

# Biostatistics

Week #10 5/05/2020



# **Ch 9 – Confidence Intervals – Part 2**



# Outline

- 9.0 Review on Sampling Distribution
- 9.1 Two-sided Confidence Intervals
- 9.2 One-sided Confidence Intervals
- 9.3 Student's  $t$  Distribution
- 9.4 Applications

## 9.3 Student's $t$ distribution



# History of $t$ -distribution

- In the English-language literature it takes its name from William Sealy Gosset's 1908 paper in Biometrika under the pseudonym "Student".
- Gosset worked at the Guinness Brewery in Dublin, Ireland, and was interested in the problems of small samples, for example the chemical properties of barley where sample sizes might be as low as 3.



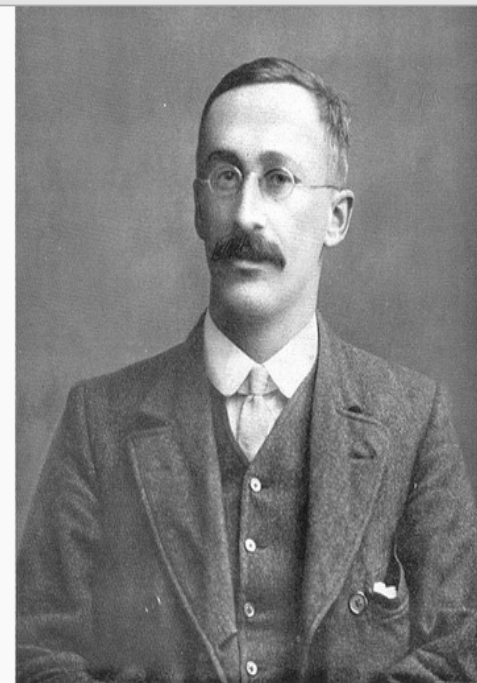
## Life and career[[edit](#)]

Born in [Canterbury, England](#) to Agnes Sealy Vidal and Colonel Frederic Gosset, Gosset attended [Winchester College](#) before reading [chemistry](#) and [mathematics](#) at [New College, Oxford](#). Upon graduating in 1899, he joined the brewery of [Arthur Guinness & Son](#) in [Dublin, Ireland](#).

As an employee of [Guinness](#), a progressive agro-chemical business, Gosset applied his statistical knowledge — both in the brewery and on the farm — to the selection of the best yielding varieties of [barley](#). Gosset acquired that knowledge by study,

ometrical  
ship. Pearson  
8 papers, but  
brewer's  
hand,  
oping small-

ntaining trade  
nfigidential  
papers



*Student in 1908*

**Born** June 13, 1876  
[Canterbury, Kent, England](#)  
**Died** October 16, 1937 (aged 61)  
[Beaconsfield, Buckinghamshire, England](#)  
**Known for** [Student's t-distribution](#)



# Cont'd

- One version of the origin of the pseudonym is that Gosset's employer preferred staff to use pen names when publishing scientific papers instead of their real name, therefore he used the name "Student" to hide his identity.
- Another version is that Guinness did not want their competitors to know that they were using the  $t$ -test to test the quality of raw material.

# Cont'd

- Gosset's paper refers to the distribution as the "frequency distribution of standard deviations of samples drawn from a normal population".
- It became well-known through the work of Ronald A. Fisher, who called the distribution "*Student's distribution*" and referred to the value as  $t$ .



- In previous lecture we mentioned using the following formula to convert the sample mean  $\bar{x}$  (which is a random variable) to  $Z$ , and from the  $Z$ -distribution (which is a standard normal distribution) to locate boundaries for CI.
- To do this, one ***needs to know*** the standard deviation  $\sigma$  of the general population.

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

- In practice, we may not know the standard deviation  $\sigma$  of the general population.
- Instead, we have only the sample standard deviation  $s$ .
- As a result, we define a new random variable  $t$  to have the following transformation.

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

- Similar to  $Z$ ,  $t$  would have its own probability density function too. Unlike  $Z$ ,  $t$  won't have a normal distribution.

# Student's $t$ distribution

- In probability and statistics, **Student's  $t$ -distribution** (or simply the  **$t$ -distribution**) is a *family of* continuous probability distributions that arises when ***estimating the mean of a normally distributed population*** in situations where:
  - the *sample size  $n$  is small*
  - *population standard deviation  $\sigma$  is unknown.*

[http://en.wikipedia.org/wiki/Student's\\_t-distribution](http://en.wikipedia.org/wiki/Student's_t-distribution)

# New random variable $t$

- This is known as Student's  $t$  distribution with  $n-1$  degree of freedom (or  $df = n-1$ ), and is often denoted by the notation  $t_{n-1}$

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

Depending on the value of  $n$ , there will be different  $t$ -distributions.

# Degree of Freedom (df)

- The number of **degrees of freedom** is the number of values in the final calculation of a statistic that are *free to vary*.
- In general, the degrees of freedom of an estimate of a parameter is equal to **the number of independent scores that go into the estimate minus the number of parameters used as intermediate steps in the estimation of the parameter itself.**

# DF – examples

- Example 1: When estimating the mean value  $\mu$  from  $n$  numbers, each number is free to change regardless the values for the other numbers. Thus  **$DF = n$** .
- Example 2: When estimating the standard deviation from  $n$  numbers, we need the value  $\bar{x}$  which is the mean value. Once  $\bar{x}$  is fixed, there are only  $n-1$  numbers free to change. The  $n^{th}$  number is dependent on those  $n-1$  numbers together with  $\bar{x}$ . As a result,  **$DF = n-1$** .



