**Social Studies 201**
**Notes for November 24, 2003**

## Issues involved in hypothesis testing

These notes for November 24 present more examples of hypothesis tests. In addition, there is a discussion of several issues involved in constructing and interpreting hypothesis tests. These are further discussed in section 9.2.4 through 9.2.9 of Chapter 9 of the text.

## Example – responses of University of Regina undergraduates to attitude questions

This question examines responses to two questions from the *Survey of Student Attitudes and Experiences* conducted in 1998 in Social Studies 306 and available in the file

`ssae.sav`

in the folder

`t:\students\public\201\`

The two questions examined are V1, "Free trade is positive for Canadians." and M5, "The government should fund festivals and special events celebrating different cultures" For each of these two statements, respondents were asked to give their view on a five-point scale, from 1 meaning strongly disagree to 5 meaning strongly disagree. Responses to each of these questions, along with the respective means and standard deviations, are given in Table 1. While the data were obtained using an ordinal five-point scale, in calculating means and standard deviations, we are treating these two variables as if they were measured at an interval level.

**Question**. For each of these two variables, test whether the mean response is on the agree side of a neutral response, that is, test whether the mean exceeds 3. Use the 0.01 level of significance. Assume this sample is a random sample of all University of Regina undergraduates in the Fall 1998 semester.

Table 1: Responses to attitude questions V1 and M5

| Response | Responses to V1 | M5 |
|---|---|---|
| 1 – strongly disagree | 55 | 91 |
| 2 – somewhat disagree | 86 | 151 |
| 3 – neutral | 301 | 199 |
| 4 – somewhat agree | 160 | 150 |
| 5 – strongly agree | 78 | 106 |
| Total | 680 | 697 |
| Mean | 3.176 | 3.042 |
| Standard deviation | 1.056 | 1.250 |

**Answer**

Before conducting the two hypotheses tests, I will explain the reason for being interested in these tests. For each of the two variables, it appears that responses are fairly evenly split between agree and disagree, with the modal response being 3, or neutral, in each case. The sample mean response does not appear very different from the neutral response of 3 for each of these two variables, although the sample mean exceeds 3 in each case. The question thus asks whether these means are sufficiently greater than 3 to argue that respondents, on average, tend to agree with the two statements, or whether there is insufficient evidence to conlude that respondents, on average, can be considered to agree.

The method of conducting each test is more or less the same. In the following notes, all the steps in conducting the first test, for V1, are provided. For the second variable and test, attitude about M5, only those items that differ are discussed.

**Hypothesis test for V1**

Let $\mu$ be the true mean level of opinion among University of Regina undergraduates about issue V1, "free trade is positive for Canadians." The steps involved in conducting the hypothesis test are as follows.

1. **Hypotheses**. Since an hypothesis test must begin with an equality for the null hypothesis, the hypothesis that makes most sense here is $\mu = 3$, that is, that undergraduates on average, have a neutral response. Then the alternative suggested in the question is that the mean may exceed 3, that is, the question asks to test whether the mean exceeds 3. This is an example of a one-tailed test, to test whether $\mu > 3$. The null and alternative hypotheses are

$$\text{Null hypothesis} \quad H_0 : \mu = 3$$

$$\text{Alternative hypothesis} \quad H_1 : \mu > 3$$

   Once the test has been conducted, the conclusion will either be that we do not reject the null hypothesis that $\mu$ is 3, or, if the sample mean is in the critical region, the conclusion will be that $\mu$ exceeds 3, or that the average response is on the agreee side.

2. **Test statistic**. The claim is about $\mu$, the mean of V1 for all undergraduate students. The sample mean, $\bar{X}$, is the test statistic.

3. **Distribution of the test statistic**. The sample is said to be a random sample of U of R undergraduates in the Fall 1998 semester, with a sample size of $n = 680$. This is a large random sample so the central limit theorem can be used. As a result,

$$\bar{X} \text{ is Nor}\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

   The sampling distribution of $\bar{X}$ is normally distributed with mean $\mu$ and standard deviation $s/\sqrt{n}$, where $s$ can be used as an estimate of the population standard deviation $\sigma$, since $n$ is large.

