Inference about a Population Mean (μ)

AAO unit 26; Baldi & Moore, Ch 17

Sahir Bhatnagar and James Hanley

EPIB 607
Department of Epidemiology, Biostatistics, and Occupational Health
McGill University

sahir.bhatnagar@mcgill.ca
https://sahirbhatnagar.com/EPIB607/

September 30, 2018

The t distribution

Inference for μ when σ is not known

Up until now, all of our calculations have relied on us knowing the value of the population standard deviation (σ). It is rare that this is the case.

We now consider methods of inference for when σ is unknown.

When σ is unknown, we must estimate it from the data using s, the sample standard deviation.

Inference for μ when σ is unknown

- When the true variance was known, we performed our calculations using the standardization

$$Z = \frac{\bar{y} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

Inference for μ when σ is unknown

- When the true variance was known, we performed our calculations using the standardization

$$Z = \frac{\bar{y} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

- We no longer can use this, so instead we use

$$t = \frac{\bar{y} - \mu}{s/\sqrt{n}} \sim t_{(n-1)}$$

  which follows a *t*-**distribution** with $n - 1$ degrees of freedom based on the n values, $y_1, ..., y_n$ in a simple random sample

Inference for μ when σ is unknown

- When the true variance was known, we performed our calculations using the standardization

$$Z = \frac{\bar{y} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

- We no longer can use this, so instead we use

$$t = \frac{\bar{y} - \mu}{s/\sqrt{n}} \sim t_{(n-1)}$$

  which follows a t-**distribution** with $n-1$ degrees of freedom based on the n values, $y_1, ..., y_n$ in a simple random sample

- There is a different t distribution for each sample size. The degrees of freedom specify which distribution we use, and are determined by the denominator used in estimating s which is $(n-1)$.

A note about the conditions

- B&M stress that the **first** of their conditions as *very important*: *we can regard* our data as a simple random sample (SRS) from the population
- The **second**, observations from the population have a *<u>Normal</u>* distribution with unknown mean parameter μ and unknown standard deviation parameter σ less so
- *In practice*, inference procedures *can accommodate some deviations from the Normality condition* when the sample is large enough. (think CLT)

t distributions

The t distribution is symmetric, but has heavier tails than the Normal distribution.

As the degrees of freedom increase (i.e., as n increases), the t-distribution becomes more and more similar to a Normal distribution.

In fact, the quantiles/area under the curve are similar for $n \geq 30$:

Quantiles

Distribution	Cumulative Probability			
	0.005	0.010	0.025	0.050
Normal	-2.58	-2.33	-1.96	-1.64
$t(50)$	-2.68	-2.40	-2.01	-1.68
$t(30)$	-2.75	-2.46	-2.04	-1.70
$t(10)$	-3.17	-2.76	-2.23	-1.81

t procedures

We can calculate CIs and perform significance tests much as before (example coming up soon).

A significance test of a single sample mean using the t-statistic is called a one-sample t-test.

Collectively, the significance tests and confidence-interval based tests using the t distribution are called t procedures. (Tests using the Z statistic and the Normal distribution are a special case of these.)

Robustness of the t procedures

A statistical procedure is said to be **robust** if it is insensitive to violations of the assumptions made.

- Results of t procedures (CIs, significance tests) are exact if the population from which the simple random sample was drawn is Normal.

Robustness of the t procedures

A statistical procedure is said to be **robust** if it is insensitive to violations of the assumptions made.

- Results of t procedures (CIs, significance tests) are exact if the population from which the simple random sample was drawn is Normal.
- t procedures are not robust against *extreme* skewness, particularly in small samples, since the procedures are based on using $\bar{x}$ and s^2 (which are sensitive to outliers).
- Recall: Unless there is a very compelling reason (e.g. known/confirmed error in the recorded data), outliers should not be discarded.

Robustness of the t procedures

A statistical procedure is said to be **robust** if it is insensitive to violations of the assumptions made.

- Results of t procedures (CIs, significance tests) are exact if the population from which the simple random sample was drawn is Normal.
- t procedures are not robust against *extreme* skewness, particularly in small samples, since the procedures are based on using $\bar{x}$ and s^2 (which are sensitive to outliers).
- Recall: Unless there is a very compelling reason (e.g. known/confirmed error in the recorded data), outliers should not be discarded.

Robustness of the t procedures

- t procedures **are** robust against other forms of non-normality and, even with considerable skew, perform well when n is large. Why?

Robustness of the t procedures

- t procedures **are** robust against other forms of non-normality and, even with considerable skew, perform well when n is large. Why?
  - When n is large, s^2 is a good estimate of σ^2 (recall that s^2 is unbiased and, like most estimates, precision improves with increasing sample size)
  - CLT: $\bar{x}$ will be Normal when n is large, even if the population data are not

- …so how large is large enough?

