Small Sample Inference

553.211 Intro to Prob. & Stats.

Johns Hopkins University

EN.553.211

Where we've been

For the past few weeks, we have known that

Fact

1. *If the population is normal distributed*
2. *Of if n are sufficiently large ($n \geq 30$), regardless of the population distributions, by CLT*

we have

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

Furthermore, if n is large enough (say $n \geq 30$), when we use S instead of σ, we have

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \overset{approx.}{\sim} N(0,1)$$

Where we've been

For the past few weeks, we have known that

Fact

1. *If the population is normal distributed*
2. *Of if n are sufficiently large ($n \geq 30$), regardless of the population distributions, by CLT*

we have

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

Furthermore, if n is large enough (say $n \geq 30$), when we use S instead of σ, we have

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \overset{approx.}{\sim} N(0,1)$$

Question: what if the population is normal distributed but n is not large enough?

Student's t-distribution

Theorem

If $X_1, X_2, \ldots, X_n$ are independent and follow $\mathcal{N}(\mu, \sigma)$ with sample mean $\overline{X}$ and sample standard deviation S.
Then we have that $\frac{\overline{X}-\mu}{S/\sqrt{n}}$ is a random variable with **Student's t distribution** *with number of* **degrees of freedom** *$df = n - 1$ (where n is the sample size). We write*

$$T_{n-1} = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim \mathcal{T}_{n-1}$$

Note: This theorem holds for any sample size n.

Student's t-distribution

- If df is large, Student's t distribution becomes $N(0,1)$. Therefore if the sample size is large enough ($n \geq 30$), we have $\frac{\overline{X}-\mu}{S/\sqrt{n}} \overset{\text{approx.}}{\sim} N(0,1)$.
- When the sample size is large, the variability of S is small, and thus S is concentrated around σ. Then $\frac{\overline{X}-\mu}{S/\sqrt{n}}$ approximately has the same distribution with $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$. But when the sample size is small, we have to consider this variability and use the Student's t distribution.

Facts about t-distribution

Fact

Assume we take an independent random sample of size n from an approximately normal distribution with mean μ.
Then $T_{n-1} = \frac{\overline{X}-\mu}{S/\sqrt{n}}$ follows $\mathcal{T}_{n-1}$ with $df = n-1$, and $\mathcal{T}_{n-1}$ is

1. *Bell shaped and symmetric about 0*
2. *has pdf given by*

$$f_{n-1}(x) = \frac{\Gamma(n/2)}{\sqrt{(n-1)\pi}\Gamma((n-1)/2)}\left(1+\frac{x^2}{n-1}\right)^{-n/2}$$

(don't have to remember it)

3. *The limiting behavior of its pdf is*

$$\lim_{n\to\infty} f_{n-1}(x) = \underbrace{\frac{1}{\sqrt{2\pi}}e^{-x^2/2}}_{\text{the pdf of } \mathcal{N}(0,1)}$$

Degrees of freedom

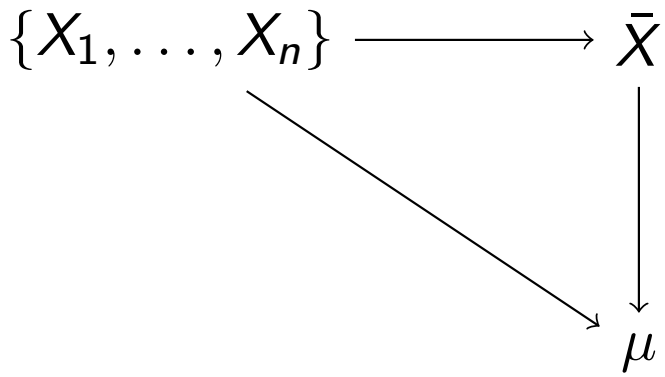
- Note that Student's t-distribution depends on the parameter df, so really, this is a family of distributions.
- When df is small, the distribution of $\mathcal{T}_{df}$ depends more heavily on df.
- As $df \to \infty$, $\mathcal{T}_{df}$ becomes $\mathcal{N}(0, 1)$

Degrees of freedom

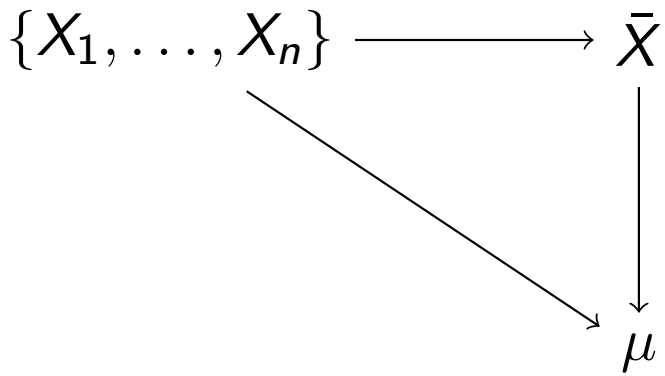
- The $df = n - 1$ comes from the sample std. dev. S (which is random and depends on the sample).
- Informally, df of a parameter can be thought of as the sample size, n, minus the number of independent parameters needed to estimate the variance within your model.
  In the case of S, we need to first estimate μ with $\overline{X}$ *before* we can calculate S. Thus, $df = n - 1$.

