# Inference about a Population Mean ($\mu$)

## 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/

October 1, 2018

# The t distribution

# Inference for $\mu$ when $\sigma$ is not known

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

We now consider methods of inference for when $\sigma$ is unknown.

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

# Inference for $\mu$ when $\sigma$ 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 $\mu$ when $\sigma$ 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 an SRS

# Inference for $\mu$ when $\sigma$ 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 an SRS
- 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 $(n - 1)$.

# $\sigma$ known vs. unknown

| $\sigma$ | known | unknown |
|---|---|---|
| Data | $\{y_1, y_2, ..., y_n\}$ | $\{y_1, y_2, ..., y_n\}$ |
| Pop'n param | $\mu$ | $\mu$ |
| Estimator | $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ | $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ |
| SD | $\sigma$ | $s = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$ |
| SEM | $\sigma/\sqrt{n}$ | $s/\sqrt{n}$ |
| $(1-\alpha)100\%$ CI | $\bar{y} \pm z^{\star}_{1-\alpha/2}(\text{SEM})$ | $\bar{y} \pm t^{\star}_{1-\alpha/2,(n-1)}(\text{SEM})$ |
| test statistic | $\frac{\bar{y}-\mu}{\text{SEM}} \sim \mathcal{N}(0,1)$ | $\frac{\bar{y}-\mu}{\text{SEM}} \sim t_{(n-1)}$ |

# $t$ distribution vs. Normal distribution

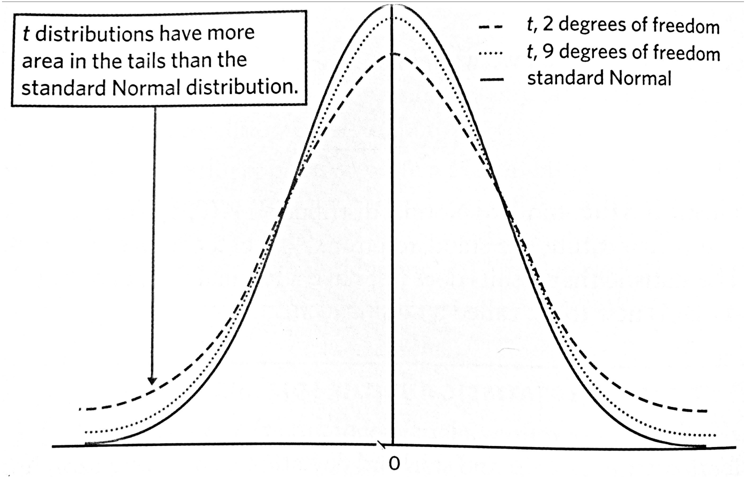
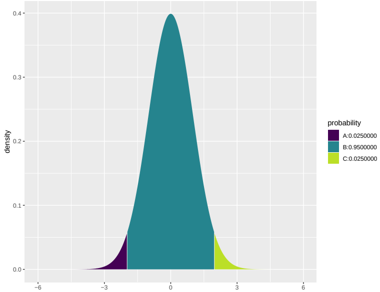



Fig.: Density curves for the $t$ distribution with 2 and 9 degrees of freedom and for the standard Normal distribution. All are symmetric with center 0. The $t$ distributions are somewhat more spread out.

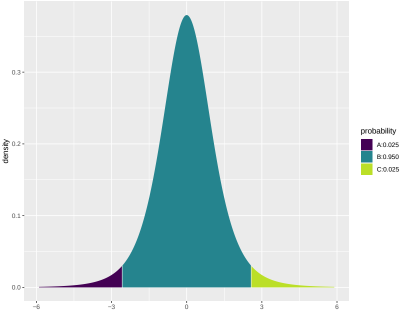
# $t_{(5)}$ distribution vs. Standard Normal distribution

```r
library(mosaic)
xqnorm(p = c(0.025, 0.975))
```




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

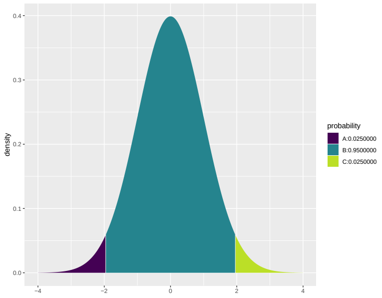
```r
library(mosaic)
xqt(p = c(0.025, 0.975), df = 5)
```




```
## [1] -2.570582  2.570582
```

