Small sample Hypothesis
Testing of means and review of chapter 7 random variables

April 13, 2020

1 Estimating μ and $\mu_1 - \mu_2$

In this section, we consider a random sample Y_1,...,Y_n with an unknown mean, but n is small. Since n is small we cannot assume $\sum_i Y_i$ is approximately normal or that S^2, the empirical sample variance, has converged to σ^2 the true sample variance. Because of these two issues we cannot hope to solve this problem in general, we saw this exact issue already when we did confidence intervals for estimation. Instead of trying to consider the general problem, we restrict ourselves to the case that the random sample is normal. Fortunately, assuming measurements come from a normal distribution is often reasonable.

****Throughout this section we will always assume Y_i is a normal random variable with unknown mean μ and variance σ^2.****

When both the μ and σ^2 are unknown, then the pivotal quantity for the estimation of μ is:

$$T = \frac{\overline{Y} - \mu}{S/\sqrt{n}}$$

where we recall $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ has a normal distribution with mean μ and variance σ^2/n and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$. The random variable $\frac{n-1}{\sigma^2}S^2$ is a χ^2 random variable with $n-1$ df. By a special property of the normal distribution, $\overline{Y}$ and S^2 are independent. The random variable T has a t-distribution with $n-1$ degrees of freedom. At the end of these notes, I have a reminder on the t-distribution.

This was the same random variable we considered with confidence intervals and is the replacement for $Z = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}}$ when the sample empirical variance has not converged.

We will use T as our test statistic. We are now ready to state the elements of a small sample size hypothesis test, we'll state the one-sided and two-sided all together.

HYPOTHESIS TEST 1.1 (Small Sample Test for μ). *Let Y_1,Y_2,...,Y_n be a random sample from a normal distribution with unknown mean μ.*

The null hypothesis is $H_0: \mu = \mu_0$.

The alternative hypothesis is one of the three $H_a: \begin{cases} \mu > \mu_0 \\ \mu < \mu_0 \\ \mu \neq \mu_0 \end{cases}$

The test statistic is $T = \frac{\overline{Y} - \mu_0}{S/\sqrt{n}}$

The rejection region is one of the three (depending on H_a) $RR: \begin{cases} T > t_{\alpha,(n-1)df} \\ T < -t_{\alpha,(n-1)df} \\ |T| > t_{\alpha/2,(n-1)df} \end{cases}$

Where $t_{\alpha,(n-1)df}$ is such that $\mathbb{P}(T > t_{\alpha,(n-1)df}) = \alpha$, where T has (as above) the t-distribution with $n-1$ degrees of freedom.

Let's look at an example (This is problem 10.66 from the book, there they give the values all the data points, I will report just the sample mean and variance, but feel free to compute them for yourself (using a computer)).

EXAMPLE 1.2. *Researchers want to study the effect of smoking on the lungs by measuring the lung's diffusing capacity (higher is better). In a study of 20 smokers an empirical mean of 90 and standard deviation of 15 is observed.*

(a) At an $\alpha = .01$ level, do you reject the null hypothesis that smoker's mean diffusing capacity is 100 (the population average) in favor the alternative hypothesis that smoker's mean diffusing capacity is lower than 100 (i.e. smoking is bad for you).

(b) What is the p-value for this test?

Solution:

(a) We begin by writing our the elements of this test. Let μ be the true smoker's mean diffusing capacity.
H_0 is $\mu = 100$, H_a is $\mu < 100$.

The test statistic is $T = \frac{\overline{Y} - \mu}{S/\sqrt{20}}$, where 20 is the sample size. We observed $\overline{Y} = 90$ and $S = 15$ so under H_0 we observed

$$T = \sqrt{20}\frac{90 - 100}{15} = -2.98$$

The RR at $\alpha = .01$ is $T < -2.537 = -t_{.01,19df}$ (This is from the table in the book the column is $t_{.010}$ and the row is 19 df).

Since $T = -2.98 < -2.537$ we reject the null Hypothesis in favor of the alternative that smoking is bad.

(b) The probability that a t-distribution random variable with 19 df is less than -2.98 is

$$\mathbb{P}(T < -2.98) = 0.00385$$

I computed this in R by typing: pt(-2.98,df=19). So the p-value is 0.00385, which is quite small (meaning we should be confident in the result). There are also lots of online calculators or other math programs that will compute this for you.

If you were on a deserted island and didn't have a computer but had your textbook and needed to bound the p-value, you would look at the table and note that 2.98 is greater than $2.861 = t_{.005,19df}$, so the p-value is smaller than .005.