degrees of freedom continued

In truth, degrees of freedom is a very nuanced idea which we won't be able to completely understand in this class. But there is some intuition as to why it shows up:



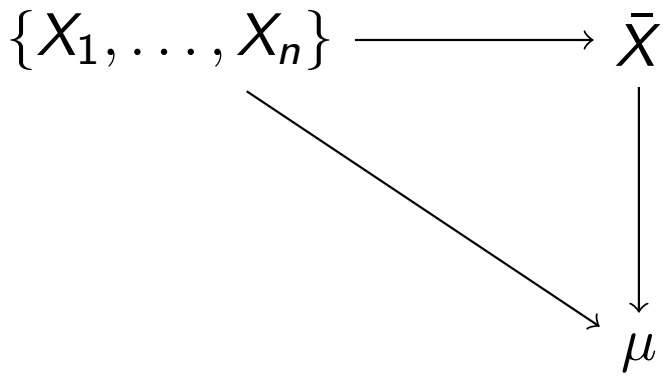
Degrees of freedom continued



The population variance of $\{X_1, \ldots, X_n\}$ is given by

$$\frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2 .$$

Degrees of freedom continued



The population variance of $\{X_1, \ldots, X_n\}$ is given by

$$\frac{1}{n} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2.$$

The variance of $\bar{X}$ we know is σ^2/n.

Degrees of freedom continued

If we assume (reasonably, though not trivial) that the population variance is given by the expected population variance of the sample plus the variance of $\bar{X}$, we get

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} E\left(X_i - \bar{X}\right)^2 + \frac{\sigma^2}{n}.$$

Degrees of freedom continued

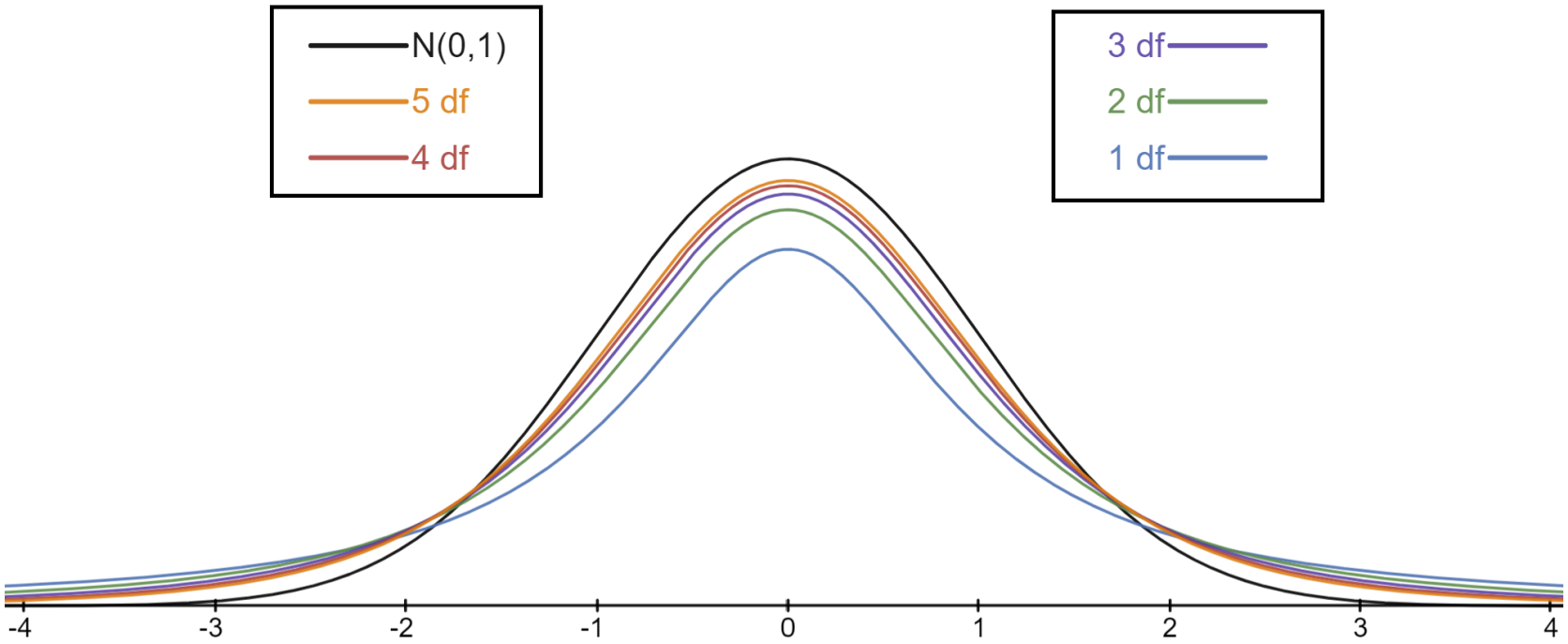
If we assume (reasonably, though not trivial) that the population variance is given by the expected population variance of the sample plus the variance of $\bar{X}$, we get