# Cont'd

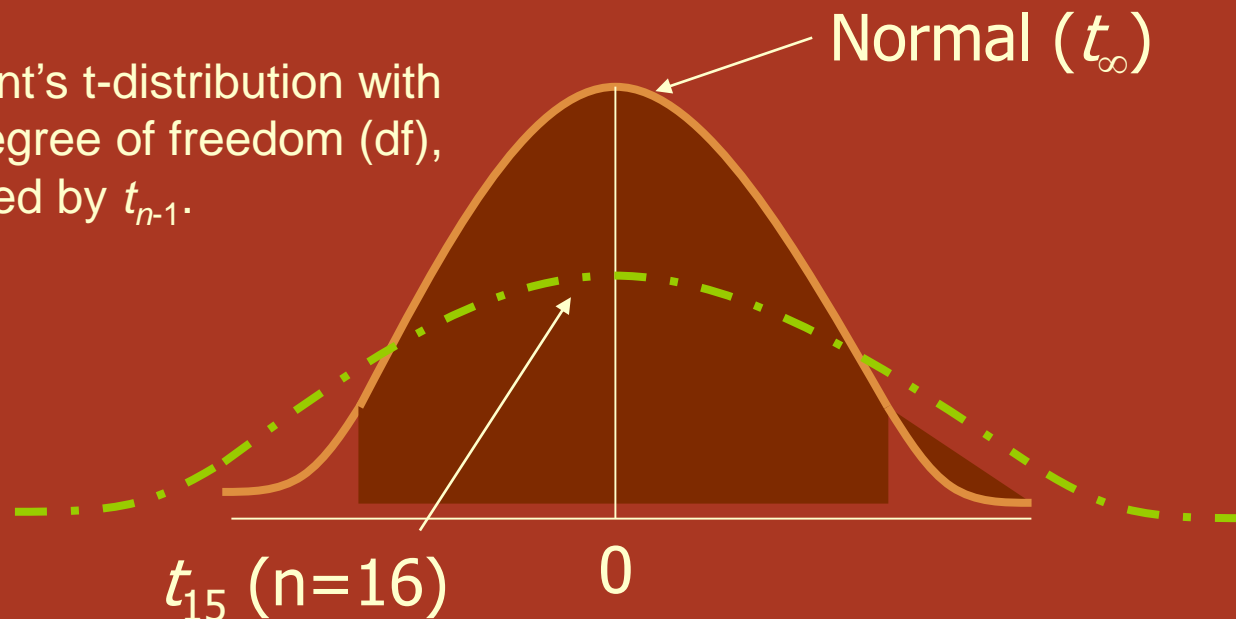
- Here, the degrees of freedom measure the amount of information available in the data that can be used to estimate  $\sigma^2$ .
- The degrees of freedom are  $n-1$  rather than  $n$  because we lose 1 degree of freedom by estimating the sample mean  $\bar{x}$ .
- Recall that  $df = n-1$  is the quantity by which we divided the sum of the squared deviations around the mean to obtain the variance.

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

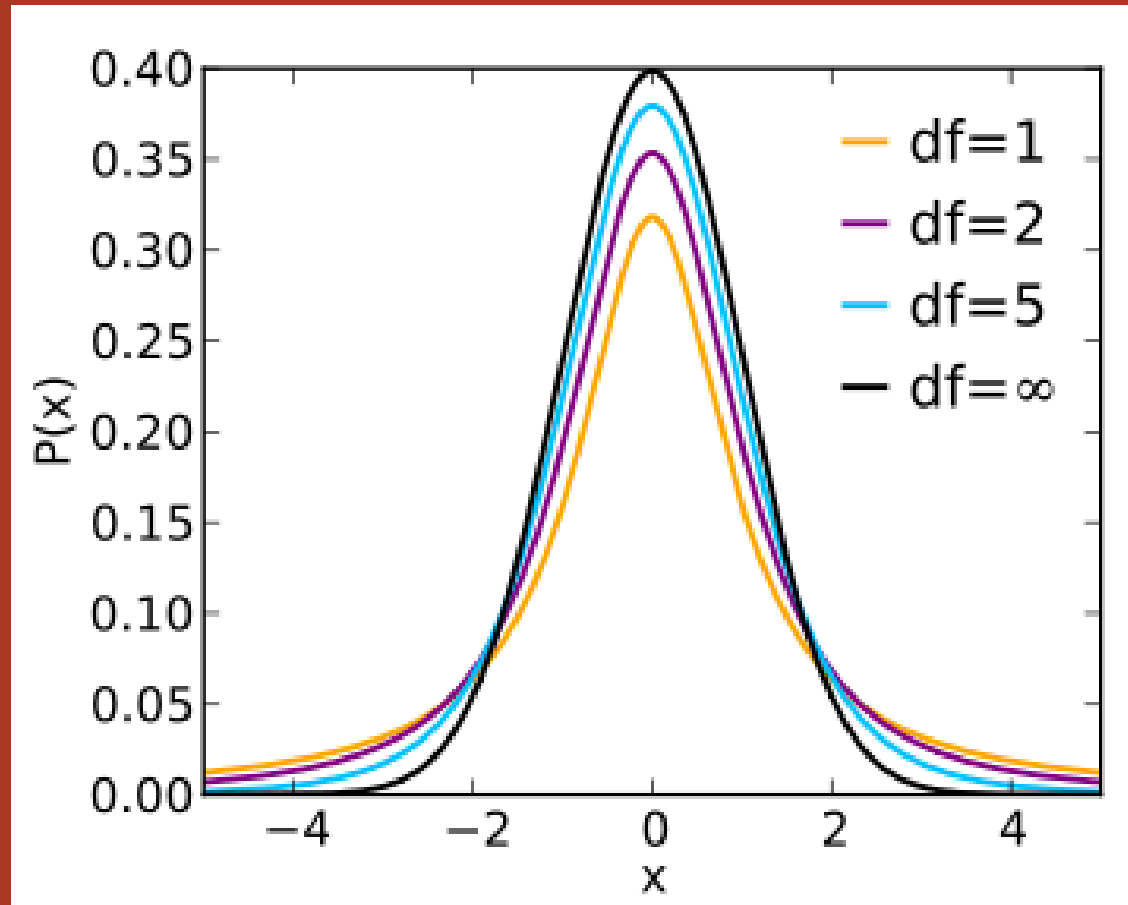
# ***t*-distribution**

- The *t*-distribution looks like a standard normal curve that has been “stepped on” – it’s a little flatter and fatter
- A *t*-distribution is solely determined by its degrees of freedom (df) – the lower the degrees of freedom, the flatter it is.

Student's *t*-distribution with  $n-1$  degree of freedom (df), denoted by  $t_{n-1}$ .



# Probability density functions (PDF) of various degrees of freedom



$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi} \Gamma(\frac{\nu}{2})} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

$\nu$  =df above.

- Recall that we used the formula to estimate the 95% CI

$$\left(\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}}\right)$$

- This is equivalent to say that the two boundaries of CI are:

$$\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

Standard error of the mean (or SEM)

Replace 1.96 with 2.58 for 99% CI.

