Lecture 7: Confidence Intervals with the t-Distribution
STAT 630, Fall 2021

The t-distribution

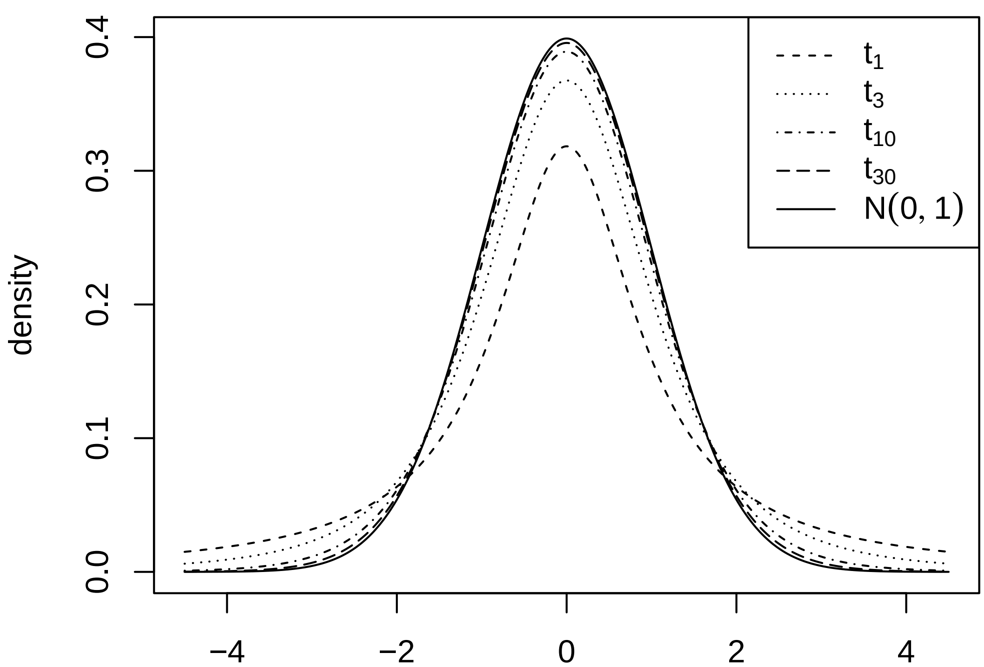
Let $X_1, X_2, \cdots, X_n$ be a random sample of size n from a normal distribution; i.e., $X_i \sim N(\mu, \sigma)$. Consider the random variable

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}},$$

were S is the sample standard deviation (also random) defined by

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

Then the random variable T is said to follow a t-distribution (or Student's t-distribution) with $n-1$ degrees of freedom. We can also use the notation $\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$.

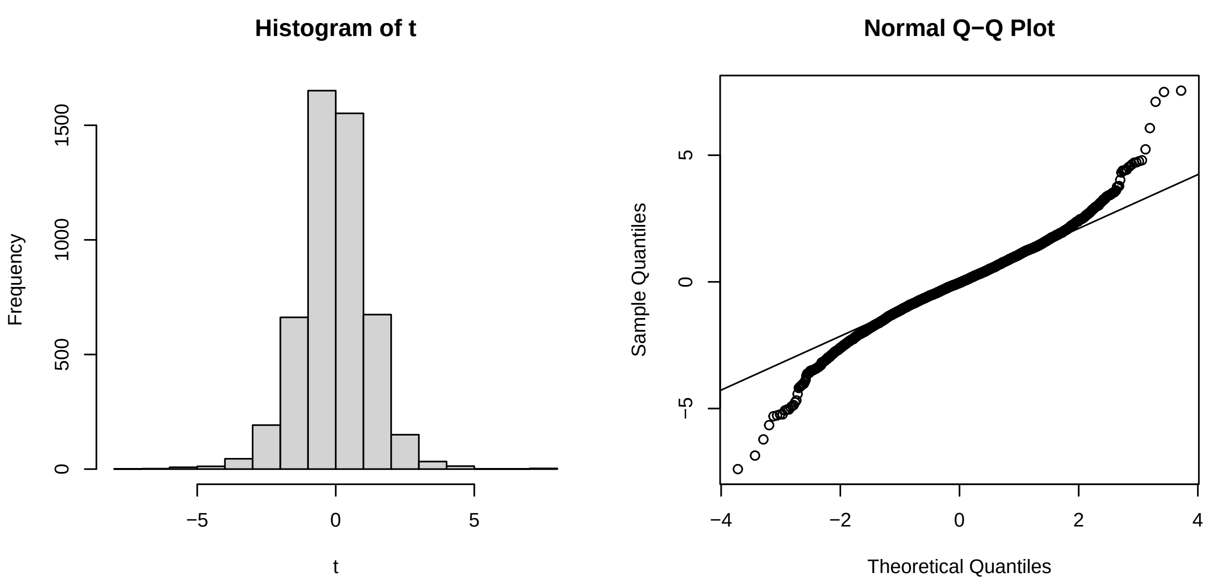


Remarks:

- Similar to the standard normal distribution, the t-distribution is bell-curve shaped, symmetric, and centered about zero.
- Remarkably, the t-score, $t = (\bar{x} - \mu)/(s/\sqrt{n})$ depends on the sample standard deviation s, not the population standard deviation σ; this is one of its most useful properties.
- The t-distribution has wider tails than the standard normal distribution.
- The t-distribution approaches the standard normal distribution as n gets large. That is, $t_{n-1} \to N(0, 1)$ as $n \to \infty$. In fact, when the degrees of freedom is about 30 or more, the t-distribution is nearly indistinguishable from the standard normal distribution.

Ex1: Simulating random numbers from t_6

```r
set.seed(999)
t <- rt(5000, df=6)
par(mfrow=c(1,2), cex=0.75)
hist(t)
qqnorm(t)
qqline(t)
```



Constructing a confidence interval for μ when σ is unknown and the population distribution is normal

Let $X_1, X_2, \cdots, X_n$ be a random sample of size n from a normal population distribution; i.e., $X_i \sim N(\mu, \sigma)$. Since the random variable $\frac{\bar{X}-\mu}{S/\sqrt{n}}$ follows a t-distribution with $n-1$ degrees of freedom we can write the following probability statement:

$$P\left(-t_{\alpha/2;n-1} < \frac{\bar{X}-\mu}{S/\sqrt{n}} < t_{\alpha/2;n-1}\right) = 1-\alpha$$

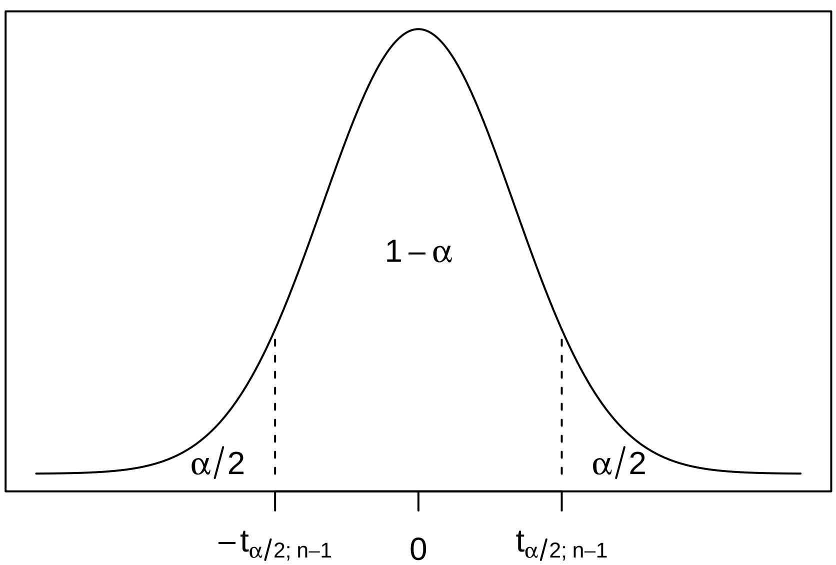
Rearranging terms in the above probability statement gives:

$$P(\bar{X} - t_{\alpha/2;n-1}S/\sqrt{n} < \mu < \bar{X} + t_{\alpha/2;n-1}S/\sqrt{n}) = 1-\alpha$$

Therefore, a $100(1-\alpha)\%$ confidence interval for μ is given by

$$\bar{x} \pm t_{\alpha/2;n-1}\frac{s}{\sqrt{n}}$$

The critical value $t_{\alpha/2;n-1}$ is defined as follows:



In R, $t_{\alpha/2;n-1}$ = `qt(1-alpha/2, df=n-1)`

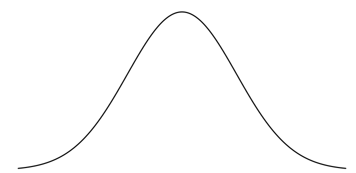
Conditions: The t-confidence interval for μ is valid if the following conditions are satisfied:

- Sample observations are independent. Generally, this is satisfied when the data come from a random sample.
- The sample size is large ($n \geq 30$), and there are no extreme outliers. This implies that the sampling distribution for $\bar{X}$ is approximately normal according to the central limit theorem.
- Otherwise, if the sample size is small ($n < 30$), the data should follow an approximate normal distribution. Graphical methods can be used to check this (box plot, histogram, normal QQ plot).