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} E\left(X_i - \bar{X}\right)^2 + \frac{\sigma^2}{n}.$$

Solving for the expected value gives

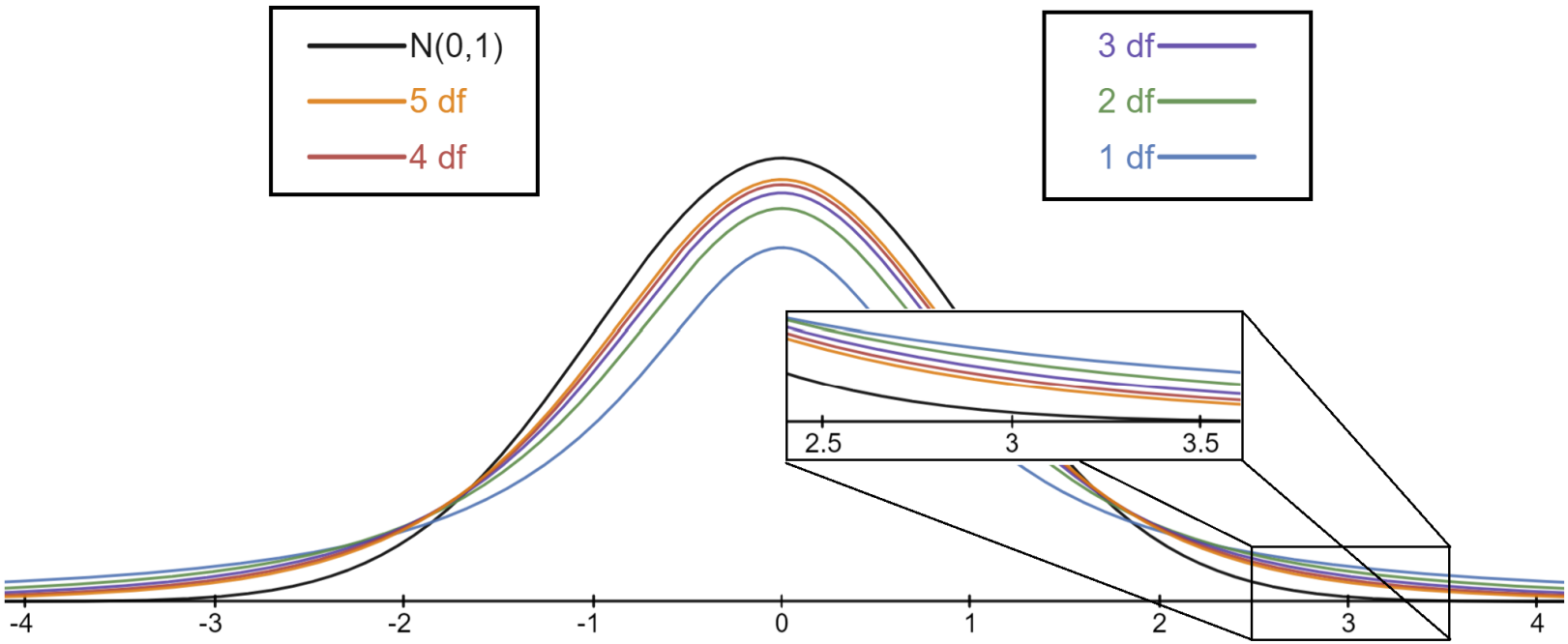
$$\sigma^2 = \frac{n}{n-1}\frac{1}{n}\sum_{i=1}^{n} E\left[\left(X_i - \bar{X}\right)^2\right] = \frac{1}{n-1}\sum_{i=1}^{n} E\left[\left(X_i - \bar{X}\right)^2\right].$$

Degrees of freedom graphically



The larger the df, the closer $\mathcal{T}_{df}$ approximates $\mathcal{N}(0,1)$. And in the limit we have $\mathcal{T}_\infty \sim \mathcal{N}(0,1)$.

Degrees of freedom graphically



The smaller the *df*, the larger the distribution "tails" are. This makes sense, as a smaller sample will have more variability in its sample standard deviation.

Sampling distribution of $T_{n-1} = \frac{\overline{X}-\mu}{S/\sqrt{n}}$ for any sample

Sometimes only a small sample is available sadly, but if we know the population distribution is approximately normal, we can apply the above theorem to have the following sampling distribution of $\frac{\overline{X}-\mu}{S/\sqrt{n}}$.

Theorem (Sampling distribution of $T_{n-1} = \frac{\overline{X}-\mu}{S/\sqrt{n}}$ for any sample)

If the population is approximately normal distributed.
Then we have

$$T_{n-1} = \frac{\overline{X}-\mu}{S/\sqrt{n}} \sim \mathcal{T}_{n-1}$$

Note: when we study a small sample, if the question doesn't mention the population distribution, we always suppose it is approximately normal.

