportion p_k of the entire population. We look at a collection of estimates $p_1^*, p_2^*, \ldots, p_k^*$ such that $p_1^* + p_2^* + \cdots + p_k^* = 1$. Instead of considering the accuracy of each individual p_i^*, we want to consider the collection together.

In the language of hypothesis tests, our null hypothesis is now a compound statement, of the form $H_0 : p_1 = p_1^*, p_2 = p_2^*, \ldots, p_k = p_k^*$. In practice, we'll use the shorthand $H_0 : p_i = p_i^* \quad 1 \le i \le k$. We test such a hypothesis with a **goodness of fit test**.

Let's set up our strategy. We collect a data set with sample size n. We let x_i represent the number of observations in category i. We are testing the null hypothesis $H_0 : p_i = p_i^* \quad 1 \le i \le k$. As usual, we

want to get some idea of the probability that we would get the data set we got if the null hypothesis were true. But if the null hypothesis were true, then we would expect to see np_1^* observations in category 1, np_2^* observations in category 2, etc. So we define $e_i = np_i^*$, the expected number of observations in category i under the null hypothesis. A goodness of fit test will measure the difference between the observed frequencies and the expected frequencies by computing a **chi-square statistic**. We'll define two different chi-square statistics, denoted X^2 and G^2.

We define $X^2 = \sum_{i=1}^{k} \frac{(x_i - e_i)^2}{e_i}$ and $G^2 = 2\sum_{i=1}^{k} x_i \ln\left(\frac{x_i}{e_i}\right)$. X^2 is known as the "Pearson chi-square statistic" and G^2 is known as the "likelihood ratio chi-square statistic". In both cases, larger values indicate a larger discrepancy, and if $x_i = e_i$ for every i, both of these statistics give us 0.

It is worth noting that these two different chi-square statistics are usually pretty close to each other, so in practice, it won't matter too much which one you use, as the reasons for choosing one or the other are beyond the scope of this course.

The reason these are called "chi-square statistics" is that we will calculate p-values for the hypothesis test by comparing the statistics to a chi-square random variable. However, these test statistics can only reasonably be compared to a chi-square random variable when each expected frequency e_i is at least 5. If we have a category such that $e_i < 5$, then we'll want to consider removing that category and grouping it with another category (or more) to ensure that the expected frequencies are all at least 5.

Our p-value given by these statistics is $P(\chi^2_{k-1} \geq X^2)$ or $P(\chi^2_{k-1} \geq G^2)$ (depending on which statistic we use).

This means that for a size α hypothesis test, we accept the null hypothesis if $X^2 \leq \chi^2_{\alpha,k-1}$ (respectively $G^2 \leq \chi^2_{\alpha,k-1}$) and we reject the null hypothesis if $X^2 > \chi^2_{\alpha,k-1}$ (respectively $G^2 > \chi^2_{\alpha,k-1}$), where $\chi^2_{\alpha,k-1}$ is the critical point of the chi-square distribution with k-1 degrees of freedom associated with α. We summarize this information in the box on the next page.

Goodness of Fit Hypothesis Tests

Consider a multinomial distribution with k categories and a set of unknown underlying probabilities $p_1, p_2, \ldots, p_k$. Based on a set of observed frequencies $x_1, x_2, \ldots, x_k$ with $x_1 + x_2 + \cdots + x_k = n$, we test the null hypothesis $H_0 : p_i = p_i^* \quad 1 \le i \le k$.

Defining $e_i = np_i^*$ for each i, we compute the test chi-square statistic

$$X^2 = \sum_{i=1}^{k} \frac{(x_i - e_i)^2}{e_i} \quad \text{or} \quad G^2 = 2\sum_{i=1}^{k} x_i \ln\left(\frac{x_i}{e_i}\right)$$

and we get p-value of

p-value $= P(\chi^2_{k-1} \ge X^2)$ or p-value $= P(\chi^2_{k-1} \ge G^2)$.

This is appropriate as long as all expected frequencies e_i are larger than 5.

For a test of size α, we accept the null hypothesis when $X^2 \le \chi^2_{\alpha,k-1}$ (or when $G^2 \le \chi^2_{\alpha,k-1}$)
and reject the null hypothesis when $X^2 > \chi^2_{\alpha,k-1}$ (or when $G^2 > \chi^2_{\alpha,k-1}$)

Goodness of Fit for Distributional Assumptions

Lastly, we can use this reasoning on numerical data to determine the plausibility that the data comes from a certain distribution. We break our data into categories based on range, and then compute our expected frequencies e_i using the assumed probability distribution. The book doesn't have too much to say on this except the one example they work out, but it's a pretty nice idea, so just go read Example 3 on pages 474-476.

That's the end of the course material!