4. **Significance level**. The level of significance requested here is 0.01, so this is $\alpha = 0.01$. Since the alternative hypothesis is that $\mu > 3$, this represents an area in only the right tail of the normal distribution.

5. **Critical region**. The critical region is the extreme area, in this case of a one-tailed or one-directional alternative hypothesis, the extreme area of $\alpha = 0.01$ is only in the right tail of the distribution. Looking through the table of the normal distribution for a B area of 0.01 in a tail of the distribution gives a $Z$-value of 2.32 or 2.33. The latter value of $Z = 2.33$ will be used here, so the critical region is all $Z$-values exceeding 2.33.

   The critical region and the associated conclusions that can be made are as follows:

   $$\text{Region of rejection of } H_0 : Z > +2.33$$

   $$\text{Area of nonrejection of } H_0 : Z \leq +2.33$$

6. **Conclusion**. In order to determine whether the sample mean $\bar{X}$ is within the critical region or not, it is necessary to determine the distance $\bar{X}$ is from the hypothesized mean $\mu$. This can be determined by obtaining the $Z$-value associated with the sample mean – that is, how many standard deviations $\bar{X} = 3.176$ is from the hypothesized mean of $\mu = 3$.

$$\begin{aligned} Z &= \frac{\bar{X} - \mu}{s/\sqrt{n}} \\ &= \frac{3.176 - 3}{1.056/\sqrt{680}} \\ &= \frac{0.176}{1.056/26.077} \\ &= \frac{0.176}{0.0405} \\ &= 4.346 > 2.33. \end{aligned}$$

   That is, the sample mean is 4.346 standard deviations above the hypothesized mean of $\mu = 3$. This is above the critical cut-off point of

+2.33, so this $Z$-value is in the critical region for the test. That is, the sample mean is 4.346 standard deviations above the hypothesized mean of 3, a great distance, and one that is extreme enough to be in the right 0.01 of the distribution.

Since this $Z$-value is in the critical region, the conclusion of the test is to reject the null hypothesis $H_0$ and accept the alternative hypothesis $H_1$. The conclusion is that the mean attitude of U of R undergraduates is on the agree side of neutral, a conclusion made at the $\alpha = 0.01$ level of significance. This provides quite strong evidence that students, on average, are not neutral on this issue but tend to agree.

**Hypothesis test for M5**

Let $\mu$ be the true mean level of opinion among all University of Regina undergraduates about issue M5, "The government should fund festivals and special events celebrating different cultures." The steps involved in conducting the hypothesis test are as follows.

1. **Hypotheses**. The null and alternative hypotheses are

$$\text{Null hypothesis} \quad H_0 : \mu = 3$$

$$\text{Alternative hypothesis} \quad H_1 : \mu > 3$$

2. **Test statistic**. The claim is about $\mu$, the mean of M5 for all undergraduate students. The sample mean, $\bar{X}$, is the test statistic.

3. **Distribution of the test statistic**. Since this is a random sample with large sample size of $n = 697$

$$\bar{X} \text{ is Nor}\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

$s$ can be used as an estimate of the population standard deviation $\sigma$, since $n$ is large.

4. **Significance level**. The level of significance requested here is 0.01, so this is $\alpha = 0.01$. Since the alternative hypothesis is that $\mu > 3$, this area represents an area in only the right tail of the normal distribution.

5. **Critical region**. The critical region and the associated conclusions that can be made are as follows:

$$\text{Region of rejection of } H_0 : Z > +2.33$$

$$\text{Area of nonrejection of } H_0 : Z \leq +2.33$$

6. **Conclusion**. For $\bar{X} = 3.042$, $s = 1.250$, $n = 697$, and hypothesized mean $\mu = 3$,

$$\begin{aligned} Z &= \frac{\bar{X} - \mu}{s/\sqrt{n}} \\ &= \frac{3.042 - 3}{1.250/\sqrt{697}} \\ &= \frac{0.042}{1.250/26.401} \\ &= \frac{0.042}{0.0473} \\ &= 0.879 < 2.33. \end{aligned}$$

The sample mean is only 0.879 standard deviations above the hypothesized mean of $\mu = 3$. This is well below the critical cut-off point of +2.33, so this $Z$-value is not in the critical region for the test. That is, while the sample mean exceeds 3, it is less than 1 standard deviation to the right of the hypothesized mean of 3, a small distance, and one that is great enough to be in the critical region in the right 0.01 of the distribution.