Sampling distribution of $T_{n-1} = \frac{\overline{X}-\mu}{S/\sqrt{n}}$ for any sample

When sampling from an approximately normal pop. with unknown σ:

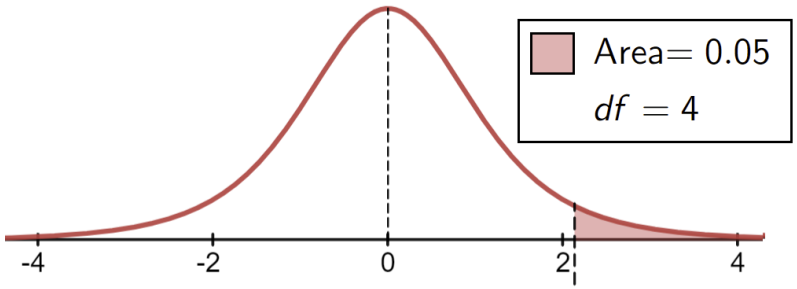
- The test statistic T_{n-1} is a *quotient* of two random variables. This makes things much more complicated.
- When the sample size is large ($n \geq 30$), these complications make very little difference. Which is why we "got away with" it the past two weeks...
- Technically, we *should* always use the t-distribution whenever σ is unknown. However, when the sample size is small, we *MUST* use the t-distribution. (When sample size is large enough, the t-distribution is very close to $N(0,1)$)

t-distribution table

We define $t_{df,\alpha}$ such that

$$P(T_{df} > t_{df,\alpha}) = \alpha$$

We have the Student's t-distribution table to find $t_{df,\alpha}$:



df	$t_{0.1}$	$t_{0.05}$	$t_{0.025}$	$t_{0.01}$	$t_{0.005}$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
⋮	⋮	⋮	⋮	⋮	⋮
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
∞	1.282	1.645	1.960	2.326	2.576

Confidence interval of μ with unknown σ and small sample

Theorem (Confidence interval of μ)

Suppose the population std. dev. σ is unknown.

- *If the population distribution is approximately normal, and the sample size is small ($n < 30$)*

We have with $(1-\alpha)\times 100\%$ confidence that

$$\mu \in CI_{(1-\alpha)\times 100\%} = \left[\overline{x} - t_{n-1,\alpha/2} \times \frac{s}{\sqrt{n}}, \overline{x} + t_{n-1,\alpha/2} \times \frac{s}{\sqrt{n}}\right]$$

Note: when we study a small sample, if the question doesn't mention the population distribution, we always suppose it is approximately normal.

Example 1

Example (Diamonds)

A new process for producing synthetic diamonds can be operated at a profitable level only if the average weight of the diamonds is greater than 0.5 karat. To evaluate the profitability of the process, six diamonds are generated with weights

$$0.46,\ 0.61,\ 0.52,\ 0.48\ ,0.57,\ 0.54$$

karat. You have reason to believe that the weights of the synthetic diamonds are normally distributed. Find a 95% confidence interval for the true average weight μ of the synthetic diamonds.

Solution

Our sample size is $n = 6 < 30$, which means we have

$$df = n - 1 = 5.$$

Solution

Our sample size is $n = 6 < 30$, which means we have

$$df = n - 1 = 5.$$

For this problem, $\alpha = 0.05$, so that

$$t_{n-1,\alpha/2} = t_{5,0.025} = 2.571.$$

Solution

Our sample size is $n = 6 < 30$, which means we have

$$df = n - 1 = 5.$$

For this problem, $\alpha = 0.05$, so that

$$t_{n-1,\alpha/2} = t_{5,0.025} = 2.571.$$

Next we compute the sample average and std. dev.

$$\overline{x} = 0.53, \quad \text{and} \quad s \approx 0.0559.$$

So the 95% CI is

$$\overline{x} \pm t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} \approx 0.53 \pm 2.571 \cdot \frac{0.0559}{\sqrt{6}} = (0.4713, 0.5887)$$

Example 2, Simulating Confidence Intervals

Let's compare our answer in Example 1 with what one would get if using $z_{\alpha/2}$:

	$z_{0.025}$	$t_{5,0.025}$
95% CI:	$(0.4853, 0.5747)$	$(0.4713, 0.5887)$

Notice that the 95% CI using the t-distribution is wider than using the normal distribution.

Example 2, Simulating Confidence Intervals

Let's compare our answer in Example 1 with what one would get if using $z_{\alpha/2}$:

	$z_{0.025}$	$t_{5,0.025}$
95% CI:	$(0.4853, 0.5747)$	$(0.4713, 0.5887)$

Notice that the 95% CI using the t-distribution is wider than using the normal distribution.