Given two different data sets we might want to not estimate their means, but rather see if one has a higher mean than the other, or if their means are the same. To do such hypothesis tests, we'll use the same test static as when making confidence intervals. In order to make a test statistic that we can compute the variance of, we need to assume the true variance of both samples is the same.

The test statistic is, as usual, a pivotal quantity that we can compute from just the data, when the null hypothesis is assumed to hold. So we take our sample means, then subtract off the true sample means and would then like to divide by the square root of the variance. Since the variance is unknown, we use our best estimate for it, the pooled sample variance:

$$S_p^2 = \frac{\sum_{i=1}^n (Y_i - \overline{Y})^2 + \sum_{i=1}^n (X_i - \overline{X})^2}{n+m-2}$$

an alternative way to write this is

$$S_p^2 = \frac{(n-1)S_Y^2 + (m-1)S_X^2}{n+m-2}$$

where

$$S_Y^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \overline{Y})^2,\ S_X^2 = \frac{1}{m-1}\sum_{i=1}^m (X_i - \overline{X})^2$$

Recall that $\frac{1}{\sigma^2}\sum_{i=1}^n (Y_i - \overline{Y})^2$ is χ^2 with $(n-1)df$ and $\frac{1}{\sigma^2}\sum_{i=1}^m (X_i - \overline{X})^2$ is χ^2 with $(m-1)df$. When we add two independent χ^2 rv's we get a new χ^2 whose df is just the sum of the original two df's, so $\frac{n+m-2}{\sigma^2}S_p^2$ is a χ^2 rv. Using the mean of a χ^2 rv we then see that S_p^2 is an unbiased estimator of σ^2.

The variance of $\overline{Y} - \overline{X}$ is

$$\mathrm{Var}(\overline{Y} - \overline{X}) = \mathrm{Var}(\overline{Y}) + \mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n} + \frac{\sigma^2}{m}$$

Putting this all together our test statistic is

$$T = \frac{\overline{Y} - \overline{X} - (\mu_Y - \mu_X)}{S_p\sqrt{1/n + 1/m}}$$

We are usually interested in the of the means being the same or bigger, but we can consider any difference, d, between the means. Below $d=0$ is the typical case, but any fixed d doesn't really change the analysis.

We now state the hypothesis test:

HYPOTHESIS TEST 1.3 (Small Sample Test for $\mu_1 - \mu_2$). *Let $Y_1, Y_2, ..., Y_n$ be a random sample from a normal distribution with unknown mean μ_Y and unknown variance σ^2.*
Let $X_1, X_2, ..., X_m$ be a random sample from a normal distribution with unknown mean μ_X and unknown

variance σ^2.

*****Note: in both examples we assume a variance of σ^2. If this is not a reasonable assumption, then this method does not work*****

Let d be a fixed constant.
The null hypothesis is $H_0: \mu_Y - \mu_X = d$.

The alternative hypothesis is one of the three $H_a: \begin{cases} \mu_Y - \mu_X > d \\ \mu_Y - \mu_X < d \\ \mu_Y - \mu_X \neq d \end{cases}$

The test statistic is $T = \frac{\overline{Y} - \overline{X} - d}{S_p\sqrt{1/n + 1/m}}$

The rejection region is one of the three (depending on H_a) $RR: \begin{cases} T > t_{\alpha,(n+m-2)df} \\ T < -t_{\alpha,(n+m-2)df} \\ |T| > t_{\alpha/2,(n+m-2)df} \end{cases}$

Where $t_{\alpha,(n+m-2)df}$ is such that $\mathbb{P}(T > t_{\alpha,(n+m-2)df}) = \alpha$, where T has (as above) the t-distribution with $n+m-2$ degrees of freedom.

So this basically works the same as estimating the mean, except we need to compute S_p^2 and the normalization is $\sqrt{1/n + 1/m}$.

EXAMPLE 1.4. *Two teaching methods are to be compared. The researchers would like to know they produce different mean scores on an exam. The following data is observed:*

	Method 1	*Method 2*
Sample size	11	14
Empirical mean	64	69
Empirical Variance	52	71

(a) What assumptions do we need to make to preform a Hypothesis test?

(b) Using these assumptions, at an $\alpha = .05$ level, do the teaching methods lead to a different mean score?

(c) What is obtained p-value from this data?

Solution:

(a) We assume that each student's score on the test is an independent normal random variable and that the normal random variable has the same variance for every student.

(b) Let X_i, $1 \leq i \leq 11$ be the i^{th} student in the first class's score and Y_i, $1 \leq i \leq 14$ be the i^{th} student in the second class's score.