Robustness of the t procedures

- When $n < 15$ and the data are symmetric and do not exhibit outliers, t procedures can be used

Robustness of the t procedures

- When $n < 15$ and the data are symmetric and do not exhibit outliers, t procedures can be used
- When $n \geq 15$ and the data do not exhibit extreme outliers or extreme skew, t procedures can be used

Robustness of the t procedures

- When $n < 15$ and the data are symmetric and do not exhibit outliers, t procedures can be used
- When $n \geq 15$ and the data do not exhibit extreme outliers or extreme skew, t procedures can be used
- When $n \geq 40$, t procedures can be used even if the data are highly skewed
  - Even when σ is not known, can use tests/CIs based on the Z statistic/Normal distribution rather than on the t statistic/t distribution. (Quantiles are very similar.)

Robustness of the t procedures

- When $n < 15$ and the data are symmetric and do not exhibit outliers, t procedures can be used
- When $n \geq 15$ and the data do not exhibit extreme outliers or extreme skew, t procedures can be used
- When $n \geq 40$, t procedures can be used even if the data are highly skewed
    - Even when σ is not known, can use tests/CIs based on the Z statistic/Normal distribution rather than on the t statistic/t distribution. (Quantiles are very similar.)

- If the data are very skewed, there are other options to consider (whether n is large or not):
    - Transformations (be careful with interpretation!)
    - Non-parametric models which do not assume any particular distribution of the data (we will study these in Part 5 of the course)

Confidence intervals for μ when σ^2 is unknown

When we must estimate the SE, the $(1-\alpha)100\%$ confidence interval now formed is:

$$\begin{aligned} \text{CI:} \quad & \text{estimate} \pm \text{margin of error} \\ & = \bar{x} \pm t_{\alpha/2} s/\sqrt{n} \end{aligned}$$

where $t_{\alpha/2}$ is the quantile of the $t(n-1)$ distribution curve such that $P(-t_{\alpha/2} < t_{n-1} < t_{\alpha/2}) = 1-\alpha$.

The CI is exactly correct when the distribution of the population data x is Normal, and approximately correct otherwise.

Recall that with highly skewed distributions and small samples, the CLT may not provide a good approximation to the distribution of $\bar{x}$.

Confidence intervals for μ when σ^2 is unknown

Exercise:
Compute a 95% and a 99% confidence interval for the mean Wechsler score of children at Lake Wobegon assuming variance is unknown.

Recall, for a 95% *Normal* CI, we use $z_{\alpha/2} = z_{0.025} = 1.96$, and for a 99% *Normal* CI, we have $z_{\alpha/2} = z_{0.005} = 2.58$. What quantile values $t_{\alpha/2}$ are used to construct CIs if we don't know the variance?

Confidence intervals for μ when σ^2 is unknown

Exercise:
Compute a 95% and a 99% confidence interval for the mean Wechsler score of children at Lake Wobegon assuming variance is unknown.

Components required to construct the CIs:

- $\bar{x} = 112.8$
- $s = \sqrt{160.4} = 12.7$
- $n = 9$
- for a 95% CI, we use $t_{\alpha/2} = t_{0.025} = 2.306$
- for a 99% CI, we use $t_{\alpha/2} = t_{0.005} = 3.355$

Inference about a population mean

(FREQUENTIST) INFERENCE for (PARAMETER) μ – the mean of an (effectively) infinite-size universe of Y values – based on the n values, $y_1, \ldots, y_n$ in an SRS from that universe/population

‘***Certain conditions apply***’ [1]

Point-estimate of μ : $\hat{\mu} = \bar{y}$

(Symmetric) Confidence Interval CI for μ : $\bar{y} \pm ME$,

where the Margin of Error (ME) is a

- z-multiple of SE [2], if n is ‘large’ AND

  the Y values in the universe have a ‘Normal’ (Gaussian’) distribution or, if not, n is large enough so that the Central Limit Theorem guarantees that the sampling distribution of possible $\bar{y}$’s of size n from this universe is well enough approximated by a Gaussian distribution
- t-multiple of SE if ‘small’ n AND

  the Y values in the universe have a ‘Normal’ (Gaussian’) distribution

[1] B&M stress that the **first** of their conditions ‘*we can regard* our data as a

t-distribution

When and why we use the t-distribution

- When σ is unknown use t distribution. but why?

