Continuing Statistical Inference IV

Comparing Means With Small Sample Size

Dealing with Small Samples

In our previous discussions for comparing two means from a two populations, we have dealt with the situation where the sample size from each population was large.

The reason for this (quick recap):

- $\bar{X}_1 \sim N(\mu_1, \sigma_1^2/n_1)$ (by CLT)
- $\bar{X}_2 \sim N(\mu_2, \sigma_2^2/n_2)$ (by CLT)
- $\bar{D} = \bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$

To get that last line, we made use of this fact:

> **FACT**
> The sum or difference of two independent normal random variables will be a normal random variable

We can then use the distribution of $\bar{D}$ to make probability statements about μ_1 and μ_2.

Dealing with Small Samples

However, our ability to make inferential statements are all based in this case on the knowledge that that $\bar{D}$ follows a normal distribution though.

If one or both of the samples has a small sample size, then we have a disruption in the logic above:

- If n_1 is small then
- $\bar{X}_1$ is not normal (can not apply central limit theorem) and
- $\bar{D} = \bar{X}_1 - \bar{X}_2$ is not normal (because $\bar{X}_1$ is not normal).

This breaks the key part of inference: we no longer have a probability distribution that connects the values we can calculate from our sample to the actual true parameters of the population.

Small n

Why small n changes things

Dealing with Small Samples

This is a problem and we can't work our way around it by using the fact that we can connect $\bar{X}_1$ and μ_1 through a t-distribution:

> **FACT**
> The sum or difference of two independent t random variables **will not be** a t random variable

Ultimately, <u>for small samples sizes</u> the distribution of $\bar{D}$ will depend completely on the distributions of the populations we are studying.

When making inferential statements, we have very different tools/methods/concerns for a feature exponentially distributed across the population vs a feature uniformly distributed across the population.

Dealing with Small Samples

Special Case: Normal population, same variance

In some cases, what we know about the populations being compared leads to nice results:

1. In both populations, the feature of interest is normally distributed across the populations members.

2. The amount of variation in the feature of interest is identical between the populations.

In other words, we are only dealing with the case where:

1. Each observation from the first population can be treated as a single value taken from a $N(\mu_1, \sigma^2)$ distribution
2. Each observation from the second population can be treated as a single value taken from a $N(\mu_2, \sigma^2)$ distribution

Small n

Why small n changes things

Assuming Normality

Dealing with Small Samples

Special Case: Normal population, same variance

Since for each sample, the sample mean will be made up of observations from a normal distribution we can now say that if $\bar{X}_1$ is the sample mean of the first population and $\bar{X}_2$ is the sample mean of the second population, then

- $\bar{X}_1 \sim N(\mu_1, \sigma^2/n_1)$ (sum of indep. normals is normal)
- $\bar{X}_2 \sim N(\mu_2, \sigma^2/n_2)$ (sum of indep. normals is normal)
- $\bar{D} = \bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \sigma^2/n_1 + \sigma^2/n_2\right)$ (sum of indep. normals is normal)

However, we are only assuming that the variance is the same value (whatever σ^2 actually is). We aren't assuming that we know that value.

Dealing with Small Samples

Special Case: Normal population, same variance (cont)

Step 1: Pooling the variance

Since we are assuming that both populations have the same variance, we need to think about how to estimate it.

Dealing with a sample from a single population, we estimate the variance using the sample variance:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

However, if we do this with samples from two populations, we will get two estimates of σ^2:

- s_1^2 from the first population's sample, and
- s_2^2 from second population's sample.

Dealing with Small Samples

Special Case: Normal population, same variance (cont)

Step 1: Pooling the variance (cont)

Should we use both s_1^2 and s_2^2 to estimate σ^2?

Conceptually, this is a problem: how can we put together a coherent solution if we are using two estimates of the same value at the same time?

It's not a well optimized use of our data: it can be proven that we get a better estimate of σ^2 by blending these different estimates.

> Side note: For two good estimates of the same value, it is generally the case that combining them will produce an even better estimate.

Dealing with Small Samples

Special Case: Normal population, same variance (cont)

Step 1: Pooling the variance (cont)

This idea of "pooling" the two estimates together leads to the following:

> **Pooled Sample Variance and Pooled Sample Standard Deviation**
> For two numerical samples of size n_1 and n_2 respectively, from populations with the same variance σ^2, using the sample variances s_1^2 and s_2^2 the **pooled sample variance** is defined as
>
> $$s_P^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}$$
>
> and the **pooled sample standard deviation** is
>
> $$s_P = \sqrt{s_P^2}$$

Dealing with Small Samples

Special Case: Normal population, same variance (cont)

Step 2: Distribution Connecting Data and Parameters

Since the sum of normal random variables is also normal and all of our observations in each sample will just be values taken from a normal distribution, then

$$\bar{X}_1 - \bar{X}_2 \sim N(\mu_1 - \mu_2, \sigma^2/n_1 + \sigma^2/n_2)$$

which means

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0, 1)$$

We could use this to estimate things if we knew σ - however, we don't know σ.

Small n

Why small n changes things

Assuming Normality

Dealing with Small Samples

Special Case: Normal population, same variance (cont)

Step 2: Distribution Connecting Data and Parameters

In order to "replace" σ with our estimate S_p we divide the equation above by S_p/σ.

But since S_p is a random variable and based on a small sample size, this messes up our distribution (we're dividing the Z from above by a new random variable).

However, we do know a about distribution based on S_p:

$$W = \frac{(n_1 + n_2 - 2)S_p^2}{\sigma^2}$$

follows a χ-squared distribution with $(n_1 + n_2 - 2)$ degrees of freedom.

Small n

Why small n changes things

Assuming Normality

Dealing with Small Samples

Special Case: Normal population, same variance (cont)

Step 2: Distribution Connecting Data and Parameters

We also know how an important relationship between Z and W:

> If Z is a standard normal random variable and W is a χ-squared random variable with ν degrees of freedom then
>
> $$T = \frac{Z}{\sqrt{W/\nu}} \sim t_\nu$$
>
> i.e., T follows a t-distribution with ν degrees of freedom.

Small n

Why small n changes things

Assuming Normality

Dealing with Small Samples

Special Case: Normal population, same variance (cont)

Step 2: Distribution Connecting Data and Parameters

Since Z (from slide 10) follows a standard normal and W (from slide 11) following a χ-squared distribution, we can "replace" the unknown σ. Notice that

$$W = \frac{(n_1 + n_2 - 2)S_P^2}{\sigma^2} \rightarrow \sqrt{\frac{W}{(n_1 + n_2 - 2)}} = \frac{S_p}{\sigma}$$

and we can create a random variable T using the relationship from slide 12:

$$T = \frac{Z}{\sqrt{\frac{W}{(n_1+n_2-2)}}} = \frac{\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}}{S_p/\sigma} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

which will follow a t-distribution with $n_1 + n_2 - 2$ degrees of freedom.

Dealing with Small Samples

Special Case: Normal population, same variance (cont)

Step 2: Distribution Connecting Data and Parameters

Notice that the only values in T that we can not calculate once we have data are μ_1 and μ_2, the parameters we want to compare.

We can now perform hypothesis tests and create confidence intervals for $\mu - \mu_2$ using a t-distribution with $n_1 + n_2 - 2$ degrees of freedom.

Dealing with Small Samples

Special Case: Normal population, same variance (cont)

Hypothesis Test Statistic

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_P\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

Confidence Intervals

$$\bar{X}_1 - \bar{X}_2 \pm t\sqrt{S_P^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

where t is a value based on a t-distribution with $n_1 + n_2 - 2$ degrees of freedom.