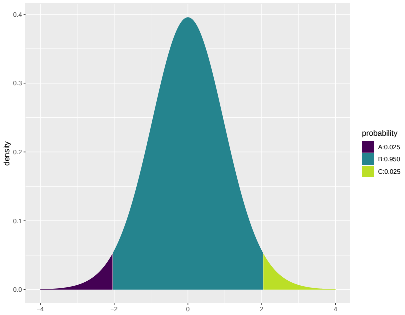
# $t_{(30)}$ distribution vs. Standard Normal distribution

```
library(mosaic)
xqnorm(p = c(0.025, 0.975))
```




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

```
library(mosaic)
xqt(p = c(0.025, 0.975), df = 30)
```




```
## [1] -2.042272  2.042272
```

# $t$ 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.

# A note about the conditions for inference about a mean

- 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

# A note about the conditions for inference about a mean

- 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 *Normal* distribution with unknown mean parameter $\mu$ and unknown standard deviation parameter $\sigma$ less so
- *In practice*, inference procedures *can accommodate some deviations from the Normality condition* when the sample is large enough. (think CLT)

# Robustness of the $t$ procedures

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

- $t$ procedures are not robust against *extreme* skewness, in small samples, since the procedures are based on using $\overline{y}$ and $s$ (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, $s$ is a good estimate of $\sigma$ (recall that $s$ is unbiased and, like most estimates, precision improves with increasing sample size)
- CLT: $\bar{y}$ will be Normal when $n$ is large, even if the population data are not

# When and why we use the $t$-distribution

- When $\sigma$ is unknown use $t$ distribution. but why?

# When and why we use the $t$-distribution

- When $\sigma$ 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 $\sigma$ is known

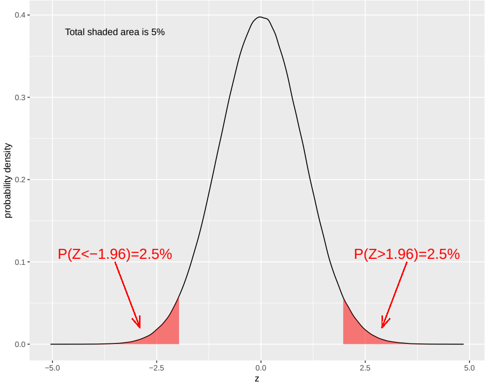
$$\underbrace{z_{0.975}}_{\text{critical value}} = 1.96 = \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}} \rightarrow \frac{1.96\sigma}{\sqrt{n}} = \bar{y} - \mu_0$$

which means that to reject $H_0$ the difference between your sample mean and $\mu_0$ needs to be greater than $\frac{1.96}{\sqrt{n}}$ standard deviations

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

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

which means that to reject $H_0$ the difference between your sample mean and $\mu_0$ needs to be greater than $\frac{1.96}{\sqrt{n}}$ standard deviations



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

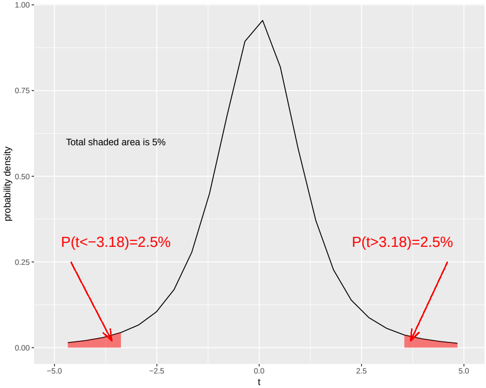
$$\underbrace{t_{0.975,df=3}}_{\text{critical value}} = 3.18 = \frac{\bar{y} - \mu_0}{s/\sqrt{n}} \to 3.18\frac{s}{\sqrt{n}} = \bar{y} - \mu_0$$

which means that to reject $H_0$ the difference between your sample mean and $\mu_0$ needs to be greater than $\frac{3.18}{\sqrt{n}}$ standard deviations

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

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

which means that to reject $H_0$ the difference between your sample mean and $\mu_0$ 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 $\sigma$

# 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 $\sigma$
- 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 $\sigma$
- 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 $\sigma$

# 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 $\sigma$
- 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 $\sigma$
- 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 $\mathcal{N}(0, 1)$