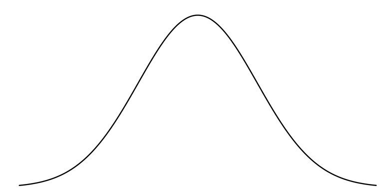
Remark: When the sample size is large ($n \geq 30$), we can use either a t or z critical value to make a confidence interval for μ, since the distributions are nearly identical.

Ex2: Let T be a random variable following a t-distribution with 9 degrees of freedom.

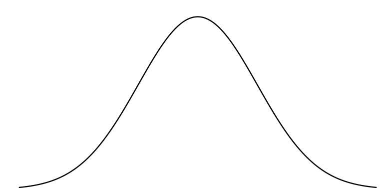
(a) Calculate $P(T < 1.5)$



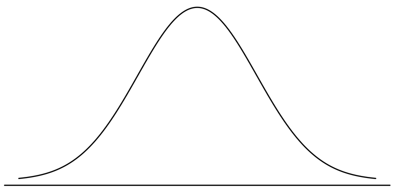
(b) Calculate $P(-0.75 < T < 1.5)$



(c) Find the value c such that $P(T > c) = 0.2$



Ex3: Compare the critical values $t_{\alpha/2;n-1}$ and $z_{\alpha/2}$ when the sample size $n = 30$ and the confidence level is 0.95.

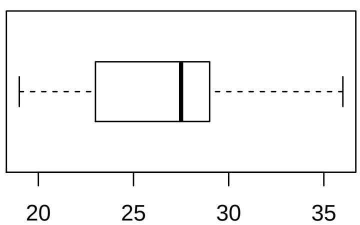


$t_{0.025;29} =$ `qt(0.975, df=29)` $= 2.045$
$z_{0.025} =$ `qnorm(0.975)` $= 1.96$

The critical values are close when $n = 30$.
The t-critical value is slightly larger.

Ex4: Below are some summary statistics and a box plot for the ages of a random sample of $n = 26$ female athletes who participated in the 2012 Olympic Games in London. Using this information, calculate and interpret a 95% confidence interval for the population mean age. Comment on whether the conditions for the interval appear satisfied.[1]

n	$\bar{x}$	s	min	max
26	26.9	4.5	19	36



The conditions for the interval are satisfied: First, the independence condition is met since the data come from a random sample. Second, the data follow an approximate normal distribution in the box plot, and there are no outliers (we need to check normality since $n < 30$).

At the 0.95 confidence level, the critical value is `qt(0.975, df=25) = 2.06`. Therefore, a 95% confidence interval for μ is given by:

$$\bar{x} \pm t_{\alpha/2;n-1} \frac{s}{\sqrt{n}} \implies 26.9 \pm 2.06 \cdot \frac{4.5}{\sqrt{26}} \implies (25.08, 28.72)$$

Interpretation: We are 95% confident that the population mean age, μ, is between 25.08 and 28.72.

[1] Data obtained from the data set `Olympics2012` in the R package `resampledata`.

Simulation Study: Compare the coverage of confidence intervals constructed using the t and z distributions when repeatedly taking samples of size $n = 5$ from a $N(\mu = 50, \sigma = 10)$ population distribution. Use a 95% confidence level.

```r
set.seed(999)
mu <- 50
count_t <- count_z <- 0
for(i in 1:1000) {
  samp <- rnorm(5, mean=50, sd=10)

  # t-interval
  tcrit <- qt(0.975, df=4)
  ci_lower <- mean(samp) - tcrit * sd(samp) / sqrt(5)
  ci_upper <- mean(samp) + tcrit * sd(samp) / sqrt(5)
  if(mu >= ci_lower & mu <= ci_upper) {
    count_t <- count_t + 1
  }

  # z-interval
  zcrit <- qnorm(0.975)
  ci_lower <- mean(samp) - zcrit * sd(samp) / sqrt(5)
  ci_upper <- mean(samp) + zcrit * sd(samp) / sqrt(5)
  if(mu >= ci_lower & mu <= ci_upper) {
    count_z <- count_z + 1
  }
}
count_t / 1000

## [1] 0.948

count_z / 1000

## [1] 0.878
```

Conclusion: The proportion of $t-$confidence intervals that contain $\mu = 50$ is 0.948, which is close to the 0.95 confidence level. However, the proportion of $z-$confidence intervals that contain $\mu = 50$ is 0.878, which is less than the 0.95 confidence level (intervals are too narrow).