Since this $Z$-value is not in the critical region, the conclusion of the test is that there is insufficient evidence to reject the null hypothesis $H_0$. The conclusion is that the mean attitude of U of R undergraduates is no different than neutral on this issue, a conclusion made at the $\alpha = 0.01$

level of significance. While the mean for all undergraduates might be above 3, the sample mean is not far enough above 3 to conclude that, on average, student views on this issue are any different than a neutral view.

**Issues in hypothesis testing**

**A. One-tailed or two-tailed test**

In the notes of November 21, there was a two-tailed test to determine whether the mean age of undergraduates was 23 years or not. In the above tests for mean attitude of students, the tests were one-tailed (greater than 3). While it is not always entirely clear whether an hypothesis test should be one- or two-tailed, some guidelines concerning this are discussed here. These one- or two-tailed tests may be referred to as one- or two-directional tests, respectively.

Note that for both one- and two-directional tests, the null hypothesis is that the population mean is equal to some specified or hypothesized value, M. That is, $H_0 : \mu = M$ is the null hypothesis in both situations. If the alternative hypothesis is $H_0 : \mu \neq M$, this is a two-tailed test. If the alternative is a one-directional test, then the alternative hypothesis can be either $H_1 : \mu < M$, if the suspicion is that the population mean is less than $M$, or $H_1 : \mu > M$, if the suspicion is that the population mean exceeds $M$.

1. If a researcher has no idea whether a population mean is greater or less than some hypothesized value, then a two-directional test is most commonly used. All the researcher may need to know is whether the sample mean supports the hypothesis or not, so he or she uses a two-directional test, with the alternative hypothesis being that $\mu$ is not equal to the value specified in the null hypothesis. The critical region, or region of rejection, is then in the two tails of the distribution.

2. If the question gives some hint that the population mean may exceed the hypothesized value, then the alternative hypothesis is $H_1 : \mu > M$ and the region of rejection is in the extreme right tail of the distribution. Similarly, when the question suggests that the population mean may fall short of the hypothesized value, then the alternative hypothesis is $H_1 : \mu < M$ and the region of rejection is in the extreme left tail of the distribution.

3. If a researcher already has some knowledge of a population, it is more common to use a one-directional test than a two-directional one. As in the examples above, a researcher may know that opinion is generally in agreement with a particular issue, and the researcher is interested in determining whether the agreement expressed in a sample is strong enough to conclude that population members as a whole agree.

   Another example could be when there is a more serious problem if a mean is less than some specified value as opposed to being greater than this value. For example, suppose a claim is made that the mean income of households in some community is below the poverty line. In this case, a researcher could sample households in the community and test whether the mean household income $\mu$ is equal to or less than the poverty line (one-tailed test). If the sample shows that the mean income is equal to or greater than the poverty line, $P$, then the researcher may not be so concerned about this issue. In this case the null hypothesis would be that $\mu = P$ and the alternative hypothesis $\mu < P$. If the data lead to the conclusion that the null hypothesis is to be rejected, this demonstrates that there likely is inadequate household income in the community.

4. For any given significance level $\alpha$, note that the $Z$-value for a one-tailed test is smaller than for the corresponding two-directional alternative. For any given significance level, this means that the statistic need not be as far from the hypothesized mean in order to reject the null hypothesis in the case of a one-tailed test, as compared with the corresponding two-tailed test. For $\alpha = 0.05$, the critical value for a one-tailed test is a $Z$-value of 1.645. In the case of the corresponding two-tailed test, the critical values are $Z = \pm 1.96$. While this may be of some consequence for deciding whether to use a one- or two-directional test, the significance level $\alpha$ that is chosen is usually regarded as a more important consideration than this relatively small difference in critical values. See the later notes on selection of significance level.

**B. Potential errors involved in hypothesis testing** – section 9.2.4, p. 580.

No hypothesis test ever leads to an absolutely certain conclusion – there is always some possibility that the conclusion is in error.