- When a *t*-correction is needed, the previous equation becomes

$$\left( \bar{x} - t \frac{s}{\sqrt{n}}, \bar{x} + t \frac{s}{\sqrt{n}} \right)$$

- This is equivalent to say that the two boundaries of CI are:

$$\bar{x} \pm t \frac{s}{\sqrt{n}}$$

Standard error of the mean (or SEM)

*The value of  $t$  will depend on  $n$  as well as on what level of CI you need to achieve.*

# Adjustment for Small Sample Sizes

## Value of $T$ for 95% Confidence Interval for Mean

(here  $T$  is the same as  $t$ )

df (n-1)	$T$	df	$T$
1	12.706	12	2.179
2	4.303	13	2.160
3	3.182	14	2.145
4	2.776	15	2.131
5	2.571	20	2.086
6	2.447	25	2.060
7	2.365	30	2.042
8	2.306	40	2.021
9	2.262	60	2.000
10	2.228	120	1.980
11	2.201	$\infty$	1.960

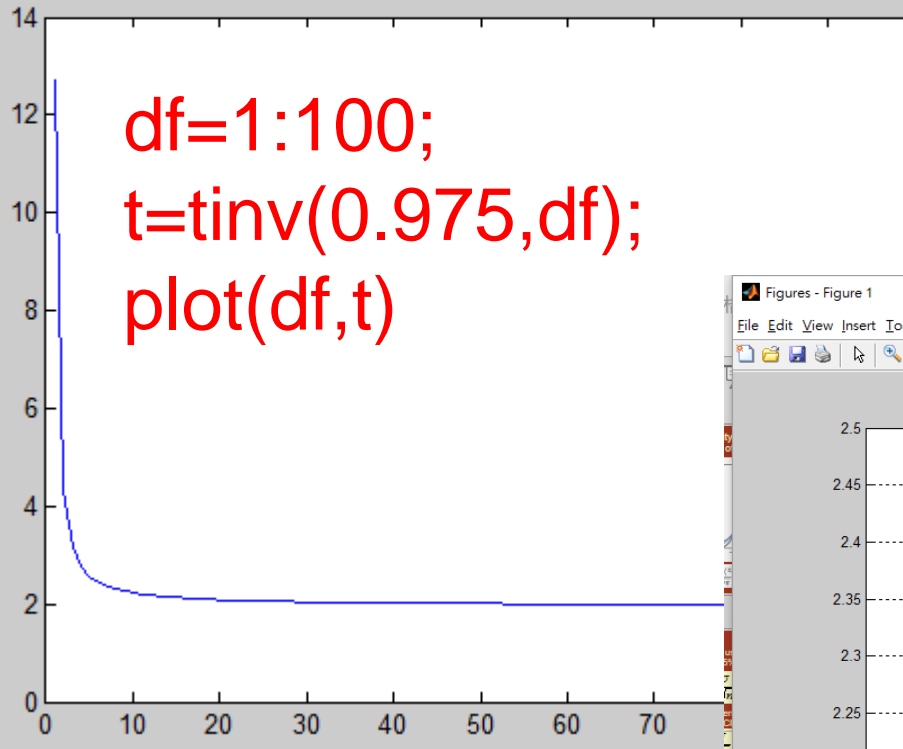
The values for  $T$  would be different when a different CI interval was used. For example, for a 90% CI or 99% CI.



df (n-1)	T	df	T
1	12.706	12	2.179
2	4.303	13	2.160
3	3.182	14	2.145
4	2.776	15	2.131
5	2.571	20	2.086
6	2.447	25	2.060
7	2.365	30	2.042
8	2.306	40	2.021
9	2.262	60	2.000
10	2.228	120	1.980
11	2.201	$\infty$	1.960

As we recall, the value of 1.960 can be obtained by the function **norminv(0.975)**. Similarly, a MATLAB function **tinva(0.975,df)** allows you obtain these T values providing df.

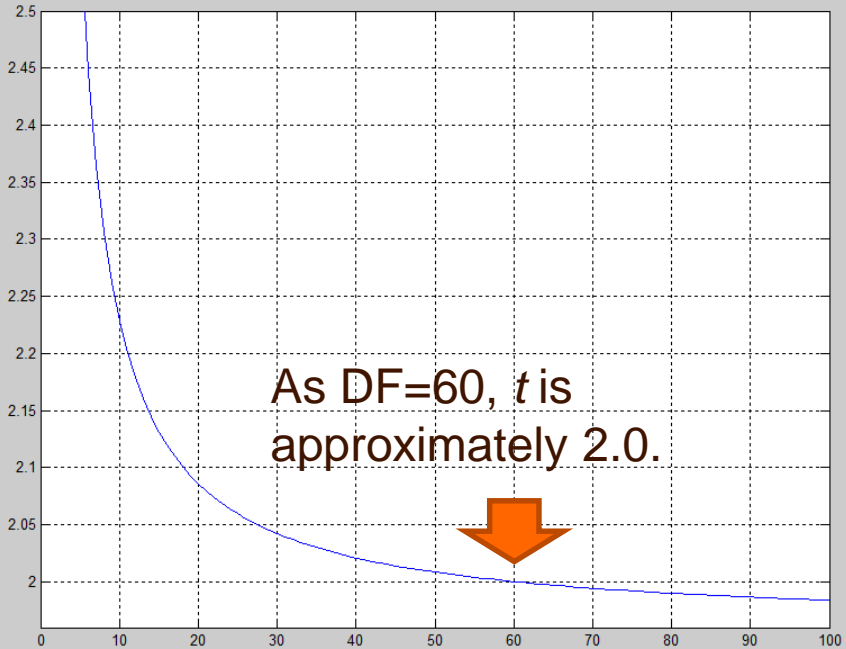
```
df=1:100;  
t=tinv(0.975,df);  
plot(df,t)
```



Q1: What will be the value of  $t_{inv}(0.975, \infty)$ ?

Q2: What will be the value of  $t_{inv}(0.5, 30)$ ?

As DF=60,  $t$  is approximately 2.0.