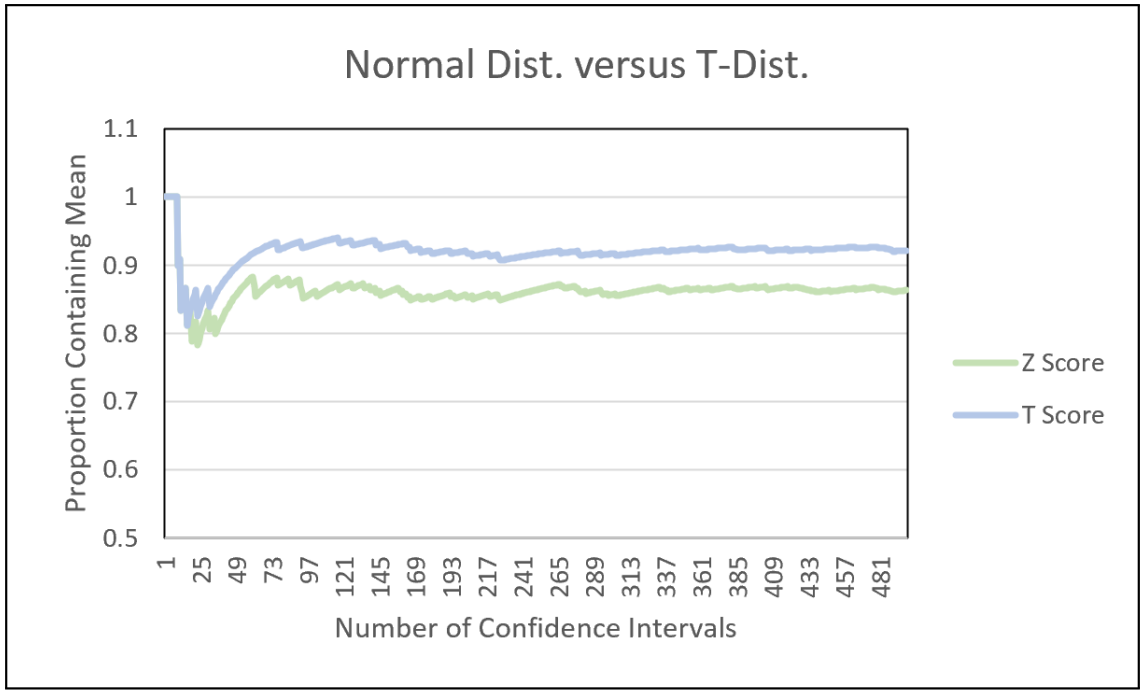
This is caused by the larger "tails" of the t-distribution. Intuitively, we have to widen the interval to get the same amount of confidence in the point estimator when we have a small sample.

Example 2, Simulating Confidence Intervals; Continued

- You might be wondering, are we being *too* safe by using t instead of z? Let's test it!

Example 2, Simulating Confidence Intervals; Continued

- You might be wondering, are we being *too* safe by using t instead of z? Let's test it!
- In the `Z_Versus_T` Excel file on Canvas, we simulate 500 random samples ($n = 6$) from a $\mathcal{N}(0, 1)$ distribution. We compare the CI given by the z-score with the one given by the t-score; Which one looks closer to 95% confidence?



Confidence interval of μ with unknown σ and small sample

Takeaways:

We do the same procedure to calculate a $(1-\alpha)100\%$ CI for small samples as we did before. Except we use the $t_{n-1,\alpha/2}$ to calculate the margin of error (with the appropriate $df = n-1$) instead of $z_{\alpha/2}$.

Hypothesis testing for μ with unknown σ and small sample

Hypothesis testing for μ

Suppose the population std. dev. σ is unknown.

- If the population distribution is approximately normal, and the sample size is small ($n < 30$)

Then the null distribution for the test statistic

$$T_{n-1} = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim \mathcal{T}_{n-1}$$

We compute the test statistic for the given sample as

$$t_{n-1} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Hypothesis testing for μ with unknown σ and small sample

(continued)
Then we obtain rejection region R_0 from $\mathcal{T}_{n-1}$

- For $H_a : \mu > \mu_0 \Rightarrow R_0 = [t_{n-1,\alpha}, \infty)$
- For $H_a : \mu < \mu_0 \Rightarrow R_0 = (-\infty, -t_{n-1,\alpha}]$
- For $H_a : \mu \neq \mu_0$
  $\Rightarrow R_0 = (-\infty, -t_{n-1,\alpha/2}] \cup [t_{n-1,\alpha/2}, \infty)$

where $t_{n-1,\alpha}$ or $t_{n-1,\alpha/2}$ can be found from t-distribution table.

Note: Unlike the z-score, the exact p-value can't be found in the t-distribution table (it can be calculated in Excel, however), so we always choose to get the rejection region.

Example 3

Example

A paint manufacturer claims that a 1-gallon can of its paint will cover 400 ft.^{2} of surface area. To avoid false advertisement complaints. The manufacturer tests a random sample of 10 1-gallon cans to paint 10 identical surfaces with the same equipment. The actual areas (in ft.^{2}) covered are given below:

310	311	412	368	447
376	303	365	410	350

Does the data present sufficient evidence to indicate that the actual average differs from 400? Use $\alpha = 0.05$.