In the case of confidence interval estimates, there is never absolute certainty that the intervals constructed contain the population mean or proportion. Similarly, a researcher cannot be certain that the rejection or non-rejection of the null hypothesis is correct. What a researcher can do though is select a larger probability of being certain. In the case of confidence intervals, a higher confidence level is associated with a greater chance that the intervals constructed will contain the true mean.

In hypothesis testing, a lower significance level is generally considered to provide a more definitive result. That is, a lower significance level means a smaller critical region, more distant from the hypothesized mean. This implies that the sample mean must be quite different from the hypothesized mean if the null hypothesis is to be rejected.

There are two types of error that are associated with a null and alternative hypothesis.

> **Type I error**. Type I error is the error of rejecting the null hypothesis $H_0$ when the null hypothesis is true. This type of error can occur when a researcher rejects a null hypothesis.

The explanation for this proceeds as follows. When an hypothesis test is conducted, the particular value for $\mu$ hypothesized in the null hypothesis is assumed to be correct. At the conclusion of the test, when the sample mean lies within the critical region this null hypothesis is rejected. In this case, the sample value is regarded as distant enough from the hypothesized mean, so that this hypothesized value can be rejected. But it is possible that the hypothesized mean is correct and the sample is one of the more unusual random samples drawn from the population. That is, a random sample could result in selecting a set of values that have a sample mean quite different than the hypothesized mean, and one so different that the sample mean is in the critical region. If this is the case, then the null hypothesis has been rejected in error.

The chance of the above is small, and is equal to the level of significance selected. That is, the critical region is of area, or probability, $\alpha$, a small value such as 0.05 or 0.01. If the null hypothesis is true, the chance of selecting a sample with a mean in the critical region is $\alpha$. But this results in Type I error. As a result, the probability of Type I error is equal to the significance level. This is stated symbolically as:

$$P(\text{type I error}) = P(\text{rejecting} H_0 / H_0 \text{is true}) = \alpha.$$

If a researcher wishes to be more certain that Type I error is not committed, then he or she can select a smaller significance level. This reduces the size of the critical region, thus reducing the chance of Type I error. The problem associated with this is that there may then be Type II error.

> **Type II error**. Type II error is the error of failing to reject the null hypothesis $H_0$ when this null hypothesis is false. This type of error can occur when the null hypothesis is not rejected.

The explanation for this is that the null hypothesis may not be correct, but the sample mean is not different enough from the hypothesized mean to reject the null hypothesis. This is sometimes termed $\beta$ (beta) error and can be stated symbolically as:

$$P(\text{type II error}) = P(\text{failing to reject} H_0 / H_0 \text{is false}) = \beta.$$

It is more difficult to calculate $\beta$ than $\alpha$ but, as explained below, the consequence of making this error is often fairly minimal.

**Types of error in above examples**. In the case of the mean age of students, the conclusion was that the mean age of all students was not equal to 23, a conclusion made at the 0.05 level of significance. While it is possible that the mean age of all undergraduate students is 23, this seems unlikely, given that the sample mean was associated with a $Z$-value of $-2.548$, well into the critical region. But it is possible that the sample was a sample with a lot of students younger than age 23, thus producing a low mean. As a result, there is at most an $\alpha = 0.05$ probability of type I error in this case.

A similar conclusion holds for the test of mean for the variable V1. The sample mean was 3.176 and, from the test, the conclusion was that the hypothesis of $\mu = 3$ could be rejected. Since this hypothesis was rejected at the 0.01 level of significance, this means that there was, at most, 0.01 chance of type I error. Given that the $Z$-value was over 4, it seems very unlikely that the hypothesis that the mean was $\mu = 3$ is correct. There is less than a 0.01 chance that the alternative hypothesis that $\mu > 3$ is an incorrect conclusion.

In the case of the variable M5, there was insufficient evidence to reject the null hypothesis that $\mu = 3$. But if all students were surveyed, it is unlikely that the population mean would be exactly 3. So there is very likely to be type II error in this case. Given that $\bar{X} = 3.042$, and $H_0 : \mu = 3$ cannot be

rejected, it seems likely that the true population mean is close to 3. While exactly 3 is unlikely to be the correct mean, it seems very likely that $\mu$ is close to 3. As a result, the consequence of making this type of error is fairly minimal – the only error is that a researcher is unable to distinguish a mean of exactly 3 from another mean very close to 3.