# Notes on the $t$ -Correction

- The value of  $t$  that you need depends on the level of confidence (e.g., 95% or 90%, etc.) you want as well as the sample size.
- With really small sample sizes ( $n < 15$ , or so), you also need to pay attention to the underlying distribution of the data in your sample
  - Needs to be “well behaved” (i.e., somewhat normal) for us to use  $\bar{x} \pm t^* \text{SEM}$  for creating confidence intervals

# Example 1

- A random sample of 10 children selected from a population of infants receiving antacids (to treat digestive disorders) that contains aluminum.
- The distribution of population plasma aluminum is approximately normal with unknown  $\mu$  and  $\sigma$ . The sample mean and standard deviation is computed as 37.2 and 7.13.
- **We'd like to estimate the 95% CI of the population mean  $\mu$ .**

# Cont'd

- Here the degree of freedom  $DF = 10 - 1 = 9$ .
- From the table below, we know the  $t$ -correction to be used would be 2.262, rather than the usually used 1.96.

df	T	df	T
1	12.706	12	2.179
2	4.303	13	2.160
3	3.182	14	2.145
4	2.776	15	2.131
5	2.571	20	2.086
6	2.447	25	2.060
7	2.365	30	2.042
8	2.306	40	2.021
9	2.262	60	2.000
10	2.228	120	1.980
11	2.201	$\infty$	1.960

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$
$$\bar{x} \pm t \times \frac{s}{\sqrt{n}}$$

# Cont'd

df	T	df	T
1	12.706	12	2.179
2	4.303	13	2.160
3	3.182	14	2.145
4	2.776	15	2.131
5	2.571	20	2.086
6	2.447	25	2.060
7	2.365	30	2.042
8	2.306	40	2.021
9	2.262	60	2.000
10	2.228	120	1.980
11	2.201	∞	1.960

$$\bar{x} \pm t \times \frac{s}{\sqrt{n}}$$
$$\bar{x} \pm 2.262 \times \frac{7.13}{\sqrt{10}}$$

(32.1, 42.3)

Based on this 10-children sample statistics, we are 95% confident that the interval (32.1, 42.3) would contain the population mean value. For example, plasma aluminum of 42.0 **would not** be considered overdosed.



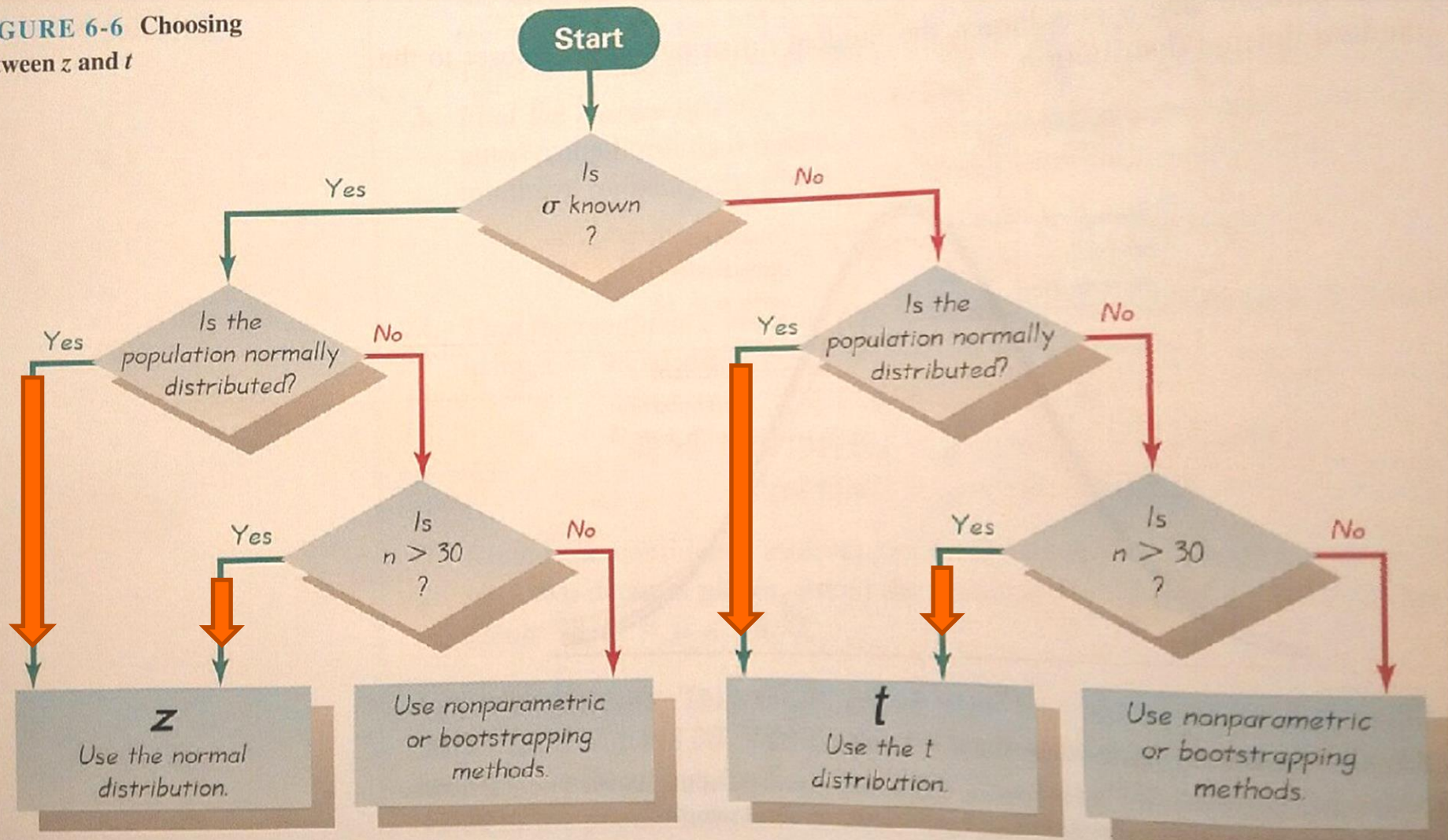
# Comments

$$\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

- Note that if the sample size increases to  $n=31$  (DF=30), the  $t$  value would become 2.042. The CI would be (32.6, 41.8).
- If the population  $\sigma$  is known to be 7.13, we would use earlier described formula to get the CI, which is (32.8, 41.6). [thus **42.0 would suggest overdosed.**]
- In both case, we see a narrower CI (because of less or nothing being corrected).
- Here we are using 2-sided CI. That is, either too much aluminum or too less are bad.

