Classwork 6: Hypothesis Testing (Part 1)

Note: this material comes from Dr. Lauren Perry's online textbook Introduction to Statistics.

In this classwork, you will get reacquainted with some statistics topics including confidence intervals and hypothesis testing. The goal is to build the fundamentals in order to understand how hypothesis testing works on regression parameter estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ in classwork 7.

Hypothesis testing is all about trying to account for **sampling error**: when we want to learn something about a population, we take a sample, but no sample is perfect. If we use the sample mean $\bar{x}$ to estimate μ_X, chances are that $\bar{x} \neq \mu_X$, although they might be close.

In classwork 2, you showed that $\bar{x}$ is an unbiased estimator for μ_X because $E[\bar{x}] = \mu_X$. You also showed that $\bar{x}$ is more efficient than other estimators because its variance is smaller: $Var(\bar{x}) = \frac{\sigma_X^2}{n}$.

1) The average living space for a detached single family home in the US is 1742 square feet

with a standard deviation of 568 square feet (the standard deviation is the square root of the variance). If we measured the living spaces of 25 random homes, find the expected value (recall that the sample mean is unbiased) and the standard deviation of that sample's mean.

2) If we instead measured the living space of 100 random homes, what is the expected value and standard deviation of that sample's mean?

3) How many homes would we have to sample if we wanted the standard deviation of the sample mean to be less than 10?

Developing Confidence Intervals: In statistics, if we want to be 95% confident about a statement, that means we want to arrive at our statement using a method that will give us a correct statement 95% of the time. A confidence interval is an interval (a, b) where we can be 95% confident that the true population parameter lies inside that interval.

In the context of using a sample mean to estimate μ_X, our best estimate for μ_X is $\bar{x}$, so that point estimate will be the center of the confidence interval. If the population standard deviation is known and the sample size is greater than 30, $\bar{x}$ is distributed approximately normally with an expected value of μ_X and a standard deviation of $\frac{\sigma_X}{\sqrt{n}}$. Then we can standardize $\bar{x}$ by subtracting μ_X and dividing by $\frac{\sigma_X}{\sqrt{n}}$.

4) If $\bar{x}$ is distributed approximately normally with an expected value of μ_X and a standard deviation of $\frac{\sigma_X}{\sqrt{n}}$, show that the distribution of $Z \equiv \frac{\bar{x}-\mu_X}{\sigma_X/\sqrt{n}}$ is standard normal: $N(0,1)$.

Actually just show that $E[Z] = 0$ and $Var(Z) = 1$.

Instead of looking up critical values in tables in the back of a book, there are R functions. `qnorm()` provides information about the quantile function for the normal distribution.

```r
qnorm(p = .025)
```

```
## [1] -1.959964
```

That is, .025 of the probability mass of the standard normal distribution is to the left of -1.96 and .025 is to the right of 1.96, which makes -1.96 and 1.96 the critical values for a 95% confidence interval.

5) Use the expression $-1.96 < Z < 1.96$, plug in for $Z \equiv \frac{\bar{x}-\mu_X}{\sigma_X/\sqrt{n}}$, and manipulate to get the formula for the 95% confidence interval bounds:

$$\bar{x} - 1.96\frac{\sigma_X}{\sqrt{n}} < \mu_X < \bar{x} + 1.96\frac{\sigma_X}{\sqrt{n}}$$

6) Calculate the 95% confidence interval for the expected value of a population characteristic, where a sample of 100 observations has mean 50 and the population standard deviation is known to be 18.

In practice, σ_X is almost never known, but the good news is that we can use the sample standard deviation $\hat{\sigma}_X$ as an approximation. So when $n \geq 30$, you can use the confidence interval formula $\bar{x} \pm z_{\alpha/2}\frac{\hat{\sigma}_X}{\sqrt{n}}$.

When $n < 30$, use the t-distribution with $n-1$ degrees of freedom: it's a lot like the normal distribution, but with slightly fatter tails depending on the degrees of freedom. Use the formula $\bar{x} \pm t_{\alpha/2, df=n-1}\frac{\hat{\sigma}_X}{\sqrt{n}}$.

7) Consider the sample {1, 3, 12, 8, 4, 6, 12, 7, 8, 9}. Find $\bar{x}$, $\hat{\sigma}_X$, and construct the 95% confidence interval for μ_X.

You can use this for the t distribution:

```
qt(.025, df = 9)
```

```
## [1] -2.262157
```

One of our goals with statistical inference is to make decisions or judgments about the value of a parameter. A confidence interval is a good place to start, but we might also want to answer questions like: Do cans of soda actually contain 12 fluid ounces? Or, is medicine A better than medicine B? A hypothesis test has two competing hypotheses:

H_0 is the null hypothesis. It's the default assumption, like $\mu_X = 12$.

H_A is the alternative hypothesis, like $\mu_X \neq 12$.

A hypothesis test helps us decide whether the null hypothesis should be rejected in favor of the alternative hypothesis. Hypothesis test results are either "reject H_0" if the result is statistically significant, or "fail to reject H_0" if the results are not statistically significant.

There are three methods for doing hypothesis tests, and they all end up being exactly equivalent. Your conclusions will never depend on which method you use.

1. Confidence intervals
2. Critical values
3. P values

The **confidence interval** approach looks like this: a confidence interval gives us a range of plausible values for μ. If the null hypothesis value μ_0 is in that interval, then μ_0 is a plausible value for μ. If not, then μ_0 is not a plausible value for μ.

8) A random sample of $n = 19$ leads to $\bar{x} = 4.4$ and $\hat{\sigma}_X = 2.3$. Is μ_X different from 2.5? Test at the .05 level of significance by building a 95% confidence interval.

Begin by writing down your null and alternative hypothesis. End by stating whether you reject H_0 or fail to reject it. You can use this:

```
qt(p = .025, df = 18)
```

```
## [1] -2.100922
```

The **critical value** approach looks like this: assuming H_0 holds, how likely are we to observe a sample like the one we have? The test statistic is the number of standard deviations away from μ_0 that $\bar{x}$ is. You can use the formula: $\frac{\bar{x}-\mu_0}{\hat{\sigma}_X/\sqrt{n}}$, which has the t-distribution when $n < 30$ with $n-1$ degrees of freedom.

Compare that test statistic to the critical value:

```r
qt(p = .025, df = 18)
```

```
## [1] -2.100922
```

If the test statistic is more extreme than the critical value, we reject the null hypothesis.

9) Repeat the hypothesis test you conducted in question 8, this time using the critical value approach.

The **p-value approach** asks if H_0 is true, what is the probability of getting a random sample that is as inconsistent with the null as the sample we got? You can use the formula: Prob(t_{df} is as extreme as the test statistic) $= 2 \times p(|t_{df}| > \frac{\bar{x}-\mu_0}{\hat{\sigma}_X/\sqrt{n}})$

And the function `pt()` is the relevant one. It tells you the proportion of the probability mass that is to the left of the argument q.

```r
pt(q = -3.6, df = 18)
```

```
## [1] 0.0010236
```

If the p-value is less than .05, you can reject the null hypothesis at the .05 significance level.

10) Again repeat the hypothesis test you conducted in question 8, this time using the p-value approach.