We are testing $H_0: \mu_X = \mu_Y$ against $H_a: \mu_X \neq \mu_Y$

Since we are given the empirical sample variance for each class we use the second formula to compute the pooled sample data

$$S_p^2 = \frac{(n-1)S_Y^2 + (m-1)S_X^2}{n+m-2} = \frac{10*52 + 13*71}{23} = 62.74$$

We can now compute T

$$T = \frac{64-69-0}{\sqrt{62.74}\sqrt{1/10+1/13}} = -1.5$$

Note: we use $\sqrt{62.74}$ because we want S_p, not S_p^2.

From the table $t_{.025,df=23} = 2.069$ (Recall we're doing 2-sided test so we compute $t_{\alpha/2}$).

Since $|-1.5| < 2.069$ we do not reject the null hypothesis.

(c) $\mathbb{P}(|T| > 1.5) = .1485$. So the p-value of this 2-side test is .1485.

The R command is: 2*pt(-1.5,df=21), I've multiplied by 2, because it's a two-sided test.

The following 3 random variables appear heavily in this and the next section, so here is a brief reminder of their definition, properties and use in this course.

2 χ^2 distribution

If Z_1,...,Z_ν are independent standard normal random variables. Then

$$W = \sum_{i=1}^{\nu} Z_i^2$$

is a χ^2 random variable with ν degrees of freedom (df). Note that it has mean ν and is always positive.

In this class, this random variable appears when we consider Y_1,...Y_n, independent normal random variables with variance σ^2.

The random variable $\frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \frac{n-1}{\sigma^2}S^2$ is a χ^2 random variable. Although, this kind of looks like the form in the definition, it is definitely different, so it is not very obvious that this random variable is be χ^2. The proof of this fact is beyond the scope of the course, the book did the proof in the case $n=2$, and even it is somewhat involved.

3 t-distribution

The most convenient way to represent a t-distributed random variable with ν degrees of freedom is as the fraction:

$$T = \frac{Z}{\sqrt{W/\nu}}$$

where Z is a standard normal random variable and W is a χ^2 random variable with ν degrees of freedom, and Z and W are independent. We could compute the pdf, but it is not as easy to work with or as useful.

The t-distribution behaves like a Normal random variable, but its pdf decays slower. In particular, its pdf is symmetric. In particular, its pdf is symmetric.

The t-distribution appears here because for normal random variables $\overline{Y}$ and S^2 are independent rv, with $\overline{Y}$ being normal and S^2 being a constant multiple of a χ^2. So properly normalizing of ratio between $\overline{Y}$ and $\sqrt{S^2}$ to get the form above gives:

$$\frac{\frac{\overline{Y}-\mu}{\sigma\sqrt{n}}}{\sqrt{\frac{S^2(n-1)}{\sigma^2(n-1)}}} = \sqrt{n}\frac{\overline{Y}-\mu}{S}$$

which is the form of our test statistic.

4 F distribution

If W_1 is χ^2 with ν_1 degrees of freedom and W_2 is χ^2 with ν_2 degrees of freedom, then

$$F = \frac{W_1/\nu_1}{W_2/\nu_2}$$

is an F-random variables with ν_1 numerator degrees of freedom and ν_2 denominator degrees of freedom. Like the t-distribution we could compute the pdf of this random variable, but this representation is the nicest to work with and most relevant for us.

We can generate this random variable if we have 2 independent random samples, each with a normal distribution. Let X_1,X_2,...,X_m be a random sample from a normal distribution with unknown variance σ_X^2 and Y_1,Y_2,...,Y_n be a random sample from a normal distribution with unknown variance σ_Y^2. We assume the X_i's and independent from the Y_i's.

As usual let $S_X^2 = \frac{1}{m-1}\sum_i^m (X_i - \overline{X})^2$ and $S_Y^2 = \frac{1}{n-1}\sum_i^n (Y_i - \overline{Y})^2$.

Since $\frac{m-1}{\sigma_X^2}S_X^2$ and $\frac{n-1}{\sigma_Y^2}S_Y^2$ are both χ^2 random variables then

$$\frac{\frac{m-1}{(m-1)\sigma_X^2}S_X^2}{\frac{n-1}{(n-1)\sigma_Y^2}S_Y^2} = \frac{S_X^2\sigma_Y^2}{S_Y^2\sigma_X^2}$$

is an F distributed random variable with $m-1$ numerator df and $n-1$ denominator df.

If F has $m-1$ numerator df and $n-1$ denominator df, then $1/F$ also has F distribution but with $n-1$ numerator df and $m-1$ denominator df.