**C. Rejection or acceptance of $H_0$**

When a null hypothesis is rejected, this is a fairly clear-cut decision, and one associated with accepting $H_1$. That is, the sample yields data inconsistent with a specific value of $\mu$ so the claim that this specific value is correct is rejected. There is, at most, a probability $\alpha$ that this conclusion is incorrect (type I error).

In contrast, when a sample yields a mean that is not in the critical region, a researcher merely concludes that the null hypothesis is not rejected. In this case, the researcher hypothesizes a specific value for the mean and the sample data is not inconsistent with this specific value. But that does not mean that the researcher is certain that the specific value hypothesized is really the population mean. There is a considerable probability of type II error and, in this case, the researcher merely concludes that the null hypothesis is not rejected. That is, $H_0$ is not necessarily accepted or regarded as exactly correct – the conclusion is that the sample hypothesized mean is not all that incorrect.

**D. Choice of significance level** – p 588.

There is no single correct or incorrect choice of a significance level. Several rules or guidelines concerning choice of a significance level are as follows.

1. **Report** $\alpha$. Regardless of what level of significance you have selected, always make sure you report the level.

2. **Default** $\alpha = 0.05$. If you are unsure what significance level to choose, the $\alpha = 0.05$ level can always be selected. It is the default or most commonly used significance level. Other common significance levels in social science research are 0.10, 0.01, and 0.001.

3. **Use level others have used**. When comparing your results with those from other researchers, use the same level or levels they have reported – then your conclusions can be compared with their results.

4. **Balance with probability of error**. If you wish to minimize type I error, select a low significance level such as 0.01 or 0.001. If you can reject $H_0$ at these levels, then this provides strong evidence that the null hypothesis is incorrect. Remember though that the resulting type II error may be large, that is, it may be difficult to reject $H_0$.

   If you are attempting to show that the null hypothesis is incorrect, you may wish to select a larger significance level, with a larger critical region. While this may allow you to reject $H_0$, the consequence of this is larger type I error, that is, you may have rejected the null hypothesis when it is actually true.

5. **Type of issue**. If you are dealing with an issue of life and death, or an issue with serious consequences if a wrong conclusion is made, then ensure that you construct the hypothesis test and select the significance level accordingly. For example, suppose that the safe level for a possibly poisonous level of a chemical in drinking water is 2 parts per million – anything over this may threaten the health of those who drink the water. In this case, you may wish to construct $H_0 : \mu = 2$ ppm, with the alternative hypothesis $H_1 : \mu < 2$ ppm. Presumably you wish to ensure that you reject $H_0$ at a significance level such as 0.001 or .0001, or even less. That is, you obtain samples and only if they provide very strong evidence that $H_0$ can be rejected, will you conclude that the water is safe. See p. 587 for a discussion of this issue.

   For the social sciences, where consequences of error may be less serious and where there may be more difficulty obtaining accurate measurement of a variable, significance levels of 0.05 or 0.01 are more common. These ensure reasonably low levels of type I error.

6. **Exact significance level**. On the computer printout for hypothesis tests, the probability associated with the $Z$-value, or exact significance level, is often reported. Or if you calculate the $Z$-value, the exact significance level is the area in the tail of the distribution beyond this value. If you report this level, the reader can then decide whether this level is low enough to reject the null hypothesis.

   This value can also be considered the probability that the $Z$-value is the size it is, given the null hypothesis. This is a conditional probability

that the sample yields a $Z$-value of the magnitude reported or greater, given that the null hypothesis is correct. If this exact level is very low then the null hypothesis is rejected. If it is not so low, the null hypothesis is not rejected. A discussion of this issue is contained in the text, pp. 595-6.

There are a number of other issues involved in hypothesis testing – see sections 9.2.4 - 9.2.9 of the text, pp. 580-606.

**Next topics**: Hypothesis tests for a mean, small sample size, and hypothesis test for a proportion.