# Choosing between $z$ and $t$ ?

FIGURE 6-6 Choosing between  $z$  and  $t$



# Choosing between z and t

- In an attempt to construct a CI for the population mean  $\mu$  for the following cases:
  - $n=150$ ,  $\bar{x}=100$ ,  $s=15$ , and the population has a skewed distribution
  - $n=8$ ,  $\bar{x}=100$ ,  $s=15$ , normal distribution
  - $n=8$ ,  $\bar{x}=100$ ,  $s=15$ , a very skewed distribution
  - $n=150$ ,  $\bar{x}=100$ ,  $\sigma=15$ , a skewed distribution (almost never occurs)
  - $n=8$ ,  $\bar{x}=100$ ,  $\sigma=15$ , a very skewed distribution (almost never occurs)

# Choosing between z and t

- In an attempt to construct a CI for the population mean  $\mu$  for the following cases:
  - $n=150$ ,  $\bar{x}=100$ ,  $s=15$ , and the population has a skewed distribution **[t]**
  - $n=8$ ,  $\bar{x}=100$ ,  $s=15$ , normal distribution **[t]**
  - $n=8$ ,  $\bar{x}=100$ ,  $s=15$ , a skewed distribution **[?]**
  - $n=150$ ,  $\bar{x}=100$ ,  $\sigma=15$ , a skewed distribution (almost never occurs) **[z]**
  - $n=8$ ,  $\bar{x}=100$ ,  $\sigma=15$ , a skewed distribution (almost never occurs) **[?]**

## 9.4 Applications of confidence intervals



## Example 2

- A drug called methylphenidate is used to treat 20 children for attention-deficit disorder (group M).
- Another 20 with the disease were given a placebo (group P).
- A test called Parent Rating Scale (PRS) is used to evaluate how their behavior and attention status was improved. A lower score means better improvement.



- The distribution of PRS scores for these children are approximately normal. The population mean and variation of scores are **not known**.
- The mean and standard variation for group M are 10.8 and 2.9. For group P they are 14 and 4.8.
- We want to estimate the 95% CI of the mean score for the group taking drug, as well as the group taking placebo.

# Solution:

Will show you how to compute this later.

- N=20, thus DF=19. The t value to use is 2.093\*.
- Standard errors of the means (SEMs) for group M and P are:

$$\frac{2.9}{\sqrt{20}}, \frac{4.8}{\sqrt{20}} \text{ or } 0.6485, 1.0733$$

Finally, the 95% CI are

$$10.8 \pm 2.093 \times 0.6485, 14.0 \pm 2.093 \times 1.0733$$

Or  $(9.44, 12.16), (11.75, 16.25)$

# A short summary

- It is clear that children taking drug are apparently having lower scores, implying improved attention.
- However, there is some overlap between the two intervals. This blurs slightly the effectiveness of this drug.

$(9.44, 12.16)$ ,     $(11.75, 16.25)$

*Taking drug*

*Taking placebo*

*Either some taking drugs may not be improved, or some not taking drugs could show improvement.*

## Example 3

- A random sample of 81 workers at a company showed that they work an average of 100 hours per month with a standard deviation of 27 hours.
- That is,  $\bar{x}=100$ ,  **$s=27$**  and  $n=81$ .
- Compute a 95% confidence interval for the mean hours per month all workers at the company work.

# Answer

- $SEM = 27/\text{sqrt}(81) = 3.$
- So 1.99×SEM=5.97. ( $t \cong 1.99$  from the table, with df= 80)
- The interval would be from  $100-5.97=94.03$  to  $100+5.97=105.97.$

df	T	df	T
1	12.706	12	2.179
2	4.303	13	2.160
3	3.182	14	2.145
4	2.776	15	2.131
5	2.571	20	2.086
6	2.447	25	2.060
7	2.365	30	2.042
8	2.236	40	2.021
9	2.262	60	2.000
10	2.228	120	1.980
11	2.201	$\infty$	1.960

# Example 4

- Consider the following random sample of 4 observations ( $n=4$ ) **25, 47, 32, 56**.
- Suppose the population is normally distributed. Please construct a 95% confidence interval for the mean.

df	T	df	T
1	12.706	12	2.179
2	4.303	13	2.160
3	3.182	14	2.145
4	2.776	15	2.131
5	2.571	20	2.086
6	2.447	25	2.060
7	2.365	30	2.042
8	2.236	40	2.021
9	2.262	60	2.000
10	2.228	120	1.980
11	2.201	$\infty$	1.960

# Answer

- First compute the mean of this sample, which is 40, followed by computing the standard deviation  $s = 14.07$ .
- Compute  $s/\sqrt{n} = 14.07/2 = 7.035$ .
- Since the sample size is small, t-correction is applicable, with  $t = \underline{3.182}$  (df=3).
- So the interval would be from  $40 - \underline{3.182} * 7.035 = 17.615$  to  $40 + \underline{3.182} * 7.035 = 62.385$ .