Solution

1. Establish null and alternative hypothesis

$$H_0 : \mu = 400, \quad \text{versus} \quad H_a : \mu \neq 400.$$

The null value $\mu_0 = 400$

2. Check the conditions and specify the null distribution of the test statistic

$n = 10 < 30$ and when we study a small sample we always suppose the population is approximately normal. Thus

$$T_{n-1} = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \sim \mathcal{T}_{n-1}$$

3. Compute the test statistic. We get the sample mean and std. dev.:

$$\overline{x} = 365.2, \quad \text{and} \quad s \approx 48.4167.$$

Then the test statistic is

$$t_{n-1} = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = \frac{365.2 - 400}{48.417/\sqrt{10}} \approx -2.27.$$

Solution; Continued

4. Find rejection region R_0 with respect to the direction of H_a

$$R_0 = (-\infty, -t_{n-1,\alpha/2}] \cup [t_{n-1,\alpha/2}, \infty)$$
$$= (-\infty, -t_{9,0.025}] \cup [t_{9,0.025}, \infty) = (-\infty, -2.262] \cup [2.262, \infty)$$

df	$t_{0.1}$	$t_{0.05}$	$t_{0.025}$	$t_{0.01}$	$t_{0.005}$
1	3.078	6.314	12.706	31.821	63.657
⋮	⋮	⋮	⋮	⋮	⋮
9	1.383	1.833	2.262	2.821	3.250
⋮	⋮	⋮	⋮	⋮	⋮

5. Decide whether the test statistic belongs to the rejection region
Since $t_{n-1} = -2.27 \in R_0$, we can reject H_0
6. Make conclusion
We reject H_0 at significance level $\alpha = 0.05$, and we have evidence to conclude that the actual average differs from 400

Difference between population means

- Like last week, we may be interested in the difference of two populations means $\mu_1 - \mu_2$ from two different populations.
- Last week, when the population std. dev. are unknown, we used the test statistic

$$\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

which has approximately $\mathcal{N}(0, 1)$ distribution for large samples $n_1 \geq 30, n_2 \geq 30$.

Difference between population means

- Like last week, we may be interested in the difference of two populations means $\mu_1 - \mu_2$ from two different populations.
- Last week, when the population std. dev. are unknown, we used the test statistic

$$\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

  which has approximately $\mathcal{N}(0, 1)$ distribution for large samples $n_1 \geq 30, n_2 \geq 30$.
- Based on what we've done so far...You might think for small samples that this will have a t-distribution.

Difference between population means

- Like last week, we may be interested in the difference of two populations means $\mu_1 - \mu_2$ from two different populations.
- Last week, when the population std. dev. are unknown, we used the test statistic

$$\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

  which has approximately $\mathcal{N}(0, 1)$ distribution for large samples $n_1 \geq 30, n_2 \geq 30$.
- Based on what we've done so far...You might think for small samples that this will have a t-distribution.
- Unfortunately, this is not necessarily the case!

Std. dev. of sample mean difference

- In order to know for sure what the distribution of the test statistic is, we must make one more assumption:

> *Both populations are normal with the same variance*
>
> $$\sigma_1 = \sigma_2$$

Std. dev. of sample mean difference

- In order to know for sure what the distribution of the test statistic is, we must make one more assumption:

> *Both populations are normal with the same variance*
>
> $$\sigma_1 = \sigma_2$$

- Given this $\uparrow$ condition, the test statistic becomes

$$\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

and it follows a t-distribution with $df = n_1 + n_2 - 2$, where the random variable S^2 is the sample estimate of the common variance σ^2 for our two samples.

Estimating the common σ^2

- Given two samples with sample sizes n_1 and n_2, and sample variances of s_1^2 and s_2^2, respectively. Now we compute s^2

Estimating the common σ^2

- Given two samples with sample sizes n_1 and n_2, and sample variances of s_1^2 and s_2^2, respectively. Now we compute s^2
- We use the *weighted average* of s_1^2 and s_2^2 to get s^2

$$s^2 = \underbrace{\frac{(n_1 - 1)}{n_1 + n_2 - 2}}_{\text{weight of } s_1^2} \times s_1^2 + \underbrace{\frac{(n_2 - 1)}{n_1 + n_2 - 2}}_{\text{weight of } s_2^2} \times s_2^2$$

$$= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Explaining the *df*

The denominator $n_1 + n_2 - 2$ is also the *df* of the test statistic. Why is there a "-2" in the *df*?

Explaining the *df*

The denominator $n_1 + n_2 - 2$ is also the *df* of the test statistic. Why is there a "-2" in the *df*?

- Recall that, informally, *df* of a parameter can be thought of as the sample size minus the number of independent parameters needed to estimate the parameter.
- S_1^2 has $df = n_1 - 1$, since the sample size is n_1, and we need to use $\overline{x}_1$ to calculate it.
- S_2^2 has $df = n_2 - 1$ for the same reasoning.
- Similarly, to calculate s^2, we have a "pooled" sample of $n_1 + n_2$ and we need to use $\overline{x}_1$ *and* $\overline{x}_2$ to compute it. Hence

$$df = n_1 + n_2 - 2.$$

Explaining the *df*

The denominator $n_1 + n_2 - 2$ is also the *df* of the test statistic. Why is there a "-2" in the *df*?

- Recall that, informally, *df* of a parameter can be thought of as the sample size minus the number of independent parameters needed to estimate the parameter.
- S_1^2 has $df = n_1 - 1$, since the sample size is n_1, and we need to use $\overline{x}_1$ to calculate it.
- S_2^2 has $df = n_2 - 1$ for the same reasoning.
- Similarly, to calculate s^2, we have a "pooled" sample of $n_1 + n_2$ and we need to use $\overline{x}_1$ *and* $\overline{x}_2$ to compute it. Hence

$$df = n_1 + n_2 - 2.$$

Notice that the *df* of the sum is the sum of the *df*, this is not a coincidence and we will see it again!