When and why we use the t-distribution

- When σ is unknown use t distribution. but why?
- the spread of the t distribution is greater than $\mathcal{N}(0, 1)$

Rejecting the Null ($H_0 : \mu = \mu_0$) when σ is known

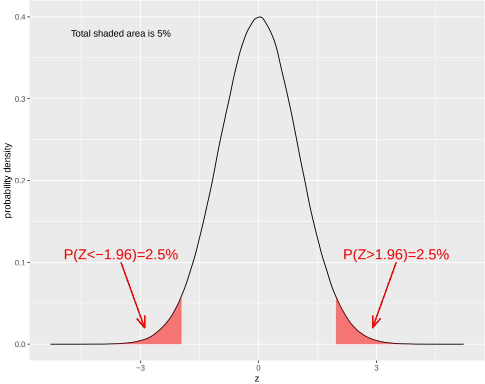
$$\underbrace{z_{0.975}}_{\text{critical value}} = 1.96 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \to \frac{1.96\sigma}{\sqrt{n}} = \bar{x} - \mu_0$$

which means that to reject H_0 the difference between your sample mean and μ_0 needs to be greater than $\frac{1.96}{\sqrt{n}}$ standard deviations

Rejecting the Null ($H_0 : \mu = \mu_0$) when σ is known

$$\underbrace{z_{0.975}}_{\text{critical value}} = 1.96 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \rightarrow \frac{1.96\sigma}{\sqrt{n}} = \bar{x} - \mu_0$$

which means that to reject H_0 the difference between your sample mean and μ_0 needs to be greater than $\frac{1.96}{\sqrt{n}}$ standard deviations



Rejecting the Null ($H_0 : \mu = \mu_0$) when σ is unknown

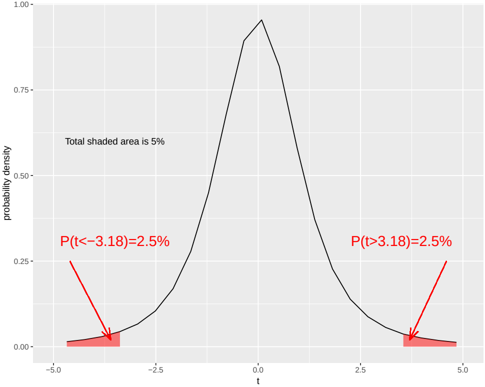
$$\underbrace{t_{0.975,df=3}}_{\text{critical value}} = 3.18 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \rightarrow 3.18\frac{s}{\sqrt{n}} = \bar{x} - \mu_0$$

which means that to reject H_0 the difference between your sample mean and μ_0 needs to be greater than $\frac{3.18}{\sqrt{n}}$ standard deviations

Rejecting the Null ($H_0 : \mu = \mu_0$) when σ is unknown

$$\underbrace{t_{0.975,df=3}}_{\text{critical value}} = 3.18 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \rightarrow 3.18\frac{s}{\sqrt{n}} = \bar{x} - \mu_0$$

which means that to reject H_0 the difference between your sample mean and μ_0 needs to be greater than $\frac{3.18}{\sqrt{n}}$ standard deviations



Summary of t distribution

- Its harder to reject the null when using the t distribution

Summary of t distribution

- Its harder to reject the null when using the t distribution
- This is due to our uncertainty about the estimated variance

Summary of t distribution

- Its harder to reject the null when using the t distribution
- This is due to our uncertainty about the estimated variance
- Larger samples lead to more accurate estimates of σ

Summary of t distribution

- Its harder to reject the null when using the t distribution
- This is due to our uncertainty about the estimated variance
- Larger samples lead to more accurate estimates of σ
- This is reflected in the fact that there is a different t distribution for each sample size

Summary of t distribution

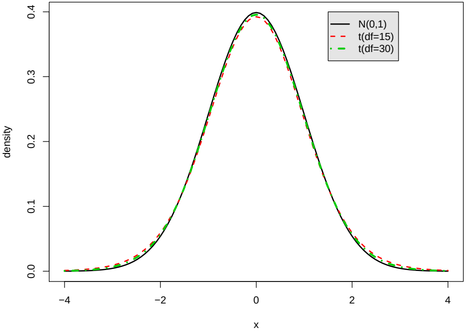
- Its harder to reject the null when using the t distribution
- This is due to our uncertainty about the estimated variance
- Larger samples lead to more accurate estimates of σ
- This is reflected in the fact that there is a different t distribution for each sample size
- As $n \to \infty$, sample variance S gets closer to σ

Summary of t distribution

- Its harder to reject the null when using the t distribution
- This is due to our uncertainty about the estimated variance
- Larger samples lead to more accurate estimates of σ
- This is reflected in the fact that there is a different t distribution for each sample size
- As $n \to \infty$, sample variance S gets closer to σ
- As degrees of freedom increase, t distribution gets closer to Normal distribution

Summary of t distribution

Sample size increases $\rightarrow$ degrees of freedom increase $\rightarrow$ t starts to look like $\mathcal{N}(0, 1)$



As df (proxy to the sample size) increases . . .

Recall that the 97.5th quantile of the $\mathcal{N}(0,1)$ is $Z = 1.96$

As df (proxy to the sample size) increases . . .

Recall that the 97.5th quantile of the $\mathcal{N}(0,1)$ is $Z = 1.96$

df	$t_{0.975,df}$
3	3.1824463
5	2.5705818
30	2.0422725
50	2.0085591
100	1.9839715
250	1.9694984
500	1.9647198