# **MATLAB function TPDF**

TPDF Probability density function (pdf)  
for Student's T distribution

**Y = TPDF(X,V)** returns the pdf of  
Student's T distribution with

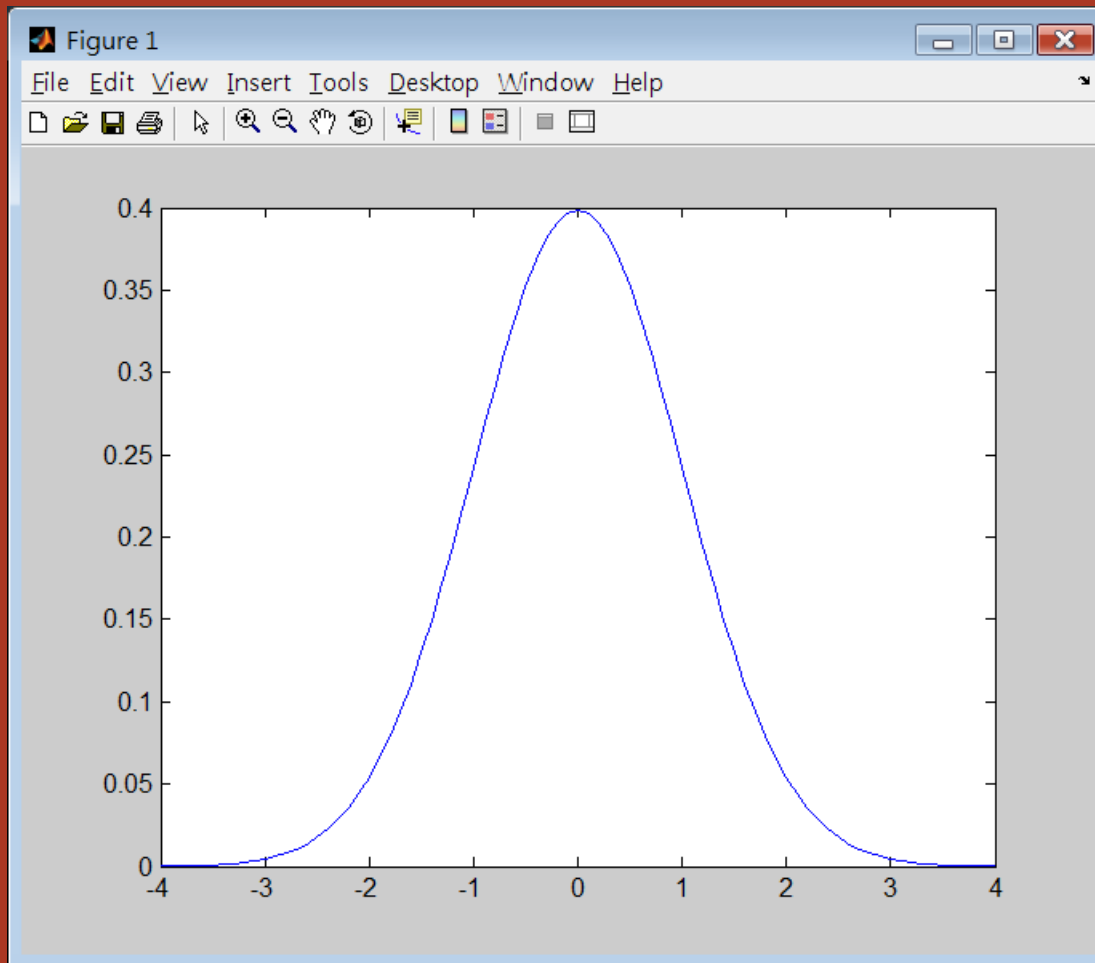
**V degrees of freedom**, at the  
values in X.



```
>> z=[-4:0.1:4];  
>> X1=1/(sqrt(2*pi))*exp(-0.5*z.^2);  
>> plot(z,X1)  
>>
```

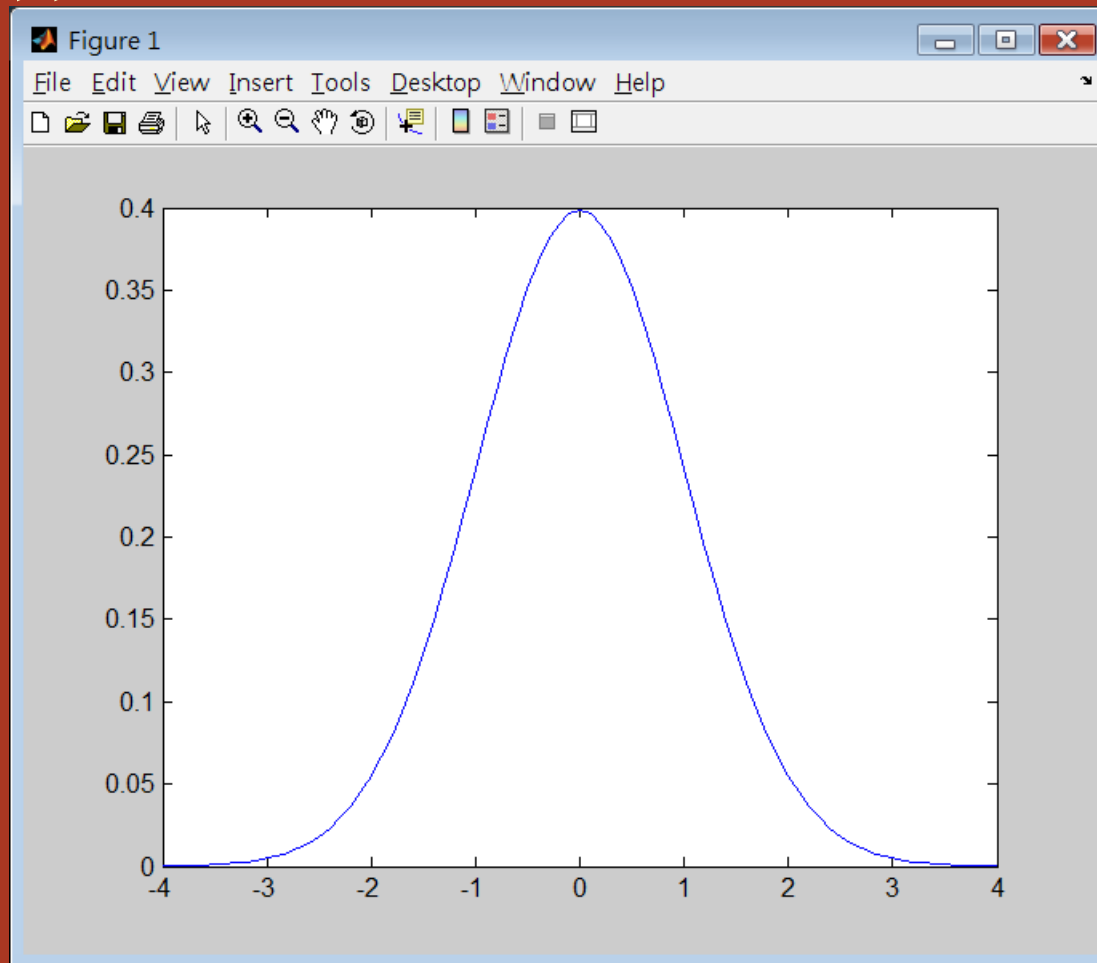
Or you can use  
 $X1 = \text{normpdf}(z)$