Sampling distribution of $\frac{(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)}{\sqrt{S^2(1/n_1+1/n_2)}}$ when $\sigma_1 = \sigma_2$

Theorem (Sampling distribution of $\frac{(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)}{\sqrt{S^2(1/n_1+1/n_2)}}$ when $\sigma_1 = \sigma_2$)

If the two populations are approximately normal with the same variance $\sigma_1 = \sigma_2$.
Then we have

$$T_{n_1+n_2-2} = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim \mathcal{T}_{n_1+n_2-2}$$

where $S = \sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}$

Confidence interval of $\mu_1 - \mu_2$ with unknown $\sigma_1 = \sigma_2$ and small samples

Theorem (Confidence interval of $\mu_1 - \mu_2$)

Suppose the population std. dev.s σ_1, σ_2 are unknown. If

- *the populations are approximately normal, and the sample sizes are small ($n_1 < 30$ or $n_2 < 30$)*
- *And we know the population variances are equal, $\sigma_1 = \sigma_2$*

We have with $(1-\alpha) \times 100\%$ confidence that

$$\mu_1 - \mu_2 \in CI_{(1-\alpha)\times 100\%}$$

$$= \left[(\overline{x}_1 - \overline{x}_2) - t_{n_1+n_2-2,\alpha/2} \times s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \right.$$

$$\left. (\overline{x}_1 - \overline{x}_2) + t_{n_1+n_2-2,\alpha/2} \times s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right]$$

where $s = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$

Rule of Thumb: How to know $\sigma_1^2 = \sigma_2^2$?

- If population variances are unknown, then how do we know when it's okay to assume they are equal?
- In the next chapter, we will see a method of testing against $H_0 : \sigma_1^2 = \sigma_2^2$ by comparing the ratio of sample variances S_1^2/S_2^2.
- For now, we will use the following rule of thumb:

> If neither sample variance is three times larger than the other, i.e.
>
> $$\frac{1}{3} < \frac{S_1^2}{S_2^2} < 3,$$
>
> we will go ahead with the assumption that $\sigma_1^2 = \sigma_2^2$.

Example 4

Example (Online Versus In Person)

A course can either be taken online or attending lectures. Suppose the two populations, scores for online group and classroom group, have the equal variances, and we take the following two random samples

Online		**Classroom**	
32	37	35	31
25	28	29	25
41	44	34	40
35	31	27	32
32		31	

Let μ_1 and μ_2 be the mean scores for online group and classroom group, respectively. Find the 90% confidence interval for $\mu_1 - \mu_2$

Solution

We have $n_1 = 9 < 30$ and $n_2 = 9 < 30$. And when we study small samples we always suppose the populations are normal. Notice the two populations have the equal variance.
We calculate:

$$\begin{array}{cc} \overline{x}_1 = 35.22 & \overline{x}_2 = 31.56 \\ s_1^2 \approx 24.4441 & s_2^2 \approx 20.0274 \end{array}$$

Then we estimate the common std. dev.

$$\begin{aligned} s^2 &= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \\ &\approx \frac{8 \cdot (24.4441 + 20.0274)}{16} \\ &\approx 22.2361. \end{aligned}$$

and $s \approx 4.716$.

Solution

With $\alpha = 0.1$, we have $t_{n_1+n_2-2,\alpha/2} = t_{16,0.05} \approx 1.746$

Finally the 90% CI is

$$(35.22 - 31.56) \pm 1.746 \times 4.716\sqrt{\frac{1}{9} + \frac{1}{9}} = (-0.22, 7.55).$$

Hypothesis testing for $\mu_1 - \mu_2$ with unknown $\sigma_1 = \sigma_2$ and small samples

Hypothesis testing for $\mu_1 - \mu_2$

Suppose the populations std. dev. are unknown. If

- the populations are approximately normal, and the sample sizes are small ($n_1 < 30$ or $n_2 < 30$)
- And the population variances are equal, $\sigma_1 = \sigma_2$

Then the null distribution for the test statistic

$$T_{n_1+n_2-2} = \frac{(\bar{X}_1 - \bar{X}_2) - \Delta_0}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim \mathcal{T}_{n_1+n_2-2}$$

We compute the test statistic for the given sample as

$$t_{n_1+n_2-2} = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Hypothesis testing for $\mu_1 - \mu_2$ with unknown σ and small sample

(continued)
Then we obtain rejection region R_0 from $\mathcal{T}_{n_1+n_2-2}$

- For $H_a : \mu_1 - \mu_2 > \Delta_0 \Rightarrow R_0 = [t_{n_1+n_2-2,\alpha}, \infty)$
- For $H_a : \mu_1 - \mu_2 < \Delta_0 \Rightarrow R_0 = (-\infty, -t_{n_1+n_2-2,\alpha}]$
- For $H_a : \mu_1 - \mu_2 \neq \Delta_0$
  $\Rightarrow R_0 = (-\infty, -t_{n_1+n_2-2,\alpha/2}] \cup [t_{n_1+n_2-2,\alpha/2}, \infty)$

where $t_{n_1+n_2-2,\alpha}$ or $t_{n_1+n_2-2,\alpha/2}$ can be found from t-distribution table.

