Worksheet 02: Proportion Test

Statistical inference can be broadly defined as the process of using a random sample to make estimates of the parameters of the distribution(s) that generated them. This is, more-or-less, the entire topic of this course. We can rougly categorize all statistical inference tasks into one of four areas: (1) point estimation, (2) confidence intervals, (3) prediction intervals, and (4) hypothesis testing. These tasks can be applied to any unknown feature of the underlying distribution. As an example here, continuing from our work last time, we will focus on inference for the expected value (μ) of the distribution $\mathcal{G}$ that was used to generate our random sample.

A **point estimate** is a random variable that serves as a best guess for an unknown parameter. Traditionally, we use a symbol of the unknown parameter with a circumflex over it ($\hat{\mu}$) to represent the value of a point estimate. These will be our primary object of study in the second unit of the course. For now, we will mostly be able to get by using the sample mean and sample variance as our point estimates for the mean and variance of an unknown distribution.

We already defined a **confidence interval** last time: it consists of two random variables L (lower) and U (upper) such that there is a minimum probability that the unknown parameter falls between these bounds. A **prediction interval** is, similarly, a set of two random variables L and U such that there is a minimum probability that a new observation from $\mathcal{G}$ will fall between the bounds. These are objects that will be most important towards the end of the semester as we investigate methods for linear and generalized regression.

Finally, an **hypothesis test** is a technique for determining whether there is sufficent evidence in the random sample to support a particular claim about the generating distribution $\mathcal{G}$. The techniques for hypothesis testing are often similar to confidence intervals, but there are cases where the two techniques diverge. Today, we will focus on hypothesis tests and the specific terminology that is used to describe them.

There are several important complexities, critiques, and interpretations of hypothesis tests. Let's start with some of the core concepts and approaches and then we can expand these as we move forward. An hypothesis test starts with a **null hypothesis** (H_0) and an **alternative hypothesis** (H_A). Our goal is to see what support the data provide for rejecting the null hypothesis in favor of the alternative hypothesis. To do this, we start with a random variable called the **test statistic** that has a known distribution under the null hypothesis. Then, we compute a **p-value** as a measurement of how extreme the value of the test statistic is under the assumption of the null hypothesis. If the p-value is sufficently small, we would interpret this as having evidence that the null hypothesis should be rejected in favor of the alternative hypothesis. If the p-value is below a pre-specified cut-off, we might say that the result is **statistically significant**. That's a lot of terminology. We can get a better sense of how this works with a concrete example.

Let's take a specific example with some simulated data. Consider the situation where UR is trying to decide whether the DHall diner hours should be moved one hour later. The adminstration wants to make this change if and only if there is compelling evidence that a majority of students want it. For data we have a random sample of 40 students who responded with the following values, where 0 means not changing the hours and 1 means moving them an hour later:

```
1, 1, 1, 0, 1, 0, 0, 1, 1, 1, 1, 1, 0, 1, 0, 1, 0, 0, 1, 0,
0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 0, 0, 1, 0, 1
```

There are 14 values of 0 and 26 values of 1. Notice that we actually known the specific form of the distribution $\mathcal{G}$ in this case: it is a Bernouilli distribution with some unknown probability p of being equal to 1. **Q01.** As a form of review, compute the mean and variance of a random variable X with a $Bernouilli(p)$ distribution. Work out the actual result; don't just copy from the table. We have:

$$\mathbb{E}X = \mathbb{P}[X = 0] \cdot 0 + \mathbb{P}[X = 1] \cdot 1 = p$$
$$\mathbb{E}X^2 = \mathbb{P}[X = 0] \cdot 0^2 + \mathbb{P}[X = 1] \cdot 1^2 = p$$

And so,

$$\text{Var}[X] = \mathbb{E}X^2 - [\mathbb{E}X]^2 = p - p^2 = p(1 - p).$$

The null and alternative hypotheses that we will use with this data for our given question are:

$$H_0 : p = 0.5$$
$$H_A : p > 0.5$$

The test statistic that we will use is the scalled value of the sample mean that we last time called Z. It's general form is:

$$Z = \frac{\bar{X} - \mu_X}{\sigma_X / \sqrt{n}}.$$

For an hypothesis test, we need to know the distribution of our test statistic under H_0. **Q02.** Using the normal approximation, what is the distribution of Z under the null hypothesis? Write a more specific form of Z by filling in the values of μ_X and σ_X. The distribution of Z will be $N(0, 1)$, just as we saw last time. The specific value of Z here is given by filling in $\mu_X = 0.5$ and $\sigma_X = \sqrt{0.5(1 - 0.5)} = 0.5$:

$$Z = \frac{\bar{X} - 0.5}{0.5 / \sqrt{n}}.$$

Q03. What is the specific value, which we will call z, of this test statistic from our example data? The mean is $26/40 = 0.65$, so:

$$z = \frac{0.65 - 0.50}{0.5 / \sqrt{40}} \approx 1.897.$$

If we observe a specific value of z by plugging in the value of the sample mean and sample size above, the p-value of this hypothesis test will be given by:

$$\text{p-value} = \mathbb{P}\left[Z > z\right]$$

Usually we would have to look up this value on a table or (better yet) use some computer software to get the exact probability. We will do that in a moment, but just for reference, know that:

$$\mathbb{P}\left[Z > 1.644\right] \approx 0.05$$
$$\mathbb{P}\left[Z > 2.323\right] \approx 0.01.$$

Q04. Based on these value, what can you say about the p-value from our example? Would you say that this is statistically significant? It should be between 0.05 and 0.01. So, the result is significant at a level 0.05 but not at a level of 0.01.

We could have used the following set of hypotheses instead of the ones that we had above:

$$H_0 : p = 0.5$$
$$H_A : p \neq 0.5$$

Q05. What would change about our analysis with these? Would the p-value be larger or smaller? The distribution of Z depends only on the null hypothesis, so its distribution is the same. However, we would need to change the definition of the p-value to be:

$$\text{p-value} = \mathbb{P}\left[|Z| > z\right]$$

This would double the p-value that we had in the our example.

The test we have shown above is called the **one-sample proportion test**. It is based on the normal approximation of $\bar{X}$ and can be easily extended to testing any null hypothesis of p being equal to a specific value. Unless we have a strong rational, it is usually recommend to use the second approach (with $H_A : p \neq p_0$), called a two-sided alternative. This is because it gives more conservative p-values and so therefore is a safer option.

An alternative test statistic to use in this example is just the value $Y = \sum_{i=1}^{n} X_i$. The rule about a test statistic is that we can use any random variable that has a known distribution under the null hypothesis. **Q06.** What is the distribution of Y under the null hypothesis? This is a sum of n Bernouilli random variables, so $Y \sim Bin(n, p)$, a binomial distribution.

The test resulting from the test statistic Y is called the **Binomial test**. Yes, that's a major hint to the previous question. This is almost always preferred to the one-sample proportion test (the latter exists primarily for its multiple-sample extensions) because it does not rely on the assumption of asymptotic normality. We have derived the proportion tests above, however, because it leads more cleanly to the test that we will explore next time.