*A standard normal distribution*

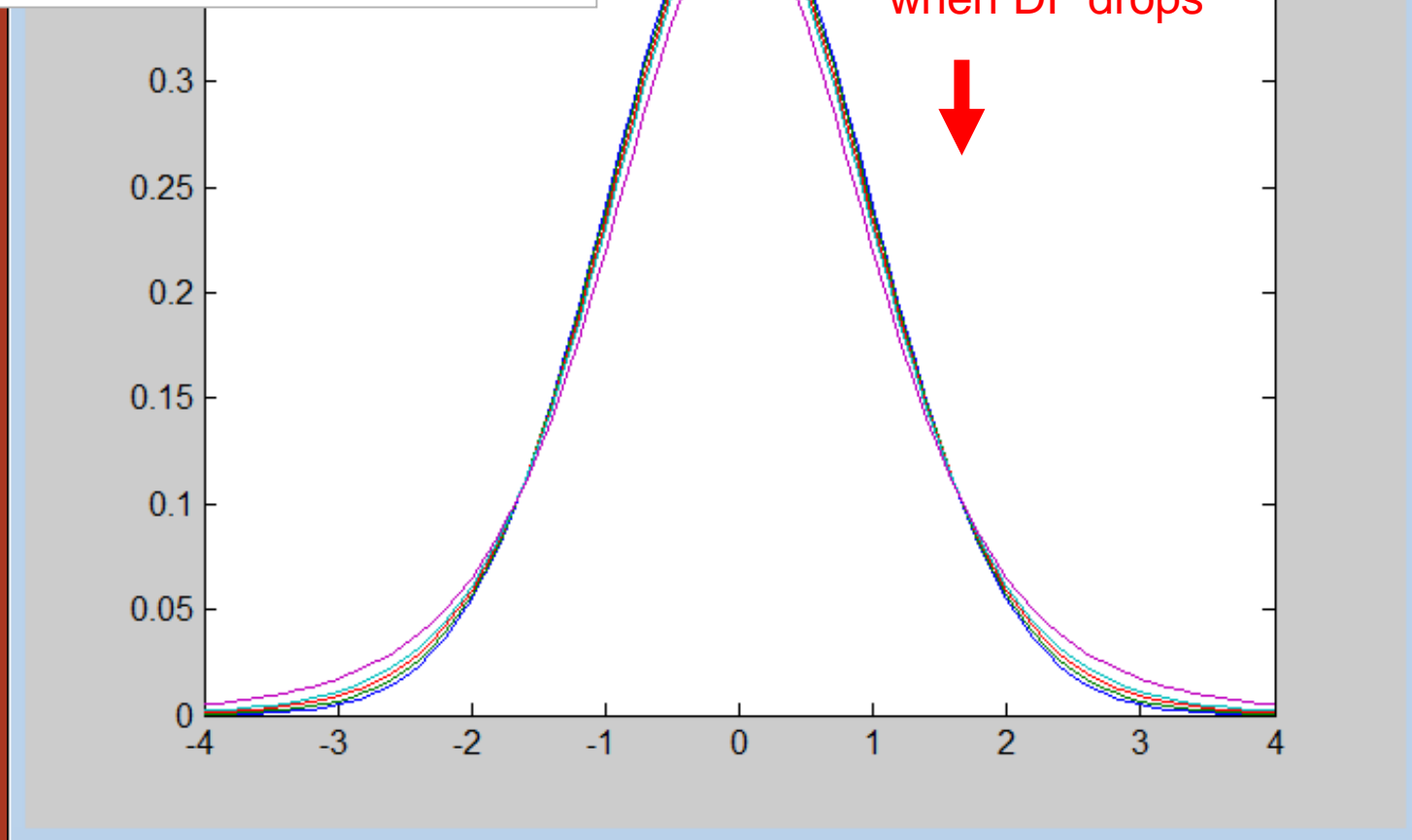


```
>> t=[-4:0.1:4];  
>> X2=tpdf(t,120);  
>> plot(t,X2)  
>>
```

*With  $DF=120$ , this is  
'nearly identical' to a  
standard normal  
distribution.*



```
>> X2=tpdf(t,120);  
>> X3=tpdf(t,30);  
>> X4=tpdf(t,15);  
>> X5=tpdf(t,10);  
>> X6=tpdf(t,5);  
>> plot(t,X2,t,X3,t,X4,t,X5,t,X6)
```



# MATLAB Function TCDF

TCDF Student's T cumulative distribution function (cdf).

$P = \text{TCDF}(X, V)$  computes the cdf for Student's T distribution with  $V$  degrees of freedom, at the values in  $X$ . The size of  $P$  is the common size of  $X$  and  $V$ . A scalar input functions as a constant matrix of the same size as the other input.

# MATLAB function TINV

TINV Inverse of Student's T cumulative distribution function (cdf).

$X = \text{TINV}(P, V)$  returns the inverse of Student's T cdf with V degrees of freedom, at the values in P.

```
>> tinv(0.95,inf)  
ans =  
1.6449
```

90% interval  
(5% each side)

```
>> tinv(0.975,inf)  
ans =  
1.9600
```

95% interval  
(2.5% each side)

```
>> tinv(0.995,inf)  
ans =  
2.5758
```

99% interval  
(0.5% each side)

# Various t values (different DFs) for 95% CI

df	T	df	T
1	12.706	12	2.179
2	4.303	13	2.160
3	3.182	14	2.145
4	2.776	15	2.131
5	2.571	20	2.086
6	2.447	25	2.060
7	2.365	30	2.042
8	2.306	40	2.021
9	2.262	60	2.000
10	2.228	120	1.980
11	2.201	$\infty$	1.960

>> tinv(0.975,8)  
ans =  
2.3060

```
>> tinv(0.975,inf)    ans = 1.9600
>> tinv(0.975,120)   ans = 1.9799
>> tinv(0.975,60)    ans = 2.0003
>> tinv(0.975,30)    ans = 2.0423
>> tinv(0.975,15)    ans = 2.1314
>> tinv(0.975,10)    ans = 2.2281
>> tinv(0.975,5)     ans = 2.5706
```

# Various t values (different DFs) for 90% CI

The following MATLAB commands produce the t values for various degrees of freedom.

```
>> tinv(0.95,inf)  
>> tinv(0.95,120)  
>> tinv(0.95,60)  
>> tinv(0.95,30)  
>> tinv(0.95,15)  
>> tinv(0.95,10)  
>> tinv(0.95,5)
```

# Example 4 - revisited

```
>> dataset=[25 47 32 56];  
>> datamean=mean(dataset)  
datamean = 40  
>> L=length(dataset)  
L = 4
```

$$\bar{x} \pm t \frac{s}{\sqrt{n}}$$

```
>> q = tinv(0.975,L-1)*std(dataset)/sqrt(L)  
q = 22.3905  
>> datamean-q  
ans = 17.6095  
>> datamean+q  
ans = 62.3905  
>>
```

*Standard Error of  
the Means (SEM)*



# MATLAB function **paramci**

**PARAMCI** Parameter confidence intervals.

CI = PARAMCI(PD) returns a 2-by-N array CI containing 95% confidence intervals for the parameters of the probability distribution PD.