Example 5

Example (Online Versus In-Person; Continued)

For example 4. At significance level $\alpha = 0.05$, is this enough to conclude that there is a difference in exam scores between the course delivery methods?

Solution

1. Establish null and alternative hypothesis

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{against} \quad H_a : \mu_1 - \mu_2 \neq 0.$$

Null value $\Delta_0 = 0$

2. Check the conditions and specify the null distribution of the test statistic

We have $n_1 = 9 < 30$ and $n_2 = 9 < 30$. And when we study small samples we always suppose the populations are normal. Notice the two populations have the equal variance. Thus

$$T_{n_1+n_2-2} = \frac{(\bar{X}_1 - \bar{X}_2) - \Delta_0}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim \mathcal{T}_{n_1+n_2-2}$$

Solution; Continued

3. Compute the test statistic

$$t_{n_1+n_2-2} = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{(35.22 - 31.56) - 0}{4.716\sqrt{\frac{1}{9} + \frac{1}{9}}} \approx 1.65.$$

4. Find rejection region R_0 with respect to the direction of H_a

$$\begin{aligned} R_0 &= (-\infty, -t_{n_1+n_2-2,\alpha/2}] \cup [t_{n_1+n_2-2,\alpha/2}, \infty) \\ &= (-\infty, -t_{16,0.025}] \cup [t_{16,0.025}, \infty) \\ &= (-\infty, -2.12] \cup [2.12, \infty) \end{aligned}$$

Solution; Continued

5. Decide whether the test statistic belongs to the rejection region

$$t_{n_1+n_2-2} = 1.65 \notin R_0$$

Thus we fail to reject H_0

6. Make conclusion

We fail to reject H_0 at significance level $\alpha = 0.05$, and we have no evidence to conclude that there is a difference in exam scores between the course delivery methods.

Paired-difference

- Suppose you want to compare the wearing qualities for two types of tires A and B.
- You place 5 type A tires on 5 different cars and drive them for a specified number of miles, then record the wear of the tires.
- You then replace all the Type A tires with Type B tires on the same 5 cars and drive them for the same number of miles, then record the wear.

Car	Tire A	Tire B
1	10.6	10.2
2	9.8	9.4
3	12.3	11.8
4	9.7	9.1
5	8.8	8.3
Mean	10.24	9.76
Variance	1.733	1.763

An issue:

If we proceeded in the same way as in Example 5, then we would have an observed test stat. of

$$\frac{10.24 - 9.76}{\sqrt{\frac{1.733+1.763}{2} \times \frac{2}{5}}} = 0.5740...$$

An issue:

If we proceeded in the same way as in Example 5, then we would have an observed test stat. of

$$\frac{10.24 - 9.76}{\sqrt{\frac{1.733+1.763}{2} \times \frac{2}{5}}} = 0.5740...$$

For $H_a : \mu_A - \mu_B \neq 0$, this has a p-value of more than 0.5 (since $t_{4,0.25} = 0.7410 > 0.574$).

An issue:

If we proceeded in the same way as in Example 5, then we would have an observed test stat. of

$$\frac{10.24 - 9.76}{\sqrt{\frac{1.733+1.763}{2} \times \frac{2}{5}}} = 0.5740...$$

For $H_a : \mu_A - \mu_B \neq 0$, this has a p-value of more than 0.5 (since $t_{4,0.25} = 0.7410 > 0.574$).
But, we should notice that *every* car had more wear on its Type A tires than Type B. If $H_0 : \mu_A - \mu_B = 0$ were true, this would occur with probability $2^{-5} = 0.03125$.

Paired-Difference

- The problem of course is that there is dependence between the cars! There is a one-to-one relationship between the two samples. This is a **paired-difference**.
- We should first compute the difference for each pair and then deal with it as only one sample from one population.
- For each car i we look at the difference

$$d_i = (\text{wear from tire } A) - (\text{wear from tire } B)$$

Then this turns the problem into a single sample test.

Example 6

Example (Tires)

Test whether there is a significant difference in the average wear between each type of tire:

Car	Tire A	Tire B	Difference
1	10.6	10.2	0.4
2	9.8	9.4	0.4
3	12.3	11.8	0.5
4	9.7	9.1	0.6
5	8.8	8.3	0.5
Mean			$\overline{d} = 0.48$
Std. Dev.			$s_d \approx 0.0837$

Solution

We can now proceed in the same way as Example 3 with hypothesis

$$H_0 : \mu_d = 0 \quad \text{against} \quad H_a : \mu_d \neq 0.$$

The null value $\mu_0 = 0$

So we have the test statistic

$$t_{n-1} = \frac{\overline{d} - \mu_0}{s_d/\sqrt{n}} = \frac{0.48 - 0}{0.0837/\sqrt{5}} \approx 12.8233$$

Certainly this is enough to reject H_0 at any reasonable significance level α.