CI = PARAMCI(PD, ALPHA) returns 100\*(1-ALPHA)% confidence intervals. The default is 0.05 for 95% confidence intervals.

# MATLAB function **fitdist**

**FITDIST** Fit probability distribution to data.

PD = FITDIST(X,DISTNAME) fits the probability distribution DISTNAME to the data in the **column vector** X, and returns an object PD representing the fitted distribution. PD is an object in a class derived from the ProbDist class.

*DISTNAME is the name for this distribution, which could be 'binomial', 'normal', 'poisson', etc.*

```
>> dataset=[25 47 32 56];
```

```
>> PD=fitdist(dataset, 'normal');
```

```
??? Error using ==> fitdist at 102
```

```
X must be a numeric column vector.
```

```
>> X=dataset'
```

```
>> PD=fitdist(X, 'normal');
```

```
>> CI=paramci(PD, 0.05)
```

```
CI =
```

<b>17.6095</b>	7.9712
<b>62.3905</b>	52.4653

```
>>
```

# Example 5

- Construct a 95% CI to estimate the mean value for the following numbers:

1903 1935 1910 2496 2108 1961 2060  
1444 1612 1316 1511

- We need to firstly assess the distribution to decide using t, or z, or none.

```
>> X=[1903 1935 1910 2496 2108 1961 2060 1444 1612  
1316 1511];
```

```
>> hist(X)
```

```
>> mean(X)
```

```
ans =
```

```
1.8415e+003
```

```
>> std(X)
```

```
ans =
```

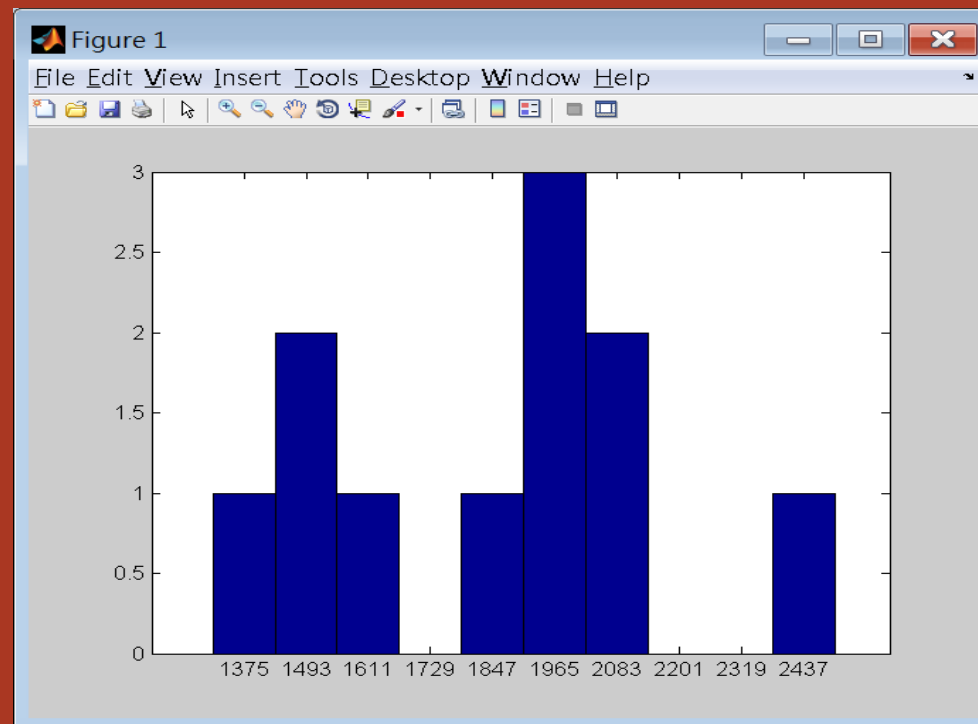
```
342.7373
```

```
>>
```

$$\bar{X} = 1841.5$$

$$s = 342.7$$

$$n = 11$$



Approximately a normal distribution.

```
>> M=mean(X);  
>> S=std(X);  
>> N=11;
```

```
>> SEM=S/sqrt(N)  
SEM = 103.3392  
>> T=tinvt(0.975, N-1)  
T = 2.2281
```

```
>> LOWER=M-T*SEM  
LOWER = 1.6112e+003  
>> UPPER=M+T*SEM  
UPPER = 2.0717e+003  
>>
```

$$\bar{x}=1841.5$$

$$s=342.7$$

$$n=11$$

$$\bar{x} \pm t \frac{s}{\sqrt{n}}$$

**Answer:**

$$CI = [1611.2 \quad 2071.7]$$

# Using fitdist and paramci

```
>> X=[1903 1935 1910 2496 2108 1961 2060 1444  
1612 1316 1511]';
```

```
>> PD=fitdist(X, 'normal');
```

```
>> CI=paramci(PD, 0.05)
```

```
CI =
```

```
1.0e+003 *
```

```
1.6112    0.2395
```

```
2.0717    0.6015
```

The same answer:

```
CI = [1611.2  2071.7]
```

```
>>
```

# Chapter Review - 1

- Explain the difference between point and interval estimation of the population mean  $\mu$ .
- What is the standard error of the mean (SEM)?
- Describe the 95% confidence interval for a population mean  $\mu$ . How is the interval interpreted?
- What are the factors that affect the length of a confidence interval for a sample mean value?
- One-sided vs two-sided CI?



# Chapter Review - 2

- Describe the similarities and differences between the  $t$  distribution and the standard normal distribution ( $z$ ).
- If you were trying to construct a CI, when would you use one rather than the other ( $z$  vs  $t$ )?